Apontamentos de

ESTATÍSTICA E DELINEAMENTO

O Modelo Linear

Jorge Cadima
Secção de Matemática (DCEB)
Instituto Superior de Agronomia
Universidade de Lisboa

2020-21

Conteúdo

Prefácio

A disciplina de Estatística e Delineamento integra-se nos programas da maioria dos Mestrados leccionados no Instituto Superior de Agronomia.

Não se trata duma disciplina introdutória de Estatística, mas sim duma disciplina de continuidade e aprofundamento de conceitos e ferramentas estatísticas. Pressupõe a frequência duma disciplina estatística prévia, semelhante às que são leccionadas na grande maioria dos primeiros ciclos nas áreas correspondentes aos cursos do ISA. Aos alunos que, por um ou outro motivo, não tenham frequentado disciplinas estatísticas introdutórias, ou que precisem de relembrar conceitos, aconselha-se vivamente a consulta de algum dos numerosos textos de introdução à Estatística. Por exemplo, podem ser consultados os materiais de apoio à disciplina de Estatística leccionada nos primeiros ciclos do ISA, disponíveis na respectiva página *web*, e que foram já editados em livro da Prof. Manuela Neves [4].

Nas disciplinas introdutórias de Estatística aborda-se sobretudo o estudo das observações univariadas, ou seja, relativas a uma única variável. Na disciplina Estatística e Delineamento estudam-se modelos que procuram explicar as observações duma dada variável à custa de outras variáveis. O fundamental do programa da disciplina Estatística e Delineamento aborda o principal modelo estatístico, o **Modelo Linear**.

Como ferramenta de apoio informático, é utilizado o **programa informático** `R` [5]. Trata-se de um *software* livre e gratuito, baseado na linguagem computacional S, especialmente concebida para aplicações estatísticas [1, 2].

Informação vária sobre o programa (manuais, respostas a perguntas frequentes, páginas de ajuda, boletim informativo) pode ser obtida através da rede, em

`http://www.r-project.org`.

O programa R pode ser descarregado gratuitamente através da Internet, a partir do *site*:

`http://cran.r-project.org`

ou em vários outros *sites* que reproduzem o conteúdo do endereço atrás referido (*mirror sites*, cujos endereços estão indicados no portal acima referido). Existem versões do programa R já compiladas para execução nos principais sistemas operativos (Linux, Macintosh, Windows).

Capítulo 1

Introdução

Nas disciplinas introdutórias de Estatística aborda-se fundamentalmente o estudo de observações *univariadas*, ou seja, relativas a uma única variável. Na disciplina Estatística e Delineamento estudam-se modelos que procuram explicar as observações duma dada variável (a *variável resposta*), à custa de outras variáveis (as *variáveis preditoras*).

1.1 Exemplo

Na nuvem de $n = 31$ pontos da Figura 1.1 pode observar-se a relação entre o *diâmetro à altura do peito* (variável DAP, em centimetros), definida como o diâmetro do tronco à altura de 1,30 m, e o *volume do tronco* (variável Volume, em metros cúbicos), medidos em 31 cerejeiras.

A nuvem de pontos da Figura 1.1 permite concluir que:

- há uma tendência de fundo relativamente forte e bem definida, o que significa que há uma *associação* entre as duas variáveis;
- essa relação é aproximadamente em linha recta, como se pode confirmar calculando o respectivo *coeficiente de correlação linear*, $r_{xy} = 0.9671$.

Nota: Parte-se do pressuposto que o conceito e as propriedades do coeficiente de correlação linear, r_{xy}, são conhecidos das disciplinas introdutórias de Estatística. Caso necessário, reveja-os.

Sempre que exista uma boa associação entre as variáveis pode procurar-se uma *equação* que descreva a *tendência de fundo* dessa relação entre as variáveis. No caso duma relação linear, a equação genérica duma linha recta (não vertical), $y = b_0 + b_1\, x$, pode ser usada para modelar a relação entre as variáveis. No exemplo acima, esse modelo permitirá estimar o volume do tronco duma cerejeira (variável resposta), a partir do seu DAP (variável preditora). A medição rigorosa do volume do tronco duma árvore é uma operação destrutiva, já que a árvore tem ser cortada e o seu tronco submergido num tanque de água, sendo o volume calculado a partir da subida do nível da água. A possibilidade de estimar o volume apenas com base na medição do DAP é vantajosa, uma vez que evita a destruição da árvore.

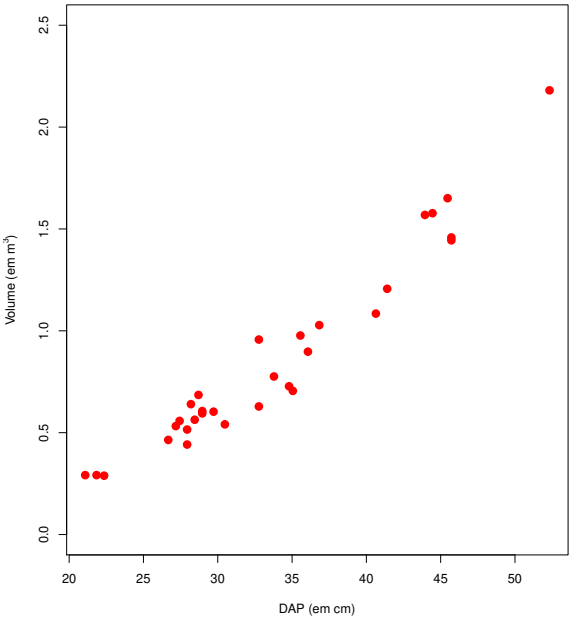


Figura 1.1: Relação entre volume do tronco (y) e DAP (diâmetro à altura do peito, x) em 31 cerejeiras. As coordenadas dos pontos são dadas pelos $n = 31$ pares de medições, $\{(x_i, y_i)\}_{i=1}^{31}$.

1.2 Ideias prévias sobre modelação em geral

Começamos por formular algumas ideias gerais sobre modelos que descrevem relações entre duas (ou mais) variáveis.

- *Todos os modelos são apenas aproximações da realidade.* No exemplo acima, é evidente que a relação entre DAP e volume do tronco não é exactamente linear. Haverá sempre erros associados às estimativas obtidas com base no modelo.
- *Existem, em geral, modelos alternativos* adequados a uma relação. Um dado modelo pode ser melhor num aspecto, mas pior noutro. No exemplo acima, os três pontos correspondentes às cerejeiras de menor porte parecem desviar-se um pouco da tendência linear. Talvez uma relação não linear (curvilínea) seja mais adequada quando se estudarem árvores de DAP pequeno. Ou pode acontecer que a introdução de novas variáveis preditoras permita diminuir acentuadamente os erros associados à previsão.
- O *princípio da parcimónia* na modelação afirma que, perante diferentes modelos considerados adequados, é preferível o *mais simples*.
- Modelos baseados em relações teóricas entre as variáveis observadas podem designar-se *modelos teóricos* ou *conceptuais*. Modelos que apenas descrevem relações observadas, mas sem recurso a relações teoricamente sustentadas, designam-se *modelos empíricos*.

1.3 Ideias prévias sobre modelos estatísticos

Nesta disciplina estudam-se um tipo particular de modelos, designados *modelos estatísticos*.

- Os modelos *estatísticos* descrevem a *tendência de fundo* entre as variáveis (que pode ser de origem teórica ou empírica). Sabe-se que existe *variabilidade* das observações em torno dessa tendência de fundo, e os modelos estatísticos incorporam essa variabilidade através de pressupostos específicos.
- Num modelo estatístico não há necessariamente uma relação de causa e efeito entre variável resposta e preditores. Há apenas *associação*. A eventual existência de uma relação de causa e efeito só pode ser justificada por argumentos teóricos exteriores à estatística.

1.3.1 Ideias prévias sobre o Modelo Linear

O *Modelo Linear* é um *caso particular* de modelação estatística. O Modelo Linear engloba um grande número de modelos específicos:

- A *Regressão Linear Simples* relaciona uma variável resposta *numérica* y e uma variável preditora igualmente *numérica* x através duma tendência de fundo linear, expressa pela equação $y = b_0 + b_1\, x$. A Regressão Linear Simples foi já motivada pelo exemplo inicial e será motivada por exemplos ulteriores na Secção 2.1.
- A *Regressão Linear Múltipla* estende a Regressão Linear Simples para relações entre uma variável resposta y e $p>1$ variáveis preditoras, x_1, x_2, ..., x_p, (todas *numéricas*), através duma equação do tipo $y = b_0 + b_1\, x_1 + b_2\, x_2 + ... + b_p\, x_p$, ou seja, escrevendo y como *combinação linear (afim)* das p variáveis preditoras.
- A *Regressão Polinomial*, em que a relação entre uma variável resposta y e uma variável preditora x é de tipo polinomial, ou seja, da forma $y = b_0 + b_1\, x + b_2\, x^2 + ... + b_p\, x^p$. Engloba igualmente modelos com equações polinomiais em várias variáveis preditoras. Como se verá, este tipo de relação pode ser estudado como um caso particular duma Regressão Linear Múltipla.
- As *Análises de Variância*, em que a variável resposta numérica y é modelada a partir de uma ou mais variáveis preditoras, que no entanto são variáveis *categóricas* (*factores*), ou seja, variáveis não numéricas, cujos valores são diferentes categorias (por exemplo, diferentes espécies, diferentes genótipos, diferentes localidades, etc.).
- As *Análises de Covariância*, em que uma variável resposta numérica y é modelada por várias variáveis, algumas das quais são numéricas e outra categóricas. Esta concretização do Modelo Linear já não faz parte do Programa da disciplina de Estatística e Delineamento.

O Modelo Linear é de estudo imprescindível, uma vez que:

- é o modelo estatístico mais frequentemente utilizado:
- é o mais completo e bem estudado tipo de modelo estatístico;

- serve de *base para numerosas extensões*, como por exemplo a Regressão Não Linear, os Modelos Lineares Generalizados, os Modelos Lineares Mistos, etc. (que não são estudados nesta disciplina).

Capítulo 2

Regressão Linear Simples

Em muitos estudos recolhem-se dados relativos a duas variáveis, havendo interesse em analisar a respectiva relação. Consideremos uma situação onde, dadas n *unidades experimentais* (por exemplo organismos, parcelas de terreno, localidades, genótipos), se observam duas variáveis *numéricas*, genericamente designadas x e y. Assim, dispõe-se de n *pares de observações* $\{(x_i, y_i)\}_{i=1}^{n}$ (correspondendo o índice i a cada unidade experimental). Trata-se de um conjunto de observações *bivariadas*.

De grande utilidade será a construção de um gráfico das n observações obtidas. Neste gráfico, cada eixo corresponde a uma das variáveis observadas, e a cada uma das n observações corresponderá um ponto, de coordenadas (x_i, y_i). Vejamos alguns exemplos.

2.1 Exemplos

2.1.1 Produção do leite de cabra em Portugal

Dados do Instituto Nacional de Estatística (INE) indicam a produção de leite de cabra em Portugal (variável y, em milhões de litros) nos anos entre 1986 e 2011 inclusive (variável x). A estes dados correspondem $n = 26$ pares de valores, $\{(x_i, y_i)\}_{i=1}^{26}$. O respectivo gráfico é mostrado na Figura 2.1.

Como se pode observar, a *tendência de fundo* é crescente e *aproximadamente linear*, com um coeficiente de correlação linear de $r_{xy} = 0.9348$. Ou seja, uma linha recta descreve bem a tendência de fundo da nuvem de pontos. De imediato coloca-se a questão de saber como identificar a melhor recta para descrever a tendência de fundo. Todas as rectas (não verticais) têm uma equação da forma $y = b_0 + b_1\, x$. O problema de determinar a melhor recta, neste contexto, será abordado posteriormente.

Neste exemplo interessa o *contexto descritivo*, ou seja, o objectivo fundamental consiste em determinar a equação da recta que melhor descreve a tendência subjacente à nuvem de pontos. A recta obtida serve para simplificar a relação entre produção de leite de cabra ao longo dos anos indicados, em Portugal, permitindo que, em vez da colecção de 26 observações bivariadas, se descreva a relação de fundo apenas à custa da equação da recta, ou seja, apenas usando os dois *parâmetros* da recta: o seu *declive* b_1 e a sua *ordenada na origem*, b_0.

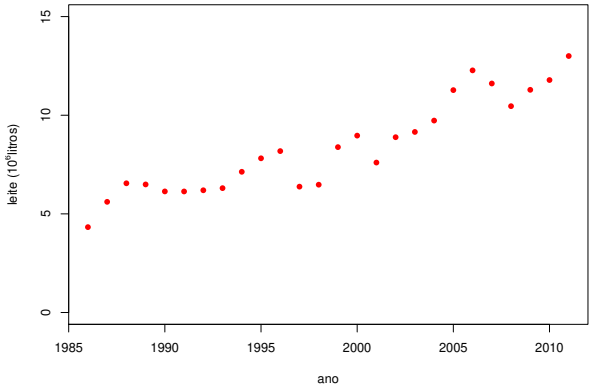


Figura 2.1: Evolução da produção do leite de cabra em Portugal, nos anos de 1986 a 2011 (dados do Instituto Nacional de Estatística, INE). O coeficiente de correlação linear é $r_{xy}=0.9348$.

2.1.2 Volume e diâmetro à altura do peito em cerejeiras

O exemplo considerado na Introdução (Capítulo 1) tem origem num conjunto de dados disponibilizado no *software* `R`, de nome `trees`. Nesse conjunto de dados existem observações de três variáveis em 31 cerejeiras (as unidades experimentais), medidas em unidades anglo-saxónicas. No entanto, utilizamos apenas as medições de duas variáveis, convertidas para o Sistema Métrico Internacional: os *diâmetros à altura do peito* (que designaremos por DAP, a variável x, convertida em centimetros) e o *volume do tronco* (variável y, em metros cúbicos) de cada cerejeira. A nuvem de pontos dos $n = 31$ pares de medições, $\{(x_i, y_i)\}_{i=1}^{31}$, é dada na Figura 1.1.

Tal como no exemplo da subsecção 2.1.1, a *tendência de fundo* é crescente e aproximadamente linear, sendo igualmente de interesse obter a melhor recta, de equação genérica $y = b_0 + b_1\, x$, para descrever a relação subjacente. No entanto, um aspecto importante distingue este exemplo do exemplo anterior. Enquanto que no exemplo da Subsecção 2.1.1 os dados disponíveis diziam respeito à totalidade da informação relativa aos 26 anos em causa, neste caso a informação disponível apenas diz respeito a um pequeno subconjunto da totalidade das árvores de cereja. Ou seja, os $n = 31$ pares de observações são apenas uma *amostra* duma *população* mais vasta. O verdadeiro objectivo, numa análise da relação entre volume do tronco e DAP, não dirá respeito apenas às 31 observações disponíveis, mas sim à relação existente na população de todas as cerejeiras.

Assim, há que admitir que existe uma *recta populacional* que descreve a relação entre volume e DAP *na população*, cuja equação será da forma $y = \beta_0 + \beta_1 x$. A recta amostral de equação $y = b_0 + b_1\, x$ obtida a partir da nossa amostra será apenas uma *estimativa* da recta populacional, mas não concidirá com a recta populacional. Diferentes amostras extraídas da população de cerejeiras irão gerar diferentes rectas estimadas. Aqui interessa o *contexto inferencial*, ou seja, saber como se pode utilizar uma amostra aleatória para, não apenas obter a recta amostral $y = b_0 + b_1\, x$, mas igualmente fazer *inferência estatística* sobre os parâmetros β_0 e β_1 da recta populacional, ou sobre os valores de y (volume) na recta populacional, dado um valor de x (DAP).

2.1.3 Peso no parto, em seres humanos

Consideremos agora um outro exemplo, em que a relação de fundo entre duas variáveis observadas *não* é de tipo linear. Os dados foram recolhidos num grande hospital português, e dizem respeito à relação entre a *idade gestacional*, ou seja, a duração duma gravidez em mulheres (variável x, em semanas) e o *peso do bebé à nascença* (variável y, em g). Há dados relativos a 251 partos, observados num dado período de tempo. Assim, dispõe-se de $n = 251$ pares de observações: $\{(x_i, y_i)\}_{i=1}^{251}$. A Figura 2.2 mostra a nuvem de pontos resultante.

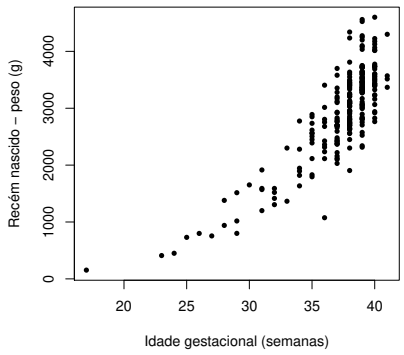


Figura 2.2: Relação entre o peso de bebés à nascença (y) e a duração da gravidez (x), em 251 partos num hospital em Portugal.

Existe uma tendência de fundo na nuvem de pontos, que é claramente crescente, mas *curvilinea*, ou seja, *não linear*. Assim, em relação aos exemplos anteriores, coloca-se uma questão adicional: saber qual a *função* que pode ser usada na equação $y = f(x)$ para descrever a relação de fundo visível na Figura 2.2. Pelo conhecimento dos gráficos de funções elementares, sabemos que uma curva crescente, como a da Figura, pode talvez estar associada a uma relação de tipo exponencial, com equação $y = f(x) = c\,\mathrm{e}^{dx}$ (com $d > 0$). Mas pode também estar associada a uma relação potência, com equação $y = f(x) = c\,x^d$, com $d > 1$ (poderia, por exemplo, tratar-se da parte ascendente duma parábola de concavidade voltada para cima). Ou talvez seja mais adequada outra forma funcional para descrever a relação entre duração da gravidez e peso dos bebés à nascença.

O ajustamento directo de relações não lineares, a chamada Regressão Não Linear, não faz parte do Programa desta disciplina. Mas, como se verá mais tarde (Secção 2.3), para muitos tipos de relações não lineares é possível identificar *transformações* de uma ou ambas as variáveis que *linearizam* a relação, ou seja, que geram uma relação linear, não entre as variveis originalmente observadas, mas entre as variáveis assim transformadas.

Em todos os exemplos anteriores, as variáveis x e y desempenham papéis que não são intercambiáveis. Em geral, haverá uma variável (associada ao eixo dos y) que se pretende modelar ou prever a partir de outra (associada ao eixo dos x). A variável que se pretende modelar costuma designar-se *variável*

resposta. A variável usada para modelar a variável resposta designa-se *variável preditora* ou *explicativa*[1].

2.2 A Regressão Linear Simples em contexto descritivo

Na disciplina de Estatística dos primeiros ciclos do ISA introduz-se o conceito de Regressão Linear, mas:

- apenas como regressão linear *simples* (uma única variável preditora); e
- apenas no contexto *descritivo* (sem preocupações inferenciais).

Vamos recordar e aprofundar o estudo da Regressão Linear Simples (RLS) em contexto meramente descritivo. Veremos como se obtém uma recta que descreva adequadamente uma relação de fundo linear evidenciada por n pares de observações de duas variáveis x e y, e estudaremos as propriedades fundamentais dessa *recta de regressão de y sobre x*.

2.2.1 O Critério dos Mínimos Quadrados

Considere de novo os exemplos das Subsecções 2.1.1 e 2.1.2. Como se pode obter uma recta $y = b_0 + b_1\, x$ que descreva bem a relação linear de fundo entre as variáveis y e x? Para justificar que uma dada recta seja 'a melhor' de todas, há que começar por definir um *critério*, que permita comparar diferentes rectas.

O critério clássico, usado na regressão linear, é o chamado *critério dos mínimos quadrados*. Para aplicar esse critério, começa-se por definir o conceito de *resíduo*. Um resíduo é a diferença entre um valor observado da variável resposta y e o correspondente valor $\hat{y}$ obtido através de um dado modelo, que se designa o *valor ajustado de y*.

Definição 2.1: Resíduos numa RLS

Seja dado um conjunto de n observações bivariadas, $\{(x_i, y_i)\}_{i=1}^{n}$, e uma qualquer recta (não vertical) relacionando y e x, de equação $y = b_0 + b_1\, x$. Designa-se por i-ésimo **resíduo** a diferença entre o i-ésimo valor observado de y, y_i, e o valor ajustado de y associado ao correspondente valor x_i do preditor, ou seja: $\hat{y}_i = b_0 + b_1\, x_i$. Assim, o resíduo associado à observação i é dado por:

$$e_i \quad = \quad y_i - \hat{y}_i \quad = \quad y_i - (b_0 + b_1 x_i)\ , \tag{2.1}$$

Nos gráficos das Figuras 2.1 e 1.1, um resíduo corresponde à *distância na vertical* entre cada ponto e a recta $y = b_0 + b_1\, x$, distância essa afectada de um *sinal* que será:

[1] A variável resposta é por vezes designada *variável dependente* e a variável preditora é designada *variável independente*. No entanto, a palavra 'independente' confunde com o conceito de independência probabilística de duas variáveis aleatórias, que nada tem a que ver com o contexto agora referido. A fim de evitar esta confusão, serão utilizadas nestes apontamentos as expressões *variável resposta* e *variável preditora*.

- positivo se o ponto se encontra acima da recta;
- negativo para pontos abaixo da recta.

O critério para definir a recta de regressão de y sobre x é o critério de minimizar a soma de quadrados dos resíduos.

Definição 2.2: Recta de Mínimos Quadrados numa RLS

Seja dado um conjunto de n observações bivariadas, $\{(x_i, y_i)\}_{i=1}^n$. A **recta de regressão de y sobre x** é a recta $y = b_0 + b_1\, x$ que minimiza a Soma de Quadrados Residual (i.e., dos resíduos), ou seja, cujos parâmetros b_0 e b_1 minimizam:

$$SQRE \;=\; \sum_{i=1}^{n} e_i^2 \;=\; \sum_{i=1}^{n} \left[y_i - (b_0 + b_1 x_i)\right]^2 \,.$$

A solução do problema agora definido dá origem a fórmulas simples para os dois parâmetros da recta: o seu *declive* b_1 e a *ordenada na origem* b_0. Essas fórmulas são dadas na seguinte Proposição.

Proposição 2.1: Fórmulas da recta de regressão

Dados n pares de observações $\{(x_i, y_i)\}_{i=1}^n$, a *recta de regressão de y sobre x* obtida a partir do Critério de Mínimos Quadrados, é a recta $y = b_0 + b_1\, x$, com

$$\text{Declive :} \qquad b_1 \;=\; \frac{cov_{xy}}{s_x^2} \tag{2.2}$$

$$\text{Ordenada na origem :} \qquad b_0 \;=\; \overline{y} - b_1 \overline{x}\ , \tag{2.3}$$

sendo:

- $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ a média das n observações de x;
- $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ a média das n observações de y;
- $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2$ a variância amostral de x; e
- $cov_{xy} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$ a covariância amostral entre x e y.

Demonstração 2.1: Proposição 2.1

Determinar valores b_0 e b_1 que minimizem $SQRE$ é um problema de minimizar uma função (aqui chamada $SQRE$) de duas variáveis (aqui chamadas b_0 e b_1). Este problema é estudado nas disciplinas de Análise

Matemática. Dada uma função de duas variáveis, $f(x,y)$, derivável em todo o seu domínio, é **condição necessária** para que a função atinja um extremo num ponto (x^*, y^*) que as duas *derivadas parciais* de f se anulem nesse ponto:

$$\frac{\partial f(x^*,y^*)}{\partial x} = 0 \quad ; \quad \frac{\partial f(x^*,y^*)}{\partial y} = 0$$

Tendo em conta a função $SQRE = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$, com variáveis b_0 e b_1, tem-se:

$$\frac{\partial SQRE(b_0,b_1)}{\partial b_0} = 2\sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)] \cdot (-1) \tag{2.4}$$

$$\frac{\partial SQRE(b_0,b_1)}{\partial b_1} = 2\sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)] \cdot (-x_i) \tag{2.5}$$

Igualando a primeira expressão a zero e dividindo por n, tem-se:

$$\begin{aligned} \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)] = 0 \quad &\Leftrightarrow \quad \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} (b_0 + b_1 x_i) \quad \Leftrightarrow \quad \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\not{n} b_0}{\not{n}} + b_1 \frac{\sum_{i=1}^{n} x_i}{n} \\ &\Leftrightarrow \quad \overline{y} = b_0 + b_1 \overline{x} \quad \Leftrightarrow \quad b_0 = \overline{y} - b_1 \overline{x} \ , \end{aligned}$$

o que prova a fórmula (2.3). Por seu lado, igualando a expressão (2.5) a zero e substituindo a fórmula acabada de obter para b_0, tem-se:

$$\begin{aligned} \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]\, x_i = 0 \quad &\Leftrightarrow \quad \sum_{i=1}^{n} y_i\, x_i = \sum_{i=1}^{n} (b_0 + b_1 x_i)\, x_i \\ &\Leftrightarrow \quad \sum_{i=1}^{n} y_i\, x_i = \sum_{i=1}^{n} [(\overline{y} - b_1\, \overline{x}) + b_1 x_i]\, x_i \\ &\Leftrightarrow \quad \sum_{i=1}^{n} y_i\, x_i = \sum_{i=1}^{n} \overline{y}\, x_i + b_1 \sum_{i=1}^{n} (x_i - \overline{x}) x_i \\ &\Leftrightarrow \quad b_1 \sum_{i=1}^{n} (x_i - \overline{x}) x_i = \sum_{i=1}^{n} (y_i - \overline{y}) x_i \ . \end{aligned} \tag{2.6}$$

Ora, pela definição de covariância amostral, tem-se

$$\begin{aligned} {}_{(n-1)}\, cov_{xy} &= \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i\, (y_i - \overline{y}) - \sum_{i=1}^{n} \overline{x}\, (y_i - \overline{y}) \\ &= \sum_{i=1}^{n} x_i\, (y_i - \overline{y}) - \underbrace{\overline{x} \sum_{i=1}^{n} (y_i - \overline{y})}_{=0} = \sum_{i=1}^{n} x_i\, (y_i - \overline{y}) \ , \end{aligned} \tag{2.7}$$

já que

$$\sum_{1=1}^{n} (y_i - \overline{y}) = \sum_{1=1}^{n} y_i - n\overline{y} = n\overline{y} - n\overline{y} = 0 \ . \tag{2.8}$$

Por contas análogas às de (2.8) mostra-se que $\sum_{i=1}^{n} (x_i - \overline{x}) = 0$ (ver também o Exercício 3 de RLS). Logo, a

partir da definição de variância amostral, tem-se:

$$\begin{aligned} {}_{(n-1)}\, s_x^2 &= \sum_{i=1}^{n}(x_i-\overline{x})^2 = \sum_{i=1}^{n}(x_i-\overline{x})(x_i-\overline{x}) = \sum_{i=1}^{n} x_i\,(x_i-\overline{x}) - \sum_{i=1}^{n}\overline{x}\,(x_i-\overline{x}) \\ &= \sum_{i=1}^{n} x_i\,(x_i-\overline{x}) - \overline{x}\underbrace{\sum_{i=1}^{n}(x_i-\overline{x})}_{=0} = \sum_{i=1}^{n} x_i\,(x_i-\overline{x})\ , \end{aligned} \tag{2.9}$$

Substituindo as expressões (2.7) e (2.9) na equação (2.6), tem-se:

$$b_1 \sum_{i=1}^{n}(x_i-\overline{x})\,x_i = \sum_{i=1}^{n}(y_i-\overline{y})x_i \quad \Leftrightarrow \quad b_1\ {}_{(n-1)}\, s_x^2 = {}_{(n-1)}\, cov_{xy} \quad \Leftrightarrow \quad b_1 = \frac{\cancel{(n-1)}\, cov_{xy}}{\cancel{(n-1)}\, s_x^2}\ ,$$

que é a fórmula (2.2). Dispensa-se a verificação da **condição suficiente** para a existência de um mínimo local (através do Critério da matriz Hessiana) no ponto crítico (b_0, b_1) agora identificado, uma vez que existe um único ponto crítico, e a natureza do problema indica que tem de ser um mínimo (isto é, tem de existir uma recta globalmente mais próxima dos n pontos).

Importa sublinhar que critérios de ajustamento diferentes produziriam rectas ajustadas diferentes. Não é difícil conceber critérios de ajustamento da recta alternativos ao critério de mínimos quadrados da Definição 2.2. Por exemplo, em vez de se procurar minimizar a soma de quadrados de distâncias na vertical, é possível usar o critério de minimizar a soma de quadrados de distâncias *na perpendicular*, ou mesmo *na horizontal*, entre pontos e recta. Da mesma forma, em vez de considerar somas de quadrados, é possível considerar somas das distâncias (valor absoluto dos resíduos), ou outros critérios[2]. Cada critério alternativo gera uma recta específica, e em geral as rectas resultantes são diferentes.

A escolha do critério de minimizar a Soma de Quadrados dos Resíduos tem, subjacente, o papel diferente das duas variáveis, sendo:

- y a *variável resposta*, que *se deseja modelar*;
- x a *variável preditora*, uma ferramenta usada na modelação de y.

O objectivo é prever y com erros globalmente o mais pequenos possível. O i-ésimo resíduo, $e_i = y_i - \hat{y}_i$, mede precisamente o desvio (com sinal) da observação y_i face à sua previsão a partir duma recta ajustada. Ao minimizar-se a Soma de Quadrados Residual, *minimiza-se a soma de quadrados dos erros de previsão* de y.

Uma vez estabelecido o critério de mínimos quadrados da Definição 2.2, *o papel das duas variáveis, x e y, não é simétrico*. Ou seja, a recta de regressão de y sobre x não é igual à recta de regressão de x sobre y. Importa, pois, considerar antecipadamente qual a variável que se deseja estudar (ou seja, a variável resposta y) e qual a variável que se deseja utilizar como ferramenta para modelar/prever y (ou seja, a variável preditora x).

[2] O critério de minimizar a soma de resíduos, afectados de sinal, não seria um bom critério: havendo resíduos com sinal diferente, resíduos positivos grandes podem cancelar resíduos negativos de grande valor absoluto, sem que isso signifique que a recta usada para calcular os resíduos esteja próxima desses pontos. Aliás, como se verá seguidamente, na recta de regressão de mínimos quadrados, a soma dos resíduos é sempre nula.

Na Figura 2.3 é mostrada a recta de regressão de volume do tronco sobre DAP, para a amostra das $n = 31$ cerejeiras discutida na Subsecção 2.1.2. Repare-se como a escolha de variável resposta e preditora é aqui evidente: o objectivo é prever volumes do tronco (cuja medição rigorosa envolve técnicas destrutivas) a partir dos DAP, e não o inverso. O declive e ordenada na origem da recta aí mostrada são dados pelas fórmulas da Proposição 2.1.

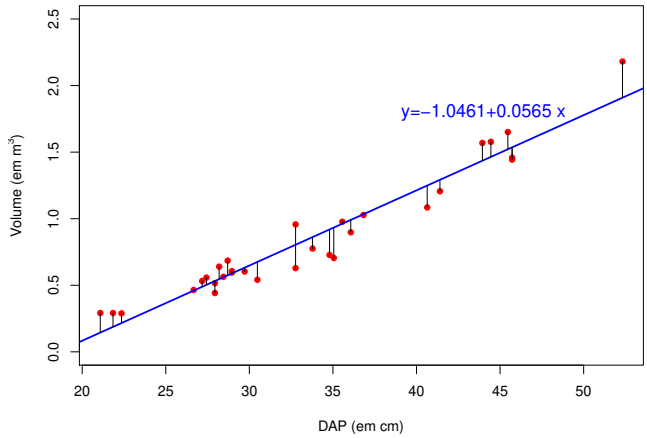


Figura 2.3: A recta de regressão entre volume do tronco (y) e diâmetro à altura do peito (x) em cerejeiras, ajustada a partir da amostra de $n = 31$ cerejeiras introduzida na Secção 1.1. Esta recta resultou de minimizar a Soma de Quadrados dos Resíduos, ou seja, minimizar a soma de quadrados das distâncias na vertical entre pontos e recta, assinaladas no gráfico.

2.2.2 Propriedades da recta de regressão

É sabido que, para qualquer recta de equação $y = b_0 + b_1\, x$,

- a *ordenada na origem* b_0 é o valor de y (na recta) associado a $x = 0$;
- o *declive* b_1 é *a variação de* y *associada a um aumento de uma unidade em* x.

No contexto duma recta de regressão de y sobre x, que apenas descreve uma tendência de fundo, estas interpretações devem ser acompanhadas do qualificativo *médio* (por exemplo, o declive será a variação *média* em y, associada a um aumento de x em uma unidade).

Estes parâmetros da recta de regressão têm *unidades de medida*, uma vez que só há coerência de unidades de medida na equação caso:

- b_0 tenha *unidades de medida iguais às de* y; e
- o declive b_1 tenha *unidades de medida iguais a* $\frac{\textit{unidades de } y}{\textit{unidades de } x}$.

Assim, a recta de regressão indicada na Figura 2.3 permite afirmar que, a cada cm adicional no DAP, o volume do tronco aumenta, em média, 0.0565 m^3. Neste contexto, é biologicamente irrelevante interpretar b_0, uma vez que não existem árvores com DAP igual a zero. Em tais casos, a ordenada na origem b_0 deve ser vista apenas como um parâmetro que permite um melhor ajustamento da recta à nuvem de pontos.

Proposição 2.2: Propriedades da recta de regressão

Seja dada uma recta de regressão de y sobre x, ajustada com base em n observações $\{(x_i, y_i)\}_{i=1}^n$. Verifica-se:

1. a recta de regressão passa no centro de gravidade da nuvem de pontos, isto é, no ponto $(\overline{x}, \overline{y})$.
2. a média dos valores observados y_i, é igual à média dos correspondentes valores ajustados, $\hat{y}_i$.
3. a soma dos resíduos e_i é igual a zero.

Demonstração 2.2: Proposição 2.2

1. É evidente a partir da fórmula para a ordenada na origem que o ponto $(\overline{x}, \overline{y})$ satisfaz a equação da recta:
$$b_0 = \overline{y} - b_1\,\overline{x} \qquad \Leftrightarrow \qquad \overline{y} = b_0 + b_1\,\overline{x}\ .$$
2. A média dos valores ajustados, $\hat{y}_i$, é dada por $\overline{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \frac{1}{n}\sum_{i=1}^{n}(b_0 + b_1\,x_i)$. Substituindo a fórmula para b_0 (2.3), vem:
$$\overline{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}\left[(\overline{y} - b_1\overline{x}) + b_1\,x_i\right] = (\overline{y} - b_1\overline{x}) + b_1\overline{x} = \overline{y}\ .$$
3. A soma dos resíduos, tendo em conta o resultado do ponto anterior, é dada por:
$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(y_i - \hat{y}_i) = n\,\overline{y} - n\,\overline{\hat{y}} = 0\ .$$

2.2.3 As três Somas de Quadrados

Já foi referido o papel da Soma de Quadrados Residual ($SQRE$) no critério que leva à definição da recta de regressão de y sobre x. Mas duas outras somas de quadrados desempenham um papel fulcral no estudo da regressão linear simples. Todas estão intimamente associadas à variância amostral de três tipos de valores: os valores observados de y, os valores ajustados de y, e os resíduos.

Sejam dadas n observações bivariadas, $\{(x_i, y_i)\}_{i=1}^n$ e a respectiva recta de regressão de y sobre x. Considerem-se as variâncias amostrais

- dos n valores observados de y: $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$;
- dos n valores ajustados de y: $s_{\hat{y}}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$; e
- dos n resíduos: $s_e^2 = \frac{1}{n-1}\sum_{i=1}^{n}(e_i - \overline{e})^2 = \frac{1}{n-1}\sum_{i=1}^{n} e_i^2$.

Nota: Na definição da variância amostral dos valores ajustados $\hat{y}$ foi tido em conta que a média desses valores é igual à média dos n valores observados de y, $\overline{y}$. De igual forma, foi tido em conta na definição da variância amostral dos resíduos, s_e^2, que a média dos resíduos é (tal como a sua soma) nula. Ambos estes resultados foram demonstrados na Proposição 2.2.

Definição 2.3: As três Somas de Quadrados

Sejam dadas n observações bivariadas, $\{(x_i, y_i)\}_{i=1}^n$ e a respectiva recta de regressão de y sobre x. Definem-se as seguintes Somas de Quadrados (repetindo-se a definição de $SQRE$):

- a **Soma de Quadrados Total,** $SQT = \sum_{i=1}^{n}(y_i - \overline{y})^2 = {}_{(n-1)}\, s_y^2$;
- a **Soma de Quadrados associada à Regressão,** $SQR = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = {}_{(n-1)}\, s_{\hat{y}}^2$;
- a **Soma de Quadrados Residual,** $SQRE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = {}_{(n-1)}\, s_e^2$.

Uma fórmula fundamental da regressão linear relaciona estas três Somas de Quadrados e, portanto, as três variâncias amostrais que lhes estão associadas. Essa fórmula fundamental afirma que a SQT se decompõe na soma de SQR e $SQRE$ (justificando assim a designação de Soma de Quadrados *Total*).

Proposição 2.3: Fórmula Fundamental da Regressão

Sejam dadas n observações bivariadas,$\{(x_i, y_i)\}_{i=1}^n$, e a respectiva recta de regressão de y sobre x. Verifica-se a seguinte relação entre as três Somas de Quadrados:

$$SQT = SQR + SQRE \quad \Leftrightarrow \quad s_y^2 = s_{\hat{y}}^2 + s_e^2$$

Demonstração 2.3: Proposição 2.3

Na expressão que define SQT vamos introduzir um par de parcelas de soma zero, que nos ajudarão

nas contas subsequentes ($-\hat{y}_i + \hat{y}_i = 0$):

$$\begin{aligned} SQT &= \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} [(y_i - \hat{y}_i) + (\hat{y}_i - \overline{y})]^2 \\ &= \sum_{i=1}^{n} \left[(y_i - \hat{y}_i)^2 + (\hat{y}_i - \overline{y})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \overline{y})\right] \\ &= \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}_{=SQRE} + \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}_{=SQR} + 2\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) \end{aligned} \tag{2.10}$$

Para que a igualdade pedida se verifique, é preciso que a última parcela na expressão (2.10) seja nula. Ora $\hat{y}_i = b_0 + b_1\, x_i$ e $b_0 = \overline{y} - b_1\, \overline{x}$. Substituindo a expressão de b_0 na definição de $\hat{y}_i$, vem: $\hat{y}_i = \overline{y} + b_1(x_i - \overline{x})$. Logo, o somatório na última parcela da equação (2.10) pode ser re-escrito como:

$$\begin{aligned} \sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) &= \sum_{i=1}^{n} \left[(y_i - \overline{y}) - b_1\,(x_i - \overline{x})\right] b_1\,(x_i - \overline{x}) \\ &= b_1 \Bigg[\underbrace{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}_{=(n-1)\,cov_{xy}} - b_1 \underbrace{\sum_{i=1}^{n} (x_i - \overline{x})^2}_{=(n-1)\,s_x^2} \Bigg] \end{aligned}$$

Tendo em conta que $b_1 = \frac{cov_{xy}}{s_x^2}$, tem-se $b_1\, s_x^2 = cov_{xy}$. Logo, a diferença acima anula-se.

A Proposição agora enunciada decompõe a Soma de Quadrados Total (e portanto a variância amostral dos valores observados de y) em duas parcelas com significado. Uma dessas parcelas, $SQRE$, está associada aos resíduos - ou seja, ao desvio dos pontos observados em relação à recta ajustada - e portanto corresponde à variabilidade da variável resposta y que a recta de regressão *não é capaz de explicar*. A outra parcela, SQR, corresponde à variabilidade resultante de substituir os valores observados de y pelos correspondentes valores ajustados pela recta - ou seja, corresponde à parte da variabilidade dos y_i que é preservada se esses valores observados forem substituídos pelos valores previstos pelo modelo, os valores $\hat{y}_i$. Pode considerar-se que se trata da parte da variabilidade dos y_i observados que é *preservada, ou explicada, pela recta de regressão.*

É evidente que uma recta de regressão é tanto mais eficaz no relacionamento entre duas variáveis, quanto maior fôr o valor de SQR em relação ao valor de $SQRE$ ou, alternativamente, quanto maior fôr o valor de SQR em relação ao total SQT. Surge assim, de forma natural, a definição daquele que é o *mais usado indicador da qualidade duma recta de regressão*, o chamado Coeficiente de Determinação, R^2:

Definição 2.4: Coeficiente de Determinação

Sejam dadas n observações bivariadas, $\{(x_i, y_i)\}_{i=1}^{n}$, e a respectiva recta de regressão de y sobre x.

Define-se o **Coeficiente de Determinação** R^2 associado à regressão, como sendo a razão:

$$R^2 = \frac{SQR}{SQT} = \frac{s_{\hat{y}}^2}{s_y^2} \qquad (s_y^2 \neq 0) \, .$$

A discussão sobre as Somas de Quadrados feita mais acima torna evidente que *uma regressão linear deve considerar-se tanto melhor quanto maior fôr o valor do Coeficiente de Determinação* R^2 que lhe está associado. Tornemos mais explícitas as propriedades deste indicador fundamental da qualidade duma recta de regressão.

Proposição 2.4: Propriedades do Coeficiente de Determinação R^2

Seja $R^2 = \frac{SQR}{SQT}$ o Coeficiente de Determinação duma recta de regressão. Verificam-se as seguintes propriedades:

1. Para qualquer conjunto de dados, $0 \leq R^2 \leq 1$;
2. R^2 mede a proporção da variabilidade total da variável resposta y que é explicada pela regressão;
3. $R^2 = 1$ se e só se os pontos correspondentes às n observações forem exactamente colineares, ou seja, estiverem todos em cima da recta de regressão;
4. $R^2 = 0$ se e só se a recta de regressão fôr horizontal, ou seja, se e só se o seu declive fôr $b_1 = 0$, o que equivale a dizer que a covariância amostral entre x e y é nula;
5. **Numa regressão linear simples**, R^2 é o quadrado do coeficiente de correlação linear entre preditor e variável resposta:

$$R^2 = r_{xy}^2 = \left(\frac{cov_{xy}}{s_x \, s_y} \right)^2 \qquad (s_y \neq 0 \, , \, s_x \neq 0) \, .$$

Demonstração 2.4: Proposição 3.4

1. Por definição, somas de quadrados não podem ser negativas. Logo $R^2 = \frac{SQR}{SQT} \geq 0$ e, pela Proposição 2.3, tem de ter-se $SQT \geq SQR$. Assim, $R^2 \leq 1$.
2. A interpretação resulta directamente da segunda expressão para R^2 na Definição 2.4: $R^2 = \frac{s_{\hat{y}}^2}{s_y^2}$.
3. $R^2 = 1$ se e só se $SQR = SQT$, ou seja, se e só se $SQRE = 0$. Ora, **qualquer soma de quadrados só pode ser nula se todas as parcelas forem nulas** (não existindo parcelas negativas, bastava que houvesse uma parcela estritamente positiva para que a soma fosse

estritamente positiva)[a]. Logo, a Soma de Quadrados dos Resíduos só pode ser nula se todos os resíduos e_i forem nulos. Pela definição de resíduo, tem então de verificar-se $y_i = \hat{y}_i$, para todo o i. Isso significa que todos os pontos estão em cima da recta de regressão.

4. $R^2 = 0$ se e só se $SQR = 0$. De novo, uma soma de quadrados é nula se e só se todas as suas parcelas forem nulas. No nosso contexto, isso significa que $\hat{y}_i = \overline{y}$ para todas as observações. Mas só é possível que todos os valores ajustados de y sejam iguais se a recta de regressão (que define esse valores ajustados de y) fôr horizontal, ou seja, tenha declive zero ($b_1 = 0$). Como o declive da recta de regressão é dado pela fórmula $b_1 = \frac{cov_{xy}}{s_x^2}$ (Proposição 2.1), o declive só pode ser nulo caso o seu numerador (cov_{xy}) seja nulo.

5. Veja-se a resolução do Exercício RLS 6.

[a] Este resultado é geral para qualquer soma de números não negativos, não dependendo do nosso contexto específico da regressão linear. Para qualquer colecção de n números reais x_i, tem-se que $\sum_{i=1}^{n} x_i^2 = 0 \Rightarrow x_i = 0$, para todo o i.

2.2.4 Regressão - um pouco de história

O critério de mínimos quadrados surge do trabalho de francês Legendre, no início do Século XIX. Assentou em trabalho anterior, motivado pelo problema de conciliar diferentes observações astronómicas e geodésicas, que se sabia estarem afectadas por erros de observação, de forma a procurar identificar a relação subjacente entre as quantidades observadas.

A designação *Regressão* tem origem num estudo posterior do inglês Francis Galton (1886), relacionando a estatura de $n = 928$ jovens adultos com a estatura (média) dos pais [3]. Galton era entusiástico defensor da *eugenia*, uma concepção então respeitável. No seu estudo, constatou que pais com alturas acima da média tinham tendência a ter filhos com altura acima da média - mas menor que os pais - enquanto pais com estatura abaixo da média tinham tendência a ter filhos com estatura abaixo da média, mas maior que os pais. Chamou ao seu artigo *Regression towards mediocrity in hereditary stature*. A expressão *regressão* ficou associada ao método devido a este acaso histórico.

Os dados de Galton estão disponíveis num objecto de nome `Galton`, disponibilizado no módulo (*package*) do `R`, chamado `HistData`. A nuvem de pontos e recta de regressão correspondentes são mostrados na Figura 2.4. O declive da recta ajustada, $b_1 = 0.65$, permite a interpretação de que, em média, a cada polegada adicional na altura média dos pais, corresponde uma altura adicional de apenas 0.65 polegadas na altura média dos filhos.

Curiosamente o exemplo de Galton tem um valor muito baixo do Coeficiente de Determinação, explicando pouco mais de 20% da variabilidade observada nas alturas médias dos filhos.

2.3 Transformações linearizantes

Após considerar as propriedades fundamentais das regressões lineares simples, em contexto descritivo, vejamos agora como a RLS pode ser útil mesmo no estudo de relações *não* lineares entre duas variáveis, como a mostrada no exemplo da Subsecção 2.1.3.

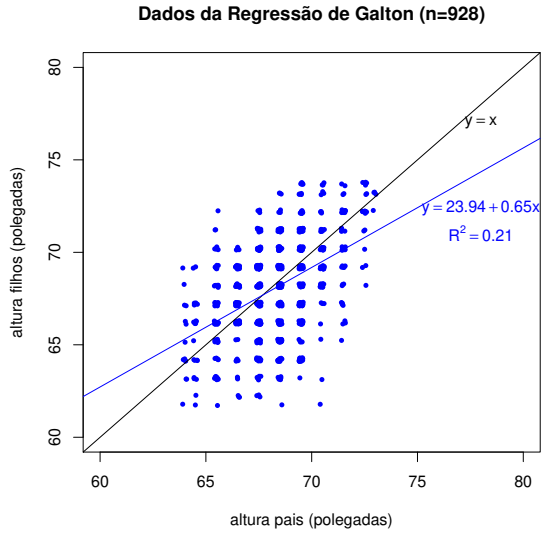


Figura 2.4: Os dados do estudo de Galton que deu origem ao nome *regressão*, com a respectiva recta de regressão e o valor (muito baixo) do seu Coeficiente de Determinação, bem como a bissectriz $y = x$. Galton arredondou os seus dados à unidade (polegadas). A fim de evidenciar que há mais observações na zona central, do que na periferia da nuvem de pontos, foi usada a função `jitter` ao produzir o gráfico, função essa que introduz pequenas perturbações nos valores observados.

Em alguns casos, felizmente frequentes embora não universais, uma relação de fundo não linear entre y e x pode ser linearizada caso se proceda a transformações adequadas numa, ou em ambas, as variáveis. Sempre que possíveis, estas *transformações linearizantes* permitem utilizar uma regressão linear simples, apesar de a relação original não ser linear.

Estudamos de seguida alguns exemplos particularmente frequentes de relações não-lineares que são linearizáveis através de transformações da variável resposta e, em alguns casos, também do preditor.

2.3.1 Relação exponencial

Definição 2.5

Uma *curva exponencial* é uma curva de equação

$$y = c\,\mathrm{e}^{d\,x}\ , \quad (\text{com } y > 0 \quad ; \quad c > 0) \tag{2.11}$$

cujo gráfico é da forma indicada na Figura 2.5.

A logaritmização de y lineariza uma relação exponencial, ou seja, uma relação de tipo exponencial torna-se

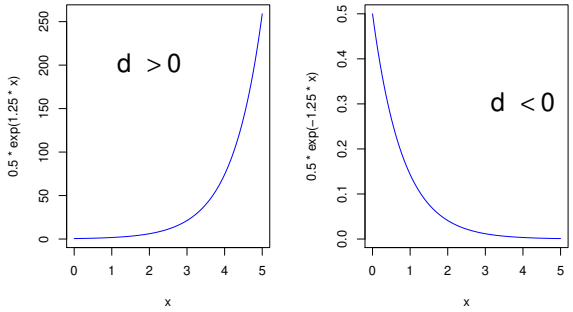


Figura 2.5: Curvas exponenciais, de equação $y = ce^{dx}$, relacionando duas variáveis, y e x. À esquerda, uma *exponencial crescente*, associada a valores positivos do parâmetro d. À direita, uma *exponencial decrescente*, associada a valores negativos de d.

uma *relação linear entre o logaritmo de* y *e* x. De facto, tomando-se logaritmos (naturais[3]), obtém-se:

$$\begin{aligned} y = ce^{dx} \quad &\Leftrightarrow \quad \ln(y) \;=\; \ln(c) + \ln(e^{dx}) \;=\; \ln(c) \;+\; d\,x \\ &\Leftrightarrow \quad\; y^* \;=\; b_0 \;+\; b_1\,x\;, \end{aligned}$$

com $y^* = \ln(y)$, $b_0 = \ln(c)$ e $b_1 = d$. Ou seja, trata-se duma *relação linear entre* $y^* = \ln(y)$ *e* x.

O *sinal do declive da recta* indica se a relação exponencial original é *crescente (quando* $b_1 > 0$*)* ou *decrescente (*$b_1 < 0$*)*.

Ilustremos a aplicação desta ideia considerando o exemplo do peso de bebés à nascença, cuja nuvem de pontos original foi dada na Figura 2.2. O gráfico de *log-pesos* dos recém-nascidos contra idade gestacional produz uma *relação de fundo linear*, como se pode observar na Figura 2.6.

Nota 2.1

Normalmente não somos capazes de identificar, perante uma curvatura como a observada na Figura 2.2, se a relação exponencial é adequada. O facto da logaritmização de y ter linearizado a relação significa que *a relação original (peso vs. idade gestacional) pode ser considerada exponencial.* Trata-se duma constatação importante, de valor geral: para saber se $y = ce^{dx}$ é uma relação adequada, construa-se o gráfico de $\ln(y)$ *vs.* x e verifique-se se a relação é aproximadamente linear.

Uma relação exponencial (2.11) é a solução duma Equação Diferencial envolvendo as variáveis y e x, como se verá na Proposição 2.5.

[3] Salvo indicação em contrário, os logaritmos usados nesta disciplina são *naturais* ou *Neperianos*, ou seja, de base e.

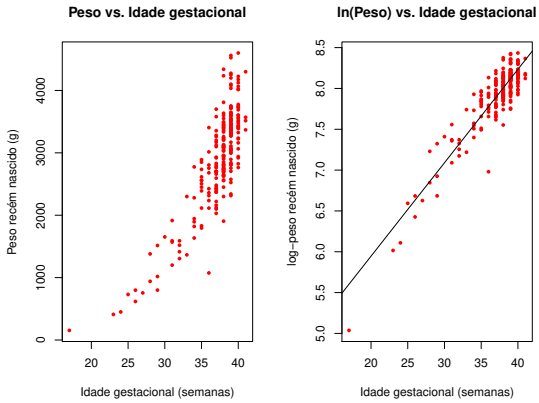


Figura 2.6: À esquerda, a relação original entre peso do bebé e duração da gravidez. À direita, a relação linearizada, com a nuvem de pontos dos log-pesos dos bebés *vs.* duração da gravidez, bem como a recta de regressão ajustada após a transformação.

Proposição 2.5: Equação Diferencial duma relação exponencial

Uma relação exponencial entre y e x resulta de admitir que y é função de x e que a *taxa de variação de* y, ou seja, a *derivada* $y'(x)$, *é proporcional a* y:

$$y'(x) = d \cdot y(x) \ . \tag{2.12}$$

De forma equivalente, pode dizer-se que a *taxa de variação* ***relativa*** *de* y, ou seja, a razão entre a derivada $y'(x)$ e $y(x)$, ***é constante***:

$$\frac{y'(x)}{y(x)} = d \ . \tag{2.13}$$

Aplicada ao exemplo acima, esta equação diferencial diz-nos que a taxa de variação do peso do bebé varia na razão directa do peso do bebé, ou seja, e de forma equivalente, que a taxa de variação relativa do peso do bebé é constante.

Demonstração 2.5: Proposição 2.5

Primitivando os dois membros da equação (2.13) em ordem a x, tem-se:

$$\ln(y(x)) \ = \ d\,x + K \qquad \Leftrightarrow \qquad y(x) \ = \ \underbrace{\mathrm{e}^{K}}_{=c} \, \mathrm{e}^{d\,x} \ = \ c\,\mathrm{e}^{d\,x}.$$

A constante de primitivação K é, no nosso contexto, a ordenada na origem da recta de regressão da relação linearizada: $K = b_0 = \ln(c)$. *O declive* $b_1 = d$ *da recta é o valor (constante) da taxa de variação relativa de* y.

Modelo exponencial de crescimento populacional

Um modelo exponencial é frequentemente usado para descrever o *crescimento de populações*, numa fase inicial onde não se faz ainda sentir a escassez de recursos limitantes.

Mas nenhum crescimento populacional exponencial é sustentável a longo prazo, ou seja, a hipótese de que a taxa de variação relativa da dimensão duma população seja constante (equação 2.13) é, a prazo, irrealista. Historicamente, a crítica a essa hipótese gerou modelos de crescimento populacional alternativos.

2.3.2 Relação logística (com 2 parâmetros)

Em 1838, Verhulst propôs uma *modelo de crescimento populacional alternativo* ao modelo exponencial, prevendo os efeitos resultantes da limitação de recursos: o *modelo logístico*. Considera-se aqui uma versão simplificada (com 2 parâmetros) dessa relação.

Definição 2.6

Seja y a variável que *mede a dimensão duma população, relativa a um máximo possível*, sendo assim uma proporção, ou seja $y \in]0,1[$. A curva de equação

$$y = \frac{1}{1+\mathrm{e}^{-(c+d\,x)}} \tag{2.14}$$

chama-se uma *curva logística*. A curva é dada na Figura 2.7, para o caso de $d > 0$ (em cujo caso se designa uma *logística crescente*).

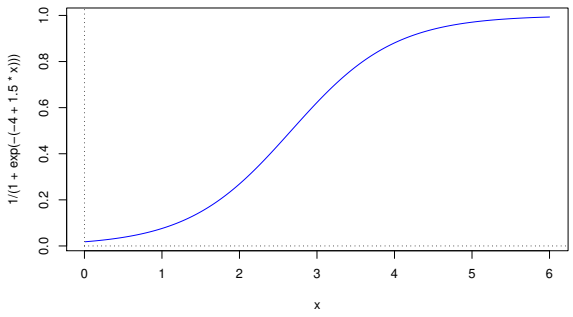


Figura 2.7: Uma curva logística crescente, gráfico da função (2.14), $y = \frac{1}{1+\mathrm{e}^{-(c+d\,x)}}$, quando $d > 0$.

A relação logística por ser linearizada através duma transformação *logit* de y, i.e., da transformação:

$$y^* = \ln\left(\frac{y}{1-y}\right) \ . \tag{2.15}$$

De facto, a partir da equação 2.14 tem-se:

$$
\begin{aligned}
y = \frac{1}{1+\mathrm{e}^{-(c+d\,x)}} \quad &\Leftrightarrow \quad 1-y = 1-\frac{1}{1+\mathrm{e}^{-(c+d\,x)}} = \frac{\cancel{1}+\mathrm{e}^{-(c+d\,x)}-\cancel{1}}{1+\mathrm{e}^{-(c+d\,x)}} = \frac{\mathrm{e}^{-(c+d\,x)}}{1+\mathrm{e}^{-(c+d\,x)}} \\
&\Leftrightarrow \quad \frac{y}{1-y} = \frac{\frac{1}{\cancel{1+\mathrm{e}^{-(c+d\,x)}}}}{\frac{\mathrm{e}^{-(c+d\,x)}}{\cancel{1+\mathrm{e}^{-(c+d\,x)}}}} = \frac{1}{\mathrm{e}^{-(c+d\,x)}} = \mathrm{e}^{c+d\,x} \\
&\Leftrightarrow \quad \underbrace{\ln\left(\frac{y}{1-y}\right)}_{=y^*} = c + d\,x\,,
\end{aligned}
$$

que é uma relação linear entre o *logit* de y e x, com ordenada na origem $b_0 = c$ e declive $b_1 = d$.

Proposição 2.6: Equação Diferencial duma relação logística

A relação logística resulta de admitir que y é função de x e que a *taxa de variação relativa de y diminui linearmente com o aumento de y*, através da equação:

$$\frac{y'(x)}{y(x)} = d \cdot [1 - y(x)]\,. \tag{2.16}$$

Para populações pequenas ($y(x) \approx 0$), a equação diferencial acima é muito próxima da equação diferencial (2.12), pelo que o crescimento populacional será próximo de um crescimento exponencial. Mas à medida que $y(x)$ cresce, diminui a taxa de variação relativa, ou seja, a população vai crescendo cada vez mais lentamente. Quando a população se aproxima do seu máximo (ou seja, quando $y(x) \approx 1$), a taxa de variação relativa será quase nula, ou seja, a população deixa praticamente de crescer, estabilizando em torno do seu valor máximo.

Demonstração 2.6: Proposição 2.6

A equação diferencial (2.16) equivale (confirme!) a:

$$\frac{y'(x)}{y(x)\cdot[1-y(x)]} = d \qquad \Leftrightarrow \qquad \frac{y'(x)}{1-y(x)} + \frac{y'(x)}{y(x)} = d$$

Primitivando (em ordem a x) os dois lados da equação, tem-se:

$$\begin{aligned} -\ln(1-y(x)) + \ln y(x) &= d\,x + K \\ \Leftrightarrow \quad \ln\left(\frac{y}{1-y}\right) &= b_1\,x + b_0\ . \end{aligned}$$

com $b_1 = d$ e $b_0 = K$. Trata-se duma relação linear entre o *logit* de y, $y^* = \ln\left(\frac{y}{1-y}\right)$, e x.

2.3.3 Relação potência

Definição 2.7

Uma curva *potência* é uma curva com a seguinte equação:

$$y = c\,x^d \qquad \text{com} \quad x, y > 0 \quad \text{e} \quad c, d > 0\ . \tag{2.17}$$

Os gráficos de curvas potência são do tipo indicado na Figura 2.8.

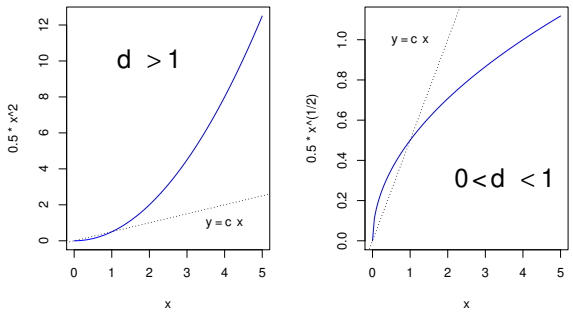


Figura 2.8: Gráficos de funções potência $y = c\,x^d$. À esquerda o caso de $d > 1$ e à direita o caso de $0 < d < 1$. Se $d = 1$ tem-se a recta $y = c\,x$, indicada com uma linha picotada.

Logaritmizando a equação (2.17), obtém-se:

$$\begin{aligned} \ln(y) &= \ln(c) + d\,\ln(x) \\ \Leftrightarrow \quad y^* &= b_0 + b_1\,x^* \end{aligned}$$

com $y^* = \ln(y)$; $x^* = \ln(x)$; $b_0 = \ln(c)$ e $b_1 = d$. Ou seja, foi obtida uma relação linear entre as transformações logarítmicas, quer de y, quer de x, em que o declive da recta (b_1) corresponde à potência d na equação potência original (2.17) e a ordenada na origem (b_0) é o logaritmo da constante multiplicativa c em (2.17).

Ilustremos uma relação potência recorrendo de novo ao exemplo da Subsecção 2.1.3. O gráfico de *log-pesos* dos recém-nascidos contra *log-idade gestacional* produz outra *relação de fundo linear*, mostrada na Figura 2.9. Esta linearização significa que a relação original (peso vs. idade gestacional) também pode ser considerada uma relação potência. Esta constatação não contradiz a afirmação feita na Subsecção 2.3.1 de que uma relação exponencial era igualmente aceitável para descrever essa mesma relação. Apenas ilustra a ideia referida na Introdução (Secção 1.2), de que pode haver mais do que uma equação adequada para modelar uma relação observada entre duas variáveis.

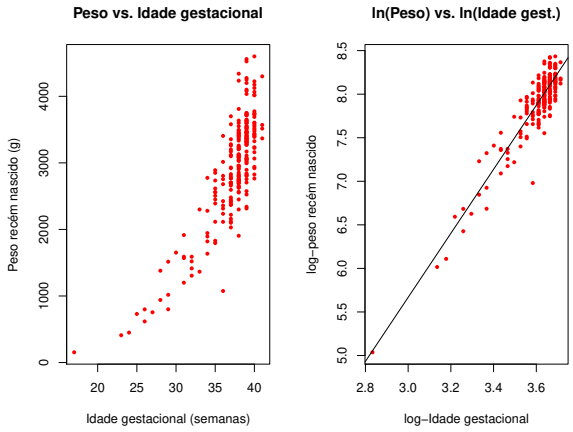


Figura 2.9: À esquerda, a relação original entre pesos dos bebés à nascença e duração da gravidez. À direita, a relação entre a log-transformação destas duas variáveis. A linearização da relação ilustra que a relação entre as variáveis originais pode ser considerada de tipo potência.

Proposição 2.7: Equação Diferencial associada à relação potência

Uma relação potência entre duas variáveis y e x surge quando se admite que ambas são funções duma terceira variável (t) e que a *taxa de variação relativa de y é proporcional à taxa de variação relativa de x*:

$$\frac{y'(t)}{y(t)} = d \cdot \frac{x'(t)}{x(t)} \ . \tag{2.18}$$

Demonstração 2.7: Proposição 2.7

Primitivando a equação (2.18) em ordem a t e depois exponenciando, tem-se:

$$\ln y \;=\; d\ln x + K \;=\; \ln x^d + K \quad\Leftrightarrow\quad y \;=\; x^d \cdot \mathrm{e}^K \quad\Leftrightarrow\quad y \;=\; c\,x^d \;,$$

com $c = \mathrm{e}^K$.

A relação potência é muito usado em estudos de *alometria*, que comparam o crescimento de partes diferentes dum organismo, ou duma parte dum organismo com o todo. A *isometria* corresponde ao valor $d=1$, ou seja, à igualdade entre as taxas de variação relativa de y e x. Diz-se que y tem uma *alometria positiva* em relação a x quando $d > 1$, ou seja, quando a taxa de variação relativa de y é maior que a de x. Diz-se que y tem uma *alometria negativa* face a x quando $d < 1$, ou seja, quando a taxa de variação relativa de y é menor. Registe-se que estas caracterizações não são simétricas, ou seja, se y tem alometria positiva face a x, então x tem alometria negativa face a y.

2.3.4 Relação de tipo hiperbólica

Definição 2.8

Chamamos *curva de tipo hiperbólico* a curvas associadas à equação (2.19).

$$y = \frac{1}{c + d\,x} \qquad \text{com} \qquad x, y > 0 \quad ; \quad c, d > 0 \;. \tag{2.19}$$

Na Figura 2.10 pode ver-se um gráfico deste tipo de curvas.

No caso particular de $c = 0$, estamos perante uma relação de proporcionalidade inversa (estudadas no Ensino Secundário), e sabemos que a essa relação correspondem curvas chamadas *hipérboles*.

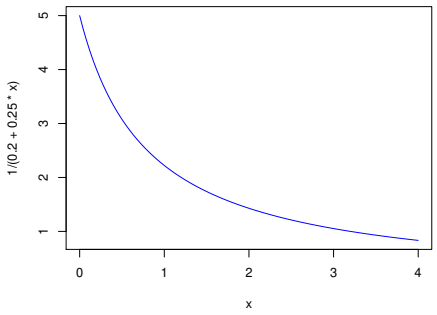


Figura 2.10: Gráficos de funções de tipo hiperbólico, dadas pela equação (2.19).

É fácil de ver que, ao considerar o *recíproco de* y, obtém-se uma *relação linear entre* $y^* = 1/y$ *e* x, e uma recta de regressão com declive $b_1 = d$ e ordenada na origem $b_0 = c$:

$$\frac{1}{y} = c + d\,x \qquad \Leftrightarrow \qquad y^* = c + d\,x\ , \tag{2.20}$$

Proposição 2.8: Equação Diferencial duma relação de tipo hiperbólico

Uma função do tipo da equação (2.19) resulta de admitir que *a taxa de variação de* y *é proporcional ao quadrado de* y ou, equivalentemente, que *a taxa de variação relativa de* y *é proporcional a* y:

$$y'(x) = -d\,y^2(x) \qquad \Leftrightarrow \qquad \frac{y'(x)}{y(x)} = -d\,y(x)\ . \tag{2.21}$$

Demonstração 2.8: Proposição 2.8

A equação diferencial (2.21) pode escrever-se como $\frac{y'(x)}{y^2(x)} = -d$. Primitivando dos dois lados, em ordem a x, tem-se $\frac{-1}{y(x)} = -dx + K$, sendo K a constante de primitivação. Tomando recíprocos, vem $y(x) = \frac{1}{dx+c}$, com $c = -K$.

Em Agronomia, este tipo de funções têm sido usadas para modelar o *rendimento por planta (y)*, como função da *densidade da cultura ou povoamento (x)*.

2.3.5 Relação Michaelis-Menten

Definição 2.9

A seguinte curva é conhecida por curva de Michaelis-Menten:

$$y = \frac{x}{c + d\,x}\ . \tag{2.22}$$

Um gráfico típico de curvas de Michaelis-Menten é dado na Figura 2.11.

A linearização duma relação de Michaelis-Menten faz-se tomando recíprocos na equação (2.22). Assim, obtém-se uma *relação linear entre* $y^* = \frac{1}{y}$ *e* $x^* = \frac{1}{x}$, com declive $b_1 = c$ e ordenada na origem $b_0 = d$.

$$\frac{1}{y} = \frac{c}{x} + d \qquad \Leftrightarrow \qquad y^* = b_0 + b_1\,x^*\ , \tag{2.23}$$

Alguns comentários sobre uma relação de Michaelis-Menten entre y e x:

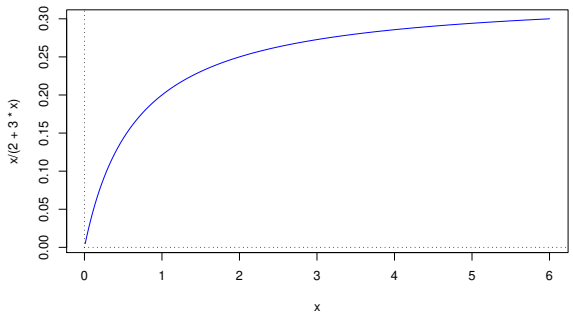


Figura 2.11: Uma típica curva de Michaelis-Menten, gráfico de funções dadas pela equação (2.22).

- A relação Michaelis-Menten é muito utilizada no estudo de *reacções enzimáticas*, relacionando a taxa da reacção (y) com a concentração do substrato (x).
- Em *modelos agronómicos de rendimento* é conhecido como modelo *Shinozaki-Kira*, com y o *rendimento total* e x a *densidade* duma cultura ou povoamento.
- Nas *pescas* é conhecido como modelo *Beverton-Holt*: y é *recrutamento* (número de novos peixes numa dada geração) e x a dimensão do *manancial* (*stock*, em inglês) de progenitores.

Proposição 2.9: Equação Diferencial da relação Michaelis-Menten

Uma relação Michaelis-Menten entre y e x resulta de admitir que *a taxa de variação de y é proporcional ao quadrado da razão entre y e x*, ou seja, que:

$$y'(x) \;=\; c\left(\frac{y(x)}{x}\right)^2. \tag{2.24}$$

Demonstração 2.9: Proposição 2.9

A equação (2.24) pode reescrever-se, dividindo tudo por $y^2(x)$, da seguinte forma: $\frac{y'(x)}{y^2(x)} \;=\; \frac{c}{x^2}$. Primitivando em ordem a x, tem-se $\frac{-1}{y(x)} = \frac{-c}{x} + K = \frac{-c+Kx}{x}$, sendo K a constante de primitivação. Tomando recíprocos, vem: $y(x) = \frac{x}{c+dx}$, com $d = -K$.

2.3.6 Advertência sobre transformações linearizantes

A regressão linear simples *não* modela directamente relações não lineares entre x e y. Caso existam transformações linearizantes, modela-se uma relação linear *entre as variáveis transformadas.*

Transformações da variável resposta y têm um impacto grande no ajustamento, uma vez que a escala dos resíduos é a escala da variável y. Logo, transformações na escala da variável resposta y *alteram a escala dos resíduos*, que é a escala que define a recta ajustada.

Assim, linearizar, obter os parâmetros b_0 e b_1 duma recta de regressão e depois desfazer a transformação linearizante *não* produz os mesmos parâmetros ajustados que resultariam de minimizar a soma de quadrados dos resíduos directamente na relação não linear. Esta última abordagem corresponde a efectuar uma *regressão não linear*, metodologia não englobada nesta disciplina.

2.3.7 Tabela de síntese das transformações linearizantes

A Tabela 2.1 sintetiza as cinco relações não lineares acima consideradas, bem como as respectivas transformações linearizantes e equações diferenciais geradoras.

Tabela 2.1: As relações não lineares consideradas, as respectivas transformações linearizantes e as equações diferenciais associadas.

Modelo	Equação	Transformações Linearizantes		Linearização $y^* = b_0 + b_1 x^*$		Equação Diferencial associada
		y^*	x^*	b_0	b_1	
Exponencial	$y = c\mathrm{e}^{dx}$	$\ln(y)$	x	$\ln(c)$	d	$\frac{y'(x)}{y(x)} = d$
Potência	$y = c\,x^d$	$\ln(y)$	$\ln(x)$	$\ln(c)$	d	$\frac{y'(t)}{y(t)} = d \cdot \frac{x'(t)}{x(t)}$
Logístico	$y = \frac{1}{1+\mathrm{e}^{-(c+d\,x)}}$	$\ln\left(\frac{y}{1-y}\right)$	x	c	d	$\frac{y'(x)}{y(x)} = d[1-y(x)]$
Hiperbólico	$y = \frac{1}{c+d\,x}$	$\frac{1}{y}$	x	c	d	$y'(x) = -d\,y^2(x)$
Michaelis-Menten	$y = \frac{x}{c+d\,x}$	$\frac{1}{y}$	$\frac{1}{x}$	d	c	$y'(x) = c\left(\frac{y(x)}{x}\right)^2$

2.4 O modelo para a inferência estatística na RLS

Até aqui a regressão linear simples foi usada apenas como *técnica descritiva*. Se as n observações usadas para ajustar a recta fossem a totalidade da população de interesse, estava dito o fundamental sobre a RLS. Mas, com frequência, as observações disponíveis são apenas uma *amostra* de uma população maior. A recta de regressão $y = b_0 + b_1\,x$ obtida com base na *amostra* não é, nesse contexto, o verdadeiro objecto de interesse. O que interessa estudar é a relação entre y e x válida para a totalidade da população. A recta amostral é apenas uma *estimativa* da recta que se supõe descrever a relação entre as duas variáveis *na população*, recta essa que se designa a *recta populacional*, e em cuja equação usamos parâmetros com a letra β ("beta") do alfabeto grego:

$$y = \beta_0 + \beta_1 x\ . \tag{2.25}$$

Coloca-se assim o problema da *inferência estatística*, ou seja, o problema de usar a informação disponível (a amostra) para extrair conclusões *sobre a recta populacional*. Como noutros contextos, *a inferência estatística assenta no pressuposto de que a amostra disponível é uma* ***amostra aleatória***, ou seja, uma

amostra extraída ao acaso da população. Será então possível usar os resultados da teoria que estuda as experiências aleatórias: a **Teoria das Probabilidades**.

Um conceito importante subjacente à Inferência Estatística é o conceito de *universo de amostras* ou *amostragem*. Uma amostra concreta de dimensão n não é única. Seria possível extrair outras amostras concretas, e cada amostra produziria outras rectas ajustadas (estimadas). No entanto, a recta populacional é única. Assim, enquanto que os parâmetros β_0 e β_1 duma recta populacional são *constantes*, os valores dos parâmetros das rectas amostrais variam de amostra concreta em amostra concreta e terão de ser descritos através do conceito de *variável aleatória*. Recorde-se que uma *variável aleatória* é o conceito que formaliza a realização de experiências aleatórias com resultado numérico, como são as observações de y resultantes duma amostra escolhida aleatoriamente. A Figura 2.12 ilustra a discussão.

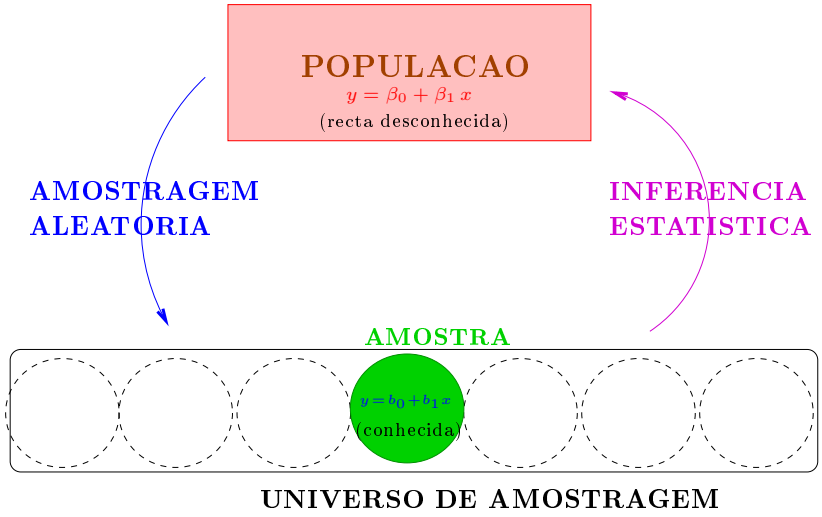


Figura 2.12: O problema da Inferência Estatística na Regressão Linear Simples, explicitando o universo das amostras. A cada amostra concreta corresponde uma recta amostral (em geral, diferente) que estima a recta populacional. A inferência basear-se-á no conhecimento da distribuição de probabilidades dos *estimadores* de β_0 e β_1 usados, ou seja, do comportamento dos valores de b_0 e b_1 ao longo do universo de amostragem.

2.4.1 O Modelo de regressão Linear Simples

Considerações iniciais. A fim de se poder fazer inferência sobre a recta populacional, admitem-se *pressupostos adicionais*. Concretamente, admite-se que:

- Y é uma variável resposta *aleatória*.
- x é uma variável preditora *não aleatória* (fixada pelo experimentador ou trabalha-se *condicionalmente* aos valores observados de x).

Nota: Mantendo a convenção usual das disciplinas introdutórias de Estatística, *variáveis aleatórias são*

indicadas por letras maiúsculas, enquanto variáveis não aleatórias, ou valores observados de variáveis aleatórias, são representados por letras minúsculas.

O modelo será ajustado com base em n pares de observações, sobre n unidades experimentais, ou seja, com base em n pares $\{(x_i, Y_i)\}_{i=1}^{n}$.

A equação de base. Vamos ainda admitir que, na população, a *relação de fundo entre as variáveis x e Y é linear*, com uma *variabilidade aleatória em torno dessa relação de fundo*, representada por uma parcela aditiva ϵ, que chamaremos **erro aleatório**. Ou seja, admitimos que cada observação da variável aleatória Y tem a seguinte estrutura:

$$\begin{array}{ccccccccc} Y_i & = & \beta_0 & + & \beta_1 & x_i & + & \epsilon_i \\ \downarrow & & \downarrow & & \downarrow & \downarrow & & \downarrow \\ \text{v.a.} & & \text{cte.} & & \text{cte.} & \text{cte.} & & \text{v.a.} \end{array}$$

para todo o $i = 1, ..., n$ (onde `v.a.` indica variável aleatória e `cte.` indica um valor não aleatório).

O erro aleatório representa *a variabilidade em torno da recta*, ou seja, a variabilidade que a relação linear de fundo entre x e Y não consegue explicar.

Pressupostos sobre os erros aleatórios. A fim de se poder fazer inferência, é ainda necessário admitir que os erros aleatórios ϵ_i verificam determinados pressupostos. No Modelo Linear admite-se que as variáveis aleatórias ϵ_i:

- *Têm valor esperado* (valor médio) *nulo*:

$$E[\epsilon_i] \; = \; 0 \; , \qquad \forall \; i = 1, ..., n \; .$$

  Não se trata dum pressuposto restritivo, uma vez que havendo um valor esperado não nulo, seria confundido com a constante aditiva β_0.

- *Têm distribuição Normal.* Trata-se dum pressuposto restritivo. No entanto, é uma hipótese bastante geral para o comportamento de erros aleatórios, após a extracção de uma tendência de fundo. As próprias origens históricas da distribuição Normal (e da curva de Gauss) assentam raízes no estudo do comportamento dos erros de medição.

- *Têm homogeneidade de variâncias*, ou seja, todos os erros aleatórios têm a mesma variância:

$$V[\epsilon_i] \; = \; \sigma^2 \; , \qquad \forall \; i = 1, ..., n \; .$$

  Trata-se dum pressuposto algo restritivo, mas conveniente: com variâncias iguais, muitas das deduções necessárias, a fim de obter os desejados resultados inferenciais, tornam-se mais simples.

- *São variáveis aleatórias independentes*, ou seja, o facto de uma observação ter um determinado erro aleatório em nada afecta o erro aleatório de outras observações. Também neste caso, trata-se duma exigência restritiva, mas conveniente, uma vez que simplifica consideravelmente o estudo. Situações

frequentes onde este pressuposto pode não se verificar dizem respeito a observações recolhidas ao longo de instantes próximos no tempo (podendo existir *auto-correlação temporal*) ou no espaço (podendo existir *auto-correlação espacial*).

Recapitulando, *para efeitos de inferência estatística*, admite-se o seguinte Modelo de Regressão Linear Simples.

Definição 2.10: Modelo de Regressão Linear Simples

Sejam dados n pares de observações $\{(x_i, Y_i)\}_{i=1}^n$. O **modelo de Regressão Linear Simples (RLS)** admite que:

1. $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \ , \quad \forall\, i = 1, ..., n.$
2. $\epsilon_i \frown \mathcal{N}(0\,,\, \sigma^2) \ , \quad \forall\, i = 1, ..., n.$
3. $\{\epsilon_i\}_{i=1}^n$ são variáveis aleatórias independentes.

A Figura 2.13 sintetiza visualmente os pressupostos agora enunciados.

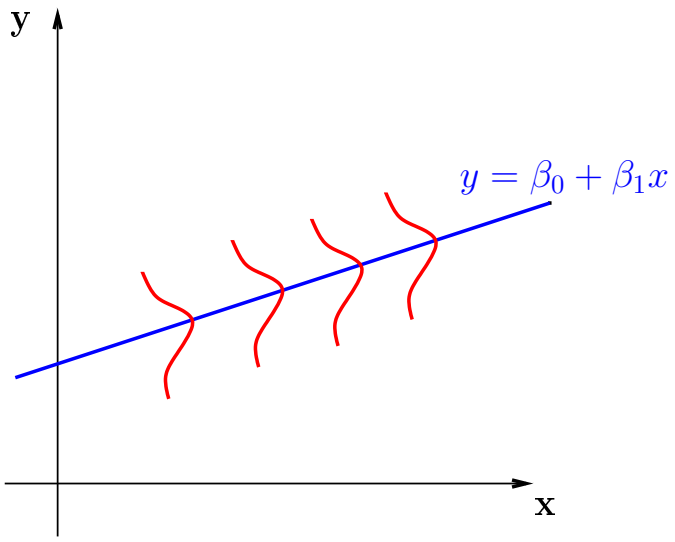


Figura 2.13: Os pressupostos do modelo de regressão linear simples admitem que, na população, existe uma relação linear de fundo entre as variáveis resposta (Y) e preditora (x), dada pela recta de equação $y = \beta_0 + \beta_1\, x$. Observações individuais distribuem-se em torno desta recta, seguindo distribuições Normais com igual valor esperado e variância comuns. Não é possível representar nesta Figura o pressuposto de independência dos erros aleatórios.

Nota 2.2

- Nesta disciplina segue-se a convenção que *o segundo parâmetro duma distribuição Normal* é a sua *variância*, e não o desvio padrão (como se convenciona na disciplina de Estatística dos primeiros ciclos do ISA).
- No Modelo RLS, os erros aleatórios são variáveis aleatórias *independentes e identicamente distribuídas (i.i.d.)*.
- A validade dos resultados inferenciais seguintes *depende da validade destes pressupostos* do modelo.

2.4.2 Uma brevíssima revisão de conceitos fundamentais

2.4.2.1 Variáveis aleatórias

No estudo de *experiências aleatórias*, e mais concretamente de experiências aleatórias cujo resultado seja *numérico*, é fundamental a ferramenta das **variáveis aleatórias**. De forma informal, podemos pensar que uma variável aleatória é uma aplicação que, a cada possível resultado duma experiência aleatória, faz corresponder um valor numérico. Pensemos por exemplo na extracção aleatória dum elemento duma população de vacas, à qual associamos a medição o respectivo peso.

Presume-se que o conceito e a caracterização de variáveis aleatórias foram estudados nas disciplinas introdutórias que o aluno terá frequentado no primeiro ciclo de estudos superiores. Estes conceitos são, por exemplo, discutidos nos apontamentos de Teoria das Probabilidades (Capítulo II) da Prof. Manuela Neves, elaborados para a disciplina Estatística, dos primeiros ciclos do ISA.

Aqui apenas se fará uma brevíssima revisão de alguns conceitos de interesse mais directo para esta disciplina. O aluno que não tenha presente estes conceitos deverá procurar nos materiais de apoio dessas disciplinas introdutórias a informação complementar necessária.

Recorde-se que as variáveis aleatórias (daqui em diante frequentemente designadas apenas pelas iniciais v.a.) podem ser:

- **Discretas**, quando podem tomar um número finito ou infinidade numerável de possíveis valores, x_i (é, por exemplo, o caso de contagens); ou
- **Contínuas**, quando podem tomar valores em intervalos, ou seja, numa infinidade não numerável de possíveis valores (como no caso da medição do peso da vaca).

existindo ainda a possibilidade de variáveis aleatórias de tipo misto.

Cada um destes tipos de v.a. tem as suas ferramentas próprias de caracterização. V.a.s discretas são caracterizadas através duma *função de probabilidades*, ou *função de massa probabilística*, ou seja, através duma função que associa a cada possível valor x_i a respectiva probabilidade p_i. Nos Exercícios Introdutórios 4 e 5 desta disciplina recordam-se dois exemplos conhecidos de distribuições de probabilidades associadas a variáveis aleatórias discretas (de contagens): a distribuição Binomial e a distribuição de Poisson.

Já uma v.a. contínua X é caracterizada através da sua *função densidade* $f(x)$, uma função não negativa que permite o cálculo da probabilidade de X tomar valores num qualquer intervalo dado, através da correspondente área por debaixo dessa função (até ao eixo dos xx), por cima do intervalo em questão:

$$P[a \leq X \leq b] = \int_a^b f(x)\, dx \ .$$

Na Figura 2.14 ilustra-se esta ideia com uma distribuição Qui-quadrado de parâmetro 6.

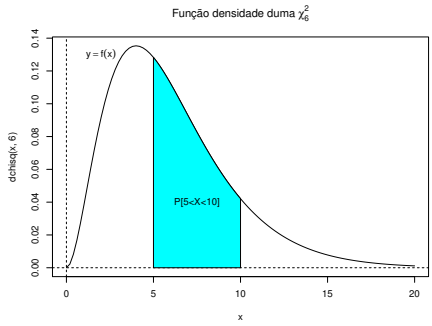


Figura 2.14: O gráfico da função densidade associada a uma distribuição Qui-quadrdao com $\nu = 6$ graus de liberdade (ou seja, uma distribuição χ^2_6). A azul claro está assinalada a região correspondente à probabilidade duma variável aleatória X com essa distribuição tomar valores no intervalo $[5, 10]$. A área total debaixo da curva é sempre 1 e corresponde à probabilidade da v.a. tomar um qualquer dos seus possíveis valores.

2.4.2.2 Valor esperado

O principal indicador *de localização* da distribuição duma variável aleatória X é o seu *valor esperado*.

Nota 2.3

O **valor esperado** ou *valor médio* duma variável aleatória X é o *centro de gravidade da sua distribuição de probabilidades* (ou seja, da sua *função de massa probabilística se X é discreta, ou função densidade se X é contínua*). A definição deste conceito é diferente para os dois tipos fundamentais de v.a.s:

- Se X é uma v.a. discreta, o seu valor esperado define-se como: $E[X] = \sum_i x_i\, p_i$;
- Se X é uma v.a. contínua, o seu valor esperado define-se como: $E[X] = \int_{-\infty}^{+\infty} x\, f(x)\, dx$.

Sejam X *e* Y *variáveis aleatórias (v.a.)* e a *e* b *constantes*. Então são válidas as seguintes *propriedades dos valores esperados* (valores médios):

- $E[X + a] = E[X] + a$.
- $E[bX] = b\,E[X]$.
- $E[X \pm Y] = E[X] \pm E[Y]$.

2.4.2.3 Variância

O principal indicador de *dispersão* da distribuição duma v.a. é a sua *variância*.

Nota 2.4

A **variância** duma v.a. X é o valor esperado do seu desvio quadrático em relação à média, ou seja:

$$V[X] \;=\; E\left[(X - E[X])^2\right] \;=\; E[X^2] - E^2[X]$$

Sejam X *e* Y *variáveis aleatórias* e a *e* b *constantes*. Então são válidas as seguintes *propriedades de variâncias* de variáveis aleatórias:

- $V[X + a] = V[X]$.
- $V[bX] = b^2\,V[X]$.
- *Se* X *e* Y *são v.a. independentes*, $V[X \pm Y] = V[X] + V[Y]$.
- Em geral, $V[X \pm Y] = V[X] + V[Y] \pm 2Cov[X, Y]$, onde $Cov[X, Y]$ é a *covariância* de X e Y.

2.4.2.4 Covariância

A **covariância** *entre duas variáveis aleatórias* X e Y mede o grau de *relacionamento linear* entre essas variáveis aleatórias.

Nota 2.5

A covariância entre X e Y define-se como:

$$Cov[X, Y] \;=\; E\left[(X - E[X])(Y - E[Y])\right] \;=\; E[XY] - E[X]\,E[Y]$$

Sejam X, Y *e* Z *variáveis aleatórias* e a *e* b *constantes*. Eis algumas propriedades da covariância:

- $Cov[X, Y] = Cov[Y, X]$ (simetria).
- $Cov[X, X] = V[X]$.
- $Cov[X + a, Y + b] = Cov[X, Y]$.

- $Cov[aX, bY] = ab\,Cov[X, Y]$.
- $Cov[X \pm Y, Z] = Cov[X, Z] \pm Cov[Y, Z]$.
- (**Desigualdade de Cauchy-Schwarz**) $|Cov[X, Y]| \leq \sqrt{V[X]\,V[Y]}$.
- *Se X, Y são variáveis aleatórias independentes*, então $Cov[X, Y] = 0$.

2.4.2.5 A distribuição Normal

A mais comum de todas as distribuições de variáveis aleatórias contínuas é a distribuição *Normal* ou *Gaussiana*. A sua função densidade está definida para qualquer valor real $x \in \mathbb{R}$:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \ ,$$

sendo μ e σ os dois parâmetros da distribuição, ou seja, duas constantes necessárias para a completa definição da distribuição. O parâmetro μ é o valor esperado (médio) da distribuição, e σ^2 é a respectiva variância. Na Figura 2.15 mostra-se a curva da função densidade duma Normal com $\mu = 0$ e $\sigma^2 = 1$, a chamada *Normal reduzida*.

Se a v.a. X tem distribuição Normal (Gaussiana), *com valor esperado μ e variância σ^2, escreve-se*:

$$X \frown \mathcal{N}(\mu, \sigma^2) \ .$$

A curva da distribuição Normal é simétrica em torno do seu valor médio, μ.

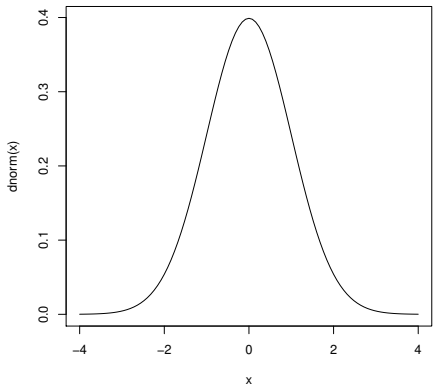


Figura 2.15: O gráfico da função densidade duma distribuição $\mathcal{N}(0, 1)$, a chamada distribuição Normal reduzida, ou Normal estandardizada, com valor médio $\mu = 0$ e desvio padrão $\sigma = 1$ (variância $\sigma^2 = 1$).

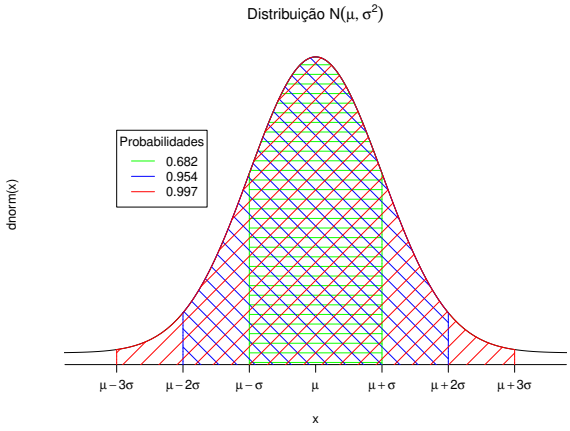
Nota 2.6

Atenção: A convenção usada nesta disciplina é que o segundo parâmetro numa distribuição Normal é a sua *variância*, e não o seu desvio padrão, como é convenção frequente noutros textos de introdução à Estatística (incluindo na disciplina de Estatística dos primeiros ciclos do ISA).

Nota 2.7

Eis algumas propriedades da distribuição Normal:

- Numa distribuição $\mathcal{N}(\mu, \sigma^2)$, cerca de 68.3% da área debaixo da curva da densidade Normal está compreendida entre $\mu - \sigma$ e $u + \sigma$ (no caso duma Normal reduzida, como na Figura 2.15, será a área debaixo da curva Gaussiana, no intervalo de x entre -1 e 1). Ao intervalo delimitado por $\mu - 2\,\sigma$ e $u + 2\,\sigma$ corresponde cerca de 95.4% da área debaixo da densidade Normal. Ao intervalo entre $\mu - 3\,\sigma$ e $u + 3\,\sigma$ corresponde cerca de 99.7% da área total. Estes resultados são ilustrados em baixo.



- Uma *transformação linear (afim) duma Normal tem distribuição Normal.* Mais concretamente, e tendo em conta as propriedades acima recordadas da esperança e variância, tem-se:

$$X \frown \mathcal{N}(\mu, \sigma^2) \text{ com } a, b \text{ constantes}, \quad \Rightarrow \quad a + b\,X \quad \frown \quad \mathcal{N}(\,a + b\,\mu\,,\, b^2\sigma^2\,)\,.$$

- Uma consequência directa da propriedade anterior é muito utilizada na leitura de tabelas da distribuição Normal, uma vez que garante que qualquer variável aleatória X com distribuição Normal pode ser *estandardizada*, ou seja, transformada numa variável aleatória com

distribuição Normal reduzida. Concretamente:

$$X \frown \mathcal{N}(\mu, \sigma^2) \quad \Rightarrow \quad \frac{X-\mu}{\sigma} \quad \frown \quad \mathcal{N}(0\,,\,1)\,.$$

- *Combinações lineares de Normais independentes têm distribuição Normal.* Mais concretamente, se X e Y são Normais *independentes* e a, b *constantes*, então $aX + bY$ é Normal (com parâmetros resultantes das propriedades recordadas nas Subsecções 2.4.2.2 e 2.4.2.3).

2.4.3 Primeiras consequências do Modelo RLS

Na formulação do Modelo de regressão linear simples, apenas se explicitaram propriedades distribucionais dos erros aleatórios ϵ_i. No entanto, implicitamente o Modelo RLS obriga a que *as observações da variável resposta, Y_i*, tenham determinadas propriedades. Concretamente, obriga a que sejam *independentes, com distribuição Normal, de valor esperado $\beta_0 + \beta_1\, x_i$ e variância σ^2*, como referido na seguinte Proposição.

Proposição 2.10: Primeiras consequências do Modelo

Dado o Modelo de Regressão Linear Simples (Definição 2.10), tem-se:

1. $E[Y_i] \;=\; \beta_0 + \beta_1\, x_i, \quad \forall i = 1, ..., n.$
2. $V[Y_i] \;=\; \sigma^2, \quad \forall i = 1, ..., n.$
3. $Y_i \;\frown\; \mathcal{N}(\beta_0 + \beta_1\, x_i\,,\, \sigma^2), \quad \forall i = 1, ..., n.$
4. $\{Y_i\}_{i=1}^n$ são variáveis aleatórias independentes.

NOTAS:

- Uma consequência do Modelo Linear é que Y *tem de ser Normal.* Assim, quando não é possível admitir que a distribuição da variável resposta Y seja pelo menos aproximadamente Normal, o Modelo Linear não é um modelo adequado.
- As observações da variável resposta Y_i *não* são identicamente distribuidas: embora sejam independentes, Normais e de variância igual, os seus valores médios são diferentes (pois dependem dos valores de $x = x_i$ associados às observações).

Demonstração 2.10: Proposição 2.10

Por aplicação directa das propriedades recordadas mais acima, tendo em conta que Y_i é uma transformação linear (afim) dos erros aleatórios ϵ_i, com β_0, β_1 e x_i constantes, e que ϵ_i é uma variável aleatória com distribuição $\mathcal{N}(0, \sigma^2)$. Concretamente, tem-se:

1. $E[Y_i] = E[\beta_0 + \beta_1\, x_i + \epsilon_i] = \beta_0 + \beta_1\, x_i + \underbrace{E[\epsilon_i]}_{=0} = \beta_0 + \beta_1\, x_i$.
2. $V[Y_i] = V[\beta_0 + \beta_1\, x_i + \epsilon_i] = \underbrace{V[\epsilon_i]}_{=\sigma^2} = \sigma^2$.
3. Y_i tem de ter distribuição Normal, já que é uma transformação linear afim da variável aleatória ϵ_i que, por pressuposto do Modelo, tem distribuição Normal. Os parâmetros da distribuição Normal de Y_i foram calculados nos dois pontos anteriores.
4. Dado um conjunto de variáveis aleatórias independentes (é o caso dos erros aleatórios ϵ_i, sob o Modelo RLS), então quaisquer suas transformações lineares comuns, como as que definem no modelo s observações da variável resposta ($Y_i = \beta_0 + \beta_1\, x_i + \epsilon_i$), preservam a independência (a independência é uma propriedade de variáveis aleatórias, não afectada por transformações que apenas envolvem constantes). Uma propriedade mais geral é dada em [4, Teorema 2.6, p.119].

2.5 Estimação dos parâmetros da recta populacional

2.5.1 Os estimadores dos parâmetros e a sua distribuição

A recta do modelo RLS tem *dois parâmetros:* β_0 *e* β_1[4]. Estes parâmetros são, respectivamente, a ordenada na origem e o declive da recta populacional, de equação $y = \beta_0 + \beta_1\, x$. A inferência sobre esses parâmetros tem de começar pela definição de **estimadores**, ou seja, de quantidades que estimem o valor de β_0 e β_1 a partir da informação disponível numa amostra aleatória.

Os estimadores de β_0 e β_1 definem-se adaptando ao contexto inferencial as expressões amostrais obtidas para b_0 e b_1 pelo Método dos Mínimos Quadrados. Recordem-se as expressões obtidas na Proposição 2.1:

$$b_1 = \frac{cov_{xy}}{s_x^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{_{(n-1)}\ s_x^2} \overset{Eq.(2.7)}{=} \frac{\sum_{i=1}^{n}(x_i - \overline{x})\, y_i}{_{(n-1)}\ s_x^2} \tag{2.26}$$

$$b_0 = \overline{y} - b_1\, \overline{x} \tag{2.27}$$

A adaptação ao contexto inferencial consiste em substituir os valores amostrais observados, y_i, pelas variáveis aleatórias correspondentes, Y_i, bem como, na equação para $\hat{\beta}_0$, substituir o valor amostral do declive, b_1, pelo estimador $\hat{\beta}_1$.

Definição 2.11: Estimadores de β_1 e β_0

[4] Infelizmente, é frequente que um mesmo conceito estatístico tenha designações diferentes e, conversamente, que uma mesma expressão designe conceitos diferentes. Nesta disciplina, entende-se por *parâmetro* uma constante necessária para especificar completamente um dado modelo. Há quem designe por parâmetros as quantidades observadas, conceito que nesta disciplina é sempre designada por *variável*.

Sejam dados n pares de observações $\{(x_i, Y_i)\}_{i=1}^n$. Os **estimadores de mínimos quadrados dos parâmetros da recta de regressão populacional**, $Y = \beta_0 + \beta_1\, x$, são:

$$\hat{\beta}_1 \;=\; \frac{\sum_{i=1}^{n}(x_i-\overline{x})\,Y_i}{_{(n-1)}\;s_x^2} \;=\; \sum_{i=1}^{n} c_i Y_i\ , \qquad \text{com} \quad c_i \;=\; \frac{(x_i-\overline{x})}{_{(n-1)}\;s_x^2} \tag{2.28}$$

$$\hat{\beta}_0 \;=\; \overline{Y} - \hat{\beta}_1\overline{x} \;=\; \frac{1}{n}\sum_{i=1}^{n} Y_i - \overline{x}\sum_{i=1}^{n} c_i Y_i \;=\; \sum_{i=1}^{n} d_i Y_i\ , \tag{2.29}$$

com

$$d_i \;=\; \frac{1}{n} - \overline{x}c_i \;=\; \frac{1}{n} - \frac{(x_i-\overline{x})\overline{x}}{_{(n-1)}\;s_x^2}.$$

Quer $\hat{\beta}_1$, quer $\hat{\beta}_0$, são *combinações lineares* das observações $\{Y_i\}_{i=1}^n$ logo, dado o Modelo RLS, *são combinações lineares de variáveis aleatórias Normais independentes.* Assim, pelas propriedades da Normal (Nota 2.7), é imediato que *ambos os estimadores têm distribuição Normal.*

Proposição 2.11: Distribuição dos estimadores dos parâmetros

Dado o Modelo de Regressão Linear Simples, os estimadores dos parâmetros da recta populacional têm as seguintes distribuições de probabilidades:

1. $\hat{\beta}_1 \;\frown\; \mathcal{N}\left(\beta_1\,,\ \frac{\sigma^2}{_{(n-1)}\,s_x^2}\right)$,
2. $\hat{\beta}_0 \;\frown\; \mathcal{N}\left(\beta_0\ ,\ \sigma^2\ \left[\frac{1}{n} + \frac{\overline{x}^2}{_{(n-1)}\,s_x^2}\right]\right)$

Demonstração 2.11: Proposição 2.11

1. Por definição (equação 2.28) tem-se: $\hat{\beta}_1 = \frac{cov_{xY}}{s_x^2} = \sum_{i=1}^{n} c_i\, Y_i$, com $c_i = \frac{x_i-\overline{x}}{(n-1)\,s_x^2}$. Logo, $\hat{\beta}_1$ é uma *combinação linear das observações de* Y, $\{Y_i\}_{i=1}^n$. Mas essas observações têm distribuição Normal e são independentes (Proposição 2.10). Como *qualquer combinação linear de Normais independentes é Normal* (Subsecção 2.4.2.5), tem-se que $\hat{\beta}_1$ *tem distribuição Normal.* Falta indicar com que parâmetros. Uma vez que o primeiro parâmetro duma Normal é o seu valor esperado, temos (tendo em conta as propriedades das Subsecções 2.4.2.2 e 2.4.2.3):

$$E[\hat{\beta}_1] \;=\; E\left[\sum_{i=1}^{n} c_i\, Y_i\right] \;=\; \sum_{i=1}^{n} c_i\, E[Y_i] \;=\; \sum_{i=1}^{n} c_i\,(\beta_0+\beta_1\, x_i) \;=\; \beta_0\sum_{i=1}^{n} c_i + \beta_1\sum_{i=1}^{n} c_i\, x_i\ .$$

Veremos de seguida que o primeiro destes somatórios tem soma nula, e o segundo tem soma

1, pelo que o valor esperado pretendido é β_1. De facto,

$$\sum_{i=1}^{n} c_i = \sum_{i=1}^{n} \frac{x_i - \overline{x}}{(n-1)\, s_x^2} = \frac{1}{(n-1)\, s_x^2} \underbrace{\sum_{i=1}^{n} (x_i - \overline{x})}_{=0} = 0 \qquad \text{(ver eq. 2.8)}$$

$$\sum_{i=1}^{n} c_i\, x_i = \sum_{i=1}^{n} \frac{(x_i - \overline{x})x_i}{(n-1)\, s_x^2} = \frac{1}{(n-1)\, s_x^2} \underbrace{\sum_{i=1}^{n} (x_i - \overline{x})x_i}_{=(n-1)\, s_x^2} = 1 \qquad \text{(ver eq. 2.9)}$$

Logo, $E[\hat{\beta}_1] = \beta_1$, pelo que o estimador é **centrado**.

Vejamos agora a expressão para o segundo parâmetro da Normal, que sabemos ser a variância. Recordem-se as propriedades das variâncias, e tenha-se presente que as observações $\{Y_i\}_{i=1}^n$ são variáveis aleatórias independentes.

$$V[\hat{\beta}_1] = V\left[\sum_{i=1}^{n} c_i\, Y_i\right] = \sum_{i=1}^{n} c_i^2 \underbrace{V[Y_i]}_{=\sigma^2\,,\,\forall\, i} = \sigma^2 \sum_{i=1}^{n} \underbrace{\frac{(x_i - \overline{x})^2}{[(n-1)\, s_x^2]^2}}_{=c_i^2} = \frac{\sigma^2}{[(n-1)\, s_x^2]^2} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

$$= \frac{\sigma^2}{[(n-1)\, s_x^2]^2} \cdot (n-1)\, s_x^2 = \frac{\sigma^2}{(n-1)\, s_x^2}\,.$$

2. Também o estimador $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\, \overline{x} = \sum_{i=1}^{n} d_i Y_i$, com $d_i = \frac{1}{n} - \overline{x}\, c_i$, é uma combinação linear de v.a.s Normais independentes (as observações Y_i), logo tem distribuição Normal. Falta determinar os respectivos parâmetros. Recordando os resultados relativos ao estimador $\hat{\beta}_1$, tem-se:

$$E\left[\hat{\beta}_0\right] = E\left[\overline{Y} - \hat{\beta}_1\, \overline{x}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] - \overline{x}\, \underbrace{E\left[\hat{\beta}_1\right]}_{=\,\beta_1} = \frac{1}{n}\sum_{i=1}^{n} \underbrace{(\beta_0 + \beta_1\, x_i)}_{=E[Y_i]} - \beta_1\, \overline{x}$$

$$= \underbrace{\frac{1}{n}\sum_{i=1}^{n} \beta_0}_{=n\,\beta_0} + \beta_1\, \underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i}_{=\overline{x}} - \beta_1\, \overline{x} = \beta_0 + \beta_1 \overline{x} - \beta_1 \overline{x} = \beta_0.$$

Tendo em conta as propriedades da variância,

$$V[\hat{\beta}_0] = V\left[\overline{Y} - \hat{\beta}_1\, \overline{x}\right] = V[\overline{Y}] + V[\hat{\beta}_1\, \overline{x}] - 2\, cov[\overline{Y}, \hat{\beta}_1\, \overline{x}]$$

$$= V\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] + \overline{x}^2 V[\hat{\beta}_1] - 2\overline{x} \underbrace{cov[\overline{Y}, \hat{\beta}_1]}_{=0\ (Ex.RLS\ 13a))}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} \underbrace{V[Y_i]}_{=\sigma^2} + \overline{x}^2 \frac{\sigma^2}{(n-1)s_x^2} = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{(n-1)\, s_x^2}\right],$$

o que completa a demonstração.

NOTAS:

1. Ambos *os estimadores são estimadores* ***centrados***, ou seja, o seu valor esperado é igual ao parâmetro que estimam: $E[\hat{\beta}_1] = \beta_1$ e $E[\hat{\beta}_0] = \beta_0$
2. Quanto maior $_{(n-1)}\, s_x^2$, menor a variabilidade dos estimadores.
3. A variabilidade de $\hat{\beta}_0$ também diminui com o aumento de n, e a maior proximidade de $\overline{x}$ a 0.

2.5.1.1 Significado das distribuições dos estimadores

Como se pode interpretar o significado prático dos resultados distribucionais da Proposição 2.11? As distribuições dos estimadores são *distribuições na amostragem*, ou seja, podem ser vistas como a distribuição de frequências dos valores de cada um dos estimadores, ao longo de todo o universo de possíveis amostras. Assim, para interpretar o resultado distribucional do estimador $\hat{\beta}_1$, podemos pensar que *se fossem recolhidas todas as possíveis amostras aleatórias (com os n valores de x_i fixados), e para cada uma calculado o declive b_1 da respectiva recta amostral, a distribuição de frequências desses declives amostrais seria a mostrada na Figura 2.16.*

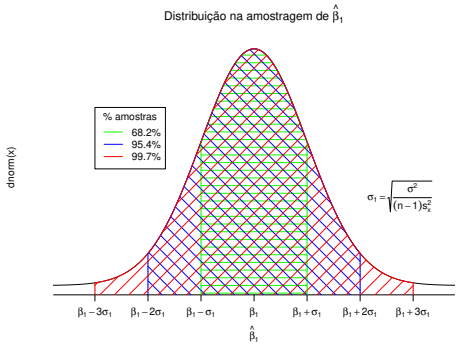


Figura 2.16: A distribuição de probabilidades de $\hat{\beta}_1$ pode ser interpretada como sendo a distribuição do conjunto de valores que b_1 tomaria, ao longo da totalidade de possíveis amostras de tamanho n (e com os valores de x_i usados na nossa amostra).

Na interpretação da Figura 2.16, podem usar-se os resultados recordados na Subsecção 2.4.2.5 para afirmar que a distância entre a estimativa b_1 e o verdadeiro valor de β_1 será:

inferior a $\sigma_{\hat{\beta}_1}$ em $\approx 68\%$ das amostras;

inferior a $2\,\sigma_{\hat{\beta}_1}$ em $\approx 95\%$ das amostras;

inferior a $3\,\sigma_{\hat{\beta}_1}$ em $\approx 99,7\%$ das amostras.

Repare-se na importância que tem, nesta interpretação, o desvio padrão associado à distribuição do estimador $\hat{\beta}_1$, ou seja,

$$\sigma_{\hat{\beta}_1} = \sqrt{V[\hat{\beta}_1]} = \sqrt{\frac{\sigma^2}{(n-1)\, s_x^2}} = \frac{\sigma}{\sqrt{(n-1)\, s_x^2}} \,. \tag{2.30}$$

O desvio padrão associado a um estimador é usualmente designado um **erro padrão** desse estimador (em inglês, *standard error*). Assim, o erro padrão de $\hat{\beta}_1$ foi dado na equação (2.30), enquanto que o erro padrão associado a $\hat{\beta}_0$ é dado na equação (2.31). Apesar da semelhança de notação, *não se deve confundir os erros padrões dos estimadores, $\sigma_{\hat{\beta}_1}$ e $\sigma_{\hat{\beta}_0}$, com o desvio padrão σ dos erros aleatórios.*

$$\sigma_{\hat{\beta}_0} = \sqrt{V[\hat{\beta}_0]} = \sqrt{\sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{(n-1)\, s_x^2}\right]} = \sigma \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{(n-1)\, s_x^2}} \,. \tag{2.31}$$

Da Proposição 2.11 sai directamente o seguinte Corolário.

Corolário 2.1

Dado o Modelo de Regressão Linear Simples, têm-se as seguintes distribuições:

1. $\frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \frown \mathcal{N}(0\,,\,1)$, com $\sigma_{\hat{\beta}_1} = \sqrt{\frac{\sigma^2}{(n-1)\, s_x^2}} = \frac{\sigma}{\sqrt{(n-1)\, s_x^2}}$
2. $\frac{\hat{\beta}_0 - \beta_0}{\sigma_{\hat{\beta}_0}} \frown \mathcal{N}(0\,,\,1)$, com $\sigma_{\hat{\beta}_0} = \sqrt{\sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{(n-1)\, s_x^2}\right]} = \sigma \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{(n-1)\, s_x^2}}$

Demonstração 2.12: Corolário 2.1

Basta aplicar directamente às distribuições da Proposição 2.11 o resultado, visto na Subsecção 2.4.2.5), de que se a uma variável aleatória com distribuição Normal subtrairmos a sua média e dividirmos pelo seu desvio padrão, obtemos uma nova v.a. com distribuição Normal reduzida, $\mathcal{N}(0,1)$.

Os resultados do Corolário anterior poderiam ser usados para fazer inferência estatística sobre os parâmetros β_0 e β_1 (e.g., construir intervalos de confiança ou efectuar testes de hipóteses), mas *apenas no caso de ser conhecida a variância dos erros aleatórios, $\sigma^2 = V[\epsilon_i]$*, que aparece nas expressões dos erros padrão (equações 2.30 e 2.31).

No entanto, a variância dos erros aleatórios, σ^2 é, na prática, desconhecida. Para que os resultados já estudados possam ter aplicabilidade, *será necessário obter um estimador da variância σ^2 dos erros aleatórios.* Vamos, na Subsecção seguinte, construir um estimador para σ^2 *a partir dos resíduos.*

2.5.2 Erros aleatórios e Resíduos

A partir da definição do Modelo de Regressão Linear Simples (Definição 2.10), cada erro aleatório é a diferença entre um valor da variável resposta, Y_i, e a ordenada, para o correspondente valor x_i do preditor, na recta de regressão *populacional*, ou seja:

$$\textit{Erros aleatórios:} \quad \epsilon_i \;=\; Y_i - (\beta_0 + \beta_1 x_i)$$

Os erros aleatórios não são conhecidos, mesmo após a extracção duma amostra, porque os parâmetros da recta populacional (β_0 e β_1) são desconhecidos. Por definição (Definição 2.1), os resíduos são as correspondentes diferenças entre um valor da variável resposta e o valor da ordenada (para igual x_i), na recta de regressão *amostral*. *No contexto inferencial, os resíduos também são variáveis aleatórias* (definidas a partir da variável aleatória Y_i e dos estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$, que também são v.a.s), pelo que serão escritos com letra maiúscula:

$$\textit{Resíduos:}\ \text{(v.a.)} \quad E_i \;=\; Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Os *resíduos são conhecíveis*, porque podem ser calculados a partir do momento em que se disponha duma amostra concreta, ou seja, quando Y_i tomar o valor concreto observado, y_i, e os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ tomarem os valores concretos das *estimativas* b_0 e b_1. Esses *resíduos observados* (que são a concretização das variáveis aleatórias E_i) são os resíduos anteriormente discutidos no contexto descritivo:

$$\textit{Resíduos:}\ \text{(observados)} \quad e_i \;=\; y_i - (b_0 + b_1 x_i)$$

Assim, os resíduos E_i são **preditores**[5] (conhecíveis) *dos erros aleatórios* (desconhecidos) ϵ_i.

É natural que, para estimar a variância σ^2 dos erros aleatórios se olhe para a variância dos preditores desses mesmos erros aleatórios. Ora, como se viu no contexto descritivo, a Soma de Quadrados Residual é *o numerador da variância amostral dos resíduos*:

$$_{(n-1)}\, s_E^2 \;=\; \sum_{i=1}^{n} E_i^2 \;=\; SQRE\ .$$

Assim, é natural que a Soma de Quadrados Residual, $SQRE$, desempenhe um papel central na estimação da variância (comum) dos erros aleatórios ϵ_i. Mas para o poder fazer, são necessários resultados relativos ao comportamento de $SQRE$ em contexto inferencial, ou seja, é necessário conhecer o comportamento que, dado o Modelo RLS, terá a Soma de Quadrados Residual ao longo do universo de amostras. Os resultados fundamentais são dados na seguinte Proposição, cuja demonstração se omite, dado ultrapassar as ferramentas disponíveis no âmbito desta disciplina.

[5] Em termos técnicos distingue-se entre um *estimador*, que é uma quantidade amostral usada para estimar *uma constante* (como são os parâmetros da recta populacional, β_0 e β_1), e um *preditor* que é uma quantidade amostral que procura prever valores duma *variável aleatória* (com são os erros aleatórios ϵ_i).

Proposição 2.12: Resultados distribucionais de SQRE

Dado o Modelo de Regressão Linear Simples (RLS), tem-se:

- $\frac{SQRE}{\sigma^2} \frown \chi^2_{n-2}$
- $SQRE$ é independente de $(\hat{\beta}_0, \hat{\beta}_1)$.

A distribuição indicada na Proposição 2.12 é a distribuição Qui-quadrado. As propriedades das distribuições χ^2 encontram-se nos apontamentos da Prof. Manuela Neves (na parte relativa à Teoria das Probabilidades). Entre essas propriedades, tem-se:

- Uma distribuição Qui-quadrado tem um único parâmetro que, por razões históricas, é designado *graus de liberdade*, escrevendo-se χ^2_ν para indicar que o parâmetro tem o valor[6] ν.
- O valor esperado duma v.a. W com distribuição $W \frown \chi^2_\nu$ é o valor do parâmetro: $E[W]=\nu$.
- A variância duma v.a. W com distribuição $W \frown \chi^2_\nu$ é o dobro do valor do parâmetro: $V[W]=2\nu$

Sai assim, a partir do primeiro ponto da Proposição 2.12 e das propriedades do valor esperado (Subsecção 2.4.2.2), que:

Corolário 2.2

Dado o Modelo de RLS, $E\left[\frac{SQRE}{n-2}\right] = \sigma^2$.

Este Corolário aponta imediatamente para um *estimador centrado de* σ^2, que é agora definido:

Definição 2.12: Quadrado Médio Residual

Define-se o **Quadrado Médio Residual** ($QMRE$) numa Regressão Linear Simples como

$$QMRE = \frac{SQRE}{n-2}$$

Resumindo:

- *O Quadrado Médio Residual, QMRE, é habitualmente usado na Regressão como estimador da variância dos erros aleatórios*, isto é, estimador de $\sigma^2=V[\epsilon_i]$:

$$\hat{\sigma}^2 = QMRE \ . \tag{2.32}$$

[6] A letra grega ν lê-se *nu* e corresponde ao nosso n.

- Como se viu, QMRE é um *estimador centrado* da variância dos erros aleatórios ϵ_i.

Iremos ver que *a substituição de σ^2 pelo seu estimador $QMRE$* no Corolário 2.1 *transforma as distribuições Normais em distribuições t-Student.*

2.5.2.1 Revisão: como surge uma distribuição *t-Student*

Recordemos *como surge uma distribuição t-Student* (ver os apontamentos da Prof. Manuela Neves, Inferência Estatística, Capítulo III).

- Seja dada uma variável aleatória Normal reduzida: $Z \frown \mathcal{N}(0,1)$;
- Seja dada uma variável aleatória com distribuição Qui-quadrado: $W \frown \chi^2_\nu$;
- Sejam essas duas v.a. Z e W *independentes.*

Então, verifica-se que a seguinte razão tem distribuição t-Student:

$$\frac{Z}{\sqrt{W/\nu}} \frown t_\nu \ . \tag{2.33}$$

2.5.3 Quantidades centrais para a inferência sobre β_0 e β_1

Toma-se $Z = \frac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}}$, $W = \frac{SQRE}{\sigma^2}$ e $\nu = n-2$, e aplicam-se os resultados da Subsecção 2.5.2.1.

Proposição 2.13: Distribuições para a inferência sobre β_0 e β_1

Dado o Modelo de Regressão Linear Simples, tem-se:

1. $\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \frown t_{n-2}$, com $\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{QMRE}{(n-1)\, s_x^2}}$;
2. $\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \frown t_{n-2}$, com $\hat{\sigma}_{\hat{\beta}_0} = \sqrt{QMRE \left[\frac{1}{n} + \frac{\overline{x}^2}{(n-1)\, s_x^2}\right]}$.

Demonstração 2.13: Proposição 2.13

1. Sabemos que, dado o Modelo RLS, a quantidade $Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}}$ tem distribuição $\mathcal{N}(0,1)$ (Corolário 2.1). Pretende-se mostrar que, substituindo $\sigma_{\hat{\beta}_1}$ pelo seu estimador $\hat{\sigma}_{\hat{\beta}_1}$, a distribuição resultante é uma t-Student. Ora, na Proposição 2.12 viu-se que $W = \frac{SQRE}{\sigma^2} \frown \chi^2_{n-2}$, sendo $SQRE$ independente de $\hat{\beta}_1$. Logo, e relembrando a forma como surgem distribuições t-Student

(Subsecção 2.5.2.1), sabemos que $\frac{Z}{\sqrt{W/(n-2)}} \frown t_{n-2}$. Mas,

$$\frac{Z}{\sqrt{W/(n-2)}} = \frac{\frac{\hat{\beta}_1-\beta_1}{\sigma_{\hat{\beta}_1}}}{\sqrt{\frac{SQRE}{\sigma^2\cdot(n-2)}}} = \frac{\frac{\hat{\beta}_1-\beta_1}{\sqrt{\frac{\cancel{\sigma^2}}{(n-1)\,s_x^2}}}}{\sqrt{\frac{QMRE}{\cancel{\sigma^2}}}} = \frac{\hat{\beta}_1-\beta_1}{\sqrt{\frac{QMRE}{(n-1)\,s_x^2}}} = \frac{\hat{\beta}_1-\beta_1}{\hat{\sigma}_{\hat{\beta}_1}} .$$

Assim, $\frac{\hat{\beta}_1-\beta_1}{\hat{\sigma}_{\hat{\beta}_1}}$ tem distribuição t-Student, com $n-2$ graus de liberdade.

2. A demonstração para o caso de β_0 é inteiramente análoga.

Este resultado é crucial, pois dá-nos os resultados que servirão de base à construção de *intervalos de confiança* e *testes de hipóteses* para os parâmetros da recta populacional, β_0 e β_1.

2.6 Intervalos de confiança para os parâmetros da recta

2.6.1 Intervalos de confiança para β_1 e β_0

Estamos agora em condições de construir os intervalos de confiança para os parâmetros duma recta populacional, ou seja, para β_0 e β_1.

Proposição 2.14: Intervalo de Confiança a $(1-\alpha)\times 100\%$ para β_1 e β_0

Dado o Modelo RLS,

1. o intervalo a $(1-\alpha)\times 100\%$ de confiança para o declive β_1 da recta de regressão populacional é:
$$\left]\ b_1 - t_{\frac{\alpha}{2}\,(n-2)}\,\hat{\sigma}_{\hat{\beta}_1} \quad , \quad b_1 + t_{\frac{\alpha}{2}\,(n-2)}\,\hat{\sigma}_{\hat{\beta}_1}\ \right[\ , \tag{2.34}$$
sendo $t_{\frac{\alpha}{2}(n-2)}$ o valor que, numa distribuição $t_{(n-2)}$, deixa *à direita* uma região de probabilidade $\alpha/2$ (ou seja, o quantil de ordem $1-\frac{\alpha}{2}$ numa distribuição t_{n-2}). As quantidades $b_1 = \frac{cov_{xy}}{(n-1)\,s_x^2}$ e $\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{QMRE}{(n-1)\,s_x^2}}$ foram definidas anteriormente.

2. o intervalo a $(1-\alpha)\times 100\%$ de confiança para a ordenada na origem, β_0, da recta de regressão populacional é dado por:
$$\left]\ b_0 - t_{\frac{\alpha}{2}\,(n-2)}\cdot\hat{\sigma}_{\hat{\beta}_0} \quad , \quad b_0 + t_{\frac{\alpha}{2}\,(n-2)}\cdot\hat{\sigma}_{\hat{\beta}_0}\ \right[\ , \tag{2.35}$$
onde $b_0 = \overline{y} - b_1\overline{x}$ e $\hat{\sigma}_{\hat{\beta}_0} = \sqrt{QMRE\cdot\left[\frac{1}{n} + \frac{\overline{x}^2}{(n-1)\,s_x^2}\right]}$ foram definidos anteriormente.

NOTAS:

- A estrutura dos dois intervalos de confiança é análoga: são ambos centrados na estimativa do

respectivo parâmetro (b_0 ou b_1), e para chegar aos dois extremos do intervalo tem de se subtrair e somar o produto do respectivo erro padrão ($\hat{\sigma}_{\hat{\beta}_0}$ ou $\hat{\sigma}_{\hat{\beta}_1}$) vezes o quantil adequado da distribuição t-Student ($t_{\frac{\alpha}{2}\,(n-2)}$);

- Assim, a amplitude (comprimento) de cada intervalo de confiança é duas vezes o produto do erro padrão vezes o quantil da distribuição t-Student: $2\,t_{\frac{\alpha}{2}\,(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_0}$ e $2\,t_{\frac{\alpha}{2}\,(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_1}$, respectivamente para o intervalo de confiança de β_0 e de β_1;
- A amplitude do intervalo de confiança (IC) para β_1 *aumenta com* $QMRE$ e *diminui com* n *e* s_x^2;
- A amplitude do IC para β_0 *aumenta com* $QMRE$ *e com* $\overline{x}^2$ e *diminui com* n *e* s_x^2;
- A amplitude de ambos os ICs *aumenta para maiores graus de confiança* $1{-}\alpha$, ou seja, para aumentar o grau de confiança do intervalo, é necessário sacrificar a sua precisão (isto é, aceitar uma amplitude maior).

Demonstração 2.14: Proposição 2.14

1. **Intervalo de confiança para** β_1. Sabemos (Proposição 2.13) que $\frac{\hat{\beta}_1-\beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \frown t_{n-2}$. Designando por $t_{\frac{\alpha}{2}(n-2)}$ o valor que, numa distribuição t_{n-2} deixa à sua direita uma região de probabilidade $\alpha/2$, e uma vez que o simétrico desse valor, $-t_{\frac{\alpha}{2}(n-2)}$, será (pela simetria da distribuição t-Student em torno de zero) o valor que deixa à sua *esquerda* uma área $\frac{\alpha}{2}$, pode-se escrever a seguinte equação:

$$P\left[-t_{\frac{\alpha}{2}(n-2)} \;<\; \frac{\hat{\beta}_1-\beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \;<\; t_{\frac{\alpha}{2}(n-2)}\right] \;=\; 1-\alpha$$

Ao substituir-se a dupla desigualdade por duplas desigualdades equivalentes, não se altera a probabilidade $1-\alpha$. Vamos assim proceder a escrever duplas desigualdades equivalentes, com o objectivo de isolar o parâmetro para o qual se pretende construir o intervalo de confiança (β_1). Comecemos por multiplicar a dupla desigualdade por $\hat{\sigma}_{\hat{\beta}_1}$, depois por -1 (recordando que, ao multiplicar desigualdades por números negativos, é necessário trocar o sentido das desigualdades) e, finalmente, vamos somar $\hat{\beta}_1$:

$$\begin{aligned}
& -t_{\frac{\alpha}{2}(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_1} \;<\; \hat{\beta}_1-\beta_1 \;<\; t_{\frac{\alpha}{2}(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_1} \\
\Leftrightarrow\quad & t_{\frac{\alpha}{2}(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_1} \;>\; \beta_1-\hat{\beta}_1 \;>\; -t_{\frac{\alpha}{2}(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_1} \\
\Leftrightarrow\quad & \hat{\beta}_1 - t_{\frac{\alpha}{2}(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_1} \;<\; \beta_1 \;<\; \hat{\beta}_1 + t_{\frac{\alpha}{2}(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_1}\,.
\end{aligned}$$

Assim, a probabilidade de o verdadeiro valor do declive β_1 da recta populacional estar contido entre os dois extremos indicados é $1-\alpha$. Mas este intervalo é um *intervalo aleatório*: os seus extremos são constituídos por variáveis aleatórias ($\hat{\beta}_1$ e $\hat{\sigma}_{\hat{\beta}_1}$), que tomam diferentes valores para cada amostra concreta que seja extraída da população. Para *uma amostra concreta* obter-se-á um *intervalo concreto* substituindo o estimador $\hat{\beta}_1$ pela estimativa concreta b_1 e substituindo o erro padrão estimado $\hat{\sigma}_{\hat{\beta}_1}$ pelo seu valor concreto (que continuamos a designar por $\hat{\sigma}_{\hat{\beta}_1}$). O intervalo assim resultante chama-se um *intervalo de confiança a* $(1-\alpha)\times 100\%$

para β_1:

$$\left] \, b_1 - t_{\frac{\alpha}{2}\,(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_1} \quad , \quad b_1 + t_{\frac{\alpha}{2}\,(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_1} \, \right[\, .$$

2. **Intervalo de confiança para** β_0. A demonstração é análoga.

Para interpretar correctamente um intervalo de confiança é necessário recordar que, *a cada possível amostra concreta (com os n valores de x_i dados) corresponde um intervalo de confiança concreto*. Cada um desses intervalos pode, ou não, conter o verdadeiro valor de β_1. Mas o resultado teórico na origem da contrução do IC (a Proposição 2.13) garante que, *nessa família de todos os intervalos concretos, $(1-\alpha) \times 100\%$ dos intervalos contêm o verdadeiro valor do declive populacional* β_1. Nunca saberemos se a amostra concreta por nós escolhida gera um intervalo que contém o verdadeiro valor de β_1 (para saber isso seria necessário conhecer o verdadeiro valor de β_1). Mas, uma vez que o intervalo gerado foi escolhido ao acaso duma família de intervalos com essa propriedade, dizemos que temos uma *confiança* $(1-\alpha) \times 100\%$ em como contém β_1.

2.6.2 Um exemplo no R: os lírios de Fisher

Um conjunto de dados associado ao nome do famoso estatístico britânico Ronald Fisher, mas recolhidos por Edgar Anderson, é constituído por medições morfométricas em $n = 150$ lírios. Disponível na *data frame* `iris`, em qualquer distribuição padrão do R, o conjunto de dados contém medições de 4 variáveis numéricas: comprimentos e larguras de sépalas, e de pétalas – veja-se `help(iris)` para mais pormenores.

A nuvem de pontos relacionando *largura* e *comprimento* das *pétalas* é dada na Figura 2.17.

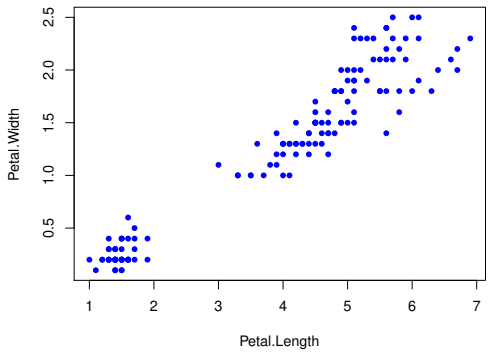


Figura 2.17: Largura contra comprimento das pétalas em 150 lírios (dados `iris`).

No R, as regressões lineares são ajustadas usando o *comando* `lm` (que são as iniciais, em inglês de ***linear model***). O comando seguinte ajusta a recta de regressão de largura das pétalas sobre comprimento das pétalas, guardando o resultado num objecto de nome `iris.lm`:

```
> iris.lm <- lm(Petal.Width ~ Petal.Length, data=iris)
```

```
> iris.lm
Call:
lm(formula = Petal.Width ~ Petal.Length, data = iris)
Coefficients:
(Intercept)  Petal.Length
    -0.3631        0.4158
```

A *recta estimada* é assim a recta de equação $y = -0.3631 + 0.4158\,x$, onde y indica a largura da pétala e x o seu comprimento.

No R, a recta pode ser sobreposta à nuvem de pontos se, após criar a nuvem com os comandos anteriores, fôr usado o comando `abline`, como indicado de seguida, produzindo o resultado da Figura 2.18.

```
> abline(iris.lm, col="red")
```

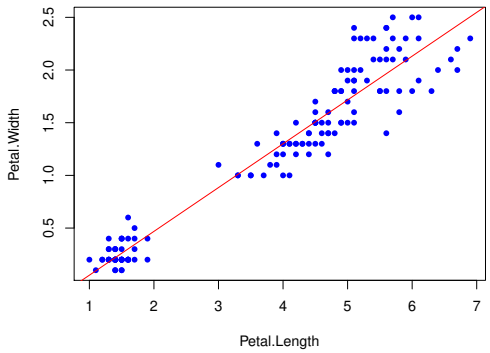


Figura 2.18: A recta de regressão ajustada, sobreposta à nuvem de pontos dos dados dos lírios com o comando `abline`.

Mais informações úteis sobre a regressão obtêm-se através do comando `summary`, aplicado à regressão ajustada:

```
> summary(iris.lm)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.363076   0.039762  -9.131  4.7e-16 ***
Petal.Length  0.415755   0.009582  43.387  < 2e-16 ***
```

Na primeira coluna da listagem de saída são indicados os valores das estimativas b_0 e b_1 (já vistos antes). Na segunda coluna são indicados os valores dos *erros padrões estimados*, para cada estimador:

$$\hat{\sigma}_{\hat{\beta}_0} = 0.039762 \qquad\qquad \hat{\sigma}_{\hat{\beta}_1} = 0.009582\,.$$

Estes valores são usados na construção dos intervalos de confiança para β_0 e β_1, usando as fórmulas da Proposição 2.14. Uma vez que $b_1 = 0.415755$ e (optando por um intervalo a 95% de confiança, ou seja, com $\frac{\alpha}{2} = 0.025$) $t_{0.025(148)} = 1.976122$, tem-se:

$$\begin{aligned} b_1 - t_{0.025(148)} \times \hat{\sigma}_{\hat{\beta}_1} &= 0.3968198 \\ b_1 + t_{0.025(148)} \times \hat{\sigma}_{\hat{\beta}_1} &= 0.4346902 \end{aligned}$$

Logo, o intervalo a 95% de confiança para o declive populacional β_1 resultante da nossa amostra é:

$$] \, 0.3968198 \, , \; 0.4346902 \, [\; .$$

Para calcular os intervalos de confiança, directamente no `R`, pode usar-se o comando `confint` sobre uma regressão ajustada:

```
> confint(iris.lm)
                 2.5 %     97.5 %
(Intercept)  -0.4416501 -0.2845010   <-- ordenada na origem
Petal.Length  0.3968193  0.4346915   <-- declive
```

Por omissão, o IC calculado é a 95% *de confiança.* Neste caso, podemos afirmar com 95% de confiança que o declive β_1 da recta populacional está no intervalo $] \, 0.397 \, , \, 0.435 \, [$ (as pequenas diferenças na séptima casa decimal resultam de erros de arredondamento nas contas acima), e que a respectiva ordenada na origem β_0 pertence ao intervalo $] \, -0.442 \, , \, -0.285 \, [$.

O *nível de confiança* pode ser mudado com o argumento `level`. Eis os intervalos a 90% de confiança:

```
> confint(iris.lm, level=0.90)
                   5 %       95 %
(Intercept)  -0.4288901 -0.2972609
Petal.Length  0.3998944  0.4316164
```

Nota 2.8

Um alerta sobre Intervalos de Confiança. Tal como na construção de outros intervalos de confiança, existem aqui duas *facetas contrastantes*:

- o *grau de confiança* em como os intervalos contêm os verdadeiros valores de β_0 ou β_1; e
- a *precisão* (amplitude) dos intervalos.

Dado um conjunto de observações, *quanto maior o grau de confiança* $(1-\alpha) \times 100\%$ *associado a um intervalo, maior será a sua amplitude, isto é, menor será a sua precisão.* Conversamente, para uma mesma amostra, um intervalo com maior precisão, ou seja, um intervalo de menor amplitude, significa um intervalo com menor grau de confiança associado.

Na próxima Secção, os mesmos resultados que serviram de base à construção dos intervalos de confiança serão usados para outro fim: efectuar testes de hipóteses a valores dos parâmetros β_0 e β_1.

2.7 Testes de hipóteses para os parâmetros da recta populacional

Vejamos agora como a Proposição 2.13, que serviu de ponto de partida para a construção dos intervalos de confiança para β_0 e β_1, é também o ponto de partida para a construção de testes de hipóteses sobre esses parâmetros da recta de regressão populacional. De facto, as quantidades indicadas nessa Proposição podem ser usadas como estatísticas de teste.

2.7.1 Breve revisão de Testes de Hipóteses

Na UC Estatística dos primeiros ciclos do ISA estudam-se Testes de Hipóteses para os seguintes indicadores quantitativos de populações:

- média μ duma população;
- variância σ^2 duma população;
- comparação de médias de duas populações $(\mu_1 - \mu_2)$;
- comparação de variâncias de duas populações $(\frac{\sigma_1^2}{\sigma_2^2})$.

No nosso contexto actual, estamos interessados em inferência sobre a recta de regressão populacional correspondente ao Modelo Linear, de equação $y = \beta_0 + \beta_1 x$. Em particular, estaremos interessados em efectuar testes de hipóteses sobre cada um dos parâmetros β_0 e β_1 da recta.

As hipóteses dizem respeito aos valores que esses parâmetros tomam, *na população*. Pretende optar-se entre *hipóteses alternativas* com base na informação proveniente duma amostra aleatória dessa população.

Num teste de Hipóteses há *cinco passos* a seguir.

1. No **primeiro passo**, *formulam-se hipóteses alternativas* em confronto. Concretamente, define-se uma *Hipótese Nula* H_0 que será confrontada com uma *Hipótese Alternativa* H_1. Por exemplo, pode tomar-se:
   - Hipótese Nula H_0: $\beta_1 = 1$
   - Hipótese Alternativa H_1: $\beta_1 \neq 1$.
2. O **segundo passo** consiste em identificar uma *estatística de teste*, que servirá para optar entre H_0 e H_1. Uma estatística de teste é:
   - uma quantidade *numérica*, cujo valor *depende apenas da amostra e de* H_0;
   - *com distribuição de probabilidades conhecida, se* H_0 *verdade*; e tal que
   - para alguns valores da estatística se considera a Hipótese Nula pouco plausível, procedendo-se a *rejeitar* H_0.

   Por exemplo, a quantidade que serviu de ponto de partida para a construção dos Intervalos de Confiança para o declive β_1 pode servir como estatística de teste para hipóteses sobre β_1. Exemplificando com as hipóteses acima indicadas, considere-se essa quantidade, e substitua-se o valor de

β_1 pelo seu valor ao abrigo de H_0, ou seja, tome-se $\beta_1 = 1$. Se esse fôr o verdadeiro valor de β_1, sabemos que a distribuição de probabilidades dessa quantidade será uma t_{n-2}:

$$T = \frac{\hat{\beta}_1 - \overbrace{\beta_{1_{|H_0}}}^{=1}}{\hat{\sigma}_{\hat{\beta}_1}} \quad \frown \quad t_{n-2} \ , \quad \text{se H}_0 \text{ verdade} \tag{2.36}$$

Dada uma amostra em que o estimador $\hat{\beta}_1$ tome um valor $b_1 \approx \beta_{1_{|H_0}} = 1$, o valor calculado da estatística T será $T_{calc} \approx 0$. Neste caso, a informação proveniente da amostra não suscita razões para duvidar de H_0. Pelo contrário, se o valor amostral b_1 fôr muito diferente de $\beta_{1_{|H_0}} = 1$, a hipótese nula é mais duvidosa. Nesse caso, o valor calculado da estatística T será (em módulo) grande (quer porque $b_1 \gg 1$, quer porque $b_1 \ll 1$). Assim, *o hipotético valor $\beta_{1_{|H_0}}$ é plausível se T_{calc} fôr próximo de zero. Quanto maior seja $|T_{calc}|$, menos plausível será* $H_0 : \beta_1 = 1$.

Mas como definir a fronteira entre os valores da estatística que levam à rejeição, ou não, de H_0? Para definir essa fronteira segue-se a abordagem de Neyman-Pearson. Há que distinguir entre:

- a *realidade* (H_0 ou H_1) que não conhecemos e não controlamos; e
- a *decisão* (H_0 ou H_1) que iremos tomar e que podemos controlar.

Do cruzamento das duas realidades possíveis com as duas decisões possíveis resultam *quatro possíveis situações*, duas das quais correspondem a decisões correctas e duas a decisões erradas:

Realidade	Decisão	
	Admitir H_0	Rejeitar H_0 (optar por H_1)
H_0 verdade	Certo	Erro (Tipo I)
H_0 falso (H_1 verdade)	Erro (Tipo II)	Certo

Ora, *não é possível reduzir simultaneamente a probabilidade dos dois erros*, uma vez que para diminuir P[Erro Tipo I]=P[Rejeitar $H_0 \,|\, H_0$ verdade], será necessário reduzir a gama de valores que levam à rejeição de H_0 (ser mais cauteloso na rejeição). Mas a diminuição da gama de valores de T_{calc} que leva à rejeição de H_0 significa que, caso H_0 seja falso, estar-se-á mais vezes a cometer o erro de não rejeição. Assim, diminuir P[Erro Tipo I] obtém-se à custa de aumentar P[Erro Tipo II].

Para sair deste dilema, considera-se que *o Erro de Tipo I é o mais grave* e opta-se por controlá-lo (deixando de controlar a probabilidade de cometer o Erro de Tipo II, considerado menos grave).

3. O **terceiro passo** dum Teste de Hipóteses consiste em definir o chamado nível de significância do teste. Designa-se *nível de significância do teste* à probabilidade de cometer o Erro de Tipo I, ou seja, à probabilidade de rejeitar H_0 quando esta hipótese é verdadeira. É habitual representar o nível de significância dum teste pela letra grega α (*alfa*):

$$\alpha \quad = \quad P[\text{ Erro de Tipo I }] \quad = \quad P[\text{ Rejeitar H}_0 \mid \text{H}_0\text{verdade }] \ .$$

Ao fixar-se o valor do nível de significância α, está-se a definir a probabilidade associada a uma gama de possíveis valores da estatística T aos quais se associa a *rejeição de* H_0, a chamada *Região Crítica* ou *Região de Rejeição* de H_0.

Assim, no exemplo acima considerado, a resposta à pergunta de como definir a fronteira entre valores de T_{calc} considerados próximos de zero e que não justificam a rejeição de H_0, e os valores de $|T_{calc}|$ que sejam demasiado grandes para se acreditar em H_0, é dada pela escolha, para Região Crítica, duma região de valores mais extremos de T_{calc}, à qual corresponda a probabilidade α.

Sendo α a probabilidade de cometer um erro, queremos que o valor de α seja pequeno. A escolha desse valor é feita pelo experimentador, e depende da gravidade das consequências de cometer o Erro de Tipo I. Em aplicações em que esteja em causa uma decisão de vida ou morte (como é frequente em muitas aplicações no campo da Medicina), é aconselhável escolher valores muito pequenos de α. Mas na generalidade das aplicações associadas aos cursos do ISA, podemos ser um pouco mais tolerantes. É habitual a utilização de valores como $\alpha = 0.05$ ou $\alpha = 0.01$.

Deve salientar-se que a abordagem de Neyman-Pearson, associada a fixar a probabilidade de rejeitar H_0 quando essa hipótese é verdade, implica que o papel das duas hipóteses em confronto deixa de ser simétrico. A decisão de controlar α =P[Erro Tipo I] implica que, na dúvida, preferimos admitir que H_0 é admissível. É habitual usar terminologia jurídica para ilustrar a situação, que tem semelhanças evidentes com o princípio de considerar o réu (o acusado num julgamento) inocente até prova em contrário. Assim, podemos afirmar que:

- a Hipótese Nula H_0 tem o *benefício da dúvida*; enquanto
- a Hipótese Alternativa H_1 tem o *ónus da prova*.

4. O **quarto passo** num Teste de Hipóteses consiste em definir a Região Crítica (ou de Rejeição), ou seja, o conjunto de valores possíveis da estatística ao qual associamos a *rejeição de* H_0 e que é constituída pelos valores "menos plausíveis", caso H_0 seja verdade. Como vimos, a Região Crítica tem de ser uma *região de probabilidade* α*, se fôr verdade* H_0.

   Dependendo do teste concreto em questão e da natureza da Hipótese Alternativa H_1, uma Região Crítica pode ser *bilateral* ou *unilateral*. No exemplo que temos vindo a considerar, faz sentido uma Região Crítica *bilateral* (associada a valores T_{calc} que, em módulo, sejam grandes uma vez que esses valores estão associados, quer a estimativas $b_1 \ll 1 = \beta_{1_{|H_0}}$, quer a estimativas $b_1 \gg 1 = \beta_{1_{|H_0}}$). Esta situação é ilustrada na Figura seguinte.

5. O **quinto e último passo** dum Teste de Hipóteses consiste em calcular o valor da estatística de teste para a nossa amostra e, com base nesse valor, retirar as conclusões do teste.

   É o único passo dum Teste de Hipóteses que exige a existência de dados. Os quatro passos anteriores poderiam ser formulados antes de realizada a extracção da amostra aleatória.

Deve salientar-se que os passos 3 a 5 podem ser substituídos pela indicação duma medida de plausibilidade de H_0, designada *valor de prova* ou *p-value*. Este é definido como sendo *a probabilidade de obter um valor tão ou mais extremo quanto o observado na estatística do teste, caso seja verdade* H_0.

Quando um *p-value* é muito pequeno, considera-se H_0 irrealista, optando-se pela sua rejeição.

Nas Subsecções seguintes resumem-se os passos associados a um Teste de Hipóteses para os valores de cada um dos parâmetros duma recta de regressão populacional, e veremos como definir o valor de prova associado a um dado teste.

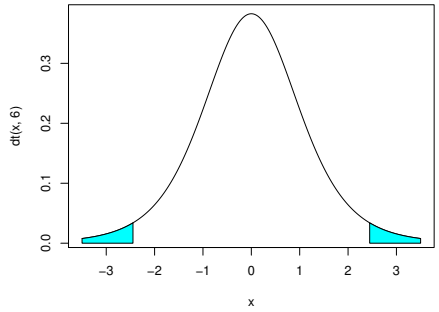


Figura 2.19: Uma Região Crítica bilateral, adequada a um teste a Hipóteses do Tipo $H_0 : \beta_1 = 1$ contra $H_1 : \beta_1 \neq 1$. A densidade é duma distribuição t-Student com $n-2$ graus de liberdade que, como se viu em (2.36), é adequada à distribuição da estatística T para testes a β_1, caso H_0 seja verdade. As regiões indicadas correspondem a probabilidades $\frac{\alpha}{2}$, pelo que, em conjunto, correspondem a uma região de probabilidade α na distribuição correspondente a H_0 verdade. Se para uma dada amostra o valor da estatística calculada recair nessa região, deve rejeitar-se H_0. Assim, deve rejeitar H_0 se $T_{calc} > t_{\frac{\alpha}{2}(n-2)}$ ou $T_{calc} < -t_{\frac{\alpha}{2}(n-2)}$ (os valores que deixam, à sua direita e à sua esquerda, respectivamente, áreas $\frac{\alpha}{2}$ na distribuição t_{n-2}). Esta dupla condição pode ser simplificada usando o valor absoluto de T_{calc}: deve rejeitar-se H_0 se $|T_{calc}| > t_{\frac{\alpha}{2}(n-2)}$.

2.7.2 Testes de hipóteses sobre o declive β_1

Nesta Subsecção resumimos o que já foi visto relativo aos testes de hipóteses para o declive da recta de regressão populacional, β_1. Serão considerados em separado os diferentes tipos possíveis de região crítica, associadas a diferentes tipos de hipóteses.

2.7.2.1 Testes de hipóteses com Região Crítica bilateral

Comecemos por considerar um teste de hipóteses a β_1, com hipóteses do tipo $H_0 : \beta_1 = c$ e $H_1 : \beta_1 \neq c$, a que corresponde uma região crítica bilateral. Sendo válido o Modelo de Regressão Linear Simples, tem-se:

Hipóteses: $H_0 : \ \beta_1 = c \quad$ vs. $\quad H_1 : \ \beta_1 \neq c.$

Estatística do Teste: $T \ = \ \frac{\hat{\beta}_1 - \overbrace{\beta_1|_{H_0}}^{=\,c}}{\hat{\sigma}_{\hat{\beta}_1}} \ \frown \ t_{n-2} \quad , \quad$ sob H_0.

Nível de significância do teste: $\alpha \ = \ P[\textit{Rejeitar } H_0 \,|\ H_0 \textit{ verdade}]$

Região Crítica ou de Rejeição: (*Bilateral*)

Calcular $T_{calc} = \frac{b_1 - c}{\hat{\sigma}_{\hat{\beta}_1}}$ e rejeitar H_0 se $|T_{calc}| > t_{\frac{\alpha}{2}(n-2)}$ (ver Figura 2.20).

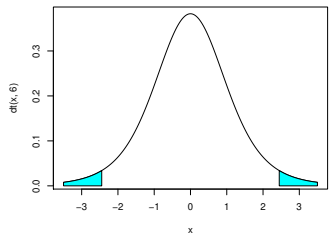


Figura 2.20: Região Crítica bilateral num teste t. A cada cauda da distribuição corresponde uma probabilidade $\frac{\alpha}{2}$, pelo que a Região de Rejeição tem probabilidade global α.

Nota 2.9

Repare-se na natureza da estatística do teste: o *valor da estatística do teste* é a quantidade de erros padrão ($\hat{\sigma}_{\hat{\beta}_1}$) a que o valor estimado (b_1) se encontra do valor de β_1 sob H_0 (c).

2.7.2.2 Teste de Hipóteses com Região Crítica Unilateral direita

Ainda relativamente a β_1 e com a mesma estatística de teste, hipóteses de tipo diferente geram regiões críticas diferentes.

Consideremos o caso em que a hipótese alternativa H_1 é da forma $H_1 : \beta_1 > c$. Neste caso, já não faz sentido incluir na região crítica a cauda esquerda da distribuição t-Student. De facto, essa cauda esquerda corresponde aos valores muito negativos de T_{calc}, valores que surgem quando o numerador da estatística de teste é muito negativo, ou seja, quando $b_1 \ll c$. Mas se o declive da recta amostral (que estima o declive β_1 da recta populacional) é muito inferior a c, não há qualquer razão para optar pela hipótese alternativa $H_1 : \beta_1 > c$. Apenas os valores na cauda direita da distribuição, que resultam de na amostra se ter $b_1 \gg c$, apontam para a veracidade de H_1. Assim, a esse tipo de hipóteses corresponde uma região crítica *unilateral direita*. Vejamos em pormenor:

Hipóteses: $H_0 : \ \beta_1 \leq c \quad$ vs. $\quad H_1 : \beta_1 > c.$

Estatística do Teste: $T \ = \ \frac{\hat{\beta}_1 - \overbrace{\beta_1|_{H_0}}^{=c}}{\hat{\sigma}_{\hat{\beta}_1}} \ \frown \ t_{n-2} \quad , \quad$ se H_0 verdade.

Nível de significância do teste: $\alpha \ = \ P[Rejeitar\ H_0 \,|\ H_0\ verdade]$

Região Crítica: (Unilateral direita)

Rejeitar H_0 se $\ T_{calc} > t_{\alpha(n-2)}$ (ver Figura 2.21)

NOTA: A hipótese nula poderia escrever-se apenas $H_0 : \beta_1 = c$, mas a fim de manter a ideia que H_0 e H_1 são hipóteses complementares, opta-se por escrever a hipótese nula como $H_0 : \beta_1 \leq c$. Assim, ao se

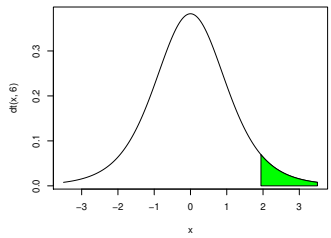


Figura 2.21: Região crítica unilateral direita num teste t. A área da região assinalada é α.

substituir, na estatística de teste, o valor de β_1 sob a hipótese nula, escolhe-se o valor fronteira, c, que corresponde ao valor de H_0 que seria mais difícil de distinguir de H_1. Se para esse valor fronteira o nível de significância α estiver garantido, a probabilidade de cometer o Erro de Tipo I para outro qualquer valor de β_1 em H_0 (ou seja, para valores $\beta_1 < c$) estará igualmente garantido. Por esta razão, o *valor fronteira entre H_0 e H_1 (o valor c) tem de pertencer a H_0*. Em termos práticos, isso significa que a desigualdade em H_1 tem de ser estrita.

2.7.2.3 Teste de Hipóteses com Região Crítica Unilateral esquerda

Trocando o sentido das desigualdades nas hipóteses, justifica-se uma Região Crítica unilateral esquerda, associada a declives amostrais $b_1 \ll c$, consentâneos com a Hipótese alternativa $H_1 : \beta_1 < c$:

Hipóteses: $H_0 : \beta_1 \geq c \quad \text{vs.} \quad H_1 : \beta_1 < c.$

Estatística do Teste: $T = \frac{\hat{\beta}_1 - \overbrace{\beta_1|_{H_0}}^{=c}}{\hat{\sigma}_{\hat{\beta}_1}} \frown t_{n-2} \quad , \quad \text{sob } H_0.$

Nível de significância do teste: $\alpha = P[\textit{Rejeitar } H_0 \mid H_0 \textit{ verdade}]$

Região Crítica: (Unilateral esquerda) *Rejeitar H_0 se* $T_{calc} < -t_{\alpha(n-2)}$

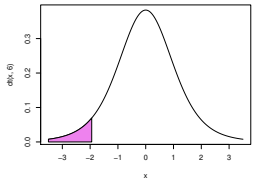


Figura 2.22: Região crítica unilateral esquerda num teste t. A região assinalada tem probabilidade α.

2.7.3 Testes de hipóteses para a ordenada na origem β_0

Vejamos agora, mais resumidamente, os testes de hipóteses à ordenada na origem da recta populacional, β_0. Sendo válido o Modelo de Regressão Linear Simples, tem-se:

Hipóteses: $H_0 : \beta_0 \begin{matrix}(\geq)\\ =\\ (\leq)\end{matrix} c$ vs. $H_1 : \beta_0 \begin{matrix}(<)\\ \neq\\ (>)\end{matrix} c$

Estatística do Teste: $T = \frac{\hat{\beta}_0 - \overbrace{\beta_0|_{H_0}}^{=c}}{\hat{\sigma}_{\hat{\beta}_0}} \frown t_{n-2}$, sob H_0.

Nível de significância do teste: $\alpha = P[Rejeitar\ H_0 \,|\, H_0\ verdade]$

Região Crítica (Região de Rejeição): Rejeitar H_0 se $T_{calc} = \frac{b_0 - c}{\hat{\sigma}_{\hat{\beta}_0}}$ verifica:

$T_{calc} < -t_{\alpha(n-2)}$	(Unilateral esquerdo)
$\lvert T_{calc}\rvert > t_{\alpha/2(n-2)}$	(Bilateral)
$T_{calc} > t_{\alpha(n-2)}$	(Unilateral direito)

2.7.4 Testes usando valores de prova (*p-values*)

Nos testes de hipóteses, quer a β_1, quer a β_0, é possível, em alternativa a fixar previamente o nível de significância α, indicar o valor de prova (*p-value*) associado ao valor calculado da estatística T.

Definição 2.13: Valor de prova (*p-value*) duma estatística de teste

O *p-value* associado ao valor calculado T_{calc} duma estatística de teste define-se como a probabilidade da estatística de teste tomar, sob H_0, valores mais extremos que T_{calc}.

O conceito de "*valor mais extremo*" depende da natureza da Região Crítica associada ao teste. Por exemplo, para um teste a β_1 ou β_0 com Região Crítica unilateral direita, onde são os valores grandes da estatística que indiciam a necessidade de rejeitar H_0, o *p-value* é calculado como a probabilidade de, na distribuição t_{n-2}, se ter um valor *maior* do que T_{calc}: $p = P[t_{n-2} > T_{calc}]$.

Assim, a forma de calcular os *p-values* difere consoante a natureza das hipóteses nula e alternativa conduza a regiões de rejeição unilaterais ou bilaterais e é indicada na Tabela 2.2

Teste Unilateral direito	$p = P[t_{n-2} > T_{calc}]$
Teste Unilateral esquerdo	$p = P[t_{n-2} < T_{calc}]$
Teste Bilateral	$p = 2\,P[t_{n-2} > \lvert T_{calc}\rvert]$.

Tabela 2.2: A forma de calcular os valores de prova (*p-values*) para os diferentes tipos de testes aos parâmetros β_0 e β_1.

Convém salientar a relação existente entre um valor de prova p e o nível de significância do teste, α. De facto, da definição do *p-value* decorre que quando $p < \alpha$, necessariamente o valor calculado da estatística pertence à Região Crítica, pelo que se rejeitaria H_0 com o nível de significância indicado. Conversamente, quando $p > \alpha$, necessariamente T_{calc} não pertence à Região Crítica, pelo que não se rejeita H_0. A situação é ilustrada na Figura 2.23.

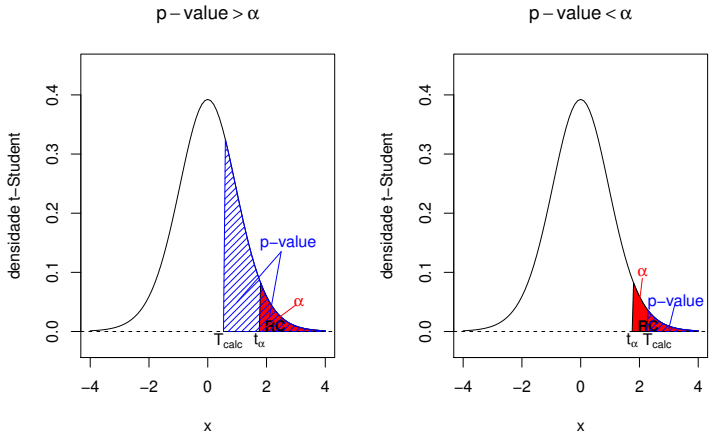


Figura 2.23: A relação entre os valores de prova (*p-values*) e os níveis de significância α, ilustrada para o caso duma Região Crítica unilateral direita.

Assim, em geral verifica-se:

- $p-value > \alpha \Rightarrow$ não rejeição de H_0 ao nível α;
- $p-value < \alpha \Rightarrow$ rejeição de H_0 ao nível α;

2.7.5 Testes de hipóteses no R: de novo o exemplo dos lírios

No R, a função `summary`, aplicada ao resultado dum comando `lm` produz a informação essencial para testes de hipóteses a β_0 e β_1. Na tabela resultante, as colunas têm a seguinte informação:

Estimate As *estimativas* b_0 *e* b_1

Std.Error As estimativas dos *erros padrões* $\hat{\sigma}_{\hat{\beta}_0}$ *e* $\hat{\sigma}_{\hat{\beta}_1}$

t value O *valor calculado das estatísticas dos testes às hipóteses*

$$H_0 : \beta_0(\beta_1) = 0 \quad vs. \quad H_1 : \beta_0(\beta_1) \neq 0 \ ,$$

ou seja,

$$T_{calc} = b_0/\hat{\sigma}_{\hat{\beta}_0} \qquad \text{e} \qquad T_{calc} = b_1/\hat{\sigma}_{\hat{\beta}_1}$$

Pr(>|t|) O *valor de prova* (*p-value*) associado ao valor de T_{calc} nos testes (bilaterais) acima referidos.

Atenção: Nas duas colunas finais, associadas aos testes t, os valores *apenas dizem respeito a testes ao valor* ***nulo*** *dos parâmetros.* Caso se pretenda efectuar testes a $\beta_1 = c \neq 0$, será necessário calcular o valor de T_{calc} (e o correspondente *p-value*) aparte.

Relembremos os resultados produzidos pelo comando `summary`, aplicado à regressão linear simples no exemplo dos lírios (Subsecção 2.6.2):

```
>  summary(iris.lm)
Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)  -0.363076   0.039762  -9.131  4.7e-16 ***
Petal.Length  0.415755   0.009582  43.387  < 2e-16 ***
```

Como se pode constatar na linha final da tabela, num teste a $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$, a estatística de teste tem valor calculado

$$T_{calc} \;=\; \frac{b_1 - \overbrace{\beta_1|_{H_0}}^{=\ 0}}{\hat{\sigma}_{\hat{\beta}_1}} \;=\; \frac{0.415755}{0.009582} \;=\; 43.387\ .$$

O respectivo valor de prova (*p-value*) é inferior à precisão da máquina ($< 2 \times 10^{-16}$), indicando uma claríssima rejeição da hipótese nula.

Para testes a valores diferentes de zero dos parâmetros β_j, será preciso completar os cálculos do valor da estatística. Por exemplo, num teste com as hipóteses $H_0 : \beta_1 = 0.5$ vs. $H_1 : \beta_1 \neq 0.5$, o valor da estatística é:

$$T_{calc} \;=\; \frac{b_1 - \overbrace{\beta_1|_{H_0}}^{=\ 0.5}}{\hat{\sigma}_{\hat{\beta}_1}} \;=\; \frac{0.415755\text{-}0.5}{0.009582} \;=\; -8.792006\ .$$

O *valor de prova* (bilateral) associado a T_{calc} calcula-se como indicado na Subsecção 2.7.4:

$$p = 2 \times P\left[t_{n-2} > |-8.792006|\right]\ .$$

Com o auxílio do `R`, calcula-se o valor desse *p-value*:

```
> 2*(1-pt(8.792006,148))
[1] 3.552714e-15
```

A claríssima rejeição de H_0 não surpreende: a estimativa $b_1 = 0.4158$ está a uma distância de $\beta_1 = 0.5$ superior a 8 vezes o erro padrão estimado $\hat{\sigma}_{\hat{\beta}_1}$.

2.8 Inferência sobre valores da variável resposta

Em muitos contextos, o interesse maior não reside no valor dos parâmetros individuais da recta populacional, mas sim nos valores da variável resposta Y correspondentes a algum valor dado do preditor,

$X = x$. Há dois aspectos diferentes em que podemos estar interessados. Por um lado, podemos querer inferir sobre o *valor esperado* (valor médio) de Y, correspondente a um dado valor do preditor, x. Ou seja, podemos querer calcular um intervalo de confiança, ou testar hipóteses, relativos ao valor de Y que, na recta populacional, corresponde ao valor $X = x$:

$$E[\, Y \mid X = x \,] = \beta_0 + \beta_1 \, x \tag{2.37}$$

A resposta a esse tipo de questão envolve a consideração simultânea da incerteza associada à estimação dos dois parâmetros β_0 e β_1 e será estudada na Subsecção 2.8.1.

Mas este tipo de problema não captura a *variabilidade de observações individuais* em torno da recta populacional. Para efectuar *inferência sobre valores individuais de* Y, dado o valor $X = x$ do preditor, será necessário ter em conta essa variabilidade suplementar. Esse tipo de problemas será considerado na Subsecção 2.8.2.

2.8.1 Inferência sobre o valor esperado de Y, dado $X = x$

Consideremos primeiro a *inferência sobre o valor esperado da variável resposta* Y*, dado um valor* x *da variável preditora*, ou seja, sobre o valor de Y *na recta populacional*, quando $X = x$, valor esse que é representado por $E[\, Y \,|\, X = x \,]$ ou, de forma mais sintética, por $\mu_{Y|x}$:

$$\mu_{Y|x} \;=\; E[\, Y \,|\, X = x \,] \;=\; \beta_0 \,+\, \beta_1 \, x \;.$$

O estimador óbvio desta quantidade é

$$\hat{\mu}_{Y|x} \quad = \quad \hat{\beta}_0 \,+\, \hat{\beta}_1 \, x \quad = \quad \sum_{i=1}^{n} d_i \, Y_i + \sum_{i=1}^{n} (c_i \, Y_i) \; x \quad = \quad \sum_{i=1}^{n} (d_i + c_i \, x) \, Y_i \;,$$

usando a notação introduzida na Definição 2.11.

Este estimador $\hat{\mu}_{Y|x}$ é, tal como os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$, uma *combinação linear das observações* Y_i que são v.a.s Normais e independentes, ao abrigo do Modelo Linear. Logo, $\hat{\mu}_{Y|x}$ tem distribuição Normal.

Proposição 2.15: Distribuição de $\hat{\mu}_{Y|x}$

Dado o Modelo de Regressão Linear Simples, tem-se

$$\hat{\mu}_{Y|x} \;=\; \hat{\beta}_0 \,+\, \hat{\beta}_1 \, x \quad \frown \quad \mathcal{N}\left(\beta_0 + \beta_1 \, x \;,\; \sigma^2 \left[\frac{1}{n} + \frac{(x - \overline{x})^2}{(n-1)\, s_x^2} \right] \right)$$

$$\Leftrightarrow \qquad \frac{\hat{\mu}_{Y|x} \;-\; \mu_{Y|x}}{\sigma_{\hat{\mu}_{Y|x}}} \quad \frown \quad \mathcal{N}(0,1) \;,$$

onde $\mu_{Y|x} = \beta_0 + \beta_1 \, x$ e $\sigma_{\hat{\mu}_{Y|x}} = \sqrt{\sigma^2 \left[\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)\, s_x^2} \right]}$.

Demonstração 2.15: Proposição 2.15

Como $\hat{\mu}_{Y|x} = \sum_{i=1}^{n} (d_i + c_i\, x)\, Y_i$ é uma combinação linear das observações Y_i que, ao abrigo do Modelo Linear são Normais e independentes, está garantida a distribuição Normal de $\hat{\mu}_{Y|x}$. Falta calcular os seus dois parâmetros que são (como para qualquer Normal), a respectiva média e variância. É imediato, tendo em conta o facto de $\hat{\beta}_0$ e $\hat{\beta}_1$ serem estimadores centrados e as propriedades do valor esperado, que

$$E[\hat{\mu}_{Y|x}] = E[\hat{\beta}_0 + \hat{\beta}_1\, x] = E[\hat{\beta}_0] + E[\hat{\beta}_1]\, x = \beta_0 + \beta_1\, x = \mu_{Y|x}\ .$$

Logo, $\hat{\mu}_{Y|x}$ também é um estimador centrado de $\mu_{Y|x}$. O cálculo da variância de $\hat{\mu}_{Y|x}$ é feito de forma análoga, tendo em atenção as propriedades operatórias da variância e ainda o resultado auxiliar demonstrado no Exercício RLS 13:

$$\begin{aligned} V[\hat{\mu}_{Y|_x}] &= V[\hat{\beta}_0 + \hat{\beta}_1\, x] = V[\hat{\beta}_0] + V[\hat{\beta}_1\, x] + 2Cov(\hat{\beta}_0, \hat{\beta}_1\, x) = V[\hat{\beta}_0] + x^2\, V[\hat{\beta}_1] + 2x\, Cov(\hat{\beta}_0, \hat{\beta}_1) \\ &= \underbrace{\sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}\right]}_{=V[\hat{\beta}_0]} + x^2 \cdot \underbrace{\frac{\sigma^2}{(n-1)s_x^2}}_{=V[\hat{\beta}_1]} + 2x \cdot \underbrace{\frac{-\sigma^2 \overline{x}}{(n-1)s_x^2}}_{=Cov[\hat{\beta}_0,\hat{\beta}_1]\ (Ex.13)} \\ &= \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2 + x^2 - 2\overline{x}\, x}{(n-1)\, s_x^2}\right] = \sigma^2 \left[\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)\, s_x^2}\right]\ . \end{aligned}$$

Tal como para as distribuições iniciais de $\hat{\beta}_0$ e $\hat{\beta}_1$ (Proposição 2.11 e Corolário 2.1), também este resultado não é ainda utilizável devido à presença da variância (desconhecida) dos erros aleatórios, σ^2. Mas esse problema pode ser resolvido de forma análoga ao que foi feito aquando do estudo dos estimadores dos parâmetros individuais, como se verá seguidamente.

Proposição 2.16: Distribuição de $\hat{\mu}_{Y|x}$, sem quantidades desconhecidas

Dado o Modelo de Regressão Linear Simples, tem-se

$$\frac{\hat{\mu}_{Y|x} - \mu_{Y|x}}{\hat{\sigma}_{\hat{\mu}_{Y|x}}} \quad \frown \quad t_{n-2}\ ,$$

onde $\hat{\sigma}_{\hat{\mu}_{Y|x}} = \sqrt{QMRE\left[\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)\, s_x^2}\right]}$.

NOTA: A justificação desta distribuição é totalmente análoga à demonstração das distribuições de $\hat{\beta}_1$ e $\hat{\beta}_0$, feita após a Proposição 2.13.

A Proposição 2.16 fornece o resultado que está na base de intervalos de confianças e/ou testes de hipóteses para $\mu_{Y|x} = E[Y|X=x] = \beta_0 + \beta_1 x$. Comecemos por ver os intervalos de confiança para $\mu_{Y|x}$.

Proposição 2.17: Intervalo de Confiança para $\mu_{Y|x} = \beta_0 + \beta_1 x$

Dado o Modelo RLS, um intervalo a $(1-\alpha)\times 100\%$ de confiança para o valor esperado de Y, dado o valor $X=x$ da variável preditora (ou seja, para $\mu_{Y|x} = E[Y|X=x] = \beta_0 + \beta_1 x$), é dado por:

$$\left] \hat{\mu}_{Y|x} - t_{\frac{\alpha}{2}(n-2)} \cdot \hat{\sigma}_{\hat{\mu}_{Y|x}} \quad , \quad \hat{\mu}_{Y|x} + t_{\frac{\alpha}{2}(n-2)} \cdot \hat{\sigma}_{\hat{\mu}_{Y|x}} \right[,$$

com $\hat{\mu}_{Y|x} = b_0 + b_1\, x$ e $\hat{\sigma}_{\hat{\mu}_{Y|x}} = \sqrt{QMRE\left[\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)\, s_x^2}\right]}$.

A dedução deste intervalo de confiança é inteiramente análoga à efectuada para construir um intervalo de confiança para β_1, partindo da Proposição 2.13, uma vez que a estrutura das quantidades que servem de ponto de partida é idêntica (nos dois casos, a variável aleatória com distribuição *t-Student* é a razão da diferença entre estimador e parâmetro estimado, a dividir pelo erro padrão).

A amplitude do intervalo de confiança *aumenta com $QMRE$ e com a distância de x a $\overline{x}$* e *diminui com n e s_x^2*. Assim, *a estimação de $\mu_{Y|x}$ é melhor para valores de x próximos de $\overline{x}$*, no sentido em que, para igual grau de confiança $(1-\alpha)\times 100\%$, os intervalos de confiança são de menor amplitude para valores de x próximos de $\overline{x}$.

2.8.1.1 Inferência sobre $\mu_{Y|x}$ no R

Valores estimados e intervalos de confiança para $\mu_{Y|x} = E[Y|X=x]$ obtêm-se no R com a função `predict`. Os novos valores da variável preditora são dados através do argument `new`, numa `data frame` onde a variável tem o mesmo nome que no ajustamento inicial. Por exemplo, no exemplo dos lírios, a largura esperada de pétalas de comprimento $x=1.85$ e de comprimento $x=4.65$, é dada por:

```
>  predict(iris.lm, new=data.frame(Petal.Length=c(1.85,4.65)))
       1        2
0.406072 1.570187
```

A omissão do argumento `new` produziria os valores ajustados de y (os $\hat{y}_i$) *associados aos valores x_i usados aquando do ajustamento* da recta de regressão (também gerados pelo comando `fitted`).

Um *intervalo de confiança* obtém-se usando, no comando `predict`, o argumento `int="conf"`. A fronteira inicial do intervalo é indicada debaixo de `lwr` (do inglês ***lower** endpoint*) e a fronteira final é indicada por `upr` (de ***upper** endpoint*). Debaixo de `fit` encontra-se o valor ajustado $\hat{\mu}_{Y|x} = b_0 + b_1\, x$, que é o ponto central do intervalo. A representação gráfica deste intervalo de confiança para $\mu_{Y|x}$ é feita na Figura 2.24.

```
> predict(iris.lm,data.frame(Petal.Length=c(4.65)), int="conf")
       fit       lwr       upr
1 1.570187  1.5328338 1.6075405
```

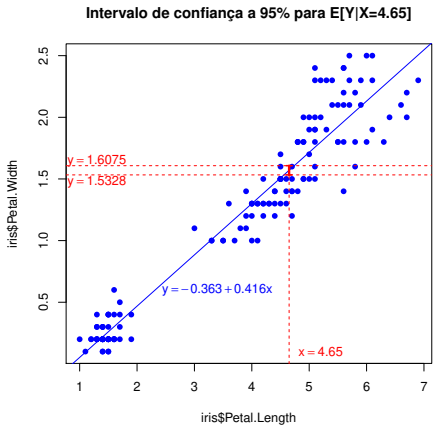


Figura 2.24: Representação gráfica do intervalo de confiança para o valor esperado de Y, dado $X=4.65$, nos dados dos lírios, ou seja, para $\mu_{Y|x} = \beta_0 + \beta_1\, x$, com $x=4.65$. Tratando-se dum intervalo de confiança relativo a valores da variável resposta, deve ser lido no eixo vertical, por cima do correspondente valor x do preditor. Tem-se 95% de confiança em como a recta de regressão *populacional* atravessa aquele intervalo vertical (indicado a vermelho).

2.8.1.2 Bandas de confiança para a recta de regressão

Fazendo variar os valores x do preditor, pode obter-se um número arbitrário de intervalos de confiança para $\mu_{Y|x}$. Unindo os extremos inferiores desses intervalos obtém-se uma *banda inferior* e unindo os extremos superiores obtém-se uma *banda superior*. Essas bandas são curvas que ladeiam a recta estimada e que contêm, com $(1-\alpha) \times 100\%$ de confiança, a verdadeira recta populacional. Na Figura 2.25 são dadas as bandas de confiança para a recta populacional no exemplo dos lírios. Com 95% de confiança, a recta populacional está contida nas bandas de confiança mostradas .

2.8.2 Inferência sobre observações individuais de Y, dado $X=x$

Os intervalos de confiança calculados na Subsecção anterior (2.8.1) dizem respeito ao *valor esperado* de Y, para um dado valor de x. Mas *uma observação individual de* Y tem associada uma *variabilidade adicional*, uma vez que as observações individuais de Y não se encontram (em geral) em cima da recta populacional.

De facto, dado o Modelo RLS, uma observação individual de Y é da forma:

$$Y \quad = \quad \beta_0 + \beta_1 x + \epsilon \quad = \quad \mu_{Y\,|x} \, + \, \epsilon.$$

Ora, $\mu_{Y|x} = E[Y|X=x] = \beta_0 + \beta_1 x$ é estimado por $\hat{\mu}_{Y|x}$, cuja variância é dada por $V[\hat{\mu}_{Y|x}] = \sigma^2 \left[\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)\, s_x^2}\right]$ (Proposição 2.15). Por outro lado, a variância dum erro aleatório ϵ é $V[\epsilon] = \sigma^2$. Toma-se a variância do preditor duma observação individual Y como sendo a soma destas duas expres-

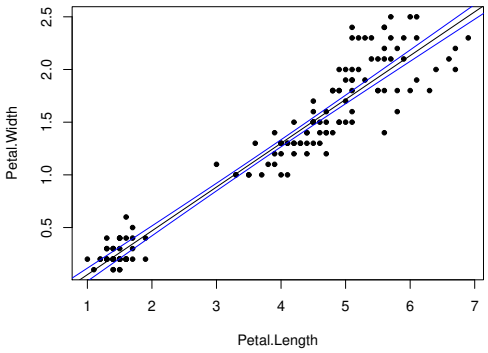


Figura 2.25: Bandas a 95% de confiança para os valores esperados de Y, fazendo variar os valores da variável preditora. Estas curvas contêm, a 95% de confiança, a recta populacional $Y = \beta_0 + \beta_1\, x$. Os intervalos de confiança para $\mu_{Y|x}$ dependem do valor de x. Terão maior amplitude quanto mais afastado x estiver da média $\overline{x}$ das observações, razão pela qual as bandas são *encurvadas*.

sões:

$$\sigma^2_{indiv} \;=\; \sigma^2 \left[\frac{1}{n} + \frac{(x-\overline{x})^2}{_{(n-1)}\, s^2_x}\right] \;+\; \sigma^2 \;=\; \sigma^2 \left[1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{_{(n-1)}\, s^2_x}\right] . \tag{2.38}$$

A variância associada à predição duma observação individual de Y depende da variância desconhecida dos erros aleatórios (σ^2), mas é estimada substituindo, na equação (2.38), σ^2 pelo seu estimador, $QMRE$:

$$\hat{\sigma}^2_{indiv} \;=\; QMRE \left[1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{_{(n-1)}\, s^2_x}\right] . \tag{2.39}$$

Obtém-se um *intervalo de predição para uma observação de Y dado $X = x$* substituindo a variância estimada de $\hat{\mu}_{Y|x}$ do intervalo de confiança para $\mu_{Y|x} = E[Y|X = x]$ (Proposição 2.17), pela variância associada à previsão duma observação individual (equação 2.39).

Proposição 2.18: Intervalo de predição para observação individual de Y

Seja dado o Modelo de Regressão Linear Simples. Um intervalo de predição para uma observação individual de Y, dado o valor $X = x$ do preditor, é dado por:

$$\Big]\quad \hat{\mu}_{Y|x} \;-\; t_{\frac{\alpha}{2}\,(n-2)} \cdot \hat{\sigma}_{indiv} \quad , \quad \hat{\mu}_{Y|x} \;+\; t_{\frac{\alpha}{2}\,(n-2)} \cdot \hat{\sigma}_{indiv} \quad \Big[\,,$$

com $\hat{\mu}_{Y|x} = b_0 + b_1 x$ e $\hat{\sigma}_{indiv} = \sqrt{QMRE \left[1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{_{(n-1)}\, s^2_x}\right]}$.

Estes intervalos de predição[7] são, para um mesmo nível $(1–\alpha)\times 100\%$, necessariamente *de maior amplitude* que os intervalos de confiança para o valor esperado (médio) de Y, $E[Y|X=x]$, vistos antes.

2.8.2.1 Intervalos de predição para Y no R

No R, um intervalo *de predição* para uma observação individual de Y obtém-se utilizando a opção `int="pred"` no comando `predict`. Tal como para o intervalo de confiança para $E[Y]$ (argumento `int="conf"`), esta função do R usa, por omissão, o grau 95%. Como se pode constatar, o intervalo de predição a 95% é] 1.16044, 1.97993 [.

```
> predict(iris.lm,data.frame(Petal.Length=c(4.65)),int="pred")
       fit          lwr       upr
1 1.570187  1.160442632 1.9799317
```

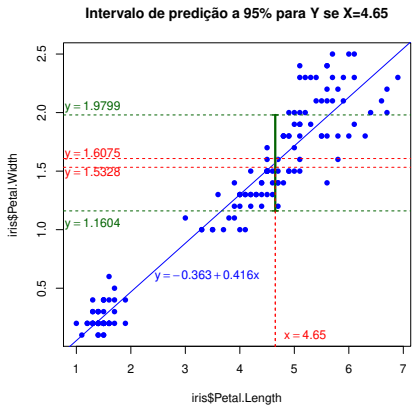


Figura 2.26: Intervalo de predição a 95% (a verde) para uma observação individual de Y, dado o valor $x = 4.65$ do preditor. A vermelho está o intervalo de confiança para $E[Y]$, dado o mesmo valor do preditor, já discutido na Figura 2.24. É visível a muito maior amplitude do intervalo de predição para uma observação individual de Y.

2.8.2.2 Bandas de predição para uma observação de Y

Tal como no caso dos intervalos de confiança para $E[Y|X=x]$, variando os valores de x obtêm-se *bandas de predição* para valores individuais de Y.

No exemplo, 95% dos valores individuais observáveis de Y deverão estar contidos entre as bandas (encurvadas) verdes indicadas na Figura 2.27 (a azul as bandas de confiança para $\mu_{Y|x}$).

[7] A designação intervalo *de predição* sublinha que se trata dum intervalo onde esperamos encontrar uma observação da *variável aleatória* Y (para $X=x$). Esta situação difere da construção de intervalos *de confiança*, que pretendem capturar os valores possíveis de *uma constante* populacional, como é $\mu_{Y|x}$.

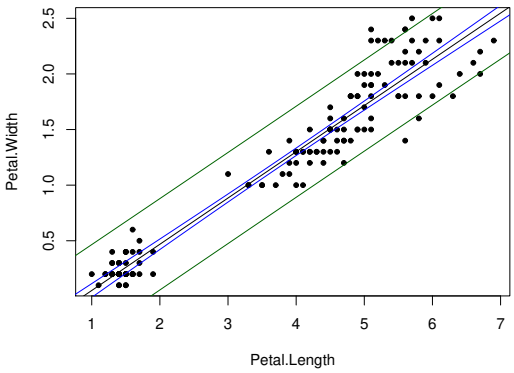


Figura 2.27: Bandas de predição a 95% (a verde) para os valores individuais das larguras das pétalas (Y), dados os comprimentos das pétalas (x), nos lírios. A azul observam-se as bandas de confiança *para o valor esperado* de Y, dado $X=x$, já mostradas na Figura 2.25.

2.9 Teste F à qualidade do ajustamento do Modelo

Em termos meramente descritivos, a qualidade do ajustamento duma regressão linear simples é medida através do *Coeficiente de Determinação,* $R^2 = \frac{SQR}{SQT}$. Num contexto inferencial, é usual também efectuar um *teste de hipótese para avaliar a qualidade global de ajustamento do Modelo.*

Um teste de ajustamento global do modelo tem a *hipótese nula de que o modelo é inútil* para prever Y a partir de X. Designando por $\mathcal{R}^2$ *o coeficiente de determinação populacional*, pode escrever-se:

$$H_0 : \mathcal{R}^2 = 0 \ , \tag{2.40}$$

Tendo em conta que o coeficiente de determinação populacional $\mathcal{R}^2 = \frac{SQR}{SQT}$ tem Somas de Quadrados calculadas para a totalidade da população, $\mathcal{R}^2 = 0$ significa que, na população, $SQR = 0$, ou seja, que os valores ajustados populacionais $\hat{y}_i$ têm variância nula. Isto só é possível se a recta populacional fôr horizontal (em cujo caso todos os valores ajustados serão iguais, não importa qual o valor de x), ou seja, se o declive da recta populacional fôr nulo: $\beta_1 = 0$. Assim, a Hipótese Nula (2.40) é equivalente a $\beta_1 = 0$. Se fôr verdade esta Hipótese Nula, a equação da recta de regressão populacional fica apenas $y = \beta_0$. O Modelo com os pressupostos do Modelo de Regressão Linear, mas cuja equação é apenas $Y = \beta_0 + \epsilon$, designa-se o *Modelo Nulo.* Corresponde a um modelo que é *inútil*, do ponto de vista de usar o preditor X para prever Y, razão pela qual só vale a pena considerar o uso dum modelo de regressão linear simples caso seja possível rejeitar a Hipótese Nula (2.40), ou seja, caso seja possível afirmar que esse modelo é significativamente diferente do Modelo Nulo.

No contexto duma Regressão Linear *Simples*, há duas formas alternativas (mas equivalentes, como se verá adiante) de efectuar um tal teste. A primeira utiliza ferramentas já conhecidas, nomeadamente, um teste t a valores do declive da recta populacional. A segunda utiliza um teste novo, conhecido por *teste F de ajustamento global do modelo.* Esta segunda abordagem, que será estudada nesta Secção, é necessária, uma vez que é a única que se estende ao estudo da Regressão Linear Múltipla.

Assim, numa Regressão Linear Simples pode testar-se a hipótese de o nosso modelo ser diferente do Modelo Nulo, de duas maneiras alternativas:

- Testar $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$, *usando o teste t* considerado na Subsecção 2.7.2.
- Efectuar o *teste F ao ajustamento global do modelo*. Este teste é descrito na Subsecção seguinte.

2.9.1 A estatística F

Ponto de partida natural para um teste à qualidade de ajustamento do Modelo será o de avaliar se SQR (o numerador de R^2) é grande. Veremos que para obter um resultado útil, será necessário comparar os valores de SQR e de QMRE.

Proposição 2.19: Estatística F na Regressão Linear Simples

Seja dado o modelo de Regressão Linear Simples. Recordem-se as definições da Soma de Quadrados da Regressão, SQR, e do Quadrado Médio Residual, $QMRE = \frac{SQRE}{n-2}$. **Se $\boldsymbol{\beta_1 = 0}$**, tem-se:

1. $\frac{SQR}{\sigma^2} \frown \chi^2_1$.
2. $F = \frac{QMR}{QMRE} = (n-2)\,\frac{R^2}{1-R^2} \frown F_{(1,n-2)}$, onde $QMR = \frac{SQR}{1}$ é o Quadrado Médio da Regressão.

Nota: A definição do Quadrado Médio da Regressão, $QMR = \frac{SQR}{1}$ parece estranha, uma vez que QMR e SQR são iguais. No entanto, na Regressão Linear Múltipla o Quadrado Médio da Regressão e a Soma de Quadrados da Regressão deixam de ser iguais.

Demonstração 2.16: Proposição 2.19

1. A partir de resultados e conhecimentos anteriores tem-se:
   - $SQR = \hat{\beta}_1^2\,(n-1)\,s_x^2$ (ver Exercício RLS 5d), adaptando a notação ao contexto inferencial em que agora nos encontramos).
   - Na Subsecção 2.5.1 viu-se que: $\frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{(n-1)\,s_x^2}}} \frown \mathcal{N}(0,1)$.
   - Recorde-se (das disciplinas introdutórias de Estatística) que o quadrado duma variável aleatória com distribuição $\mathcal{N}(0,1)$ tem distribuição Qui-quadrado com um único grau de liberdade, ou seja, que $Z \frown \mathcal{N}(0,1) \Rightarrow Z^2 \frown \chi^2_1$. Logo,
   $$\frac{(\hat{\beta}_1 - \beta_1)^2}{\sigma^2 / \left[(n-1)\,s_x^2\right]} = \frac{(\hat{\beta}_1 - \beta_1)^2\,(n-1)\,s_x^2}{\sigma^2} \frown \chi^2_1 .$$
   - Então, se $\beta_1 = 0$, tem-se: $\frac{SQR}{\sigma^2} \frown \chi^2_1$.
2. A quantidade SQR/σ^2 cuja distribuição agora se conhece, quando $\beta_1 = 0$, não pode ainda ser

usada como estatística dum teste à Hipótese Nula (2.40), uma vez que depende da incógnita σ^2. Mas temos forma de tornear o problema.

- Sabemos (Proposição 2.12) que, dado o Modelo RLS, $\frac{SQRE}{\sigma^2} \frown \chi^2_{n-2}$.
- Sabemos (da disciplina de Estatística introdutória) que as distribuições F surgem a partir da razão de duas variáveis aleatórias Qui-quadrado a dividir pelos seus graus de liberdade, caso sejam independentes. Ou seja, sabemos que:

$$\left.\begin{array}{l} W \frown \chi^2_{\nu_1} \\ V \frown \chi^2_{\nu_2} \\ W, V \text{ independentes} \end{array}\right\} \quad \Rightarrow \quad \frac{W/\nu_1}{V/\nu_2} \frown F_{\nu_1,\nu_2} \ .$$

- É possível mostrar que $SQRE$ e SQR são v.a. independentes (demonstração omitida, por exceder o âmbito da disicplina).
- Logo, $\frac{\frac{SQR}{\cancel{\sigma^2}_1}}{\frac{SQRE}{\cancel{\sigma^2}_{(n-2)}}} = \frac{QMR}{QMRE}$ tem distribuição $F_{(1,n-2)}$, caso seja verdade a hipótese nula $\beta_1 = 0$.

A expressão alternativa da estatística de teste resulta de:

$$\frac{QMR}{QMRE} \ = \ \frac{\frac{SQR}{1}}{\frac{SQRE}{n-2}} \ = \ \frac{n-2}{1}\frac{SQR}{SQRE} \ = \ (n-2)\frac{SQR}{SQT-SQR} \ = \ (n-2)\frac{R^2}{1-R^2} \ ,$$

após, na última passagem, se dividir numerador e denominador por SQT.

A Proposição 2.19 diz-nos qual o comportamento esperado para a estatística $F = \frac{QMR}{QMRE}$, caso seja verdade a hipótese nula $\beta_1 = 0$ que, como se viu, é equivalente à hipótese nula $\mathcal{R}^2 = 0$ (equação 2.40). Mas que tipo de Região Crítica é natural associar a esta estatística? Ou seja, que tipo de valores seria de esperar para a estatística F caso $\beta_1 \neq 0$? Ora, quanto maior fôr $\hat{\beta}_1^2$, mais duvidoso será $H_0 : \beta_1 = 0$. Ao mesmo tempo, maior será $SQR = \hat{\beta}_1^2 \, {}_{(n-1)} \, s_x^2$, pelo que maior será a estatística $F = \frac{QMR}{QMRE}$. Assim, *valores elevados da estatística F sugerem a opção por $H_1 : \beta_1 \neq 0$.* Ou seja, a Região Crítica adequada é uma região *unilateral direita.*

2.9.2 Formulações alternativas do teste F de ajustamento global

Sendo válido o Modelo de Regressão Linear Simples, pode efectuar-se o seguinte teste de hipóteses.

Hipóteses: $H_0 : \ \beta_1 = 0 \quad$ vs. $\quad H_1 : \ \beta_1 \neq 0.$

Estatística do Teste: $F \ = \ \frac{QMR}{QMRE} \ \frown \ F_{(1,n-2)} \quad$ se H_0 verdade.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{\alpha(1,n-2)}$.

Como se viu, podem reformular-se as hipóteses e/ou a estatística do teste, usando *Coeficientes de Determinação*:

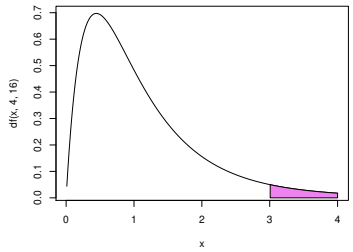


Figura 2.28: Região Crítica (unilateral direita) nos testes F de ajustamento global dum Modelo de Regressão Linear Simples. Note-se que a forma das funções densidade duma distribuição F_{ν_1,ν_2} é, em geral, a indicada nesta Figura, embora para $\nu_1 = 1$ (como no nosso caso), a densidade seja uma curva estritamente decrescente. Opta-se por usar aqui a representação mais genérica das distribuições F.

Hipóteses: $H_0 \ : \ \mathcal{R}^2 = 0 \quad \text{vs.} \quad H_1 \ : \ \mathcal{R}^2 > 0.$

Estatística do Teste: $F \ = \ (n-2)\,\frac{R^2}{1-R^2} \ \frown \ F_{(1,n-2)} \quad \text{se } H_0.$

Nível de significância do teste: α

Região Crítica (Região de Rejeição): (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{\alpha(1,n-2)}$

É fácil de constatar que estatística F é uma função crescente do coeficiente de determinação amostral, R^2, ou seja, quanto maior R^2, maior será o valor calculado de F. Também por essa via, e já que $F = (n-2)\frac{R^2}{1-R^2}$, torna-se natural que a Região de Rejeição de $H_0 : \mathcal{R}^2 = 0$ seja constituída pelos maiores valores de F, isto é, que seja uma Região Crítica unilateral direita.

2.9.3 O teste F no R

A informação essencial para efectuar um teste F ao ajustamento global de um modelo de regressão também se obtém através do comando `summary`, aplicado a uma regressão ajustada através do comando `lm`. Vejamos um exemplo de aplicação, com a regressão relacionando as larguras (y) e comprimentos (x) das pétalas dos lírios. A parte final da listagem produzida pelo comando `summary` é indicada de seguida:

```
> summary(iris.lm)
(...)
Residual standard error: 0.2065 on 148 degrees of freedom
Multiple R-Squared: 0.9271,  Adjusted R-squared: 0.9266
F-statistic:  1882 on 1 and 148 DF,  p-value: < 2.2e-16
```

A informação relevante para o teste F, encontra-se na linha final, nomeadamente:

F-statistic indica o valor calculado da estatística, $F_{calc} = \frac{QMR}{QMRE} = 1882$, e os graus de liberdade (DF, de d*egrees of* f*reedom*, em inglês) na distribuição F correspondente (ou seja, 1 e $n-2=148$).

p-value valor de prova de F_{calc} no teste de ajustamento global do modelo que, tendo em conta a natureza unilateral direita da região crítica, se define como $p = P[F_{1,n-2} > F_{calc}]$. No nosso exemplo, o valor de prova é inferior à precisão de máquina (2.2×10^{-16}), ou seja, é indistinguível de zero, pelo que se tem uma claríssima rejeição de H_0. Assim, rejeita-se de forma enfática a hipótese de que o nosso modelo seja o Modelo Nulo.

Nas penúltima e ante-penúltima linhas dos resultados acima indicados, são também dados os valores de:

Residual Standard error: Estimativa do desvio padrão σ dos erros aleatórios ϵ_i, ou seja, de

$$\hat{\sigma} \;=\; \sqrt{QMRE} \;=\; \sqrt{\frac{SQRE}{n-2}} \qquad [\,=\; 0.2065]$$

Multiple R-squared: O *Coeficiente de Determinação*:

$$R^2 \;=\; \frac{SQR}{SQT} \;=\; \frac{s^2_{\hat{y}}}{s^2_y} \;=\; 1 - \frac{SQRE}{SQT} \qquad [\,=\; 0.9271]$$

Adjusted R-squared: O R^2 *modificado* (que será melhor estudado na Regressão Linear Múltipla):

$$R^2_{mod} \;=\; 1 - \frac{QMRE}{QMT} \;=\; 1 - \frac{\hat{\sigma}^2}{s^2_y}\,, \quad (QMT = SQT/(n-1)) \qquad [\,=\; 0.9266]$$

2.10 Validação do Modelo

Toda a inferência feita até aqui admitiu a validade do Modelo Linear, e em particular, a validade dos pressupostos relativos aos *erros aleatórios*: Normais, de média zero, variância homogénea e independentes. A validade dos intervalos de confiança e testes de hipóteses atrás referidos (incluindo do teste F de ajustamento global) *depende da validade desses pressupostos.*

Uma análise de regressão fica incompleta sem uma *validação destes pressupostos do modelo* (*model checking* ou *validation*, em inglês). Uma vez que os erros aleatórios não são observáveis (mesmo após a extracção da amostra - como se viu na Subsecção 2.5.2), *a validação dos pressupostos relativos aos erros aleatórios faz-se através dos seus preditores, os resíduos.*

2.10.1 A distribuição dos Resíduos no Modelo RLS

Para se estudar a validade dos pressupostos do modelo através dos resíduos, é necessário conhecer o comportamento desses resíduos *quando o Modelo é válido*, o que é enunciado na Proposição 2.20 (ver também o Exercício RLS 22).

Proposição 2.20: Distribuição dos Resíduos no Modelo RLS

Dado o Modelo de Regressão Linear Simples, tem-se:

$$E_i \;\frown\; \mathcal{N}\left(0\,,\,\sigma^2\,(1-h_{ii})\right)\;,\quad \text{onde } h_{ii}=\frac{1}{n}+\frac{(x_i-\overline{x})^2}{{}_{(n-1)}\,s_x^2}\;.$$

Nota: Recordar que o modelo de Regressão Linear Simples admite que os erros aleatórios tenham distribuição $\epsilon_i \;\frown\; \mathcal{N}(0,\sigma^2)$. Esta Proposição indica que os preditores desses mesmos erros aleatórios (os resíduos) têm uma distribuição parecida, mas não idêntica aos erros aleatórios. Sendo igualmente Normais e de média zero, *os resíduos E_i têm variâncias diferentes das dos erros aleatórios e diferentes entre si*: $V[E_i]=\sigma^2(1-h_{ii})$. Assim, os resíduos *não* são identicamente distribuídos. Também *não* são independentes: já se viu no Exercício RLS 5 que a sua soma é zero, pelo que dados $n-1$ resíduos, o último está totalmente especificado, não podendo assim haver independência.

Demonstração 2.17: Proposição 2.20

Um *resíduo* também é uma *combinação linear dos Y_i*:

$$E_i \;=\; Y_i-\hat{Y}_i \;=\; Y_i-(\hat{\beta}_0+\hat{\beta}_1 x_i) \;=\; Y_i-\sum_{j=1}^{n}(d_j+c_j x_i)Y_j \;=\; \sum_{j=1}^{n}k_j Y_j\;,$$

$$\text{com}\quad k_j=\begin{cases}-(d_j+x_i c_j) & \text{se}\quad j\neq i\\ 1-(d_i+x_i c_i) & \text{se}\quad j=i\end{cases}$$

Sendo cada resíduo E_i uma combinação linear de Normais independentes, tem distribuição Normal. Falta determinar os respectivos parâmetros, ou seja, o seu valor esperado e variância. Comecemos pelo valor esperado (e não confundir na expressão seguinte o valor esperado com o resíduo E_i, apesar de ambos serem indicados pela letra E):

$$E[E_i] \;=\; E[Y_i-(\hat{\beta}_0+\hat{\beta}_1\,x_i)] \;=\; E[Y_i]-E[\hat{\beta}_0]-E[\hat{\beta}_1]\,x_i \;=\; (\beta_0+\beta_1\,x_i)-\beta_0-\beta_1\,x_i \;=\; 0\;.$$

A variância de cada resíduo, $V[E_i]$, é calculada no Exercício RLS 22 (ver a respectiva resolução).

2.10.1.1 Diferentes tipos de resíduos

Os resíduos usuais, $E_i = Y_i - \hat{Y}_i$ são quantidades com as unidades de medida da variável resposta Y. Assim, não é possível afirmar se um resíduo é grande ou pequeno, sem ter em conta as dimensões e a variabilidade dos valores de Y. Ora, a Proposição 2.20 permite definir novos tipos de resíduos, sem unidades de medida e cujos valores já sejam mais facilmente identificáveis como sendo, ou não, de magnitude importante. Decorre imediatamente dessa Proposição que, sendo válido o Modelo de Regressão Linear Simples, se verifica:

$$\frac{E_i}{\sqrt{\sigma^2(1-h_{ii})}} \;\frown\; \mathcal{N}(0,1)\;. \tag{2.41}$$

Caso fosse possível calcular estes *resíduos normalizados* com distribuição Normal reduzida, valores fora do intervalo $[-3,3]$ deveriam ser considerados extremos. No entanto, o cálculo dos resíduos na equação

(2.41) é impossível, dada a presença da variância desconhecida dos erros aleatórios, σ^2.

Na literatura dos modelos de regressão linear encontra-se referência a *três variantes de resíduos*, duas das quais correspondem a resíduos *estandardizados*, que aproximam os valores dos resíduos normalizados (eq. 2.41) através de duas diferentes formas de estimar σ^2.

Definição 2.14: Três tipos de resíduos

Resíduos usuais : $E_i = Y_i - \hat{Y}_i$;

Resíduos (internamente) estandardizados : $R_i = \frac{E_i}{\sqrt{QMRE \cdot (1-h_{ii})}}$. Este tipo de resíduos resulta de normalizar os resíduos usuais (a partir da sua distribuição, dada na Proposição 2.20), e depois substituir a variância desconhecida σ^2 pela sua estimativa amostral $QMRE$.

Resíduos Studentizados (ou *externamente estandardizados*): $T_i = \frac{E_i}{\sqrt{QMRE_{[-i]} \cdot (1-h_{ii})}}$, sendo $QMRE_{[-i]}$ o valor de $QMRE$ resultante de um ajustamento da Regressão *excluíndo* a i-ésima observação. Usando $QMRE_{[-i]}$, procura-se tornar a estimativa de σ^2 independente da observação i, associada ao resíduo E_i que se procura normalizar.

Nota: É possível mostrar que os resíduos Studentizados e (internamente) estandardizados estão directamente relacionados pela fórmula: $T_i = R_i \sqrt{\frac{n-3}{n-2-R_i^2}}$.

2.10.2 Como analisar os resíduos

O facto de os resíduos não serem independentes (nem identicamente distribuídos) torna difícil estudar pressupostos como a Normalidade através dos habituais testes de hipóteses, como seja o teste de Shapiro-Wilks. De facto, os testes de Normalidade usuais exigem observações independentes e identicamente distribuídas o que já vimos não o caso dos resíduos E_i. Assim, é hábito validar os pressupostos do Modelo de Regressão através duma discussão mais subjectiva e rudimentar, baseada em *gráficos* de resíduos.

Vejamos seguidamente os principais gráficos de resíduos para este estudo. Registe-se que, em regressões lineares simples, alguns destes gráficos são dispensáveis, uma vez que a informação que transmitem também será visível directamente nas nuvens de pontos da variável resposta y sobre a variável preditora x. No entanto, a leitura destes gráficos é mais fácil de perceber neste contexto, para depois ser aplicada no caso de regressões lineares múltiplas, onde a nuvem das observações deixa de ser, em geral, visualizável.

2.10.2.1 Gráficos de resíduos e_i vs. valores ajustados $\hat{y}_i$

Um gráfico indispensável é o de *Resíduos* (em geral, os resíduos usuais e_i, mas poderia usar-se uma das outras variantes de resíduos) contra os *valores ajustados de* Y. Neste tipo de gráfico, e quando são válidos os pressupostos do Modelo RLS, *os resíduos devem dispor-se aproximadamente numa banda horizontal em torno de zero, sem qualquer padrão especial.* De facto, sendo válido o Modelo RLS, a correlação entre os valores que definem cada eixo são nulos: $cor(E_i, \hat{Y}_i) = 0$ (veja-se o Exercício RLS 22).

No exemplo dos lírios, o gráfico em questão pode ser construído pelo comando seguinte, com os resultados

apresentados na Figura 2.29. A leitura do gráfico sugere que pode haver alguma maior dispersão dos resíduos para as observações mais à direita no gráfico (o que pode indiciar problemas com o pressuposto de homogeneidade de variâncias).

```
> plot(fitted(iris.lm),residuals(iris.lm))
```

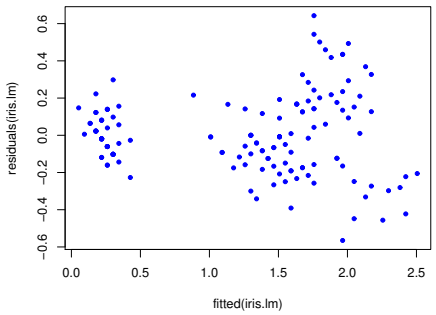


Figura 2.29: Gráfico de resíduos (usuais), no eixo vertical, contra valores ajustados $\hat{y}_i$, no eixo horizontal, para o exemplo dos lírios (Exercícios RLS 8 e 15). Quando se verificam os pressupostos do modelo de regressão linear simples, este tipo de gráfico não deve apresentar qualquer padrão especial, devendo os pontos aparecer numa banda horizontal em torno do valor zero (o valor médio dos resíduos).

Possíveis padrões indicativos de problemas Num gráfico de E_i vs. $\hat{Y}_i$ podem surgir alguns padrões indicativos de problemas. Eis os principais:

- **Curvatura na disposição dos resíduos**, como na Figura 2.30. Indica violação da hipótese de linearidade entre x e y.
- **Gráfico em forma de funil**, como na Figura 2.30, indicia a violação da hipótese de homogeneidade de variâncias, ou seja, sugere que as variâncias dos erros aleatórios, $V[\epsilon_i]$, não são iguais.
- **Um ou mais resíduos muito destacados, ou disposição dos resíduos em banda oblíqua**. Indica possíveis observações atípicas. Um possível efeito da presença de observações atípicas nos dados do Exercício RLS 9 (os dinossáurios em `Animals`) sobre este tipo de gráfico é ilustrado na Figura 2.31.

2.10.2.2 Gráficos para estudar a hipótese de normalidade.

Como foi visto na Proposição 2.20, e dado o Modelo RLS, tem-se $\frac{E_i}{\sqrt{\sigma^2\,(1-h_{ii})}} \frown \mathcal{N}(0,1)$.

Embora os resíduos estandardizados, $R_i = \frac{E_i}{\sqrt{QMRE\,(1-h_{ii})}}$ não tenham exactamente a distribuição $\mathcal{N}(0,1)$, para grandes amostras desvios importantes à Normalidade neste tipo de resíduos indiciam a

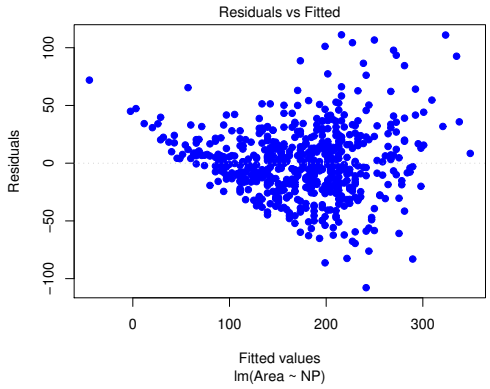


Figura 2.30: Um exemplo de resíduos em *forma de funil*, e sugerindo alguma *curvatura* na relação entre as duas variáveis. Os dados são referentes à regressão linear da variável `Area` (área foliar de $n=600$ folhas de videira) sobre a variável `NP` (comprimento da nervura principal das folhas). Estes dados constam da *data frame* `videiras` e são usados em Exercícios da Regressão Linear Múltipla.

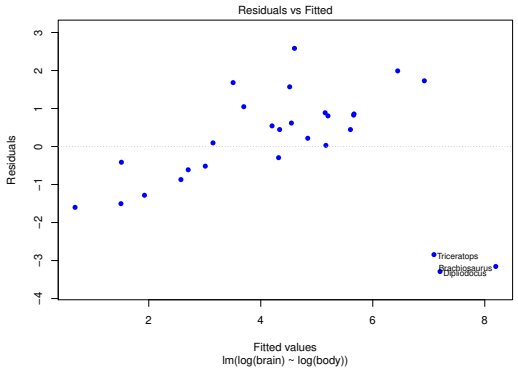


Figura 2.31: Um exemplo de resíduos em *banda oblíqua*, resultantes da presença de observações atípicas. Os dados são os da *data frame* `Animals`, referidos no Exercício RLS 9.

violação do pressuposto de Normalidade dos erros aleatórios ϵ_i. É hábito investigar a validade do pressuposto de erros aleatórios Normais através de gráficos como:

- Um histograma dos resíduos standardizados; ou
- um gráfico de quantis contra quantis, ou *qq-plot*, que compara os *quantis empíricos* dos n resíduos standardizados com os *quantis teóricos* numa distribuição $\mathcal{N}(0,1)$. Recorde-se que, para qualquer conjunto de n valores (que no nosso contexto são os n resíduos estandardizados R_i) o quantil empírico de ordem $\frac{i}{n}$ é definido como o valor que, após ordenação por ordem crescente, fica na posição i. O correspondente quantil teórico é o valor que, na distribuição $\mathcal{N}(0,1)$, deixa à sua

esquerda um intervalo de probabilidade $\theta = \frac{i}{n}$.

Um qq-plot indicativo de concordância com a hipótese de Normalidade dos erros aleatórios deverá ter os pontos aproximadamente em cima de uma linha recta. O *qq-plot* da Figura 2.32 diz respeito aos dados dos lírios considerados nos Exercícios RLS 8 e 15.

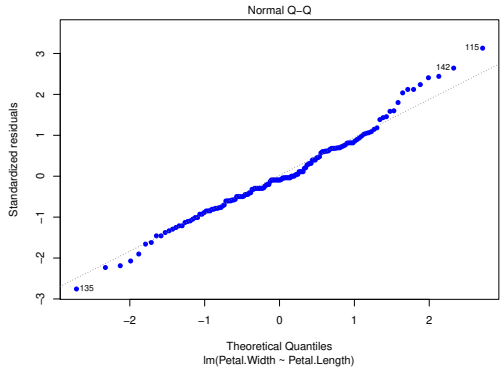


Figura 2.32: Um *qq-plot*, comparando quantis empíricos (no eixo vertical) e os quantis teóricos duma distribuição $\mathcal{N}(0,1)$ (no eixo horizontal), no caso da regressão linear simples das variáveis `Petal.Width` sobre `Petal.Length`, nos dados dos Exercícios RLS 8 e 15 (lírios). Quando os pontos resultantes se afastam muito duma relação linear, o pressuposto de Normalidade dos erros aleatórios é questionável. Apesar de algum afastamento na parte mais à direita, nem esse afastamento é muito acentuado, nem afecta muitos pontos. Assim, nesta regressão a hipótese de Normalidade dos erros aleatórios parece admissível.

2.10.2.3 Estudo de resíduos no `R`

No `R`, os três tipos de resíduos obtêm-se com outras tantas funções:

Resíduos usuais (E_i): `residuals`

Resíduos estandardizados (R_i): `rstandard`

Resíduos Studentizados (T_i): `rstudent`

O comando `plot`, aplicado a uma regressão linear ajustada pelo comando `lm`, pode produzir até seis gráficos, sendo os dois primeiros os gráficos referidos acima. Para o exemplo dos lírios que tem estado a ser discutido, o seguinte comando gera esses dois gráficos, com os resultados na Figura 2.33.

```
> plot(iris.lm, which=1:2)
```

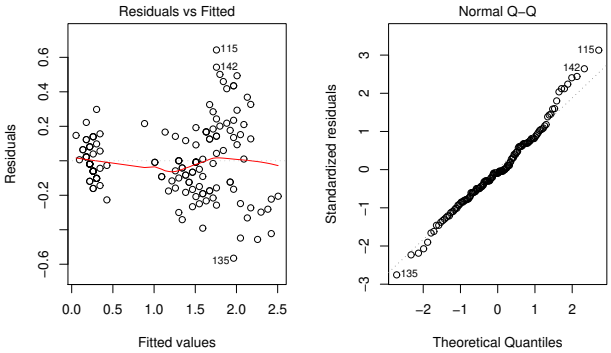


Figura 2.33: Os gráficos de resíduos produzidos pelo comando `plot` do `R` quando aplicado a uma regressão linear, neste caso à regressão linear simples da largura sobre os das pétalas dos $n = 150$ lírios (dados na *data frame* `iris` do `R`, usados nos Exercícios 8 e 15). Os gráficos foram criados com o comando `plot(iris.lm, which=c(1,2))`, sendo `iris.lm` a regressão referida. À esquerda, o gráfico de resíduos usuais, e_i, contra valores ajustados, $\hat{y}_i$ (`which=1`). Este é o gráfico já mostrado na Figura 2.29. À direita, o *qq-plot* correspondente (`which=2`). Este gráfico é o mostrado na Figura 2.32.

2.10.2.4 Gráficos para o estudo de independência

Dependência entre erros aleatórios pode surgir com observações que sejam sequenciais no tempo como resultado, por exemplo, de um "tempo de retorno" de um aparelho de medição, ou de outro fenómeno associado a *correlação temporal*. Pode também surgir associado a *correlação espacial*.

Em casos onde se suspeite de correlação no tempo, ou no espaço, será útil inspeccionar um *gráfico de resíduos vs. ordem de observação* ou *posição no espaço*, para verificar se existem padrões que sugiram falta de independência.

Este tipo de situação corresponde a uma violação do pressuposto de independência e exige a utilização de outros tipos de modelos, como modelos para sucessões cronológicas (*time series*, em inglês) ou modelos para dados espaciais, que ultrapassam o âmbito desta disciplina.

2.10.3 Outro tipo de diagnósticos

Além dos resíduos, existem outros tipos de indicadores de diagnóstico que é útil estudar numa regressão linear simples, a fim de identificar eventuais observações cujo comportamento seja, de alguma forma, de destacar. Trata-se de observações que merecem ulterior análise e que podem ter um grande impacto no ajustamento do modelo.

2.10.3.1 Observações atípicas

Nota 2.10: Observações atípicas

O conceito de **observações atípicas** (*outliers* em inglês) não tem uma definição rigorosa. Procura designar *observações que se distanciam da relação linear de fundo entre Y e a variável preditora.*

Muitas vezes, observações atípicas surgem associadas a resíduos grandes (em módulo). Em particular, e como os resíduos estandardizados ou Studentizados têm distribuição aproximadamente $\mathcal{N}(0,1)$ para n grande, observações para as quais $|R_i| > 3$ ou $|T_i| > 3$ podem ser classificadas como atípicas.

Mas é preciso ter cautela: por vezes, observações distantes da tendência geral *podem afectar de tal forma o próprio ajustamento do modelo*, que deixam de ser facilmente identificáveis a partir do valor do seu resíduo. Pode ser o caso duma observação muito afastada da nuvem de pontos, mas numa direcção diferente da direcção que caracteriza a tendência linear de fundo dos restantes pontos. Em casos extremos, essa observação individual pode ser de tal forma importante na determinação da direcção da recta ajustada, que acaba por ter um resíduo relativamente pequeno.

2.10.3.2 O efeito alavanca (*leverage*)

Definição 2.15: Efeito alavanca duma observação

Na Regressão Linear Simples, o efeito alavanca (*leverage*, em inglês) associado à i-ésima observação define-se como:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{(n-1)\, s_x^2} \ . \tag{2.42}$$

Observações com um elevado valor do efeito alavanca (*leverage points* em inglês) são observações que tendem a "atrair" a recta de regressão. A escolha de h_{ii} como indicador de diagnóstico resulta da sua presença na expressão da variância do i-ésimo resíduo E_i (ver Proposição 2.20): $V[E_i] = \sigma^2\,(1 - h_{ii})$. Assim, *Se h_{ii} é elevado*, a variância do resíduo E_i é baixa, logo o resíduo tende a estar próximo do seu valor médio (zero), ou seja, *a recta de regressão tende a passar próximo desse ponto.*

É evidente a partir da equação (2.42), que numa RLS, quanto mais afastado estiver o valor x_i da média $\overline{x}$, maior será o efeito alavanca. Outras propriedades úteis na interpretação dum efeito alavanca na RLS são dadas de seguida.

Proposição 2.21: Propriedades do valor do efeito alavanca

Os valores do efeito alavanca verificam as seguintes propriedades:

1. *Para qualquer observação i*, verifica-se:

$$\frac{1}{n} \ \leq \ h_{ii} \ \leq 1 \ .$$

2. O *valor médio* das observações alavanca numa regressão linear simples é a razão entre o número de parâmetros do modelo e o número de observações:

$$\overline{h} = \frac{2}{n} \ .$$

3. Se existirem r observações com o mesmo valor x_i do preditor, o efeito alavanca de qualquer delas não pode exceder $\frac{1}{r}$. Assim, *repetir observações de Y para os mesmos valores da variável preditora é uma forma de impedir que os efeitos alavanca sejam excessivos.*

Demonstração 2.18: Proposição 2.21

1. A primeira desigualdade é imediata, já que a segunda parcela na definição (eq. 2.42) não pode ser negativa. A segunda desigualdade tem de ser verdadeira, pois caso contrário a variância do i-ésimo resíduo, $V[E_i] = \sigma^2 (1 - h_{ii})$ (Proposição 2.20) seria negativa, o que é impossível.

2. A soma dos n efeitos alavanca é $\sum_{i=1}^{n} h_{ii} = \sum_{i=1}^{n} \left[\frac{1}{n} + \frac{(x_i - \overline{x})^2}{(n-1)\, s_x^2}\right] = 1 + \frac{\overbrace{\sum_{i=}^{n} (x_i - \overline{x})^2}^{=(n-1)\, s_x^2}}{(n-1)\, s_x^2} = 2$. Logo, a média dos n valores h_{ii} será $\overline{h} = \frac{2}{n}$.

3. Omite-se a demonstração.

Observações com um efeito alavanca elevado podem, ou não, estar dispostas com a mesma tendência de fundo que as restantes observações (i.e., *podem, ou não, ser atípicas*). A forma de identificar estas diferentes situações é discutida na Subsecção seguinte.

2.10.3.3 Observações influentes.

Observações influentes são observações que, se retiradas do conjunto de dados, geram mudanças assinaláveis nos parâmetros estimados, b_0 e b_1, e portanto na recta e nos valores ajustados de Y. A medida mais frequente para a *influência da observação i* é a **distância de Cook**, que é uma versão normalizada da soma de quadrados entre os valores $\hat{y}_j$ ajustados a partir das duas rectas: a recta obtida a partir das n observações, e a recta obtida sem a observação i.

Definição 2.16: Distância de Cook

$$D_i = \frac{\sum_{j=1}^{n} \left[\hat{y}_j - \hat{y}_{j_{(-i)}}\right]^2}{2 \cdot QMRE} \ , \tag{2.43}$$

sendo $\hat{y}_j$ o j-ésimo valor ajustado pela recta de regressão com as n observações e $\hat{y}_{j_{(-i)}}$ o correspondente valor, ajustado com base na recta obtida sem a observação i.

Uma expressão equivalente para a distância de Cook, utiliza o resíduo estandardizado R_i e o efeito alavanca h_{ii}:

$$D_i = R_i^2 \left(\frac{h_{ii}}{1 - h_{ii}} \right) \frac{1}{2} \, . \tag{2.44}$$

Quanto maior a distância de Cook D_i, maior é a influência da i-ésima observação. Sugere-se $D_i > 0.5$ como *critério de observação influente*. Repare-se que se $D_i > 0.5$, a soma de quadrados no numerador da definição de D_i excede o Quadrado Médio Residual, ou seja, a soma de quadrados das diferenças entre os valores ajustados $\hat{y}_j$, obtidos com as duas rectas, é maior do que a variabilidade dos pontos em torno da recta, estimada por $QMRE$.

2.10.4 Um exemplo com o auxílio do R

Observações atípicas, influentes ou alavanca, embora podendo estar relacionadas, não são o mesmo conceito. Por exemplo, uma observação com resíduo (internamente) estandardizado grande e h_{ii} elevado, tem de ter uma distância de Cook grande, logo ser influente. Se tiver R_i^2 grande e h_{ii} pequeno (ou viceversa), pode, ou não, ser influente, consoante a grandeza relativa desses dois valores.

Ilustramos essas diferenças recorrendo a um *subconjunto* de 23, de entre as 28 espécies animais estudadas no Exercício 9 de Regressão Linear Simples, a que corresponde o gráfico de log-peso do corpo *vs.* log-peso do cérebro dado na Figura 2.34. Como se pode observar, há duas espécies mais distantes da nuvem de pontos, mas com um afastamento de características diferentes: enquanto o *rato* se dispõe na mesma tendência de fundo, o *Triceratops* não.

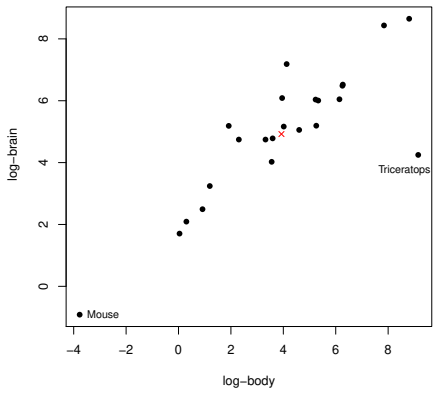


Figura 2.34: Gráfico de log-peso do cérebro (y) contra log-peso do corpo (x), para algumas (23 das 28) espécies terrestres referidas na *data frame* `Animals`, do módulo `MASS` (usada no Exercício RLS 9). A cruz (x) indica o *centro de gravidade* $(\overline{x}, \overline{y})$ da nuvem de pontos.

Foi ajustada uma regressão linear simples, e foram calculados os resíduos (internamente) estandardizados

(R_i), as distâncias de Cook (D_i) e os valores do efeito alavanca (h_{ii}) para este subconjunto de dados. Os valores estão indicados de seguida. Como se pode verificar, a espécie rato tem uma distância de Cook ($D_{18} = 0.355$) bastante menor do que a espécie *Triceratops* ($D_{15} = 1.431$), ou seja, é menos influente do que esta última espécie. Isso reflecte o facto de uma exclusão da espécie rato do conjunto de dados afectar menos o ajustamento da recta do que a exclusão do *Triceratops*. No entanto, a espécie rato está associada ao maior efeito alavanca ($h_{18,18} = 0.341$) de qualquer das 23 espécies, cerca do dobro do efeito alavanca do *Triceratops* ($h_{15,15} = 0.180$). Isso reflecte o facto de que o log-peso do corpo dos ratos (o seu valor x^*, que pelo gráfico se verifica ser próximo de -4) se afastar mais da média desses log-pesos (pelo gráfico, próximo de 4) do que o log-peso do dinossáurio (pelo gráfico, próximo de 9). É de novo o *Triceratops* que tem o maior valor absoluto de resíduo (internamente) estandardizado, R_i ($|R_{15}| = 3.610$).

```
                   R_i   D_i  h_ii
Mountain beaver  -0.547 0.018 0.109
Cow              -0.201 0.001 0.068
Grey wolf         0.057 0.000 0.044
Goat              0.168 0.001 0.045
Guinea pig       -0.754 0.039 0.119
Asian elephant    1.006 0.069 0.120
Donkey            0.276 0.002 0.052
Horse             0.121 0.001 0.071
Potar monkey      0.711 0.015 0.057
Cat              -0.006 0.000 0.081
Giraffe           0.145 0.001 0.071
Gorilla           0.195 0.001 0.053
Human             1.850 0.078 0.044
African elephant  0.688 0.046 0.163
Triceratops      -3.610 1.431 0.180   <-- D_i muito grande; h_ii nem por isso
Rhesus monkey     1.306 0.058 0.064
Kangaroo         -0.578 0.008 0.044
Mouse            -1.172 0.355 0.341   <-- h_ii mais elevado; D_i nem por isso
Rabbit           -0.519 0.013 0.089
Sheep             0.163 0.001 0.044
Jaguar           -0.243 0.001 0.046
Chimpanzee        0.992 0.022 0.043
Pig              -0.471 0.006 0.052
```

A função `plot`, aplicada a uma regressão ajustada pelo comando `lm`, produz também gráficos com alguns dos diagnósticos considerados acima.

A opção `which=4` gera um diagrama de barras das distâncias de Cook associadas a cada observação. Um exemplo destes gráfico de diagnóstico, para os dados *completos* do Exercício 9 de Regressão Linear Simples (`Animals`) é dado no gráfico da esquerda da Figura 2.35. Há uma espécie (a espécie 26, o *Brachiosaurus*), cuja distância de Cook é próxima de 0.6, ou seja, excede o valor (0.5) habitualmente usado para salientar uma observação como sendo muito influente. Outra espécie de dinossáurio (a espécie 6, o *Dipliodocus*) tem uma distância de Cook inferior a 0.5, mas já considerável ($D_6 \approx 0.35$). Estes valores elevados de distância de Cook reflectem o distanciamento das espécies de dinossáurios da tendência geral das outras espécies, com os efeitos estudados no Exercício RLS 9. Mas deve ser sublinhado que no cálculo das distâncias de Cook apenas se exclui *uma* observação, pelo que o facto de haver nestes dados *três* observações atípicas

mitiga um pouco o valor das respectivas distâncias de Cook (no cálculo de D_{26} continuam presentes nos dados as duas outras espécies de dinossáurios, com o seu efeito de atracção da recta ajustada).

A opção `which=5` produz um gráfico de Resíduos estandardizados (R_is) no eixo vertical contra valores de h_{ii} (*leverages*) no eixo horizontal, traçando linhas de igual distância de Cook (para os níveis 0.5 e 1, por omissão), que destacam eventuais observações influentes. Estas linhas resultam de substituir o valor $D_i = 0.5$ na equação 2.44, de onde resulta (após manipulação algébrica) as equações $R_i = \pm\sqrt{\frac{1}{h_{ii}} - 1}$, que correspondem às curvas assinaladas. É visível que a observação 26 ultrapassa uma destas curvas, reflectindo o facto de a sua distância de Cook ser superior a 0.5. No gráfico é ainda visível que nenhum resíduo estandardizado tem valor absoluto digno de registo (os maiores são pouco superiores a 2) e que também os maiores valores do efeito alavanca são relativamente modestos (embora dois desses valores sejam próximos de 0.2).

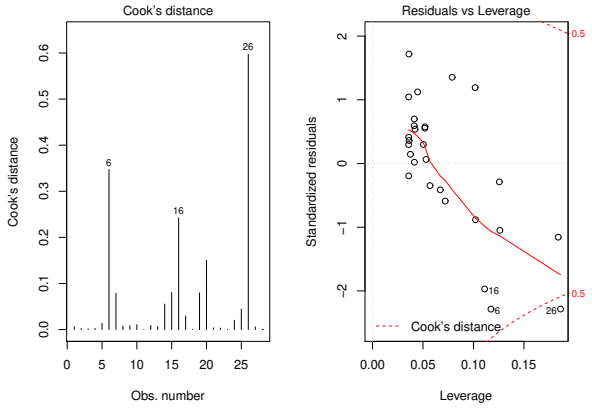


Figura 2.35: Estes gráficos foram obtidos com o comando `plot(Animals.lm, which=c(4,5))`, onde `Animals.lm` indica a regressão linear simples de log-pesos do cérebro sobre log-pesos do corpo das 28 espécies do conjunto de dados `Animals`, do módulo `MASS`. À esquerda, o diagrama de barras das distâncias de Cook (`which=4`). À direita, o gráfico dos resíduos estandardizados R_i contra os valores do efeito alavanca h_{ii} (`which=5`). Nos cantos inferior e superior direito deste último gráfico são visíveis as curvas correspondentes à distância de Cook 0.5.

2.10.5 Ainda as transformações de variáveis

Perante violações graves de pressupostos do Modelo, como o pressuposto da Normalidade dos erros aleatórios ou da homogeneidade de variâncias, torna-se necessário ultrapassar os problemas antes de proceder à utilização do modelo. Há muitas sugestões na literatura relativas à utilização de *transformações de variáveis* com este objectivo. Algumas transformações aconselhadas para estabilizar a variância, quando esta é proporcional a determinadas potências do valor médio, são indicadas na tabela seguinte.

Relação entre a variância e a média	Transformação aconselhada
$var(Y_i) \propto E[Y_i]$	$Y \longrightarrow \sqrt{Y}$
$var(Y_i) \propto (E[Y_i])^2$	$Y \longrightarrow \ln Y$
$var(Y_i) \propto (E[Y_i])^4$	$Y \longrightarrow 1/Y$

Existe uma família inteira de potenciais transformações, a *família Box-Cox de transformações*, aconselhada na tentativa de ultrapassar problemas com a Normalidade dos dados. A família Box-Cox define-se da seguinte forma, para qualquer valor real do parâmetro λ:

$$Y \longrightarrow \begin{cases} \frac{Y^\lambda - 1}{\lambda} & , \lambda \neq 0 \\ \ln(Y) & , \lambda = 0 \end{cases}$$

Nota 2.11: Prevenções

A utilização de transformações, sobretudo quando afectam a variável resposta Y, deve ser *feita com cautela*.

- Uma transformação de variáveis *também altera a relação de base entre as variáveis originais*;
- Uma transformação que "corrija" um problema (e.g., variâncias heterogéneas) *pode gerar outro* (e.g., não-normalidade);
- Existe o perigo de usar transformações que resolvam o problema duma amostra específica, mas *não tenham qualquer generalidade*.

Recomenda-se especial atenção ao impacto que a utilização de transformações de variáveis tem sobre a relação entre x e y (já estudado em contexto meramente descritivo – ver Subsecção 2.3). Às prevenções feitas em contexto descritivo sobre o uso de transformações linearizantes (ou seja, o facto de os estimadores que minimizam a soma de quadrados dos resíduos nas relações linearizadas *não* serem os que minimizam a soma de quadrados de resíduos na relação não-linear original), acrescentam-se mais duas prevenções, específicas do contexto inferencial agora sob consideração:

- *As transformações consideradas em contexto descritivo não levaram em conta os erros aleatórios.*
- As hipóteses do Modelo Linear (ou seja, erros aleatórios aditivos, Normais, de variância homogénea, média zero e independentes) *terão de ser válidas para as relações lineares entre as variáveis transformadas*, ou seja, aquando da aplicação do Modelo Linear.

Capítulo 3

Regressão Linear Múltipla

Um modelo linear com uma única variável preditora pode não se mostrar adequado. Nesse caso, pode haver interesse em considerar o uso de *mais do que uma variável preditora* para modelar a variável resposta de interesse. A escolha desses preditores adicionais pode ser feita, ou por considerações teóricas, ou por considerações empíricas.

3.1 Um exemplo motivador: os dados Tinta Francisca

Num estudo sobre uma população experimental de clones da casta Tinta Francisca, realizado no Tabuaço em 2003, foram medidos os valores das seguintes variáveis em 24 videiras:

- teor de antocianas (variável `antoci`, em mg/dm^3);
- teor de fenóis totais (variável `fentot`);
- pH (variável `pH`).

O teor de antocianas é uma variável de medição difícil, logo há interesse em modelar essa variável resposta, uma vez que um bom modelo evitaria a sua medição rigorosa. O teor de fenóis totais e pH serão usadas como variáveis preditoras.

As $n = 24$ observações em três variáveis descrevem uma *nuvem de 24 pontos em* $\mathbb{R}^3$, o que levanta dificuldades na visualização dos dados. Com o auxílio do módulo `rggobi`, que permite usar o *software* gráfico `Ggobi` a partir do `R`, foi construída essa nuvem de pontos em $\mathbb{R}^3$. À primeira vista, a relação nada tem de especial (ver Figura 3.1, gráfico da esquerda). Utilizando a possibilidade que o *software* oferece de rodar a nuvem de pontos, encontra-se um outro ângulo de visão, onde se torna evidente que os pontos se dispersam aproximadamente em torno de *um plano*, ou seja, duma *superfície linear*, em $\mathbb{R}^3$, como se constata na Figura 3.1, gráfico da direita.

Ora, qualquer plano em $\mathbb{R}^3$, no sistema $x0y0z$, tem equação

$$Ax + By + Cz + D = 0 \ .$$

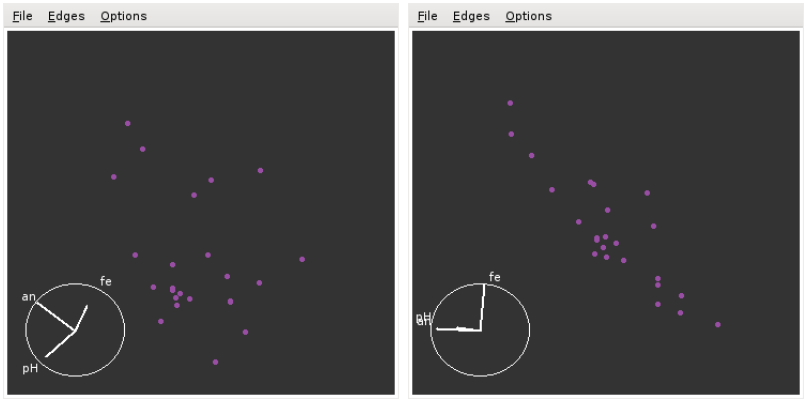


Figura 3.1: Visão da nuvem de pontos a três dimensões, dos dados da Tinta Francisca, vista de dois ângulos diferentes. A imagem da direita mostra que os pontos se dispõem aproximadamente em torno de um plano em $\mathbb{R}_3$ (que se prolonga em profundidade). Os gráficos forma construídos com o auxílio do módulo `rggobi`, que permite a interface entre o *software* gráfico `Ggobi` e o `R`. É necessário instalar previamente, quer o `Ggobi`, quer o módulo (*package*) do `R` de nome `rggobi`.

No nosso contexto, e associando ao eixo vertical (z) a variável resposta Y; a um dos eixos horizontais (x), um preditor X_1; ao terceiro eixo (y), o outro preditor X_2, a equação fica (no caso geral de planos não verticais, com $C \neq 0$):

$$\begin{aligned} Ax_1 + Bx_2 + Cy + D = 0 \quad &\Leftrightarrow \quad y = -\frac{D}{C} - \frac{A}{C}x_1 - \frac{B}{C}x_2 \\ &\Leftrightarrow \quad y = b_0 + b_1 x_1 + b_2 x_2 \end{aligned}$$

Esta equação generaliza a equação da recta, para o caso de haver dois preditores. A Figura 3.2 representa graficamente a situação associada ao ajustamento dum plano de equação $y = b_0 + b_1 x_1 + b_2 x_2$ num espaço tri-dimensional $\mathbb{R}^3$ ($x_1 0 x_2 0 y$).

3.2 Regressão Linear Múltipla em contexto descritivo

A equação do plano mostrado na Figura 3.2 pode ser *ajustada pelo mesmo critério que na Regressão Linear Simples*, ou seja, *minimizar a Soma de Quadrados dos Resíduos, SQRE*. O plano assim obtido será o *plano de regressão linear*, ou *plano de mínimos quadrados*.

3.2.1 O caso geral: p preditores

Caso se pretenda modelar uma variável resposta, y, com base em p *variáveis preditoras*, x_1, x_2, ..., x_p, serão necessários n conjuntos de observações nestas $p+1$ variáveis:

$$\left\{ \left(x_{1(i)}, x_{2(i)}, ... x_{p(i)}, y_i \right) \right\}_{i=1}^{n} \ . \tag{3.1}$$

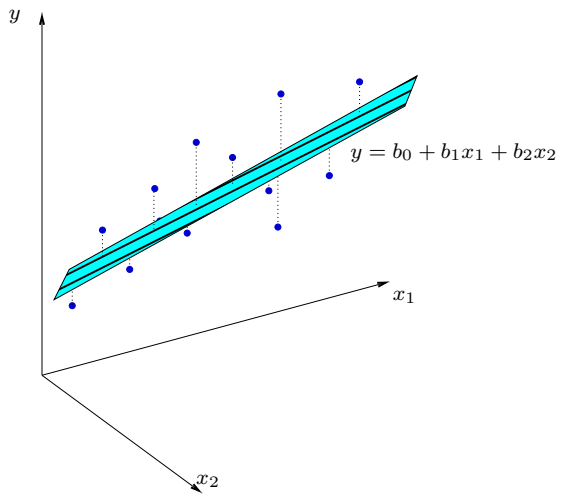


Figura 3.2: Nuvem de pontos genérica em $\mathbb{R}^3$, com um plano como tendência de fundo. A tracejado encontram-se as distâncias na vertical entre valores de y observados e valores correspondentes de y, ajustados pelo plano de equação $y = b_0 + b_1x_1 + b_2x_2$. Essas distâncias (afectadas de sinal) correspondem aos resíduos $e_i = y_i - \hat{y}_i$.

A nuvem de pontos resultantes já não é visualizável. A *representação gráfica usual* da nuvem de n pontos observados *exige* $p+1$ *eixos*: um para y e um para cada uma das p variáveis preditoras. Para $p > 2$, são necessários mais de três eixos e *a visualização torna-se impossível*. Será necessário usar a nossa intuição geométrica para nos ajudar na compreensão do que se pretende fazer. Neste contexto, torna-se tentador recorrer a uma forma de representação alternativa dos dados, mais adequada aos fins estatísticos em questão, o que será feito na Subsecção 3.2.2. Para já, aprofundemos alguns conceitos relativos à forma tradicional de representar n observações em p variáveis.

A representação tradicional gera uma *nuvem de n pontos num espaço de dimensão* $p+1$, ao:

- associar um eixo a cada *variável* observada (logo, $p+1$ *eixos*).
- representar cada *indivíduo (unidade experimental)* observado por um ponto, cujas coordenadas são os $p+1$ valores observados para esse indivíduo (ver (3.1)).

Uma generalização da equação de regressão linear simples admite que *os valores de y oscilam em torno duma combinação linear (afim) das p variáveis preditoras*:

$$y = b_0 + b_1x_1 + b_2x_2 + ... + b_px_p \ . \tag{3.2}$$

Trata-se da equação dum *hiperplano em* $\mathbb{R}^{p+1}$, que define a relação de fundo entre a variável resposta e os p preditores.

Sendo impossível visualizar uma nuvem de n pontos em $\mathbb{R}^{p+1}$, podem no entanto obter-se *visões parciais*, como sejam as nuvens de pontos definidas por cada par de variáveis. Por exemplo, e considerando os dados dos lírios disponíveis na *data frame* `iris`, para as $n = 150$ observações em 4 variáveis obter-se-iam

os gráficos dados na Figura 3.3. Estes gráficos de *pares* de variáveis são as *projecções ortogonais* da nuvem de n pontos *sobre cada plano coordenado de* $\mathbb{R}^{p+1}$.

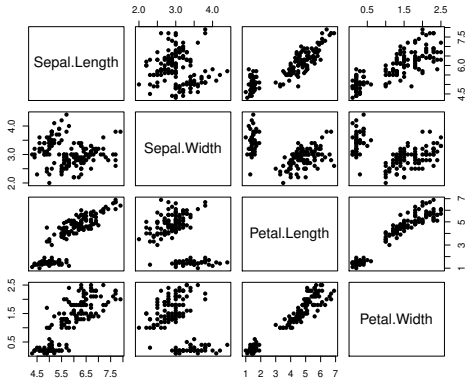


Figura 3.3: Nuvens de pontos dos dados `iris`, em todos os possíveis pares de variáveis. Os nomes das variáveis são indicados nos quadrados da diagonal principal. Em cada coluna encontram-se os gráficos com uma mesma variável no eixo horizontal. Em cada linha encontram-se os gráficos com uma mesma variável no eixo vertical. Cada um destes gráficos corresponde à projecção da nuvem de $n=150$ pontos em $\mathbb{R}^4$ sobre um dos $\binom{4}{2}=6$ planos coordenados definidos por cada par de eixos (note-se que cada plano coordenado aparece duas vezes na Figura, trocando os eixos vertical e horizontal).

A projecção da nuvem de n pontos nos planos coordenados *nem sempre permite verificar a hipótese básica de linearidade*, isto é, a hipótese de que os pontos se dispersam em torno de um hiperplano. *Tal hipótese pode ser válida, mesmo que não se verifique linearidade em qualquer das nuvens de pontos de* y *contra um preditor individual,* x_j.

3.2.2 Uma representação gráfica alternativa: o espaço das variáveis

A representação gráfica em $\mathbb{R}^{p+1}$ das n observações de Y e das p variáveis preditoras não é a única possível. Uma outra representação dos dados é concebível, representação que *casa conceitos geométricos e conceitos estatísticos* e será *útil na determinação dos parâmetros ajustados.*

As n observações da variável resposta y definem um *vector em* $\mathbb{R}^n$:

$$\vec{\mathbf{y}} \;=\; (y_1, y_2, y_3, ..., y_n)\ .$$

Da mesma forma, *as* n *observações de cada variável preditora* definem um *vector de* $\mathbb{R}^n$:

$$\vec{\mathbf{x_j}} \;=\; (x_{j_{(1)}}, x_{j_{(2)}}, x_{j_{(3)}}, ..., x_{j_{(n)}}) \quad (j = 1, 2, ..., p).$$

Assim, *todas as variáveis podem ser representadas por vectores em* $\mathbb{R}^n$, pelo que se diz que esta é a representação alternativa *no espaço das variáveis.* O *vector de* n *uns*, representado por $\vec{\mathbf{1}}_n$, também será útil. Nesta representação, que é ilustrada na Figura 3.4,

- *cada eixo* corresponde a um *indivíduo* observado;
- *cada vector* corresponde a uma *variável*.

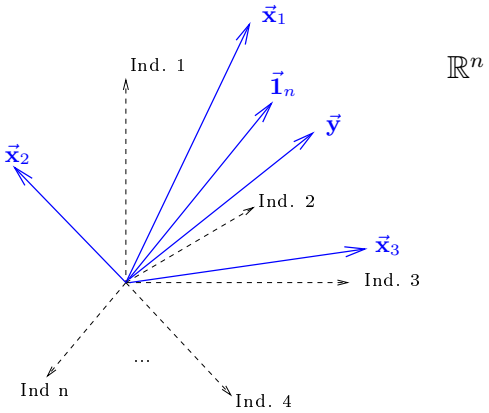


Figura 3.4: A representação de observações de $p+1$ variáveis sobre n indivíduos, no espaço das variáveis. Cada eixo corresponde a um indivíduo e cada variável define um vector em $\mathbb{R}^n$.

Dada a equação do hiperplano (3.2), os valores ajustados $\hat{y}_i$ obtêm-se usando para cada indivíduo i os valores correspondentes dos preditores, ou seja:

$$\hat{y}_i \;=\; b_0 + b_1 x_{1_{(i)}} + b_2 x_{2_{(i)}} + ... + b_p x_{p_{(i)}} \;. \tag{3.3}$$

O vector com os n valores ajustados, que designaremos por $\vec{\hat{\mathbf{y}}}$ também é um *vector de* $\mathbb{R}^n$. Como se verá, o vector dos valores ajustados, $\vec{\hat{\mathbf{y}}}$, é uma *combinação linear dos vectores* $\vec{\mathbf{1}}_n$, $\vec{\mathbf{x}}_1$, $\vec{\mathbf{x}}_2$, ..., $\vec{\mathbf{x}}_p$:

$$\begin{aligned}
\vec{\hat{\mathbf{y}}} \;&=\; \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \vdots \\ \hat{y}_n \end{bmatrix} \;=\; \begin{bmatrix} b_0 + b_1 x_{1_{(1)}} + b_2 x_{2_{(1)}} + ... + b_p x_{p_{(1)}} \\ b_0 + b_1 x_{1_{(2)}} + b_2 x_{2_{(2)}} + ... + b_p x_{p_{(2)}} \\ b_0 + b_1 x_{1_{(3)}} + b_2 x_{2_{(3)}} + ... + b_p x_{p_{(3)}} \\ \cdots \\ b_0 + b_1 x_{1_{(n)}} + b_2 x_{2_{(n)}} + ... + b_p x_{p_{(n)}} \end{bmatrix} \\
&=\; b_0 \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + b_1 \begin{bmatrix} x_{1_{(1)}} \\ x_{1_{(2)}} \\ x_{1_{(3)}} \\ \vdots \\ x_{1_{(n)}} \end{bmatrix} + ... + b_p \begin{bmatrix} x_{p_{(1)}} \\ x_{p_{(2)}} \\ x_{p_{(3)}} \\ \vdots \\ x_{p_{(n)}} \end{bmatrix} \\
&=\; b_0 \vec{\mathbf{1}}_n + b_1 \vec{\mathbf{x}}_1 + b_2 \vec{\mathbf{x}}_2 + ... + b_p \vec{\mathbf{x}}_p
\end{aligned}$$

3.2.3 A matriz do modelo e o seu subespaço de colunas

Recordemos alguns conceitos básicos de *Álgebra Linear* (leccionados nos primeiros ciclos do ISA).

- O conjunto de *todas* as combinações lineares de um dado conjunto de vectores de $\mathbb{R}^n$ chama-se o **subespaço gerado** por esses vectores.
- Se o conjunto de vectores forem as *colunas* duma matriz $\mathbf{X}$ chamamos ao subespaço gerado por esses vectores o **subespaço das colunas da matriz X**, que é representado por $\mathcal{C}(\mathbf{X}) \subseteq \mathbb{R}^n$.

Definição 3.1: A matriz do modelo X e o seu subespaço das colunas $\mathcal{C}(\mathbf{X})$

No contexto duma regressão linear múltipla, define-se:

- A **matriz do modelo X**, de dimensão $n \times (p+1)$, como a matriz cujas $p+1$ colunas são os vectores $\vec{\mathbf{1}}_n$, $\vec{\mathbf{x}}_1$, ..., $\vec{\mathbf{x}}_p$.
- O **subespaço das colunas $\mathcal{C}(\mathbf{X})$** é o conjunto de todas as combinações lineares dos vectores $\vec{\mathbf{1}}_n$, $\vec{\mathbf{x}}_1$, ..., $\vec{\mathbf{x}}_p$.

O subespaço das colunas da matriz do modelo, $\mathcal{C}(\mathbf{X})$, é de *dimensão* $p+1$ se os vectores forem *linearmente independentes*, isto é, se nenhum dos vectores se puder escrever como combinação linear dos restantes.

O vector $\hat{\vec{\mathbf{y}}}$ dos valores ajustados é, por definição, um vector do subespaço $\mathcal{C}(\mathbf{X})$.

Um produto matricial $\mathbf{X}\vec{\mathbf{a}}$. *Qualquer combinação linear dos vectores coluna da matriz* $\mathbf{X}$ é dada por um produto da forma $\mathbf{X}\vec{\mathbf{a}}$, onde $\vec{\mathbf{a}} = (a_0, a_1, a_2, ..., a_p)$ *é o vector dos coeficientes* que define a combinação linear. De facto,

$$
\begin{aligned}
\mathbf{X}\vec{\mathbf{a}} &= \begin{bmatrix} 1 & x_{1_{(1)}} & x_{2_{(1)}} & \cdots & x_{p_{(1)}} \\ 1 & x_{1_{(2)}} & x_{2_{(2)}} & \cdots & x_{p_{(2)}} \\ 1 & x_{1_{(3)}} & x_{2_{(3)}} & \cdots & x_{p_{(3)}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1_{(n)}} & x_{2_{(n)}} & \cdots & x_{p_{(n)}} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} \\
&= \begin{bmatrix} a_0 + a_1 x_{1_{(1)}} + a_2 x_{2_{(1)}} + ... + a_p x_{p_{(1)}} \\ a_0 + a_1 x_{1_{(2)}} + a_2 x_{2_{(2)}} + ... + a_p x_{p_{(2)}} \\ a_0 + a_1 x_{1_{(3)}} + a_2 x_{2_{(3)}} + ... + a_p x_{p_{(3)}} \\ \cdots \\ a_0 + a_1 x_{1_{(n)}} + a_2 x_{2_{(n)}} + ... + a_p x_{p_{(n)}} \end{bmatrix} \\
&= a_0 \vec{\mathbf{1}}_n + a_1 \vec{\mathbf{x}}_1 + a_2 \vec{\mathbf{x}}_2 + ... + a_p \vec{\mathbf{x}}_p
\end{aligned}
$$

- Assim, cada escolha possível de coeficientes $\vec{\mathbf{a}} = (a_0, a_1, a_2, ..., a_p)$ corresponde a uma combinação linear $\mathbf{X}\vec{\mathbf{a}}$ no subespaço $\mathcal{C}(\mathbf{X})$.

- Essa escolha de coeficientes é *única* caso as colunas de $\mathbf{X}$ sejam *linearmente independentes*, isto é, se *não houver dependência linear (multicolinearidade) entre as variáveis* $\vec{\mathbf{x}}_1, ..., \vec{\mathbf{x}}_p, \vec{\mathbf{1}}_n$. Dito de outra forma, se nenhuma coluna de $\mathbf{X}$ se puder escrever como combinação linear das restantes, então $\mathbf{X}\vec{\mathbf{a}} = \mathbf{X}\vec{\mathbf{b}}$ implica que $\vec{\mathbf{a}} = \vec{\mathbf{b}}$.

- *Um dos pontos/vectores do subespaço* $\mathcal{C}(\mathbf{X})$ é a combinação linear dada pelo vector de coeficientes $\vec{\mathbf{b}} = (b_0, b_1, ..., b_p)$ que *minimiza* a Soma dos Quadrados dos Resíduos:

$$SQRE \quad = \quad \sum_{i=}^{n} e_i^2 \quad = \quad \sum_{i=}^{n} (y_i - \hat{y}_i)^2$$

  onde os y_i são os valores observados da variável resposta e $\hat{y}_i = b_0 + b_1\, x_{1(i)} + b_2\, x_{2(i)} + ... + b_p\, x_{p(i)}$ os valores ajustados. *É a combinação linear que desejamos determinar.*

Mas *como identificar esse ponto/vector?* Vamos usar *argumentos geométricos*, aproveitando a representação dos dados no espaço das variáveis ($\mathbb{R}^n$).

- Temos *um vector de n observações de* $\vec{\mathbf{y}}$ *que* está em $\mathbb{R}^n$ mas, em geral, *não está no subespaço* $\mathcal{C}(\mathbf{X})$.

- Queremos *aproximar esse vector por outro* vector, $\hat{\vec{\mathbf{y}}} = b_0\vec{\mathbf{1}}_n + b_1\vec{\mathbf{x}}_1 + ... + b_p\vec{\mathbf{x}}_p$ que, por definição, *pertence ao subespaço* $\mathcal{C}(\mathbf{X})$.

- Considerações meramente geométricas sugerem aproximar o vector de observações $\vec{\mathbf{y}}$ pelo vector $\hat{\vec{\mathbf{y}}}$ do subespaço $\mathcal{C}(\mathbf{X})$ que esteja *mais próximo de* $\vec{\mathbf{y}}$.

Solução: *Tomar a projecção ortogonal de* $\vec{\mathbf{y}}$ *sobre* $\mathcal{C}(\mathbf{X})$: o vector de $\mathcal{C}(\mathbf{X}) \subset \mathbb{R}^n$ mais próximo dum vector $\vec{\mathbf{y}} \in \mathbb{R}^n$ é o vector $\hat{\vec{\mathbf{y}}}$ que resulta de projectar ortogonalmente $\vec{\mathbf{y}}$ sobre $\mathcal{C}(\mathbf{X})$ (ver Figura 3.5).

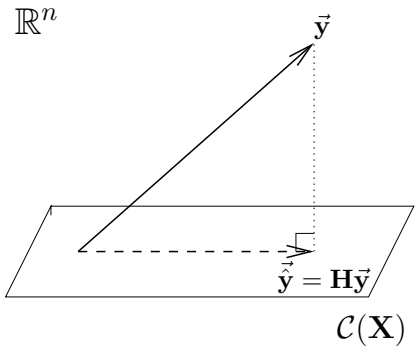


Figura 3.5: A projecção ortogonal de $\vec{\mathbf{y}}$ sobre $\mathcal{C}(\mathbf{X})$. O vector de $\mathcal{C}(\mathbf{X}) \subset \mathbb{R}^n$ mais próximo dum vector $\vec{\mathbf{y}} \in \mathbb{R}^n$ é o vector $\hat{\vec{\mathbf{y}}}$ que resulta de projectar ortogonalmente $\vec{\mathbf{y}}$ sobre $\mathcal{C}(\mathbf{X})$, criando um triângulo rectângulo como ilustrado.

Esse critério *minimiza a Soma dos Quadrados dos Resíduos, SQRE*. De facto, o vector $\vec{\hat{\mathbf{y}}}$ que minimiza a distância ao vector de observações $\vec{\mathbf{y}}$ minimiza também o *quadrado dessa distância*, que é dado por:

$$dist^2(\vec{\mathbf{y}}, \vec{\hat{\mathbf{y}}}) \;=\; \|\vec{\mathbf{y}} - \vec{\hat{\mathbf{y}}}\|^2 \;=\; \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \;=\; SQRE\ . \tag{3.4}$$

Nota 3.1

Recorde-se que:

1. dado um qualquer vector $\vec{\mathbf{x}} \in \mathbb{R}^n$, a sua norma define-se como a raíz quadrada da soma de quadrados dos seus elementos: $\|\vec{\mathbf{x}}\| = \sqrt{\sum_{i=1}^{n} x_i^2}$. A norma dum vector é o seu *comprimento*.
2. dados dois quaisquer vectores, $\vec{\mathbf{x}}, \vec{\mathbf{y}} \in \mathbb{R}^n$, a distância entre eles é dada por:

$$dist(\vec{\mathbf{x}}, \vec{\mathbf{y}}) = \|\vec{\mathbf{x}} - \vec{\mathbf{y}}\| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}\ .$$

Assim, o critério geométrico de minimizar a distância em $\mathbb{R}^n$ entre $\vec{\mathbf{y}}$ e $\vec{\hat{\mathbf{y}}}$ corresponde a *minimizar a soma de quadrados dos resíduos* $e_i = y_i - \hat{y}_i$, como ilustrado na Figura 3.6. Trata-se do *mesmo critério que foi usado na Regressão Linear Simples*.

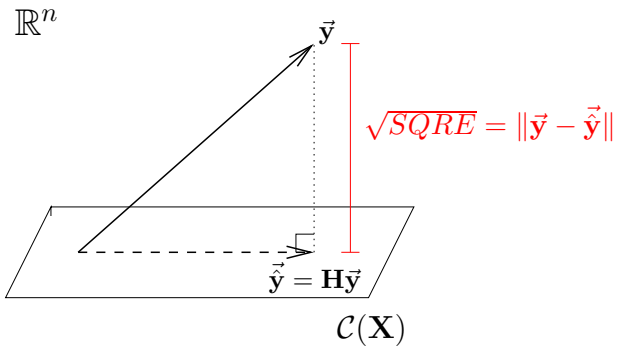


Figura 3.6: O quadrado da distância de $\vec{\mathbf{y}}$ à sua projecção ortogonal sobre $\mathcal{C}(\mathbf{X})$, $\vec{\hat{\mathbf{y}}} = \mathbf{H}\vec{\mathbf{y}}$, é $SQRE$, a soma dos quadrados dos resíduos.

Vejamos agora como proceder à projecção ortogonal dum vector de $\mathbb{R}^n$ sobre um subespaço gerado pelas colunas duma dada matriz $\mathbf{X}$.

3.2.4 Projecções ortogonais e os parâmetros ajustados

Definição 3.2: A matriz de projecção ortogonal sobre o subespaço $\mathcal{C}(\mathbf{X})$

A projecção ortogonal de um vector $\vec{\mathbf{y}} \in \mathbb{R}^n$ sobre o subespaço $\mathcal{C}(\mathbf{X})$ gerado pelas colunas (linearmente independentes) de $\mathbf{X}$ faz-se pré-multiplicando $\vec{\mathbf{y}}$ pela **matriz de projecção ortogonal sobre $\mathcal{C}(\mathbf{X})$**:

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t \ . \tag{3.5}$$

Logo, o vector dos valores ajustados, $\hat{\vec{\mathbf{y}}}$, é gerado da seguinte forma:

$$\begin{aligned} & \hat{\vec{\mathbf{y}}} = \mathbf{H}\vec{\mathbf{y}} \\ \Leftrightarrow \quad & \hat{\vec{\mathbf{y}}} = \mathbf{X}\underbrace{(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{y}}}_{=\vec{\mathbf{b}}} \end{aligned}$$

Acabámos assim de mostrar o resultado enunciado na Proposição 3.1.

Proposição 3.1: Os parâmetros ajustados na RL Múltipla

O vector $\vec{\mathbf{b}}$ dos parâmetros que minimizam a Soma de Quadrados dos Resíduos é dado por:

$$\vec{\mathbf{b}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{y}} \ . \tag{3.6}$$

Notas:

1. A equação (3.6) gera o vector $\vec{\mathbf{b}}$, de dimensão $p+1$, cujos elementos são os parâmetros $b_0, b_1, b_2, ..., b_p$ resultantes do critério de minimizar a Soma dos Quadrados Residual.
2. No caso de haver apenas um preditor ($p=1$), ou seja, de estarmos perante uma regressão linear simples, a fórmula (3.6) produz as fórmulas já estudadas no Capítulo 2, ou seja o vector $\vec{\mathbf{b}}=(b_0, b_1)$, com $b_1 = \frac{cov_{xy}}{s^2}$ e $b_0 = \overline{y} - b_1\overline{x}$ (ver o Exercício RLM 3).
3. Ao contrário do que acontece na Regressão Linear Simples não é, em geral, possível desdobrar a fórmula matricial/vectorial (3.6) em fórmulas individuais para cada b_j. Apenas esta fórmula matricial/vectorial nos permite obter os valores de cada parâmetro através do vector $\vec{\mathbf{b}}$.
4. Os parâmetros b_j obtidos definem, na representação tradicional dos dados em $\mathbb{R}^{p+1}$, o hiperplano que melhor se ajusta à nuvem de n pontos (no sentido de minimizar $SQRE$).

Vejamos agora propriedades que são verificadas por qualquer matriz de projecção ortogonal.

Proposição 3.2: Propriedades de matrizes de projecção ortogonal

Qualquer matriz de projecção ortogonal, como é a matriz $\mathbf{H}$, verifica as seguintes propriedades:

1. $\mathbf{H}$ é **simétrica**, ou seja: $\mathbf{H}^t = \mathbf{H}$;
2. $\mathbf{H}$ é **idempotente**, ou seja: $\mathbf{H}\,\mathbf{H} = \mathbf{H}$;
3. $\mathbf{H}$ deixa invariantes os vectores que já pertencem ao subespaço sobre o qual projecta, ou seja: se $\mathbf{z}$ pertence a $\mathcal{C}(\mathbf{X})$, então $\mathbf{Hz}=\mathbf{z}$.

Demonstração 3.1: da Proposição 3.2

Consideremos a matriz de projecção ortogonal sobre $\mathcal{C}(\mathbf{X})$, dada por $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$, onde $\mathbf{X}$ é a matriz do modelo, de dimensão $n \times (p+1)$. Ora,

1. Tendo em conta que $(\mathbf{X}^t\mathbf{X})^{-1}$ é a matriz inversa de $\mathbf{X}^t\mathbf{X}$, o produto $(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X} = \mathbf{I}_{p+1}$, a matriz identidade de dimensão $(p+1) \times (p+1)$. Logo:

$$\mathbf{H}\,\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\,\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\cancel{\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}}\mathbf{X}^t = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t = \mathbf{H}\ .$$

2. A simetria resulta de três propriedade conhecidas de matrizes: a transposta duma matriz transposta é a matriz original $((\mathbf{A}^t)^t = \mathbf{A})$; a transposta dum produto de matrizes é o produto das correspondentes transpostas, pela ordem inversa $((\mathbf{AB})^t = \mathbf{B}^t\mathbf{A}^t)$; e a transposta duma matriz inversa é a inversa da transposta $((\mathbf{A}^{-1})^t = (\mathbf{A}^t)^{-1})$. Tem-se:

$$\mathbf{H}^t = [\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t]^t = [\mathbf{X}^t]^t[(\mathbf{X}^t\mathbf{X})^{-1}]^t\mathbf{X}^t = \mathbf{X}[(\mathbf{X}^t\mathbf{X})^t]^{-1}\mathbf{X}^t = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t = \mathbf{H}\ .$$

3. Como foi visto na página 88, qualquer vector do subespaço das colunas da matriz $\mathbf{X}$, $\mathcal{C}(\mathbf{X}) \subset \mathbb{R}^n$, se pode escrever como $\mathbf{X}\vec{\mathbf{a}}$, onde $\vec{\mathbf{a}} \in \mathbb{R}^{p+1}$ é o vector dos $p+1$ coeficientes na combinação linear das colunas de $\mathbf{X}$. Ora, a projecção ortogonal deste vector sobre o subespaço $\mathcal{C}(\mathbf{X})$ (que já o contém) é dada por

$$\mathbf{H}\mathbf{X}\vec{\mathbf{a}} \;=\; \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t(\mathbf{X}\vec{\mathbf{a}}) \;=\; \mathbf{X}\cancel{(\mathbf{X}^t\mathbf{X})^{-1}(\mathbf{X}^t\mathbf{X})}\vec{\mathbf{a}} \;=\; \mathbf{X}\vec{\mathbf{a}}\ .$$

Assim, o vector $\mathbf{X}\vec{\mathbf{a}}$ fica igual após a projecção.

3.2.5 As três Somas de Quadrados

Na Regressão Linear Múltipla definem-se três Somas de Quadrados, de forma idêntica ao que se fez na Regressão Linear Simples.

Definição 3.3: As três Somas de Quadrados

Considere uma regressão linear múltipla, ajustada com n observações $\{(x_{1(i)}, x_{2(i)}, ..., x_{p(i)}, y_i\}_{i=1}^n$, e sejam $\hat{y}_i = b_0 + b_1\,x_{1(i)} + ... + b_p\,x_{p(i)}$ os correspondentes valores ajustados. Definem-as as três Somas de Quadrados de forma análoga ao que foi visto na RLS:

SQRE – Soma de Quadrados dos Resíduos: $SQRE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

SQT – Soma de Quadrados Total: $SQT = \sum_{i=1}^{n}(y_i - \overline{y})^2$

SQR – Soma de Quadrados associada à Regressão: $SQR = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$

Nota 3.2

Como se verá na Proposição 3.4 (pg.97), também na Regressão Linear Múltipla se verifica que *os y observados (y_i) e os y ajustados ($\hat{y}_i$) têm a mesma média*, logo que a média dos n resíduos é nula. Isso permite re-escrever SQT e SQR duma forma que será útil adiante. Tem-se:

$$SQT = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i^2 + \overline{y}^2 - 2\,y_i\,\overline{y}) = \sum_{o=1}^{n} y_i^2 + n\,\overline{y}^2 - 2\,\overline{y}\underbrace{\sum_{i=1}^{n} y_i}_{=n\,\overline{y}} = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2 \quad (3.7)$$

e ainda:

$$SQR = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i^2 + \overline{y}^2 - 2\,\hat{y}_i\,\overline{y}) = \sum_{i=1}^{n} \hat{y}_i^2 + n\,\overline{y}^2 - 2\,\overline{y}\underbrace{\sum_{i=1}^{n} \hat{y}_i}_{=n\,\overline{y}} = \sum_{i=1}^{n} \hat{y}_i^2 - n\overline{y}^2 \quad (3.8)$$

3.2.6 Propriedades duma Regressão Linear Múltipla descritiva

3.2.6.1 O Teorema de Pitágoras e a Fórmula Fundamental da Regressão Linear

O *Teorema de Pitágoras* é válido em qualquer espaço euclideano, como é o espaço das variáveis $\mathbb{R}^n$. Como sabemos, esse Teorema afirma que, num triângulo rectângulo, o quadrado do comprimento da hipotenusa (o lado oposto ao ângulo recto) é igual à soma dos quadrados dos comprimentos dos catetos (os lado adjacentes ao ângulo recto). Aplicado ao triângulo rectângulo da Figura 3.6 (pg.90), cuja hipotenusa é dada por $\vec{\mathbf{y}}$, e que tem o cateto $\vec{\hat{\mathbf{y}}}$ em $\mathcal{C}(\mathbf{X})$ e outro cateto dado por $\vec{\mathbf{y}} - \vec{\hat{\mathbf{y}}}$, produz a relação $\|\vec{\mathbf{y}}\|^2 = \|\vec{\hat{\mathbf{y}}}\|^2 + \|\vec{\mathbf{y}} - \vec{\hat{\mathbf{y}}}\|^2$. Como veremos na Proposição 3.3, esta igualdade equivale a afirmar que SQT=SQR+$SQRE$.

Proposição 3.3: Fórmula Fundamental da Regressão

No contexto duma Regressão Linear Múltipla, tem-se:

$$SQT = SQR + SQRE\ .$$

Demonstração 3.2: da Proposição 3.3

Aplicando o Teorema de Pitágoras ao triângulo rectângulo da Figura 3.6, e tendo em conta as expressões para SQT e SQR obtidas nas equações (3.7) e (3.8), tem-se:

$$\begin{aligned} \|\vec{\mathbf{y}}\|^2 &= \|\hat{\vec{\mathbf{y}}}\|^2 + \|\vec{\mathbf{y}} - \hat{\vec{\mathbf{y}}}\|^2 \\ \Leftrightarrow \quad \sum_{i=1}^{n} y_i^2 &= \sum_{i=1}^{n} \hat{y}_i^2 + \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}_{= SQRE} \\ \Leftrightarrow \quad \sum_{i=1}^{n} y_i^2 - n\overline{y}^2 &= \sum_{i=1}^{n} \hat{y}_i^2 - n\overline{y}^2 + SQRE \\ \Leftrightarrow \quad SQT &= SQR + SQRE \end{aligned}$$

Assim, a relação fundamental da Regressão Linear, $SQT = SQR + SQRE$, resulta duma aplicação do Teorema de Pitágoras. Mas foi necessário introduzir a subtracção de $n\overline{y}^2$ dos dois lados da equação, duma forma algo artificial. Um outro triângulo rectângulo é estatisticamente mais interessante.

3.2.6.2 Um triângulo rectângulo alternativo e o Coeficiente de Determinação R^2

Vamos contruir um novo triângulo rectângulo, a partir do vector *centrado* das observações de y e a sua projecção ortogonal sobre $\mathcal{C}(\mathbf{X})$.

Considere-se o *vector centrado* das observações da variável resposta, isto é, o vector cujo elemento genérico é $y_i - \overline{y}$. Este vector, que será representado por $\vec{\mathbf{y}}^c$, obtém-se subtraíndo a $\vec{\mathbf{y}}$ o vector que repete n vezes a média $\overline{y}$:

$$\vec{\mathbf{y}}^c = \begin{pmatrix} y_1 - \overline{y} \\ y_2 - \overline{y} \\ \vdots \\ y_n - \overline{y} \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} \overline{y} \\ \overline{y} \\ \vdots \\ \overline{y} \end{pmatrix} = \vec{\mathbf{y}} - \overline{y}\,\vec{\mathbf{1}}_n \ . \tag{3.9}$$

A *norma* deste vector é

$$\|\vec{\mathbf{y}}^c\| = \sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2} = \sqrt{SQT} \tag{3.10}$$

Ora, a projecção ortogonal do vector $\vec{\mathbf{y}}^c$ sobre o subespaço $\mathcal{C}(\mathbf{X})$ gera o vector $\mathbf{H}\vec{\mathbf{y}}^c$:

$$\begin{aligned} \mathbf{H}\vec{\mathbf{y}}^c &= \mathbf{H}\left(\vec{\mathbf{y}} - \overline{y}\,\vec{\mathbf{1}}_n\right) \\ \Leftrightarrow \quad \mathbf{H}\vec{\mathbf{y}}^c &= \mathbf{H}\vec{\mathbf{y}} - \overline{y}\,\mathbf{H}\vec{\mathbf{1}}_n \\ \Leftrightarrow \quad \mathbf{H}\vec{\mathbf{y}}^c &= \hat{\vec{\mathbf{y}}} - \overline{y}\,\vec{\mathbf{1}}_n \end{aligned}$$

já que $\mathbf{H}\vec{\mathbf{1}}_n = \vec{\mathbf{1}}_n$, pois o vector $\vec{\mathbf{1}}_n$ *já pertence ao subespaço* $\mathcal{C}(\mathbf{X})$, *logo fica invariante quando projectado nesse mesmo subespaço* (ver Proposição 3.2).

Esta última expressão significa que o vector $\mathbf{H}\vec{\mathbf{y}}^c$ tem elemento genérico $\hat{y}_i - \overline{y}$:

$$\mathbf{H}\vec{\mathbf{y}}^c \;=\; \vec{\hat{\mathbf{y}}} - \overline{y}\,\vec{\mathbf{1}}_n \;=\; \begin{pmatrix} \hat{y}_1 - \overline{y} \\ \hat{y}_2 - \overline{y} \\ \vdots \\ \hat{y}_n - \overline{y} \end{pmatrix} \tag{3.11}$$

Logo, a sua norma é:

$$\|\mathbf{H}\vec{\mathbf{y}}^c\| = \sqrt{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2} = \sqrt{SQR}\;. \tag{3.12}$$

Tendo em conta as expressões para $\vec{\mathbf{y}}^c$ (eq. 3.9) e $\mathbf{H}\vec{\mathbf{y}}^c$ (eq. 3.11), a distância entre o vector $\vec{\mathbf{y}}^c$ e a sua projecção ortogonal sobre $\mathcal{C}(\mathbf{X})$ continua a ser $\sqrt{SQRE}$:

$$\begin{aligned} & \vec{\mathbf{y}}^c - \mathbf{H}\vec{\mathbf{y}}^c \;=\; (\vec{\mathbf{y}} - \overline{y}\,\vec{\mathbf{1}}_n) - (\vec{\hat{\mathbf{y}}} - \overline{y}\,\vec{\mathbf{1}}_n) \\ \Leftrightarrow \quad & \vec{\mathbf{y}}^c - \mathbf{H}\vec{\mathbf{y}}^c \;=\; \vec{\mathbf{y}} - \vec{\hat{\mathbf{y}}} \end{aligned}$$

pelo que

$$\|\vec{\mathbf{y}}^c - \mathbf{H}\vec{\mathbf{y}}^c\| \;=\; \|\vec{\mathbf{y}} - \vec{\hat{\mathbf{y}}}\| \;=\; \sqrt{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \;=\; \sqrt{SQRE}\;.$$

Assim, *a fórmula fundamental da Regressão Linear, $SQT = SQR + SQRE$, é uma aplicação directa do Teorema de Pitágoras ao triângulo definido por $\vec{\mathbf{y}}^c$ e a sua projecção ortogonal sobre $\mathcal{C}(\mathbf{X})$*, como ilustrado na Figura 3.7.

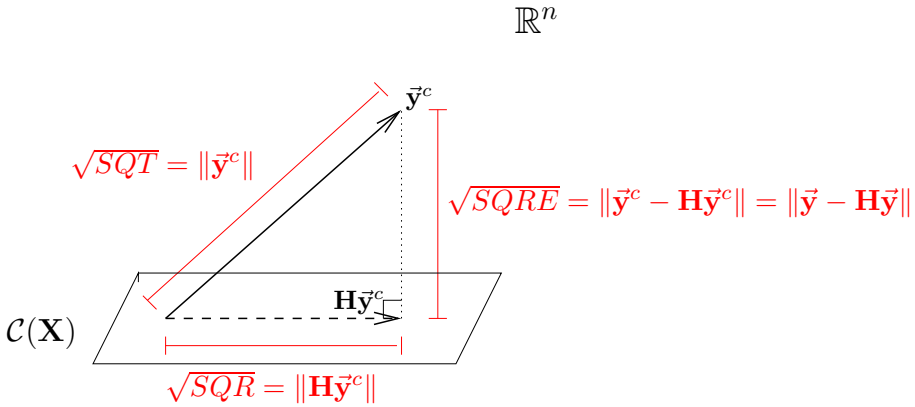


Figura 3.7: O triângulo rectangulo definido a partir do vector $\vec{\mathbf{y}}^c$ das observações *centradas* de y, $y_i - \overline{y}$, e da sua projecção ortogonal sobre $\mathcal{C}(\mathbf{X})$. A fórmula fundamental da regressão sai directamente da aplicação do Teorema de Pitágoras a este rectângulo.

Neste novo triângulo rectangulo, construído a partir da projecção ortogonal de $\vec{\mathbf{y}}^c$ sobre $\mathcal{C}(\mathbf{X})$, o Coeficiente de Determinação $R^2 = \frac{SQR}{SQT}$ também tem uma interpretação geométrica evidente: é o quadrado do cosseno do ângulo entre $\vec{\mathbf{y}}^c$ e $\mathbf{H}\vec{\mathbf{y}}^c$, como ilustrado na Figura 3.8 e afirmado na Definição 3.4.

Definição 3.4

Numa Regressão Linear, o Coeficiente de Determinação R^2 define-se como:

$$R^2 \quad = \quad \frac{SQR}{SQT} \quad = \quad \cos^2\theta$$

sendo θ o ângulo, no espaço das variáveis, entre o vector centrado das observações da variável resposta, $\vec{\mathbf{y}}^c$, e a sua projecção ortogonal no subespaço das colunas de $\mathbf{X}$, $\mathbf{H}\vec{\mathbf{y}}^c$ (ver Figura 3.8).

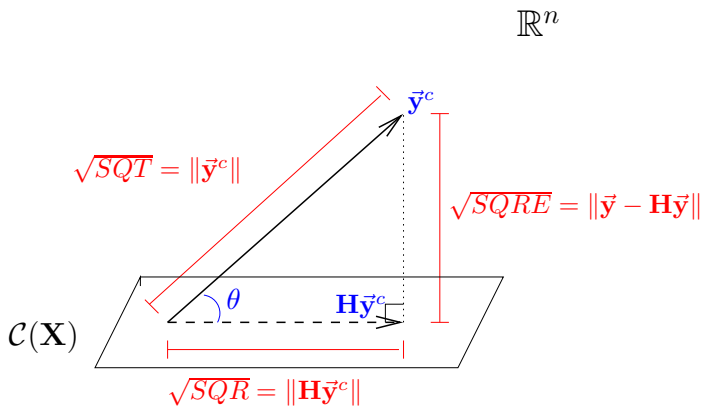


Figura 3.8: O Coeficiente de Determinação na Regressão Linear, $R^2 = \frac{SQR}{SQT}$, é o quadrado do cosseno do ângulo entre $\vec{\mathbf{y}}^c$ e $\mathbf{H}\vec{\mathbf{y}}^c$. Ou seja, $R^2 = \frac{SQR}{SQT} = \cos^2\theta$, sendo θ o ângulo entre $\vec{\mathbf{y}}^c$ e $\mathbf{H}\vec{\mathbf{y}}^c$.

3.2.6.3 Propriedades do Coeficiente de Determinação

A abordagem geométrica confirma que, também na Regressão Linear Múltipla, são válidas as seguintes propriedades do Coeficiente de Determinação, já conhecidas da Regressão Linear Simples:

- Sendo o quadrado dum cosseno, R^2 *toma valores entre* 0 *e* 1.
- *Quanto mais próximo de* 1 *estiver* R^2, menor o ângulo θ, e portanto *melhor será a correspondência entre o vector (centrado) das observações,* $\vec{\mathbf{y}}^c$, *e o seu ajustamento em* $\mathcal{C}(\mathbf{X})$, ou seja, mais próximos estarão, em geral os valores observados y_i e os correspondentes valores ajustados, $\hat{y}_i$.
- *Se* $R^2 \approx 0$, o vector $\vec{\mathbf{y}}^c$ é quase perpendicular ao subespaço $\mathcal{C}(\mathbf{X})$ onde se pretende aproximá-lo, e a projecção vai quase anular todas os elementos do vector projectado. De facto, se $\vec{\mathbf{y}}^c$ fosse exactamente perpendicular a $\mathcal{C}(\mathbf{X})$, a sua projecção $\mathbf{H}\vec{\mathbf{y}}^c$ seria o vector origem, com todos os elementos iguais a zero. Isso significaria, pela equação 3.11, que todos os valores ajustados seriam iguais ($\hat{y}_i = \overline{y}$, para qualquer i). $R^2 \approx 0$ corresponde ao *resultado ser de má qualidade*, uma vez que se perde quase toda a variabilidade nos valores ajustados de y ($SQR \approx 0$).

3.2.6.4 Propriedades do hiperplano de regressão

Numa regressão linear múltipla verificam-se ainda as propriedades da Proposição 3.4

Proposição 3.4: Propriedades do hiperplano de regressão

Sejam dados n conjuntos de observações $\{(x_{1_{(i)}}, x_{2_{(i)}}, ..., x_{p_{(i)}}, y_i)\}_{i=1}^n$ e seja ajustada a Regressão Linear Múltipla de y sobre as p variáveis preditoras $X_1, X_2, ..., X_p$, obtendo-se o hiperplano ajustado em $\mathbb{R}^{p+1}$, de equação $y = b_0 + b_1\, x_1 + b_2\, x_2 + ... + b_p\, x_p$. Verificam-se as seguintes propriedades.

1. *a média dos valores observados de* Y, $\{y_i\}_{i=1}^n$, *é igual à média dos respectivos valores ajustados,* $\{\hat{y}_i\}_{i=1}^n$.
2. *O hiperplano ajustado contém o centro de gravidade da nuvem de pontos*, i.e., o ponto de coordenadas $(\overline{x}_1, \overline{x}_2, ..., \overline{x}_p, \overline{y})$.
3. *os coeficientes* $\{b_j\}_{j=1}^p$ *que multiplicam variáveis preditoras interpretam-se como a variação (média) em* Y, *associada a aumentar a variável preditora correspondente em uma unidade, mantendo os restantes preditores constantes.*
4. o valor do coeficiente de determinação R^2 *numa regressão múltipla não pode ser inferior ao valor de* R^2 *que se obteria excluindo do modelo um qualquer subconjunto de preditores.* Em particular, não pode ser inferior ao R^2 das regressões lineares simples de Y sobre cada preditor individual.

Demonstração 3.3: da Proposição 3.4

1. A média $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ dos valores observados de Y pode ser escrita como $\frac{1}{n}\vec{\mathbf{1}}_n{}^t\vec{\mathbf{y}}$, já que o produto interno $\vec{\mathbf{1}}_n{}^t\vec{\mathbf{y}}$ calcula a soma dos elementos y_i do vector $\vec{\mathbf{y}}$ (confirme). Por definição, o vector dos valores ajustados é dado por $\vec{\hat{\mathbf{y}}} = \mathbf{H}\vec{\mathbf{y}}$. Ora, a média desses valores ajustados, $\overline{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i$, também pode ser calculada tomando o produto interno do vector $\vec{\mathbf{1}}_n$ de n uns com o vector $\vec{\hat{\mathbf{y}}}$, que faz a soma dos elementos de $\vec{\hat{\mathbf{y}}}$. Assim, a média dos valores ajustados é $\overline{\hat{y}} = \frac{1}{n}\vec{\mathbf{1}}_n{}^t\vec{\hat{\mathbf{y}}} = \frac{1}{n}\vec{\mathbf{1}}_n{}^t\mathbf{H}\vec{\mathbf{y}} = \frac{1}{n}(\mathbf{H}\vec{\mathbf{1}}_n)^t\vec{\mathbf{y}} = \frac{1}{n}\vec{\mathbf{1}}_n{}^t\vec{\mathbf{y}}$, uma vez que $\mathbf{H}\vec{\mathbf{1}}_n = \vec{\mathbf{1}}_n$, como se viu na Proposição 3.2. Mas a expressão final obtida, $\frac{1}{n}\vec{\mathbf{1}}_n{}^t\vec{\mathbf{y}}$, é a média $\overline{y}$ dos valores observados.

2. Partindo da igualdade da média dos valores observados e ajustados, tem-se:

$$\begin{aligned}
\overline{y} &= \overline{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \frac{1}{n}\sum_{i=1}^{n}(b_0 + b_1 x_{1_{(i)}} + b_2 x_{2_{(i)}} + b_3 x_{3_{(i)}} + \ldots + b_p x_{p_{(i)}}) \\
\Leftrightarrow \quad \overline{y} &= \frac{1}{n}\sum_{i=1}^{n} b_0 + \frac{1}{n}\sum_{i=1}^{n} b_1 x_{1_{(i)}} + \frac{1}{n}\sum_{i=1}^{n} b_2 x_{2_{(i)}} + \frac{1}{n}\sum_{i=1}^{n} b_3 x_{3_{(i)}} + \ldots + \frac{1}{n}\sum_{i=1}^{n} b_p x_{p_{(i)}} \\
\Leftrightarrow \quad \overline{y} &= \frac{1}{\cancel{n}}\cdot \cancel{n} b_0 + b_1 \underbrace{\frac{1}{n}\sum_{i=1}^{n} x_{1_{(i)}}}_{=\overline{x}_1} + b_2 \underbrace{\frac{1}{n}\sum_{i=1}^{n} x_{2_{(i)}}}_{=\overline{x}_2} + b_3 \underbrace{\frac{1}{n}\sum_{i=1}^{n} x_{3_{(i)}}}_{=\overline{x}_3} + \ldots + + b_p \underbrace{\frac{1}{n}\sum_{i=1}^{n} x_{p_{(i)}}}_{=\overline{x}_p} \\
\Leftrightarrow \quad \overline{y} &= b_0 + b_1\,\overline{x}_1 + b_2\,\overline{x}_2 + b_3\overline{x}_3 + \ldots + b_p\overline{x}_p
\end{aligned}$$

   Esta última equação significa que o ponto de coordenadas $(\overline{x}_1, \overline{x}_2, \overline{x}_3, ..., \overline{x}_p, \overline{y})$ satisfaz a equação do hiperplano, ou seja, o centro de gravidade da nuvem de pontos pertence ao hiperplano de regressão.

3. O valor de Y no hiperplano, quando $X_1 = x_1$, $X_2 = x_2$, $X_3 = x_3$, ... , $X_p = x_p$ (ou seja o valor médio de y para esses valores dos preditores), é dado por $y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + ... + b_p x_p$. O valor de Y no hiperplano, se aumentarmos um preditor, digamos X_2, em uma unidade, mantendo os restantes preditores constantes, é $y^* = b_0 + b_1 x_1 + b_2(x_2 + 1) + b_3 x_3 + ... + b_p x_p$. A diferenças destes dois valores é dada por:

$$\begin{array}{rcl}
y^* & = & \cancel{b_0} + \cancel{b_1 x_1} + \overbrace{b_2(x_2+1)}^{=\cancel{b_2 x_2} + b_2} + \cancel{b_3 x_3} + ... + \cancel{b_p x_p} \\
- \quad y & = & \cancel{b_0} + \cancel{b_1 x_1} + \cancel{b_2 x_2} + \cancel{b_3 x_3} + ... + \cancel{b_p x_p} \\
\hline
y^* - y & = & b_2
\end{array}$$

   Logo, b_2 representa o acréscimo em Y associado a aumentar X_2 em uma unidade, mantendo iguais os valores dos restantes preditores. Esta interpretação aplica-se naturalmente a um aumento de uma unidade no valor de outro preditor qualquer (mantendo os restantes iguais).

4. A combinação linear $\vec{\mathbf{y}} = \mathbf{H}\vec{\mathbf{y}} = \mathbf{X}\vec{\mathbf{b}} = b_0\vec{\mathbf{1}}_n + b_1\vec{\mathbf{x}}_1 + b_2\vec{\mathbf{x}}_2 + ... + b_p\vec{\mathbf{x}}_p$ é a que minimiza a Soma de Quadrados dos Resíduos, e por conseguinte maximiza R^2, ao longo de *todas* as possíveis combinações lineares. Entre essas possíveis combinações lineares encontram-se as que correspondem a ignorar alguns dos preditores (basta associar a esse preditores ignorados coeficientes $b_j = 0$). Assim, o valor de R^2 nunca pode ser inferior ao que se obteria ignorando alguns dos preditores.

3.2.7 A Regressão Múltipla no `R`

Uma Regressão Múltipla no `R` estuda-se através do mesmo comando `lm` usado para a regressão linear simples, apenas adaptando a fórmula. A fórmula através da qual se indica a variável resposta y e as variáveis preditoras $x_1, ..., x_p$ deve ser escrita separando, à direita do til, os nomes das variáveis preditoras por um sinal de adição. Por exemplo, se a variável resposta se chama `y` e existirem três preditores de

nomes `x1`, `x2` e `x3`, a fórmula que indica a relação será:

$$y \quad \sim \quad x1 \; + \; x2 \; + \; x3$$

O comando correspondente no `R` será:

```
>  lm ( y ~  x1 + x2 + x3 , data=dados)
```

O resultado produzido por este comando é o vector das estimativas dos $p+1$ parâmetros do modelo, b_0, b_1, ..., b_p, resultante de usar a fórmula da Proposição 3.1.

Exemplo 3.1: Dados dos lírios (`iris`)

Exemplifique-se de novo com os dados dos lírios. Pretende-se prever a variável resposta largura da pétala, não apenas a partir do comprimento da pétala, mas também das duas medições (largura e comprimento) das sépalas.

```
> iris2.lm <- lm(Petal.Width ~ Petal.Length + Sepal.Length + Sepal.Width , data=iris)
> iris2.lm
(...)
Coefficients:
 (Intercept)  Petal.Length  Sepal.Length   Sepal.Width
-0.2403         0.5241       -0.2073        0.2228
```

O *hiperplano ajustado*, em $\mathbb{R}^{p+1}$ tem assim a seguinte equação:

$$y \;=\; -0.2403 + 0.5241\, x_1 - 0.2073\, x_2 + 0.2228\, x_3 \;,$$

onde y indica a variável resposta `Petal.Width` , x_1 o primeiro preditor da fórmula, `Petal.Length`, x_2 o segundo preditor da formula, `Sepal.Length`, e x_3 o último preitor, `Sepal.Width`.

O Coeficiente de Determinação é $R^2 = 0.9379$, só ligeiramente maior que o valor $R^2 = 0.9271$ do modelo RLS (ver Subsecção 2.9.3).

3.3 Contexto inferencial: o Modelo RLM

Até aqui, apenas se considerou o problema descritivo: dados n conjuntos de observações da variável resposta Y e de p preditores $X_1, X_2, ..., X_p$, ou seja as n observações $\{(x_{1(i)}, x_{2(i)}, ..., x_{p(i)}, y_i)\}_{i=1}^{n}$, determinar os $p+1$ coeficientes $\vec{\mathbf{b}} = (b_0, b_1, b_2, ..., b_p)$ que minimizam a soma de quadrados de resíduos

$$SQRE \;=\; \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \;=\; \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_{1(i)} + b_2 x_{2(i)} + ... + b_p x_{p(i)})]^2.$$

Viu-se que a solução que minimiza $SQRE$ é dada pelo vector de coeficientes $\vec{\mathbf{b}} = (\mathbf{X}^t\mathbf{X})^{-1}\,\mathbf{X}^t\vec{\mathbf{y}}$.

Mas, tal como na Regressão Linear Simples, coloca-se o *problema inferencial* quando as n observações representam uma amostra aleatória de uma população mais vasta. É a relação populacional entre Y e as p variáveis preditoras que se pretende conhecer. Para esse fim, será necessário *admitir alguns pressupostos adicionais*.

3.3.1 O Modelo RLM para observações individuais

Vamos admitir que, na população, existe um hiperplano em $\mathbb{R}^{p+1}$ que descreve a relação de fundo entre a variável resposta Y e as p variáveis preditoras X_1, X_2, ... , X_p. A equação desse hiperplano é semelhante à equação amostral (3.2), mas com um conjunto de parâmetros que designaremos por β_j ($j=0,1,2,...,p$):

$$y \;=\; \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \;. \tag{3.13}$$

Admite-se ainda que as n observações da variável resposta, Y_i, são *aleatórias* e podem ser modeladas pela equação seguinte, que incorpora variabilidade aleatória de cada observação em torno do hiperplano populacional:

$$Y_i \;=\; \beta_0 + \beta_1 x_{1(i)} + \beta_2 x_{2(i)} + ... + \beta_p x_{p(i)} \;+\; \epsilon_i \;, \qquad i = 1, ..., n$$

Admitem-se válidos os restantes pressupostos do modelo RLS, relativos aos erros aleatórios ϵ_i.

Definição 3.5: O Modelo da Regressão Linear Múltipla - RLM

Considerem-se n observações $\{(x_{1(i)}, x_{2(i)}, ..., x_{p(i)}, Y_i)\}_{i=1}^n$, em que Y_i representa uma observação da variável aleatória resposta e os restantes valores são fixados pelo experimentador. O **Modelo de Regressão Linear Múltipla (RLM)** verifica os seguintes pressupostos:

1. $Y_i \;=\; \beta_0 + \beta_1 x_{1_{(i)}} + \beta_2 x_{2_{(i)}} + ... + \beta_p x_{p_{(i)}} + \epsilon_i, \quad \forall i = 1, ..., n.$
2. $\epsilon_i \;\frown\; \mathcal{N}(0\,,\,\sigma^2), \quad \forall i = 1, ..., n.$
3. $\{\epsilon_i\}_{i=1}^n$ v.a. independentes.

Por um raciocínio análogo ao da Proposição 3.4, a constante β_j ($j = 1, 2, ..., p$) que multiplica a variável X_j pode ser interpretada como a *variação esperada em Y, associada a aumentar X_j em uma unidade, mantendo as restantes variáveis constantes.*

A notação matricial/vectorial. As equações do modelo para as n observações podem ser escritas como *uma única equação* utilizando notação vectorial/matricial:

$$\begin{array}{rcccccc}
Y_1 & = & \beta_0 + \beta_1 x_{1(1)} + \beta_2 x_{2(1)} + \cdots + \beta_p x_{p(1)} & + & \epsilon_1 \\
Y_2 & = & \beta_0 + \beta_1 x_{1(2)} + \beta_2 x_{2(2)} + \cdots + \beta_p x_{p(2)} & + & \epsilon_2 \\
Y_3 & = & \beta_0 + \beta_1 x_{1(3)} + \beta_2 x_{2(3)} + \cdots + \beta_p x_{p(3)} & + & \epsilon_3 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
Y_n & = & \beta_0 + \beta_1 x_{1(n)} + \beta_2 x_{2(n)} + \cdots + \beta_p x_{p(n)} & + & \epsilon_n
\end{array}$$

As n equações correspondem a *uma única equação matricial*:

$$\vec{\mathbf{Y}} \;=\; \mathbf{X}\vec{\boldsymbol{\beta}} \;+\; \vec{\boldsymbol{\epsilon}}\,,$$

onde

$$\vec{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix}, \; \mathbf{X} = \begin{bmatrix} 1 & x_{1_{(1)}} & x_{2_{(1)}} & \cdots & x_{p_{(1)}} \\ 1 & x_{1_{(2)}} & x_{2_{(2)}} & \cdots & x_{p_{(2)}} \\ 1 & x_{1_{(3)}} & x_{2_{(3)}} & \cdots & x_{p_{(3)}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1_{(n)}} & x_{2_{(n)}} & \cdots & x_{p_{(n)}} \end{bmatrix}, \; \vec{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \; \vec{\boldsymbol{\epsilon}} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Nesta equação, tem-se:

$\vec{\mathbf{Y}}$ é um *vector aleatório* das n variáveis aleatórias *resposta*;

$\mathbf{X}$ é a *matriz do modelo* (*não aleatória*) de dimensões $n \times (p+1)$ cujas colunas são dadas pelas observações de cada variável preditora (e por uma coluna inicial de uns, associada a constante aditiva do modelo);

$\vec{\boldsymbol{\beta}}$ é um vector (*não aleatório*) dos $p+1$ parâmetros do modelo;

$\vec{\boldsymbol{\epsilon}}$ é um *vector aleatório* dos n *erros aleatórios*.

Representa-se um vector de n *valores observados* de Y com uma letra minúscula: $\vec{\mathbf{y}}$.

Com alguns *conceitos adicionais* podemos escrever também os pressupostos relativos aos erros aleatórios em notação vectorial/matricial.

3.3.2 Ferramentas para vectores aleatórios

3.3.2.1 Propriedades operatórias de vectores esperados e matrizes de (co)variâncias

O vector de n componentes $\vec{\mathbf{Y}}$, tal como o vector dos n erros aleatórios, $\vec{\boldsymbol{\epsilon}}$, constituem *vectores aleatórios*.

Definição 3.6: Vector esperado e matriz de (co)variâncias dum vector aleatório

Para qualquer *vector aleatório* $\vec{\mathbf{W}} = (W_1, W_2, ..., W_k)^t$, define-se:

- O **vector esperado** de $\vec{\mathbf{W}}$, constituído pelos *valores esperados de cada componente*:

$$E[\vec{\mathbf{W}}] \;=\; \begin{bmatrix} E[W_1] \\ E[W_2] \\ \vdots \\ E[W_k] \end{bmatrix}.$$

- a **matriz de variâncias-covariâncias** de $\vec{\mathbf{W}}$ é constituída pelas (co)variâncias de cada par

de componentes:

$$V[\vec{\mathbf{W}}] = \begin{bmatrix} V[W_1] & Cov[W_1,W_2] & Cov[W_1,W_3] & \dots & Cov[W_1,W_k] \\ Cov[W_2,W_1] & V[W_2] & Cov[W_2,W_3] & \dots & Cov[W_2,W_k] \\ Cov[W_3,W_1] & Cov[W_3,W_2] & V[W_3] & \dots & Cov[W_3,W_k] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Cov[W_k,W_1] & Cov[W_k,W_2] & Cov[W_k,W_3] & \dots & V[W_k] \end{bmatrix}$$

A matriz de (co)variâncias dum vector aleatório é necessariamente uma *matriz simétrica*, já que para qualquer i e j se verifica $Cov[W_i, W_j] = Cov[W_j, W_i]$.

Tal como para o caso de variáveis aleatórias, também o vector esperado de um vector aleatório $\vec{\mathbf{W}}_{k\times 1}$ tem propriedades operatórias simples. A Proposição seguinte enuncia essas propriedades.

Proposição 3.5: Propriedades de vectores esperados

Para qualquer *vector aleatório* $\vec{\mathbf{W}} = (W_1, W_2, ..., W_k)^t$ verificam-se as seguintes propriedades.

- Se α é um escalar *não aleatório*, $E[\alpha\vec{\mathbf{W}}] = \alpha\, E[\vec{\mathbf{W}}]$.
- Se $\vec{\mathbf{a}}_{k\times 1}$ é um vector *não aleatório*, $E[\vec{\mathbf{W}} + \vec{\mathbf{a}}] = E[\vec{\mathbf{W}}] + \vec{\mathbf{a}}$.
- Se $\vec{\mathbf{a}}_{k\times 1}$ é um vector *não aleatório*, $E[\vec{\mathbf{a}}^t\, \vec{\mathbf{W}}] = \vec{\mathbf{a}}^t\, E[\vec{\mathbf{W}}]$.
- Se $\mathbf{B}_{m\times k}$ é uma matriz *não aleatória*, $E[\mathbf{B}\vec{\mathbf{W}}] = \mathbf{B}\, E[\vec{\mathbf{W}}]$.
- Se $\vec{\mathbf{W}}_{k\times 1}$, $\vec{\mathbf{U}}_{k\times 1}$ são vectores *aleatórios*, $E[\vec{\mathbf{W}} + \vec{\mathbf{U}}] = E[\vec{\mathbf{W}}] + E[\vec{\mathbf{U}}]$.

Demonstração 3.4: Proposição 3.5

Tendo em conta a definição de vector esperado e de matriz de variâncias, bem como as propriedades de valores esperados, variâncias e covariâncias de variáveis aleatórias unidimensionais, tem-se:

1. Por definição, $\alpha\vec{\mathbf{W}} = (\alpha W_1, \alpha W_2, ..., \alpha W_k)^t$. Logo,

$$E[\alpha\vec{\mathbf{W}}] = \begin{bmatrix} E[\alpha\, W_1] \\ E[\alpha\, W_2] \\ \vdots \\ E[\alpha\, W_k] \end{bmatrix} = \begin{bmatrix} \alpha\, E[W_1] \\ \alpha\, E[W_2] \\ \vdots \\ \alpha\, E[W_k] \end{bmatrix} = \alpha \begin{bmatrix} E[W_1] \\ E[W_2] \\ \vdots \\ E[W_k] \end{bmatrix} = \alpha\, E[\vec{\mathbf{W}}]\ .$$

2. Por definição, $\vec{\mathbf{W}} + \vec{\mathbf{a}} = (W_1 + a_1, W_2 + a_2, ..., W_k + a_k)^t$. Logo,

$$E[\vec{\mathbf{W}} + \vec{\mathbf{a}}] = \begin{bmatrix} E[W_1 + a_1] \\ E[W_2 + a_2] \\ \vdots \\ E[W_k + a_k] \end{bmatrix} = \begin{bmatrix} E[W_1] + a_1 \\ E[W_2] + a_2 \\ \vdots \\ E[W_k] + a_k \end{bmatrix} = \begin{bmatrix} E[W_1] \\ E[W_2] \\ \vdots \\ E[W_k] \end{bmatrix} + \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix} = E[\vec{\mathbf{W}}] + \vec{\mathbf{a}} \, .$$

3. O produto interno $\vec{\mathbf{a}}^t \vec{\mathbf{W}}$ é, por definição, $\vec{\mathbf{a}}^t \vec{\mathbf{W}} = a_1 W_1 + a_2 W_2 + ... + a_k W_k$. Logo,

$$E[\vec{\mathbf{a}}^t \vec{\mathbf{W}}] = E[a_1 W_1 + a_2 W_2 + ... + a_k W_k] = a_1 E[W_1] + a_2 E[W_2] + ... + a_k E[W_k] = \vec{\mathbf{a}}^t E[\vec{\mathbf{W}}] \, .$$

4. O produto matricial $\mathbf{B}_{m \times k} \vec{\mathbf{W}}_{k \times 1}$ é possível e resulta num vector de dimensão $m \times 1$. Cada elemento desse vector resulta do produto interno da linha i da matriz $\mathbf{B}$ (que designaremos $\vec{\mathbf{b}}^t_{[i]}$) com o vector $\vec{\mathbf{W}}$, ou seja, $\vec{\mathbf{b}}^t_{[i]} \vec{\mathbf{W}}$. Pela alínea anterior, o vector esperado de $\mathbf{B}\vec{\mathbf{W}}$ terá na posição i o produto interno $\vec{\mathbf{b}}^t_{[i]} E[\vec{\mathbf{W}}]$. Logo, o vector $E[\mathbf{B}\vec{\mathbf{W}}] = \mathbf{B} E[\vec{\mathbf{W}}]$.

5. Por definição, a soma dos vectores $\vec{\mathbf{W}} + \vec{\mathbf{U}}$ é um vector (de dimensão igual à de $\vec{\mathbf{W}}$ e $\vec{\mathbf{U}}$), que resulta de somar os elementos correspondentes de cada vector. Logo, na posição i de $\vec{\mathbf{W}} + \vec{\mathbf{U}}$ tem-se o elemento $W_i + U_i$. Assim,

$$E[\vec{\mathbf{W}} + \vec{\mathbf{U}}] = \begin{bmatrix} E[W_1 + U_1] \\ E[W_2 + U_2] \\ \vdots \\ E[W_k + U_k] \end{bmatrix} = \begin{bmatrix} E[W_1] + E[U_1] \\ E[W_2] + E[U_2] \\ \vdots \\ E[W_k] + E[U_k] \end{bmatrix} = \begin{bmatrix} E[W_1] \\ E[W_2] \\ \vdots \\ E[W_k] \end{bmatrix} + \begin{bmatrix} E[U_1] \\ E[U_2] \\ \vdots \\ E[U_k] \end{bmatrix} = E[\vec{\mathbf{W}}] + E[\vec{\mathbf{U}}] \, .$$

Na Proposição 3.6 enunciam-se algumas propriedades operatórias das *matrizes de variâncias-covariâncias* de vectores aleatórios.

Proposição 3.6: Propriedades das matrizes de (co)variâncias

- Se α é um escalar *não aleatório*, $V[\alpha \vec{\mathbf{W}}] = \alpha^2 \, V[\vec{\mathbf{W}}]$.
- Se $\vec{\mathbf{a}}_{k \times 1}$ é um vector *não aleatório*, $V[\vec{\mathbf{W}} + \vec{\mathbf{a}}] = V[\vec{\mathbf{W}}]$.
- Se $\vec{\mathbf{a}}_{k \times 1}$ é um vector *não aleatório*, $V[\vec{\mathbf{a}}^t \, \vec{\mathbf{W}}] = \vec{\mathbf{a}}^t \, V[\vec{\mathbf{W}}] \, \vec{\mathbf{a}}$.
- Se $\mathbf{B}_{m \times k}$ é uma matriz *não aleatória*, $V[\mathbf{B}\vec{\mathbf{W}}] = \mathbf{B} \, V[\vec{\mathbf{W}}] \, \mathbf{B}^t$.
- Se $\vec{\mathbf{W}}_{k \times 1}$ e $\vec{\mathbf{U}}_{k \times 1}$ forem vectores *aleatórios independentes*, $V[\vec{\mathbf{W}} + \vec{\mathbf{U}}] = V[\vec{\mathbf{W}}] + V[\vec{\mathbf{U}}]$.

Demonstração 3.5: Proposição 3.6

1. Sabemos que, para *variáveis* aleatórias, constantes multiplicativas saltam para fora das va-

riâncias elevadas ao quadrado ($V[\alpha X] = \alpha^2 V[X]$). Em covariâncias entre variáveis aleatórias, constantes multiplicativas em cada argumento saltam para fora ($Cov[\alpha X, \beta Y] = \alpha\beta\, Cov[X,Y]$). Logo,

$$V[\alpha\vec{\mathbf{W}}] = \begin{bmatrix} V[\alpha W_1] & Cov[\alpha W_1, \alpha W_2] & \cdots & Cov[\alpha W_1, \alpha W_k] \\ Cov[\alpha W_2, \alpha W_1] & V[\alpha W_2] & \cdots & Cov[\alpha W_2, \alpha W_k] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[\alpha W_k, \alpha W_1] & Cov[\alpha W_k, \alpha W_2] & \cdots & V[\alpha W_k] \end{bmatrix}$$

$$= \begin{bmatrix} \alpha^2 V[W_1] & \alpha^2 Cov[W_1, W_2] & \cdots & \alpha^2 Cov[W_1, W_k] \\ \alpha^2 Cov[W_2, W_1] & \alpha^2 V[W_2] & \cdots & \alpha^2 Cov[W_2, W_k] \\ \vdots & \vdots & \ddots & \vdots \\ \alpha^2 Cov[W_k, W_1] & \alpha^2 Cov[W_k, W_2] & \cdots & \alpha^2 V[W_k] \end{bmatrix} = \alpha^2\, V[\vec{\mathbf{W}}]$$

2. Por definição, $\vec{\mathbf{W}} + \vec{\mathbf{a}} = (W_1 + a_1, W_2 + a_2, ..., W_k + a_k)^t$. Sabemos ainda que, para a e b constantes, $V[X + a] = V[X]$ e $Cov[X + a, Y + b] = Cov[X, Y]$. Logo,

$$V[\vec{\mathbf{W}}+\vec{\mathbf{a}}] = \begin{bmatrix} V[W_1+a_1] & Cov[W_1+a_1, W_2+a_2] & \cdots & Cov[W_1+a_1, W_k+a_k] \\ Cov[W_2+a_2, W_1+a_1] & V[W_2+a_2] & \cdots & Cov[W_2+a_2, W_k+a_k] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[W_k+a_k, W_1+a_1] & Cov[W_k+a_k, W_2+a_2] & \cdots & V[W_k+a_k] \end{bmatrix}$$

$$= \begin{bmatrix} V[W_1] & Cov[W_1, W_2] & \cdots & Cov[W_1, W_k] \\ Cov[W_2, W_1] & V[W_2] & \cdots & Cov[W_2, W_k] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[W_k, W_1] & Cov[W_k, W_2] & \cdots & V[W_k] \end{bmatrix} = V[\vec{\mathbf{W}}]$$

3. Apenas se considera o caso mais simples ($k = 2$), ou seja, dum vector aleatório $\vec{\mathbf{W}} = \begin{bmatrix} W_1 \\ W_2 \end{bmatrix}$ e um vector constante $\vec{\mathbf{a}} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$. Tem-se $\vec{\mathbf{a}}^t\vec{\mathbf{W}} = a_1 W_1 + a_2 W_2$. Logo (e como $V[X + Y] = V[X] + V[Y] + 2\, Cov[X, Y]$),

$$\begin{aligned} V[\vec{\mathbf{a}}^t\vec{\mathbf{W}}] &= V[a_1 W_1 + a_2 W_2] = V[a_1 W_1] + V[a_2 W_2] + 2\, Cov[a_1 W_1, a_2 W_2] \\ &= a_1^2\, V[W_1] + a_2^2\, V[W_2] + 2\, a_1 a_2\, Cov[W_1, W_2]\ . \end{aligned}$$

Mas o produto do lado direito do resultado produz a mesma expressão:

$$\begin{aligned}
\vec{\mathbf{a}}^t V[\vec{\mathbf{W}}]\vec{\mathbf{a}} &= \begin{bmatrix} a_1 & a_2 \end{bmatrix} \begin{bmatrix} V[W_1] & Cov[W_1,W_2] \\ Cov[W_2,W_1] & V[W_2] \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \\
&= \begin{bmatrix} a_1 & a_2 \end{bmatrix} \begin{bmatrix} a_1\,V[W_1] + a_2\,Cov[W_1,W_2] \\ a_1\,Cov[W_2,W_1] + a_2\,V[W_2] \end{bmatrix} \\
&= a_1^2\,V[W_1] + a_1 a_2\,Cov[W_1,W_2] + a_2 a_1\,Cov[W_2,W_1] + a_2^2\,V[W_2] \\
&= a_1^2\,V[W_1] + a_2^2\,V[W_2] + 2\,a_1 a_2\,Cov[W_1,W_2]
\end{aligned}$$

.

A demonstração para o caso de vectores de dimensão $k > 2$ é análoga.

4. Omite-se a demonstração.

5. Por definição, $\vec{\mathbf{W}} + \vec{\mathbf{U}} = (W_1 + U_1, W_2 + U_2, ..., W_k + U_k)^t$. Sabemos ainda que, para X e Y variáveis aleatórias *independentes*, $V[X + Y] = V[X] + V[Y]$, uma vez que $Cov(X, Y) = 0$. Logo,

$$V[\vec{\mathbf{W}}{+}\vec{\mathbf{U}}] = \begin{bmatrix} V[W_1{+}U_1] & Cov[W_1{+}U_1, W_2{+}U_2] & \cdots & Cov[W_1{+}U_1, W_k{+}U_k] \\ Cov[W_2{+}U_2, W_1{+}U_1] & V[W_2{+}U_2] & \cdots & Cov[W_2{+}U_2, W_k{+}U_k] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[W_k{+}U_k, W_1{+}U_1] & Cov[W_k{+}U_k, W_2{+}U_2] & \cdots & V[W_k{+}U_k] \end{bmatrix}$$

A independência dos vectores $\vec{\mathbf{W}}$ e $\vec{\mathbf{U}}$ significa que, para qualquer componente i, W_i e U_i são independentes, logo os elementos diagonais da matriz simplificam: $V[W_i{+}U_i]{=}V[W_i]{+}V[U_i]$.

Por outro lado, as expressões fora da diagonal (envolvendo as covariâncias) também se podem simplificar. Aplicando repetidamente a quinta propriedade da covariância (Subsecção 2.4.2.4), que afirma que $Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)$, tem-se:

$$\begin{aligned}
Cov(W_i + U_i, W_j + U_j) &= Cov(W_i, W_j + U_j) + Cov(U_i, W_j + U_j) \\
&= Cov(W_i, W_j) + Cov(W_i, U_j) + Cov(U_i, W_j) + Cov(U_i, U_j) \\
&= Cov(W_i, W_j) + Cov(U_i, U_j)\ ,
\end{aligned}$$

pois, pela independência de $\vec{\mathbf{W}}$ e $\vec{\mathbf{U}}$, $Cov(W_i, U_j){=}Cov(U_i, W_j){=}0$. Assim, $V[\vec{\mathbf{W}}{+}\vec{\mathbf{U}}]$ fica:

$$\begin{bmatrix} V[W_1]{+}V[U_1] & Cov[W_1,W_2]{+}Cov[U_1,U_2] & \cdots & Cov[W_1,W_k]{+}Cov[U_1,U_k] \\ Cov[W_2,W_1]{+}Cov[U_2,U_1] & V[W_2]{+}V[U_2] & \cdots & Cov[W_2,W_k]{+}Cov[U_2,U_k] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[W_k,W_1]{+}Cov[U_k,U_1] & Cov[W_k,W_2]{+}Cov[U_k,U_2] & \cdots & V[W_k]{+}V[U_k] \end{bmatrix}$$

Donde:

$$
\begin{aligned}
V[\vec{\mathbf{W}}+\vec{\mathbf{U}}] &= \begin{bmatrix} V[W_1] & Cov[W_1,W_2] & \cdots & Cov[W_1,W_k] \\ Cov[W_2,W_1] & V[W_2] & \cdots & Cov[W_2,W_k] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[W_k,W_1] & Cov[W_k,W_2] & \cdots & V[W_k] \end{bmatrix} + \\
&\quad + \begin{bmatrix} V[U_1] & Cov[U_1,U_2] & \cdots & Cov[U_1,U_k] \\ Cov[U_2,U_1] & V[U_2] & \cdots & Cov[U_2,U_k] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[U_k,U_1] & Cov[U_k,U_2] & \cdots & V[U_k] \end{bmatrix} \\
&= V[\vec{\mathbf{W}}] + V[\vec{\mathbf{U}}]
\end{aligned}
$$

3.3.2.2 A distribuição Multinormal e suas propriedades

Vectores aleatórios têm também *distribuições (multivariadas) de probabilidades*. Para vectores aleatórios *contínuos* $\vec{\mathbf{W}}_{k\times 1}$, a distribuição é caracterizada por uma *função densidade conjunta*, uma função real de k variáveis: $f:\mathbb{R}^k \to \mathbb{R}$. A distribuição multivariada de vectores aleatórios mais usada é a *Multinormal* ou *Normal multivariada*, que generaliza a distribuição Normal para vectores aleatórios.

Definição 3.7: Distribuição Normal Multivariada

O vector aleatório k-dimensional $\vec{\mathbf{W}}$ tem **distribuição Multinormal**, com *parâmetros* dados pelo vector $\vec{\boldsymbol{\mu}} \in \mathbb{R}^k$ e a matriz $\boldsymbol{\Sigma}_{k\times k}$, se a sua função densidade conjunta fôr:

$$
f(\vec{\mathbf{w}}) = \frac{1}{(2\pi)^{k/2}\sqrt{det(\boldsymbol{\Sigma})}}\, \mathrm{e}^{-\frac{1}{2}(\vec{\mathbf{w}}-\vec{\boldsymbol{\mu}})^t\, \boldsymbol{\Sigma}^{-1}(\vec{\mathbf{w}}-\vec{\boldsymbol{\mu}})} \quad , \quad \vec{\mathbf{w}} \in \mathbb{R}^k. \tag{3.14}
$$

Se $\vec{\mathbf{W}}$ tem essa distribuição Normal multivariada, escreve-se: $\vec{\mathbf{W}} \frown \mathcal{N}_k(\vec{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$.

Para um vector aleatório com k componentes, $\vec{\mathbf{W}} = (W_1, W_2, ..., W_k)^t$, o gráfico da densidade Normal multivariada é uma hipersuperfície em $\mathbb{R}^{k+1}$ (sendo necessários k eixos para indicar os k valores das componentes de $\vec{\mathbf{W}}$ e mais um eixo para indicar o valor da função densidade conjunta nesse ponto). Assim, apenas é possível visualizar graficamente a densidade duma Normal bivariada, ou seja, um vector Multinormal com $k=2$ componentes, dado por uma superfície em $\mathbb{R}^3$, que é representada na Figura 3.9. É característica desta superfície que, sendo cortada por planos verticais, obtêm-se curvas Normais.

Enunciemos agora as propriedades fundamentais das distribuições Normais Multivariadas.

Proposição 3.7: Propriedades da Multinormal

Se $\vec{\mathbf{W}} \frown \mathcal{N}_k(\vec{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$, verificam-se as seguintes propriedades:

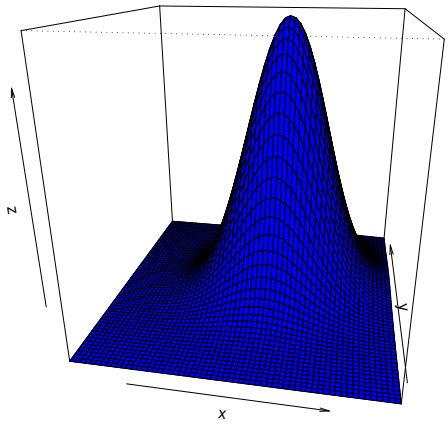


Figura 3.9: A densidade Binormal (Multinormal com $k = 2$).

1. O *vector esperado* de $\vec{\mathbf{W}}$ é $E[\vec{\mathbf{W}}] = \vec{\boldsymbol{\mu}}$.
2. A *matriz de (co)variâncias* de $\vec{\mathbf{W}}$ é $V[\vec{\mathbf{W}}] = \boldsymbol{\Sigma}$.
3. Se duas componentes de $\vec{\mathbf{W}}$ têm covariância nula, são independentes: $Cov[W_i, W_j] = 0 \Rightarrow W_i$, W_j independentes.
4. Todas as distribuições marginais de $\vec{\mathbf{W}}$ são (multi)normais. Em particular, *cada componente W_i é normal com média μ_i e variância $\boldsymbol{\Sigma}_{(i,i)}$*: $W_i \frown \mathcal{N}(\mu_i, \boldsymbol{\Sigma}_{(i,i)})$.
5. Se $\vec{\mathbf{a}}$ um vector (não-aleatório) $k \times 1$, então $\vec{\mathbf{W}} + \vec{\mathbf{a}} \frown \mathcal{N}_k(\vec{\boldsymbol{\mu}} + \vec{\mathbf{a}}, \boldsymbol{\Sigma})$.
6. *Combinações lineares das componentes dum vector multinormal são Normais*: $\vec{\mathbf{a}}^t\vec{\mathbf{W}} = a_1 W_1 + a_2 W_2 + ... + a_k W_k \frown \mathcal{N}(\vec{\mathbf{a}}^t\vec{\boldsymbol{\mu}}, \vec{\mathbf{a}}^t\boldsymbol{\Sigma}\vec{\mathbf{a}})$.
7. Se $\mathbf{B}$ é matriz $m \times k$ (não aleatória, de característica $m \leq k$), então $\mathbf{B}\vec{\mathbf{W}} \frown \mathcal{N}_m(\mathbf{B}\vec{\boldsymbol{\mu}}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^t)$.

Notas:

1. Omite-se a demonstração.
2. Nas disciplinas introdutórias de Estatística mostra-se que se X, Y são variáveis aleatórias independentes, então $cov[X, Y] = 0$. Agora sabemos que, *quando a distribuição conjunta de X e Y é Multinormal, tem-se também a implicação contrária.*
3. Qualquer elemento nulo numa matriz de (co)variâncias duma Multinormal indica que as compo-

nentes correspondentes são independentes.

3.3.3 Modelo Regressão Linear Múltipla - versão matricial

Definição 3.8: O Modelo RLM em notação matricial

Sejam dados os vectores aleatórios de observações da variável resposta, $\vec{\mathbf{Y}}$, e dos erros aleatórios, $\vec{\boldsymbol{\epsilon}}$, bem como a matriz do modelo $\mathbf{X}$ e o vector dos parâmetros, $\vec{\boldsymbol{\beta}}$, como definidos na Subsecção 3.3.1. Então o Modelo RLM consiste em admitir que:

1. $\vec{\mathbf{Y}} = \mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\epsilon}}$.
2. $\vec{\boldsymbol{\epsilon}} \frown \mathcal{N}_n(\vec{\mathbf{0}}, \sigma^2 \mathbf{I}_n)$, sendo $\mathbf{I}_n$ a *matriz identidade* $n \times n$.

Na segunda destas hipóteses são feitas quatro afirmações (tendo em conta as propriedades da Multinormal, referidas atrás):

- Cada erro aleatório individual ϵ_i tem distribuição Normal.
- Cada erro aleatório individual tem média zero: $E[\epsilon_i] = 0$.
- Cada erro aleatório individual tem variância igual: $V[\epsilon_i] = \sigma^2$.
- Erros aleatórios diferentes são independentes, porque $Cov[\epsilon_i, \epsilon_j] = 0$ se $i \neq j$ e, numa Multinormal, isso implica a independência.

3.3.3.1 A distribuição das observações $\vec{\mathbf{Y}}$ da variável resposta

O seguinte Teorema é consequência directa de aplicar as propriedades da Proposição 3.7 ao Modelo de Regressão Linear Múltipla.

Proposição 3.8: Primeiras Consequências do Modelo

Dado o Modelo de Regressão Linear Múltipla, tem-se:

$$\vec{\mathbf{Y}} \frown \mathcal{N}_n(\mathbf{X}\vec{\boldsymbol{\beta}}, \sigma^2 \mathbf{I}_n).$$

Demonstração 3.6: Proposição 3.8

Segundo o Modelo RLM (Definição 3.8), o vector aleatório $\vec{\mathbf{Y}}$ das observações da variável resposta é a soma dum vector não aleatório ($\mathbf{X}\vec{\boldsymbol{\beta}}$) e um vector aleatório dos erros ($\vec{\boldsymbol{\epsilon}}$), sendo este último um vector Multinormal. Assim, tem a natureza das somas $\vec{\mathbf{a}} + \vec{\mathbf{W}}$ que foram estudados nas Proposições

3.5, 3.6 e 3.7 da Subsecção 3.3.2.1.

Tendo em conta as propriedades da distribuição Multinormal (Proposição 3.7), ao somar um vector aleatório Multinormal ($\vec{\mathbf{W}} = \vec{\boldsymbol{\epsilon}}$) e um vector não aleatório ($\vec{\mathbf{a}} = \mathbf{X}\vec{\boldsymbol{\beta}}$), mantém-se a Multinormalidade. Falta determinar os seus dois parâmetros: o vector esperado e a matriz de (co-)variâncias. As propriedades operatórias de vectores esperados (Proposição 3.5) e matrizes de (co-)variâncias (Proposição 3.6) indicam que se tem:

$$E[\vec{\mathbf{Y}}] = E[\mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\epsilon}}] = \mathbf{X}\vec{\boldsymbol{\beta}} + \underbrace{E[\vec{\boldsymbol{\epsilon}}]}_{=\vec{\mathbf{0}}} = \mathbf{X}\vec{\boldsymbol{\beta}} \ , \tag{3.15}$$

$$V[\vec{\mathbf{Y}}] = V[\mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\epsilon}}] = V[\vec{\boldsymbol{\epsilon}}] = \sigma^2\mathbf{I}. \tag{3.16}$$

Logo, fica provada a distribuição do vector $\vec{\mathbf{Y}}$ quando é válido o Modelo RLM.

Tendo em conta as propriedades da Multinormal, a Proposição 3.8 implica as seguintes conclusões:

- Cada observação individual Y_i tem distribuição Normal.
- Cada observação individual Y_i tem média dada pelo elemento i de $\mathbf{X}\vec{\boldsymbol{\beta}}$, ou seja, $E[Y_i] = \beta_0 + \beta_1 x_{1(i)} + \beta_2 x_{2(i)} + ... + \beta_p x_{p(i)}$.
- Cada observação individual tem variância igual (homogeneidade de variâncias): $V[Y_i] = \sigma^2$.
- Observações diferentes de Y são independentes, porque $Cov[Y_i, Y_j] = 0$ se $i \neq j$ e numa Multinormal isso implica a independência.

3.4 O estimador $\vec{\hat{\boldsymbol{\beta}}}$ dos parâmetros do modelo e a sua distribuição

Tal como na regressão linear simples, os estimadores dos parâmetros β_j $(j = 0, 1, 2, ..., p)$ obtêm-se adaptando a expressão matricial resultante de minimizar $SQRE$ no contexto descritivo (equação 3.6).

Definição 3.9: Estimador dos parâmetros populacionais

O vector $\vec{\hat{\boldsymbol{\beta}}}$ que estima o vector $\vec{\boldsymbol{\beta}}$ dos parâmetros populacionais é dado por:

$$\vec{\hat{\boldsymbol{\beta}}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\vec{\mathbf{Y}} \ ,$$

onde $\vec{\mathbf{Y}}$ e $\mathbf{X}$ são o vector e a matriz definidos nas Subsecções 3.2.2 e 3.2.3.

O vector $\vec{\hat{\boldsymbol{\beta}}}$ é de dimensão $p+1$. O seu primeiro elemento é o estimador de β_0, o seu segundo elemento é

o estimador de β_1, e por aí fora. Assinale-se o desfasamento dos índices, resultante da presença de β_0 na primeira posição: *o estimador de β_j está na posição $j+1$ do vector $\vec{\hat{\boldsymbol{\beta}}}$.*

3.4.0.1 A distribuição do vector de estimadores $\vec{\hat{\boldsymbol{\beta}}}$

Qualquer inferência sobre os parâmetros β_j exige o conhecimento da distribuição de probabilidades do vector dos respectivos estimadores. A Proposição 3.9 fornece o resultado necessário.

Proposição 3.9: Distribuição do estimador $\vec{\hat{\boldsymbol{\beta}}}$

Dado o Modelo de Regressão Linear Múltipla, tem-se:

$$\vec{\hat{\boldsymbol{\beta}}} \frown \mathcal{N}_{p+1}\left(\vec{\boldsymbol{\beta}}\,,\, \sigma^2\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\right) \ .$$

Demonstração 3.7: Proposição 3.9

O vector de estimadores $\vec{\hat{\boldsymbol{\beta}}}$ resulta do produto duma matriz não aleatória, $\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t$, e um vector aleatório, $\vec{\mathbf{Y}}$. Assim, tem a natureza dos produtos $\mathbf{B}\vec{\mathbf{W}}$ que foram estudados nas Proposições 3.5, 3.6 e 3.7, aquando do estudo das propriedades de vectores esperados, matrizes de (co-)variâncias e de vectores Multinormais, respectivamente. Pela alínea 7 da Proposição 3.7, tem-se que a distribuição do estimador $\vec{\hat{\boldsymbol{\beta}}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\vec{\mathbf{Y}}$ é Multinormal, com parâmetros que, aplicando as propriedades das Proposições 3.5, 3.6, 3.8 e as propriedades de produtos matriciais, são:

$$\begin{aligned}
E[\vec{\hat{\boldsymbol{\beta}}}] &= E[\underbrace{\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t}_{=\text{“}\mathbf{B}''}\ \underbrace{\vec{\mathbf{Y}}}_{=\text{“}\vec{\mathbf{W}}''}] = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\ \underbrace{E[\vec{\mathbf{Y}}]}_{=\mathbf{X}\vec{\boldsymbol{\beta}}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\left(\mathbf{X}^t\mathbf{X}\right)\vec{\boldsymbol{\beta}} = \mathbf{I}_{p+1}\,\vec{\boldsymbol{\beta}} = \vec{\boldsymbol{\beta}} \\
V[\vec{\hat{\boldsymbol{\beta}}}] &= V[\underbrace{\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t}_{=\text{“}\mathbf{B}''}\ \underbrace{\vec{\mathbf{Y}}}_{=\text{“}\vec{\mathbf{W}}''}] = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\ \underbrace{V[\vec{\mathbf{Y}}]}_{=\sigma^2\mathbf{I}_n}\ \left[\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\right]^t \\
&= \sigma^2\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\left[\mathbf{X}^t\right]^t\left[\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\right]^t = \sigma^2\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\left(\mathbf{X}^t\mathbf{X}\right)\left[\left(\mathbf{X}^t\mathbf{X}\right)^t\right]^{-1} \\
&= \sigma^2\,\mathbf{I}_{p+1}\left[\left(\mathbf{X}^t(\mathbf{X}^t)^t\right)\right]^{-1} = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1} \ .
\end{aligned}$$

3.4.0.2 Propriedades de estimadores individuais, $\hat{\beta}_j$

Tendo ainda em conta as propriedades da Multinormal (Proposição 3.7), os resultados da Proposição 3.9 implicam as seguintes conclusões, relativas aos estimadores $\hat{\beta}_j$ $(j = 0, 1, 2, ..., p)$:

- Cada estimador individual $\hat{\beta}_j$ tem distribuição Normal.
- Cada estimador tem média $E[\hat{\beta}_j] = \beta_j$, logo é um estimador centrado.

- Cada estimador tem variância $V[\hat{\beta}_j] = \sigma^2 \left(\mathbf{X}^t\mathbf{X}\right)^{-1}_{(j+1,j+1)}$ (note-se o desfasamento nos índices).
- Estimadores de diferentes parâmetros β_j *não* são (em geral) independentes, porque a matriz $(\mathbf{X}^t\mathbf{X})^{-1}$ não é, em geral, uma matriz diagonal. As covariâncias entre diferentes estimadores são dadas pelos elementos não diagonais de $\sigma^2 \left(\mathbf{X}^t\mathbf{X}\right)^{-1}$, e concretamente, $Cov[\hat{\beta}_i, \hat{\beta}_j] = \sigma^2 \left(\mathbf{X}^t\mathbf{X}\right)^{-1}_{(i+1,j+1)}$.

Assim, tem-se, para qualquer $j = 0, 1, ..., p$:

$$\begin{aligned} & \hat{\beta}_j \ \frown \ \mathcal{N}\left(\beta_j \ , \ \sigma^2 \left(\mathbf{X}^t\mathbf{X}\right)^{-1}_{(j+1,j+1)}\right) \\ \Leftrightarrow \quad & \frac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}} \ \frown \ \mathcal{N}(0,1) \ , \end{aligned} \tag{3.17}$$

onde $\sigma_{\hat{\beta}_j} = \sqrt{\sigma^2 \left(\mathbf{X}^t\mathbf{X}\right)^{-1}_{(j+1,j+1)}}$. Este resultado generaliza os relativos à Regressão Linear Simples.

3.4.0.3 O problema de σ^2 desconhecido

O resultado distribucional agora obtido (equação 3.17) permitiria construir intervalos de confiança ou fazer testes a hipóteses sobre os parâmetros $\vec{\boldsymbol{\beta}}$, não fosse a existência de um problema já familiar: o desconhecimento da variância σ^2 dos erros aleatórios.

Procedemos de forma análoga ao que se fez na Regressão Linear Simples:

- obtem-se um estimador para σ^2; e
- vê-se o efeito de substituir σ^2 pelo seu estimador, na distribuição acima indicada.

Proposição 3.10: Resultados distribucionais de SQRE

Dado o Modelo RLM, de Regressão Linear Múltipla, tem-se:

- $\frac{SQRE}{\sigma^2} \ \frown \ \chi^2_{n-(p+1)}$.
- $SQRE$ é independente de $\vec{\hat{\boldsymbol{\beta}}}$.

NOTA: Omite-se a demonstração.

NOTA: Os graus de liberdade associados a $SQRE$ são o *número de observações (n)* menos o *número de parâmetros do modelo (p+1)*.

Corolário 3.1

Dado o Modelo de RLM, $\quad E\left[\frac{SQRE}{n-(p+1)}\right] \ = \ \sigma^2$.

Demonstração 3.8: Corolário 3.1

Como o valor esperado duma variável aleatória com distribuição Qui-quadrado é igual ao parâmetro (graus de liberdade) dessa distribuição, e tendo em conta que σ^2 e $n-(p+1)$ são ambas constantes, tem-se:

$$E\left[\frac{SQRE}{\sigma^2}\right] = n-(p+1) \quad \Leftrightarrow \quad \frac{1}{\sigma^2}E\left[SQRE\right] = n-(p+1)$$
$$\Leftrightarrow \frac{1}{n-(p+1)}E\left[SQRE\right] = \sigma^2 \quad \Leftrightarrow \quad E\left[\frac{SQRE}{n-(p+1)}\right] = \sigma^2$$

O Quadrado Médio Residual na Regressão Múltipla.

Definição 3.10: Quadrado Médio Residual

Define-se o **Quadrado Médio Residual** ($QMRE$) numa Regressão Linear Múltipla como

$$QMRE \;=\; \frac{SQRE}{n-(p+1)}$$

- O QMRE é usado na Regressão Linear como estimador da variância dos erros aleatórios, isto é, toma-se $\hat{\sigma}^2 = QMRE$. A expressão agora definida é a expressão geral: na regressão linear simples há um único preditor, pelo que $p=1$ e tem-se a particularização da definição dada naquele contexto.
- Como se viu no acetato anterior, QMRE é um *estimador centrado* de σ^2.

Vejamos agora o efeito de substituir σ^2 pelo estimador $QMRE$, na distribuição indicada no início desta Subsecção.

Proposição 3.11: Distribuições para a inferência sobre β_j, $j = 0, 1, ..., p$

Dado o Modelo de Regressão Linear Múltipla, tem-se

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \;\frown\; t_{n-(p+1)} \;,$$

com $\hat{\sigma}_{\hat{\beta}_j} = \sqrt{QMRE \cdot (\mathbf{X}^t\mathbf{X})^{-1}_{(j+1,j+1)}}$.

Demonstração 3.9: Proposição 3.11

Vimos na Subsecção 3.4.0.2 que cada estimador $\hat{\beta}_j$ verifica:

$$Z = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 \cdot (\mathbf{X}^t\mathbf{X})^{-1}_{(j+1,j+1)}}} \frown \mathcal{N}(0,1) \ .$$

Temos ainda:

$$W = \frac{SQRE}{\sigma^2} \frown \chi^2_{n-(p+1)} \quad \text{e} \quad Z, W \text{ v.a. independentes} \ .$$

Logo (ver também a Subsecção 2.5.2.1):

$$\frac{Z}{\sqrt{W/(n-(p+1))}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{QMRE \cdot (\mathbf{X}^t\mathbf{X})^{-1}_{(j+1,j+1)}}} \frown t_{n-(p+1)} \ .$$

Esta Proposição dá-nos os resultados que servem de base à construção de intervalos de confiança e testes de hipóteses para os parâmetros β_j do modelo populacional.

NOTA: A quantidade fulcral da Proposição 3.11 tem uma estrutura totalmente análoga aos resultados correspondentes na RLS. Assim, *os intervalos de confiança e testes de hipóteses a parâmetros β_j individuais são, na Regressão Linear Múltipla, análogos aos da regressão linear simples*, dispensando a repetição de justificações que são idênticas às já descritas na regressão linear simples.

3.5 Intervalos de confiança para um β_j individual

Proposição 3.12: Intervalo de Confiança a $(1-\alpha)\times 100\%$ para β_j

Dado o Modelo de Regressão Linear Múltipla, um intervalo a $(1-\alpha)\times 100\%$ de confiança para o parâmetro β_j do modelo é:

$$\left] \ b_j - t_{\frac{\alpha}{2}\,[n-(p+1)]} \cdot \hat{\sigma}_{\hat{\beta}_j} \ , \ b_j + t_{\frac{\alpha}{2}\,[n-(p+1)]} \cdot \hat{\sigma}_{\hat{\beta}_j} \ \right[\ ,$$

com $\hat{\sigma}_{\hat{\beta}_j} = \sqrt{QMRE \cdot (\mathbf{X}^t\mathbf{X})^{-1}_{(j+1,j+1)}}$, e sendo $t_{\frac{\alpha}{2}[n-(p+1)]}$ o valor que na distribuição $t_{n-(p+1)}$ deixa *à direita* uma região de probabilidade $\frac{\alpha}{2}$. O valor b_j é o elemento $j+1$ do vector das estimativas $\vec{\mathbf{b}}$ (Proposição 3.1).

NOTA: A amplitude do intervalo de confiança aumenta com o valor de $QMRE$ e o valor diagonal da matriz $(\mathbf{X}^t\mathbf{X})^{-1}$ associado ao parâmetro β_j em questão.

3.5.1 Intervalos de confiança para β_j no R

A informação necessária para a construção de intervalos de confiança para cada parâmetro β_j obtém-se, no R, a partir das tabelas produzidas pela função `summary`. No exemplo de regressão linear múltipla com

os dados dos lírios (a *data frame* `iris`), já considerado na Subsecção 3.2.7, tem-se:

```
> summary(iris2.lm)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.24031    0.17837  -1.347     0.18
Petal.Length  0.52408    0.02449  21.399  < 2e-16 ***
Sepal.Length -0.20727    0.04751  -4.363 2.41e-05 ***
Sepal.Width   0.22283    0.04894   4.553 1.10e-05 ***
```

As estimativas b_j encontram-se na coluna de nome *Estimate*, enquanto que os erros padrões $\hat{\sigma}_{\hat{\beta}_j}$ encontram-se na coluna de nome *Std. Error*. Assim, estima-se que em média a largura da pétala diminui $0.20727cm$ por cada aumento de $1\,cm$ no comprimento da sépala (mantendo-se as outras medições constantes). O erro padrão associado a esta estimativa é 0.04751.

Como o quantil de ordem 0.975 numa distribuição *t-Student* com $n-(p+1)=150-4=146$ graus de liberdade é dado por $t_{0.025(146)}=1.976346$, o intervalo a 95% de confiança para β_2 é:

$$]\,(-0.20727)-(1.976346)(0.04751)\;,\;(-0.20727)+(1.976346)(0.04751)\,[\quad=\quad]\,-0.3012\;,\;-0.1134\,[$$

Alternativamente, é possível usar a função `confint` no objecto `iris2.lm`, resultante de ajustar a regressão, para obter os intervalos de confiança para cada β_j individual:

```
> confint(iris2.lm)
                 2.5 %     97.5 %
(Intercept)  -0.5928277  0.1122129
Petal.Length  0.4756798  0.5724865
Sepal.Length -0.3011547 -0.1133775
Sepal.Width   0.1261101  0.3195470
```

Como na regressão linear simples, é possível controlar o grau de confiança associado ao intervalo através do argumento `level` (cujos valores devem corresponder a uma proporção, entre 0 e 1). Eis os intervalos a 99% de confiança para os vários β_j:

```
> confint(iris2.lm,level=0.99)
                  0.5 %      99.5 %
(Intercept)  -0.70583864  0.22522386
Petal.Length  0.46016260  0.58800363
Sepal.Length -0.33125352 -0.08327863
Sepal.Width   0.09510404  0.35055304
```

3.6 Testes de Hipóteses sobre os parâmetros individuais β_j

O mesmo resultado usado para construir intervalos de confiança serve para construir testes a hipóteses para cada β_j individual, dado o Modelo de Regressão Linear Múltipla.

3.6.1 Testes bilaterais para β_j

Tal como no caso das regressões lineares simples, a natureza das hipóteses determina o tipo de Região Crítica associado. Comecemos por ver como lidar com testes a que correspondem regiões críticas bilaterais.

Hipóteses: $H_0: \quad \beta_j = c \quad$ vs. $\quad H_1: \quad \beta_j \neq c$

Estatística do Teste: $T = \frac{\hat{\beta}_j - \overbrace{\beta_j|_{H_0}}^{=c}}{\hat{\sigma}_{\hat{\beta}_j}} \frown t_{n-(p+1)}$

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Rejeitar H_0 se $|T_{calc}| > t_{\frac{\alpha}{2}[n-(p+1)]}$.

3.6.2 Testes unilaterais esquerdos para β_j

No caso da Hipótese Alternativa ser do tipo $\beta_j < c$, a região crítica deverá ser unilateral esquerda.

Hipóteses: $H_0: \quad \beta_j \geq c \quad$ vs. $\quad H_1: \quad \beta_j < c$

Estatística do Teste: $T = \frac{\hat{\beta}_j - \overbrace{\beta_j|_{H_0}}^{=c}}{\hat{\sigma}_{\hat{\beta}_j}} \frown t_{n-(p+1)}$

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Rejeitar H_0 se $T_{calc} < -t_{\alpha[n-(p+1)]}$.

3.6.3 Testes unilaterais direitos para β_j

Finalmente, no caso da Hipótese Alternativa ser do tipo $\beta_j > c$, a região crítica será unilateral direita.

Hipóteses: $H_0: \quad \beta_j \leq c \quad$ vs. $\quad H_1: \quad \beta_j > c$

Estatística do Teste: $T = \frac{\hat{\beta}_j - \overbrace{\beta_j|_{H_0}}^{=c}}{\hat{\sigma}_{\hat{\beta}_j}} \frown t_{n-(p+1)}$

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Rejeitar H_0 se $T_{calc} > t_{\alpha[n-(p+1)]}$.

Note-se que, ao estimar σ^2 com base no Quadrado Médio Residual, a matriz de (co-)variâncias do vector de estimadores $\vec{\hat{\boldsymbol{\beta}}}$, $V[\vec{\hat{\boldsymbol{\beta}}}] = \sigma^2 (\mathbf{X}^t\mathbf{X})^{-1}$ passa a ser estimada pela matriz indicada na Nota 3.3.

Nota 3.3: Matriz estimada de (co-)variâncias de $\vec{\hat{\boldsymbol{\beta}}}$

No estudo do modelo de regressão linear múltipla, a matriz de (co-)variâncias do vector de estimadores $\vec{\hat{\boldsymbol{\beta}}}$ é estimada pela matriz:

$$\widehat{V[\vec{\hat{\boldsymbol{\beta}}}]} \;=\; QMRE \cdot (\mathbf{X}^t\mathbf{X})^{-1} \;. \tag{3.18}$$

3.7 Inferência sobre combinações lineares dos parâmetros

Seja $\vec{\mathbf{a}} = (a_0, a_1, ..., a_p)^t$ um vector de constantes (não aleatório) em $\mathbb{R}^{p+1}$. O produto interno $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$ define uma *combinação linear dos parâmetros* do modelo:

$$\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} \;=\; \left[\begin{array}{ccccc} a_0 & a_1 & a_2 & ... & a_p \end{array}\right] \left[\begin{array}{c} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{array}\right] \;=\; a_0\beta_0 + a_1\beta_1 + a_2\beta_2 + ... + a_p\beta_p \;. \tag{3.19}$$

Eis alguns *casos particulares importantes* de combinações lineares dos parâmetros β_j (que não esgotam todas as possibilidades de combinações lineares):

um β_j individual: se $\vec{\mathbf{a}} = (0\ 0\ ...\ \underbrace{1}_{=a_j}\ ...\ 0)^t$ é um vector com um único elemento não-nulo, de valor 1, na posição $j+1$ (correspondente a a_j), a combinação linear (3.19) reduz-se a um único parâmetro, mais concretamente β_j: $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} = \beta_j$. Este caso particular já foi considerado antes e não introduz novidades.

a soma ou diferença de dois parâmetros: seja $\vec{\mathbf{a}} = (0\ 0\ ...\ 0\ \underbrace{1}_{=a_i}\ 0\ ...\ \underbrace{\pm 1}_{=a_j}\ ...\ 0)^t$ um vector com apenas dois elementos não-nulos, 1 na posição $i+1$ e ± 1 na posição $j+1$. Nesse caso, a combinação linear será a *soma* ou a *diferença* de dois parâmetros: $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} = \beta_i \pm \beta_j$.

o valor esperado de Y, para valores dados das variáveis preditoras: caso $\vec{\mathbf{a}} = (1, x_1, x_2, ..., x_p)$, sendo x_j um possível valor da variável preditora X_j, então $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$ representa o *valor esperado de Y associado aos valores indicados das variáveis preditoras*:

$$\begin{aligned} \vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \\ &= E[Y \mid X_1 = x_1, X_2 = x_2, ..., X_p = x_p] \;=\; \mu_{Y|\vec{\mathbf{x}}}. \end{aligned}$$

Na notação mais sintética $\mu_{Y|\vec{\mathbf{x}}}$, $\vec{\mathbf{x}} = (x_1, x_2, ..., x_p)$ é o vector dos valores das variáveis preditoras.

Obtendo-se resultados inferenciais para qualquer combinação linear $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$ dos parâmetros obtêm-se automaticamente resultados inferenciais para estes (e outros) casos particulares.

Para estimar $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} = a_0\beta_0 + a_1\beta_1 + a_2\beta_2 + ... + a_p\beta_p$, usa-se o estimador:

$$\vec{\mathbf{a}}^t\vec{\hat{\boldsymbol{\beta}}} \;=\; a_0\hat{\beta}_0 + a_1\hat{\beta}_1 + a_2\hat{\beta}_2 + ... + a_p\hat{\beta}_p \;. \tag{3.20}$$

Resultados inferenciais sobre as combinações lineares $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$ dos parâmetros exigem o conhecimento duma distribuição de probabilidades associada a este estimador $\vec{\mathbf{a}}^t\vec{\hat{\boldsymbol{\beta}}}$.

3.7.1 Quantidade fulcral para a inferência sobre $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$

Na Proposição 3.13 tem-se o resultado que serve de base à construção de *intervalos de confiança* e *testes de hipóteses* para quaisquer combinações lineares dos parâmetros β_j do modelo.

Proposição 3.13: Distribuições para combinações lineares dos βs

Dado o Modelo de Regressão Linear Múltipla, tem-se

$$\frac{\vec{\mathbf{a}}^t\vec{\hat{\boldsymbol{\beta}}} - \vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}}{\hat{\sigma}_{\vec{\mathbf{a}}^t\vec{\hat{\boldsymbol{\beta}}}}} \;\frown\; t_{n-(p+1)} \;,$$

com $\hat{\sigma}_{\vec{\mathbf{a}}^t\vec{\hat{\boldsymbol{\beta}}}} = \sqrt{QMRE \cdot \vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}}$.

Demonstração 3.10: Proposição 3.13

A multinormalidade de $\vec{\hat{\boldsymbol{\beta}}}$ (Proposição 3.9) implica a normalidade de qualquer vector que seja uma combinação linear das suas componentes (lembrar as propriedades de vectores Multinormais, vistas na Proposição 3.7). Assim, e tendo em conta as restantes propriedades dessa mesma Proposição, tem-se:

- $\vec{\hat{\boldsymbol{\beta}}} \frown \mathcal{N}_{p+1}\left(\vec{\boldsymbol{\beta}}, \sigma^2\,(\mathbf{X}^t\mathbf{X})^{-1}\right)$;
- Combinações lineares de vectores Multinormais têm distribuição Normal:
  $\vec{\mathbf{a}}^t\vec{\hat{\boldsymbol{\beta}}} \frown \mathcal{N}(\,\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}\,, \sigma^2\,\vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}\,)$;
- Subtraindo a média e dividindo pelo desvio padrão de $\vec{\mathbf{a}}^t\vec{\hat{\boldsymbol{\beta}}}$ obtém-se uma Normal reduzida:
  $\mathbf{Z} \;=\; \frac{\vec{\mathbf{a}}^t\vec{\hat{\boldsymbol{\beta}}} - \vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}}{\sqrt{\sigma^2\,\vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}}} \;\frown\; \mathcal{N}(0,1)$;
- Por um raciocínio análogo ao usado aquando dos βs individuais, tem-se então

$$\frac{\vec{\mathbf{a}}^t\vec{\hat{\boldsymbol{\beta}}} - \vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}}{\sqrt{QMRE \cdot \vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}}} \;\frown\; t_{n-(p+1)} \;.$$

NOTA: Repare-se na analogia da estrutura desta quantidade fulcral com os resultados anteriores, relativos a β_js individuais (Proposição 3.11): a quantidade fulcral é sempre a razão entre a diferença do estimador e a quantidade que se pretende estimar, a dividir pelo erro padrão associado à estimação. Esta analogia significa que a forma de construção de intervalos de confiança ou testes de hipóteses às combinações lineares $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$ dos parâmetros do modelo é análoga à que foi usada na construção do mesmo tipo de ferramentas inferenciais, aquando do estudo dos parâmetros β_j individuais.

3.7.2 Intervalos de confiança para $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$

Comecemos por ver intervalos de confiança para uma combinação linear genérica, $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} = a_0\beta_0 + a_1\beta_1 + a_2\beta_2 + ... + a_p\beta_p$.

Proposição 3.14: Intervalo de Confiança a $(1-\alpha)\times 100\%$ para $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$

Dado o Modelo de Regressão Linear Múltipla, um intervalo a $(1-\alpha)\times 100\%$ de confiança para a combinação linear dos parâmetros, $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} = a_0\beta_0 + a_1\beta_1 + ... + a_p\beta_p$, é:

$$\left]\ \vec{\mathbf{a}}^t\vec{\mathbf{b}} - t_{\frac{\alpha}{2}\,[n-(p+1)]}\cdot\hat{\sigma}_{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}}\ ,\ \vec{\mathbf{a}}^t\vec{\mathbf{b}} + t_{\frac{\alpha}{2}\,[n-(p+1)]}\cdot\hat{\sigma}_{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}}\ \right[\ ,$$

com $\vec{\mathbf{a}}^t\vec{\mathbf{b}} = a_0b_0 + a_1b_1 + ... + a_pb_p$ e $\hat{\sigma}_{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}} = \sqrt{QMRE\cdot\vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}}$.

Demonstração 3.11: Proposição 3.14

Vamos construir o intervalo de confiança a $(1-\alpha)\times 100\%$ para $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$, a partir da distribuição indicada no enunciado. Sendo $t_{\frac{\alpha}{2}}$ o valor que, numa distribuição $t_{n-(p+1)}$, deixa à direita uma região de probabilidade $\frac{\alpha}{2}$, temos a seguinte afirmação probabilística, na qual trabalhamos a dupla desigualdade até deixar a combinação linear (para a qual se quer o intervalo de confiança) isolada no meio:

$$\begin{aligned}
P\left[-t_{\frac{\alpha}{2}} < \frac{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}} - \vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}}{\hat{\sigma}_{a^t\hat{\vec{\boldsymbol{\beta}}}}} < t_{\frac{\alpha}{2}}\right] &= 1-\alpha \\
P\left[-t_{\frac{\alpha}{2}}\cdot\hat{\sigma}_{a^t\hat{\vec{\boldsymbol{\beta}}}} < \vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}} - \vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} < t_{\frac{\alpha}{2}}\cdot\hat{\sigma}_{a^t\hat{\vec{\boldsymbol{\beta}}}}\right] &= 1-\alpha \\
P\left[t_{\frac{\alpha}{2}}\cdot\hat{\sigma}_{a^t\hat{\vec{\boldsymbol{\beta}}}} > \vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} - \vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}} > -t_{\frac{\alpha}{2}}\cdot\hat{\sigma}_{a^t\hat{\vec{\boldsymbol{\beta}}}}\right] &= 1-\alpha \qquad \text{(multiplicando por } -1) \\
P\left[\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}} - t_{\frac{\alpha}{2}}\cdot\hat{\sigma}_{a^t\hat{\vec{\boldsymbol{\beta}}}} < \vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} < \vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}} + t_{\frac{\alpha}{2}}\cdot\hat{\sigma}_{a^t\hat{\vec{\boldsymbol{\beta}}}}\right] &= 1-\alpha
\end{aligned}$$

Assim, calculando o valor $\vec{\mathbf{a}}^t\vec{\mathbf{b}} = a_0b_0 + a_1b_1 + ... + a_pb_p$ do estimador $\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}$ e o erro padrão $\hat{\sigma}_{a^t\hat{\vec{\boldsymbol{\beta}}}}$, para

a nossa amostra, temos o intervalo a $(1-\alpha)\times 100\%$ de confiança para $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} = a_0\beta_0 + a_1\beta_1 + ... + a_p\beta_p$:

$$\left] \ \vec{\mathbf{a}}^t\vec{\mathbf{b}} - t_{\frac{\alpha}{2}[n-(p+1)]} \cdot \hat{\sigma}_{a^t\hat{\vec{\boldsymbol{\beta}}}} \quad , \quad \vec{\mathbf{a}}^t\vec{\mathbf{b}} + t_{\frac{\alpha}{2}[n-(p+1)]} \cdot \hat{\sigma}_{a^t\hat{\vec{\boldsymbol{\beta}}}} \ \right[$$

3.7.3 Testes de Hipóteses sobre os parâmetros

Dado o Modelo de Regressão Linear Múltipla, tem-se:

3.7.3.1 Testes de Hipóteses bilateral para $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$

Hipóteses: $H_0: \ \vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} = c$ vs. $H_1: \ \vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} \neq c$.

Estatística do Teste: $T = \frac{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}} - \overbrace{\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}_{|H_0}}^{=c}}{\hat{\sigma}_{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}}} \frown t_{n-(p+1)}$, se H_0 verdade.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Rejeitar H_0 se $|T_{calc}| > t_{\frac{\alpha}{2}[n-(p+1)]}$.

3.7.3.2 Testes de Hipóteses unilateral esquerdo para $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$

Hipóteses: $H_0: \ \vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} \geq c$ vs. $H_1: \ \vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} < c$.

Estatística do Teste: $T = \frac{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}} - \overbrace{\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}_{|H_0}}^{=c}}{\hat{\sigma}_{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}}} \frown t_{n-(p+1)}$, se H_0 verdade.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Rejeitar H_0 se $T_{calc} < -t_{\alpha[n-(p+1)]}$.

3.7.3.3 Testes de Hipóteses unilateral direito para $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$

Hipóteses: $H_0: \ \vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} \leq c$ vs. $H_1: \ \vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} > c$.

Estatística do Teste: $T = \frac{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}} - \overbrace{\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}_{|H_0}}^{=c}}{\hat{\sigma}_{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}}} \frown t_{n-(p+1)}$, se H_0 verdade.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Rejeitar H_0 se $T_{calc} > t_{\alpha[n-(p+1)]}$.

3.7.4 Comentários sobre casos particulares

Viram-se três casos particulares importantes de combinações lineares dos parâmetros. Eis algumas considerações sobre a aplicação nesses casos particulares dos resultados inferenciais gerais para combinações lineares dos parâmetros:

- No caso de β_j individuais ($\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} = \beta_j$), os intervalos e testes acabados de enunciar são idênticos aos dados nas Subsecções 3.5 e 3.6.
- No caso de somas ou diferenças de βs individuais ($\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} = \beta_j \pm \beta_j$), o erro padrão pode calcular-se através duma fórmula alternativa, que tem a sua origem nas expressões para a variância da soma ou diferença de duas variáveis aleatórias: $V[X \pm Y] = V[X] + V[Y] \pm 2Cov[X, Y]$. De facto, considerando o erro padrão $\hat{\sigma}_{\vec{\mathbf{a}}^t\vec{\hat{\boldsymbol{\beta}}}} = \hat{\sigma}_{\hat{\beta}_i \pm \hat{\beta}_j}$, vem:

$$\begin{aligned} \hat{\sigma}_{\hat{\beta}_i \pm \hat{\beta}_j} &= \sqrt{\widehat{V[\hat{\beta}_i \pm \hat{\beta}_j]}} &&= \sqrt{\widehat{V[\hat{\beta}_i]} + \widehat{V[\hat{\beta}_j]} \pm 2 \cdot \widehat{Cov[\hat{\beta}_i, \hat{\beta}_j]}} \\ &&&= \sqrt{QMRE \cdot \left[(\mathbf{X}^t\mathbf{X})^{-1}_{(i+1,i+1)} + (\mathbf{X}^t\mathbf{X})^{-1}_{(j+1,j+1)} \pm 2(\mathbf{X}^t\mathbf{X})^{-1}_{(i+1,j+1)} \right]} \end{aligned}$$

  Os valores necessários para calcular as três parcelas debaixo da raíz quadrada encontram-se na matriz de (co-)variâncias estimada de $\vec{\hat{\boldsymbol{\beta}}}$, explicitada na Nota 3.3.
- No caso de $\vec{\mathbf{a}}$ conter os valores das variáveis preditoras *usados na i-ésima observação* com que o modelo foi ajustado, o vector $\vec{\mathbf{a}}^t$ será a linha i da matrix $\mathbf{X}$ e o vector $\vec{\mathbf{a}}$ é a coluna i da matriz $\mathbf{X}^t$. Nesse caso, $\vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}$ será o elemento da linha i, coluna i, da matriz de projecções ortogonal $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$, que designaremos h_{ii}. Uma inferência sobre $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$ envolvendo um tal vector $\vec{\mathbf{a}}$ terá assim erro padrão associado dado por:

$$\hat{\sigma}_{\vec{\mathbf{a}}^t\vec{\hat{\boldsymbol{\beta}}}} = \sqrt{QMRE \cdot \vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}} = \sqrt{QMRE \cdot h_{ii}} \; .$$

3.7.5 Exemplo de intervalo de confiança para a soma de dois parâmetros

Considere-se de novo o exemplo dos lírios, já analisado nas Subsecções 3.2.7 e 3.5.1. Vamos exemplificar a construção dum intervalo de confiança para a soma $\beta_2 + \beta_3$. Particularizando o resultado geral obtido na Proposição 3.14 para qualquer combinação linear dos parâmetros β_j, um intervalo de confiança para $\beta_2 + \beta_3$ é da forma:

$$\left] \; (b_2 + b_3) \; - \; t_{\frac{\alpha}{2}\,[n-(p+1)]} \cdot \hat{\sigma}_{\hat{\beta}_2 + \hat{\beta}_3} \quad , \quad (b_2 + b_3) \; + \; t_{\frac{\alpha}{2}\,[n-(p+1)]} \cdot \hat{\sigma}_{\hat{\beta}_2 + \hat{\beta}_3} \; \right[\tag{3.21}$$

Na listagem da Subsecção 3.5.1 viu-se que $b_2 = -0.20727$ e $b_3 = 0.22283$. Logo $b_2 + b_3 = -0.20727 + 0.22283 = 0.01556$. Este será o ponto central do intervalo de confiança. Na mesma Subsecção viu-se ainda que $t_{0.025(146)} = 1.976346$. Falta determinar o erro padrão estimado de $\hat{\beta}_2 + \hat{\beta}_3$, ou seja, $\hat{\sigma}_{\hat{\beta}_2 + \hat{\beta}_3}$. Para tal, e como se viu acima, será necessária a matriz das (co)variâncias estimadas dos estimadores $\vec{\hat{\boldsymbol{\beta}}}$, $\widehat{V[\vec{\hat{\boldsymbol{\beta}}}]} = QMRE \cdot (\mathbf{X}^t\mathbf{X})^{-1}$ (equação 3.18). No `R`, esta matriz é calculada pela função `vcov`, aplicada a uma regressão já ajustada. Assim,

```
> vcov(iris2.lm)
              (Intercept)  Petal.Length Sepal.Length  Sepal.Width
(Intercept)   0.031815766  0.0015144174 -0.005075942 -0.002486105
Petal.Length  0.001514417  0.0005998259 -0.001065046  0.000802941
Sepal.Length -0.005075942 -0.0010650465  0.002256837 -0.001344002
Sepal.Width  -0.002486105  0.0008029410 -0.001344002  0.002394932
```

O erro padrão estimado de $\hat{\beta}_2 + \hat{\beta}_3$ pode ser calculado da forma discutida na Subsecção anterior:

$$\begin{aligned} \hat{\sigma}_{\hat{\beta}_2+\hat{\beta}_3} &= \sqrt{\widehat{V[\hat{\beta}_2]} + \widehat{V[\hat{\beta}_3]} + 2\,\widehat{Cov[\hat{\beta}_2, \hat{\beta}_3]}} \\ &= \sqrt{0.002256837 + 0.002394932 + 2(-0.001344002)} = 0.04431439 \,. \end{aligned}$$

Substituindo na expressão (3.21), tem-se o seguinte intervalo a 95% de confiança para $\beta_2 + \beta_3$:

$$\left] \, -0.07202057 \,, 0.1031406 \, \right[\,.$$

Entre os valores admissíveis para $\beta_2 + \beta_3$ encontra-se o valor zero, pelo que é admissível que $\beta_2 = -\beta_3$.

Na próxima Secção será considerada a inferência relativa a valores da variável resposta.

3.8 Inferência sobre valores de Y

3.8.1 Intervalos de confiança para $\mu_{Y|\vec{\mathbf{x}}}$

Já se viu na Secção 3.7 que se $\vec{\mathbf{a}}$ é um vector cujo primeiro elemento seja 1 e os restantes sejam valores das p variáveis preditoras, a combinação linear $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$ será a o valor esperado de Y, dado esse conjunto de valores das variáveis preditoras: $\mu_{Y|\vec{\mathbf{x}}} = E[\,Y\,|X_1 = x_1, X_2 = x_2, ..., X_p = x_p]$. Nesse caso, o intervalo de confiança da Subsecção 3.7.2 particulariza-se da seguinte forma:

$$\left] \quad \hat{\mu}_{Y|\vec{\mathbf{x}}} - t_{\frac{\alpha}{2}\,[n-(p+1)]} \cdot \hat{\sigma}_{\hat{\mu}_{Y|\vec{\mathbf{x}}}} \quad , \quad \hat{\mu}_{Y|\vec{\mathbf{x}}} + t_{\frac{\alpha}{2}\,[n-(p+1)]} \cdot \hat{\sigma}_{\hat{\mu}_{Y|\vec{\mathbf{x}}}} \quad \right[\tag{3.22}$$

sendo $\vec{\mathbf{x}} = (x_1, x_2, ..., x_p)$ o vector dos valores dos preditores, e

- $\hat{\mu}_{Y|\vec{\mathbf{x}}} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_p x_p$; e
- $\hat{\sigma}_{\hat{\mu}_{Y|\vec{\mathbf{x}}}} = \sqrt{QMRE \cdot \vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}}$, com $\vec{\mathbf{a}} = (1, x_1, x_2, ..., x_p)$.

3.8.2 Intervalos de predição para observações individuais de Y

Pode também obter-se, de forma análoga ao que foi visto na regressão linear simples, um intervalo de predição para uma observação individual de Y, associada aos valores $X_1 = x_1, X_2 = x_2, ..., X_p = x_p$ das variáveis preditoras.

Tal como se fez na regressão linear simples, a medida de variabilidade associada a estes intervalos é aumentada em $QMRE$ unidades. Concretamente, toma-se como estimativa da variância associada a uma

observação individual de Y a soma:

$$\hat{\sigma}^2_{indiv} = \hat{\sigma^2}_{\hat{\mu}_{Y|\vec{\mathbf{x}}}} + \hat{\sigma}^2 = QMRE \cdot \vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}} + QMRE = QMRE \cdot \left[\vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}} + 1\right] \ . \tag{3.23}$$

Assim, o intervalo de predição ($(1-\alpha)\times 100\%$) para uma observação individual de Y é:

$$\left] \ \hat{\mu}_{Y|\vec{\mathbf{x}}} - t_{\frac{\alpha}{2}\,[n-(p+1)]} \cdot \hat{\sigma}_{indiv} \quad , \quad \hat{\mu}_{Y|\vec{\mathbf{x}}} + t_{\frac{\alpha}{2}\,[n-(p+1)]} \cdot \hat{\sigma}_{indiv} \ \right[\tag{3.24}$$

onde $\vec{\mathbf{x}} = (x_1, x_2, ..., x_p)^t$ indica o vector dos valores dos preditores e

- $\hat{\mu}_{Y|\vec{\mathbf{x}}} = b_0 + b_1x_1 + b_2x_2 + ... + b_px_p$; e
- $\hat{\sigma}_{indiv} \ = \ \sqrt{QMRE\ [1 + \vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}]}$ com $\vec{\mathbf{a}} = (1, x_1, x_2, ..., x_p)$.

3.8.3 Inferência sobre valores de Y no R

É possível obter o intervalo de confiança referido na Subsecção 3.8.1 através do comando `predict`, tal como na regressão linear simples. No exemplo dos lírios, um intervalo a 95% de confiança para a largura esperada de pétalas de flores com `Petal.Length=2`, `Sepal.Length=5` e `Sepal.Width=3.1` corresponde a usar o vector $\vec{\mathbf{a}} = (1,\ 2,\ 5,\ 3.1)^t$. No R, esse intervalo de confiança é obtido da seguinte forma:

```
> predict(iris2.lm, new=data.frame(Petal.Length=2, Sepal.Length=5,  Sepal.Width=3.1),
+             int="conf")
          fit       lwr       upr
[1,] 0.462297 0.4169203 0.5076736
```

O intervalo a 95% de confiança para $E[Y|X_1{=}2,\ X_2{=}5,\ X_3{=}3.1]$ é assim] 0.4169 , 0.5077 [.

É possível obter um intervalo de predição no R, através do comando `predict` usando o argumento `int="pred"`, tal como na regressão linear simples. Eis, na regressão linear múltipla que tem estado a ser considerada com os dados dos lírios, o intervalo de predição para a largura da pétala, num lírio cujo comprimento de pétala seja 2 e com sépala de comprimento 5 e largura 3.1:

```
> predict(iris2.lm, new=data.frame(Petal.Length=2, Sepal.Length=5,  Sepal.Width=3.1),
+                       int="pred")
          fit       lwr       upr
[1,] 0.462297 0.08019972 0.8443942
```

O intervalo de predição pedido é] 0.0802 , 0.8444 [. Trata-se dum intervalo de enorme amplitude, provavelmente de reduzido interesse práctico.

3.9 Avaliando a qualidade do ajustamento: o teste F global

Numa Regressão Linear Simples, se $\beta_1{=}0$, a equação do modelo reduz-se a $Y{=}\beta_0{+}\epsilon$, o chamado Modelo Nulo. Neste caso, o conhecimento do preditor X em nada contribui para o conhecimento de Y (o Modelo Nulo não tira partido da informação do preditor).

Numa Regressão Linear Múltipla o Modelo Nulo, de equação $Y_i = \beta_0 + \epsilon_i$, corresponde a admitir que *todas* as variáveis preditoras têm coeficiente β_j nulo. As hipóteses que queremos testar são então:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$
(Inexistência de relação linear)
vs.
$$H_1 : \exists j = 1, ..., p \quad \text{tal que} \quad \beta_j \neq 0$$
(Existência de relação linear)

Nota 3.4

1. Repare-se que β_0 *não intervém nas hipóteses.*
2. A Hipótese Nula exige que *em simultâneo* os parâmetros β_j $(j > 0)$ que multiplicam variáveis preditores sejam todos nulos. Esta hipótese **não** pode ser estudada através de p testes *t-Student* a que cada β_j seja nulo, uma vez que em cada um desses testes admite-se que os restantes parâmetros são livres de tomar qualquer valor.

3.9.1 O Teste F de ajustamento global do Modelo

Vamos obter uma estatística de teste para optar entre as hipóteses acima referidas. Tal como na discussão no contexto da regressão linear simples, é natural supôr que a Soma de Quadrados associada à Regressão, SQR, seja útil, uma vez que se trata do numerador do Coeficiente de Determinação $R^2 = \frac{SQR}{SQT}$, um indicador que, sendo válido o Modelo Nulo, deveria tomar valores próximos de zero.

3.9.1.1 Distribuição associada a SQR

A Soma de Quadrados associada à Regressão define-se como $SQR = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$. Tem-se o seguinte resultado (cuja demonstração se omite por exceder o âmbito da disciplina):

Proposição 3.15: Distribuição associada a SQR

Dado o Modelo de Regressão Linear Múltipla,

- $\frac{SQR}{\sigma^2} \frown \chi^2_p$, se $\beta_1 = \beta_2 = ... = \beta_p = 0$.
- SQR e $SQRE$ são variáveis aleatórias independentes.

Este resultado vai permitir obter a estatística do teste para optar entre as Hipóteses formuladas no início da Secção 3.9.

3.9.1.2 A estatística do teste F

Comecemos por definir o Quadrado Médio associado à Regressão, no contexto duma regressão linear múltipla.

Definição 3.11: Quadrado Médio da Regressão

Define-se o **Quadrado Médio associado à Regressão** como sendo:

$$QMR = \frac{SQR}{p} \ .$$

Podemos agora definir a estatística do teste F de ajustamento global.

Proposição 3.16: A estatística do teste F de ajustamento global

Seja dado o Modelo de Regressão Linear Múltipla com p variáveis preditoras e ajustada com n observações. Se $\beta_1 = \beta_2 = ... = \beta_p = 0$, tem-se:

$$F \ = \ \frac{QMR}{QMRE} \ = \ \frac{n-(p+1)}{p} \cdot \frac{R^2}{1-R^2} \ \frown \ F_{p,n-(p+1)} \ .$$

Demonstração 3.12: Proposição 3.16

Temos (veja também a Subsecção 2.9.1), que se H_0 fôr verdade, ou seja, se $\beta_j = 0 \, , \forall i = 1 : p$, então:

$$\left.\begin{array}{l} W \ = \ \frac{SQR}{\sigma^2} \frown \chi^2_p \\ V \ = \ \frac{SQRE}{\sigma^2} \frown \chi^2_{n-(p+1)} \\ W, V \text{ independentes} \end{array}\right\} \ \Rightarrow \ \frac{W/p}{V/_{n-(p+1)}} = \frac{QMR}{QMRE} \frown F_{p,n-(p+1)} \ ,$$

sendo $QMR = \frac{SQR}{p}$ e $QMRE = \frac{SQRE}{n-(p+1)}$.

A segunda expressão para a estatística F obtém-se a partir das definições de cada Quadrado Médio, seguindo passos análogos aos usados na demonstração da proposição 2.19.

Sendo válido o Modelo RLM, pode efectuar-se o seguinte *Teste F de ajustamento global do modelo RLM.*

3.9.1.3 Os passos do teste F de ajustamento global

A segunda expressão da estatística F (Proposição 3.16) torna evidente que, também numa Regressão Linear Múltipla, a estatística F é uma função crescente do Coeficiente de Determinação amostral, R^2. Assim, quanto maior o valor de R^2, maior o valor de F_{calc}, e menos plausível se torna a Hipótese Nula

(ou seja, menos plausível é a inexistência duma relação linear entre a variável resposta e os preditores). Desta forma, também aqui é adequada uma Região Crítica unilateral direita. Eis os passos do teste:

Hipóteses: $H_0 \ : \ \beta_1 = \beta_2 = ... = \beta_p \ = \ 0$

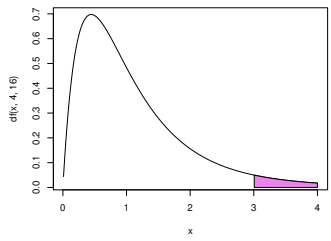
vs.

$H_1 \ : \ \exists \ j = 1, ..., p \quad \text{tal que} \quad \beta_j \ \neq \ 0.$

Estatística do Teste: $F \ = \ \frac{QMR}{QMRE} \ \frown \ F_{p,n-(p+1)} \quad \text{se } H_0.$

Nível de significância do teste: α

Região Crítica (Região de Rejeição): (Unilateral direita) Rejeitar H_0 se $F_{calc} \ > \ f_{\alpha[p,n-(p+1)]}$.



3.9.1.4 Formulação alternativa do Teste F de ajustamento global

As hipóteses do teste podem escrever-se usando o Coeficiente de Determinação populacional, $\mathcal{R}^2$:

$$H_0 : \mathcal{R}^2 = 0 \qquad \text{vs.} \qquad H_1 : \mathcal{R}^2 > 0 \ .$$

A hipótese $H_0 : \mathcal{R}^2 = 0$ indica ausência de relação linear entre Y e o conjunto dos preditores. Corresponde ao Modelo Nulo, sem qualquer papel para os preditores. A sua rejeição não garante um bom ajustamento do modelo. Apenas permite dizer que o nosso modelo difere significativamente desse Modelo Nulo.

O teste de ajustamento global pode também ser escrito usando estas formulações alternativas das hipóteses e/ou da estatística do teste:

Hipóteses: $H_0 \ : \ \mathcal{R}^2 \ = \ 0 \quad \text{vs.} \quad H_1 \ : \ \mathcal{R}^2 \ > \ 0.$

Estatística do Teste: $F \ = \ \frac{n-(p+1)}{p} \cdot \frac{R^2}{1-R^2} \ \frown \ F_{[p,n-(p+1)]} \quad \text{se } H_0.$

Nível de significância do teste: α

Região Crítica (Região de Rejeição): (Unilateral direita) Rejeitar H_0 se $F_{calc} \ > \ f_{\alpha(p,n-(p+1))}$

- A hipótese nula $H_0 \ : \ \mathcal{R}^2 \ = \ 0$ afirma que, na população, o coeficiente de determinação é nulo.
- Com esta formulação, torna-se evidente a opção por uma região crítica unilateral direita: quanto maior o valor amostral de R^2, mais improvável se torna a hipótese nula $\mathcal{R}^2 = 0$.

3.9.1.5 A tabela de síntese do ajustamento global

Pode sintetizar-se a informação usada num teste de ajustamento global num quadro-resumo da regressão:

Fonte	graus de liberdade	Somas de Quadrados	Quadrados Médios	F_{calc}
Regressão	p	$SQR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	$QMR = \frac{SQR}{p}$	$\frac{QMR}{QMRE}$
Resíduos	$n-(p+1)$	$SQRE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$QMRE = \frac{SQRE}{n-p-1}$	
Total	$n-1$	$SQT = \sum_{i=1}^{n}(y_i - \bar{y})^2$	–	–

3.10 Modelo e Submodelos: o teste F parcial

Um problema que surge em modelos de regressão linear múltipla (e que não fazia sentido nas regressões lineares simples) é o de saber se um dado modelo que se considera ter um ajustamento adequado, pode eventualmente ser simplificado. Recordemos o *princípio da parcimónia* na modelação: entre dois modelos que descrevam adequadamente uma dada variável resposta, preferimos o que seja mais simples (mais parcimonioso).

A aplicação deste princípio no nosso contexto traduz-se em saber se, caso se disponha de um modelo de Regressão Linear Múltipla com um ajustamento considerado adequado, *será possível obter um modelo com menos variáveis preditoras, sem perder significativamente em termos de qualidade de ajustamento.*

Exemplifiquemos a ideia partindo de um modelo de Regressão Linear Múltipla com cinco variáveis preditoras. A equação de base é assim:

$$Y \;=\; \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \;.$$

Chamamos *submodelo* a um modelo de regressão linear múltipla contendo apenas algumas das variáveis preditoras, como seria por exemplo o seguinte modelo de regressão linear múltipla, apenas com os preditores x_2 e x_5:

$$Y \;=\; \beta_0 \qquad + \beta_2 x_2 \qquad\qquad + \beta_5 x_5 \;.$$

Note-se que se opta por manter a indexação que as variáveis preditoras tinham no modelo completo original, a fim de evitar confusões. Podemos identificar o submodelo através do conjunto dos índices, $\mathcal{S}$, das variáveis preditoras que pertencem ao submodelo. No exemplo acima, ter-se-ia $\mathcal{S} = \{2, 5\}$.

O modelo completo e um seu submodelo são idênticos caso $\beta_j = 0$ para todas as variáveis preditoras X_j cujo índice *não* pertença a $\mathcal{S}$. Aqui, apenas se consideram índices j maiores ou iguais a 1, que são os índices correspondentes a variáveis preditoras. A constante aditiva β_0 não intervem nesta discussão.

3.10.1 O teste F parcial, para comparar um modelo e submodelo

Para avaliar se um dado modelo difere significativamente dum seu submodelo (identificado pelo conjunto $\mathcal{S}$ dos índices das suas variáveis), precisamos de optar entre as hipóteses:

$$H_0 : \beta_j = 0\,, \quad \forall j \notin \mathcal{S} \qquad \text{vs.} \quad H_1 : \exists\, j \notin \mathcal{S} \quad \text{tal que } \beta_j \neq 0.$$

(SUBMODELO = MODELO) (SUBMODELO PIOR)

Admite-se que β_0 faz sempre parte dos submodelos considerados. Por isso, a constante aditiva β_0 não faz parte das hipóteses em confronto. Note-se que β_0 não é relevante do ponto de vista da parcimónia: a sua presença não implica trabalho adicional de recolha de dados, nem de interpretação do modelo (ao mesmo tempo que permite um melhor ajustamento do modelo).

3.10.1.1 Estatística de teste para comparar modelo e submodelo

A estatística de teste envolve a comparação das Somas de Quadrados Residuais do:

- *modelo completo* (referenciado pelo índice C e com p preditores); e do
- *submodelo* (referenciado pelo índice S e com k preditores).

Proposição 3.17: A estatística do teste F parcial

Seja dado um Modelo de Regressão Linear Múltipla com p preditores, e um seu submodelo com apenas $k < p$ preditores. Caso $\beta_j = 0$, para todas as variáveis preditoras X_j não pertencentes ao submodelo, tem-se o seguinte resultado:

$$F \;=\; \frac{\frac{SQRE_S - SQRE_C}{p-k}}{\frac{SQRE_C}{n-(p+1)}} \quad \frown \quad F_{p-k\,,\,n-(p+1)}\ ,$$

onde $SQRE_C$ e $SQRE_S$ indicam as Somas de Quadrados Residuais, respectivamente, do modelo completo e do submodelo.

Notas:

- Omite-se a demonstração deste resultado.
- A condição na qual a distribuição da quantidade F é $F_{p-k,n-(p+1)}$ corresponde à Hipótese Nula acima referida.
- O denominador de F é o Quadrado Médio Residual do modelo completo.
- Verifica-se necessariamente que $SQRE_S \geq SQRE_C$, uma vez que a variabilidade explicada pelo modelo completo nunca pode aumentar se se deixarem alguns preditores fora do modelo. Mas enquanto as duas Somas de Quadrados Residuais forem próximas em valor (ou seja, enquanto o valor de F fôr próximo de zero), não há razões para duvidar de que o modelo e o submodelo difiram. Quanto maior fôr $SQRE_S$ em relação a $SQRE_C$, mais duvidosa será a Hipótese Nula. Assim, são os valores grandes da estatística que levantam dúvidas sobre a Hipótese Nula, H_0, o que aponta para uma *região crítica unilateral direita* do teste.

Tem-se então o seguinte teste de hipóteses para comparar um modelo completo de regressão linear múltipla com um seu submodelo, teste este que é conhecido por *teste F parcial* ou *teste F a modelos encaixados* (*nested models*, em inglês).

3.10.1.2 Os passos do teste F parcial

Hipóteses:

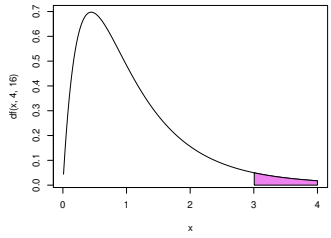
$H_0 : \beta_j = 0\,, \quad \forall j \notin \mathcal{S} \quad \text{vs.} \quad H_1 : \exists\, j \notin \mathcal{S} \quad \text{tal que} \quad \beta_j \neq 0.$

Estatística do Teste:

$F = \frac{(SQRE_S - SQRE_C)/(p-k)}{SQRE_C/[n-(p+1)]} \quad \frown \quad F_{p-k\,,\,n-(p+1)}$, sob H_0.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{\alpha[p-k\,,\,n-(p+1)]}$.



3.10.1.3 Expressões alternativas para as Hipóteses e a estatística do teste

A estatística do teste F de comparação de um modelo completo com p preditores, e um seu submodelo com apenas k preditores pode ser escrita numa forma alternativa, envolvendo os Coeficientes de Determinação amostrais do modelo completo (R^2_C) e do submodelo (R^2_S).

Assinale-se que a Soma de Quadrados Total $SQT = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = (n-1)\, S^2_Y$ não depende do modelo ajustado, sendo assim igual para o modelo e submodelo (desde que eles sejam ajustados com as mesmas observações da variável resposta). Esta mesma quantidade SQT é decomposta de formas diferentes no modelo e submodelo, gerando assim também diferentes valores dos coeficientes de determinação:

	Decomposição	Coeficiente de Determinação
Modelo Completo	$SQT = SQR_C + SQRE_C$	$R^2_C = \frac{SQR_C}{SQT}$
Submodelo	$SQT = SQR_S + SQRE_S$	$R^2_S = \frac{SQR_S}{SQT}$

Usando estas decomposições alternativas, tem-se o seguinte resultado.

Proposição 3.18: Expressão alternativa da estatística do teste F parcial

Uma expressão alternativa da estatística do teste F parcial é dada por:

$$F = \frac{n-(p+1)}{p-k} \cdot \frac{R_C^2 - R_S^2}{1-R_C^2} \ . \tag{3.25}$$

Demonstração 3.13: Proposição 3.18

A estatística do teste F parcial pode ser re-escrita da seguinte forma:

$$\begin{aligned} F &= \frac{n-(p+1)}{p-k} \frac{SQRE_S - SQRE_C}{SQRE_C} = \frac{n-(p+1)}{p-k} \frac{(\cancel{SQT} - SQR_S) - (\cancel{SQT} - SQR_C)}{SQT - SQR_C} \\ &= \frac{n-(p+1)}{p-k} \frac{SQR_C - SQR_S}{SQT - SQR_C} \cdot \underbrace{\frac{\frac{1}{SQT}}{\frac{1}{SQT}}}_{=1} = \frac{n-(p+1)}{p-k} \frac{\frac{SQR_C}{SQT} - \frac{SQR_S}{SQT}}{\frac{SQT}{SQT} - \frac{SQR_C}{SQT}} \\ &= \frac{n-(p+1)}{p-k} \cdot \frac{R_C^2 - R_S^2}{1-R_C^2} \ . \end{aligned}$$

As hipóteses do teste também se podem escrever como

$$H_0 : \mathcal{R}_C^2 = \mathcal{R}_S^2 \qquad \text{vs.} \qquad H_1 : \mathcal{R}_C^2 > \mathcal{R}_S^2 \ ,$$

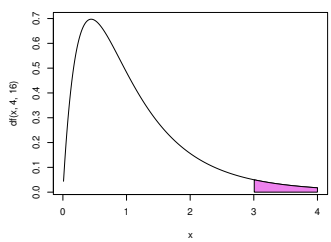
A hipótese H_0 indica que o grau de relacionamento linear entre Y e o conjunto dos preditores é idêntico no modelo e no submodelo. Juntando as hipóteses assim expressas à expressão alternativa da estatística (equação 3.25), tem-se a seguinte formulação alternativa do teste F parcial.

Hipóteses: $H_0 : \mathcal{R}_C^2 = \mathcal{R}_S^2$ vs. $H_1 : \mathcal{R}_C^2 > \mathcal{R}_S^2$.

Estatística do Teste: $F = \frac{n-(p+1)}{p-k} \cdot \frac{R_C^2 - R_S^2}{1-R_C^2} \ \frown \ F_{p-k\,,\,n-(p+1)}$, sob H_0.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{\alpha[p-k\,,\,n-(p+1)]}$



Caso não se rejeite a hipótese nula H_0, os modelo e submodelo não podem ser considerados significativamente diferentes. Nesse caso, o princípio da parcimónia sugere a opção pelo submodelo (mais parcimonioso). Caso se rejeite H_0, o modelo deve ser considerado significativamente melhor do que o submodelo pelo que, do ponto de vista estatístico, é aconselhada a escolha do modelo completo. Como sempre, estas conclusões podem ser condicionadas por considerações de outro tipo, mas não devem ignorar a discussão de base estatística agora descrita.

3.10.2 O teste F parcial a submodelos no R

A informação necessária para um teste F parcial obtém-se no R através da função `anova`, com dois argumentos: os objectos `lm` resultantes de ajustar o modelo completo e o submodelo sob comparação. No conjunto de dados dos lírios, a comparação entre o modelo completo de regressão linear múltipla ajustado na Subsecção 3.2.7 e o submodelo de regressão linear simples ajustado na Subsecção 2.6.2 produz os seguintes resultados:

```
> anova(iris.lm, iris2.lm)
  Analysis of Variance Table
Model 1: Petal.Width ~ Petal.Length
Model 2: Petal.Width ~ Petal.Length + Sepal.Length + Sepal.Width
  Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
1    148 6.3101
2    146 5.3803   2    0.9298 12.616 8.836e-06 ***
```

Os valores indicados na coluna `RSS` correspondem às Somas de Quadrados Residuais (*Residual Sums of Squares*, em inglês) de cada modelo. À esquerda, na coluna de nome `Res.Df`, encontram-se os respectivos graus de liberdade: $n-(p+1)$ no caso do modelo completo e $n-(k+1)$ no caso do submodelo. Na coluna de nome `Df` encontra-se a diferença entre estes graus de liberdade (ou seja, $p-k$) e na coluna de nome `Sum of Sq` encontra-se a diferença das Somas de Quadrados Residuais (ou seja, a diferença $SQRE_S - SQRE_C$). Finalmente, na coluna de nome F está o valor da estatística do teste F parcial descrito na Subsecção 3.10.1.1 e, ao lado, o respectivo valor de prova (*p-value*).

O valor calculado da estatística é $F_{calc} = 12.616$ e o respectivo *p-value* é $p = 8.836 \times 10^{-6}$, pelo que se rejeita a hipótese nula de igualdade de modelo e submodelo: o ajustamento do modelo completo deve ser considerado significativamente melhor do que o ajustamento do submodelo.

Deve sublinhar-se que este teste F parcial *não* equivale a efectuar testes *t-Student* separados às hipóteses de que cada preditor que distingue o modelo e o submodelo esteja *individualmente* associado a um coeficiente β_j nulo. Exemplifiquemos considerando os dados relativos ao Exercício RLM 2, com os dados `brix` (relativos a framboesas). A tabela associada à regressão da variável *Brix* sobre as restantes é:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.08878    1.00252   6.073 0.000298 ***
Diametro     1.27093    0.51219   2.481 0.038030 *
Altura      -0.70967    0.41098  -1.727 0.122478
Peso        -0.20453    0.14096  -1.451 0.184841
```

```
pH          0.51557    0.33733   1.528 0.164942
Acucar      0.08971    0.03611   2.484 0.037866 *
```

Nas duas colunas finais (de nomes `t value` e `Pr(>|t|)`) está a informação relativa aos testes t a hipóteses do tipo $\beta_j = 0$ (valor calculado da estatística T e respectivo *p-value*). Essa informação pode ser usada para identificar eventuais preditores que não desempenhem um papel significativo na previsão da variável resposta Y. De facto, se $\beta_j = 0$ (com $j \geq 1$), a parcela $\beta_j x_j$ na equação do modelo será sempre nula, independentemente dos valores de x_j. Nesse caso, a variável x_j não participa na definição dos valores da variável resposta Y. Ora, pela tabela acima, conclui-se que qualquer das variáveis preditoras *Altura*, *Peso* ou *pH* pode ser *individualmente* excluída do modelo sem afectar de forma significativa a qualidade do modelo (os respectivos *p-values* são todos maiores do que os níveis usuais de α, pelo que não se rejeita a hipótese nula $\beta_j = 0$). Mas *não* é legítimo concluir que *Altura*, *Peso* e *pH* são *simultaneamente* dispensáveis, como se pode verificar procedendo a um teste F parcial comparando as regressões lineares múltiplas com, e sem, esses preditores:

```
> anova(brix2.lm,brix.lm)
Analysis of Variance Table
Model 1: Brix ~ Diametro + Acucar
Model 2: Brix ~ Diametro + Altura + Peso + pH + Acucar
  Res.Df     RSS Df Sum of Sq    F  Pr(>F)
1     11 0.42743
2      8 0.14925  3   0.27818 4.97 0.03104 *
```

3.10.3 Relação entre os testes-t a parâmetros individuais e o teste F parcial

Caso o modelo e submodelo difiram num único preditor, X_j, o teste F parcial descrito nesta Secção é equivalente ao teste *t-Student* às hipóteses $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$. Nesse caso, não apenas as hipóteses dos dois testes são iguais, como a estatística do teste F parcial é o quadrado da estatística do teste t referido. Tem-se $p - k = 1$, e como é sabido (ver os apontamentos da disciplina de Estatística dos primeiros ciclos do ISA), se uma variável aleatória T tem distribuição t_ν, então o seu quadrado, T^2 tem distribuição $F_{1,\nu}$.

Assim, o teste F parcial para comparar o modelo de regressão linear múltipla da variável `Brix` sobre todos os restantes preditores, com o modelo de regressão linear múltipla de `Brix` sobre os quatro preditores que resultam de excluir o preditor `Altura` será (a menos de erros de arredondamento) $F_{calc} = T^2_{calc} = (-1.727)^2 = 2.9825$. O respectivo valor de prova tem de ser igual ao do teste a que $\beta_2 = 0$, ou seja, $p = 0.122478$. Ilustremos, ajustando esse modelo de quatro preditores e efectuando o teste F parcial.

```
> brix4.lm <- lm(Brix ~ Diametro + Peso + pH + Acucar, data=brix)
> anova(brix4.lm, brix.lm)
Analysis of Variance Table
Model 1: Brix ~ Diametro + Peso + pH + Acucar
Model 2: Brix ~ Diametro + Altura + Peso + pH + Acucar
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1      9 0.20489
```

```
2      8 0.14925  1  0.055631 2.9818 0.1225
```

3.10.4 Uma nota a propósito do teste F parcial *

(*) A matéria desta Subsecção não é avaliada

No contexto do estudo do modelo de regressão linear múltipla, o teste F parcial foi apresentado como ferramenta para comparar um modelo completo, com p variáveis preditoras, e um seu submodelo, em que apenas se retêm k dos p preditores originais, ambos ajustados com o mesmo conjunto de n observações.

Na realidade, o teste F parcial é de aplicação mais geral. O teste é aplicável na comparação de dois modelos para os quais os subespaços de $\mathbb{R}^n$ gerados pelas colunas das respectivas matrizes do modelo, $\mathbf{X}$, estejam contidos um no outro. Em concreto, considere-se um modelo de RLM com p preditores, cuja matriz associada é $\mathbf{X}_c$, e outro modelo, com k preditores, cuja matriz associada é $\mathbf{X}_s$. Se o espaço das colunas de $\mathbf{X}_s$ estiver contido no espaço das colunas de $\mathbf{X}_c$, ou seja, se $\mathcal{C}(\mathbf{X}_s) \subset \mathcal{C}(\mathbf{X}_c)$, então pode aplicar-se o teste F parcial para testar a hipótese nula de que os dois modelos coincidem (contra a alternativa de que não coincidem). A estatística do teste será dada pela mesmas expressões vistas atrás:

$$F \;=\; \frac{\frac{SQRE_s - SQRE_c}{p-k}}{\frac{SQRE_c}{n-(p+1)}} \;=\; \frac{n-(p+1)}{p-k} \cdot \frac{R_c^2 - R_s^2}{1-R_c^2} \;.$$

Caso os dois modelos sejam equivalentes, esta estatística tem uma distribuição $F_{(p-k,n-(p+1))}$[1].

No caso das colunas da matriz do modelo $\mathbf{X}_s$ serem um subconjunto das colunas da matriz do modelo $\mathbf{X}_c$ (o caso discutido nas aulas, correspondente a ter-se um submodelo constituído apenas por algumas das variáveis preditoras do modelo completo), a condição $\mathcal{C}(\mathbf{X}_s) \subset \mathcal{C}(\mathbf{X}_c)$ verifica-se sempre, uma vez que qualquer combinação linear das colunas de $\mathbf{X}_s$ ($\mathbf{X}_s\vec{\mathbf{a}}$) também se pode escrever como combinação linear das colunas de $\mathbf{X}_c$, bastando associar às colunas da matriz $\mathbf{X}_c$ que não sejam colunas de $\mathbf{X}_s$ o coeficiente zero, e às colunas comuns às duas matrizes os mesmos coeficientes (dados pelos elementos do vector $\vec{\mathbf{a}}$). Mas a condição $\mathcal{C}(\mathbf{X}_s) \subset \mathcal{C}(\mathbf{X}_c)$ é de aplicação mais geral, como se verá de seguida.

Vamos exemplificar esta generalização ilustrando uma forma alternativa de resolver a alínea f) do Exercício RLM 9, onde se pede para testar a igualdade de dois parâmetros β_j num modelo de regressão linear múltipla. Seguidamente, veremos como se pode usar a mesma ideia para estudar a hipótese de igualdade entre três ou mais parâmetros β_j.

1. No Exercício 9 de regressão linear múltipla estudam-se os dados relativos a $n=600$ folhas de videira, nas quais se observam a área foliar (variável resposta, `Area`, em cm^2) e os comprimentos de três nervuras (as variáveis preditoras): a nervura principal (`NP`), a nervura lateral esquerda (`NLesq`) e a nervura lateral direita (`NLdir`), todas em cm. A equação do modelo é:

$$Area \;=\; \beta_0 + \beta_1\, NP + \beta_2\, NLesq + \beta_3\, NLdir + \epsilon. \tag{3.26}$$

   Assim, a matriz do modelo $\mathbf{X}_c$ é composta por quatro colunas: uma coluna de n uns, uma coluna com os n valores observados da variável `NP`, uma terceira coluna com os n valores observados de

[1] Na realidade, $p-k$ indica a diferença nas dimensões dos subespaços encaixados, $\mathcal{C}(\mathbf{X}_c)$ e $\mathcal{C}(\mathbf{X}_s)$.

NLesq, e uma coluna final com os n valores observados de NLdir. O modelo ajustado tinha um coeficiente de determinação $R^2_c = 0.8649$.

Na alínea f) do Exercício 9 pede-se para estudar a hipótese $H_0 : \beta_2 = \beta_3$. Esse estudo foi feito considerando a hipótese equivalente $H_0 : \beta_2 - \beta_3 = 0$, e utilizando os testes t relativos a combinações lineares $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$ dos parâmetros do modelo. Usando a estatística de teste $T = \frac{(\hat{\beta}_2 - \hat{\beta}_3) - 0}{\hat{\sigma}_{\hat{\beta}_2 - \hat{\beta}_3}} \frown t_{(n-(p+1))}$, obteve-se o valor calculado $T_{calc} = -0.3636027$. O valor de prova respectivo pode ser calculado (dado tratar-se dum teste com Região Crítica bilateral, e dum valor de T_{calc} na parte esquerda da distribuição) como $p = 2 \times P[T_{596} < T_{calc}]$. Com o auxílio do R, obtém-se:

```
> 2*pt(-0.3636027, 596)
[1] 0.7162836
```

2. No entanto, poder-se-ia proceder da seguinte forma alternativa. A hipótese nula $H_0 : \beta_2 = \beta_3$ corresponde ao modelo de regressão linear múltipla de equação:

$$Area \;=\; \beta_0 + \beta_1\, NP + \beta_2\, (NLesq + NLdir) + \epsilon. \tag{3.27}$$

Trata-se dum modelo com $k = 2$ variáveis preditoras, as variáveis NP e *a soma* das variáveis NLesq e NLdir. A matriz deste modelo, $\mathbf{X}_s$, tem três colunas: uma coluna de n uns, uma coluna com os n valores observados da variável NP, e uma coluna final com as n somas de valores das duas nervuras laterais, NLesq+NLdir. Ora, qualquer combinação linear destas três colunas se pode escrever também como combinação linear das quatro colunas da matriz $\mathbf{X}_c$, bastando igualar, nesta última, os coeficientes individuais de NLesq e NLdir. Assim, o subespaço das colunas da matriz $\mathbf{X}_s$ está contido no subespaço das colunas da matriz $\mathbf{X}_c$, ou seja, $\mathcal{C}(\mathbf{X}_s) \subset \mathcal{C}(\mathbf{X}_c)$. Será então possível efectuar um teste F parcial para comparar os modelos (3.26) e (3.27). Com o auxílio do R, tem-se:

```
> videiras.lm <- lm(Area ~ NP + NLesq + NLdir, data=videiras)
> vid2Betas.lm <- lm(Area ~ NP + I(NLesq+NLdir), data=videiras)
> anova(vid2Betas.lm, videiras.lm)
Analysis of Variance Table

Model 1: Area ~ NP + I(NLesq + NLdir)
Model 2: Area ~ NP + NLesq + NLdir
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    597 365391
2    596 365310  1    81.001 0.1322 0.7163
```

Ou seja, a estatística tem valor calculado $F_{calc} = 0.1322$, com valor de prova $p = 0.7163$. Não é uma coincidência que o valor de prova seja o mesmo que foi obtido na resolução alternativa baseada no teste *t-Student*. Tal como não é uma coincidência que o quadrado do valor então calculado da estatística T seja o valor agora calculado da estatística F: $T^2_{calc} = (-0.3636027)^2 = 0.1322069$. Esta relação ilustra que também se generaliza a relação que sabíamos existir na aplicação dum teste F parcial para comparar um modelo com p preditores e um seu submodelo com apenas p–1 preditores, ou seja, resultante da exclusão dum único preditor.

3. Consideremos agora o exemplo de se querer testar a igualdade de três ou mais coeficientes β_j num modelo RLM. Este problema já não poderia ser estudado considerando a teoria de combinações lineares dos β_j dada nas aulas. Mas pode ser abordado através dum teste F parcial, de forma

análoga à acima ilustrada. Continuemos com o exemplo dos dados do Exercício 7, e consideremos a única hipótese deste tipo possível, a hipótese de que, no modelo 3.26, os coeficientes populacionais dos comprimentos das três nervuras sejam iguais, ou seja, $H_0 : \beta_1 = \beta_2 = \beta_3$. A essa hipótese corresponde um novo modelo, de equação:

$$Area = \beta_0 + \beta_1 (NP + NLesq + NLdir) + \epsilon. \tag{3.28}$$

Neste novo modelo há apenas $k = 1$ preditor: a soma dos comprimentos das três nervuras. A matriz do modelo $\mathbf{X}_s$ correspondente tem agora apenas duas colunas: a coluna de n uns, e a coluna destas n somas das três nervuras. Qualquer combinação linear destas duas colunas pode também escrever-se como combinação linear das quatro colunas da matriz $\mathbf{X}_c$ do modelo original, bastando usar o coeficiente de `NP+NLesq+NLdir` nas três colunas de $\mathbf{X}_c$ correspondentes a estas três variáveis individuais. Logo, de novo, $\mathcal{C}(\mathbf{X}_s) \subset \mathcal{C}(\mathbf{X}_c)$. Podemos efectuar um teste F parcial para testar a igualdade dos modelos (3.26) e (3.28). Com o auxílio do `R`:

```
> vid3Betas.lm <- lm(Area ~ I(NP+NLesq+NLdir), data=videiras)
> anova(vid3Betas.lm, videiras.lm)
Analysis of Variance Table

Model 1: Area ~ I(NP + NLesq + NLdir)
Model 2: Area ~ NP + NLesq + NLdir
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    598 365766
2    596 365310  2    456.13 0.3721 0.6895
```

Também neste caso, não se rejeita H_0 para nenhum dos níveis de significância habituais, pelo que se considera admissível a hipótese $\beta_1 = \beta_2 = \beta_3$.

Nota: Em todos os exemplos considerados, não se discute o problema da curvatura que parece existir na relação de fundo, e que é visível nos gráficos de resíduos estudados na alínea h) do Exercício RLM 9. Esse problema pode ser ultrapassado de várias formas. Uma envolve uma regressão linear sobre as transformações logarítmicas das variáveis, como considerado nas alíneas i) e j) do Exercício RLM 9. Outra envolve uma regressão polinomial, como será exemplificado mais adiante. Esta discussão ilustra a ideia de que podem existir diferentes modelos concorrentes para modelar um mesmo conjunto de dados.

3.11 A escolha dum submodelo

O teste F parcial (teste aos modelos encaixados) permite-nos optar entre um modelo e um seu submodelo. Por vezes, um submodelo pode ser sugerido por:

- *razões de índole teórica*, sugerindo que determinadas variáveis preditoras não sejam, na realidade, importantes para influenciar os valores de Y.
- *razões de índole prática*, como a dificuldade, custo ou volume de trabalho associado à recolha de observações para determinadas variáveis preditoras.

Nestes casos, pode ser claro que submodelo(s) se deseja testar. Veja-sa o Exercício RLM 11 g) (dados relativos à produção de milho) para um exemplo.

Mas em muitas situações não é evidente qual o subconjunto de variáveis preditoras que se deseja considerar no submodelo. Pretende-se apenas verificar se o modelo é simplificável, e em caso afirmativo, escolher um submodelo mais simples cuja qualidade não difira de forma significativa da qualidade do modelo completo original. Nestes casos, a identificação de um tal submodelo não é um problema fácil. De facto, dadas p variáveis preditoras, o número de subconjuntos de preditores, de qualquer cardinalidade excepto 0 (conjunto vazio) e p (o modelo completo), que é possível escolher *é dado por* $2^p - 2$. A tabela seguinte indica o número desses subconjuntos para $p = 5, 10, 15, 20, 30$.

p	$2^p - 2$
5	30
10	1 022
15	32 766
20	1 048 574
30	1 073 741 822

Para pequenos valores de p, é viável analisar todos os possíveis submodelos (ou seja, os modelos de regressão linear com todos os possíveis subconjuntos de preditores). Com algoritmos e rotinas informáticas adequadas, essa avaliação completa de todos os subconjuntos ainda é possível para valores de p até $p \approx 35$. Mas para p muito grande, uma pesquisa exaustiva é computacionalmente inviável.

Registe-se que não é legítimo usar os testes t à significância individual de cada coeficiente β_j, no modelo completo, para decidir sobre a exclusão de vários preditores *em simultâneo*, como já se viu na Subsecção 3.10.2. Os testes t aos coeficientes individuais β_j partem do princípio que todas as restantes variáveis pertencem ao modelo. A exclusão de um qualquer preditor altera os valores estimados b_j e os respectivos erros padrão associados às variáveis que permanecem no submodelo. Pode acontecer que um preditor seja dispensável num modelo completo, mas deixe de o ser num submodelo, ou viceversa.

3.11.1 Algoritmos de pesquisas exaustivas

Para um número p de preditores pequeno ou médio, e dispondo de algoritmos e rotinas informáticas adequadas, é possível efectuar uma pesquisa exaustiva, que assegure a identificação do subconjunto de k preditores com o maior valor de R^2 (ou outro critério de qualidade do submodelo). O algoritmo *leaps and bounds*, de Furnival e Wilson [2] é um algoritmo computacionalmente eficiente que permite identificar o melhor subconjunto de preditores, de uma dada cardinalidade k.

Uma rotina implementando o algoritmo encontra-se disponível no `R`, num módulo (*package*) de nome `leaps` (comando com o mesmo nome). Outra rotina análoga encontra-se na função `eleaps` do módulo `subselect`.

[2] Furnival, G.W and Wilson, R.W.,Jr. (1974) Regressions by leaps and bounds, *Technometrics*, **16**, 499-511.

3.11.1.1 Um exemplo de pesquisa exaustiva

Aploque-se a rotina `leaps`, acima referida, aos dados do Exercício RLM 2 (dados `brix`). O módulo do `R` que contém esta rotina (igualmente de nome `leaps`) terá de ter sido previamente instalado. Admitindo que esse módulo já tenha sido instalado, eis os passos a seguir:

```
> library(leaps)   <--- carregar o módulo (tem de estar previamente instalado)
> colnames(brix)
[1] "Diametro" "Altura"   "Peso"     "Brix"     "pH"       "Acucar"

> leaps(y=brix$Brix, x=brix[,-4], method="r2", nbest=1)  <--- y resposta, x preditores

$which  <--- matriz de valores lógicos, indicando resultados (cada coluna um preditor,
      1     2     3     4     5         cada linha uma cardinalidade de subconjunto)
1 FALSE FALSE FALSE FALSE  TRUE <--- k=1 ; melhor preditor individual: Acucar
2  TRUE  TRUE FALSE FALSE FALSE <--- k=2 ; melhor par de preditores: Diametro e Altura
3  TRUE  TRUE FALSE FALSE  TRUE <--- k=3 ; melhor trio: Diametro, Altura, Acucar
4  TRUE  TRUE FALSE  TRUE  TRUE
5  TRUE  TRUE  TRUE  TRUE  TRUE
[...]

$r2     <--- Coef. Determinação da melhor solução com o no. k=1,2,3,4,5 de preditores
[1] 0.5091325 0.6639105 0.7863475 0.8083178 0.8482525
```

Repare-se como o melhor submodelo (R^2 mais elevado) com dois preditores *não é* o submodelo com os preditores `Diametro` e `Acucar`, como sugerido pelos *p-values* do ajustamento do modelo completo. Repare-se ainda como *não* é verdade que o melhor subconjunto de k preditores tenha de estar contido no melhor subconjunto de $k+1$ preditores (veja-se para $k=1$).

3.11.2 Algoritmos de pesquisa sequencial

Caso não esteja disponível *software* apropriado, ou se o número p de preditores fôr demasiado grande, pode recorrer-se a algoritmos de pesquisa heurísticos, que simplificam uma regressão linear múltipla *sem a garantia de obter os melhores subconjuntos*, trocando a garantia duma pesquisa exaustiva por tempos computacionais viáveis.

Considere-se um algoritmo que, em cada passo, exclui uma variável preditora, até alcançar uma condição de paragem considerada adequada, ou seja, um *algoritmo de exclusão sequencial* (*backward elimination*, em inglês).

Eis os passos dum **algoritmo de exclusão sequencial**, baseado nos testes t às hipóteses de que $\beta_j=0$:

1. definir um nível de significância α para os testes de hipóteses;

2. considerar inicialmente o modelo completo, com os p preditores;

3. ajustar o modelo;

4. para todas as variáveis rejeita-se a hipótese $\beta_j = 0$?

   - Se sim, não é possível simplificar o modelo: passar ao ponto 5.
   - Se não, qualquer das variáveis em que *não* se rejeita $H_0 : \beta_j = 0$ é candidata a sair do modelo.
     - se apenas existe uma variável candidata a sair, excluir essa variável;
     - se existir mais do que uma variável candidata a sair, excluir a variável associada ao maior *p-value* (isto é, ao valor da estatística t mais próxima de zero);

   Regressar ao ponto 3.

5. Quando não existirem variáveis candidatas a sair, ou quando sobrar um único preditor, o algoritmo pára. Tem-se então o *submodelo final*.

Existem variantes deste algoritmo, que não serão estudadas nesta disciplina, nomeadamente:

- o *algoritmo de inclusão sequencial* (*forward selection*, em inglês), cuja ideia geral é de proceder de baixo para cima, começando por considerar a melhor regressão linear simples e depois, em passos sucessivos, avaliar se se justifica a inclusão de algum outro preditor e, em caso afirmativo, qual.
- algoritmos de *exclusão/inclusão alternada* (*stepwise selection*), que combinam, alternadamente, passos de inclusão e de exclusão de variáveis preditoras.

3.11.2.1 Um exemplo de aplicação do algoritmo de exclusão sequencial

Consideremos os dados `brix`, relativos a medições de 6 variáveis em 14 framboesas e descritos no enunciado do Exercício RLM 2. Eis o ajustamento da regressão linear múltipla da variável resposta `Brix` sobre os restantes $p=5$ preditores.

```
> summary(lm(Brix ~ Diametro + Altura + Peso + pH + Acucar, data=brix))
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.08878    1.00252   6.073 0.000298 ***
Diametro     1.27093    0.51219   2.481 0.038030 *
Altura      -0.70967    0.41098  -1.727 0.122478
Peso        -0.20453    0.14096  -1.451 0.184841
pH           0.51557    0.33733   1.528 0.164942
Acucar       0.08971    0.03611   2.484 0.037866 *
```

Fixando o nível de significância $\alpha = 0.05$, verifica-se que há três variáveis preditoras para as quais o teste *t-Student* não rejeita a hipótese $H_0 : \beta_j = 0$, nomeadamente no caso dos preditores `Altura` (*p-value* $0.122478 > 0.05 = \alpha$), `Peso` ($p = 0.184841$) e `pH` ($p = 0.164942$). Assim, qualquer destas variáveis preditoras poderia (individualmente!) ser excluída do modelo sem afectar significativamente a qualidade do ajustamento. No entanto, *não* há justificação estatística para proceder à exclusão *simultânea* destes três preditores, como já se viu. Podendo excluir-se apenas um preditor, procede-se a escolher aquele para

a qual a estatística calculada (no teste a $\beta_j = 0$) está mais longe da região crítica, ou seja, tem o maior *p-value* associado. Neste caso, trata-se da variável preditora `Peso`.

Procede-se então a ajustar o modelo de regressão linear múltipla resultante da exclusão do preditor `Peso`.

```
> summary(lm(Brix ~ Diametro + Altura + pH + Acucar, data=brix))
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.25964    1.05494   5.934 0.000220 ***
Diametro     1.40573    0.53373   2.634 0.027189 *
Altura      -1.06413    0.35021  -3.039 0.014050 *  <-- Passou a ser significativo
pH           0.33844    0.33322   1.016 0.336316
Acucar       0.08481    0.03810   2.226 0.053031 .  <-- Deixou de ser significativo
```

O ajustamento deste submodelo tem mudanças importantes em relação ao ajustamento do modelo completo original. Todas as estimativas, quer da constante aditiva, quer dos coeficientes dos preditores, são diferentes dos obtidos no modelo completo, tal como os respectivos erros padrão. Em dois casos, essas diferenças alteram qualitativamente a conclusão dos testes a $\beta_j = 0$ (para o nível $\alpha = 0.05$ considerado): o preditor `Altura`, que era considerado dispensável no modelo completo, passou a ser considerado indispensável após a exclusão do preditor `Peso`, tendo agora um *p-value* $p = 0.014050 < 0.05 = \alpha$. Em sentido contrário foi o preditor `Acucar`, cujo *p-value* (0.053031) é agora (embora muito ligeiramente) superior a $\alpha = 0.05$ e que portanto, ao não se rejeitar a hipótese do seu coeficiente β_j ser nulo, passa a ser considerado dispensável no modelo.

Nota: Em parte, estes resultados reflectem o facto de haver relativamente poucas observações ($n = 14$) para o número de parâmetros a estimar. Mas o exemplo ilustra que este tipo de comportamento é possível, sublinhando a complexidade do problema de escolher submodelos duma regressão linear múltipla.

Com o submodelo de $k = 4$ preditores agora ajustado, há duas variáveis preditoras cuja exclusão (individualmente) do modelo pode ser considerada, por terem $p > \alpha$: `pH` e `Acucar`. Sendo apenas possível excluir um preditor em cada passo, procede-se à exclusão do preditor `pH`, cujo *p-value* é o mais elevado.

Ajusta-se agora o modelo com os três preditores `Diametro`, `Altura` e `Acucar`.

```
> summary(lm(Brix ~ Diametro + Altura + Acucar, data=brix))
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.97183    0.78941   8.832  4.9e-06 ***
Diametro     1.57932    0.50642   3.119  0.01090 *
Altura      -1.11589    0.34702  -3.216  0.00924 **
Acucar       0.09039    0.03776   2.394  0.03771 *    <-- Voltou a ser significativo
```

Não havendo agora preditores para os quais a hipótese $\beta_j = 0$ não é rejeitada, todas as variáveis preditoras são consideradas indispensáveis para não piorar significativamente a qualidade de ajustamento do modelo. *O algoritmo pára aqui*, escolhendo este submodelo final com 3 preditores.

Refira-se que, do ponto de vista do coeficiente de determinação R^2, a simplificação do modelo registou uma baixa do valor inicial $R^2 = 0.8483$ no modelo com cinco preditores, para um valor final de $R^2 = 0.7863$ no modelo com três preditores. Sempre que o algoritmo envolva mais do que uma exclusão, deixa de haver

a garantia de que o submodelo final e o modelo completo original não possam diferir significativamente, para o mesmo valor de α fixado. Caso se considere necessário, pode comparar-se o submodelo final com o modelo completo, através dum teste F parcial. No caso do exemplo agora considerado, os modelos inicial e final *não* são considerados significativamente diferentes para qualquer dos níveis de significância usuais:

```
> anova(brix3.lm, brix.lm)
Analysis of Variance Table
Model 1: Brix ~ Diametro + Altura + Acucar
Model 2: Brix ~ Diametro + Altura + Peso + pH + Acucar
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     10 0.21014
2      8 0.14925  2  0.060888 1.6318 0.2545
```

3.11.3 O Critério de Informação de Akaike e algoritmos com base no AIC

O R disponibiliza funções para automatizar pesquisas sequenciais de submodelos, semelhantes à que aqui foi enunciada, mas em que o critério de exclusão duma variável em cada passo se baseia no *Critério de Informação de Akaike (AIC)*.

O AIC é uma medida geral da qualidade de ajustamento de modelos, baseada no conceito de *verosimilhança* (em inglês *likelihood*), que extravasa o âmbito da disciplina. Mas no contexto duma Regressão Linear Múltipla com k variáveis preditoras, a definição geral do AIC traduz-se na seguinte fórmula.

Definição 3.12: AIC dum modelo de regressão linear múltipla

Seja dado um modelo de Regressão Linear Múltipla com k variáveis preditoras e ajustado com base em n observações. Então, o respectivo Critério de Informação de Akaike (AIC) define-se como:

$$AIC \;=\; n \cdot \ln\left(\frac{SQRE_k}{n}\right) + 2(k+1)\;, \tag{3.29}$$

onde $SQRE_k$ é a Soma de Quadrados Residual do modelo.

Nota: O AIC *pode tomar valores negativos*, uma vez que $SRQE_k$ pode ser inferior ao número de observações n, em cujo caso a primeira parcela é negativa.

Interpretando o AIC

- a primeira parcela na definição do AIC é uma função crescente de $SQRE_k$. Logo, é uma medida da *qualidade do ajustamento* e *quanto melhor o ajustamento, mais pequena a primeira parcela*;
- a segunda parcela na definição do AIC é uma medida da *complexidade* do modelo, uma vez que k+1 é o número de parâmetros do modelo. Tendo em conta o princípio da parcimónia, *quanto mais pequena a segunda parcela, melhor*.

Assim, *o AIC mede simultaneamente a qualidade do ajustamento e a complexidade do modelo*, e quanto menor fôr o valor do AIC, melhor. *Valores do AIC para diferentes modelos de regressão linear são comparáveis, desde que nesses modelos a variável resposta Y seja a mesma e desde que os modelos sejam ajustados com os mesmos dados.* Assinale-se que esta comparação é possível *mesmo que não se trate de modelos encaixados*, ou seja, mesmo que ambos os modelos contenham preditores que não fazem parte do outro modelo. Também nesse caso, *um modelo é considerado melhor que outro se tiver um AIC menor.*

3.11.3.1 Algoritmos sequenciais com base no AIC

Pode definir-se uma variante do algoritmo de exclusão sequencial, de espírito análogo ao que já foi considerado, mas com base no critério AIC. Eis os passos deste algoritmo.

1. começar com o modelo completo.
2. ajustar o modelo e calcular o respectivo AIC.
3. ajustar cada submodelo com menos *uma* variável e calcular os respectivos AICs em cada um desses submodelos.
    - Se nenhum dos AICs dos submodelos considerados fôr inferior ao AIC do modelo anterior, o algoritmo termina, sendo o modelo anterior o modelo final.
    - Caso alguma das exclusões reduza o AIC, efectua-se a exclusão da variável preditora que mais reduz o AIC e regressa-se ao ponto 2.

Deve assinalar-se que, ao considerar-se diferentes submodelos com igual número k de preditores, o submodelo com menor AIC será sempre aquele que tiver menor $SQRE$. Assim, num dado passo do algoritmo, a comparação será sempre entre o modelo inicialmente ajustado e o submodelo (com menos um preditor) a que corresponda o melhor ajustamento, ou seja, que tiver excluído a variável cujo teste a $\beta_j = 0$ tem maior *p-value*.

Assim, o algoritmo de exclusão sequencial baseado nos testes t e o algoritmo baseado no AIC *coincidem na ordenação das variáveis a excluir*. Apenas podem diferir na decisão de efectuar ou não a exclusão de um preditor, ou seja, apenas podem diferir no critério de paragem do algoritmo.

Em geral, um algoritmo de exclusão sequencial baseado no AIC é mais 'cauteloso' na decisão de excluir um preditor do que um algoritmo baseado nos testes t, sobretudo se o valor de α usado nesses testes t fôr baixo. A fim de contrariar um excesso de propensão para excluir preditores, é aconselhável usar nos algoritmos de exclusão baseados nos testes t, valores relativamente elevados de α, como por exemplo $\alpha = 0.10$.

3.11.3.2 Algoritmos de pesquisa sequencial no `R`

A função `step` do `R` corre algoritmos de selecção de submodelos baseados no Critério de Informação de Akaike (AIC). A função `step` permite, através do argumento `dir` (ou, por extenso, `direction`), indicar se se deseja um argumento de exclusão sequencial (`dir="backward"`), de inclusão sequencial

(`dir="forward"`) ou de exclusão/inclusão alternadas (a opção por omissão, ou explicitando-se o argumento `dir`, com a opção `dir="both"`). Exemplifique-se o algoritmo de exclusão sequencial, de novo usando os dados `brix` do Exercício RLM 2.

```
> brix.lm <- lm(Brix ~ Diametro + Altura + Peso + pH + Acucar, data=brix)
> step(brix.lm, dir="backward")
Start:  AIC=-51.58
Brix ~ Diametro + Altura + Peso + pH + Acucar
           Df Sum of Sq     RSS     AIC
<none>                  0.14925 -51.576
- Peso      1  0.039279 0.18853 -50.306
- pH        1  0.043581 0.19284 -49.990
- Altura    1  0.055631 0.20489 -49.141
- Diametro  1  0.114874 0.26413 -45.585
- Acucar    1  0.115132 0.26439 -45.572
```

Os vários modelos ensaiados são *ordenados por ordem crescente de AIC*, pelo que os melhores modelos considerados em cada passo do algoritmo surgirão sempre à cabeça da lista de modelos ensaiados. Neste caso, o melhor modelo é o modelo identificado por `<none>` que corresponde ao modelo em que *não se exclui qualquer variável*, ou seja ao modelo inicial. De facto, e como se pode observar na coluna de nome `AIC`, o AIC do modelo inicial é inferior ao de qualquer submodelo resultante de excluir uma variável. Registe-se que os submodelos são identificados no início de cada linha através da indicação do nome da variável preditora cuja exclusão (sinal `-`) é experimentada. Assim, por exemplo, a exclusão da variável `Peso` provocaria um *aumento* do AIC, que passaria de -51.576 para -50.306. Se, em vez do preditor `Peso` fossem excluído algum outro preditor, os respectivos valores de AIC seriam sempre maiores. Assim, com esta variante do algoritmo de exclusão sequencial baseado no Critério de Informação de Akaike, o *submodelo final é o modelo completo inicial.*

Este exemplo ilustra o comentário feito acima: embora o melhor submodelo de quatro preditores corresponda à exclusão do preditor `Peso`, que foi o preditor excluído no primeiro passo do algortimo baseado nos testes t a que $\beta_j = 0$, no caso do critério AIC essa exclusão não chega a ser feita, obtendo-se um submodelo final com mais preditores do que no caso da variante baseada nos testes t.

Uma advertência sobre algoritmos de pesquisa heurísticos (ou seja, que não garantem uma pesquisa exaustiva entre todos os possíveis submodelos), como é o caso do algoritmo de exclusão sequencial (nas duas variantes agora consideradas). Estas heurísticas *não* garantem a identificação do melhor submodelo com um dado número k de preditores. Apenas identificam, de forma que não é computacionalmente muito pesada, submodelos bons.

Outra advertência final, relativa a processos de selecção de submodelos duma regressão linear múltipla: os processos agora descritos devem ser usados com bom senso e eventuais submodelos obtidos devem ser cruzados com outras considerações (como por exemplo, o custo ou dificuldade de obtenção de cada variável, ou o papel que a teoria relativa ao problema em questão reserva a cada preditor).

3.12 A Regressão Polinomial

Um *caso particular de relação não-linear*, mesmo que envolvendo apenas uma variável preditora e a variável resposta, pode ser facilmente tratada no âmbito duma regressão linear múltipla: o caso de *relações polinomiais* entre Y e um ou mais preditores.

Consideremos os dados do Exercício RLM 9, relativos a medições sobre $n = 600$ folhas de videira, e a relação entre a variável resposta área foliar (`Area`) e um único preditor, o comprimento da nervura principal (`NP`). A nuvem de pontos respectiva é indicada na Figura 3.10, e sobre ela foi traçada a recta de regressão. Há curvilinearidade na nuvem de pontos, um fenómeno presente também no modelo de regressão linear múltipla de `Area` sobre o comprimento das três nervuras, estudado no Exercício RLM 9. Poderá essa forma encurvada da nuvem de pontos ser bem descrita por uma parábola? Esta hipótese é sugerida pela constatação que o preditor `NP` é uma medida de comprimento (dimensão um), enquanto a variável resposta é uma área (dimensão dois).

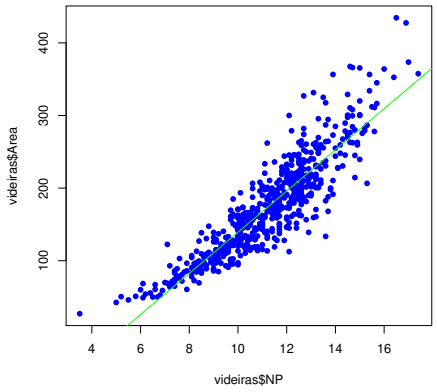


Figura 3.10: A nuvem de pontos correspondente às medições, sobre $n = 600$ folhas de videira, de duas variáveis: área foliar (varivel `Area`, em cm^2), no eixo vertical e comprimento da nervura principal (variável `NP`, em cm) no eixo horizontal. Sobre a nuvem foi traçada a recta de regressão linear. A curvilinearidade da nuvem de pontos indica que a recta está a subestimar as áreas foliares das folhas com menores, e com maiores, comprimentos de nervura, ou seja, das folhas mais pequenas e das folhas maiores.

Qualquer parábola, com equação

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 \,, \tag{3.30}$$

pode ser ajustada e estudada como se se tratasse duma regressão linear entre a variável resposta Y e *duas* variáveis preditoras: a variável X original, que passamos a designar como variável $X_1 = X$, e a variável preditora definida pelos *quadrados de* X, que passamos a designar o preditor $X_2 = X^2$. Ajustemos um modelo assim definido aos dados do Exercício RLM 9 (videiras). Assinale-se a utilização da função `I` na fórmula associada ao comando `lm`, para *i*nibir a interpretação do símbolo `^` como um operador especial na notação das fórmulas e obrigando à sua utilização como de facto se pretende, ou seja, como operador aritmético de cálculo duma potência.

```
> summary(lm(Area ~ NP + I(NP^2), data=videiras))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.5961    22.0431   0.345    0.731
NP           -0.2172     4.0125  -0.054    0.957
I(NP^2)       1.2941     0.1801   7.187 1.98e-12 ***
---
Residual standard error: 28.86 on 597 degrees of freedom
Multiple R-squared: 0.8162,  Adjusted R-squared: 0.8155
F-statistic:  1325 on 2 and 597 DF,  p-value: < 2.2e-16
```

A equação da *parábola ajustada* resulta de utilizar os três coeficientes ajustados pelo método de mínimos quadrados na equação 3.30, ou seja, é dada por

$$y \;=\; b_0 + b_1 x_1 + b_2 x_2 \;=\; 7.5961 - 0.2172\, x + 1.2941\, x^2 \;.$$

A parábola ajustada pode ser vista na Figura 3.11.

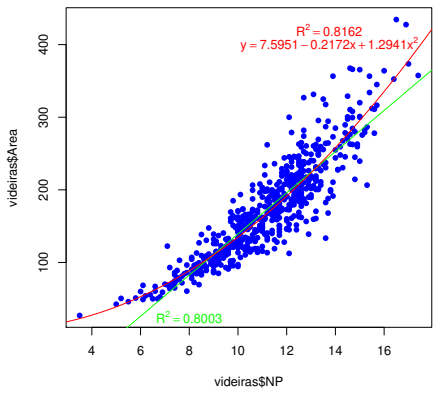


Figura 3.11: A nuvem de $n=600$ pontos relacionando `Area` e `NP`, com sobreposta a parábola de mínimos quadrados, ajustada como indicado no texto.

Repare-se que o modelo de regressão linear simples, cuja equação de base é $Y=\beta_0+\beta_1\, x$, é um submodelo do modelo parabólico agora ajustado, correspondente a ter-se $\beta_2=0$ na equação 3.30. No teste t a essa hipótese, obtém-se um valor de prova $p=1.98\times 10^{-12}$, logo inferior a qualquer dos valores usuais de α, levando assim à rejeição clara da hipótese $H_0 : \beta_2=0$. Pode assim afirmar-se que esta parábola tem um ajustamento significativamente melhor que a recta de regressão linear simples de `Area` sobre `NP`, ou seja, que o submodelo da equação 3.30 resultante de admitir que $\beta_2=0$.

É legítimo afirmar que o modelo de regressão polinomial (quadrático) agora ajustado explica $R^2=81.62\%$ da variabilidade nas áreas foliares observadas, uma vez que *não houve transformação da variável resposta* Y. Trata-se duma pequena melhoria face ao valor (comparável) $R^2=0.8004$ da regressão linear simples,

mas uma diferença que é, não apenas significativa (como o teste *t-Student* atrás realizado confirma), mas sobretudo reflectindo um modelo que tende a aproximar melhor as áreas foliares de folhas de qualquer tamanho, nomeadamente as folhas mais pequenas e maiores.

O argumento acima ilustrado é extensível a qualquer polinómio de qualquer grau, e em qualquer número de variáveis. Consdieremos dois exemplos:

- Um polinómio de grau p numa única variável é da forma

$$Y = \beta_0 + \beta_1 \underbrace{x}_{=x_1} + \beta_2 \underbrace{x^2}_{=x_2} + \beta_3 \underbrace{x^3}_{=x_3} + ... + \beta_p \underbrace{x^p}_{=x_p} ,$$

  e pode assim ser ajustado como se fosse uma regressão linear múltipla com p 'variáveis preditoras' (as primeiras p potências da única variável preditora x).

- Um polinómio de grau 2 em duas variáveis é da forma

$$Y = \beta_0 + \beta_1 \underbrace{x}_{=x_1} + \beta_2 \underbrace{x^2}_{=x_2} + \beta_3 \underbrace{z}_{=x_3} + \beta_4 \underbrace{z^2}_{=x_4} + \beta_5 \underbrace{xz}_{=x_5} ,$$

  e corresponderia a ajustar uma regressão linear múltipla nas cinco variáveis preditoras acima indicadas.

3.13 O R^2 modificado

O R^2 *modificado* (*adjusted* R^2, em inglês) é uma variante do Coeficiente de Determinação que visa avaliar a qualidade do ajustamento levando em conta a relação entre o número de observações disponíveis (n) e o número $(p+1)$ de parâmetros β_j que é necessário estimar.

Comecemos por relembrar a definição do R^2 usual, e por reescrevê-la com base na Soma de Quadrados Residual:

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQRE}{SQT} .$$

O R^2 modificado resulta de substituir as Somas de Quadrados, nesta última expressão, por Quadrados Médios, como indicado na Definição seguinte.

Definição 3.13: R^2 modificado

Seja dada uma regressão linear múltipla, com p variáveis preditoras, ajustada com base em n observações. Então define-se o R^2 modificado, R^2_{mod}, como sendo:

$$R^2_{mod} = 1 - \frac{QMRE}{QMT} ,$$

onde $QMT = \frac{SQT}{n-1} = s^2_y$ e $QMRE = \frac{SQRE}{n-(p+1)}$.

Esta definição merece várias considerações.

1. Recordando que o Quadrado Médio Residual é o estimador (centrado) da variância dos erros aleatórios, σ^2, que é simultaneamente a variância das observações Y_i em torno da superfície linear que relaciona Y e os seus p preditores, verifica-se que o R^2_{mod} pode ser escrito como $R^2_{mod} = 1 - \frac{\hat{\sigma}^2}{s^2_y}$. A segunda parcela é a razão entre a *variância de Y em torno da superfície linear que a relaciona com os p preditores* ($\hat{\sigma}^2 = QMRE$) e a *variância das n observações de Y, sem referência a qualquer modelo explicativo* (s^2_y). Assim, a fracção $\frac{QMRE}{QMT}$ mede a redução na variabilidade inexplicada de Y, antes e depois de explicar parte dessa variabilidade através da relação linear com os preditores. O valor $R^2_{mod} = 1 - \frac{\hat{\sigma}^2}{s^2_y} = \frac{s^2_y - \hat{\sigma}^2}{s^2_y}$ mede assim a *redução relativa na variabilidade inexplicada de Y*.

2. Pode ainda deduzir-se uma relação directa entre o valor de R^2_{mod} e o valor do R^2 usual:

$$R^2_{mod} = 1 - \frac{QMRE}{QMT} = 1 - \frac{SQRE}{SQT} \cdot \frac{n-1}{n-(p+1)} = 1 - (1 - R^2) \cdot \frac{n-1}{n-(p+1)} \,. \tag{3.31}$$

3. Tem-se sempre $n-1 > n-(p+1)$, pelo que, a partir da relação no ponto anterior, se verifica sempre a seguinte relação: $R^2_{mod} < R^2$.

4. Quando $n \gg p+1$ (ou seja, quando há muito mais observações que parâmetros no modelo), o valor do factor $\frac{n-1}{n-(p+1)}$ é muito próximo de 1, pelo que as duas variantes de R^2 têm valores aproximadamente iguais: $R^2 \approx R^2_{mod}$.

5. Se n é pouco maior que o número de variáveis preditoras, então $\frac{n-1}{n-(p+1)}$ é grande, e R^2_{mod} vem bastante inferior a R^2, excepto quando R^2 fôr muito próximo de 1. Assim, o R^2_{mod} penaliza ajustamentos de modelos em que o número de observações não seja muito maior que o número de parâmetros do modelo (excepto se o valor de R^2 inicial fôr já muito próximo de 1).

6. Algumas características importantes a que estamos habituados no R^2 deixam de ser verdade no R^2_{mod}, uma das quais é o facto de R^2_{mod} poder tomar *valores negativos*. No Exercício RLM 22 mostra-se que a condição para que isso aconteça é que o R^2 usual seja inferior à razão entre o número de variáveis preditoras (p) e o número de observações menos 1 ($n-1$).

Exemplifiquemos o uso do Coeficiente de Determinação modificado com os dados `brix`, introduzidos no Exercício RLM 2 e já considerados anteriormente. Trata-se dum conjunto de dados com $n = 14$ observações e em que a regressão linear múltipla completa tem $p = 5$ variáveis preditoras e $p+1 = 6$ parâmetros. O valor do Coeficiente de Determinação usual é $R^2 = 0.8483$. Mas o facto de o número de observações não chegar ao dobro do número de parâmetros do modelo significa que o factor de penalização na expressão para R^2_{mod} dada na equação (3.31) é $\frac{n-1}{n-(p+1)} = \frac{13}{8} = 1.625$. Este factor vai aumentar em 62.5% a proporção de variabilidade não explicada pelo modelo ($1 - R^2 = 0.1517$), elevando-o para 0.2465. O valor final de R^2_{mod} é a diferença deste valor para a unidade, ou seja, $1 - 0.2465 = 0.7535$. Confirmemos (sem os erros de arredondamento), com o auxílio do `R` que, na penúltima linha da listagem produzida pelo comando `summary` aplicado a uma regressão linear, fornece os valores das duas variantes de R^2.

```
> summary(brix.lm)
[...]
Multiple R-squared:  0.8483, Adjusted R-squared:  0.7534
```

Outra chamada de atenção importante é que, ao contrário do que sucede com o R^2 usual, *um submodelo pode ter R^2_{mod} maior do que um modelo completo*. Ilustremos esta ideia recorrendo a outro conjunto

de dados analisado nas aulas práticas, ou seja os dados `milho` do exercício RLM 11. Aproveita-se este exemplo para também chamar a atenção de que a já referidajá referidajá referida rotina `leaps`, que faz a pesquisa exaustiva de submodelos com $k < p$ preditores, também aceita (através do argumento `method`) o R^2_{mod} como critério a optimizar. Em baixo invoca-se este comando no estudo dos dados do Exercício RLM 11.

```
> library(leaps)
> leaps(y=milho$y , x=milho[,-10], method="adjr2", nbest=1)
[...]
$adjr2         <--- o maior R2 modificado é no submodelo com k=4 preditores
[1] 0.5493014 0.6337329 0.6544835 0.6807418 0.6798986 0.6779395 0.6745412
[8] 0.6633467 0.6488148
```

3.14 Análise de Resíduos e outros diagnósticos

Tal como na Regressão Linear Simples, uma análise de regressão linear múltipla não fica completa sem o estudo dos resíduos e de alguns outros diagnósticos. Este estudo adquire uma importância ainda maior na Regressão Linear Múltipla onde, em geral, deixa de ser possível visualizar a nuvem de pontos original à procura de características como curvaturas na relação de fundo, heterogeneidade de variâncias dos erros aleatórios, etc.

Grande parte do que se disse sobre resíduos na Regressão Linear Simples mantém-se válido numa Regressão Linear Múltipla, havendo apenas que proceder a alguns ajustamentos de pormenor em certas definições.

Comecemos por relembrar que o objectivo fundamental do estudo de resíduos é o de validar os pressupostos do Modelo Linear, nomeadamente a linearidade como relação de fundo e os pressupostos relativos aos erros aleatórios ϵ_i: Normalidade, média zero, variância constante σ^2 e independência. Tal como na regressão linear simples, não é possível estudar estes pressupostos directamente sobre os erros aleatórios, uma vez que eles não são conhecíveis, mesmo após a recolha de uma amostra. De facto, e a partir da equação do modelo (primeiro ponto na Definição 3.5), tem-se:

$$\epsilon_i = Y_i - (\beta_0 + \beta_1 x_{1(i)} + \beta_2 x_{2(i)} + ... + \beta_p x_{p(i)}) \ .$$

Substituindo os parâmetros β_j desconhecidos pelos seus estimadores $\hat{\beta}_j$ obtêm-se os *resíduos* (enquanto variáveis aleatórias), como indicado na Definição seguinte.

Definição 3.14

Seja dado o Modelo de Regressão Linear Múltipla (Definição 3.5). Definem-se as variáveis aleatórias resíduos de cada observação como sendo:

$$E_i \ = \ Y_i - \hat{Y}_i \ = \ Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1(i)} + \hat{\beta}_2 x_{2(i)} + ... + \hat{\beta}_p x_{p(i)}) \ .$$

Após a selecção dum *amostra concreta*, estas variáveis aleatórias tomam os *valores numéricos*:

$$e_i \;=\; y_i - \hat{y}_i \;=\; y_i - (b_0 + b_1 x_{1(i)} + b_2 x_{2(i)} + ... + b_p x_{p(i)}) \;.$$

3.14.1 Propriedades dos Resíduos sob o Modelo RLM

O modelo de Regressão Linear Múltipla admite que

$$\epsilon_i \;\frown\; \mathcal{N}(0\,,\,\sigma^2) \qquad \forall\, i = 1, ..., n \;.$$

Sob o modelo RLM, as variáveis aleatórias *resíduos* têm a seguinte distribuição:

$$E_i \;\frown\; \mathcal{N}\left(0\,,\,\sigma^2\,(1 - h_{ii})\right) \qquad \forall i = 1, ..., n \;, \tag{3.32}$$

onde h_{ii} é o i-ésimo elemento diagonal da matriz $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$, de projecção ortogonal sobre o subespaço $\mathcal{C}(\mathbf{X})$.

Este resultado será demonstrado determinando a distribuição, ao abrigo do Modelo RLM, do *vector* dos n resíduos, ou seja, do vector:

$$\vec{\mathbf{E}} \;= \vec{\mathbf{Y}} - \vec{\hat{\mathbf{Y}}} \;=\; \vec{\mathbf{Y}} - \mathbf{H}\vec{\mathbf{Y}} \;=\; (\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}} \;. \tag{3.33}$$

Na Proposição seguinte indica-se a distribuição de probabilidades deste vector aleatório.

Proposição 3.19: Distribuição dos Resíduos no modelo RLM

Dado o Modelo de Regressão Linear Múltipla, o vector dos resíduos $\vec{\mathbf{E}} = (\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}$ tem distribuição:

$$\vec{\mathbf{E}} \;\frown\; \mathcal{N}_n\left(\vec{\mathbf{0}}\,,\, \sigma^2(\mathbf{I}_n - \mathbf{H})\right) \;.$$

Demonstração 3.14: Proposição 3.19

O vector dos resíduos $\vec{\mathbf{E}} = \vec{\mathbf{Y}} - \vec{\hat{\mathbf{Y}}} = \vec{\mathbf{Y}} - \mathbf{H}\vec{\mathbf{Y}} = (\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}$ é o produto duma matriz não aleatória $(\mathbf{I}_n - \mathbf{H})$ e um vector aleatório $(\vec{\mathbf{Y}})$, sendo que este vector aleatório $\vec{\mathbf{Y}}$ tem distribuição Multinormal. Assim, $\vec{\mathbf{E}}$ também tem distribuição Multinormal, tendo em conta a última propriedade da Proposição 3.7 (página 106). Falta determinar os dois parâmetros dessa distribuição Multinormal, ou seja, o vector esperado e a matriz de (co-)variâncias de $\vec{\mathbf{E}}$.

O vector esperado de $\vec{\mathbf{E}}$ resulta das propriedades da Proposição 3.5 e da distribuição de $\vec{\mathbf{Y}}$ (Proposição 3.8):

$$E[\vec{\mathbf{E}}] \;=\; E[(\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}] \;=\; (\mathbf{I}_n - \mathbf{H})\underbrace{E[\vec{\mathbf{Y}}]}_{=\mathbf{X}\vec{\boldsymbol{\beta}}} \;=\; (\mathbf{I}_n - \mathbf{H})\mathbf{X}\vec{\boldsymbol{\beta}} \;=\; \mathbf{X}\vec{\boldsymbol{\beta}} - \underbrace{\mathbf{H}\mathbf{X}\vec{\boldsymbol{\beta}}}_{=\mathbf{X}\vec{\boldsymbol{\beta}}} \;=\; \mathbf{X}\vec{\boldsymbol{\beta}} - \mathbf{X}\vec{\boldsymbol{\beta}} \;=\; \vec{\mathbf{0}} \;,$$

pois o vector $\mathbf{X}\vec{\boldsymbol{\beta}} \in \mathcal{C}(\mathbf{X})$ é uma combinação linear das colunas de $\mathbf{X}$, logo permanece invariante sob a acção da matriz de projecção $\mathbf{H}$: $\mathbf{H}\mathbf{X}\vec{\boldsymbol{\beta}} = \mathbf{X}\vec{\boldsymbol{\beta}}$ (veja-se também o Exercício RLM 4).

No que respeita à matriz de (co-)variâncias do vector aleatório dos resíduos, $\vec{\mathbf{E}}$, calcula-se a partir das propriedades dessas matrizes (Proposição 3.6) e do facto de a matriz de projecção ortogonal ser (veja-se o Exercício RLM 4) simétrica ($\mathbf{H}^t = \mathbf{H}$) e idempotente ($\mathbf{H}^2 = \mathbf{H}\,\mathbf{H} = \mathbf{H}$):

$$\begin{aligned} V[\vec{\mathbf{E}}] &= V[(\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}] = (\mathbf{I}_n - \mathbf{H}) \underbrace{V[\vec{\mathbf{Y}}]}_{=\sigma^2\mathbf{I}_n} (\mathbf{I}_n - \mathbf{H})^t = \sigma^2\,(\mathbf{I}_n - \mathbf{H})\mathbf{I}_n(\mathbf{I}_n^t - \mathbf{H}^t) \\ &= \sigma^2\,(\mathbf{I}_n - \mathbf{H})(\mathbf{I}_n - \mathbf{H}) = \sigma^2(\mathbf{I}_n - \mathbf{H} - \cancel{\mathbf{H}} + \cancel{\mathbf{H}\mathbf{H}}) = \sigma^2(\mathbf{I}_n - \mathbf{H})\ . \end{aligned}$$

Notas:

1. Como elementos individuais dum vector Multinormal têm distribuição Normal, está garantida a Normalidade de cada resíduo E_i. O respectivo valor esperado tem de ser nulo (como são todos os elementos do vector esperado $E[\vec{\mathbf{E}}] = \vec{\mathbf{0}}$). A respectiva variância é dada pelo i-ésimo elemento diagonal da matriz de (co-)variâncias de $\vec{\mathbf{E}}$, ou seja, por $\sigma^2\,(1 - h_{ii})$. Assim, e como indicado inicialmente, tem-se $E_i \frown \mathcal{N}\left(0, \sigma^2\,(1 - h_{ii})\right)$, para qualquer resíduo E_i.

2. A distribuição dos resíduos E_i é análoga à já considerada na regressão linear simples (embora não haja, ao contrário do que acontecia na regressão linear simples, uma fórmula alternativa para h_{ii}). Assim, grande parte da discussão sobre resíduos num Regressão Linear Múltipla será análoga à que já foi considerada na Regressão Linear Simples.

3. Embora no modelo RLM os erros aleatórios sejam independentes, *os resíduos não são variáveis aleatórias independentes*, pois as covariâncias entre resíduos diferentes não são (em geral), nulas:

$$cov[E_i, E_j] = -\sigma^2\,h_{ij}\ , \quad \text{se } i \neq j\ ,$$

onde h_{ij} indica o elemento da linha i e coluna j da matriz $\mathbf{H}$. Recorde-se que dois elementos dum vector Multinormal (como são E_i e E_j) são independentes se e só se tiverem covariância nula.

3.14.2 Análise dos resíduos e outros diagnósticos

3.14.2.1 Vários tipos de resíduos

Tal como na Regressão Linear Simples, definem-se diferentes tipos de resíduos:

Definição 3.15: Três tipos de resíduos

Resíduos usuais : $E_i = Y_i - \hat{Y}_i$;

Resíduos (internamente) estandardizados : $R_i = \frac{E_i}{\sqrt{QMRE\,(1-h_{ii})}}$.

Resíduos Studentizados (ou externamente estandardizados): $T_i = \frac{E_i}{\sqrt{QMRE_{[-i]}\,(1-h_{ii})}}$, sendo $QMRE_{[-i]}$ o valor de $QMRE$ resultante de um ajustamento da Regressão excluíndo a i-ésima observação (associada ao resíduo E_i).

3.14.2.2 Principais gráficos de resíduos

Tal como para a Regressão Linear Simples, também em regressões múltiplas se avalia a validade dos pressupostos do modelo através de *gráficos de resíduos*. Mas estes gráficos são agora *mais importantes do que na RLS*, dada a impossibilidade de visualização de nuvens de pontos em espaços de alta dimensionalidade.

Os gráficos mais usuais são os já considerados na RLS e a sua leitura faz-se de forma análoga:

gráfico de E_is vs. $\hat{Y}_i$s: os pontos devem-se dispor numa banda horizontal, centrada no valor zero, sem outro padrão especial. Curvaturas questionam a hipótese de linearidade e efeitos de tipo funil questionam o pressuposto de homogeneidade de variâncias.

***qq-plot* dos resíduos estandardizados:** a linearidade neste gráfico sustenta o pressuposto de Normalidade dos erros aleatórios.

gráfico de resíduos vs. ordem de observação: para investigar eventuais faltas de independência dos erros aleatórios.

3.14.2.3 O efeito alavanca

Como na Regressão Linear Simples, outras ferramentas de diagnóstico visam identificar observações individuais que merecem ulterior análise. Mas importa adaptar as definições ao contexto de Regressão Múltipla.

Definição 3.16: Efeito alavanca

Numa Regressão Linear Múltipla, o *valor de efeito alavanca* (*leverage*, em inglês) é o valor h_{ii} do elemento diagonal da matriz de projecção ortogonal $\mathbf{H}$, correspondente à observação i.

Continua a ser verdade que tem de ter-se $\frac{1}{n} \leq h_{ii} \leq 1$. O conceito tem uma interpretação análoga ao que tinha na regressão linear simples: uma vez que a variância dum resíduo é dada por $V[E_i] = \sigma^2(1-h_{ii})$ (equação 3.32), valores de h_{ii} muito grandes (próximos de 1) indicam que os resíduos têm, além de média zero, também variância próxima de zero, ou seja, serão necessariamente próximos de zero. Esse facto indica que a observação i 'obriga' a superfície ajustada a passar próximo dela. Tal facto pode ainda ser analisado de outra forma: uma vez que o vector dos valores *ajustados* de Y, é dado por $\vec{\hat{\mathbf{y}}} = \mathbf{H}\vec{\mathbf{y}}$, o i-ésimo valor ajustado é dado pelo produto interno da linha i da matriz $\mathbf{H}$ com o vector das observações $\vec{\mathbf{Y}}$, ou seja,

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j \ , \tag{3.34}$$

onde h_{ij} indica o elemento que está na linha i, coluna j, da matriz $\mathbf{H}$. Assim, cada valor ajustado de Y é uma combinação linear das n observações de Y. Como a soma dos coeficientes tem de ser 1 (pois $\mathbf{H}\vec{\mathbf{1}}_n = \vec{\mathbf{1}}_n$), valores de h_{ii} próximos de 1 significam que o i-ésimo valor ajustado $\hat{y}_i$ depende sobretudo da própria observação y_i, e pouco das restantes observações. Assim, tem-se $\hat{y}_i \approx y_i$, implicando que para o i-ésimo resíduo se tem $e_i = \hat{y}_i - y_i \approx 0$.

No entanto, a expressão do *valor médio* das observações alavanca numa RLM é agora dado por

$$\overline{h} = \frac{p+1}{n} \ ,$$

ou seja, pela razão entre o número de parâmetros e o número de observações.

3.14.2.4 Influência

Definição 3.17: Distância de Cook

A *distância de Cook* para avaliar a influência da observação i define-se agora como:

$$D_i = \frac{\sum_{j=1}^{n} \left[\hat{y}_j - \hat{y}_{j_{(-i)}}\right]^2}{(p+1)\, QMRE} \ ,$$

onde $\hat{y}_j$ o j-ésima valor ajustado, usando o hiperplano ajustado com as n observações, e $\hat{y}_{j_{(-i)}}$ o valor ajustado da j-ésima observação, mas usando o hiperplano ajustado sem a observação i.

Expressão equivalente é (sendo R_i o correspondente resíduo estandardizado) dada por:

$$D_i = R_i^2 \left(\frac{h_{ii}}{1-h_{ii}}\right) \frac{1}{p+1} \ .$$

Os restantes aspectos da discussão são análogos aos duma RL Simples, podendo contruir-se gráficos com estes diagnósticos.

3.14.2.5 Um exemplo de gráficos de diagnóstico

Um exemplo de gráficos de diagnósticos no contexto duma Regressão Linear Múltipla pode ser dado com os dados `brix`, do Exercício RLM 2, como mostrado na Figura 3.12. Estes gráficos foram obtidos aplicando o comando `plot` ao modelo de regressão linear ajustado, dando ao argumento `which` os valores `4` e `5`, respectivamente:

```
> plot(brix.lm, which=c(4,5))
```

Os valores muito grandes de várias distâncias de Cook e efeitos alavanca h_{ii} deste exemplo reflectem o reduzido número de observações ($n=14$) usado para ajustar um modelo com muitos parâmetros ($p+1=6$).

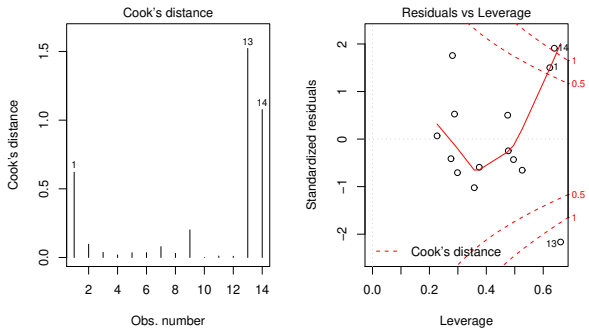


Figura 3.12: À esquerda, um diagrama de barras com as distâncias de Cook de cada observação. À direita, o gráfico de resíduos (internamente) estandardizados (eixo vertical) contra valores do efeito alavanca (h_{ii}) e, nos cantos do lado direito, as isolinhas correspondentes às distâncias de Cook 0.5 e 1. Repare-se nas três observações com influência muito grande ($D_i > 0.5$) e efeito alavanca elevado, facto que é também reflexo do número relativamente pequeno de observações disponíveis para ajustar este modelo.

3.15 Advertências finais

Para encerrar esta discussão de Regressões Lineares Múltiplas, deixemos algumas advertências:

1. Podem surgir problemas associados à *multicolinearidade* das variáveis preditoras, ou seja, ao facto das colunas da matriz $\mathbf{X}$ serem linearmente dependentes (multicolinearidade exacta) ou quase (multicolinearidade aproximada).

   No caso de multicolinearidade exacta nas colunas da matriz do modelo $\mathbf{X}$, não existe a inversa $(\mathbf{X}^t\mathbf{X})^{-1}$, que desempenha um papel fundamental na definição, quer da matriz de projecções ortogonais $\mathbf{H}$ (logo dos valores ajustados $\vec{\mathbf{Y}}$), quer do vector de estimadores $\vec{\boldsymbol{\beta}}$. Mesmo no caso de apenas existir multicolinearidade aproximada, têm-se vários problemas:

   - podem surgir problemas numéricos no cálculo de $(\mathbf{X}^t\mathbf{X})^{-1}$, logo no ajustamento do modelo e na estimação dos parâmetros.
   - haverá tendência para a existência de *variâncias muito grandes de alguns estimadores* $\hat{\beta}_i s$ (ou seja, de elementos diagonais grandes na matriz $\sigma^2\,(\mathbf{X}^t\mathbf{X})^{-1}$), o que significa muita incerteza na inferência (por exemplo, intervalos de confiança de muito grande amplitude).

   Um exemplo frequente de multicolinearidade exacta surge quando se utilizam variáveis preditoras que correspondem a diferentes percentagens cuja soma seja necessariamente 100% (como por exemplo, na textura dos solos). Nesse caso, a soma das colunas da matriz $\mathbf{X}$ correspondentes a esses preditores será igual a $100\,\vec{\mathbf{1}}_n$, pelo que existirá uma dependência linear exacta (ou aproximada, no caso de erros de arredondamento) nas colunas de $\mathbf{X}$.

   É possível eliminar multicolinearidades exactas ou aproximadas entre os preditores, através da exclusão de uma ou mais variáveis preditoras que sejam responsáveis pela dependência linear (exacta ou aproximada) dos preditores.

2. Tal como na Regressão Linear Simples, podem ser usadas transformações da variável resposta e/ou de (algumas ou todas) as variáveis preditoras. Em particular, podem ser úteis transformações que linearizem a relação entre Y e os preditores X_1, X_2, ..., X_p. Tal como na regressão linear simples, tais *transformações linearizantes* podem permitir estudar relações de tipo não-linear através de relações lineares entre as variáveis transformadas.

   Considerem-se por exemplo os dados do Exercício RLM 15, que correspondem a uma relação não linear, de tipo dupla potência, entre a variável resposta Y e dois preditores x_1 e x_2:

   $$y \;=\; a\, x_1^b\, x_2^c$$

   Logaritmizando, obtém-se uma relação linear entre $y^* = \ln(y)$, $x_1^* = \ln(x_1)$ e $x_2^* = \ln(x_2)$ (com $b_0 = \ln(a)$, $b_1 = b$ e $b_2 = c$):

   $$\ln(y) = \ln(a) + b\,\ln(x_1) + c\,\ln(x_2) \qquad \Leftrightarrow \qquad y^* = b_0 + b_1\, x_1^* + b_2\, x_2^*.$$

3. *Não se deve confundir a existência de uma relação linear entre preditores X_1, X_2, ..., X_p e variável resposta Y, com uma relação de causa e efeito.* Sendo possível que exista uma relação de causa e efeito, poderão também verificar-se outras situações, entre as quais:

   - Uma relação de *associação*, ou seja de *variação conjunta* dessas variáveis, mas não de tipo causal. Tome-se, por exemplo, variáveis morfométricas em que é frequente que animais ou plantas com uma medição maior numa característica sejam igualmente maiores nas restantes características (correspondendo a indivíduos globalmente maiores), mas sem que se possa dizer que um caule maior, por exemplo, *provoca* raízes maiores. Por vezes, quer os preditores, quer a variável resposta, são influenciadas por causas comuns subjacentes.
   - Uma relação totalmente *espúria*, de coincidência numérica.

   A existência duma relação *causal* (ou seja, de causa e efeito) só pode ser afirmada com base em teoria própria do fenómeno sob estudo, e não com base na relação linear estabelecida estatisticamente.

Capítulo 4

Análise de Variância

A Regressão Linear visa modelar uma variável resposta numérica (quantitativa), à custa de uma ou mais variáveis preditoras, igualmente numéricas. Mas uma variável resposta *numérica* pode depender de variáveis *qualitativas (categóricas)*, ou seja, de um ou mais **factores**. A **Análise de Variância (ANOVA)** é uma metodologia estatística para lidar com este tipo de situações. Foi desenvolvida nos anos 30 do Século XX, na Estação Experimental Agrícola de Rothamstead (Inglaterra), por R.A. Fisher.

4.1 Dois exemplos: os lírios por espécie

Considere-se de novo o conjunto de dados `iris`. Até aqui ignorou-se o facto de os 150 lírios para os quais existem informações pertencerem a três diferentes espécies: *Iris setosa*, *Iris versicolor* e *Iris virginica*.

Figura 4.1: As três espécies de lírios nos dados `iris`: à esquerda uma *Iris setosa*; a meio uma *Iris versicolor*; à direita uma *Iris virginica*.

É natural perguntar se os valores médios de cada característica morfométrica *diferem consoante as espécies*. Uma inspecção dos diagramas de extremos e quartis das variáveis morfométricas (numéricas) *por espécie*, pode sugerir respostas.

A Figura 4.2 sugere que a largura média das pétalas difere entre as espécies de lírios consideradas. No que respeita às larguras das sépalas, essas diferenças são menos pronunciadas. Mas, em qualquer caso, e uma vez que os diagramas de extremos e quartis foram construídos com apenas 50 observações de cada espécie,

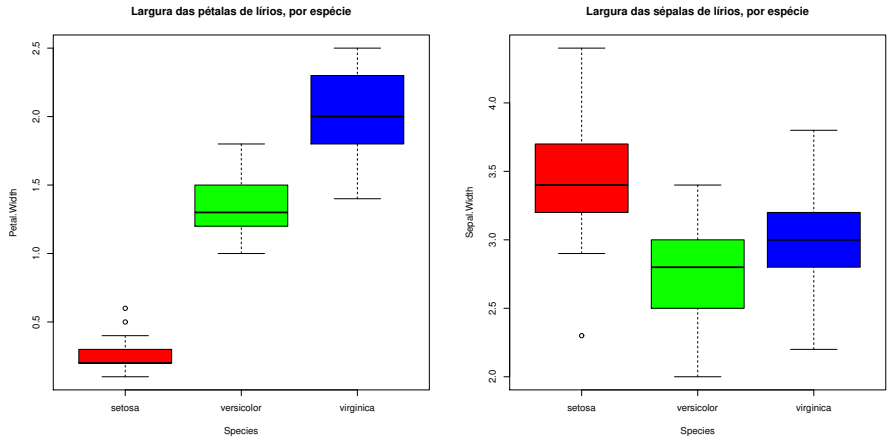


Figura 4.2: Os diagramas de extremos e quartis, por espécie de lírios, nos dados `iris`. À esquerda, relativos às larguras das pétalas. À direita, relativos às larguras das sépalas.

surge de forma natural um *problema inferencial*: pode afirmar-se que as diferenças observadas reflectem verdadeiras diferenças nos valores médios populacionais de cada espécie? Ou estamos perante diferenças apenas nos valores amostrais, e que resultam da variabilidade associada a qualquer amostragem?

4.2 A ANOVA como caso particular do Modelo Linear

Embora a Análise de Variância tenha historicamente surgido como método autónomo, quer a Análise de Variância, quer a Regressão Linear, são particularizações do Modelo Linear. Introduzir a ANOVA como um caso particular do Modelo Linear permite aproveitar a teoria estudada aquando da consideração das Regressões Lineares. Procurar-se-á enquadrar o mais possível o estudo da ANOVA no contexto geral analisado anteriormente.

4.2.1 Terminologia e notação

Fixemos terminologia e notação adequada ao contexto. Designa-se:

Variável resposta Y: uma variável *numérica* (quantitativa), que se pretende estudar e modelar.

Factor: uma variável preditora *categórica (qualitativa)*;

Níveis do factor: as diferentes categorias ("valores") do factor.

Os *níveis do factor* podem ser encarados como as diferentes *situações experimentais* onde se efectuam observações de Y.

Nos exemplos acima considerados, a variável resposta poderá ser a largura da pétala, ou a largura da sépala. O factor preditor é dado pela *espécie* de lírios, que é um factor com $k=3$ *níveis*.

A expressão **delineamento experimental** designa a forma como foi organizada a experiência, indicando aspectos como o número de observações da variável resposta que correspondem a cada nível de um dado factor. Muitas das considerações que serão feitas sobre delineamentos experimentais são genéricas para qualquer experiência que envolva a recolha de dados a serem usados em modelos estatísticos. Mas algumas considerações serão específicas das ANOVAs, razão pela qual se optou por apenas fazer a discussão deste tema nesta parte do programa.

4.3 A ANOVA a um Factor

No mais simples de todos os modelos ANOVA, a modelação da variável resposta baseia-se numa única variável preditora (categórica). Esse modelo designa-se *ANOVA a um Factor*, ou ainda, ANOVA a 1 factor totalmente casualizado (*one-way ANOVA* em inglês). Designamos por k o número de níveis do factor. No exemplo dos lírios, tem-se $k=3$ níveis do único factor preditor: as espécies.

Será conveniente adequar a notação ao contexto em apreço. Como de costume, admitimos que existem ao todo n *observações independentes de* Y, mas designaremos por n_i (com $i = 1, ..., k$) o número de observações correspondentes ao nível i do factor. Logo, $\sum_{i=1}^{k} n_i = n$.

O caso de *igual número de observações em cada nível* é importante e merece uma designação própria.

Definição 4.1: Delineamento equilibrado

Um delineamento a um factor, onde existe o mesmo número de observações associadas a qualquer nível do factor chama-se um **delineamento equilibrado**, podendo designar-se por n_c o *número comum de observações* em todos os níveis do factor:

$$n_1 = n_2 = n_3 = \cdots = n_k \qquad (= n_c) \, .$$

Os delineamentos equilibrados são aconselháveis, por várias razões que adiante se discutem. *Nos delineamentos equilibrados, existe igual quantidade de informação associada a cada uma das situações experimentais* que, no caso de um delineamento com um único factor, correspondem aos k níveis do factor.

4.3.1 A dupla indexação de Y

Na regressão linear indexavam-se as n observações de Y com um único índice (i), variando de 1 a n. Neste novo contexto, é preferível utilizar *dois índices para indexar as observações de* Y:

- um primeiro índice (i) indica o nível do factor a que a observação corresponde;
- um segundo índice (j) permite distinguir as observações num mesmo nível, também designadas por *repetições* nesse nível.

Assim, a j-ésima observação de Y, no i-ésimo nível do factor, é representada por Y_{ij} , (com $i=1,...,k$ e $j=1,...,n_i$).

4.3.2 A equação do modelo ANOVA a um factor

A equação do modelo será mais simples do que na regressão, reflectindo a natureza mais pobre da informação disponível para modelar a variável resposta Y. De facto, numa ANOVA a um único factor, a modelação Y_{ij} assenta apenas no facto de essa observação corresponder ao nível i do factor. Não há informação no modelo para explicar diferentes valores de Y em repetições num mesmo nível do factor. Assim, toda a variação de Y *no seio dum dado nível* será considerada variação aleatória, não explicada pelo factor. Esta variação aleatória será associada, como nas regressões lineares, a *erros aleatórios* aditivos, que também serão indicados pela dupla indexação (ϵ_{ij}) a fim de os associar a uma dada observação Y_{ij}.

Uma primeira equação do modelo pode ser a seguinte:

$$Y_{ij} \;=\; \mu_i + \epsilon_{ij} \qquad , \qquad \text{com} \quad E[\epsilon_{ij}] = 0 \;,$$

onde μ_i representa o valor esperado das observações Y_{ij} efectuadas no nível i do factor. Esta interpretação de μ_i resulta da exigência de que $E[\epsilon_{ij}] = 0$, já que: $\mu_i = E[Y_{ij}]$, ou seja, o valor esperado de qualquer observação no nível i do factor.

Para poder enquadrar a ANOVA na teoria do Modelo Linear já estudada, é conveniente re-escrever as médias de nível numa forma diferente, fazendo surgir uma constante aditiva comum a todas as observações.

$$E[Y_{ij}] \;=\; \mu_i \;=\; \mu + \alpha_i \;.$$

O parâmetro μ é comum a todas as observações, enquanto os parâmetros α_i são específicos para cada nível (i) do factor. Cada α_i é designado o **efeito do nível** i e admite-se que seja uma *constante* (o que por vezes leva a que este modelo seja chamado *de efeitos fixos*).

Tal como nos modelos de regressão linear, admite-se que as observações Y_{ij} oscilam aleatoriamente em torno do seu valor médio:

$$Y_{ij} \;=\; \mu + \alpha_i + \epsilon_{ij} \;, \tag{4.1}$$

com $E[\epsilon_{ij}] = 0$. Nesta forma, não é imediatamente evidente que estas equações sejam um caso particular da equação do Modelo Linear, uma vez que não se explicitam variáveis preditoras. Mas veremos em seguida que é possível re-escrever a equação (4.1) salientando a presença implícita de variáveis preditoras duma natureza especial.

4.3.2.1 As variáveis indicatrizes

A equação geral (4.1) significa que as n_1 observações efectuadas no nível $i = 1$ ficam:

$$Y_{1j} \;=\; \mu + \alpha_1 + \epsilon_{1j} \;,$$

as n_2 observações efectuadas no nível $i = 2$ ficam:

$$Y_{2j} \;=\; \mu + \alpha_2 + \epsilon_{2j} \;,$$

e por aí adiante, até $i=k$. Este conjunto de k equações pode ser escrita como uma única equação geral, introduzindo as chamadas **variáveis indicatrizes** de pertença duma dada observação a cada nível do factor. De facto, defina-se a variável $\boldsymbol{\mathcal{I}}_m$ (onde $m \in \{1, 2, ..., k\}$) que, para uma dada observação Y_{ij} toma valor 1 se a observação foi feita no nível m do factor, e 0 caso contrário, ou seja:

$$\boldsymbol{\mathcal{I}}_{m_{ij}} = \begin{cases} 1 & \text{se} \quad i = m \\ 0 & \text{se} \quad i \neq m \end{cases} \tag{4.2}$$

Com a ajuda destas variáveis indicatrizes, as equações (4.1), para qualquer nível i, podem escrever-se como uma única equação, do tipo das equações do Modelo Linear, com os vectores das variáveis indicatrizes a desempenharem o papel de variáveis preditoras:

$$Y_{ij} = \mu + \alpha_1 \boldsymbol{\mathcal{I}}_{1_{ij}} + \alpha_2 \boldsymbol{\mathcal{I}}_{2_{ij}} + ... + \alpha_k \boldsymbol{\mathcal{I}}_{k_{ij}} + \epsilon_{ij} \ ,$$

Definindo os vectores $\vec{\boldsymbol{\mathcal{I}}}_m$ com os n valores da variável indicatriz de cada nível m do factor, pode re-escrever-se esta equação do modelo em notação matricial/vectorial (à semelhança do que foi feito na Regressão Linear):

$$\vec{\mathbf{Y}} = \mu \, \vec{\mathbf{1}}_n + \alpha_1 \vec{\boldsymbol{\mathcal{I}}}_1 + \alpha_2 \vec{\boldsymbol{\mathcal{I}}}_2 + \alpha_3 \vec{\boldsymbol{\mathcal{I}}}_3 + ... + \alpha_k \vec{\boldsymbol{\mathcal{I}}}_k + \vec{\boldsymbol{\epsilon}} \tag{4.3}$$

$$\Leftrightarrow \quad \vec{\mathbf{Y}} = \mathbf{X} \vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\epsilon}} \ , \tag{4.4}$$

sendo as colunas da matriz do modelo $\mathbf{X}$ dadas pelo vector $\vec{\mathbf{1}}_n$ e os vectores das indicatrizes $\vec{\boldsymbol{\mathcal{I}}}_m$. São estas variáveis indicatrizes que desempenham agora o papel das variáveis preditoras $\vec{\mathbf{x}}_j$ numa regressão linear múltipla. A matriz do modelo $\mathbf{X}$ permite identificar as observações de cada nível do factor. O vector dos parâmetros $\vec{\boldsymbol{\beta}}$ tem elementos: μ e os efeitos α_i.

Exemplo Vejamos um pequeno exemplo em que se admite haver um factor com $k=3$ níveis e apenas $n=9$ observações, assim distribuidas pelos níveis: $n_1 = 3$, $n_2 = 4$ e $n_3 = 2$. Admitamos que as observações estão agrupadas e ordenadas pelos respectivos níveis. Tem-se a seguinte equação do modelo:

$$\underbrace{\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \\ Y_{31} \\ Y_{32} \end{bmatrix}}_{=\vec{\mathbf{Y}}} = \underbrace{\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}}_{=\mathbf{X}} \cdot \underbrace{\begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}}_{=\vec{\boldsymbol{\beta}}} + \underbrace{\begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{24} \\ \epsilon_{31} \\ \epsilon_{32} \end{bmatrix}}_{=\vec{\boldsymbol{\epsilon}}} \tag{4.5}$$

4.3.2.2 O problema do excesso de parâmetros

Existe um problema "técnico" com a equação do modelo definida na equação (4.3): *as colunas da matriz do modelo* $\mathbf{X}$ *assim construída são linearmente dependentes*, já que a soma de todas as variáveis indicatrizes

é uma coluna de uns, ou seja a primeira coluna de $\mathbf{X}$ pode ser escrita como uma soma das restantes. Esse facto implica que a matriz $\mathbf{X}^t\mathbf{X}$ não é invertível, pelo que o vector de estimadores não está bem definido. Pode afirmar-se que o modelo tem um *excesso de parâmetros* já que, em última análise o problema resulta de haver $k+1$ parâmetros para estimar, mas apenas k situações experimentais onde se podem estimar.

O problema pode ser resolvido de várias formas, uma vez que é possível impôr várias restrições alternativas que façam desaparecer a dependência linear das colunas de $\mathbf{X}$. Entre as soluções possíveis para este problema encontram-se:

Solução 1: retirar o parâmetro μ do modelo. Esta solução:

- corresponde a retirar a coluna de uns da matriz $\mathbf{X}$;
- cada α_i passa a ser equivalente à média do nível, μ_i.

Mas,

- esta solução não se pode generalizar a situações mais complexas, com mais do que um factor;
- e esta solução é mais difícil de encaixar na teoria já dada do Modelo Linear, uma vez que desapareceria a constante aditiva comum a todas as observações (o parâmetro β_0 dos modelos de regressão linear).

Solução 2: impôr uma restrição do tipo $\sum_{i=1}^{k} \alpha_i = 0$. Repare-se que esta equação corresponde a reduzir o número de parâmetros α_i livres, já que conhecendo o valor de $k-1$ desses parâmetros, o último é obrigado a tomar o valor que assegure a soma nula.

- Esta é a solução clássica, ainda hoje frequente em livros de ANOVA; mas
- é mais difícil de encaixar na teoria geral do Modelo Linear, pelo que não será utilizada.

Solução 3: tomar $\alpha_1 = 0$: *será a solução utilizada.*

- corresponde a excluir do modelo (e de $\mathbf{X}$) a variável indicatriz do primeiro nível;
- permite aproveitar a teoria do Modelo Linear e é generalizável.

Cada solução tem implicações na forma de interpretar os parâmetros. Admitir a terceira opção, ou seja, exigir que $\alpha_1 = 0$, corresponde a excluir a primeira variável indicatriz ($\vec{\boldsymbol{\mathcal{I}}}_1$) na equação 4.3. No exemplo acima, a condição $\alpha_1 = 0$ permite re-escrever a equação (4.5) da seguinte forma:

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \\ Y_{31} \\ Y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{24} \\ \epsilon_{31} \\ \epsilon_{32} \end{bmatrix}$$

Agora, o parâmetro μ é o valor médio das observações do nível $i = 1$ (e será doravante escrito como $\mu = \mu_1$):

$$\begin{array}{rcl} Y_{1j} = \mu + \epsilon_{1j} & \Rightarrow & \mu_1 = E[Y_{1j}] = \mu \quad , \forall\, j = 1, ..., n_1 \\ Y_{2j} = \mu_1 + \alpha_2 + \epsilon_{2j} & \Rightarrow & \mu_2 = E[Y_{2j}] = \mu_1 + \alpha_2 \quad , \forall\, j = 1, ..., n_2 \\ Y_{3j} = \mu_1 + \alpha_3 + \epsilon_{3j} & \Rightarrow & \mu_3 = E[Y_{3j}] = \mu_1 + \alpha_3 \quad , \forall\, j = 1, ..., n_3 \\ \vdots \quad \vdots & & \vdots \\ Y_{kj} = \mu_1 + \alpha_k + \epsilon_{kj} & \Rightarrow & \mu_k = E[Y_{kj}] = \mu_1 + \alpha_k \quad , \forall\, j = 1, ..., n_k \end{array}$$

4.3.2.3 Os efeitos de nível α_i

No modelo para uma ANOVA a um factor, cada α_i ($i > 1$) representa o *acréscimo* que transforma a média do primeiro nível na média do nível i:

$$\begin{array}{rcl} \alpha_1 & = & 0 \\ \alpha_2 & = & \mu_2 - \mu_1 \\ \alpha_3 & = & \mu_3 - \mu_1 \\ \vdots & \vdots & \vdots \\ \alpha_k & = & \mu_k - \mu_1 \end{array}$$

A *igualdade de todas as médias populacionais de nível* μ_i equivale a que todos os efeitos de nível sejam nulos: $\alpha_i = 0\ , \quad \forall\ i$. Esta será a Hipótese Nula que vamos querer testar contra a Hipótese Alternativa de que exista pelo menos um nível i para o qual $\alpha_i \neq 0$, o que equivale a dizer que existem pelo menos dois níveis i e i' tais que as respectivas médias populacionais diferem: $\mu_i \neq \mu_{i'}$.

4.3.2.4 Estimadores dos parâmetros

Consideremos agora os estimadores dos parâmetros μ e α_i ($i = 2, ..., k$) acima referidos. Uma vez que se escreveu a equação do modelo na forma típica de um Modelo Linear, é possível afirmar que o vector com os parâmetros ajustados pelo critério dos Mínimos Quadrados é dado pela fórmula geral já conhecida:

$$\vec{\hat{\boldsymbol{\beta}}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{Y}}\ .$$

No entanto, e como se viu, na ANOVA a um factor o papel das variáveis preditoras é desempenhado pelas variáveis indicatrizes, variáveis que apenas podem tomar os valores 1 e 0. As k colunas da matriz do modelo $\mathbf{X}$ são os vectores $\vec{\mathbf{1}}_n, \vec{\boldsymbol{I}}_2, \vec{\boldsymbol{I}}_3, \ldots, \vec{\boldsymbol{I}}_k$.

Dada esta natureza especial da matriz $\mathbf{X}$, volta a ser possível (tal como na Regressão Linear Simples, mas ao contrário do que acontecia na Regressão Linear Múltipla) ter fórmulas para cada estimador individual de um dos parâmetros. As fórmulas dos parâmetros ajustados, geram *estimadores* dos parâmetros

populacionais que são as *quantidades amostrais análogas* às que os parâmetros definem. Vejamos essas fórmulas. Sendo $\overline{Y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$ a média das n_i observações de Y no nível i, tem-se:

$$\begin{array}{rcll} & \mu_1 & & \longrightarrow \quad \hat{\mu}_1 = \overline{Y}_{1\cdot} \\ \alpha_2 = & \mu_2 - \mu_1 & & \longrightarrow \quad \hat{\alpha}_2 = \overline{Y}_{2\cdot} - \overline{Y}_{1\cdot} \\ \alpha_3 = & \mu_3 - \mu_1 & & \longrightarrow \quad \hat{\alpha}_3 = \overline{Y}_{3\cdot} - \overline{Y}_{1\cdot} \\ & \vdots & & \qquad\qquad\quad \vdots \quad \vdots \quad \vdots \\ \alpha_k = & \mu_k - \mu_1 & & \longrightarrow \quad \hat{\alpha}_k = \overline{Y}_{k\cdot} - \overline{Y}_{1\cdot} \end{array}$$

4.3.2.5 Os valores ajustados $\hat{Y}_{ij}$

Do que foi visto, decorre que qualquer observação tem valor ajustado:

$$\hat{Y}_{ij} \;=\; \hat{\mu}_i \;=\; \hat{\mu}_1 + \hat{\alpha}_i \;=\; \overline{Y}_{i\cdot}\;. \tag{4.6}$$

Ou seja, os valores ajustados $\hat{Y}_{ij}$ são iguais para todas as observações num mesmo nível i do factor, e são dadas pela média amostral das observações nesse nível.

Tal como na Regressão, estes valores ajustados de Y resultam de projectar ortogonalmente o vector $\vec{\mathbf{Y}}$ dos valores observados da variável resposta, sobre o subespaço de $\mathbb{R}^n$ gerado pelas colunas da matriz $\mathbf{X}$: $\hat{\vec{\mathbf{Y}}} = \mathbf{H}\vec{\mathbf{Y}}$. Numa ANOVA a um factor, o subespaço $\mathcal{C}(\mathbf{X})$ tem natureza especial: todos os vectores de $\mathcal{C}(\mathbf{X})$ têm de ter o mesmo valor nas posições correspondentes a observações dum mesmo nível do factor.

$$a_1\,\vec{\mathbf{1}}_n + a_2\vec{\boldsymbol{\mathcal{I}}}_2 + a_3\vec{\boldsymbol{\mathcal{I}}}_3 + ... + a_k\vec{\boldsymbol{\mathcal{I}}}_k = \begin{bmatrix} a_1 \\ ... \\ a_1 \\ \hline a_1+a_2 \\ ... \\ a_1+a_2 \\ \hline a_1+a_3 \\ ... \\ a_1+a_3 \\ \hline (...) \\ \hline a_1+a_k \\ ... \\ a_1+a_k \end{bmatrix}$$

O vector $\hat{\vec{\mathbf{Y}}}$ pertence a $\mathcal{C}(\mathbf{X})$, logo tem esta natureza. No caso do vector $\hat{\vec{\mathbf{Y}}}$, o valor comum às observações de cada nível dado pela *média amostral desse nível.*

4.3.3 O modelo ANOVA a um factor

Para se poder fazer inferência no modelo ANOVA a um factor, admite-se ainda que os erros aleatórios ϵ_{ij} têm as mesmas propriedades que no modelo de regressão linear. O Modelo ANOVA a um Factor

completo fica assim da seguinte forma.

Definição 4.2: Modelo ANOVA a um factor, com k níveis

Seja dada uma variável resposta Y, que será observada de forma independente em k níveis de um factor. Existem n observações, Y_{ij}, das quais n_i estão associadas ao nível i $(i = 1, ..., k)$ do factor. Tem-se:

1. $Y_{ij} = \mu_1 + \alpha_i + \epsilon_{ij}$, $\forall\, i = 1, ..., k$, $\forall\, j = 1, ..., n_i$ (com a restrição $\alpha_1 = 0$).
2. $\epsilon_{ij} \frown \mathcal{N}(0\,,\,\sigma^2)$, $\forall\, i, j$
3. $\{\epsilon_{ij}\}_{i,j}$ v.a.s independentes.

O modelo ANOVA a um factor tem k *parâmetros*: a média de Y no primeiro nível do factor, μ_1, e os $k-1$ acréscimos α_i $(i > 1)$ que geram as médias de cada um dos $k-1$ restantes níveis do factor.

4.3.4 O modelo ANOVA a um factor - notação vectorial

Alternativamente, o modelo ANOVA a um factor pode ser escrito, de forma equivalente, usando notação vectorial. O vector dos parâmetros do modelo é o vector:

$$\vec{\boldsymbol{\beta}} = (\mu_1\,,\,\alpha_2\,,\,\alpha_3\,,\,\cdots\,,\alpha_k)^t\ .$$

Definição 4.3: Modelo ANOVA a um factor - notação vectorial

Seja dada uma variável resposta Y, que será observada de forma independente em k níveis de um factor. O vector $\vec{\mathbf{Y}}$ das n observações verifica:

1. $\vec{\mathbf{Y}} = \mu_1\,\vec{\mathbf{1}}_n + \alpha_2\vec{\boldsymbol{\mathcal{I}}}_2 + \alpha_3\vec{\boldsymbol{\mathcal{I}}}_3 + ... + \alpha_k\vec{\boldsymbol{\mathcal{I}}}_k + \vec{\boldsymbol{\epsilon}} = \mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\epsilon}}$, sendo
   - $\vec{\mathbf{1}}_n$ o vector de n uns; $\vec{\boldsymbol{\mathcal{I}}}_2$, $\vec{\boldsymbol{\mathcal{I}}}_3$, ..., $\vec{\boldsymbol{\mathcal{I}}}_k$ as variáveis indicatrizes dos níveis indicados;
   - $\mathbf{X} = \left[\ \vec{\mathbf{1}}_n \mid \vec{\boldsymbol{\mathcal{I}}}_2 \mid \vec{\boldsymbol{\mathcal{I}}}_3 \mid \cdots \mid \vec{\boldsymbol{\mathcal{I}}}_k\ \right]$ a matriz do modelo;
   - $\vec{\boldsymbol{\beta}} = (\mu_1, \alpha_2, \alpha_3, \cdots, \alpha_k)^t$.
2. O vector dos erros aleatórios tem distribuição $\vec{\boldsymbol{\epsilon}} \frown \mathcal{N}_n(\vec{\mathbf{0}}\,,\,\sigma^2\,\mathbf{I}_n)$, sendo $\mathbf{I}_n$ a matriz identidade $n \times n$.

Trata-se dum modelo análogo a um modelo de Regressão Linear Múltipla, diferindo apenas na natureza das variáveis preditoras, que são aqui variáveis indicatrizes dos níveis 2 a k do factor. Este facto permite aproveitar os resultados inferenciais já estudados. Mas permite igualmente concretizações específicas do contexto ANOVA, para as quais viramos agora a atenção.

4.3.5 O teste F aos efeitos do factor

Já se viu que a hipótese de que nenhum dos níveis do factor afecte a média da variável resposta corresponde à hipótese

$$\begin{aligned} & \alpha_2 = \alpha_3 = ... = \alpha_k = 0 \\ \Leftrightarrow \quad \mu_1 = & \mu_2 = \mu_3 = \cdots = \mu_k \end{aligned}$$

Dado o paralelismo com os modelos de Regressão Linear, esta hipótese corresponde a dizer que todos os coeficientes das "variáveis preditoras" (na ANOVA, as variáveis indicatrizes $\vec{\boldsymbol{I}}_i$) são nulos.

Logo, *é possível testar esta hipótese, através dum teste F de ajustamento global do modelo* (ver Secção 3.9). Trata-se dum caso particular do modelo linear, mas neste contexto há notação e fórmulas específicas associadas a este teste.

4.3.5.1 Notação e graus de liberdade

Numa ANOVA a um factor, utilizaremos SQF em vez de SQR, para indicar a Soma de Quadrados associada à variabilidade explicada pelo Modelo, ou seja, aos efeitos do Factor. A definição desta Soma de Quadrados é idêntica ao visto no contexto da regressão.

Numa ANOVA a um factor, o *número de preditores do modelo* (as variáveis indicatrizes dos níveis $2, 3, ..., k$) é $p = k-1$ e o *número de parâmetros do modelo* é $p+1 = k$. Logo, os graus de liberdade associados a cada Soma de Quadrados são:

SQ	g.l.
SQF	$k-1$
SQRE	$n-k$

Os *Quadrados Médios* continuam a ser os quocientes das Somas de Quadrados a dividir pelos respectivos graus de liberdade:

$$QMF = \frac{SQF}{k-1} \quad ; \quad QMRE = \frac{SQRE}{n-k} \, . \tag{4.7}$$

4.3.5.2 O Teste F

Sendo válido o Modelo de ANOVA a um factor, tem-se então o seguinte *teste F à existência de efeitos do Factor*:

Proposição 4.1: Teste F aos efeitos do factor

Hipóteses: $H_0 : \alpha_i = 0 \, , \quad \forall \, i = 2, ..., k \quad$ vs. $\quad H_1 : \exists i = 2, .., k$ t.q. $\alpha_i \neq 0$.

[FACTOR NÃO AFECTA] vs. [FACTOR AFECTA Y]

Estatística do Teste: $F = \frac{QMF}{QMRE} \frown F_{(k-1,n-k)}$ se H_0 é verdade.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{\alpha(k-1,n-k)}$

A forma da região crítica é indicada na Figura 4.3.

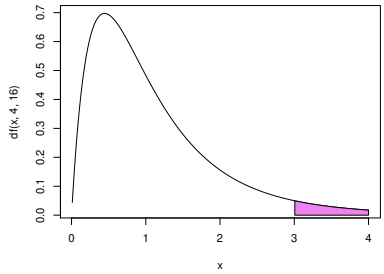


Figura 4.3: Uma região crítica unilateral direita para o teste F à existência dos efeitos do Factor.

No contexto duma ANOVA a um Factor, as Somas de Quadrados e Quadrados Médios têm fórmulas específicas, que permitem o seu cálculo a partir dos indicadores simples (médias e variâncias amostrais) das observações de cada nível do factor. Estas fórmulas são analisadas na subsecção seguinte.

4.3.6 Os resíduos, SQRE e QMRE

Viu-se na equação (4.6) que os valores esperados de Y, numa ANOVA a um factor, são dados pela média amostral das observações de Y nesse factor: $\hat{Y}_{ij} = \hat{\mu}_i = \overline{Y}_{i\cdot}$. Assim, *o resíduo da observação Y_{ij} é dado pela sua diferença em relação à média amostral de nível*:

$$E_{ij} \;=\; Y_{ij} - \hat{Y}_{ij} \;=\; Y_{ij} - \overline{Y}_{i\cdot} \;. \tag{4.8}$$

A *Soma de Quadrados dos Resíduos* é, como sempre, o somatório do quadrado dos resíduos ao longo de todas as n observações. No entanto, a nossa opção por usar uma dupla indexação para referenciar cada observação individual implica que para calcular tais somas ao longo das n observações seja necessário usar um *duplo somatório*:

$$SQRE \;=\; \sum_{i=1}^{k}\sum_{j=1}^{n_i} E_{ij}^2 \;=\; \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i\cdot}\right)^2 \;=\; \sum_{i=1}^{k} (n_i - 1)\, S_i^2 \;, \tag{4.9}$$

onde $S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot})^2$ é a *variância amostral das n_i observações de Y no i-ésimo nível do factor.*

Assim, $SQRE$ mede a *variabilidade no seio dos k níveis*, variabilidade que não é explicada pelo modelo.

O Quadrado Médio Residual é uma *média ponderada* das variâncias amostrais de nível dos Y_{ij}, ou seja, dos S_i^2:

$$QMRE \;=\; \frac{SQRE}{n-k} \;=\; \sum_{i=1}^{k} \frac{n_i - 1}{n-k}\, S_i^2 \;, \tag{4.10}$$

(tendo-se $\sum_{i=1}^{k}(n_i - 1) = n - k$).

4.3.6.1 Fórmulas para SQRE e QMRE em delineamentos equilibrados

As expressões para $SQRE$ e $QMRE$ simplificam ulteriormente no caso de um *delineamento equilibrado*, ou seja, de um delineamento em que todos os níveis do factor têm o mesmo número n_c de observações. De facto, se $n_1 = n_2 = ... = n_k (= n_c)$, tem-se $n = n_c \cdot k$, e:

$$SQRE \quad = \quad (n_c - 1) \sum_{i=1}^{k} S_i^2$$

$$QMRE \quad = \quad \frac{n_c - 1}{n-k} \sum_{i=1}^{k} S_i^2 \quad = \quad \frac{\cancel{n_c - 1}}{k(\cancel{n_c - 1})} \sum_{i=1}^{k} S_i^2 \quad = \quad \frac{1}{k} \sum_{i=1}^{k} S_i^2 \;.$$

Assim, *em delineamentos equilibrados, o Quadrado Médio Residual, QMRE, é a média simples das k variâncias de nível*, nos valores da variável resposta Y.

4.3.7 A Soma de Quadrados e Quadrado Médio associados ao Factor

A Soma de Quadrados associada à Regressão toma, neste contexto, a designação *Soma de Quadrados associada ao Factor* e será representada por SQF. De novo, para somar os quadrados das diferenças entre os valores ajustados de Y e a média global de todos esses valores, será necessário usar duplos somatórios. A média da totalidade das n observações é dada por:

$$\overline{Y}_{..} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij} \;.$$

A Soma de Quadrados associada ao Factor é dada por:

$$\begin{aligned} SQF \;&=\; \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\hat{Y}_{ij} - \overline{Y}_{..}\right)^2 \;=\; \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\overline{Y}_{i.} - \overline{Y}_{..}\right)^2 \\ \Leftrightarrow \qquad SQF \;&=\; \sum_{i=1}^{k} n_i \left(\overline{Y}_{i.} - \overline{Y}_{..}\right)^2 \;. \end{aligned}$$

Assim, a Soma de Quadrados associada ao Factor, SQF, mede *variabilidade entre as médias amostrais de cada nível.*

4.3.7.1 Fórmulas para SQF e QMF em delineamentos equilibrados

No caso de um *delineamento equilibrado*, com $n_1 = n_2 = ... = n_k (= n_c)$, tem-se:

$$SQF \;=\; n_c \sum_{i=1}^{k} (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2 \;=\; n_c(k-1) \cdot S^2_{\overline{Y}_{i\cdot\cdot}} \;,$$

onde $S^2_{\overline{Y}_{i\cdot\cdot}} = \frac{1}{k-1} \sum_{i=1}^{k} (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2$ indica a variância amostral das k médias de nível amostrais. Dividindo SQF pelos seus graus de liberdade $(k-1)$, obtém-se o Quadrado Médio associado ao Factor:

$$QMF \quad = \quad \frac{SQF}{k-1} \quad = \quad n_c \cdot S^2_{\overline{Y}_{i\cdot\cdot}} \;.$$

Assim, *em delineamentos equilibrados, o Quadrado Médio associado aos efeitos do Factor, QMF, é proporcional à variância das k médias de nível* da variável Y.

4.3.8 A relação entre Somas de Quadrados

A relação fundamental entre as três Somas de Quadrados, já estudada no contexto geral de Modelos Lineares, assume no caso da ANOVA a um Factor (mesmo com delineamentos não equilibrados) um significado particular:

$$\begin{array}{ccccc} SQT & = & SQF & + & SQRE \\ \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{\cdot\cdot})^2 & = & \sum_{i=1}^{k} n_i \, (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2 & + & \sum_{i=1}^{k} (n_i - 1) \, S_i^2 \;. \end{array}$$

onde:

$SQT = (n-1)s_y^2$ mede a *variabilidade total* das n observações de Y;

SQF mede a variabilidade entre diferentes níveis do factor (*variabilidade inter-níveis*);

$SQRE$ mede a variabilidade no seio de cada nível - a *variabilidade intra-níveis* que não é explicada pelo factor.

Esta é a origem histórica do nome *Análise da Variância*: a variância de Y é decomposta ("analisada", no seu significado original) em parcelas associadas a diferentes causas. No caso duma ANOVA a um Factor, as causas podem ser apenas o efeito do *factor*, ou outras não explicadas pelo modelo (residuais).

4.3.9 A tabela de síntese da ANOVA a um Factor

É usual resumir toda esta informação numa *tabela-resumo da ANOVA a um Factor*:

Fonte	g.l.	SQ	QM	f_{calc}
Factor	$k-1$	$SQF = \sum_{i=1}^{k} n_i \cdot (\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2$	$QMF = \frac{SQF}{k-1}$	$\frac{QMF}{QMRE}$
Resíduos	$n-k$	$SQRE = \sum_{i=1}^{k} (n_i - 1)\, s_i^2$	$QMRE = \frac{SQRE}{n-k}$	
Total	$n-1$	$SQT = (n-1)\, s_y^2$	–	–

Nota: Quer na coluna das Somas de Quadrados, quer na coluna dos graus de liberdade, o valor relativo à linha Total é dado pela somas das correspondentes parcelas nas linhas do Factor e Residual. Esta relação *não* é válida para a coluna dos Quadrados Médios.

4.3.10 A ANOVA a um Factor no R

No R define-se uma *estrutura de dados específica para variáveis qualitativas (categóricas)*, designada `factor`. Um *factor* é criado pelo comando com o mesmo nome, `factor`, aplicado a um vector contendo os nomes dos vários níveis.

Um vector com os nomes dos níveis é frequentemente criado usando o comando `rep`, que repete os elementos de um vector de nomes, num número de vezes dado por um outro vector de valores numéricos. Exemplifiquemos a criação dum factor correspondente às 150 observaçoes de lírios, e que indica que as primeiras 50 observações da espécie *setosa*, seguidas de 50 observações da espécie *versicolor*, e finalmente 50 observações da espécie *virginica*.

```
> especie <- factor( rep( c("setosa","versicolor","virginica") , c(50,50,50) ) )
> especie
[1] setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa
[9] setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa
[17] setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa
[25] setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa
[33] setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa
[41] setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa
[49] setosa     setosa     versicolor versicolor versicolor versicolor versicolor versicolor
[57] versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor
[65] versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor
[73] versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor
[81] versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor
[89] versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor
[97] versicolor versicolor versicolor versicolor virginica  virginica  virginica  virginica
[105] virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica
[113] virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica
[121] virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica
[129] virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica
[137] virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica
```

```
[145] virginica  virginica  virginica  virginica  virginica  virginica
Levels: setosa versicolor virginica
```

A última linha mostrada acima, começada pela palavra `Levels`, é característica dos objectos da classe `factor`. Nessa linha, listam-se os nomes dos níveis do factor. O comando `summary`, aplicado a um factor, indica o número de observações em cada nível do factor. Por exemplo, no objecto `iris`, a coluna `Species` é um factor (igual ao factor `especie` acima criado), enquanto as restantes são variáveis numéricas. Vejamos como a função `summary` lida com factores:

```
> summary(iris)
 Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

Para efectuar uma ANOVA a um Factor no `R`, os dados devem ser dados numa `data.frame` com *duas colunas*:

1. uma coluna para os valores (numéricos) da variável resposta;
2. outra coluna para o factor (com a indicação dos seus níveis); esta coluna deve ter a classe *factor*.

As fórmulas usadas no `R` para especificar uma ANOVA a um factor são semelhantes às usadas na regressão linear, mas indicando um `factor` como variável preditora. Por exemplo, para efectuar uma ANOVA de larguras das pétalas sobre espécies, nos dados dos $n = 150$ lírios, a fórmula é:

$$\texttt{Petal.Width} \sim \texttt{Species}$$

Embora seja possível usar o comando `lm` para efectuar uma ANOVA (a ANOVA é caso particular do Modelo Linear), existe outro comando que produz informação sob a forma mais tradicional numa ANOVA: o comando `aov`.

O comando `aov` produz uma espécie de mini-tabela resumo da ANOVA (um resultado diferente do resultado do comando `lm`) contendo as Somas de Quadrados e graus de liberdade para os dois tipos de variabilidade: a explicada pelo Factor e a Residual. A função `summary`, quando aplicada ao resultado do comando `aov`, produz o quadro-resumo completo da ANOVA.

Para obter as estimativas dos parâmetros μ_1, α_2, α_3, ..., α_k, pode aplicar-se a função `coef` à ANOVA ajustada.

As médias da variável resposta, global e por nível do factor, podem ser directamente obtidas através da função `model.tables`, com o argumento `type="means"`.

Por omissão, o `R` ordena os níveis de um factor por ordem alfabética, independentemente da ordem pela qual surja a primeira instância de cada nível no conjunto de dados.

Ilustremos a aplicação destes comandos com um exemplo.

Exemplo 4.1: ANOVA das larguras das sépalas sobre as espécies de lírios

Na Secção 4.1 (pg. 153) motivou-se a ANOVA com dois exemplos, no segundo dos quais se consideravam os diagramas de extremos e quartis das larguras das sépalas, por espécie de lírio. Vamos agora verificar se as diferenças entre valores médios de largura de sépalas por espécie se podem considerar significativas.

A ANOVA da largura de sépalas sobre espécies para os lírios invoca-se da seguinte forma:

```
> aov(Sepal.Width ~ Species, data=iris)
Call:
   aov(formula = Sepal.Width ~ Species, data = iris)
Terms:
                Species Residuals
Sum of Squares  11.34493  16.96200
Deg. of Freedom        2       147

Residual standard error: 0.3396877
```

O quadro de síntese completo desta ANOVA obtém-se com o comando `summary`:

```
> iris.aov <- aov(Sepal.Width ~ Species , data=iris)
> summary(iris.aov)
             Df Sum Sq Mean Sq F value Pr(>F)
Species       2  11.35   5.672   49.16 <2e-16 ***
Residuals   147  16.96   0.115
```

Neste caso, rejeita-se claramente a hipótese de que os acréscimos de nível, α_i, sejam todos nulos, pelo que se rejeita a hipótese de larguras médias de sépalas iguais em todas as espécies. Conclusão: o factor (espécie) afecta a variável resposta (largura da sépala).

No exemplo dos lírios, têm-se os seguintes valores estimados dos parâmetros μ, α_2 e α_3:

```
> coef(iris.aov)
(Intercept) Speciesversicolor  Speciesvirginica
      3.428            -0.658            -0.454
```

Assim:

- $\hat{\mu}_1 = 3.428$: é a média amostral de larguras de sépalas *setosa*;

- $\hat{\alpha}_2 = -0.658$: é o acréscimo (como $\hat{\alpha}_2$ é negativo, trata-se dum decréscimo) que, somado à média amostral das *setosa*, dá a média amostral das larguras de sépalas *versicolor*;
- $\hat{\alpha}_3 = -0.454$: é o acréscimo que, somado à média amostral das *setosa*, dá a média amostral das larguras de sépalas *virginica*.

Eis as médias da variável resposta, global e por nível do factor, obtidas com a função `model.tables`:

```
> model.tables(iris.aov , type="means")
Tables of means
Grand mean
3.057333

 Species
Species
    setosa versicolor  virginica
     3.428      2.770      2.974
```

Também é possível estudar uma ANOVA através do comando geral para os modelos lineares, o comando `lm`. Este comando não produz a habitual tabela resumo das ANOVAs. Produz os resultados das listagens já vistas aquando do estudo do Modelo Linear, mas com a importante diferença que, sendo o preditor um factor, os parâmetros do modelo serão os indicados no modelo ANOVA (Definição 4.2). Estes resultados podem ser úteis para a construção de intervalos de confiança ou testes de hipóteses, relativos ao parâmetro μ_1 e aos efeitos do factor, α_i $(i > 1)$. Eis um exemplo, relativo à ANOVA da largura das pétalas sobre as espécies de lírios:

```
> summary(lm(Petal.Width ~ Species , data=iris))
Call: lm(formula = Petal.Width ~ Species, data = iris)
(...)
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.24600    0.02894    8.50 1.96e-14
Speciesversicolor  1.08000    0.04093   26.39  < 2e-16
Speciesvirginica   1.78000    0.04093   43.49  < 2e-16
---
Residual standard error: 0.2047 on 147 degrees of freedom
Multiple R-squared: 0.9289,  Adjusted R-squared: 0.9279
F-statistic:   960 on 2 and 147 DF,  p-value: < 2.2e-16
```

Os valores estimados dos parâmetros ($\hat{\mu} = 0.246$, $\hat{\alpha}_2 = 1.08$ e $\hat{\alpha}_3 = 1.78$) e respectivos erros padrão podem servir para construir intervalos de confiança ou efectuar testes de hipóteses relativos aos valores populacionais dos parâmetros, ou de suas combinações lineares, usando a teoria geral dos modelos lineares estudada na Secção 3.3.

4.3.11 A exploração ulterior de H_1: as comparações múltiplas de Tukey

A Hipótese Nula, no teste F numa ANOVA a 1 Factor, afirma que todos os níveis do factor têm efeito nulo, isto é, que a média da variável resposta Y é igual nos k níveis do Factor:

$$\begin{aligned} & \alpha_2 = \alpha_3 = ... = \alpha_k = 0 \\ \Leftrightarrow \quad \mu_1 = & \mu_2 = \mu_3 = \cdots = \mu_k \end{aligned}$$

A Hipótese Alternativa diz que pelo menos um dos níveis do factor tem uma média μ_i diferente da média μ_1 do primeiro nível, ou seja, que:

$$\begin{aligned} & \exists\, i \quad \text{tal que} \quad \alpha_i \neq 0 \\ \Leftrightarrow \quad & \exists\, i \quad \text{tal que} \quad \mu_1 \neq \mu_i \end{aligned}$$

Um eventual opçao por H_1 indica que nem todas as médias de nível de Y são iguais, mas (a não ser no caso simples de haver apenas dois níveis, ou seja, $k = 2$), não explicita quais pares de médias de nível devem ser considerados diferentes. Assim, a opção por H_1 comporta a necessidade de aprofundar ulteriormente a análise de precisamente quais são os pares de níveis do factor em que as médias de Y se devem considerar diferentes. Mesmo com apenas $k=3$ níveis do factor, a rejeição de H_0 pode dever-se a situações diferentes, nomeadamente:

$$\begin{aligned} \mu_1 = \mu_2 \neq \mu_3 \quad & \text{i.e.,} \quad \alpha_2 = 0 \; ; \; \alpha_3 \neq 0 \\ \mu_1 = \mu_3 \neq \mu_2 \quad & \text{i.e.,} \quad \alpha_3 = 0 \; ; \; \alpha_2 \neq 0 \\ \mu_1 \neq \mu_2 = \mu_3 \quad & \text{i.e.,} \quad \alpha_2 = \alpha_3 \neq 0; \\ \mu_i \text{ todos diferentes} \quad & \text{i.e.,} \quad \alpha_2 \neq \alpha_3 \text{ e } \alpha_2, \alpha_3 \neq 0. \end{aligned}$$

Como optar entre estas diferentes alternativas? Uma possibilidade consiste em efectuar testes aos valores dos efeitos α_i, com base na teoria já estudada anteriormente (recorde-se que um modelo ANOVA é um modelo linear). Assim, por exemplo, um teste à Hipótese $H_0 : \alpha_2 = 0$ dir-nos-á se $\mu_1 = \mu_2$ é admissível. Por outro lado, um teste à hipótese $H_0 : \alpha_2 = \alpha_3$ dir-nos-á se é admissível considerar que $\mu_2 = \mu_3$.

No entanto, esta abordagem comporta um problema. Exige $k-1$ testes a cada α_i individual, mais $\binom{k-1}{2}$ testes a comparar pares de diferentes α_is. Assim, o número total de testes a efectuar é $\frac{k(k-1)}{2}$, e se o número de níveis k não fôr muito pequeno, o número de testes *t-Student* a efectuar será grande. Um grande número de testes t, cada um dos quais realizados ao nível de significância α, não permite controlar do nivel de significância *global* para o conjunto de todos os testes. Por exemplo, num factor com $k = 10$ níveis, haverá 45 testes de hipóteses a realizar, e a probabilidade de *no conjunto de todos* estes testes de hipóteses se ter alguma vez incorrectamente rejeitado a hipótese da igualdade de médias μ_i sob comparação deixa de ser apenas o valor de α com que cada teste individual foi realizado.

Procurando ultrapassar este problema, foram desenvolvidas abordagens alternativas, cuja preocupação é avaliar a igualdade de todos os $\frac{k(k-1)}{2}$ pares de médias ($\mu_i = \mu_j$, para todos os pares (i, j) de níveis), mas podendo no final afirmar que a probabilidade de se ter rejeitado uma tal hipótese num qualquer par, quando em todos ela era verdadeira, é um valor de significância α controlável pelo experimentador.

Alternativamente, pode optar-se por construir um conjunto de intervalos de confiança para todas as diferenças de médias populacionais, $\mu_i - \mu_j$, com um grau *global* de confiança $1-\alpha$ em como qualquer todas as diferenças de médias populacionais pertencem aos respectivos intervalos. Esse tipo de abordagens designam-se de *comparações múltiplas*. A mais utilizada de entre elas será estudada na próxima Subsecção.

4.3.11.1 Intervalos de confiança e testes de Tukey

As comparações múltiplas de Tukey baseiam-se num resultado inferencial geral, mas cuja aplicação é de particular utilidade neste contexto.

Proposição 4.2: Distribuição de Tukey

Sejam $\{W_i\}_{i=1}^k$ variáveis aleatórias independentes, com distribuição Normal, e com os mesmos parâmetros: $W_i \frown \mathcal{N}(\mu_W, \sigma_W^2)$, $\forall\, i = 1, ..., k$.

- Seja $R_W = \max_i W_i - \min_i W_i$ a *amplitude total amostral*.
- Seja S_W^2 um estimador da variância comum σ_W^2, tal que $\frac{\nu\, S_W^2}{\sigma_W^2} \frown \chi_\nu^2$.
- Sejam S_w e R_w variáveis aleatórias independentes.

Então, *a amplitude Studentizada, $\frac{R_W}{S_W}$, tem uma distribuição de Tukey*, que depende de *dois parâmetros*: k (dimensão da amostra) e ν (graus de liberdade da χ^2 associada à estimação de σ_W^2).

Notas:

1. Omite-se a demonstração.
2. Repare-se que este é um resultado relativamente geral, que permite conhecer a distribuição de probabilidades na amostragem duma amplitude Studentizada associada a qualquer amostra aleatória $(W_1, W_2, ..., W_k)$ duma (mesma) população Normal. Assim, é um resultado que pode ser aplicado em contextos diferentes do contexto da ANOVA, em que agora estamos interessados.

A utilidade da distribuição de Tukey para o problema da comparação múltipla de médias resulta do facto de poder ser utilizado tomando as variáveis aleatórias Normais W_i como sendo a diferença entre as médias amostral e populacional para cada nível do factor. De facto, numa ANOVA a um factor, tem-se:

$$\overline{Y}_{i\cdot} \frown \mathcal{N}\left(\mu_i\,,\,\frac{\sigma^2}{n_i}\right) \qquad \Leftrightarrow \qquad \overline{Y}_{i\cdot} - \mu_i \frown \mathcal{N}\left(0\,,\,\frac{\sigma^2}{n_i}\right)$$

Se o delineamento fôr equilibrado, isto é, $n_1 = n_2 = ... = n_k\ (= n_c)$, as k diferenças $\overline{Y}_{i\cdot} - \mu_i$ terão a mesma distribuição $\mathcal{N}\left(0\,,\,\frac{\sigma^2}{n_c}\right)$, e podem ser consideradas as variáveis W_i da Proposição 4.2. Para construir a amplitude Studentizada destas variáveis, é necessário um estimador da variância comum σ^2/n_c. Uma vez que n_c é conhecido, apenas será preciso um estimador da variância dos erros aleatórios σ^2, estimador esse de que já dispomos: o Quadrado Médio Residual. Assim, o estimador de $\frac{\sigma^2}{n_c}$ é dado por $S_W^2 = \frac{QMRE}{n_c}$.

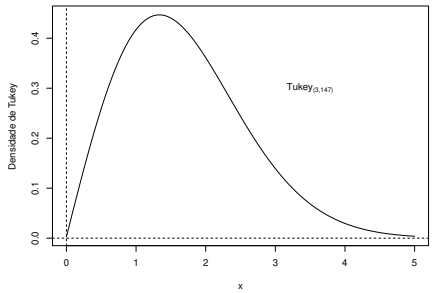


Figura 4.4: Densidade duma distribuição Tukey com parâmetros $k=3$ e $\nu=147$.

Este estimador S_W^2 verifica a exigência da Proposição 4.2 de que $\frac{\nu\, S_W^2}{\sigma^2}$ tenha distribuição χ^2_ν, uma vez que sabemos (Proposição 3.10) que $\frac{SQRE}{\sigma^2} \frown \chi^2_{n-(p+1)}$, logo (e tendo em conta que $p+1=k$ no contexto da ANOVA a um factor) tem-se $\frac{\frac{(n-k)\,QMRE}{n_c}}{\frac{\sigma^2}{n_c}} = \frac{(n-k)S_W^2}{\sigma_W^2} \frown \chi^2_{n-k}$. A restante condição da Proposição 4.2 verifica-se, pelo que a seguinte quantidade tem a distribuição de Tukey, com parâmetros k e $n-k$ (sendo estes últimos os graus de liberdade de $QMRE$)

$$\frac{R_W}{S_W} \;=\; \frac{\max\limits_i(\overline{Y}_{i\cdot}-\mu_i) - \min\limits_j(\overline{Y}_{j\cdot}-\mu_j)}{\sqrt{\frac{QMRE}{n_c}}} \;\frown\; Tukey_{(k,n-k)} \tag{4.11}$$

O quociente $\frac{R_W}{S_W}$ não pode ser negativo, por definição das quantidades envolvidas na fracção.

Intervalos de Confiança para $\mu_i - \mu_j$. Seja $q_{\alpha\,(k,n-k)}$ o valor que numa distribuição de Tukey com parâmetros k e $n-k$, deixa à direita uma região de probabilidade α. Então, por definição:

$$P\left[\frac{R_W}{S_W} \;<\; q_{\alpha\,(k,n-k)}\right] \;=\; 1-\alpha\ .$$

Logo, *um intervalo de confiança* (unilateral) a $(1-\alpha)\times 100\%$ *para a amplitude total* R_W é dado por:

$$R_W \quad < \quad q_{\alpha\,(k,n-k)} \cdot S_W \quad = \quad q_{\alpha\,(k,n-k)} \cdot \sqrt{\frac{QMRE}{n_c}}\ .$$

Mas a amplitude

$$R_W \;=\; \max_i\,(\overline{Y}_{i\cdot}-\mu_i) - \min_j\,(\overline{Y}_{j\cdot}-\mu_j)$$

é a *maior de todas as diferenças*, para qualquer $i,j=1,...,k$, do tipo:

$$\left|(\overline{Y}_{i\cdot}-\mu_i)-(\overline{Y}_{j\cdot}-\mu_j)\right| \;=\; \left|(\overline{Y}_{i\cdot}-\overline{Y}_{j\cdot})-(\mu_i-\mu_j)\right|,$$

Logo, *para todos os pares de níveis i e j, tem-se, com grau de confiança global* $(1-\alpha)\times 100\%$,

$$\left|(\overline{y}_{i\cdot}-\overline{y}_{j\cdot})-(\mu_i-\mu_j)\right| \quad \leq \quad R_W \quad < \quad q_{\alpha\,(k,n-k)}\cdot\sqrt{\tfrac{QMRE}{n_c}}$$

$$\Leftrightarrow \quad -q_{\alpha\,(k,n-k)}\sqrt{\tfrac{QMRE}{n_c}} \;<\; (\mu_i-\mu_j)-(\overline{y}_{i\cdot}-\overline{y}_{j\cdot}) \;<\; q_{\alpha\,(k,n-k)}\sqrt{\tfrac{QMRE}{n_c}}$$

Assim, tem-se $(1-\alpha)\times 100\%$ de confiança em como *todas* as diferenças de médias de nível $\mu_i-\mu_j$ estão em intervalos da forma:

$$\left]\quad (\overline{y}_{i\cdot}-\overline{y}_{j\cdot})-q_{\alpha\,(k,n-k)}\sqrt{\tfrac{QMRE}{n_c}}\quad,\quad (\overline{y}_{i\cdot}-\overline{y}_{j\cdot})+q_{\alpha\,(k,n-k)}\sqrt{\tfrac{QMRE}{n_c}}\quad\right[\tag{4.12}$$

Se para qualquer par (i,j) de níveis, o intervalo correspondente *não* contém o valor zero, então $\mu_i=\mu_j$ *não* é admissível.

Testes de Hipóteses para $\mu_i-\mu_j=0\ ,\forall\, i,j$. Alternativamente, a partir do resultado da equação (4.11) é possível testar a Hipótese Nula de que *todas* as diferenças de pares de médias de nível, $\mu_i-\mu_j$, sejam nulas, em cujo caso

$$\left|\overline{Y}_{i\cdot}-\overline{Y}_{j\cdot}\right| \quad < \quad q_{\alpha\,(k,n-k)}\cdot\sqrt{\tfrac{QMRE}{n_c}}\ , \quad \forall\ i,j \tag{4.13}$$

com probabilidade $(1-\alpha)$. Assim, qualquer diferença de médias amostrais de nível, $\overline{Y}_{i\cdot}-\overline{Y}_{j\cdot}$, cujo módulo exceda o limiar

$$q_{\alpha\,(k,n-k)}\cdot\sqrt{\tfrac{QMRE}{n_c}} \tag{4.14}$$

indica que, para esse par de níveis i,j, se deve considerar $\mu_i\neq\mu_j$. O nível (global) de significância de todas estas comparações é α, ou seja, há probabilidade α de se concluir que $\mu_i\neq\mu_j$ para *algum* par i,j, se em *todos* os casos $\mu_i=\mu_j$.

Sintetizando o que foi dito acima, temos o *Teste de Tukey às diferenças de médias de nível*, numa ANOVA a um Factor:

Proposição 4.3: Testes de Tukey às diferenças de pares de médias

Hipóteses: $H_0:\mu_i=\mu_j\ ,\ \forall\ i,j$ vs. $H_1:\exists i,j$ t.q. $\mu_i\neq\mu_j$.
[FACTOR NÃO AFECTA] vs. [FACTOR AFECTA Y]

Estatística do Teste: $\frac{R_W}{S_W}\ \frown\ Tukey_{(k,n-k)}$ se H_0.

Nível de significância do teste: α

Regra de rejeição: Para qualquer par (i,j), rejeita-se $\mu_i=\mu_j$ se $\left|\overline{Y}_{i\cdot}-\overline{Y}_{j\cdot}\right|>q_{\alpha\,(k,n-k)}\sqrt{\tfrac{QMRE}{n_c}}$

A natureza da estatística $\frac{R}{S}$ permite não apenas rejeitar H_0 globalmente, como identificar o(s) par(es) (i,j) responsáveis pela rejeição (a diferença das correspondentes médias amostrais excede o termo de comparação), *permitindo assim conclusões sobre diferenças significativas em cada par de médias.*

4.3.12 Comparações múltiplas de médias no R

As comparações múltiplas de médias de nível, com base no resultado de Tukey, podem ser facilmente efectuadas no R. Por um lado, os valores da função distribuição cumulativa e os quantis $q_{\alpha\,(k,n-k)}$ duma distribuição de Tukey podem ser calculados através das funções `ptukey` e `qtukey`, respectivamente. Para se obter o *termo de comparação* nos testes de hipóteses a que $\mu_i - \mu_j = 0$, o quantil de ordem $1-\alpha$ na distribuição de Tukey é obtido indicando os valores numéricos de α, k e $n-k$ no comando `qtukey`:

$>$ `qtukey`(1-α, k, $n-k$)

O valor de $\sqrt{QMRE}$ é produzido pelo comando `aov`, sob a designação "*Residual standard error*", e o valor de $QMRE$ consta da tabela de síntese da ANOVA, produzido pelo comando `summary` aplicado a uma ANOVA ajustada.

Pelo seu lado, o comando `TukeyHSD` calcula os intervalos de confiança a $(1-\alpha)\times 100\%$ para as diferenças de médias.

Exemplo 4.2: Intervalos de confiança de Tukey nos dados dos lírios

Voltando à ANOVA das larguras das sépalas sobre as espécies de lírios, tem-se:

```
> TukeyHSD(iris.aov)
Tukey multiple comparisons of means
95% family-wise confidence level
$Species
                      diff         lwr        upr     p adj
versicolor-setosa    -0.658 -0.81885528 -0.4971447 0.0000000
virginica-setosa     -0.454 -0.61485528 -0.2931447 0.0000000
virginica-versicolor  0.204  0.04314472  0.3648553 0.0087802
```

Assim, o intervalo a 95% de confiança (o nível de confiança por omissão) para $\mu_2 - \mu_1$ (`versicolor-setosa`) é $]\ -0.8189\ ,\ -0.4971\ [$. Uma vez que não inclui o valor zero, não é admissível que $\mu_2 - \mu_1 = 0$, ou seja, rejeita-se a igualdade de μ_2 e μ_1.

Neste exemplo, nenhum dos intervalos de confiança para diferenças de pares de médias de nível inclui o valor zero, pelo que consideramos que $\mu_i \neq \mu_j$, para qualquer $i \neq j$, ou seja, todas as médias de espécie são diferentes entre si.

O *valor de prova* indicado (`p adj`) deve ser interpretado como sendo o valor de α para o qual cada diferença de médias, $\overline{y}_{i.} - \overline{y}_{j.}$, seria, pela primeira vez, considerada não significativa. Assim, a diferença de médias amostrais para as espécies *virginica* e *versicolor* apenas seria considerada não significativa para um nível de significância $\alpha = 0.00878$. Ou seja, apenas intervalos de Tukey a mais de $(1-\alpha)\times 100\% = 99.122\%$ de confiança, para essa diferença de médias, conteriam o valor zero.

Representação gráfica das comparações múltiplas. O R disponibiliza ainda um auxiliar gráfico para visualizar as comparações das médias de nível, através da função `plot`, aplicada ao resultado da

função `TukeyHSD`. O resultado de aplicar esse comando ao exemplo dos lírios é mostrado na Figura 4.5.

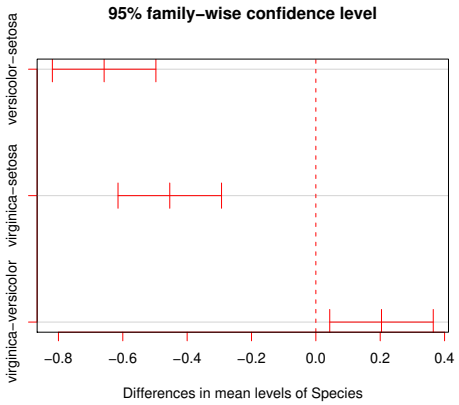


Figura 4.5: Intervalos (a 95%) de confiança de Tukey para as diferenças das médias populacionais de nível, $\mu_i - \mu_j$ das larguras das sépalas (dados dos lírios). Nenhum dos intervalos de confiança contém o valor zero (assinalado pela linha vertical a tracejado), pelo que se rejeita a hipótese de igualdade para todos os pares de médias.

4.3.12.1 Delineamentos não equilibrados

Nota: Quando o delineamento da ANOVA a um Factor não é equilibrado (isto é, quando existe um número diferente de observações nos vários níveis do factor), os testes ou intervalos de confiança de Tukey agora enunciados não são, em rigor, válidos. Mas, para delineamentos em que o desequilíbrio no número de observações em cada nível não seja muito acentuado, é possível um resultado aproximado, baseado no número médio de observações por nível, que a função `TukeyHSD` do `R` incorpora.

4.3.13 Análise de Resíduos e diagnósticos na ANOVA a 1 Factor

Em geral, a validade dos pressupostos do modelo estuda-se de forma idêntica ao que foi visto na Regressão Linear. Mas há *algumas particularidades* do contexto específico da ANOVA, que importa sublinhar.

Nos gráficos de resíduos usuais (e_{ij}) contra valores ajustados de y ($\hat{y}_{ij}$) para uma ANOVA a um factor, os resíduos aparecem sempre empilhados em k colunas. Tal particularidade *não* é um padrão indicativo de problemas, e resulta do facto (visto na equação 4.6) de qualquer valor ajustado $\hat{y}_{ij} = \overline{y}_{i.}$ ser igual para observações num mesmo nível do factor. Assim, todos os pontos correspondentes a um mesmo nível do factor terão, no gráfico referido, a mesma coordenada no eixo horizontal, criando as k "colunas" referidas. Como foi indicado, este padrão *não* corresponde a qualquer violação dos pressupostos do modelo e poderia também surgir no contexto duma Regressão Linear, caso houvesse repetições de observações com valores idênticos nas variáveis preditoras. Este tipo de gráfico continua a ser útil, porque permite inspeccionar visualmente indícios de diferença na variabilidade dos resíduos em cada nível do factor. Caso essa seja

uma característica evidente no gráfico, suscita dúvidas sobre a validade do pressuposto de homogeneidade de variâncias dos erros aleatórios.

A Figura 4.6 ilustra este padrão de resíduos na ANOVA a 1 Factor correspondente aos dados dos lírios, com a variável resposta `Sepal.Width` e o factor preditor `Species`. A semelhança na dispersão dos pontos em cada coluna indicia que não há problemas com o pressuposto de homogeneidade de variâncias dos erros aleatórios.

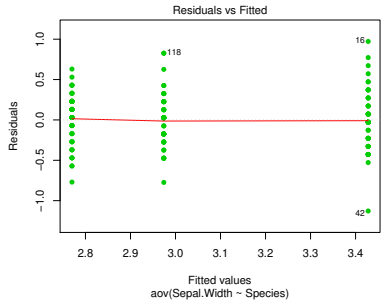


Figura 4.6: Gráfico de resíduos (usuais) contra valores ajustados de Y, na ANOVA da largura das sépalas sobre o factor espécies. O padrão das colunas é expectável e não indicia qualquer violação de pressupostos do modelo. A dispersão semelhante dos pontos em cada uma das colunas é compatível com o pressuposto de variâncias homogéneas dos erros aleatórios.

Outra particularidade do contexto da ANOVA a um factor diz respeito aos indicadores de diagnóstico, e mais concretamente aos efeitos alavanca. De facto, *todas as observações dum mesmo nível i do factor terão idêntico efeito alavanca, igual a* $\frac{1}{n_i}$. No caso de delineamentos equilibrados, os efeitos alavanca de *todas* as observações serão assim iguais. Logo, os efeitos alavanca deixam de ser um diagnóstico útil no contexto das ANOVAs. Como consequência evidente a partir da expressão final na Definição 3.17, as distâncias de Cook são também menos úteis, reflectindo sobretudo o valor de cada resíduo estandardizado R_{ij}. No entanto, o limiar $D_{ij} > 0.5$ continua a ser útil na identificação de observações influentes.

Em compensação, o contexto específico da ANOVA a um factor, com as suas n_i repetições em cada um dos k níveis do factor, permite fazer algo que não é, em geral, possível nas regressões lineares: testar formalmente se as variâncias dos erros aleatórios diferem entre os níveis do factor. Entre os testes usualmente propostos para este fim encontram-se os testes de Bartlett ou de Levene, que no entanto não são matéria para avaliação nesta disciplina.

4.3.13.1 Testando a homogeneidade de variâncias: teste de Bartlett *

(*) A matéria desta Subsecção não é avaliada

O *Teste de Bartlett* confronta as hipóteses

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

vs.

$$H_1 : \exists i, i' \quad \text{t.q.} \quad \sigma_i^2 \neq \sigma_{i'}^2 \ ,$$

sendo σ_i^2 a *variância comum dos erros aleatórios* ϵ_{ij} *do nível* i *do factor*. A hipótese nula indica que o pressuposto de igualdade de todas as variâncias de erros aleatórios é admissível, enquanto que a hipótese alternativa viola esse pressuposto, uma vez que afirma que em níveis diferentes do factor, os erros aleatórios serão diferentes. Repare-se que em nenhum momento se considera a possibilidade de diferentes erros aleatórios no seio dum mesmo nível do factor terem variâncias diferentes (o que também violaria o pressuposto do modelo sobre a existência de homogeneidade de variâncias). Nem seria possível testar essa possibilidade (tal como não existe um teste de hipóteses para a homogeneidade de variâncias dos erros aleatórios numa regressão linear). Mas a repetição de observações num mesmo nível permite efectuar o teste agora indicado, o que é já um contributo importante, tanto mais que a causa mais plausível para heterogeneidade nas variâncias dos erros aleatórios corresponde à diferença de situações experimentais associada aos níveis do factor.

A estatística do teste de Bartlett compara as médias aritmética e geométrica das k variâncias amostrais de nível de Y, $S_i^2 = \frac{1}{n_i-1} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i\cdot}\right)^2$. Vejamos de seguida a descrição dos passos num teste de Bartlett à homogeneidade de variâncias dos erros, numa ANOVA a um factor.

Hipóteses: $H_0: \ \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$ vs. $H_1: \ \exists i, i' \quad \text{t.q.} \quad \sigma_i^2 \neq \sigma_{i'}^2$
[Variâncias homogéneas] vs. [Variâncias heterogéneas]

Estatística do Teste:

$$K^2 = \frac{(n-k)\ln QMRE - \sum_{i=1}^{k}(n_i-1)\ln S_i^2}{C} \sim \chi^2_{k-1},$$

onde $C = 1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k}\frac{1}{n_i-1} - \frac{1}{n-k}\right]$.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): (Unilateral direita) Rejeitar H_0 se $K^2_{calc} > \chi^2_{\alpha(k-1)}$.

Duas precauções são necessárias na utilização do teste de Bartlett:

- O teste de Bartlett é *fortemente dependente da Normalidade das observações subjacentes.*
- *A distribuição* χ^2 *é apenas assintótica.* Uma regra comum para a admissibilidade desta distribuição assintótica é considerar que *o teste apenas deve ser usado caso* $n_i \geq 5$, $\forall\, i = 1, .., k$.

O Teste de Bartlett no R. No R, o teste de Bartlett numa ANOVA a um factor é invocado pelo comando `bartlett.test`, tendo por argumento uma fórmula análoga à usada no comando `aov` para indicar a variável resposta e o factor. Assim, por exemplo, para pedir o teste de Bartlett na ANOVA de `Sepal.Width` sobre `Species` (nos dados dos lírios), utilizar-se-ia o seguinte comando:

```
> bartlett.test(Sepal.Width ~ Species, data=iris)

        Bartlett test of homogeneity of variances
```

```
data:  Sepal.Width by Species
Bartlett's K-squared = 2.0911, df = 2, p-value = 0.3515
```

O valor calculado da estatística é $K^2_{calc} = 2.0911$, e o correspondente *p-value* numa distribuição χ^2 com $k-1=2$ graus de liberdade é $p = 0.3515$. Assim, o teste de Bartlett indica a não rejeição de H_0, ou seja, é admissível a hipótese de igualdade nas variâncias em cada nível do factor. Essa conclusão permite admitir a homogeneidade das variâncias dos erros aleatórios.

4.3.13.2 Advertências finais

Eventuais violações aos pressupostos do modelo ANOVA a um Factor não têm sempre igual gravidade. Estudos efectuados (desde logo, na obra clássica de Scheffé [6]), permitem alguns comentários gerais a este respeito.

- O teste F da ANOVA e as comparações múltiplas de Tukey são *relativamente robustos a desvios à hipótese de normalidade*.
- As *violações ao pressuposto de variâncias homogéneas são em geral menos graves no caso de delineamentos equilibrados*, mas podem ser graves em delineamentos não equilibrados.
- A *falta de independência entre erros aleatórios é a violação mais grave dos pressupostos* e deve ser evitada, o que é em geral possível com um delineamento experimental adequado.

Refira-se ainda que na *formulação clássica do modelo ANOVA a um Factor*, e a partir da equação-base do modelo

$$Y_{ij} \;=\; \mu + \alpha_i + \epsilon_{ij} \;, \qquad \forall\, i,j$$

em vez de se impor a restrição $\alpha_1 = 0$ (ver Subsecção 4.3.2.2), *impõe-se a restrição alternativa* $\sum_{i=1}^{k} \alpha_i = 0$.

Esta condição alternativa:

- Também resolve o problema de excesso de parâmetros no modelo, uma vez que, além do parâmetro μ, apenas deixa $k-1$ parâmetros livres α_i (o conhecimento de $k-1$ parâmetros α_i implica o conhecimento do restante, já que têm de somar zero).
- Muda a forma de interpretar os parâmetros: μ é agora uma espécie de *média geral de* Y (para qualquer nível do factor) e α_i será o desvio da média do nível i em relação a essa média geral.
- Muda as fórmulas dos estimadores dos parâmetros.
- Mas *não* muda o resultado do teste F à existência de efeitos do factor, nem os resultados dos testes de comparaçãoes múltiplas de Tukey.

Assim, a formulação da restrição alternativa não afecta os resultados globais da ANOVA a um Factor, mas apenas os aspectos ligados à interpretação, estimação e inferência dos parâmetros do modelo. Recorde-se que a nossa abordagem (ou seja, a restrição $\alpha_1 = 0$), foi introduzida pois permite aproveitar directamente os resultados do Modelo Linear, já estudados no Capítulo 3. Essa vantagem ultrapassa qualquer desvantagem que se possa apontar.

4.4 Delineamentos e Unidades experimentais

No *delineamento das experiências* para posterior análise através duma ANOVA (ou regressão linear), as n observações da variável resposta correspondem a n diferentes *unidades experimentais* (indivíduos, parcelas de terreno, locais, etc.).

4.4.1 Os princípios gerais da casualização e repetição

Princípios gerais, já conhecidos, na selecção destas unidades experimentais são:

1. a *casualização*, ou seja, a *aleatoriedade* na recolha de observações. No contexto duma ANOVA a um factor, essa casualização significa muitas vezes a casualização da escolha de diferentes tratamentos (níveis do factor) a aplicar a diferentes unidades experimentais, como sejam parcelas de terreno, plantas, ou quantidades padronizadas de um produto (vinho, queijo, etc.). Essa casualização é fundamental para:
   - se poder *trabalhar com a Teoria de Probabilidades*; e
   - se *evitar enviesamentos* (mesmo inconscientes) que muitas vezes estão associados a escolhas de unidades experimentais às quais associar um determinado nível do factor (como seja, um tratamento).
2. Outro princípio importante do delineamento, já conhecido, é a *repetição* de observações *independentes*, que é necessária para:
   - *estimar a variabilidade associada à estimação (erros padrões).*
   - minorar o impacto duma eventual observação atípica.

   Convém a este respeito distinguir entre *repetições* e *pseudo-repetições*. Exemplifiquemos a diferença entre estes dois conceitos, considerando um estudo sobre frutos do tomateiro. Do ponto de vista experimental não é a mesma coisa:
   - seleccionar frutos *dum mesmo tomateiro*; ou
   - seleccionar frutos de *tomateiros diferentes*.

   As repetições querem-se *independentes*. Mas as características genotípicas, fenotípicas e ambientais, são idênticas (ou muito semelhantes) para frutos *duma mesma planta*. Pelo que nesse caso, estamos perante *pseudo-repetições*, que embora repliquem valores diferentes duma qualquer característica, dificilmente podem ser consideradas repetições independentes. Já no caso de frutos de tomateiros diferentes, as características genotípicas e fenotípicas (e talvez também as ambientais) são diferentes, havendo mais verdade no pressuposto de que se trata de observações independentes.

   Pseudo-repetições podem ter utilidade. Substituindo cada grupo de pseudo-repetições por uma única observação média pode-se *diminuir a variabilidade entre diferentes observações* (ou seja, entre as médias de grupos de pseudo-repetições), uma vez que esbate o impacto de uma ou outra observação atípica. E estas médias de pseudo-repetições já podem ser consideradas independentes umas das outras. A redução da variabilidade entre observações independentes torna a inferência mais precisa.

4.4.2 Heterogeneidade nas unidades experimentais

Num Modelo Linear, toda a variabilidade nas unidades experimentais *não atribuível aos preditores* é considerada *variação inexplicada* e contemplada nos *erros aleatórios*. Assim, heterogeneidade não controlada nas unidades experimentais contribui para *aumentar a variância σ^2 dos erros aleatórios*, e por conseguinte, *aumentar o valor do seu estimador, $QMRE$*. Por sua vez, aumentar $QMRE$ significa, nos testes F, *diminuir o valor calculado da estatística F*, afastando-a da região crítica. Assim, a *heterogeneidade não controlada* nas unidades experimentais,

- **numa ANOVA**, contribui para *esconder a presença de eventuais efeitos do(s) factor(es)* (veja-se a Subsecção 4.5.9 para um exemplo).
- **numa Regressão Linear**, contribui para *piorar a qualidade de ajustamento do modelo*, diminuindo o seu Coeficiente de Determinação.

Na prática, é quase sempre impossível tornar as unidades experimentais totalmente homogéneas: a natural variabilidade de plantas, animais, terrenos, localidades geográficas, células, etc. significa que existe variabilidade não controlável entre unidades experimentais. Mesmo que seja possível ter unidades experimentais (quase) homogéneas, isso tem uma *consequência indesejável: restringir a validade dos resultados ao tipo de unidades experimentais com as características utilizadas na experiência.*

Caso se saiba que existe um factor de variabilidade importante nas unidades experimentais, a melhor forma de controlar os seus efeitos consiste em *contemplar a existência desse factor de variabilidade no delineamento e no modelo*, de forma a *filtrar os seus efeitos.*

4.4.2.1 Um exemplo

Imagine-se que se pretende analisar o rendimento de 5 diferentes variedades de trigo (variável resposta). Os rendimentos são também afectados pelos *tipo de solos* usados, que podem fazer os rendimentos diferir substancialmente, mesmo numa única variedade. Assim, admita-se que em duas parcelas de terrenos diferentes e onde se usaram variedades diferentes, se observam rendimentos muito diferentes. Fica em aberto a dúvida se as causas dessa variabilidade se devem ao factor variedade que se pretende estudar, ou se resultam de terem sido observadas em terrenos de tipo diferente, havendo *confusão* entre essas duas possíveis causas de variabilidade.

Nem sempre é possível ter terrenos homogéneos numa experiência. Mesmo que seja possível, terá um efeito indesejável, uma vez que as consequências que se poderiam retirar duma experiência assim concebida iriam limitar a validade dos resultados a um único tipo de solos.

Pode lidar-se com esta situação organizando (delineando) uma experiência em que se admite a existência dum factor terrenos. Assim, admita-se que estamos interessados em estudar os rendimentos em *quatro terrenos com diferentes tipos de solos.* Cada terreno pode ser dividido em cinco parcelas viáveis para o trigo, o que cria, ao todo, 20 *unidades experimentais.* Em vez de repartir aleatoriamente as 5 variedades pelas 20 parcelas, é preferível forçar cada tipo de terreno a conter uma parcela com cada variedade. Apenas dentro dos terrenos haverá casualização.

A situação agoras descrita é ilustrada da seguinte forma:

Terreno 1	Var.1	Var.3	Var.4	Var.5	Var.2
Terreno 2	Var.4	Var.3	Var.5	Var.1	Var.2
Terreno 3	Var.2	Var.4	Var.1	Var.3	Var.5
Terreno 4	Var.5	Var.2	Var.4	Var.1	Var.3

A associação de variedades de terreno às 20 unidades experimentais não foi feita de forma totalmente aleatória. Houve uma *restrição à casualização total*: dentro de cada terreno há casualização, mas obriga-se cada terreno a ter uma parcela associada a cada nível do factor *variedade*.

O delineamento agora exemplificado é um caso particular de um *delineamento factorial a dois factores*, sendo um dos factores a variedade de trigo e o outro o tipo de solos. Este tipo de delineamentos será estudado na Secção seguinte.

A existência de mais do que um factor pode resultar de:

- pretender-se estudar eventuais efeitos de mais do que um factor sobre a variável resposta; ou
- a tentativa de controlar a variabilidade experimental, como no exemplo desta Subsecção 4.4.2.1.

Historicamente, a segunda situação ficou associada à designação **blocos**, e na primeira fala-se apenas em factores, mas são *situações análogas*.

4.5 Delineamento factorial a 2 factores: modelo sem interacção

Quando existem dois ou mais factores, um delineamento experimental pode contemplar diferentes tipos de relações entre esses factores. Nesta disciplina apenas estudaremos delineamentos com dois factores, e veremos dois diferentes tipos de relações entre eles: os chamados *delineamento factoriais* e os *delineamentos hierarquizados*.

Definição 4.4: Delineamento factorial (a 2 factores)

Um *delineamento factorial a dois factores* é um delineamento com observações da variável resposta *em todas as possíveis combinações de níveis de cada factor*.

É habitual representar um delineamento factorial através duma esquematização em grelha, como indicado na Figura 4.1. Sublinhe-se que essa representação *não* descreve a disposição no terreno duma eventual experiência, sendo apenas uma representação esquemática das relaçoes entre os níveis de cada factor.

		Factor B				
	Níveis	B_1	B_2	B_3	...	B_b
	A_1	× × ×	× × ×	× × ×	...	× × ×
	A_2	× × ×	× × ×	× × ×	...	× × ×
Factor A	A_3	× × ×	× × ×	× × ×	...	× × ×
	⋮	⋮	⋮	⋮	⋱	⋮
	A_a	× × ×	× × ×	× × ×	...	× × ×

Tabela 4.1: Representação esquemática dum delineamento factorial a dois factores, um Factor A com a níveis, e um Factor B com b níveis. Para o delineamento ser factorial tem de haver observações da variável resposta Y em todas as situações experimentais (células da tabela), resultantes de cruzar todos os níveis de um e outro factor. Nesta representação sugere-se a existência de $n_c = 3$ observações em todas as ab células, em cujo caso o delineamento é *equilibrado*.

4.5.1 Notação e terminologia

Fixemos alguma notação relativa aos delineamentos factoriais a dois factores. Admita-se a existência de:

- uma *variável resposta* Y;
- um *Factor A*, com a *níveis*;
- Um *Factor B*, com b *níveis*;
- n *observações*, com pelo menos uma em cada uma das ab situações experimentais criadas pela combinação de cada um dos a níveis do factor A com cada um dos b níveis do Factor B.

Cada cruzamento dum nível dum Factor com um nível doutro Factor correspondem a uma diferente *situação experimental*, ou *célula*. Num delineamento factorial a dois factores haverá assim ab diferentes situações experimentais. Cada situação experimental (célula) pode ser identificada por dois índices (i, j), onde i indica o nível do factor A (a linha da grelha, na representação da Figura 4.1) e j indica o nível do factor B (coluna da grelha).

O número de observações na célula (i, j) é representado por n_{ij}. Tem-se $\sum_{i=1}^{a} \sum_{j=1}^{b} n_{ij} = n$.

Definição 4.5: Delineamento factorial equilibrado

Se o número de observações fôr igual em todas as células, ou seja, se $n_{ij} = n_c$, para todo i, j (sendo n_c o *número comum* de observações em cada célula), falamos num *delineamento equilibrado*.

Cada observação da variável resposta Y será agora identificada através de *três índices* (Y_{ijk}), onde:

- o primeiro índice, i, indica o *nível i do Factor A*;

- o segundo índice, j, indica o *nível j do Factor B*;
- o terceiro índice, k, indica a *repetição k, no seio da célula* (i, j).

Nos delineamentos a um factor, estudámos um único modelo ANOVA. Diferentemente, para delineamentos factoriais a dois factores, consideramos *dois diferentes modelos* ANOVA (*two-way ANOVA* em inglês).

4.5.2 A equação do Modelo

Um *primeiro modelo* prevê a existência de dois diferentes tipos de efeitos condicionando os valores de Y: os efeitos associados aos níveis do Factor A, e os efeitos associados aos níveis do Factor B.

Neste primeiro modelo, admite-se que o valor esperado de cada observação é da forma:

$$E[Y_{ijk}] \quad = \quad \mu_{ij} \quad = \quad \mu + \alpha_i + \beta_j \quad , \quad \forall\, i, j, k \ .$$

onde μ, α_i e β_j são constantes.

O parâmetro μ é comum a todas as observações. Cada parâmetro α_i funciona como um acréscimo que pode diferir entre níveis do Factor A, e é designado o *efeito do nível i do factor A*. Cada parâmetro β_j funciona como um acréscimo que pode diferir entre níveis do Factor B, e é designado o *efeito do nível j do factor B*. Este tipo de modelos, em que os efeitos α_i e β_j se admitem constantes são também conhecidos por *modelos de efeitos fixos*.

Tal como em anteriores modelos lineares, admite-se que a variação de Y_{ijk} em torno do seu valor médio é aleatória, e é representada por uma parcela aditiva designada *erro aleatório*. Cada erro aleatório tem a mesma tripla indexação que a observação a que corresponde, ou seja, o erro aleatório associado à observação Y_{ijk} é representado por ϵ_{ijk}. Assim, a equação de base para cada observação de Y é da forma:

$$Y_{ijk} \ = \ \mu + \alpha_i + \beta_j + \epsilon_{ijk} \ , \tag{4.15}$$

exigindo-se $E[\epsilon_{ijk}] = 0$, a fim de garantir que o valor médio de Y_{ijk} seja $E[Y_{ijk}] = \mu_{ij} = \mu + \alpha_i + \beta_j$.

4.5.3 A equação-base em notação vectorial

A equação de base do modelo ANOVA a dois factores (sem interacção) pode ser vista como um caso particular de Modelo Linear utilizando *variáveis indicatrizes de pertença a cada nível do Factor A e a cada nível do Factor B*. A ideia já foi estudada em pormenor no modelo ANOVA a um único Factor. Será aqui discutida considerando já a equação do modelo escrita na forma vectorial. Seja

$\vec{\mathbf{Y}}$ o vector n-dimensional com a totalidade das observações da variável resposta;

$\vec{\mathbf{1}}_n$ o vector de n uns;

$\vec{\boldsymbol{\mathcal{I}}}_{A_i}$ o vector da *variável indicatriz de pertença ao nível i do Factor A*;

$\vec{\boldsymbol{\mathcal{I}}}_{B_j}$ o vector da *variável indicatriz de pertença ao nível j do Factor B*;

$\vec{\boldsymbol{\epsilon}}$ o *vector dos n erros aleatórios.*

Se se admitem efeitos para *todos* os níveis de ambos os factores, a equação de base em notação vectorial será:

$$\vec{\mathbf{Y}} = \mu\vec{\mathbf{1}}_n + \alpha_1\vec{\boldsymbol{\mathcal{I}}}_{A_1} + \alpha_2\vec{\boldsymbol{\mathcal{I}}}_{A_2} + ... + \alpha_a\vec{\boldsymbol{\mathcal{I}}}_{A_a} + \beta_1\vec{\boldsymbol{\mathcal{I}}}_{B_1} + \beta_2\vec{\boldsymbol{\mathcal{I}}}_{B_2} + ... + \beta_b\vec{\boldsymbol{\mathcal{I}}}_{B_b} + \vec{\boldsymbol{\epsilon}}$$

A matriz do modelo $\mathbf{X}$ definida com base nesta equação teria uma primeira coluna de uns, seguida de a colunas com as indicatrizes de todos os níveis do Factor A, e finalmente b indicatrizes de todos os níveis do Factor B, como ilustrado de seguida:

$$\mathbf{X} = \left[\begin{array}{c|cccc|cccc}
1 & 1 & 0 & ... & 0 & 1 & 0 & ... & 0 \\
1 & 1 & 0 & ... & 0 & 1 & 0 & ... & 0 \\
1 & 1 & 0 & ... & 0 & 0 & 1 & ... & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 1 & 0 & ... & 0 & 0 & 0 & ... & 1 \\
1 & 1 & 0 & ... & 0 & 0 & 0 & ... & 1 \\
-- & -- & -- & -- & -- & -- & -- & -- & -- \\
1 & 0 & 1 & ... & 0 & 1 & 0 & ... & 0 \\
1 & 0 & 1 & ... & 0 & 1 & 0 & ... & 0 \\
1 & 0 & 1 & ... & 0 & 1 & 0 & ... & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & 1 & ... & 0 & 0 & 0 & ... & 1 \\
1 & 0 & 1 & ... & 0 & 0 & 0 & ... & 1 \\
-- & -- & -- & -- & -- & -- & -- & -- & -- \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
-- & -- & -- & -- & -- & -- & -- & -- & -- \\
1 & 0 & 0 & ... & 1 & 1 & 0 & ... & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & 0 & ... & 1 & 0 & 0 & ... & 1 \\
1 & 0 & 0 & ... & 1 & 0 & 0 & ... & 1 \\
\uparrow & \uparrow & \uparrow & & \uparrow & \uparrow & \uparrow & & \uparrow \\
\vec{\mathbf{1}}_n & \vec{\boldsymbol{\mathcal{I}}}_{A_1} & \vec{\boldsymbol{\mathcal{I}}}_{A_2} & ... & \vec{\boldsymbol{\mathcal{I}}}_{A_a} & \vec{\boldsymbol{\mathcal{I}}}_{B_1} & \vec{\boldsymbol{\mathcal{I}}}_{B_2} & ... & \vec{\boldsymbol{\mathcal{I}}}_{B_b}
\end{array}\right]$$

Uma tal matriz tem dependências lineares por duas diferentes razões:

- a soma das indicatrizes $\vec{\boldsymbol{\mathcal{I}}}_{A_i}$ do Factor A é igual à coluna dos uns, $\vec{\mathbf{1}}_n$;
- a soma das indicatrizes $\vec{\boldsymbol{\mathcal{I}}}_{B_j}$ do Factor B também dá a coluna dos uns, $\vec{\mathbf{1}}_n$.

Será necessário introduzir restrições aos parâmetros, não podendo estimar-se parâmetros α_i e β_j para *todos* os níveis de cada Factor. Mas, ao contrário do que acontecia no delineamento a um único factor, há agora *duas* causas de dependência linear, sendo necessárias *duas restrições.* A exclusão da coluna de n uns (ou seja, do parâmetro μ que lhe está associado) já não será agora suficiente para tornar as colunas

de $\mathbf{X}$ linearmente independentes, uma vez que a soma das a indicatrizes do Factor A (que é o vector dos n uns) continuaria a ser igual à soma das indicatrizes dos b níveis do Factor B.

As duas restrições, necessárias para eliminar as duas dependências lineares referidas, serão introduzidas estendendo a ideia já usada aquando do estudo do modelo ANOVA a um único Factor. Doravante, admitimos que são *excluídas da equação do modelo (4.15) as parcelas associadas ao primeiro nível de cada Factor*, isto é, impõem-se as restrições:

$$\alpha_1 = 0 \qquad \text{e} \qquad \beta_1 = 0 \; .$$

A equação de base vectorial do modelo ANOVA a 2 Factores, sem interacção, fica assim:

$$\vec{\mathbf{Y}} \;=\; \mu\vec{\mathbf{1}}_n \;+\; \alpha_2\vec{\boldsymbol{\mathcal{I}}}_{A_2} \;+\; \ldots \;+\; \alpha_a\vec{\boldsymbol{\mathcal{I}}}_{A_a} \;+\; \beta_2\vec{\boldsymbol{\mathcal{I}}}_{B_2} \;+\; \ldots \;+\; \beta_b\vec{\boldsymbol{\mathcal{I}}}_{B_b} + \vec{\boldsymbol{\epsilon}} \tag{4.16}$$

Esta opção corresponde a *excluir as colunas* $\vec{\boldsymbol{\mathcal{I}}}_{A_1}$ *e* $\vec{\boldsymbol{\mathcal{I}}}_{B_1}$ *da matriz* $\mathbf{X}$, que fica com o seguinte aspecto:

$$\mathbf{X} = \left[\begin{array}{c|ccc|ccc}
1 & 0 & \ldots & 0 & 0 & \ldots & 0 \\
1 & 0 & \ldots & 0 & 0 & \ldots & 0 \\
1 & 0 & \ldots & 0 & 1 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & \ldots & 0 & 0 & \ldots & 1 \\
1 & 0 & \ldots & 0 & 0 & \ldots & 1 \\
\hline
1 & 1 & \ldots & 0 & 0 & \ldots & 0 \\
1 & 1 & \ldots & 0 & 0 & \ldots & 0 \\
1 & 1 & \ldots & 0 & 0 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
1 & 1 & \ldots & 0 & 0 & \ldots & 1 \\
1 & 1 & \ldots & 0 & 0 & \ldots & 1 \\
\hline
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\hline
1 & 0 & \ldots & 1 & 0 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & \ldots & 1 & 0 & \ldots & 1 \\
1 & 0 & \ldots & 1 & 0 & \ldots & 1 \\
\hline
\uparrow & \uparrow & & \uparrow & \uparrow & & \uparrow \\
\vec{\mathbf{1}}_n & \vec{\boldsymbol{\mathcal{I}}}_{A_2} & \ldots & \vec{\boldsymbol{\mathcal{I}}}_{A_a} & \vec{\boldsymbol{\mathcal{I}}}_{B_2} & \ldots & \vec{\boldsymbol{\mathcal{I}}}_{B_b}
\end{array}\right]$$

Com estas restrições, o parâmetro μ é o valor esperado das observações na primeira célula: $E[Y_{11k}] = \mu_{11}$. Por isso, daqui em diante esse parâmetro será escrito com a notação μ_{11}.

4.5.4 O Modelo ANOVA a dois Factores, sem interacção

Juntando os pressupostos sobre os erros aleatórios, necessários à inferência, obtém-se o *Modelo ANOVA a dois factores, sem interacção.*

Definição 4.6: Modelo ANOVA a dois factores, sem interacção

Admitimos um delineamento factorial a dois factores, o Factor A com a níveis e o Factor B com b níveis. Admite-se que existem n observações, Y_{ijk}, n_{ij} das quais associadas à célula (i,j) $(i=1,..,a;\ j=1,..,b)$. Tem-se:

1. $Y_{ijk} \ = \ \mu_{11} + \alpha_i + \beta_j + \epsilon_{ijk}$, $(k=1,...,n_{ij})$ com $\alpha_1=0$ e $\beta_1=0$.
2. $\epsilon_{ijk} \ \frown \ \mathcal{N}(0\,,\,\sigma^2), \quad \forall\, i,j,k.$
3. $\{\epsilon_{ijk}\}_{i,j,k}$ é um conjunto de variáveis aleatórias independentes.

O modelo ANOVA a dois Factores, sem interacção, tem um total de $a+b-1$ *parâmetros* desconhecidos:

- o parâmetro μ_{11};
- os $a-1$ efeitos de nível do Factor A, α_i $(i>1)$; e
- os $b-1$ efeitos de nível do Factor B, β_j $(j>1)$.

Mais adiante interpretar-se-á o significado dos efeitos de nível α_i e β_j, de uma forma mais precisa.

4.5.5 O modelo ANOVA a dois factores, sem interacção - notação vectorial

Alternativamente, este modelo ANOVA sem efeitos de interacção, pode ser escrito usando notação vectorial.

Definição 4.7: Modelo ANOVA a dois factores sem interacção - notação vectorial

Seja dada uma variável resposta Y, com n observações Y_{ijk} nas ab células definidas pelo cruzamento factorial de dois factores, A e B.

1. O vector $\vec{\mathbf{Y}}$ das n observações Y_{ijk} verifica:

$$\begin{aligned} \vec{\mathbf{Y}} &= \mu_{11}\,\vec{\mathbf{1}}_n + \alpha_2\vec{\boldsymbol{\mathcal{I}}}_{A_2} + ... + \alpha_a\vec{\boldsymbol{\mathcal{I}}}_{A_a} + \beta_2\vec{\boldsymbol{\mathcal{I}}}_{B_2} + ... + \beta_b\vec{\boldsymbol{\mathcal{I}}}_{B_b} + \vec{\boldsymbol{\epsilon}} \\ &= \mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\epsilon}} \quad , \end{aligned}$$

   sendo

   - $\vec{\mathbf{1}}_n$ o vector de n uns; $\vec{\boldsymbol{\mathcal{I}}}_{A_i}$ indicatriz do nível i nível do Factor A; e $\vec{\boldsymbol{\mathcal{I}}}_{B_j}$ indicatriz do

nível j do Factor B;

- $\mathbf{X} = \left[\vec{\mathbf{1}}_n \mid \vec{\boldsymbol{\mathcal{I}}}_{A_2} \mid \cdots \mid \vec{\boldsymbol{\mathcal{I}}}_{A_a} \mid \vec{\boldsymbol{\mathcal{I}}}_{B_2} \mid \cdots \mid \vec{\boldsymbol{\mathcal{I}}}_{B_b} \right]$ a matriz do modelo;
- $\vec{\boldsymbol{\beta}} = (\mu_1, \alpha_2, \cdots, \alpha_k, \beta_2, \cdots, \beta_b)^t$.

2. O vector dos erros aleatórios tem distribuição $\vec{\boldsymbol{\epsilon}} \frown \mathcal{N}_n(\vec{\mathbf{0}}, \sigma^2 \mathbf{I}_n)$.

4.5.6 Os dois testes F

Um teste de ajustamento global do modelo tem como hipótese nula que *todos* os efeitos, quer do factor A, quer do Factor B, sejam simultaneamente nulos, mas não distingue entre os efeitos de cada factor. Uma vez que se admitiu a existência de dois diferentes factores, será mais útil *testar separadamente a existência dos efeitos de cada factor*.

Assim, serão necessários dois testes, para duas diferentes Hipóteses Nulas:

- Teste I: $H_0 : \alpha_i = 0\,, \quad \forall i = 2, ..., a \quad$ vs. $\quad H_1 : \exists\, i$ tal que $\alpha_i \neq 0$;
- Teste II: $H_0 : \beta_j = 0\,, \quad \forall j = 2, ..., b \quad$ vs. $\quad H_1 : \exists\, j$ tal que $\beta_j \neq 0$.

4.5.6.1 O teste F aos efeitos do Factor B

Como se viu, o modelo do ANOVA a 2 Factores, sem interacção, tem equação de base (4.16) vectorial dada por:

$$\vec{\mathbf{Y}} \;=\; \mu\vec{\mathbf{1}}_n \;+\; \alpha_2\vec{\boldsymbol{\mathcal{I}}}_{A_2} \;+\; ... \;+\; \alpha_a\vec{\boldsymbol{\mathcal{I}}}_{A_a} \;+\; \beta_2\vec{\boldsymbol{\mathcal{I}}}_{B_2} \;+\; ... \;+\; \beta_b\vec{\boldsymbol{\mathcal{I}}}_{B_b} + \vec{\boldsymbol{\epsilon}}\,.$$

O facto de ser um Modelo Linear permite aplicar a teoria já conhecida para este tipo de modelos, para testar as hipóteses

$$H_0 : \beta_j = 0\,, \quad \forall j = 2, ..., b \qquad \text{vs.} \qquad H_1 : \exists\, j \quad \text{tal que } \beta_j \neq 0\,.$$

Uma vez que se trata duma hipótese Nula da igualdade a zero dum conjunto de parâmetros (os parâmetros β_j) que multiplicam variáveis preditoras (as variáveis indicatrizes $\vec{\boldsymbol{\mathcal{I}}}_{B_j}$), a ferramenta necessária é um *teste F parcial* (Ver a Subsecção 3.10.1). Concretamente, ter-se-á de comparar o *modelo completo*, de equação

$$\vec{\mathbf{Y}} \;=\; \mu\vec{\mathbf{1}}_n \;+\; \alpha_2\vec{\boldsymbol{\mathcal{I}}}_{A_2} \;+\; ... \;+\; \alpha_a\vec{\boldsymbol{\mathcal{I}}}_{A_a} \;+\; \beta_2\vec{\boldsymbol{\mathcal{I}}}_{B_2} \;+\; ... \;+\; \beta_b\vec{\boldsymbol{\mathcal{I}}}_{B_b} + \vec{\boldsymbol{\epsilon}}\,,$$

com o submodelo correspondente a tomar todos os $\beta_j = 0$, e que é o modelo a um único Factor, A:

$$\vec{\mathbf{Y}} \;=\; \mu\vec{\mathbf{1}}_n \;+\; \alpha_2\vec{\boldsymbol{\mathcal{I}}}_{A_2} \;+\; ... \;+\; \alpha_a\vec{\boldsymbol{\mathcal{I}}}_{A_a} \;+\; \vec{\boldsymbol{\epsilon}}\,.$$

Podemos também escrever as equações dos modelos sem a notação vectorial. Designaremos o modelo completo por Modelo M_{A+B}:

$$\text{(Modelo } M_{A+B}) \qquad Y_{ijk} \;=\; \mu_{11} + \alpha_i + \beta_j + \epsilon_{ijk}\,,$$

e o *submodelo* por Modelo M_A:

$$\text{(Modelo } M_A) \qquad Y_{ijk} \;=\; \mu_{11} + \alpha_i + \epsilon_{ijk}\,.$$

Registe-se que o submodelo M_A é um *modelo ANOVA a 1 Factor* (só com o factor A).

Os passos completos para a realização deste teste F parcial envolvem:

- O ajustamento do modelo completo M_{A+B} e do submodelo M_A, com base nas respectivas matrizes do modelo $\mathbf{X}$ que definem as duas matrizes de projecção ortogonal $\mathbf{H}$, e por conseguinte, os valores ajustados $\hat{\vec{\mathbf{Y}}}=\mathbf{H}\vec{\mathbf{Y}}$ por cada um dos modelos.
- Obter as respectivas Somas de Quadrados Residuais, $SQRE_{A+B}$ e $SQRE_A$, dados pela soma dos quadrados dos elementos dos vectores de resíduos $\vec{\mathbf{E}}=\vec{\mathbf{Y}}-\hat{\vec{\mathbf{Y}}}$ em cada modelo.
- Efectuar o teste F parcial indicado. Repare-se que, neste caso, a diferença do número de parcelas do modelo completo (M_{A+B}) e Submodelo (M_A) é dado pelo número de parâmetros associados aos efeitos do Factor B, ou seja (e após a restrição $\beta_1 = 0$), $b-1$. Analogamente, os graus de liberdade associados à Soma de Quadrados Residual é, como em qualquer Modelo Linear, o número de observações menos o número de parâmetros do modelo, ou seja $n-(a+b-1)$. A diferença nas Somas de Quadrados do Submodelo e do Modelo a dois factores, que aparece no numerador do numerador da estatística, ou seja $SQRE_A - SQRE_{A+B}$ passa a designar-se a Soma de Quadrados associada aos efeitos do Factor B, SQB:

  $$SQB \;=\; SQRE_A - SQRE_{A+B}\ .$$

  Assim, a estatística de teste será da forma:

  $$\text{(Teste aos Efeitos Factor B)} \qquad F \;=\; \frac{\overbrace{\frac{SQRE_A - SQRE_{A+B}}{b-1}}^{=SQB}}{\frac{SQRE_{A+B}}{n-(a+b-1)}} \;=\; \frac{QMB}{QMRE} \tag{4.17}$$

  definindo o Quadrado Médio associado aos efeitos do Factor B,

  $$QMB \;=\; \frac{SQB}{b-1} \;=\; \frac{SQRE_A - SQRE_{A+B}}{b-1} \tag{4.18}$$

  O $QMRE$ no denominador refere-se ao Quadrado Médio Residual *do Modelo completo*, M_{A+B}.
- Tratando-se dum teste F parcial, esta estatística F tem distribuição $F_{[\,b-1\,,\,n-(a+b-1)\,]}$ sob a Hipótese Nula de igualdade do Modelo M_{A+B} e o Submodelo M_A, ou seja, no caso de todos os efeitos de nível do factor B serem nulos ($H_0 : \beta_j = 0,\ \forall\, j$).

Coleccionando os passos do teste, tem-se o **Teste aos Efeitos do Factor B**.

Proposição 4.4: Teste a efeitos do Factor B no modelo sem efeitos de interacção

Hipóteses: $H_0 : \beta_j = 0\,, \quad \forall\, j = 2, ..., b$ vs. $H_1 : \exists j = 2, .., b$ tal que $\beta_j \neq 0$.
[Factor B NÃO AFECTA Y] vs. [Factor B AFECTA Y]

Estatística do Teste: $F \;=\; \frac{QMB}{QMRE} \;\frown\; F_{(\,b-1\,,\,n-(a+b-1)\,)}$ se H_0.

Nível de significância do teste: α

Região Crítica (Unilateral direita): Rejeitar H_0 se $F_{calc} > f_{\alpha(b-1,n-(a+b-1))}$.

4.5.6.2 O teste F aos efeitos do Factor A

Consideremos também um *teste aos efeitos do Factor A*. Este teste já não será exactamente um teste F parcial, embora a sua estatística de teste tenha também uma distribuição F caso seja verdade a hipótese nula. Neste teste, considera-se o ajustamento do Modelo apenas ao Factor A, o Modelo M_A referido na Subsecção anterior, e a respectiva Soma de Quadrados do Factor, SQF_A será agora designada a Soma de Quadrados associada ao Factor A, SQA. Define-se o Quadrado Médio associado aos efeitos do Factor A, de forma identica ao QMF do Modelo só com o Factor A, ou seja, dividindo $SQA = SQF_A$ por $a-1$ graus de liberdade (que são número de parâmetros α_i, correspondentes aos efeitos de nível do Factor A, após a introdução da restrição $\alpha_1 = 0$). Este Quadrado Médio do Factor A (que designaremos QMA) também será comparado com o Quadrado Médio Residual do Modelo completo a dois Factores (sem interacção) original (M_{A+B}). Concretamente, definem-se:

- $SQA = SQF_A$, a *Soma de Quadrados do Factor no Modelo* M_A;
- $QMA = \frac{SQA}{a-1}$, o *Quadrado Médio do Factor no Modelo* M_A;
- As Somas de Quadrados e Quadrado Médio Residuais do modelo completo a dois Factores (sem interacção), $SQRE_{A+B}$ e $QMRE = \frac{SQRE_{A+B}}{n-(a+b-1)}$.

É possível provar que, caso todos os efeitos do Factor A no Modelo M_{A+B} sejam nulos ($\alpha_i = 0, \forall i = 2, ..., a$), a estatística

$$F = \frac{QMA}{QMRE} = \frac{\frac{SQA}{a-1}}{\frac{SQRE_{A+B}}{n-(a+b-1)}} \tag{4.19}$$

tem distribuição $F_{(a-1,n-(a+b-1))}$.

Assim, sendo válido o Modelo de ANOVA a dois factores, sem interacção, tem-se o seguinte **Teste F aos efeitos do Factor A**.

Proposição 4.5: Teste a efeitos do Factor A no modelo sem efeitos de interacção

Hipóteses: $H_0 : \alpha_i = 0\ ,\quad \forall i = 2, ..., a$ vs. $H_1 : \exists\ i = 2, .., a$ tal que $\alpha_i \neq 0$.
[Factor A NÃO AFECTA Y] vs. [Factor A AFECTA Y]

Estatística do Teste: $F = \frac{QMA}{QMRE} \frown F_{(a-1,n-(a+b-1))}$ se H_0 verdade.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{\alpha(a-1,n-(a+b-1))}$.

Nos dois testes as regiões críticas são unilaterais direitas, como indicado na Figura 4.7.

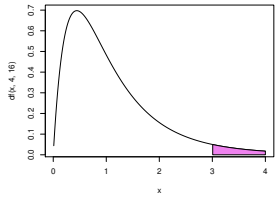


Figura 4.7: Regiões críticas para os dois testes aos efeitos dos factores num delineamento factorial, num modelo ANOVA sem efeitos de interacção.

4.5.7 A decomposição de SQT

Tendo em conta as Somas de Quadrados acima definidas, tem-se uma nova decomposição da Soma de Quadrados Total, $SQT = (n-1)\, s_y^2$. De facto, recorde-se as definições das Somas de Quadrados associadas aos efeitos de cada Factor:

$$SQB = SQRE_A - SQRE_{A+B} \tag{4.20}$$

$$SQA = SQF_A = SQT - SQRE_A \tag{4.21}$$

Somando estas Somas de Quadrados à Soma de Quadrados Residual do Modelo a dois Factores (sem interacção), $SQRE_{A+B}$, obtém-se:

$$SQRE_{A+B} + SQB + SQA = \cancel{SQRE_{A+B}} + (\cancel{SQRE_A} - \cancel{SQRE_{A+B}}) + (SQT - \cancel{SQRE_A}) \ ,$$

ou seja,

$$SQT = SQA + SQB + SQRE_{A+B} \tag{4.22}$$

que é uma *nova decomposição de SQT*, agora em três parcelas, associadas ao facto de haver agora dois factores com efeitos previstos no modelo, e ainda a variabilidade residual.

4.5.8 ANOVA a dois Factores sem interacção no `R`

Para efectuar uma ANOVA a dois Factores (sem interacção) no `R`, convém organizar os dados numa *data frame* com três colunas:

1. uma para os valores (numéricos) da variável resposta;
2. outra para o factor A (com a indicação dos seus níveis);
3. outra para o factor B (com a indicação dos seus níveis).

A fórmula utilizada no `R` para indicar uma ANOVA a dois Factores, sem interacção, é semelhante à usada numa Regressão Linear com dois preditores, devendo o nome dos dois factores (digamos `fA` e `fB`) ser separado pelo símbolo '+':

```
y  ~  fA + fB
```

Tal como numa ANOVA a um Factor, desde que os preditores tenham sido definidos como objecto de classe `factor`, o comando `aov` procede à construção das variáveis indicatrizes necessárias, que serão colocadas nas colunas da matriz do modelo, $\mathbf{X}$.

4.5.9 Um exemplo

Vejamos agora um exemplo clássico[1] dum modelo a dois factores, sem efeitos de interacção, e que também ilustra como ignorar heterogeneidade nas unidades experimentais pode esconder a existência de efeitos dum factor.

Exemplo 4.3: Os dados do rendimento de cevada

O *rendimento* de cinco variedades de cevada (*manchuria, svansota, velvet, trebi* e *peatland*) foi registado em seis diferentes localidades. Em cada localidade foi semeada uma e uma só parcela com cada variedade (havendo casualização das parcelas associadas às variedades, em cada localidade). Foi ajustada uma ANOVA com a variável resposta rendimento (`Y1`), e os efeitos dos dois Factores, *variedade* (`Var`) e *localidade* (`Loc`), que produziu a seguinte tabela de síntese.

```
> summary(aov(Y1 ~ Var + Loc, data=immer))
            Df  Sum Sq Mean Sq F value    Pr(>F)
Var          4  2756.6   689.2  4.2309   0.01214 *
Loc          5 17829.8  3566.0 21.8923 1.751e-07 ***
Residuals   20  3257.7   162.9
```

Num teste aos efeitos de localidade, há uma clara rejeição de H_0, enquanto que no teste aos efeitos de variedade, há rejeição de H_0 ao nível de significância $\alpha=0.05$, mas não ao nível $\alpha=0.01$. Assim, há alguma indicação de efeitos significativos entre variedades, e muita entre localidades.

Vale a pena comparar estes resultados com o resultado de, aos mesmos dados, ajustar um modelo com apenas o factor variedade (o Modelo M_A acima referido), e tratando todas as parcelas (quer da mesma localidade, quer de localidades diferentes) como se fossem repetições. Este modelo corresponde a ignorar eventuais efeitos de localidade (de bloco). Como se pode constatar pelos resultados abaixo indicados, nesse modelo os efeitos de variedade são considerados *não* significativos (para qualquer α razoável).

```
> summary(aov(Y1 ~ Var, data=immer))
            Df  Sum Sq Mean Sq F value Pr(>F)
Var          4  2756.6   689.2   0.817 0.5264
Residuals   25 21087.6   843.5
```

[1] Immer, Hayes e LeRoy Powers, Statistical adaptation of barley varietal adaptation, Journal of the American Society for Agronomy, 26, 403-419, 1934. Os dados estão disponíveis no módulo `MASS` do `R`, numa *data frame* de nome `immer`.

Uma análise detalhada das duas tabelas de resumo das ANOVAs mostra a razão de ser deste resultado qualitativamente diferente. É visível que a Soma de Quadrados associada aos efeitos do factor *variedade* e respectivos graus de liberdade são iguais nas duas tabelas. Assim tinha de ser, pela definição de $SQA = SQF_A$ usada no modelo a dois factores. Uma vez que a Soma de Quadrados Total também é igual nos dois casos (já que não depende do modelo ajustado, mas apenas da variâncias dos valores observados da variável resposta, que são os mesmos nos dois casos), torna-se evidente que tem de ter-se $SQRE_A = SQRE_{A+B} + SQB$ (numericamente, e a menos de arredondamentos, $21087.6 = 17829.8 + 3257.7$), o que confirma numericamente a fórmula dada na equação (4.20). Assim, ignorar os eventuais efeitos do Factor B (`Loc`) equivale a torná-los efeitos não explicados pelo modelo, logo efeitos que irão inflacionar a Soma de Quadrados Residual e (na medida em que esse aumento é considerável) também o Quadrado Médio Residual, que passa de 162.9 no modelo com os dois factores, para 843.5 no modelo apenas com o factor `Var`. Esta inflação do Quadrado Médio Residual implica que um valor idêntico de $QMA = 689.2$ deixa de ser considerado significativo (ao nível $\alpha = 0.05$) e passa a ser claramente não significativo. A ilação geral é a que atrás se referiu: uma grande variabilidade inexplicada tende a mascarar a importância de eventuais efeitos de um factor.

4.5.10 Uma decomposição alternativa de SQT

Um aspecto importante para o qual é necessário chamar a atenção diz respeito ao facto de ser possível trocar o papel dos factores A e B (factores que são, na realidade, arbitrários) na discussão anterior. Essa troca de papéis levaria a definir as Somas de Quadrados de cada factor de forma diferente.

Designando por M_B o modelo ANOVA a um factor, mas apenas com o factor que temos chamado B, é possível considerar que um teste aos efeitos do Factor A corresponderia a um teste F parcial comparando os modelos M_{A+B} e M_B. Construindo uma estatística de teste F para os efeitos do Factor B de forma análoga ao que foi feito na Subsecção 4.5.6.2 (mas trocando o papel dos factores A e B) resultaria nas seguintes definições para as Somas de Quadrados associadas a cada Factor:

$$\begin{aligned} SQB' &= SQF_B = SQT - SQRE_B \\ SQA' &= SQRE_B - SQRE_{A+B} \ . \end{aligned}$$

Continua a ser verdade que SQT se pode decompor na forma

$$SQT = SQA' + SQB' + SQRE_{A+B} \ ,$$

embora as definições de SQA e SQB sejam agora diferentes. Mas *as duas formas alternativas de definir SQA e SQB apenas produzem resultados iguais no caso de delineamentos equilibrados.* Ou seja, embora em ambos os casos se tem uma única definição de $SQRE_{A+B}$, apenas no caso de delineamentos equilibrados será verdade que:

$$SQF_A = SQA = SQA' = SQRE_B - SQRE_{A+B}$$

e:

$$SQF_B = SQB' = SQB = SQRE_A - SQRE_{A+B} \ .$$

Assim, *só no caso de delineamentos equilibrados é que a ordem dos factores é arbitrária.* O Exercício ANOVA 9 ilustra esta afirmação.

Para delineamentos não equilibrados, a variabilidade de Y associada a só ter um dos factores (por exemplo A) não é igual à variabilidade adicional que esse mesmo Factor A explica, depois de se ter inicialmente introduzido o Factor B.

Em todo o caso, e para qualquer decomposição, os graus de liberdade associados a cada uma destas Somas de Quadrados é dado pelo número de parâmetros do tipo respectivo que sobram, após a introdução das restrições $\alpha_1 = 0$ e $\beta_1 = 0$, ou seja, $a-1$ e $b-1$, respectivamente, para SQA e SQB. E continua a ser verdade que estatísticas de teste da forma $\frac{QMxx}{QMRE}$ têm distribuição F, caso sejam verdadeiras as respectivas hipóteses Nulas.

4.5.11 Fórmulas para delineamentos equilibrados

Também neste caso é possível obter fórmulas para os estimadores de cada parâmetro individual, embora essas fórmulas não sejam tão simples como no caso duma ANOVA a um Factor, razão pela qual apenas serão consideradas as fórmulas para o caso de delineamentos equilibrados.

Comecemos por definir três diferentes tipos de médias: a média global das n observações e as médias de nível de cada Factor, ou seja, as médias de cada linha e de cada coluna na grelha que esquematiza o delineamento factorial (Tabela 4.1, pg. 182). Repare-se como a tripla indexação usada neste modelo obriga a utilizar um triplo somatório para somar ao longo de todas as observações.

$\overline{Y}_{...}$ a média amostral da totalidade das $n = a\,b\,n_c$ observações: $\overline{Y}_{...} = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_c} Y_{ijk}$.

$\overline{Y}_{i..}$ a média amostral das $b\,n_c$ observações do nível i do Factor A: $\overline{Y}_{i..} = \frac{1}{b\,n_c}\sum_{j=1}^{b}\sum_{k=1}^{n_c} Y_{ijk}$.

$\overline{Y}_{.j.}$ a média amostral das $a\,n_c$ observações do nível j do Factor B: $\overline{Y}_{.j.} = \frac{1}{a\,n_c}\sum_{i=1}^{a}\sum_{k=1}^{n_c} Y_{ijk}$.

A Soma de Quadrados do Factor A, definida na Subsecção 4.5.6.2 é a Soma de Quadrados do (único) Factor no Modelo M_A, apenas com o Factor A (SQF_A). Nesse modelo, os valores ajustados são as médias das observações no mesmo nível do Factor A. Tendo em conta a notação de tripla indexação introduzida nos delineamentos factoriais a dois factores, trata-se das médias calculadas para um valor fixo do primeiro índice (i), somando ao longo de todos os possíveis valores dos outros dois índices (j e k), ou seja das médias $\overline{Y}_{i..}$ acima indicadas. Assim, os valores ajustados são da forma $\hat{Y}_{ijk} = \overline{Y}_{i..}$. Logo, *num delineamento equilibrado*, e indicando por $\overline{Y}_{...}$ a média global das n observações de Y, a Soma de Quadrados de A é sempre dada (qualquer que seja a decomposição de SQT usada) por:

$$SQF_A = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_c}(\hat{Y}_{ijk}-\overline{Y}_{...})^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_c}(\overline{Y}_{i..}-\overline{Y}_{...})^2 = b\,n_c\cdot\sum_{i=1}^{a}(\overline{Y}_{i..}-\overline{Y}_{...})^2 = SQA\,. \quad (4.23)$$

Da mesma forma, *num delineamento equilibrado*, SQB é a Soma de Quadrados do Factor (SQF_B) do Modelo M_B, apenas com o Factor B. Nesse modelo, os valores ajustados são as médias de todas as observações no nível do Factor B correspondente à observação, ou seja, $\hat{Y}_{ijk} = \overline{Y}_{.j.}$ acima indicadas. Logo,

a Soma de Quadrados do Factor B será dada por:

$$SQF_B = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_c}(\hat{Y}_{ijk}-\overline{Y}_{...})^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_c}(\overline{Y}_{.j.}-\overline{Y}_{...})^2 = a\,n_c\cdot\sum_{j=1}^{b}(\overline{Y}_{.j.}-\overline{Y}_{...})^2 = SQB\,. \quad (4.24)$$

É ainda possível mostrar que os estimadores de Mínimos Quadrados de cada parâmetro, *se o delineamento é equilibrado*, ou seja, se $n_{ij} = n_c$ em todas as células (i, j), são os seguintes:

- $\hat{\mu}_{11} = \overline{Y}_{1..} + \overline{Y}_{.1.} - \overline{Y}_{...}$
- $\hat{\alpha}_i = \overline{Y}_{i..} - \overline{Y}_{1...}$
- $\hat{\beta}_j = \overline{Y}_{.j.} - \overline{Y}_{.1..}$

Tendo em conta estas fórmulas e a equação base do Modelo, tem-se que *os valores ajustados de cada observação dependem das médias dos respectivos níveis em cada factor e da média geral de todas as observações*:

$$\hat{Y}_{ijk} = \hat{\mu}_{11} + \hat{\alpha}_i + \hat{\beta}_j = \overline{Y}_{i..} + \overline{Y}_{.j.} - \overline{Y}_{...} \quad , \quad \forall\, i, j, k\ .$$

Assim, cada resíduo é dado por:

$$E_{ijk} = Y_{ijk} - \hat{Y}_{ijk} = Y_{ijk} - (\overline{Y}_{i..} + \overline{Y}_{.j.} - \overline{Y}_{...})\ . \quad (4.25)$$

A Soma de Quadrdados Residual é a soma dos quadrados das parcelas indicadas na equação (4.25).

Aviso: Ao contrário do que sucede na ANOVA a um factor, numa ANOVA a dois Factores, sem efeitos de interacção, os valores ajustados $\hat{Y}_{ijk}$ *não* são a média das observações de Y na mesma situação experimental, ou seja, na célula (i, j).

Usando estas fórmulas (que constam também do formulário da disciplina) obtém-se o quadro de síntese da ANOVA a 2 Factores (sem interacção) para um delineamento equilibrado, que é dado na Tabela 4.2.

4.5.12 A interpretação dos parâmetros e a rigidez do modelo

O significado dos parâmetros do modelo depende da convenção usada para ultrapassar o problema da multicolinearidade das colunas da matriz $\mathbf{X}$. Vejamos a interpretação dos parâmetros resultante de usar as restrições $\alpha_1 = \beta_1 = 0$.

O parâmetro μ_{11} corresponde ao *valor esperado da variável resposta* Y *na célula cujas indicatrizes foram excluídas da matriz do delineamento.* De facto, tendo em conta a expressão da equação do modelo, para uma *observação de* Y *efectuada na célula* $(1, 1)$, correspondente ao cruzamento do primeiro nível de cada factor, tem-se:

$$Y_{11k} = \mu_{11} + \epsilon_{11k} \quad \Longrightarrow \quad E[Y_{11k}] = \mu_{11}\ .$$

Por outro lado, *o parâmetro* α_i, que se designa o *efeito do nível* i *do factor* A, corresponde ao *acréscimo no valor esperado da variável resposta* Y *associado a observações do nível* $i > 1$ *do Factor* A (relativamente às observações do primeiro nível do Factor A), *quando* $j = 1$. De facto, uma observação de Y efectuada

Fonte	g.l.	SQ	QM	f_{calc}
Factor A	$a-1$	$SQA = b\,n_c \cdot \sum_{i=1}^{a} (\overline{y}_{i..} - \overline{y}_{...})^2$	$QMA = \frac{SQA}{a-1}$	$\frac{QMA}{QMRE}$
Factor B	$b-1$	$SQB = a\,n_c \cdot \sum_{j=1}^{b} (\overline{y}_{.j.} - \overline{y}_{...})^2$	$QMB = \frac{SQB}{b-1}$	$\frac{QMB}{QMRE}$
Resíduos	$n-(a+b-1)$	$SQRE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_c} (y_{ijk} - (\overline{y}_{i..} + \overline{y}_{.j.} - \overline{y}_{...}))^2$	$QMRE = \frac{SQRE}{n-(a+b-1)}$	
Total	$n-1$	$SQT = (n-1)\,s_y^2$	–	–

Tabela 4.2: Tabela de síntese duma ANOVA a dois Factores, sem efeitos de interacção, para um delineamento equilibrado.

na célula $(i,1)$, com $i>1$, correspondente ao cruzamento dum nível do factor A diferente do primeiro, com o primeiro nível do Factor B ($j=1$) será da forma:

$$Y_{i1k} = \mu_{11} + \alpha_i + \epsilon_{i1k} \quad \Longrightarrow \quad \mu_{i1} = E[Y_{i1k}] = \mu_{11} + \alpha_i \quad \Longleftrightarrow \quad \alpha_i = \mu_{i1} - \mu_{11}\ .$$

Esta interpretação dos parâmetros α_i é sistematizada na Tabela 4.3.

		Factor B				
	Níveis	B_1	B_2	B_3	$\ldots$	B_b
FACTOR A	A_1	μ_{11}	μ_{12}	μ_{13}	$\ldots$	μ_{1b}
	A_2	$\mu_{21} = \mu_{11}+\alpha_2$	μ_{22}	μ_{23}	$\ldots$	μ_{2b}
	A_3	$\mu_{31} = \mu_{11}+\alpha_3$	μ_{32}	μ_{33}	$\ldots$	μ_{3b}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	A_a	$\mu_{a1} = \mu_{11}+\alpha_a$	μ_{a2}	μ_{a3}	$\ldots$	μ_{ab}

Tabela 4.3: Interpretação dos efeitos α_i no modelo ANOVA a dois factores, sem interacção, com as restrições $\alpha_1 = \beta_1 = 0$.

Finalmente, o parâmetro β_j corresponde ao *acréscimo no valor esperado da variável resposta Y associado a observações do nível j do Factor B* (relativamente às observações do primeiro nível do Factor B), *quando* $i=1$. Designa-se o *efeito do nível j do factor B*. De facto, uma observação de Y efectuada na célula $(1,j)$, com $j>1$, correspondente ao cruzamento do primeiro nível do factor A com um nível do Factor B diferente do primeiro será da forma:

$$Y_{1jk} = \mu_{11} + \beta_j + \epsilon_{1jk} \quad \Longrightarrow \quad \mu_{1j} = E[Y_{1jk}] = \mu_{11} + \beta_j \quad \Longleftrightarrow \quad \beta_j = \mu_{1j} - \mu_{11}\ .$$

Esta interpretação dos parâmetros β_j é sistematizada na Tabela 4.4.

		Factor B				
	Níveis	B_1	B_2	B_3	$\ldots$	B_b
FACTOR A	A_1	μ_{11}	$\mu_{12}=\mu_{11}+\beta_2$	$\mu_{13}=\mu_{11}+\beta_3$	$\ldots$	$\mu_{1b}=\mu_{11}+\beta_b$
	A_2	μ_{21}	μ_{22}	μ_{23}	$\ldots$	μ_{2b}
	A_3	μ_{31}	μ_{32}	μ_{33}	$\ldots$	μ_{3b}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	A_a	μ_{a1}	μ_{a2}	μ_{a3}	$\ldots$	μ_{ab}

Tabela 4.4: Interpretação dos efeitos β_j no modelo ANOVA a dois factores, sem interacção, com as restrições $\alpha_1 = \beta_1 = 0$.

Assim, já foram utilizados todos os $a + b - 1$ parâmetros do modelo, apenas para definir as médias populacionais correspondentes às células associadas aos primeiros níveis do Factor A e do Factor B. *O Modelo ANOVA a um Factor revela-se pouco flexível*: não existem mais parâmetros e *os valores esperados nas restantes células já estão pré-determinados*, porque essas médias populacionais das restantes células dependem dos parâmetros já introduzidos. De facto, *observações de Y efectuadas numa célula genérica (i, j), com $i > 1$ e $j > 1$*, correspondente ao cruzamento de níveis diferentes do primeiro, quer no Factor A, quer no Factor B, verificam:

$$Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \epsilon_{ijk} \qquad \Longrightarrow \qquad E[Y_{ijk}] = \mu_{11} + \alpha_i + \beta_j.$$

Os valores esperados de Y são acrescidos em relação ao valor esperado duma observação na célula de referência (célula $(1, 1)$) pelas parcelas α_i e β_j (já discutidas), mas *não há flexibilidade para descrever situações específicas de células com $i > 1$ e $j > 1$*. A implicação desse facto é que a existência de particularidades associadas a uma combinação de níveis dos dois factores (com $i > 1$ e $j > 1$), como por exemplo a ocorrência de médias especialmente elevadas ou baixas nessas células, não poderá ser adequadamente estudada com este modelo. É esse tipo de situações que se designam *interacções*, e a introdução dos *efeitos de interacção* será discutida na próxima Secção.

4.6 Delineamento factorial a 2 factores: modelo com interacção

Um modelo ANOVA a 2 Factores, *sem interacção*, para um *delineamento factorial*, isto é, em que se cruzam todos os níveis de um e outro factor, foi estudado na Secção 4.5. Mas, como se viu, trata-se dum modelo pouco flexível, que não permite total liberdade na estimação das médias populacionais de situação experimental (célula).

Um modelo sem efeitos de interacção é utilizado sobretudo quando existe *uma única observação em cada célula*, isto é, quando $n_{ij} = 1$, $\forall\ i, j$. Na presença de repetições nas células, a forma mais natural de modelar um delineamento com dois factores é a de prever a existência de *um terceiro tipo de efeitos*: os chamados **efeitos de interacção**, específicos das células (situações experimentais).

4.6.1 A equação do Modelo a dois factores, com interacção

A ideia é incorporar na equação base do modelo para Y_{ijk} uma parcela, que denotaremos $(\alpha\beta)_{ij}$, e que permita que em cada célula haja um *efeito específico associado à combinação dos níveis i do Factor A e j do Factor B*:

$$Y_{ijk} \;=\; \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \;. \tag{4.26}$$

Também no contexto deste modelo torna-se necessário admitir restrições, de forma a assegurar que a matriz do modelo $\mathbf{X}$ resultante não tenha dependências lineares nas suas colunas. Vamos admitir as seguintes *restrições aos parâmetros*:

$$\alpha_1 = 0 \quad ; \quad \beta_1 = 0 \quad ; \quad (\alpha\beta)_{1j} = 0 \;,\; \forall\, j \quad ; \quad (\alpha\beta)_{i1} = 0 \;,\; \forall\, i. \tag{4.27}$$

Estas restrições podem-se sintetizar afirmando que *qualquer efeito em que pelo menos um dos seus índices tome o valor 1 é considerado nulo.* Registe-se que as restrições acima indicadas para os efeitos de interacção podem igualmente ser escritas como $(\alpha\beta)_{ij}=0$ se $i=$ ou $j=1$.

Tem-se, a partir da equação geral (4.26) e das restrições (4.27), as seguintes expressões para as médias de célula μ_{ij}:

- Para a *primeira célula* $(i = j = 1)$: $\mu_{11} = E[Y_{11k}] = \mu$.
- Nas *restantes células* $(1, j)$ *do primeiro nível do Factor A*: $\mu_{1j} = E[Y_{1jk}] = \mu_{11} + \beta_j$.
- Nas *restantes células* $(i, 1)$ *do primeiro nível do Factor B*: $\mu_{i1} = E[Y_{i1k}] = \mu_{11} + \alpha_i$.
- Nas *células genéricas* (i, j), com $i > 1$ e $j > 1$: $\mu_{ij} = E[Y_{ijk}] = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij}$.

Como se pode constatar, cada média populacional de célula é livre de tomar qualquer valor, uma vez que existe pelo menos um parâmetro livre nas fórmulas de qualquer dessas médias.

Os efeitos α_i e β_j designam-se agora efeitos *principais* dos níveis de cada Factor, uma vez que os efeitos de interacção também são efeitos que dependem dos valores de i e j.

4.6.2 A equação vectorial do modelo

A versão vectorial do modelo com interacção associa os novos efeitos $(\alpha\beta)_{ij}$ a *variáveis indicatrizes de cada célula, excluíndo as células associadas ao primeiro nível de qualquer dos factores.*

A **equação-base do modelo ANOVA a 2 Factores, com interacção**, é:

$$\begin{aligned}\vec{\mathbf{Y}} \;=\; & \mu\vec{\mathbf{1}}_n \;+\; \alpha_2\vec{\boldsymbol{\mathcal{I}}}_{A_2} \;+\; \ldots \;+\; \alpha_a\vec{\boldsymbol{\mathcal{I}}}_{A_a} \;+\; \beta_2\vec{\boldsymbol{\mathcal{I}}}_{B_2} \;+\; \ldots \;+\; \beta_b\vec{\boldsymbol{\mathcal{I}}}_{B_b} + \\ & + (\alpha\beta)_{22}\vec{\boldsymbol{\mathcal{I}}}_{A_2:B_2} \;+\; (\alpha\beta)_{23}\vec{\boldsymbol{\mathcal{I}}}_{A_2:B_3} \;+\; \ldots \;+\; (\alpha\beta)_{ab}\vec{\boldsymbol{\mathcal{I}}}_{A_a:B_b} \;+\; \vec{\boldsymbol{\epsilon}}\end{aligned}$$

onde $\vec{\boldsymbol{\mathcal{I}}}_{A_i:B_j}$ representa a *variável indicatriz da célula* correspondente ao nível i do Factor A e nível j do factor B. Cada indicatriz de célula é da forma $\vec{\boldsymbol{\mathcal{I}}}_{A_i:B_j} = \vec{\boldsymbol{\mathcal{I}}}_{A_i} \star \vec{\boldsymbol{\mathcal{I}}}_{B_j}$, com o operador $\star$ a indicar uma multiplicação, elemento a elemento, entre dois vectores.

Neste modelo, que designamos modelo M_{A*B}, existem ao todo *ab parâmetros* desconhecidos, que são:

- a 1 média da célula de referência, μ_{11};
- os $a-1$ acréscimos α_i $(i>1)$;
- os $b-1$ acréscimos β_j $(j>1)$; e
- os $(a-1)(b-1)$ efeitos de interacção $(\alpha\beta)_{ij}$, para $i>1$, $j>1$.

O ajustamento deste modelo faz-se de forma análoga ao ajustamento de modelos anteriores. A matriz $\mathbf{X}$ do delineamento é agora constituída por ab colunas:

- *uma* coluna de uns, $\vec{\mathbf{1}}_n$, associada ao parâmetro μ_{11}.
- $a-1$ colunas de indicatrizes de nível do factor A, $\vec{\boldsymbol{\mathcal{I}}}_{A_i}$, $(i>1)$, associadas aos parâmetros α_i.
- $b-1$ colunas de indicatrizes de nível do factor B, $\vec{\boldsymbol{\mathcal{I}}}_{B_j}$, $(j>1)$, associadas aos parâmetros β_j.
- $(a-1)(b-1)$ indicatrizes de célula, $\vec{\boldsymbol{\mathcal{I}}}_{A_i:B_j}$, $(i,j>1)$, associadas a efeitos de interacção $(\alpha\beta)_{ij}$.

Como em modelos anteriores, o vector dos valores ajustados de Y, $\vec{\hat{\mathbf{Y}}}$, é obtido pré-multiplicando o vector dos valores observados, $\vec{\mathbf{Y}}$, pela matriz $\mathbf{H}$ de projecção ortogonal sobre $\mathcal{C}(\mathbf{X})$, construída a partir dessa matriz do modelo: $\vec{\hat{\mathbf{Y}}} = \mathbf{H}\vec{\mathbf{Y}}$. E como habitualmente, a Soma de Quadrados Residual é da forma: $SQRE_{A*B} = \|\vec{\mathbf{Y}} - \vec{\hat{\mathbf{Y}}}\|^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(Y_{ijk} - \hat{Y}_{ijk})^2$.

4.6.3 O modelo ANOVA a dois factores, com interacção

Juntando os pressupostos necessários à inferência, obtém-se o **Modelo ANOVA a dois factores, com interacção**, que será representado em curto como o **Modelo** M_{A*B}. Admite-se que existem n observações, Y_{ijk}, das quais n_{ij} correspondem à célula (i,j) $(i=1,...,a;\ j=1,...,b)$, e que:

1. $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$, $(i{=}1,...,a\ ;\ j{=}1,...,b\ ;\ k{=}1,...,n_{ij})$
   com as restrições $\alpha_1 = 0$; $\beta_1 = 0$; $(\alpha\beta)_{1j} = 0$, $\forall\, j$; $(\alpha\beta)_{i1} = 0$, $\forall\, i$.
2. $\epsilon_{ijk} \frown \mathcal{N}(0\,,\,\sigma^2)$, para todo o i, j e k.
3. $\{\epsilon_{ijk}\}_{i,j,k}$ são um conjunto de variáveis aleatórias independentes.

4.6.4 Os três testes ANOVA

Neste modelo, em cuja equação de base (4.26) existem três tipos de efeitos, desejamos fazer um *teste à existência de cada um desses três tipos de efeitos*:

Teste I: à existência de efeitos de interacção: $H_0 : (\alpha\beta)_{ij} = 0\,, \quad \forall\, i = 2,...,a\ ,\ \forall j = 2,...,b\ ;$

Teste II: à existência de efeitos principais do Factor A: $H_0 : \alpha_i = 0\,, \quad \forall i = 2,...,a\ ;$ e

Teste III: à existência de efeitos principais do Factor B: $H_0 : \beta_j = 0\,, \quad \forall j = 2,...,b\ .$

As estatísticas de teste para cada um destes testes obtêm-se a partir da decomposição da Soma de Quadrados Total em parcelas convenientes.

4.6.4.1 A decomposição de SQT

Para testar a existência de efeitos de interacção, com hipótese Nula correspondente à inexistência desses efeitos,

$$H_0 : (\alpha\beta)_{ij} = 0\,, \quad \forall\, i = 2, ..., a\ ,\ \forall j = 2, ..., b\ ,$$

pode efectuar-se *um teste F parcial* comparando o *modelo* agora considerado,

$$\text{(Modelo } M_{A*B}) \qquad Y_{ijk} \;=\; \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}\ ,$$

com o *submodelo* correspondente ao Modelo a dois factores, mas *sem efeitos de interacção*:

$$\text{(Modelo } M_{A+B}) \qquad Y_{ijk} \;=\; \mu_{11} + \alpha_i + \beta_j + \epsilon_{ijk}\ ,$$

A *Soma de Quadrados associada à interacção* corresponde à diferença das Somas de Quadrados Residuais desses dois Modelos, que surgirá no numerador da estatística desse teste F parcial:

$$SQAB \;=\; SQRE_{A+B} - SQRE_{A*B} \tag{4.28}$$

Para testar os efeitos principais do Factor B, com a Hipótese Nula $H_0 : \beta_j = 0\,,\ \forall j = 2, ..., b$ (que, mais uma vez, corresponde à inexistência dos referidos efeitos), pode considerar-se o Modelo a dois Factores, sem efeitos de interacção, e o Modelo com um único Factor, o Factor A, cujas equações são:

$$\begin{aligned} \text{(Modelo } M_{A+B}) \qquad Y_{ijk} &= \mu_{11} + \alpha_i + \beta_j + \epsilon_{ijk} \\ \text{(Modelo } M_{A}) \qquad Y_{ijk} &= \mu_{11} + \alpha_i + \epsilon_{ijk}\ , \end{aligned}$$

e definir Somas de Quadrados associadas a cada um dos Factores, de forma igual ao que foi feito aquando do estudo do modelo sem efeitos de interacção:

$$\begin{aligned} SQB &= SQRE_A - SQRE_{A+B} \\ SQA &= SQF_A \;=\; SQT - SQRE_A \end{aligned}$$

Assim, definiram-se as três Somas de Quadrados associadas aos três tipos de efeitos previstos na equação do Modelo:

$$\begin{aligned} SQAB &= SQRE_{A+B} - SQRE_{A*B} \\ SQB &= SQRE_A - SQRE_{A+B} \\ SQA &= SQF_A \;=\; SQT - SQRE_A \end{aligned}$$

Somando estas Somas de Quadrados à Soma de Quadrados Residual, $SQRE_{A*B}$, obtém-se:

$$SQRE_{A*B} + SQAB + SQA + SQB = SQT \tag{4.29}$$

Esta decomposição de SQT gera as quantidades nas quais se baseiam as estatísticas dos três testes associados ao Modelo M_{A*B}.

A cada uma das Somas de Quadrados associam-se *graus de liberdade* de acordo com as seguintes regras (análogas às de outros modelos ANOVA):

- os graus de liberdade *associados a cada um dos tipos de efeitos* são dados pelo *número de parâmetros desse tipo, após a imposição das restrições*: $a-1$ para os efeitos principais do Factor A; $b-1$ para os efeitos principais do Factor B; e $(a-1)(b-1)$ para os efeitos de interacção.
- os graus de liberdade *residuais* são o número de observações (n) menos o número de parâmetros do modelo (ab).

Como é hábito definem-se Quadrados Médios dividindo cada uma das Somas de Quadrados pelos respectivos graus de liberdade. E, também como noutras ANOVAs, as estatísticas de cada um dos três testes reultarão de dividir o Quadrado Médio do tipo de efeito que se pretende testar, pelo Quadrado Médio Residual. Vejamos agora em pormenor cada um dos três testes F para este Modelo.

4.6.4.2 O Teste F aos efeitos de interacção

Sendo válido o Modelo ANOVA a dois factores, com interacção, o **Teste F aos efeitos de interacção** é definido pelos seguintes passos:

Hipóteses: $H_0 : (\alpha\beta)_{ij} = 0\ , \quad \forall\, i,j \quad$ vs. $\quad H_1 : \exists i,j$ tal que $(\alpha\beta)_{ij} \neq 0$.
[NÃO HÁ INTERACÇÃO] vs. [HÁ INTERACÇÃO]

Estatística do Teste: $F = \frac{QMAB}{QMRE} \underset{\cap}{\sim} F_{[\,(a-1)(b-1)\ ,\ n-ab\,]}$ se H_0.

Nível de significância do teste: α

Região Crítica (Unilateral direita): Rejeitar H_0 se $F_{calc} > f_{\alpha(\,(a-1)(b-1)\,,\,n-ab\,)}$

4.6.4.3 O Teste F aos efeitos principais do factor A

Sendo válido o Modelo ANOVA a 2 factores com interacção, o Teste F aos efeitos principais do factor A define-se da forma seguinte:

Hipóteses: $H_0 : \alpha_i = 0\ , \quad i = 2,...,a \quad$ vs. $\quad H_1 : \exists i = 2,..,a$ tal que $\alpha_i \neq 0$.
[NÃO HÁ EFEITOS DE A] vs. [HÁ EFEITOS DE A]

Estatística do Teste: $F = \frac{QMA}{QMRE} \underset{\cap}{\sim} F_{(\,a-1\,,\,n-ab\,)}$ se H_0.

Nível de significância do teste: α

Região Crítica (Unilateral direita): Rejeitar H_0 se $F_{calc} > f_{\alpha(a-1,n-ab)}$

4.6.4.4 O Teste F aos efeitos principais do factor B

Sendo válido o Modelo ANOVA a 2 factores com interacção, o Teste F aos efeitos principais do factor B

Hipóteses: $H_0 : \beta_j = 0\ , \quad \forall\, j = 2,...,b \quad$ vs. $\quad H_1 : \exists j = 2,..,b$ tal que $\beta_j \neq 0$.
[NÃO HÁ EFEITOS DE B] vs. [HÁ EFEITOS DE B]

Estatística do Teste: $F = \frac{QMB}{QMRE} \underset{H_0}{\sim} F_{(b-1,\,n-ab)}$ se H_0.

Nível de significância do teste: α

Região Crítica (Unilateral direita): Rejeitar H_0 se $F_{calc} > f_{\alpha(b-1,n-ab)}$

A informação relevante para esses testes pode ser coleccionada num quadro-resumo, como nos Modelos ANOVA anteriores.

4.6.4.5 O quadro de síntese

Com base na decomposição da equação (4.29) pode-se construir o *quadro de síntese da ANOVA a 2 Factores, com interacção*, dada na Tabela 4.5

Fonte	g.l.	SQ	QM	f_{calc}
Factor A	$a-1$	$SQA = SQF_A$	$QMA = \frac{SQA}{a-1}$	$\frac{QMA}{QMRE}$
Factor B	$b-1$	$SQB = SQRE_A - SQRE_{A+B}$	$QMB = \frac{SQB}{b-1}$	$\frac{QMB}{QMRE}$
Interacção	$(a-1)(b-1)$	$SQAB = SQRE_{A+B} - SQRE_{A*B}$	$QMAB = \frac{SQAB}{(a-1)(b-1)}$	$\frac{QMAB}{QMRE}$
Resíduos	$n-ab$	$SQRE = SQRE_{A*B}$	$QMRE = \frac{SQRE}{n-ab}$	
Total	$n-1$	$SQT = (n-1)\,s_y^2$	–	–

Tabela 4.5: O quadro resumo duma ANOVA a dois factores, com efeitos de interacção, válido quer o delineamento seja, ou não, equilibrado.

4.6.5 ANOVA a dois Factores com interacção no R

Para efectuar uma ANOVA a dois Factores, com interacção, no R, organizam-se os dados de forma igual à usada para o modelo sem interacção, ou seja, numa `data.frame` com três colunas:

1. uma coluna para a variável resposta;
2. outra coluna, de classe `factor`, para o factor A;
3. uma terceira coluna, também de classe `factor`, para o factor B.

As fórmulas utilizadas no R para indicar uma ANOVA a dois Factores, com interacção, recorrem ao símbolo '$*$' :

```
y  ~  fA * fB
```

sendo `y` o nome da variável resposta e `fA` e `fB` os nomes dos factores.

4.6.6 A necessidade de repetições nas células

Importa sublinhar que, para se poder estudar o modelo a dois Factores com efeitos de interacção, é necessário que haja repetições nas células. Uma forma fácil de ver que assim é, consiste em observar que os graus de liberdade do $SQRE$ neste modelo são $n-ab$. Se houver uma única observação em cada célula, tem-se $n = ab$, ou seja, tantos parâmetros quantas as observações existentes. Nesse caso, nem sequer será possível definir o Quadrado Médio Residual, $QMRE$, cujo denominador seria, neste caso, zero.

Assim, *num delineamento com uma única observação por célula é obrigatório optar por um modelo sem interacção*, uma vez que não existe informação suficiente (ou seja, observações suficientes) para estudar os efeitos de interacção. É sobretudo por esta razão que se justifica o interesse no modelo a dois factores, sem efeitos de interacção.

Havendo repetições, é mais natural considerar um modelo com interacção e deixar que a conclusão sobre a existência, ou não, desse tipo de efeitos resulte do estudo do modelo. Mas a realização de repetições pode ser demasiado dispendiosa e trabalhosa, em cujo caso um delineamento com uma única observação por célula pode tornar-se convidativa e o modelo sem os efeitos de interacção a única possibilidade viável de estudar os efeitos de cada factor em separado. Nesses casos, *a eventual existência de interacção*, que não foi possível estudar no modelo, irá *inflacionar a variabilidade residual*, não explicada pelo modelo.

4.6.7 Algumas fórmulas de interesse

4.6.7.1 Médias de Y

Às médias já definidas no estudo do modelo a dois Factores, sem efeitos de interacção, ou seja as médias global ($\overline{Y}_{...}$), de cada nível do Factor A ($\overline{Y}_{i..}$) e de cada nível do Factor B ($\overline{Y}_{.j.}$), acrescentam-se agora as médias de cada célula:

$$\overline{Y}_{ij.} \;=\; \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} Y_{ijk} \;. \tag{4.30}$$

4.6.7.2 Valores ajustados de Y

Os *valores ajustados* $\hat{Y}_{ijk}$ são *iguais para todas as observações numa mesma célula*, e são dados pela *média amostral da célula*:

$$\hat{Y}_{ijk} \;=\; \overline{Y}_{ij.} \;. \tag{4.31}$$

4.6.7.3 Estimadores dos parâmetros

Os estimadores dos parâmetros num modelo ANOVA a 2 Factores, com interacção, são, tal como numa ANOVA a um Factor, as quantidades amostrais correspondentes ao significado que, na população, tem cada um dos parâmetros. Como foi visto na Subsecção 4.6.1, tem-se:

- μ_{11} é a média populacional para a *primeira célula* $(i = j = 1)$.
- α_i (com $i>1$) é a diferença da média populacional das células $(i, 1)$ e $(1, 1)$: $\alpha_i = \mu_{i1} - \mu_{11}$.

- β_j (com $j>1$) é a diferença da média populacional das células $(1,j)$ e $(1,1)$: $\beta_j = \mu_{1j} - \mu_{11}$.
- Os efeitos de interacção nas *células genéricas* (i,j), com $i > 1$ e $j > 1$ podem ser escritos como: $(\alpha\beta)_{ij} = \mu_{ij} - \mu_{11} - \alpha_i - \beta_j = \mu_{ij} - \mu_{i1} - \mu_{1j} + \mu_{11}$.

Assim, os estimadores destes parâmetros resultam de substituir as médias populacionais nestas expressões pelas correspondentes médias amostrais:

- $\hat{\mu}_{11} = \overline{Y}_{11.}$
- $\hat{\alpha}_i = \overline{Y}_{i1.} - \overline{Y}_{11.} \qquad (i > 1)$
- $\hat{\beta}_j = \overline{Y}_{1j.} - \overline{Y}_{11.} \qquad (j > 1)$
- $\widehat{(\alpha\beta)}_{ij} = (\overline{Y}_{ij.} + \overline{Y}_{11.}) - (\overline{Y}_{i1.} + \overline{Y}_{1j.}) \qquad (i,j > 1).$

Intervalos de confiança ou testes de hipóteses para qualquer dos parâmetros individuais, ou combinações lineares desses parâmetros, podem ser efectuados utilizando a teoria geral do Modelo Linear, ou seja, através de testes t.

4.6.7.4 A Soma de Quadrados Residual

Como os valores ajustados correspondem às medias amostrais da célula onde se efectuaram as observações, $\hat{Y}_{ijk} = \overline{Y}_{ij.}$, tem-se:

$$SQRE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} (Y_{ijk} - \hat{Y}_{ijk})^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} (Y_{ijk} - \overline{Y}_{ij.})^2 = \sum_{i=1}^{a}\sum_{j=1}^{b} (n_{ij}-1) S_{ij}^2 \ , \qquad (4.32)$$

sendo S_{ij}^2 a variância amostral das observações da célula (i,j).

Num delineamento equilibrado, tem-se $n = n_c ab$, e o Quadrado Médio Residual será então a média simples das variâncias amostrais de célula, S_{ij}^2:

$$QMRE = \frac{SQRE}{n-ab} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b} S_{ij}^2 \ .$$

4.6.7.5 SQA e SQB em delineamentos equilibrados

Para delineamentos equilibrados (com n_c observações por célula) é possível obter igualmente fórmulas simples para as Somas de Quadrados associadas aos efeitos principais de cada factor. Estas fórmulas correspondem (tal como no modelo sem efeitos de interacção) às Somas de Quadrados associadas a cada factor, caso se ajustasse (aos mesmos dados) um modelo ANOVA apenas com esse factor:

$$SQA = bn_c \sum_{i=1}^{a} (\overline{Y}_{i..} - \overline{Y}_{...})^2$$

$$SQB = an_c \sum_{j=1}^{b} (\overline{Y}_{.j.} - \overline{Y}_{...})^2$$

4.6.8 Comparações múltiplas de médias de células

O número potencialmente grande de comparações possíveis entre pares de *médias de célula* aconselha a utilização de *métodos de comparação múltipla*, que permitam controlar globalmente o nível de significância do conjunto de testes de hipóteses (ou grau de confiança do conjunto de intervalos de confiança).

O mais utilizado dos métodos de comparação múltipla está associado ao nome de Tukey, e foi já introduzido aquando do estudo de delineamentos a um Factor. Adapta-se facilmente à comparação múltipla de médias de células.

Admite-se que o delineamento é *equilibrado*, com $n_c > 1$ repetiçoes em todas as ab células. Sendo $q_{\alpha\,(ab,n-ab)}$ o valor que deixa à direita uma região de probabilidade α numa distribuição de Tukey com parâmetros $k = ab$ (o número total de médias de célula) e $\nu = n - ab$ (os graus de liberdade associados ao $QMRE$), têm-se os seguintes resultados, relativos às abordagens alternativas via testes de hipóteses ou intervalos de confiança.

4.6.8.1 Testes a Hipóteses sobre $\mu_{ij} - \mu_{i'j'}$

Rejeita-se a igualdade das médias das células (i,j) *e* (i',j'), a favor da hipótese $\mu_{ij} \neq \mu_{i'j'}$, se

$$|\overline{Y}_{ij\cdot} - \overline{Y}_{i'j'\cdot}| \quad > \quad q_{\alpha\,(ab,n-ab)} \cdot \sqrt{\frac{QMRE}{n_c}}\ , \tag{4.33}$$

O nível de significância α é global, ou seja, relativo à totalidade das comparações de cada par de níveis.

4.6.8.2 Intervalos de Confiança para $\mu_{ij} - \mu_{i'j'}$

Com grau de confiança global $(1-\alpha)\times 100\%$, todas as diferenças de médias de pares de células, $\mu_{ij} - \mu_{i'j'}$, estão em intervalos da forma:

$$\Big]\ \left(\overline{y}_{ij\cdot} - \overline{y}_{i'j'\cdot}\right) - q_{\alpha\,(ab,n-ab)}\sqrt{\tfrac{QMRE}{n_c}} \quad , \quad \left(\overline{y}_{ij\cdot} - \overline{y}_{i'j'\cdot}\right) + q_{\alpha\,(ab,n-ab)}\sqrt{\tfrac{QMRE}{n_c}}\ \Big[$$

Conclui-se que duas médias populacionais de célula são diferentes, $\mu_{ij} \neq \mu_{i'j'}$, se o intervalo de confiança correspondente a este par de células não contém o valor zero.

4.6.8.3 Tukey no `R`

A obtenção dos Intervalos de Confiança de Tukey no `R`, para a diferença da média de células, no caso de um delineamento a dois Factores, é análogo ao caso de um único factor:

$>$ `TukeyHSD(aov(y ~ fA * fB, data=`*`dados`*`))`

O comando produz também intervalos de confiança para as *médias de nível* de cada Factor isoladamente.

É possível representar graficamente estes Intervalos de Confiança encaixando o comando anterior na função `plot`.

4.6.9 Análise dos Resíduos

A validade dos pressupostos do Modelo relativos aos erros aleatórios pode ser estudada de forma análoga ao que foi visto para um delineamento a 1 Factor. Também neste caso, convém sublinhar algumas especifidades deste estudo, para o contexto de ANOVAs a dois Factores, com efeitos de interacção.

Os resíduos relativos a uma mesma célula aparecem em ab colunas verticais num gráfico de E_{ijk} vs. $\hat{Y}_{ijk}$.

A hipótese de heterogeneidade de variâncias entre diferentes células pode ser testada recorrendo a testes de hipóteses (como o Teste de Bartlett), mas essa matéria não será leccionada.

4.6.9.1 O Teste de Bartlett para delineamentos a dois factores *

(*) O Teste de Bartlett não é avaliado

Um Teste de Bartlett visa estudar a homogeneidade de variâncias em cada célula, sendo a existência dessa homogeneidade a Hipótese Nula do teste.

Hipóteses: $H_0: \sigma^2_{11} = \sigma^2_{12} = ... = \sigma^2_{ab}$ vs. $H_1: \exists_{i,j,i',j'} : \sigma^2_{ij} \neq \sigma^2_{i'j'}$
[Variâncias homogéneas] vs. [Variâncias heterogéneas]

Estatística do Teste:

$$K^2 = \frac{(n-ab)\ln QMRE - \sum_{i=1}^{a}\sum_{j=1}^{b}(n_{ij}-1)\ln S^2_{ij}}{C} \sim \chi^2_{ab-1},$$

onde $C = 1 + \frac{1}{3(ab-1)}\left[\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{1}{n_{ij}-1} - \frac{1}{n-ab}\right]$

Nível de significância do teste: α

Região Crítica (Unilateral direita): Rejeitar H_0 se $K^2_{calc} > \chi^2_{\alpha(ab-1)}$

Tal como no Modelo a um Factor, a distribuição da estatística do Teste de Bartlett é apenas assintótica, pelo que o teste exige amostras de grande dimensão. Além disso, é um teste fortemente dependente da Normalidade dos erros aleatórios.

4.6.10 Uma advertência

Na formulação clássica do modelo ANOVA a dois Factores, com interacção, e a partir da equação-base $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$, em vez de impor as restrições $\alpha_1 = \beta_1 = (\alpha\beta)_{i1} = (\alpha\beta)_{1j} = 0$ $(\forall\, i, j)$, admite-se a existência de acréscimos de todos os tipos para qualquer valor de i e j e impõe-se as condições:

- $\sum_i \alpha_i = 0$;
- $\sum_j \beta_j = 0$;

- $\sum_i (\alpha\beta)_{ij} = 0\ , \qquad \forall\, j;$
- $\sum_j (\alpha\beta)_{ij} = 0\ , \qquad \forall\, i.$

Tal como em Modelos ANOVA anteriores, estas condições alternativas:

- mudam a forma de interpretar os parâmetros;
- mudam os estimadores dos parâmetros;
- *não* mudam o resultado dos testes F à existência de efeitos.

4.6.11 Visualização gráfica de efeitos de interacção

Efeitos de interacção em delineamentos factoriais a dois factores podem ser visualizados em gráficos onde:

- o eixo horizontal é associado aos níveis de um dos factores (por exemplo, o factor f_A);
- no eixo vertical são indicados os valores médios da variável resposta Y, em cada célula;
- para cada célula, indica-se um ponto cujas coordenadas são determinadas pelo nível do primeiro factor e respectiva média de célula da variável resposta;
- unem-se com segmentos de recta os pontos correspondentes a um mesmo nível do segundo factor (por exemplo, o factor f_B).

A utilização destes *gráficos de interacção* é ilustrada na Figura 4.8.

A inexistência de interacção significativa produz, nestes gráficos, linhas aproximadamente "paralelas". Havendo interacção, as linhas estarão longe de qualquer paralelismo. A confirmação da significância dos efeitos de interacção exige sempre que se efectue o respectivo teste F.

A cada problema correspondem sempre *dois possíveis gráficos de interacção*, uma vez que é arbitrária a escolha de qual o factor associado ao eixo horizontal.

Serão em seguida estudados um tipo de delineamentos a dois factores, mas *não* factoriais.

4.7 Delineamentos hierarquizados

Delineamentos que, superficialmente, podem confundir-se com os delineamentos factoriais são delineamentos com dois (ou mais) factores, mas em que *os níveis de um dos factores diferem consoante os níveis do outro factor.*

4.7.1 Um exemplo

Considere-se o seguinte exemplo: pretende-se estudar o índice de desempenho (variável resposta), em várias tarefas, de três tractores, de diferentes modelos (factor A), cada um dos quais é conduzidos por

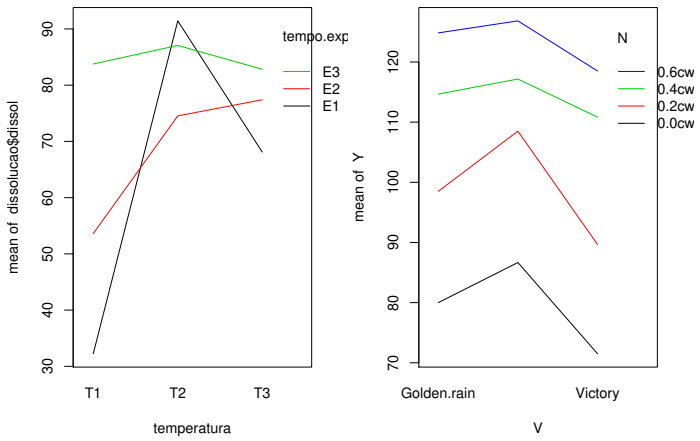


Figura 4.8: Dois exemplos de gráficos de interacção. À esquerda, os segmentos de rectas correspondentes aos vários níveis do factor estão longe de qualquer 'paralelismo', pelo que há fortes indícios da existência de efeitos significativos de interacção. À direita há um maior 'paralelismo', que sugere que a interacção não é significativa. Apenas a realização do teste F à existência de efeitos de interacção permite tornar estes indícios em conclusões estatisticamente sustentadas.

quatro tractoristas (factor B). Se os mesmos 4 tractoristas conduzirem os 3 tractores, estamos perante um delineamento factorial e aplicam-se os modelos antes considerados. Mas se para cada modelo de tractor existir um grupo de quatro diferentes tractoristas especializados (ao todo 12 pessoas), o delineamento não é factorial, mas antes *hierarquizado*: só é possível identificar os tractoristas (níveis do factor B), após especificar o tractor (nível do factor A). Assim, existe uma *hierarquia* dos factores: só identificamos os níveis de um dos factores (o *factor subordinado*) após ter identificado o nível do outro factor (*factor dominante*). A situação pode ser representada esquematicamente da seguinte forma, que utiliza um tipo de grelha semelhante às já consideradas no estudo de delineamento factoriais.

	Tractor A_1	Tractor A_2	Tractor A_3
Tractorista $A_1 1$	×	-	-
Tractorista $A_1 2$	×	-	-
Tractorista $A_1 3$	×	-	-
Tractorista $A_1 4$	×	-	-
Tractorista $A_2 1$	-	×	-
Tractorista $A_2 2$	-	×	-
Tractorista $A_2 3$	-	×	-
Tractorista $A_2 4$	-	×	-
Tractorista $A_3 1$	-	-	×
Tractorista $A_3 2$	-	-	×
Tractorista $A_3 3$	-	-	×
Tractorista $A_3 4$	-	-	×

Um delineamento hierarquizado pode assim ser visto como um *delineamento factorial* muito *incompleto*.

Mas uma representação alternativa, em forma de dendrograma, pode ser mais útil para transmitir a noção da dependência entre os níveis dos dois factores. Essa representação alternativa, para o exemplo em consideração, é dada na Figura 4.9.

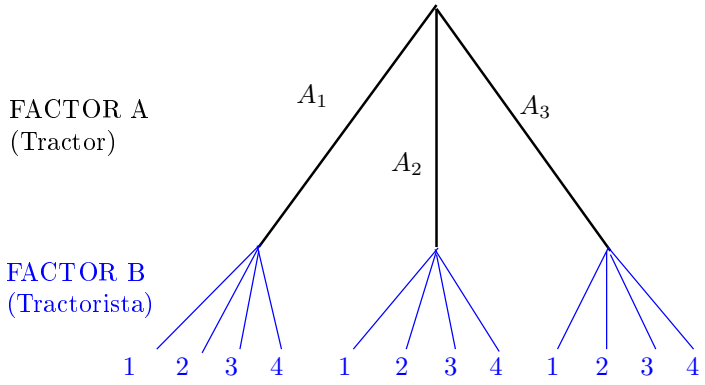


Figura 4.9: Dendrograma representativo do delineamento hierarquizado relacionando tractores e tractoristas. Cada 'folha' na ponta de cada 'ramo terminal' do dendrograma corresponde a uma diferente situação experimental.

Um tal delineamento diz-se *hierarquizado* (*nested*, em inglês). Nos delineamentos hierarquizados deixa de fazer sentido falar em efeitos de interacção entre os níveis de cada Factor, uma vez que os níveis do factor subordinado são específicos de um dado nível do factor dominante.

Note-se que nada obriga a que o número de níveis do factor subordinado seja igual, nos vários níveis do factor dominante. Assim, seria possível que, no exemplo acima considerado, houvesse um número diferente de tractoristas a guiar dois diferentes tractores.

4.7.2 A equação do Modelo a dois factores hierarquizados

O Factor dominante num delineamento hierarquizado a dois factores será genericamente designado o Factor A, com a níveis. O factor subordinado será designado Factor B. Mas, para cada nível i do factor dominante, pode existir um número b_i de níveis do factor dominado, que não tem de ser sempre igual, para os vários valores de i.

Tal como nos modelos para delineamentos factoriais a dois factores, cada observação é representada por uma variável aleatória com *três índices*, Y_{ijk}:

i nível do factor dominante ($i = 1, ..., a$);

j nível do factor subordinado ($j = 1, ..., b_i$);

k repetição para a célula (i, j), com $k = 1, ..., n_{ij}$.

A equação base do modelo inclui *efeitos de nível do Factor A*, α_i, e *efeitos de nível do factor B (subordinado)*, que serão representados por $\beta_{j(i)}$, para salientar que o nível j de B é contado no seio do nível i do factor dominante A. Com esta notação, a equação geral do modelo será da forma:

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijk} \ . \tag{4.34}$$

Mais uma vez, é necessário impôr restrições, de forma a garantir que as colunas da matriz do modelo $\mathbf{X}$ sejam linearmente independentes. Neste caso, além da exigência de que seja nulo o efeito associado ao primeiro nível do Factor dominante, A ($\alpha_1 = 0$), exigir-se-á que no primeiro nível do factor subordinado (Factor B), em *todos* os níveis do Factor A, o efeito correspondente seja igualmente nulo: $\beta_{1(i)} = 0$, $\forall\, i$, tal como ilustrado na Figura 4.10. Sublinhe-se que os efeitos de nível $j > 1$ no Factor B são livres, mesmo

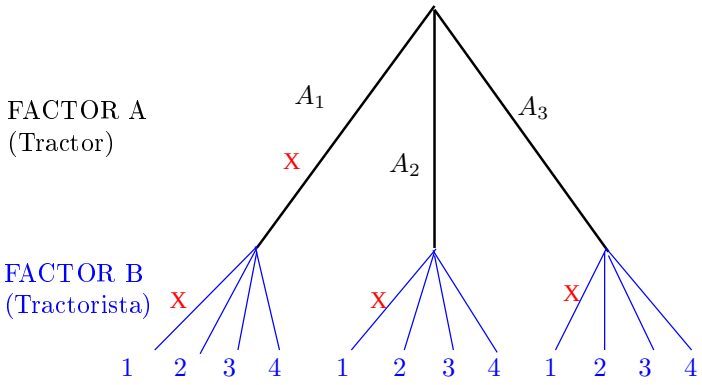


Figura 4.10: Dendrograma indicando as restrições num modelo para um delineamento hierarquizado. As cruzes a vermelho indicam os ramos (terminais ou não) aos quais está associado um efeito nulo.

no primeiro nível do Factor A dominante ($i = 1$). Com estas restrições, a constante comum a todas as observações, μ, será a média correspondente à primeira situação experimental, em que simultaneamente $i = 1$ e $j = 1$, ou seja, $\mu = \mu_{11}$.

Nos delineamentos hierarquizados, não faz sentido falar em efeitos do nível j do Factor B, sem especificar qual o nível do Factor A a que nos referimos. Nem faz sentido falar em efeitos de interacção.

4.7.3 Particularidades do Modelo

4.7.3.1 Variáveis indicatrizes

Tal como em modelos anteriores, a cada parâmetro associa-se uma variável indicatriz das observações correspondentes. Assim:

- o parâmetro μ_{11} está associado à coluna de uns, $\vec{\mathbf{1}}_n$.
- cada um dos $(a-1)$ parâmetros α_i está associado a uma indicatriz $\vec{\boldsymbol{\mathcal{I}}}_{A_i}$ de pertença ao nível $i > 1$ do Factor A.

- cada um dos $\sum_{i=1}^{a}(b_i-1)$ parâmetros $\beta_{j(i)}$ está associado a uma indicatriz $\vec{\boldsymbol{1}}_{B_{j(i)}}$ de pertença ao nível $j>1$ do Factor B (para $i=1,...,a$).

4.7.3.2 Parâmetros

O *número total de parâmetros* é *igual ao número de situações experimentais*:

$$1+(a-1)+\sum_{i=1}^{a}(b_i-1) \quad = \quad \cancel{1}+(\cancel{a}-\cancel{1})+\sum_{i=1}^{a}b_i-\cancel{a} \quad = \quad \sum_{i=1}^{a}b_i$$

No caso de haver sempre o mesmo número de níveis do Factor subordinado B, em qualquer nível i de A ($b=b_i$ para qualquer i), haverá ab parâmetros no modelo.

4.7.3.3 Os valores esperados de Y_{ijk}

As restrições acabadas de discutir implicam que os valores esperados de observações de Y em cada situação experimental podem ser interpretadas da seguinte forma:

- Para a primeira célula ($i=j=1$), $E[Y_{ijk}]=\mu=\mu_{11}$.
- Nas *restantes células* do primeiro nível do Factor A ($i=1; j>1$), $\mu_{1j}=E[Y_{ijk}]=\mu_{11}+\beta_{j(1)}$.
- Nos restantes primeiros níveis do factor B ($i>1; j=1$), $\mu_{i1}=E[Y_{ijk}]=\mu_{11}+\alpha_i$.
- Nas células genéricas (i,j), com $i>1$ e $j>1$, $\mu_{ij}=E[Y_{ijk}]=\mu_{11}+\alpha_i+\beta_{j(i)}$.

Os efeitos α_i e $\beta_{j(i)}$ designam-se *efeitos dos níveis de cada Factor*.

4.7.4 O modelo ANOVA a dois factores, hierarquizados.

Juntando os pressupostos necessários à inferência, tem-se o **Modelo ANOVA a dois factores, hierarquizados**, que será doravante representado por **Modelo** $M_{A/B}$.

Seja A o Factor dominante e B o Factor subordinado. Admitem-se que existem n observações, Y_{ijk}, das quais n_{ij} associadas à célula (i,j) ($i=1,...,a\ ;\ j=1,...,b_i$). Tem-se:

1. $Y_{ijk} \ = \ \mu_{11}+\alpha_i+\beta_{j(i)}+\epsilon_{ijk}\ , \quad \forall\, i=1,...,a\ ;\ j=1,...,b_i\ ;\ k=1,...,n_{ij}$
   com $\alpha_1=0\ ;\ \beta_{1(i)}=0\ ,\ \forall i.$
2. $\epsilon_{ijk} \ \frown \ \mathcal{N}(0\,,\,\sigma^2)\ , \qquad \forall\, i,j,k$
3. $\{\epsilon_{ijk}\}_{i,j,k}$ variáveis aleatórias independentes.

4.7.5 Os dois testes ANOVA

Neste caso, tal como noutros modelos ANOVA, pretende-se *testar a existência de cada um dos dois tipos de efeitos previstos no modelo*, havendo por isso lugar a dois testes:

Teste I: aos efeitos de nível do Factor A (dominante), com $H_0 : \alpha_i = 0\,,\ \forall i = 2, ..., a$; e

Teste II: aos efeitos do conjunto dos níveis do Factor B (subordinado), com $H_0 : \beta_{j(i)} = 0\,,\ \forall i = 1, ..., a$ e $j = 2, ..., b_i$.

As estatísticas de teste para cada um destes testes obtêm-se a partir da *decomposição da Soma de Quadrados Total* em *três parcelas*, correspondentes aos dois tipos de efeito e à variabilidade residual. As Somas de Quadrados associadas a cada tipo de efeito definem-se de forma análoga à usada em delineamentos anteriores.

4.7.5.1 A decomposição de SQT

Para efectuar a decomposição da Soma de Quadrados Total, começamos por definir a diferença nas Somas de Quadrados do Modelo agora introduzido, e do Modelo M_A a um único Factor (o Factor A), já que essa diferença de Somas de Quadrados Residuais apareceria num teste F parcial comparando estes dois Modelos, que serviria para estudar se os efeitos do factor subordinado são, ou não, signficativos (Teste II). Assim, tomem-se os Modelos:

$$\begin{aligned} \text{(Modelo } M_{A/B}) \qquad & Y_{ijk} \;=\; \mu_{11} + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}\ , \\ \text{(Modelo } M_A) \qquad & Y_{ijk} \;=\; \mu_{11} + \alpha_i + \epsilon_{ijk}\ , \end{aligned}$$

Designa-se *Soma de Quadrados associada aos efeitos de B* a

$$SQB(A) \;=\; SQRE_A - SQRE_{A/B}$$

e *Soma de Quadrados associada aos efeitos de A* a

$$SQA \;=\; SQF_A \;=\; SQT - SQRE_A\ .$$

Juntamente com $SQRE_{A/B}$, tem-se:

$$SQT \;=\; SQA + SQB(A) + SQRE_{A/B}\ .$$

Graus de liberdade: Os *graus de liberdade* associados a cada tipo de efeito são dados por:

- $g.l.(SQA) = a - 1$, o número de parâmetros associados aos efeitos de nível de A.
- $g.l.[SQB(A)] = \sum_{i=1}^{a} (b_i - 1)$, o número de parâmetros associados aos efeitos de nível de B.
- $g.l.(SQRE) = n - \sum_{i=1}^{a} b_i$, o número de observações menos o número total de parâmetros do modelo.

4.7.5.2 Quadro-resumo da ANOVA a 2 Factores hierarquizados

Fonte	g.l.	SQ	QM	f_{calc}
Factor A	$a-1$	$SQA = SQF_A = SQT - SQRE_A$	$QMA = \frac{SQA}{a-1}$	$\frac{QMA}{QMRE}$
Factor B(A)	$\sum_{i=1}^{a}(b_i-1)$	$SQB(A) = SQRE_A - SQRE_{A/B}$	$QMB(A) = \frac{SQB(A)}{\sum_{i=1}^{a}(b_i-1)}$	$\frac{QMB(A)}{QMRE}$
Resíduos	$n - \sum_{i=1}^{a} b_i$	$SQRE = SQRE_{A/B}$	$QMRE = \frac{SQRE}{n-\sum_{i=1}^{a} b_i}$	
Total	$n-1$	$SQT = (n-1)\,S_y^2$	–	–

4.7.5.3 O Teste F aos efeitos do factor A (dominante)

Sendo válido o Modelo de ANOVA a 2 factores hierarquizados, tem-se o Teste F aos efeitos do factor A (dominante)

Hipóteses: $H_0 : \alpha_i = 0\ , \quad \forall\, i = 2, ..., a$ vs. $H_1 : \exists\, i = 2, .., a$ tal que $\alpha_i \neq 0$.
[FACTOR A NÃO AFECTA Y] vs. [FACTOR A AFECTA Y]

Estatística do Teste: $F = \frac{QMA}{QMRE} \underset{\cdot}{\sim} F_{(a-1\,,\, n-\sum_i b_i)}$ se H_0.

Nível de significância do teste: α

Região Crítica (Unilateral direita): Rejeitar H_0 se $F_{calc} > f_{\alpha(\,a-1\,,\, n-\sum_i b_i)}$

4.7.5.4 O Teste F aos efeitos do factor B (subordinado)

Sendo válido o Modelo de ANOVA a dois factores hierarquizado, tem-se o Teste F aos efeitos do factor B (subordinado):

Hipóteses: $H_0 : \beta_{j(i)} = 0\ , \quad \forall\, j = 2, ..., b_i\ ,\ i = 1, ..., a$ vs. $H_1 : \exists\, i, j$ tais que $\beta_{j(i)} \neq 0$.
[FACTOR B NÃO AFECTA Y] vs. [FACTOR B AFECTA Y]

Estatística do Teste: $F = \frac{QMB(A)}{QMRE} \underset{\cdot}{\sim} F_{(\sum_i (b_i-1)\,,\, n-\sum_i b_i)}$ se H_0.

Nível de significância do teste: α

Região Crítica (Unilateral direita): Rejeitar H_0 se $F_{calc} > f_{\alpha(\sum_i (b_i-1)\,,\, n-\sum_i b_i)}$.

4.7.6 ANOVA a dois Factores hierarquizados no R

Para efectuar uma ANOVA a dois Factores hierarquizados no R, organizam-se os dados como nos anteriores modelos com dois factores, ou seja, numa `data.frame` com três colunas:

1. uma coluna com a variável resposta;

2. outra coluna com o factor A;
3. uma terceira coluna com o factor B.

A *fórmula* utilizada no `R` para indicar uma ANOVA a dois Factores hierarquizados é semelhante às anteriores, mas com o nome dos dois factores separado pelo símbolo '`/`'. Se o factor dominante tem nome fA, a fórmula terá o seguinte aspecto:

```
y  ~  fA / fB
```

No exemplo de tractores/tractoristas, a tabela-resumo produzida pelo comando `aov` é a seguinte:

```
> summary(aov(indice ~ tractor/tractorista, data=tractores))
                    Df Sum Sq Mean Sq F value   Pr(>F)
tractor              2   1696   847.8   35.92 2.90e-10 ***
tractor:tractorista  9   2272   252.5   10.70 6.99e-09 ***
Residuals           48   1133    23.6
```

Neste caso, há efeitos significativos dos diferentes tipos de tractores sobre a variável resposta, e também efeitos significativos dos tractoristas que conduzem os tractores.

4.7.7 Comparações múltiplas de médias

Caso se conclua pela existência de efeitos do factor subordinado, é natural querer comparar médias da variável resposta nas $\sum_{j=1}^{a} b_i$ diferentes situações experimentais.

As *comparações múltiplas de Tukey* podem ser efectuadas, caso o delineamento seja *equilibrado*, isto é, se houver o mesmo número n_c de observações em cada situação experimental. A ideia é semelhante à das aplicações da teoria de Tukey feita em modelos anteriores.

Neste caso, os parâmetros da distribuição de Tukey serão

- o número de situações experimentais, $k = \sum_{i=1}^{a} b_i$; e
- os graus de liberdade associados ao $QMRE$, $\nu = n - \sum_{i=1}^{a} b_i$.

Numa abordagem de tipo Testes de Hipóteses, duas médias populacionais de diferentes situações experimentais, μ_{ij} e $\mu_{i'j'}$, devem ser consideradas diferentes caso as respectivas médias amostrais, $\overline{Y}_{ij.}$ e $\overline{Y}_{i'j'.}$ difiram em mais do que o termo de comparação, ou seja, se

$$|\overline{Y}_{ij.} - \overline{Y}_{i'j'.}| > q_{\alpha\left(\sum_{i=1}^{a} b_i \ , \ n-\sum_{i=1}^{a} b_i\right)} \sqrt{\frac{QMRE}{n_c}} \ . \tag{4.35}$$

Numa abordagem do tipo Intervalos de Confiança, cada um dos intervalos é centrado na respectiva diferença de médias amostrais, às quais se subtrai, e soma, o termo de comparação, para alcançar os extremos do intervalo.

4.7.8 Análise de resíduos

Também no que respeita à análise de resíduos para validar os pressupostos do modelo, a situação é análoga à vista no estudo de anteriores modelos.

Pode efectuar-se um *teste de Bartlett* para testar a hipótese que as variâncias populacionais são iguais em cada uma das $k = \sum_{i=1}^{a} b_i$ diferentes situações experimentais. A estatística de teste e os graus de liberdade da respectiva distribuição assintótica são iguais aos casos anteriores, com este valor de k. **O teste de Bartlett não é matéria para avaliação**.

4.8 Comentários finais sobre ANOVA

4.8.1 ANOVAs como comparação de k amostras

Alguns testes F ANOVA generalizam os testes t de comparação de médias de duas populações, estudados nas disciplinas introdutórias de Estatística, onde são dados testes t para comparar as médias de duas populações,

- através de *amostras independentes* (admitindo a igualdade de variâncias); e
- com *amostras emparelhadas*.

Ora, verifica-se que:

- O *quadrado da estatística* t à diferença de médias, no caso de *amostras independentes*, é a *estatística* F do teste aos efeitos do factor, num *modelo ANOVA a 1 Factor com* $k = 2$ *níveis*.
- O *quadrado da estatística* t à diferença de médias, no caso de *amostras emparelhadas*, é a *estatística* F do teste aos efeitos do Factor, num *modelo ANOVA a dois factores* - um dos quais introduzido para definir o emparelhamento das unidades experimentais - *sem interacção* e *com uma única observação por célula*, *quando* $a = 2$.

4.8.2 Comparações múltiplas alternativas na ANOVA

A comparação múltipla de médias, que abordámos pela teoria de Tukey, tem numerosas alternativas.

A alternativa mais conceituada baseia-se na teoria de Scheffé [6]. Tem tendência a produzir intervalos de confiança maiores (ao mesmo nível $(1 - \alpha) \times 100\%$ de confiança) do que os intervalos de Tukey.

Quer Tukey, quer Scheffé, podem ser generalizados para obter testes/intervalos de confiança sobre *combinações lineares genéricas das médias* de nível ou de células. Nesse caso, a teoria de Scheffé tem melhor desempenho.

4.8.3 Delineamentos factoriais com vários factores

Um delineamento factorial (isto é, com observações para todas as combinações de níveis de cada factor) pode ser definido com qualquer número de factores.

Num delineamento *factorial a três factores* – designados por A, B e C – cada observação da variável resposta indexa-se com *quatro índices*: Y_{ijkl} indica a observação l no nível i do Factor A, nível j do Factor B e nível k do Factor C. A equação de base para Y_{ijkl} prevê a existência de *sete tipos de efeitos*:

- três *efeitos principais de cada factor*, α_i, β_j e γ_k.
- três *efeitos de interacção dupla* associados a cada combinação de níveis de dois Factores diferentes: $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$ e $(\beta\gamma)_{jk}$.
- um *efeito de tripla interacção* para as células onde se cruzam níveis dos três factores: $(\alpha\beta\gamma)_{ijk}$.

O modelo factorial a três factores tem assim a seguinte equação de base do modelo:

$$Y_{ijkl} \;=\; \mu_{111} + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl} \;,$$

excluíndo-se efeitos sempre que um dos índices fôr 1.

Este modelo tem um total de abc parâmetros. A Soma de Quadrados Total é agora decomposta em *oito parcelas*: SQA, SQB, SQC, $SQAB$, $SQAC$, $SQBC$, $SQABC$ e $SQRE$. As sete SQs associadas a efeitos são definidas pela diferença das Somas de Quadrados Residuais de modelos onde se vão sucessivamente omitindo os efeitos correspondentes.

Os *graus de liberdade* associados a cada tipo de efeito generalizam conceitos anteriores:

- Para as SQs de efeitos principais de factor, são os números de níveis, menos um: $a-1$, $b-1$ e $c-1$.
- para as interacções duplas, são o produto dos graus de liberdade de cada factor: $(a-1)(b-1)$, $(a-1)(c-1)$ e $(b-1)(c-1)$.
- para as interacções triplas, são o produto dos graus de liberdade dos três efeitos principais: $(a-1)(b-1)(c-1)$.
- para o residual, o número de observações menos o número de parâmetros, $n-abc$.

Haverá agora *sete Testes de Hipóteses*: um para cada tipo de efeitos. As estatísticas desses sete testes são todas do tipo $\frac{QMx}{QMRE}$, onde x designa o tipo de efeitos em questão. As estatísticas desses testes terão, caso seja verdade a respectiva Hipótese Nula H_0, distribuição F com graus de liberdade dados pelos g.l. do numerador e do denominador, respectivamente.

4.8.4 Outros tipos de delineamentos experimentais

Apenas foi aflorada a teoria dos delineamentos experimentais. Existem numerosos outros delineamentos mais complexos.

Alguns delineamentos visam reduzir o número de situações experimentais que seria necessário estudar (objectivo que também pode motivar um delineamento hierarquizado). Entre estes, refiram-se:

- os *quadrados latinos*; ou
- os *delineamentos em blocos incompletos.*

Outros delineamentos visam ultrapassar dificuldades práticas na execução de uma experiência, como é o caso dos delineamentos em *parcelas divididas (split plots, em inglês).*

4.8.5 Métodos não paramétricos de tipo ANOVA

Uma forma alternativa de estudar problemas análogos aos objectivos de ANOVAs resulta da utilização de *métodos não paramétricos.* Nestes métodos, não se exigem pressupostos tão fortes como os métodos clássicos, (por exemplo, o pressuposto de normalidade). A sua maior generalidade tem como contrapartida uma menor capacidade de rejeitar as hipóteses nulas caso elas sejam falsas (i.e., têm menor *potência*), quando os pressupostos adicionais dos métodos clássicos são válidos.

Com grande frequência, embora nem sempre, os métodos não paramétricos substituem os valores observados da variável resposta pelas *ordens (ranks)* dessas observações. As estatísticas de teste são então funções dessas ordens.

O **teste de Kruskal-Wallis** é uma *alternativa não paramétrica à ANOVA a 1 Factor*, em que:

- Cada observação é substituída pela sua ordem;
- A estatística de teste compara as ordens médias em cada nível do factor com a ordem média global.
- A hipótese nula é que nos vários níveis do factor as observações seguem a mesma distribuição.
- A hipótese alternativa é que a distribuição dos vários níveis difere apenas nas suas localizações (medianas).

O **teste de Friedman** é uma *alternativa não paramétrica à ANOVA com um factor e blocos*, ou seja, *a dois factores, sem interacção, nem repetições nas células*, em que:

- Cada observação é substituída pela sua ordem *no seio do seu bloco*;
- A estatística de teste compara as ordens médias em cada nível do factor com a ordem média global.
- A hipótese nula é que nos vários níveis do factor as observações seguem a mesma distribuição, excepto devido a translações associadas a cada bloco.
- A hipótese alternativa é que a distribuição dos vários níveis difere também devido a translações associadas aos níveis do factor.

Em ambos estes casos, as estatísticas de teste são funções das Somas de Quadrados usuais, *aplicadas às ordens*, em vez de aos valores observados de Y.

Os métodos não paramétricos são uma alternativa viável quando haja violação grave dos pressupostos dos modelos ANOVA clássicos. No entanto, para delineamentos mais complexos a existência de métodos não paramétricos é menos frequente.

4.8.6 Efeitos aleatórios em modelos tipo ANOVA

Nos modelos ANOVA estudados até aqui, admitiu-se sempre que as parcelas de efeitos nas equações dos modelos eram *constantes*. Este tipo de modelos dizem-se *de efeitos fixos*. Uma outra grande classe de modelos alternativos designam-se *modelos de efeitos aleatórios* ou ainda *modelos mistos*. Não sendo, em rigor, modelos lineares do tipo considerado até aqui, têm pontos de contacto importantes, em particular no caso dum modelo a um único factor.

Se um factor tem um número muito grande, ou mesmo uma infinidade, de possíveis níveis, não sendo possível estudar todos, pode optar-se por estudar apenas uma *amostra aleatória de níveis do factor*, na tentativa de extrair conclusões para o factor na sua totalidade.

Esta situação surge com frequência quando os níveis de um factor são terrenos, genótipos ou outras entidades para as quais se admite variabilidade, mas em que não é possível estudar a totalidade dos possíveis casos (níveis do factor). Nesses casos, *os efeitos dos níveis seleccionados aleatoriamente para o estudo* são melhor descritos por *variáveis aleatórias*, e não por constantes. *Efeitos de blocos, ou de factores hierarquizados subordinados* são, com muita frequência, mais correctamente descritos por efeitos aleatórios.

Por exemplo, a equação base de um *modelo a um factor com efeitos aleatórios*, com k níveis do factor, será

$$Y_{ij} \;=\; \mu + \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_{ij} \;,$$

sendo $\boldsymbol{\alpha}_i$ uma *variável aleatória que indica o efeito do nível que vier a ser aleatoriamente seleccionado como nível i* do factor.

A existência de novas variáveis aleatórias (além dos erros aleatórios) na equação de base de um modelo com efeitos aleatórios exige *novos pressupostos* para possibilitar o estudo do modelo. Os pressupostos usuais em modelos com efeitos aleatórios são que os efeitos aleatórios do tipo $\boldsymbol{\alpha}_i$:

- têm distribuição Normal;
- têm média zero;
- têm variância σ^2_α;
- são independentes entre si e independentes dos erros aleatórios.

Estas hipóteses correspondem a *admitir que a distribuição dos efeitos de nível do factor é* $\boldsymbol{\alpha}_i \frown \mathcal{N}(0, \sigma^2_\alpha)$ e que os níveis amostrados são seleccionados de forma independente.

Um *teste à existência de efeitos do factor* tem as hipóteses:

$$H_0 :\; \sigma^2_\alpha \;=\; 0 \qquad \text{vs.} \qquad H_1 : \sigma^2_\alpha \neq 0 \;.$$

Sendo verdade a Hipótese Nula, a variável aleatória $\boldsymbol{\alpha}_i$ toma sempre o valor nulo, e diz-se que o Factor A não afecta a variável resposta. Em caso de rejeição de H_0, a favor de $H_1 : \sigma^2_\alpha \neq 0$, a variável aleatória $\boldsymbol{\alpha}_i$ toma diferentes valores, e conclui-se que o factor afecta a variável resposta.

Embora este modelo a um factor não seja um Modelo Linear do mesmo tipo que o modelo de efeitos fixos antes estudado, o teste envolve uma estatística equivalente.

Em geral, com delineamentos mais complexos, testes à existência de efeitos aleatórios envolvem quocientes de Quadrados Médios, com distribuição F sob H_0, mas nem sempre as estatísticas dos testes são iguais aos correspondentes casos de efeitos fixos.

Podem ser considerados modelos com vários factores em que todos, ou apenas alguns, são de efeitos aleatórios. *Um modelo com factores de efeitos fixos e outros de efeitos aleatórios diz-se um* **modelo misto**.

Bibliografia

[1] R.A. Becker, J.M. Chambers, and A.R. Wilks. *The S Language*. Wadsworth & Brooks/Cole, 1988.

[2] J.M. Chambers and T. Hastie. *Statistical Models in S*. Wadsworth & Brooks/Cole, 1992.

[3] F. Galton. Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.

[4] Manuela Neves. *Introdução à Estatística e à Probabilidade com utilização do R*. ISA Press, 2017.

[5] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[6] H. Scheffé. *The Analysis of Variance*. John Wiley & Sons, 1959. COTA BISA: U10 484.