

## II.2. Regressão Linear Múltipla

Por vezes, é necessário **mais do que uma variável preditora** para modelar a variável resposta de interesse.

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

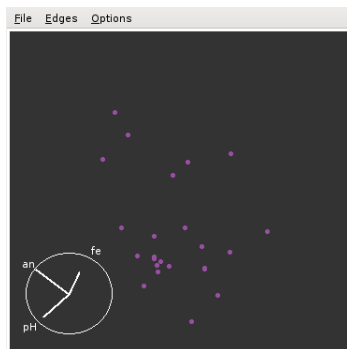
- teor de antocianas (variável **antoci**, em  $mg/dm^3$ );
- fenóis totais (variável **fentot**);
- pH (variável **pH**).

Há interesse em estudar a relação entre o teor de antocianas (variável resposta) e o teor de fenóis totais e pH.

# A nuvem de pontos - uma perspectiva

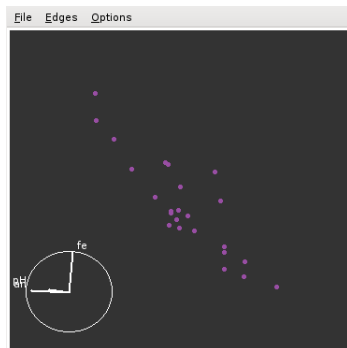
As  $n = 24$  observações em três variáveis descrevem agora uma nuvem de 24 pontos em  $\mathbb{R}^3$ .

Neste ângulo de visão, a nuvem de pontos em  $\mathbb{R}^3$  nada tem de especial.



# A nuvem de pontos - outra perspectiva

Noutro ângulo de visão percebe-se que os pontos se dispersam aproximadamente em torno de **um plano**.



## Plano em $\mathbb{R}^3$

Qualquer plano em  $\mathbb{R}^3$ , no sistema  $xOyz$ , tem equação

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

No nosso contexto, e colocando:

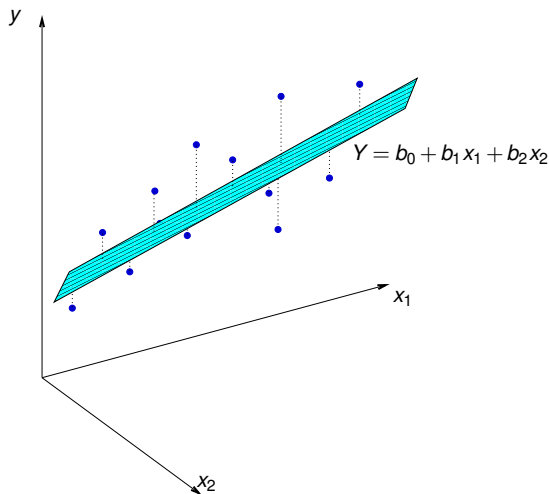
- no eixo vertical ( $z$ ) a variável resposta  $Y$ ;
- noutro eixo ( $x$ ) um preditor  $X_1$ ;
- no terceiro eixo ( $y$ ) o outro preditor  $X_2$ ,

A equação fica (no caso geral de planos não verticais, com  $C \neq 0$ ):

$$\begin{aligned} Ax_1 + Bx_2 + Cy + D = 0 &\Leftrightarrow y = -\frac{D}{C} - \frac{A}{C}x_1 - \frac{B}{C}x_2 \\ &\Leftrightarrow y = b_0 + b_1x_1 + b_2x_2 \end{aligned}$$

Esta equação generaliza a equação da recta, para o caso de haver dois preditores.

## Regressão Múltipla - representação gráfica ( $p = 2$ )



$Y = b_0 + b_1x_1 + b_2x_2$  é a equação dum plano em  $\mathbb{R}^3$  ( $x_1 \geq 0, x_2 \geq 0, y$ ).

Podem ser ajustados pelo mesmo critério que na RLS: minimizar SQRE.

## O caso geral: $p$ preditores

Caso se pretenda modelar uma variável resposta,  $Y$ , com base em  $p$  **variáveis preditoras**,  $x_1, x_2, \dots, x_p$ , 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**:

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_px_p .$$

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

Tal como na Regressão Linear Simples, **admite-se que dispomos de  $n$  conjuntos de observações para ajustar este hiperplano**:

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

# As dificuldades na representação gráfica

A representação gráfica usual da nuvem de  $n$  pontos observados exige  $p + 1$  eixos: um para  $Y$  e um para cada um dos  $p$  preditores.

Para  $p > 2$ , são necessários mais de três eixos e a visualização torna-se impossível.

As características fundamentais dessas representações são:

- Existem  $p + 1$  eixos – um para cada variável em questão.
- Existem  $n$  pontos – um para cada indivíduo (unidade experimental) observado.
- Tem-se uma nuvem de  $n$  pontos num espaço  $(p + 1)$ -dimensional.

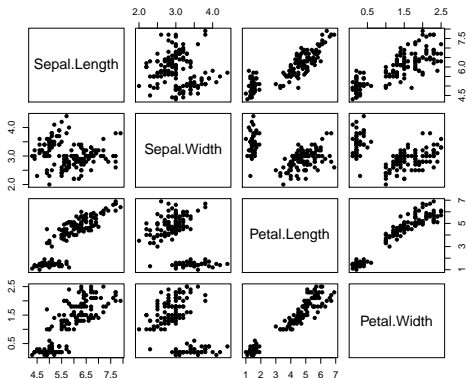
Na regressão linear múltipla admite-se que os pontos se dispõem em torno de um hiperplano em  $\mathbb{R}^{p+1}$ , de equação

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_px_p.$$

# Visualizações parciais da nuvem de pontos

A impossibilidade de visualizar as nuvens de  $n$  pontos em  $\mathbb{R}^{p+1}$  sugere a consideração de **visões parciais**, como sejam as nuvens de pontos definidas por cada par de variáveis.

E.g., para as  $n = 150$  observações de lírios em 4 variáveis:





# Advertência

Estes gráficos de pares de variáveis são as **projeções ortogonais** da nuvem de  $n$  pontos em cada plano coordenado de  $\mathbb{R}^{p+1}$ .

A projecção da nuvem de  $n$  pontos nos planos coordenados não é uma solução ideal.

Em particular, 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  $y$  vs. um preditor individual,  $x_j$ .

## Outra representação gráfica

A representação gráfica em  $\mathbb{R}^{p+1}$  das  $n$  observações de  $Y$  e as  $p$  variáveis preditoras não é a única possível.

Há **outra representação possível** dos dados, que **casa conceitos geométricos e conceitos estatísticos** e é útil na determinação dos parâmetros ajustados.

As  $n$  observações de  $Y$  definem um **vector em  $\mathbb{R}^n$** :

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

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

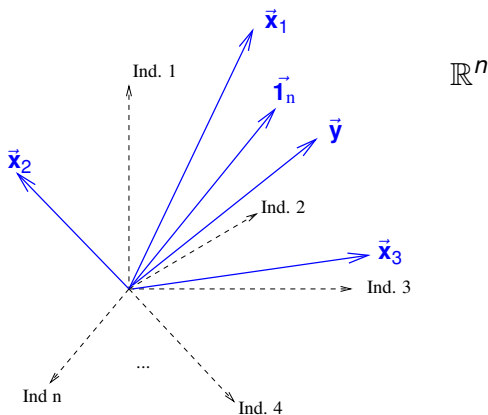
$$\vec{x}_j = (x_{j(1)}, x_{j(2)}, x_{j(3)}, \dots, x_{j(n)}) \quad (j = 1, 2, \dots, p).$$

Podemos representar todas as variáveis por vectores **em  $\mathbb{R}^n$** .

# A representação em $\mathbb{R}^n$ , o espaço das variáveis

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

O vector de  $n$  uns, representado por  $\vec{1}_n$ , também é útil.



# Vantagens da representação gráfica alternativa

Os  $n$  valores ajustados  $\vec{y}_j$  também definem um vector de  $\mathbb{R}^n$ ,  $\vec{\hat{y}}$ , que é uma combinação linear dos vectores  $\vec{1}_n, \vec{x}_1, \vec{x}_2, \dots, \vec{x}_p$ :

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

# A matriz $\mathbf{X}$ e o seu subespaço de colunas

Recordemos alguns conceitos dados na UC **Álgebra Linear** dos 1os. ciclos do ISA.

- O conjunto de **todas** as combinações lineares de  $p+1$  **vetores**  $\vec{\mathbf{1}}_n, \vec{\mathbf{x}}_1, \dots, \vec{\mathbf{x}}_p$  chama-se o **subespaço gerado** por esses vetores.
- Colocando os vetores  $\vec{\mathbf{1}}_n, \vec{\mathbf{x}}_1, \dots, \vec{\mathbf{x}}_p$  nas **colunas** duma **matriz  $\mathbf{X}$** , de dimensão  $n \times (p+1)$ , podemos chamar a este subespaço o **subespaço das colunas da matriz  $\mathbf{X}$** ,  $\mathcal{C}(\mathbf{X}) \subset \mathbb{R}^n$ .
- É um subespaço de **dimensão  $p+1$**  (se os vetores forem **linearmente independentes**, isto é, nenhum se pode escrever como combinação linear dos restantes).
- **Qualquer combinação linear** dos vetores coluna da matriz  $\mathbf{X}$  é dada por  $\mathbf{X}\vec{\mathbf{a}}$ , onde  $\vec{\mathbf{a}} = (a_0, a_1, a_2, \dots, a_p)$  é o **vector dos coeficientes** que define a combinação linear.

## Um produto matricial $\mathbf{X}\vec{\mathbf{a}}$

O produto da matriz  $\mathbf{X}_{n \times (p+1)}$  por um vector  $\vec{\mathbf{a}} \in \mathbb{R}^{p+1}$  é uma combinação linear das colunas de  $\mathbf{X}$ :

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

# Os parâmetros

- Cada escolha possível de coeficientes  $\vec{\mathbf{a}} = (a_0, a_1, a_2, \dots, a_p)$  corresponde a um ponto/vector 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  $\vec{\mathbf{x}}_1, \dots, \vec{\mathbf{x}}_p, \vec{\mathbf{1}}_n$ .
- Um dos pontos/vectores do subespaço é a combinação linear dada pelo vector de coeficientes  $\vec{\mathbf{b}} = (b_0, b_1, \dots, b_p)$  que minimiza:

$$SQRE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

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

Como identificar esse ponto/vector?

# Geometria

Vamos usar argumentos geométricos.

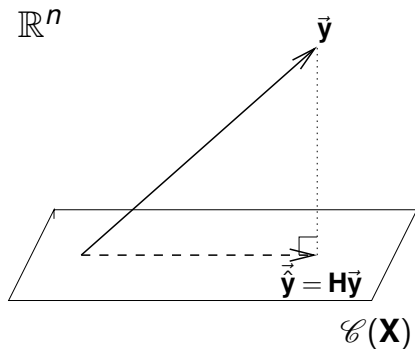
- Dispomos de um vector de  $n$  observações de  $\vec{y}$  que está em  $\mathbb{R}^n$  mas, em geral, não está no subespaço  $\mathcal{C}(\mathbf{X})$ .
- Queremos aproximar esse vector por outro vector,  $\vec{\hat{y}} = b_0 \vec{1}_n + b_1 \vec{x}_1 + \dots + b_p \vec{x}_p$ , que está no subespaço  $\mathcal{C}(\mathbf{X})$ .
- Vamos aproximar o vector de observações  $\vec{y}$  pelo vector  $\vec{\hat{y}}$  do subespaço  $\mathcal{C}(\mathbf{X})$  que está mais próximo de  $\vec{y}$ .

## SOLUÇÃO:

Tomar a projecção ortogonal de  $\vec{y}$  sobre  $\mathcal{C}(\mathbf{X})$ :  $\vec{\hat{y}} = \mathbf{H}\vec{y}$ .



## A projecção ortogonal de $\vec{y}$ sobre $\mathcal{C}(\mathbf{X})$



O vector de  $\mathcal{C}(\mathbf{X}) \subset \mathbb{R}^n$  mais próximo dum vector  $\vec{y} \in \mathbb{R}^n$  é o vector  $\hat{\vec{y}}$  que resulta de projectar ortogonalmente  $\vec{y}$  sobre  $\mathcal{C}(\mathbf{X})$ .

## O critério minimiza *SQRE*

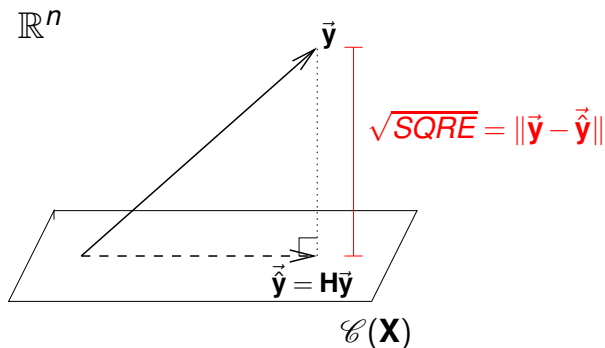
O vector  $\vec{\hat{y}}$  que minimiza a distância ao vector de observações  $\vec{y}$  minimiza também o **quadrado dessa distância**, que é dado por:

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

Ou seja, o critério **minimiza a soma de quadrados dos resíduos**.

Trata-se do **mesmo critério** que foi usado na Regressão Linear Simples.

# SQRE na projecção ortogonal



O quadrado da distância de  $\vec{y}$  a  $\vec{\hat{y}}$  é  $SQRE$ , a soma dos quadrados dos resíduos.

## A projecção ortogonal

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

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

Logo, temos:

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

A combinação linear dos vectores  $\vec{\mathbf{1}}_n, \vec{\mathbf{x}}_1, \dots, \vec{\mathbf{x}}_p$  que gera o vector mais próximo de  $\vec{y}$  tem coeficientes dados pelos elementos do vector  $\vec{\mathbf{b}}$ :

Os parâmetros ajustados na RL Múltipla

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

## 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:

**SQRE** – Soma de Quadrados dos Resíduos (já definida):

$$SQRE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 .$$

**SQT** – Soma de Quadrados Total:

$$SQT = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2 .$$

**SQR** – Soma de Quadrados associada à Regressão:

$$SQR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^n \hat{y}_i^2 - n\bar{y}^2 .$$

Nota: Também aqui os  $y$  observados ( $y_i$ ) e os  $y$  ajustados ( $\hat{y}_i$ ) têm a mesma média (ver Exercício 4 da RLM).

# Pitágoras e a Regressão

O Teorema de Pitágoras aplica-se em qualquer espaço euclidiano  $\mathbb{R}^n$ .

Aplicado ao triângulo rectângulo do acetato 214 produz a seguinte relação:

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

## Revisitando Pitágoras

Vimos que a relação fundamental da Regressão Linear ( $SQT = SQR + SQRE$ ) resulta duma aplicação do Teorema de Pitágoras. Mas foi necessário introduzir a subtracção de  $n\bar{y}^2$ .

Um outro triângulo rectângulo é estatisticamente mais interessante.

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

$$\vec{y}^c = \vec{y} - \bar{y}\mathbf{1}_n = (y_1 - \bar{y}, y_2 - \bar{y}, \dots, y_n - \bar{y})^t.$$

A norma deste vector é  $\|\vec{y}^c\| = \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2} = \sqrt{SQT}$ .

## Revisitando Pitágoras (cont.)

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

$$\begin{aligned}\mathbf{H}\vec{y}^c &= \mathbf{H}(\vec{y} - \bar{y}\vec{1}_n) \\ \Leftrightarrow \mathbf{H}\vec{y}^c &= \mathbf{H}\vec{y} - \bar{y}\mathbf{H}\vec{1}_n \\ \Leftrightarrow \mathbf{H}\vec{y}^c &= \vec{y} - \bar{y}\vec{1}_n\end{aligned}$$

já que  $\mathbf{H}\vec{1}_n = \vec{1}_n$ , pois o vector  $\vec{1}_n$  já pertence ao subespaço  $\mathcal{C}(\mathbf{X})$ , logo fica invariante quando projectado nesse mesmo subespaço (ver exercício 4b).

O vector  $\mathbf{H}\vec{y}^c$  tem elemento genérico  $\hat{y}_i - \bar{y}$ , e a sua norma é

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



## Revisitando Pitágoras (cont.)

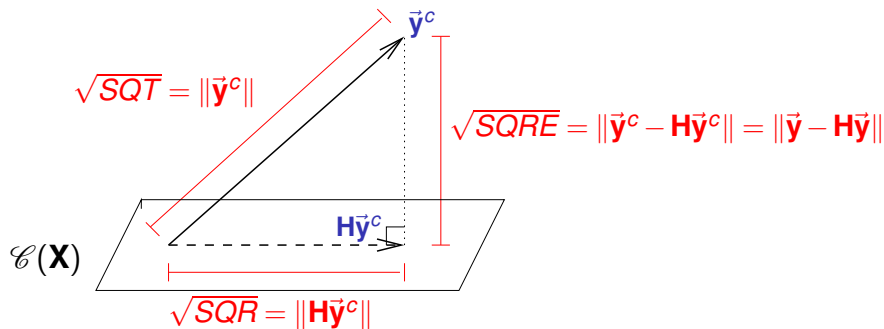
A distância entre o vector  $\vec{y}^c$  e a sua projecção ortogonal sobre  $\mathcal{C}(\mathbf{X})$  continua a ser  $\sqrt{SQRE}$ :

$$\begin{aligned}\vec{y}^c - \mathbf{H}\vec{y}^c &= (\vec{y} - \cancel{\bar{y}\vec{1}_n}) - (\hat{\vec{y}} - \cancel{\bar{y}\vec{1}_n}) \\ \Leftrightarrow \vec{y}^c - \mathbf{H}\vec{y}^c &= \vec{y} - \hat{\vec{y}}\end{aligned}$$

pelo que

$$\|\vec{y}^c - \mathbf{H}\vec{y}^c\| = \|\vec{y} - \hat{\vec{y}}\| = \sqrt{\sum_{i=1}^n (y_i - \hat{y}_i)^2} = \sqrt{SQRE} .$$

## Revisitando Pitágoras (cont.)

 $\mathbb{R}^n$ 

A fórmula fundamental da Regressão Linear,  $SQT = SQR + SQRE$ , é uma aplicação directa do Teorema de Pitágoras ao triângulo definido por  $\vec{y}^c$  e a sua projecção ortogonal sobre  $\mathcal{C}(\mathbf{X})$ .

# Pitágoras e o Coeficiente de Determinação

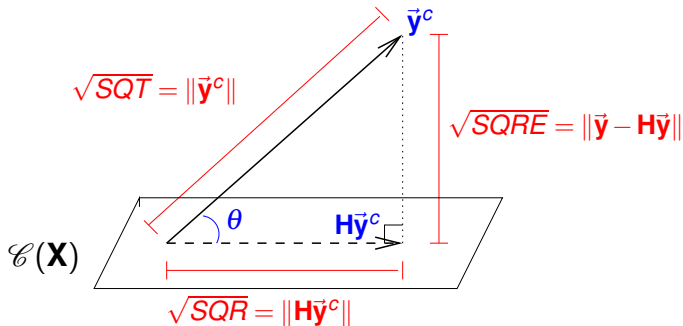
O acetato 221 torna evidente outra relação importante entre a geometria e a estatística da Regressão Linear:

Definindo o **coeficiente de determinação** da forma usual,  $R^2 = \frac{SQR}{SQT}$ , este resulta ser o cosseno ao quadrado do ângulo entre o vector centrado das observações da variável resposta,  $\vec{y}^c$ , e a sua projecção ortogonal sobre o subespaço  $\mathcal{C}(\mathbf{X})$ :

$$\cos^2(\theta) = \frac{SQR}{SQT} = R^2,$$

onde  $\theta$  é o ângulo entre os vectores  $\vec{y}^c$  e  $\mathbf{H}\vec{y}^c$ .

# Pitágoras e o Coeficiente de Determinação (cont.)

 $\mathbb{R}^n$ 

O Coeficiente de Determinação na Regressão Linear,  $R^2 = \frac{SQR}{SQT}$ , é o cosseno ao quadrado do ângulo entre  $\vec{y}^c$  e  $H\vec{y}^c$ .

# Propriedades do Coeficiente de Determinação

A abordagem geométrica confirma que, também na Regressão Linear Múltipla, são válidas as propriedades (já conhecidas da Regressão Linear Simples) do Coeficiente de Determinação:


- $R^2$  toma valores entre 0 e 1.
- Quanto mais próximo de 1 estiver  $R^2$ , menor o ângulo  $\theta$ , e portanto melhor será a correspondência entre o vector (centrado) das observações,  $\vec{y}^c$ , e o seu ajustamento em  $\mathcal{L}(\mathbf{X})$ .
- Se  $R^2 \approx 0$ , o vector  $\vec{y}^c$  é quase perpendicular ao subespaço  $\mathcal{L}(\mathbf{X})$  onde se pretende aproximá-lo, e a projecção vai quase anular todas os elementos do vector projectado. O resultado será de má qualidade, uma vez que se perde quase toda a variabilidade nos valores de  $y$ .

# Algumas propriedades dos hiperplanos ajustados

Numa regressão linear múltipla verifica-se:

- O hiperplano ajustado em  $\mathcal{R}^{p+1}$  contém o centro de gravidade da nuvem de pontos, i.e., o ponto de coordenadas  $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p, \bar{y})$ .
- a média dos valores observados de  $Y$ ,  $\{y_i\}_{i=1}^n$ , é igual à média dos respectivos valores ajustados,  $\{\hat{y}_i\}_{i=1}^n$  (ver Exercício RLM 4c).
- 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.
- 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.

## A Regressão Múltipla no

Uma Regressão Múltipla no  estuda-se através do mesmo comando `lm` usado para a regressão linear simples. A indicação de qual a variável resposta  $y$  e quais as variáveis preditoras  $x_1, \dots, x_p$  faz-se de forma semelhante à da RLS.

Por exemplo, se a variável resposta se chama  $y$  e existirem três preditores de nome  $x_1$ ,  $x_2$  e  $x_3$ , a fórmula que indica a relação será:

$$y \sim x_1 + x_2 + x_3$$

O comando correspondente no R será:

```
> lm ( y ~ x1 + x2 + x3 , data=dados)
```

O resultado produzido por este comando será o **vector das estimativas dos  $p+1$  parâmetros do modelo,  $b_0, b_1, \dots, b_p$ .**

## A Regressão Múltipla no (cont.)

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 é:

$$PW = -0.2403 + 0.5241 PL - 0.2073 SL + 0.2228 SW$$

O coeficiente de determinação é  $R^2 = 0.9379$ , só ligeiramente maior que o valor  $R^2 = 0.9271$  do modelo RLS (acetato 172).



## O contexto inferencial

Até aqui, apenas se considerou o **problema descritivo**:

dados  $n$  conjuntos de observações  $\{(x_{1(i)}, x_{2(i)}, \dots, x_{p(i)}, y_{(i)})\}_{i=1}^n$ ,  
determinar os  $p+1$  coeficientes  $\vec{\mathbf{b}} = (b_0, b_1, b_2, \dots, b_p)$  que  
minimizam a soma de quadrados de resíduos

$$SQRE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n [y_i - (b_0 + b_1 x_{1(i)} + b_2 x_{2(i)} + \dots + b_p x_{p(i)})]^2$$

$$\Rightarrow \text{SQRE mínimo se } \vec{\mathbf{b}} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \vec{\mathbf{y}}.$$

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**.

# O Modelo RLM

Na Regressão Linear Múltipla admite-se que as  $n$  observações da variável resposta  $Y$  são aleatórias e podem ser modeladas como

$$Y_i = \beta_0 + \beta_1 x_{1(i)} + \beta_2 x_{2(i)} + \dots + \beta_p x_{p(i)} + \varepsilon_i, \quad i = 1, \dots, n$$

Admitem-se válidos pressupostos semelhantes aos do modelo RLS:

## Definição (O Modelo da Regressão Linear Múltipla - RLM)

- 1  $Y_i = \beta_0 + \beta_1 x_{1(i)} + \beta_2 x_{2(i)} + \dots + \beta_p x_{p(i)} + \varepsilon_i, \quad \forall i = 1, \dots, n.$
- 2  $\varepsilon_i \cap \mathcal{N}(0, \sigma^2), \quad \forall i = 1, \dots, n.$
- 3  $\{\varepsilon_i\}_{i=1}^n$  v.a. independentes.

A constante  $\beta_j$  ( $j = 1, 2, \dots, p$ ) 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  $n$  equações do modelo, válidas para as  $n$  observações, podem ser escritas como uma única equação, utilizando notação vectorial/matricial:

$$\begin{array}{rcll} Y_1 & = & \beta_0 + \beta_1 x_{1(1)} + \beta_2 x_{2(1)} + \cdots + \beta_p x_{p(1)} & + \varepsilon_1 \\ Y_2 & = & \beta_0 + \beta_1 x_{1(2)} + \beta_2 x_{2(2)} + \cdots + \beta_p x_{p(2)} & + \varepsilon_2 \\ Y_3 & = & \beta_0 + \beta_1 x_{1(3)} + \beta_2 x_{2(3)} + \cdots + \beta_p x_{p(3)} & + \varepsilon_3 \\ \vdots & \vdots & \vdots & \vdots \\ Y_n & = & \beta_0 + \beta_1 x_{1(n)} + \beta_2 x_{2(n)} + \cdots + \beta_p x_{p(n)} & + \varepsilon_n \end{array}$$

## A notação matricial (cont.)

As  $n$  equações correspondem a **uma única equação matricial**:

$$\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\varepsilon},$$

onde

$$\vec{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & X_{1(1)} & X_{2(1)} & \cdots & X_{p(1)} \\ 1 & X_{1(2)} & X_{2(2)} & \cdots & X_{p(2)} \\ 1 & X_{1(3)} & X_{2(3)} & \cdots & X_{p(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1(n)} & X_{2(n)} & \cdots & X_{p(n)} \end{bmatrix}, \vec{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \vec{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Nesta equação,  $\vec{Y}$  e  $\vec{\varepsilon}$  são **vetores aleatórios**,  $\mathbf{X}$  é uma matriz não aleatória e  $\vec{\beta}$  um vector não-aleatório.

## A notação matricial (cont.)

Na equação matricial  $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$ , tem-se:

$\vec{Y}$  – vector aleatório das  $n$  variáveis aleatórias resposta;

$\mathbf{X}$  – 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 de uns, associada a constante aditiva do modelo);

$\vec{\beta}$  – vector (não aleatório) dos  $p + 1$  parâmetros do modelo;

$\vec{\epsilon}$  – vector aleatório dos  $n$  erros aleatórios.

Representa-se um vector de  $n$  observações de  $Y$  por  $\vec{y}$ .

Com alguns conceitos adicionais podemos escrever também os pressupostos relativos aos erros aleatórios em notação vectorial/matricial.

# Ferramentas para vectores aleatórios

O vector de  $n$  componentes  $\vec{Y}$ , tal como o vector dos  $n$  erros aleatórios,  $\vec{\epsilon}$ , constituem **vectores aleatórios**.

Para qualquer **vector aleatório**  $\vec{W} = (W_1, W_2, \dots, W_k)^t$ , define-se:

- O **vector esperado** de  $\vec{W}$ , constituído pelos **valores esperados de cada componente**:

$$E[\vec{W}] = \begin{bmatrix} E[W_1] \\ E[W_2] \\ \vdots \\ E[W_k] \end{bmatrix} .$$

## Ferramentas para vectores aleatórios (cont.)

- a **matriz de variâncias-covariâncias** de  $\vec{W}$  é constituída pelas (co)variâncias de cada par de componentes:

$$V[\vec{W}] = \begin{bmatrix} V[W_1] & C[W_1, W_2] & C[W_1, W_3] & \dots & C[W_1, W_k] \\ C[W_2, W_1] & V[W_2] & C[W_2, W_3] & \dots & C[W_2, W_k] \\ C[W_3, W_1] & C[W_3, W_2] & V[W_3] & \dots & C[W_3, W_k] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C[W_k, W_1] & C[W_k, W_2] & C[W_k, W_3] & \dots & V[W_k] \end{bmatrix}$$

É necessariamente uma **matriz simétrica**.

## Propriedades do vector esperado

Tal como para o caso de variáveis aleatórias, também o vector esperado de um vector aleatório  $\vec{W}_{k \times 1}$  tem propriedades simples:

- Se  $b$  é um escalar não aleatório,  $E[b\vec{W}] = b E[\vec{W}]$ .
- Se  $\vec{a}_{k \times 1}$  é um vector não aleatório,  $E[\vec{W} + \vec{a}] = E[\vec{W}] + \vec{a}$ .
- Se  $\vec{a}_{k \times 1}$  é um vector não aleatório,  $E[\vec{a}^t \vec{W}] = \vec{a}^t E[\vec{W}]$ .
- Se  $\mathbf{B}_{m \times k}$  é uma matriz não aleatória,  $E[\mathbf{B}\vec{W}] = \mathbf{B} E[\vec{W}]$ .

Também o vector esperado da soma de dois vectors aleatórios tem uma propriedade operatória simples:

- Se  $\vec{W}_{k \times 1}$ ,  $\vec{U}_{k \times 1}$  são vectores aleatórios,  $E[\vec{W} + \vec{U}] = E[\vec{W}] + E[\vec{U}]$ .



# Propriedades da matriz de (co)variâncias

Eis algumas propriedades operatórias das matrizes de variâncias-covariâncias de vectores aleatórios:

- Se  $b$  é um escalar não aleatório,  $V[b\vec{W}] = b^2 V[\vec{W}]$ .
- Se  $\vec{a}_{k \times 1}$  é um vector não aleatório,  $V[\vec{W} + \vec{a}] = V[\vec{W}]$ .
- Se  $\vec{a}_{k \times 1}$  é um vector não aleatório,  $V[\vec{a}^t \vec{W}] = \vec{a}^t V[\vec{W}] \vec{a}$ .
- Se  $\mathbf{B}_{m \times k}$  é uma matriz não aleatória,  $V[\mathbf{B}\vec{W}] = \mathbf{B} V[\vec{W}] \mathbf{B}^t$ .

A matriz de variâncias-covariâncias da soma de dois vectores aleatórios tem uma propriedade operatória simples se os vectores aleatórios forem independentes:

- Se  $\vec{W}_{k \times 1}$  e  $\vec{U}_{k \times 1}$  forem vectores aleatórios independentes,  $V[\vec{W} + \vec{U}] = V[\vec{W}] + V[\vec{U}]$ .

# A distribuição Normal Multivariada

Vectores aleatórios têm também distribuições (multivariadas) de probabilidades. Para vectores aleatórios contínuos  $\vec{W}_{k \times 1}$ , a distribuição pode ser caracterizada por uma função densidade conjunta  $f: \mathbb{R}^k \rightarrow \mathbb{R}$ . A mais frequente distribuição multivariada para vectores aleatórios é a **Multinormal**:

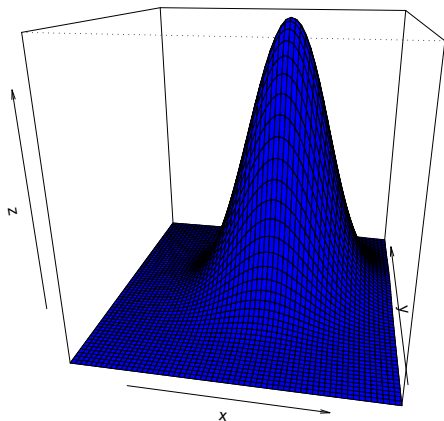
## Definição (Distribuição Normal Multivariada)

O vector aleatório  $k$ -dimensional  $\vec{W}$  tem **distribuição Multinormal**, com **parâmetros** dados pelo vector  $\vec{\mu}$  e a matriz  $\Sigma$  se a sua função densidade conjunta for:

$$f(\vec{w}) = \frac{1}{(2\pi)^{k/2} \sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(\vec{w}-\vec{\mu})^t \Sigma^{-1}(\vec{w}-\vec{\mu})}, \quad \vec{w} \in \mathbb{R}^k. \quad (3)$$

Notação:  $\vec{W} \sim \mathcal{N}_k(\vec{\mu}, \Sigma)$ .

## A densidade Binormal (Multinormal com $k = 2$ )



# Algumas propriedades da distribuição Multinormal

## Teorema (Propriedades da Multinormal)

Se  $\vec{W} \sim \mathcal{N}_k(\vec{\mu}, \Sigma)$ :

- 1 O vector esperado de  $\vec{W}$  é  $E[\vec{W}] = \vec{\mu}$ .
- 2 A matriz de (co)variâncias de  $\vec{W}$  é  $V[\vec{W}] = \Sigma$ .
- 3 Se duas componentes de  $\vec{W}$  têm covariância nula, são independentes:  $Cov[W_i, W_j] = 0 \Rightarrow W_i, W_j$  independentes.

**Nota:** Nas disciplinas introdutórias de Estatística dá-se que  $X, Y$  independentes  $\Rightarrow cov[X, Y] = 0$ . Agora sabemos que, quando a distribuição conjunta de  $X$  e  $Y$  é Multinormal, tem-se também a implicação contrária.

**Nota:** Qualquer elemento nulo numa matriz de (co)variâncias duma Multinormal indica que as componentes correspondentes são independentes.

## Propriedades da Multinormal (cont.)

### Teorema (Propriedades da Multinormal)

Se  $\vec{W} \cap \mathcal{N}_k(\vec{\mu}, \Sigma)$ :

- 4 Todas as distribuições marginais de  $\vec{W}$  são (multi)normais. Em particular, *cada componente  $W_i$  é normal com média  $\mu_i$  e variância  $\Sigma_{(i,i)}$* :  $W_i \cap \mathcal{N}(\mu_i, \Sigma_{(i,i)})$ .
- 5 Se  $\vec{a}$  um vector (não-aleatório)  $k \times 1$ , então  $\vec{W} + \vec{a} \cap \mathcal{N}_k(\vec{\mu} + \vec{a}, \Sigma)$ .
- 6 *Combinações lineares das componentes dum vector multinormal são Normais*:  $\vec{a}^t \vec{W} = a_1 W_1 + a_2 W_2 + \dots + a_k W_k \cap \mathcal{N}(\vec{a}^t \vec{\mu}, \vec{a}^t \Sigma \vec{a})$ .
- 7 Se  $\mathbf{B}$  é matriz  $m \times k$  (não aleatória, de característica  $m \leq k$ ), então  $\mathbf{B}\vec{W} \cap \mathcal{N}_m(\mathbf{B}\vec{\mu}, \mathbf{B}\Sigma\mathbf{B}^t)$ .

# Modelo Regressão Linear Múltipla - versão matricial

**Definição** (O Modelo em notação matricial)

1  $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\varepsilon}.$

2  $\vec{\varepsilon} \cap \mathcal{N}_n(\vec{0}, \sigma^2 \mathbf{I}_n),$  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  $\varepsilon_i$  tem distribuição Normal.
- Cada erro aleatório individual tem média zero:  $E[\varepsilon_i] = 0.$
- Cada erro aleatório individual tem variância igual:  $V[\varepsilon_i] = \sigma^2.$
- Erros aleatórios diferentes são independentes, porque  $Cov[\varepsilon_i, \varepsilon_j] = 0$  se  $i \neq j$  e, numa Multinormal, isso implica a independência.

# A distribuição das observações $\vec{Y}$ da variável resposta

O seguinte Teorema é consequência directa de aplicar as propriedades dos acetatos 239 e 240 ao Teorema do acetato 241.

## Teorema (Primeiras Consequências do Modelo)

*Dado o Modelo de Regressão Linear Múltipla, tem-se:*

$$\vec{Y} \sim \mathcal{N}_n(\mathbf{X}\vec{\beta}, \sigma^2 \mathbf{I}_n).$$

Tendo em conta as propriedades da Multinormal:

- Cada observação individual  $Y_i$  tem distribuição Normal.
- Cada observação individual  $Y_i$  tem média  
 $E[Y_i] = \beta_0 + \beta_1 x_{1(i)} + \beta_2 x_{2(i)} + \dots + \beta_p x_{p(i)}.$
- Cada observação individual tem variância igual:  $V[Y_i] = \sigma^2.$
- Observações diferentes de  $Y$  são independentes, porque  $Cov[Y_i, Y_j] = 0$  se  $i \neq j$  e, numa Multinormal, isso implica a independência.

## O estimador dos parâmetros do Modelo

Tal como na Regressão Linear Simples, os estimadores dos parâmetros  $\beta_j$  do modelo ( $j = 0, 1, 2, \dots, p$ ) obtêm-se adaptando a expressão matricial resultante de minimizar *SQRE* (acetato 215).

O vector  $\vec{\hat{\beta}}$  que estima o vector  $\vec{\beta}$  dos parâmetros populacionais é:

**Definição** (Estimador dos parâmetros populacionais)

$$\vec{\hat{\beta}} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \vec{\mathbf{Y}},$$

onde  $\mathbf{X}$  e  $\vec{\mathbf{Y}}$  são a matriz e o vector definidos no acetato 231.

O vector  $\vec{\hat{\beta}}$  é de dimensão  $p + 1$ . O seu primeiro elemento é o estimador de  $\beta_0$ , o seu segundo elemento é o estimador de  $\beta_1$ , etc..

Em geral, o estimador de  $\beta_j$  está na posição  $j + 1$  do vector  $\vec{\hat{\beta}}$ .



# A distribuição do vector de estimadores $\vec{\hat{\beta}}$

**Teorema** (Distribuição do estimador  $\vec{\hat{\beta}}$ )

*Dado o Modelo de Regressão Linear Múltipla, tem-se:*

$$\vec{\hat{\beta}} \sim \mathcal{N}_{p+1}(\vec{\beta}, \sigma^2 (\mathbf{X}^t \mathbf{X})^{-1}).$$

Tendo em conta as propriedades da Multinormal (acetatos 239 e 240):

- $E[\vec{\hat{\beta}}] = \vec{\beta}$  e  $V[\vec{\hat{\beta}}] = \sigma^2 (\mathbf{X}^t \mathbf{X})^{-1}$ .
- Cada estimador individual  $\hat{\beta}_j$  tem distribuição **Normal**.
- Cada estimador individual tem média  $E[\hat{\beta}_j] = \beta_j$  (logo, é **centrado**).
- Cada estimador individual tem variância  $V[\hat{\beta}_j] = \sigma^2 (\mathbf{X}^t \mathbf{X})^{-1}_{(j+1,j+1)}$ .  
(Note-se o desfasamento nos índices).
- Estimadores individuais diferentes **não são** (em geral) independentes, porque a matriz  $(\mathbf{X}^t \mathbf{X})^{-1}$  não é, em geral, uma matriz diagonal.  $Cov[\hat{\beta}_i, \hat{\beta}_j] = \sigma^2 (\mathbf{X}^t \mathbf{X})^{-1}_{(i+1,j+1)}$ .

# A distribuição dum estimador individual

Como se viu no acetato anterior, tem-se,  $\forall j = 0, 1, \dots, p$ :

$$\hat{\beta}_j \cap \mathcal{N} \left( \beta_j, \sigma^2 (\mathbf{X}^t \mathbf{X})_{(j+1, j+1)}^{-1} \right)$$
$$\Leftrightarrow \frac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}} \cap \mathcal{N}(0, 1),$$

onde  $\sigma_{\hat{\beta}_j} = \sqrt{\sigma^2 (\mathbf{X}^t \mathbf{X})_{(j+1, j+1)}^{-1}}$ .

Este resultado generaliza os relativos à Regressão Linear Simples.

## O problema de $\sigma^2$ desconhecido

O resultado distribucional indicado no acetato anterior permitiria construir intervalos de confiança ou fazer testes a hipóteses sobre os parâmetros  $\vec{\beta}$ , não fosse a existência de um problema já familiar: o desconhecimento da variância  $\sigma^2$  dos erros aleatórios.

Procedemos de forma análoga ao que se fez na Regressão Linear Simples:

- obter um estimador para  $\sigma^2$ ; e
- ver o que acontece à distribuição do acetato anterior quando  $\sigma^2$  é substituído pelo seu estimador.

# SQRE na Regressão Múltipla

## Teorema (Resultados distribucionais de SQRE)

Dado o Modelo de Regressão Linear Múltipla (RLM), tem-se:

- $\frac{SQRE}{\sigma^2} \cap \chi^2_{n-(p+1)}$
- *SQRE é independente de  $\vec{\beta}$ .*

**NOTA:** Omite-se a demonstração

## Corolário

Dado o Modelo de RLM,  $E \left[ \frac{SQRE}{n-(p+1)} \right] = \sigma^2$ .

**NOTA:** Os graus de liberdade associados a *SQRE* são o número de observações ( $n$ ) menos o número de parâmetros do modelo ( $p+1$ ).

# O Quadrado Médio Residual na Regressão Múltipla

## Definição (Quadrado Médio Residual)

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

$$QMRE = \frac{SQRE}{n - (p + 1)}$$

- O QMRE é habitualmente usado na Regressão como estimador da variância dos erros aleatórios, isto é, toma-se

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

- Como se viu no acetato anterior, QMRE é um **estimador centrado**.

## Revisitando o estimador de $\beta_j$

Vimos (acetato 245) que cada estimador  $\hat{\beta}_j$  verifica:

$$Z = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 \cdot (\mathbf{X}^t \mathbf{X})_{(j+1, j+1)}^{-1}}} \cap \mathcal{N}(0, 1).$$

Temos ainda:

$$W = \frac{SQRE}{\sigma^2} \cap \chi_{n-(p+1)}^2 \quad \text{e} \quad Z, W \text{ v.a. independentes.}$$

Logo (ver também o acetato 136):

$$\frac{Z}{\sqrt{W/(n-(p+1))}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{QMRE \cdot (\mathbf{X}^t \mathbf{X})_{(j+1, j+1)}^{-1}}} \cap t_{n-(p+1)}.$$

## Quantidades fulcrais para a inferência sobre $\beta_j$

**Teorema** (Distribuições para a inferência sobre  $\beta_j$ ,  $j = 0, 1, \dots, p$ )

*Dado o Modelo de Regressão Linear Múltipla, tem-se*

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \cap t_{n-(p+1)},$$

com  $\hat{\sigma}_{\hat{\beta}_j} = \sqrt{QMRE \cdot (\mathbf{X}^t \mathbf{X})_{(j+1, j+1)}^{-1}}$ .

Este Teorema dá-nos os resultados que servem de base à construção de **intervalos de confiança** e **testes de hipóteses** para os parâmetros  $\beta_j$  do modelo populacional.

**NOTA:** A quantidade fulcral acima é totalmente análoga aos resultados correspondentes na RLS. Assim, **os ICs e testes de hipóteses a parâmetros individuais, na RLM, serão análogos aos da RLS.**

## Intervalo de confiança para $\beta_j$

**Teorema** (Intervalo de Confiança a  $(1 - \alpha) \times 100\%$  para  $\beta_j$ )

*Dado o Modelo de Regressão Linear Múltipla, um intervalo a  $(1 - \alpha) \times 100\%$  de confiança para o parâmetro  $\beta_j$  do modelo é:*

$$\left[ b_j - t_{\alpha/2[n-(p+1)]} \cdot \hat{\sigma}_{\hat{\beta}_j}, \quad b_j + t_{\alpha/2[n-(p+1)]} \cdot \hat{\sigma}_{\hat{\beta}_j} \right],$$

com  $\hat{\sigma}_{\hat{\beta}_j} = \sqrt{QMRE \cdot (\mathbf{X}^t \mathbf{X})_{(j+1,j+1)}^{-1}}$ , e sendo  $t_{\alpha/2[n-(p+1)]}$  o valor que na distribuição  $t_{n-(p+1)}$  deixa à direita uma região de probabilidade  $\alpha/2$ . O valor  $b_j$  é o elemento  $j+1$  do vector das estimativas  $\vec{\mathbf{b}}$  (acetato 215).

**NOTA:** A amplitude do IC  **aumenta com o valor de  $QMRE$  e o valor diagonal da matriz  $(\mathbf{X}^t \mathbf{X})^{-1}$  associado ao parâmetro  $\beta_j$  em questão.**



## Intervalos de confiança para $\beta_j$ no

A informação básica para a construção de intervalos de confiança para cada parâmetro  $\beta_j$  obtém-se, no R, a partir das tabelas produzidas pela função `summary`. No exemplo do acetato 227:

```
> 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	***

Assim, estima-se que em média a largura da pétala diminui  $0.20727\text{ cm}$  por cada aumento de  $1\text{ cm}$  no comprimento da sépala (mantendo-se as outras medições constantes). Como  $t_{0.025(146)} = 1.976346$ , o IC a 95% para  $\beta_2$  é

$$\begin{aligned} & ] (-0.20727) - (1.976346)(0.04751), (-0.20727) + (1.976346)(0.04751) [ \\ & \Leftrightarrow ] -0.3012, -0.1134 [ \end{aligned}$$

## Intervalos de confiança para $\beta_j$ no (cont.)

Alternativamente, é possível usar a função `confint` no objecto resultante de ajustar a regressão para obter os intervalos de confiança para cada  $\beta_j$  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

```
> 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

## Testes de Hipóteses sobre os parâmetros

O mesmo resultado (acetato 250) usado para construir intervalos de confiança serve para construir testes a hipóteses para cada  $\beta_j$  individual. Dado o Modelo de Regressão Linear Múltipla,

### Testes de Hipóteses a $\beta_j$ (Regressão Linear Múltipla)

$$\text{Hipóteses: } H_0 : \beta_j \begin{matrix} \geq \\ \leq \end{matrix} c \quad \text{vs.} \quad H_1 : \beta_j \begin{matrix} < \\ > \end{matrix} c$$

$$\text{Estatística do Teste: } T = \frac{\hat{\beta}_j - \overbrace{\beta_j}_{=c} |_{H_0}}{\hat{\sigma}_{\hat{\beta}_j}} \cap t_{n-(p+1)}$$

Nível de significância do teste:  $\alpha$

Região Crítica (Região de Rejeição): **Rejeitar  $H_0$  se**

$$\begin{array}{ll} T_{calc} < -t_{\alpha[n-(p+1)]} & \text{(Unilateral esquerdo)} \\ |T_{calc}| > t_{\alpha/2[n-(p+1)]} & \text{(Bilateral)} \\ T_{calc} > t_{\alpha[n-(p+1)]} & \text{(Unilateral direito)} \end{array}$$

## Combinações lineares dos parâmetros

Seja  $\vec{\mathbf{a}} = (a_0, a_1, \dots, a_p)^t$  um vector 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:

$$\vec{\mathbf{a}}^t \vec{\boldsymbol{\beta}} = a_0 \beta_0 + a_1 \beta_1 + a_2 \beta_2 + \dots + a_p \beta_p .$$

Casos particulares importantes nas aplicações são:

- Se  $\vec{\mathbf{a}}$  tem um único elemento não-nulo, de valor 1, na posição  $j+1$ ,  $\vec{\mathbf{a}}^t \vec{\boldsymbol{\beta}} = \beta_j$ .
- Se  $\vec{\mathbf{a}}$  tem apenas dois elementos não-nulos, 1 na posição  $i+1$  e  $\pm 1$  na posição  $j+1$ ,  $\vec{\mathbf{a}}^t \vec{\boldsymbol{\beta}} = \beta_i \pm \beta_j$ .
- Se  $\vec{\mathbf{a}} = (1, x_1, x_2, \dots, x_p)$ , onde  $x_j$  indica 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 predictoras**:

$$\begin{aligned} \vec{\mathbf{a}}^t \vec{\boldsymbol{\beta}} &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p \\ &= E[Y | X_1 = x_1, X_2 = x_2, \dots, X_p = x_p] = \mu_{Y|\vec{\mathbf{x}}} . \end{aligned}$$

## Inferência sobre combinações lineares dos $\beta_j$ s

A multinormalidade do vector de estimadores  $\vec{\hat{\beta}}$  implica a normalidade de qualquer vector que seja combinação linear das suas componentes (acetato 240, ponto 4). Mais concretamente,

- Sabemos que  $\vec{\hat{\beta}} \cap \mathcal{N}_{p+1}(\vec{\beta}, \sigma^2(\mathbf{X}^t\mathbf{X})^{-1})$  (acetato 244);
- Logo,  $\vec{\mathbf{a}}^t\vec{\hat{\beta}} \cap \mathcal{N}(\vec{\mathbf{a}}^t\vec{\beta}, \sigma^2\vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}})$  (acetato 240, ponto 4);
- Ou seja,  $\mathbf{Z} = \frac{\vec{\mathbf{a}}^t\vec{\hat{\beta}} - \vec{\mathbf{a}}^t\vec{\beta}}{\sqrt{\sigma^2\vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}}} \cap \mathcal{N}(0, 1)$ ;
- Por um raciocínio análogo ao usado aquando dos  $\beta$ s individuais, tem-se então

$$\frac{\vec{\mathbf{a}}^t\vec{\hat{\beta}} - \vec{\mathbf{a}}^t\vec{\beta}}{\sqrt{QMRE \cdot \vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}}} \cap t_{n-(p+1)}.$$

## Quantidade fulcral para a inferência sobre $\vec{a}^t \vec{\beta}$

**Teorema** (Distribuições para combinações lineares dos  $\beta$ s)

*Dado o Modelo de Regressão Linear Múltipla, tem-se*

$$\frac{\vec{a}^t \hat{\vec{\beta}} - \vec{a}^t \vec{\beta}}{\hat{\sigma}_{\vec{a}^t \vec{\beta}}} \cap t_{n-(p+1)},$$

com  $\hat{\sigma}_{\vec{a}^t \vec{\beta}} = \sqrt{QMRE \cdot \vec{a}^t (\mathbf{X}^t \mathbf{X})^{-1} \vec{a}}$ .

Neste Teorema temos o resultado que serve de base à construção de **intervalos de confiança** e **testes de hipóteses** para quaisquer combinações lineares dos parâmetros  $\beta_j$  do modelo.

**NOTA:** Repare-se na analogia da estrutura desta quantidade fulcral com os resultados anteriores, relativos a  $\beta_j$ s individuais (acetato 250).

## Intervalo de confiança para $\vec{a}^t \vec{\beta}$

**Teorema** (Intervalo de Confiança a  $(1 - \alpha) \times 100\%$  para  $\vec{a}^t \vec{\beta}$ )

Dado o Modelo de Regressão Linear Múltipla, um intervalo a  $(1 - \alpha) \times 100\%$  de confiança para a combinação linear dos parâmetros,  $\vec{a}^t \vec{\beta} = a_0 \beta_0 + a_1 \beta_1 + \dots + a_p \beta_p$ , é:

$$\left[ \vec{a}^t \vec{b} - t_{\alpha/2[n-(p+1)]} \cdot \hat{\sigma}_{\vec{a}^t \vec{\beta}} \quad , \quad \vec{a}^t \vec{b} + t_{\alpha/2[n-(p+1)]} \cdot \hat{\sigma}_{\vec{a}^t \vec{\beta}} \right] ,$$

com  $\vec{a}^t \vec{b} = a_0 b_0 + a_1 b_1 + \dots + a_p b_p$  e  $\hat{\sigma}_{\vec{a}^t \vec{\beta}} = \sqrt{QMRE \cdot \vec{a}^t (\mathbf{X}^t \mathbf{X})^{-1} \vec{a}}$ .

# Testes de Hipóteses sobre os parâmetros

Dado o Modelo de Regressão Linear Múltipla,

## Testes de Hipóteses a $\vec{a}^t \vec{\beta}$ (Regressão Linear Múltipla)

$$\text{Hipóteses: } H_0 : \vec{a}^t \vec{\beta} \begin{matrix} \geq \\ = \\ \leq \end{matrix} c \quad \text{vs.} \quad H_1 : \vec{a}^t \vec{\beta} \begin{matrix} < \\ \neq \\ > \end{matrix} c$$

$$\text{Estatística do Teste: } T = \frac{\overbrace{\vec{a}^t \vec{\beta} - \vec{a}^t \vec{\beta}}^{=c} |_{H_0}}{\hat{\sigma}_{\vec{a}^t \vec{\beta}}} \cap t_{n-(p+1)} .$$

Nível de significância do teste:  $\alpha$

Região Crítica (Região de Rejeição): **Rejeitar  $H_0$  se**

$$T_{calc} < -t_{\alpha[n-(p+1)]} \quad (\text{Unilateral esquerdo})$$

$$|T_{calc}| > t_{\alpha/2[n-(p+1)]} \quad (\text{Bilateral})$$

$$T_{calc} > t_{\alpha[n-(p+1)]} \quad (\text{Unilateral direito})$$



## De novo os casos particulares

No acetato 255 viram-se três casos particulares importantes de combinações lineares dos parâmetros.

- No caso de  $\vec{\mathbf{a}}^t \vec{\beta} = \beta_j$ , os intervalos e testes acabados de ver são idênticos aos dados nos acetatos 251 e 254.
- No caso de  $\vec{\mathbf{a}}^t \vec{\beta} = \beta_j \pm \beta_j$ , tem-se  $\hat{\sigma}_{\vec{\mathbf{a}}^t \vec{\beta}} = \hat{\sigma}_{\hat{\beta}_i \pm \hat{\beta}_j}$ , com:

$$\begin{aligned}\hat{\sigma}_{\hat{\beta}_i \pm \hat{\beta}_j} &= \sqrt{\widehat{V[\hat{\beta}_i \pm \hat{\beta}_j]}} = \sqrt{\widehat{V[\hat{\beta}_i]} + \widehat{V[\hat{\beta}_j]} \pm 2 \cdot \widehat{Cov[\hat{\beta}_i, \hat{\beta}_j]}} \\ &= \sqrt{QMRE \cdot \left[ (\mathbf{X}^t \mathbf{X})_{(i+1, i+1)}^{-1} + (\mathbf{X}^t \mathbf{X})_{(j+1, j+1)}^{-1} \pm 2(\mathbf{X}^t \mathbf{X})_{(i+1, j+1)}^{-1} \right]}\end{aligned}$$

- No caso de  $\vec{\mathbf{a}}$  conter os valores das variáveis preditoras **usados na  $i$ -ésima observação**,  $\vec{\mathbf{a}}$  será a linha  $i$  da matrix  $\mathbf{X}$ . Nesse caso,

$$\hat{\sigma}_{\vec{\mathbf{a}}^t \vec{\beta}} = \sqrt{QMRE \cdot \vec{\mathbf{a}}^t (\mathbf{X}^t \mathbf{X})^{-1} \vec{\mathbf{a}}} = \sqrt{QMRE \cdot h_{ij}},$$

onde  $h_{ij}$  indica o  $i$ -ésimo elemento diagonal da matriz de projecções ortogonal  $\mathbf{H} = \mathbf{X}(\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t$ .

## ICs para combinações lineares no $\mathbb{R}$

Para construir um intervalo de confiança para  $\vec{a}'\vec{\beta}$ , será necessário conhecer a matriz das (co)variâncias estimadas dos estimadores  $\vec{\hat{\beta}}$ ,

$$\widehat{V[\vec{\hat{\beta}}]} = \hat{\sigma}_{\vec{\hat{\beta}}}^2 = QMRE \cdot (\mathbf{X}'\mathbf{X})^{-1}.$$

No  $\mathbb{R}$ , esta matriz obtém-se através da função `vcov`.

E.g., a matriz das (co)variâncias estimadas no exemplo dos lírios é:

```
> 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  $\hat{\beta}_2 + \hat{\beta}_3$  é:

$$\hat{\sigma}_{\hat{\beta}_2 + \hat{\beta}_3} = \sqrt{0.002256837 + 0.002394932 + 2(-0.001344002)} = 0.04431439.$$

## Intervalos de confiança para $\mu_{Y|\vec{x}}$

Se a combinação linear dos  $\beta$ s que se deseja corresponde ao **valor esperado de  $Y$ , dado um conjunto de valores das variáveis preditoras**, isto é, a  $\mu_{Y|\vec{x}} = E[Y|X_1 = x_1, X_2 = x_2, \dots, X_p = x_p]$ , então o intervalo de confiança do acetato 258 particulariza-se da seguinte forma:

$$\left[ \hat{\mu}_{Y|\vec{x}} - t_{\alpha/2[n-(p+1)]} \cdot \hat{\sigma}_{\hat{\mu}_{Y|\vec{x}}} \quad , \quad \hat{\mu}_{Y|\vec{x}} + t_{\alpha/2[n-(p+1)]} \cdot \hat{\sigma}_{\hat{\mu}_{Y|\vec{x}}} \right]$$

sendo  $\vec{x} = (x_1, x_2, \dots, x_p)$  o vector dos valores dos preditores,  
 $\vec{a} = (1, x_1, x_2, \dots, x_p)$  o vector obtido acrescentando o valor inicial 1,

$$\hat{\mu}_{Y|\vec{x}} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p$$

e

$$\hat{\sigma}_{\hat{\mu}_{Y|\vec{x}}} = \sqrt{QMRE \cdot \vec{a}^t (\mathbf{X}^t \mathbf{X})^{-1} \vec{a}}.$$

## Intervalos de confiança para $\mu_{Y|\bar{x}}$ no

É possível obter o intervalo de confiança referido no acetato anterior através do comando `predict`, tal como na RLS.

No exemplo dos lírios, um IC a 95% 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  $\vec{a} = (1, 2, 5, 3.1)^t$ . No R é obtido assim:

```
> predict(iris2.lm, data.frame(Petal.Length=c(2), Sepal.Length=c(5),
+                               Sepal.Width=c(3.1)), int="conf")
      fit      lwr      upr
[1,] 0.462297 0.4169203 0.5076736
```

O IC para  $E[Y|X_1=2, X_2=5, X_3=3.1]$  é:  `] 0.4169 , 0.5077 [`.

## Intervalos de predição para $Y$

Podem também obter-se, de forma análoga ao que foi visto na RLS, **intervalos de predição para uma observação individual de  $Y$** , associada aos valores  $X_1 = x_1, \dots, X_p = x_p$  das variáveis preditoras.

Nestes intervalos, a estimativa da variância associada a uma observação individual de  $Y$  é acrescida em  $QMRE$  unidades:

$$\left[ \hat{\mu}_{Y|\vec{x}} - t_{\alpha/2[n-(p+1)]} \cdot \hat{\sigma}_{indiv} \quad , \quad \hat{\mu}_{Y|\vec{x}} + t_{\alpha/2[n-(p+1)]} \cdot \hat{\sigma}_{indiv} \right]$$


onde  $\vec{x} = (x_1, x_2, \dots, x_p)^t$  indica o vector dos valores dos preditores e

$$\hat{\mu}_{Y|\vec{x}} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p$$

e

$$\hat{\sigma}_{indiv} = \sqrt{QMRE [1 + \vec{a}^t (\mathbf{X}^t \mathbf{X})^{-1} \vec{a}]} \quad \text{com } \vec{a} = (1, x_1, x_2, \dots, x_p).$$

## Intervalos de predição para $Y$ no R

No , é possível obter um intervalo de predição através do comando `predict` com o argumento `int="pred"`, tal como na RLS.

Eis, na RLM 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, data.frame(Petal.Length=c(2), Sepal.Length=c(5),  
+   Sepal.Width=c(3.1)), int="pred")
```

```
           fit           lwr           upr  
[1,] 0.462297 0.08019972 0.8443942
```

O intervalo de predição pedido é: ] 0.0802 , 0.8444 [.

## Avaliando a qualidade do ajustamento global

Numa Regressão Linear **Simples**, se  $\beta_1 = 0$ , a equação do modelo é apenas  $Y = \beta_0 + \varepsilon$ . Neste caso, o conhecimento do preditor  $X$  em nada contribui para o conhecimento de  $Y$  (o Modelo Nulo não tira partido da informação dos preditores).

Numa **Regressão Linear Múltipla**, o **modelo Nulo**  $Y_i = \beta_0 + \varepsilon_i$ , corresponde a admitir que **todas** as variáveis preditoras têm coeficiente nulo. As hipóteses que queremos confrontar são:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$$

[MODELO INÚTIL]

vs.

$$H_1 : \exists j = 1, \dots, p \quad \text{t.q.} \quad \beta_j \neq 0$$

[MODELO NÃO INÚTIL]

**NOTA:** repare que  $\beta_0$  não intervém nas hipóteses.

## Distribuição associada a $SQR$

De novo, o ponto de partida para uma estatística de teste será a Soma de Quadrados associada à Regressão,  $SQR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ .

Tem-se (sem demonstração):

### Teorema

*Dado o Modelo de Regressão Linear Múltipla,*

- $\frac{SQR}{\sigma^2} \cap \chi_p^2$ , *se  $\beta_1 = \beta_2 = \dots = \beta_p = 0$ .*
- $SQR$  e  $SQRE$  *são variáveis aleatórias independentes.*

Defina-se o **Quadrado Médio** associado à Regressão,  $QMR = \frac{SQR}{p}$ .



# A estatística do teste de ajustamento global

Temos (veja também o acetato 168), se  $\beta_j = 0, \forall i = 1 : p$

$$\left. \begin{array}{l} W = \frac{SQR}{\sigma^2} \cap \chi_p^2 \\ V = \frac{SQRE}{\sigma^2} \cap \chi_{n-(p+1)}^2 \\ W, V \text{ independentes} \end{array} \right\} \Rightarrow \frac{W/p}{V/n-(p+1)} = \frac{QMR}{QMRE} \cap F_{p, n-(p+1)} .$$

sendo  $QMR = \frac{SQR}{p}$  e  $QMRE = \frac{SQRE}{n-(p+1)}$  .

# O Teste $F$ de ajustamento global do Modelo

Sendo válido o Modelo RLM, pode efectuar-se o seguinte

## Teste $F$ de ajustamento global do modelo RLM

Hipóteses:  $H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$

vs.

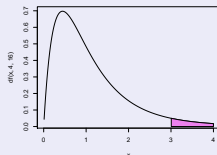
$H_1 : \exists j = 1, \dots, p$  tal que  $\beta_j \neq 0$ .

Estatística do Teste:  $F = \frac{QMR}{QMRE} \cap F_{p, n-(p+1)}$  se  $H_0$ .

Nível de significância do teste:  $\alpha$

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar  $H_0$  se  $F_{calc} > f_{\alpha[p, n-(p+1)]}$



## Expressões alternativas no teste $F$ global

A estatística do teste  $F$  de ajustamento global do modelo numa Regressão Linear Múltipla pode ser escrita na forma alternativa:

$$F = \frac{n - (p + 1)}{p} \cdot \frac{R^2}{1 - R^2} .$$

Tal como na Regressão Linear Simples, a estatística  $F$  é uma função crescente do Coeficiente de Determinação,  $R^2$ .

As hipóteses do teste também se podem escrever como

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

A hipótese  $H_0 : \mathcal{R}^2 = 0$  indica ausência de relação linear entre  $Y$  e o conjunto dos preditores. Corresponde a um ajustamento “péssimo” do modelo. A sua rejeição não garante um bom ajustamento, mas apenas a capacidade de o distinguir do Modelo Nulo.

## Outra formulação do Teste $F$ de ajustamento global

### Teste $F$ de ajustamento global do modelo RLM (alternativa)

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

Estatística do Teste:  $F = \frac{n-(p+1)}{p} \cdot \frac{R^2}{1-R^2} \cap F_{(p, n-(p+1))}$  se  $H_0$ .

Nível de significância do teste:  $\alpha$

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar  $H_0$  se  $F_{calc} > f_{\alpha(p, n-(p+1))}$

- A estatística  $F$  é uma função crescente do coeficiente de determinação amostral,  $R^2$ .
- A hipótese nula  $H_0 : \mathcal{R}^2 = 0$  afirma que, na população, o coeficiente de determinação é nulo.

# O Quadro-resumo do ajustamento global

Frequentemente, sintetiza-se a informação usada num teste de ajustamento global num **quadro-resumo da regressão**:

Fonte	g.l.	SQ	QM	$f_{calc}$
Regressão	$p$	$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	$\frac{SQR}{p}$	$\frac{QMR}{QMRE}$
Resíduos	$n - (p + 1)$	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	$\frac{SQRE}{n - p - 1}$	
Total	$n - 1$	$\sum_{i=1}^n (y_i - \bar{y})^2$	—	—

# O princípio da parcimónia na RLM

Recordemos o **princípio da parcimónia** na modelação: queremos um modelo que descreva adequadamente a relação entre as variáveis, mas que **seja o mais simples (parcimonioso) possível**.

Caso se disponha de um modelo de Regressão Linear Múltipla com um ajustamento considerado adequado, a aplicação deste princípio traduz-se em saber se **será possível obter um modelo com menos variáveis preditoras, sem perder significativamente em termos de qualidade de ajustamento**.

## Modelo e Submodelos

Se dispomos de um modelo de Regressão Linear Múltipla, com relação de base

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 ,$$

chamamos **submodelo** a um modelo de regressão linear múltipla contendo apenas algumas das variáveis preditoras, e.g.,

$$Y = \beta_0 + \beta_2 x_2 + \beta_5 x_5 ,$$

Podemos identificar o submodelo pelo conjunto  $\mathcal{S}$  das variáveis preditoras que pertencem ao submodelo. No exemplo,  $\mathcal{S} = \{2, 5\}$ . O modelo e o submodelo são idênticos se  $\beta_j = 0$  para qualquer variável  $x_j$  cujo índice não pertença a  $\mathcal{S}$ .

## Comparando modelo e submodelos

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:

$$H_0 : \beta_j = 0, \quad \forall j \notin \mathcal{S} \quad \text{vs.} \quad H_1 : \exists j \notin \mathcal{S} \quad \text{tal que } \beta_j \neq 0.$$

[SUBMODELO = MODELO]

[SUBMODELO PIOR]

NOTA: Esta discussão só envolve coeficientes  $\beta_j$  de variáveis predictoras. O coeficiente  $\beta_0$  faz sempre parte dos submodelos. Este coeficiente  $\beta_0$  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).



# Estatística de teste para comparar modelo/submodelo

A estatística de teste envolve a comparação das Somas de Quadrados Residuais do:

- **modelo completo** (referenciado pelo índice  $C$ ); e do
- **submodelo** (referenciado pelo índice  $S$ )

Seja  $k$  o número de preditores do submodelo ( $k + 1$  parâmetros).

Tem-se:

$$F = \frac{(SQRE_S - SQRE_C)/(p - k)}{SQRE_C/[n - (p + 1)]} \cap F_{p-k, n-(p+1)},$$

sob  $H_0$ , isto é, caso  $\beta_j = 0$ , para todas as variáveis  $x_j$  que não pertençam ao submodelo. São os valores grandes da estatística que levantam dúvidas sobre  $H_0$ .

## O teste a um submodelo (teste $F$ parcial)

### Teste $F$ de comparação dum modelo com um seu submodelo

Dado o Modelo de Regressão Linear Múltipla,

Hipóteses:

$$H_0 : \beta_j = 0, \quad \forall j \notin \mathcal{S} \quad \text{vs.} \quad H_1 : \exists j \notin \mathcal{S} \quad \text{tal que} \quad \beta_j \neq 0.$$

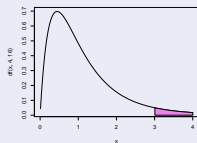
Estatística do Teste:

$$F = \frac{(SQRE_S - SQRE_C)/(p-k)}{SQRE_C/[n-(p+1)]} \cap F_{p-k, n-(p+1)}, \text{ sob } H_0.$$

Nível de significância do teste:  $\alpha$

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar  $H_0$  se  $F_{calc} > f_{\alpha|p-k, n-(p+1)}$



## Expressão alternativa para 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 na forma alternativa:

$$F = \frac{n - (p + 1)}{p - k} \cdot \frac{R_C^2 - R_S^2}{1 - R_C^2}.$$

**NOTA:** Assinale-se que a Soma de Quadrados Total,  $SQT$ , apenas depende dos valores observados da variável resposta  $Y$ , e não de qual o modelo ajustado. Assim,  $SQT$  é igual no modelo completo e no submodelo.

# Expressão alternativa para as hipóteses do teste

As hipóteses do teste também se podem escrever como

$$H_0 : \mathcal{R}_C^2 = \mathcal{R}_S^2 \quad \text{vs.} \quad H_1 : \mathcal{R}_C^2 > \mathcal{R}_S^2 ,$$

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.

Caso não se rejeite  $H_0$ , opta-se pelo submodelo (mais parcimonioso).  
Caso se rejeite  $H_0$ , opta-se pelo modelo completo (ajusta-se significativamente melhor).

# Teste $F$ parcial: formulação alternativa

## Teste $F$ de comparação dum modelo com um seu submodelo

Dado o Modelo de Regressão Linear Múltipla,

Hipóteses:

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

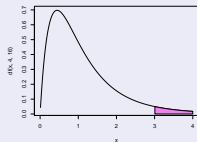
Estatística do Teste:

$$F = \frac{n-(p+1)}{p-k} \cdot \frac{R_C^2 - R_S^2}{1 - R_C^2} \quad \cap \quad F_{p-k, n-(p+1)}, \text{ sob } H_0 .$$


Nível de significância do teste:  $\alpha$

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar  $H_0$  se  $F_{calc} > f_{\alpha[p-k, n-(p+1)]}$



## O teste a submodelos no

A informação necessária para um teste  $F$  parcial obtem-se no , através da função `anova`, com dois argumentos: os objectos `lm` resultantes de ajustar o modelo completo e o submodelo sob comparação.

Nos exemplos dos lírios (acetatos 140 e 252), temos:

```
> 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	***

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.

## Relação entre os testes- $t$ e o teste $F$ parcial

Caso o modelo e submodelo difiram num único preditor,  $X_j$ , o teste  $F$  parcial descrito nos acetatos anteriores é equivalente ao teste  $t$  (acetato 254) com as 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  $t_v$ , então o seu quadrado,  $T^2$  tem distribuição  $F_{1,v}$ .

## Como escolher um 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.

**Nota:** Veja-sa o Exercício RLM 9e) para um exemplo.



## Como escolher um submodelo? (cont.)

Mas em muitas situações não é evidente qual o subconjunto de variáveis preditoras que se deseja considerar no submodelo.

Pretende-se apenas ver se o modelo é simplificável. Nestes casos, a opção por um submodelo não é um problema fácil.

Dadas  $p$  variáveis preditoras, o número de subconjuntos, 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

## Cuidado com exclusões simultâneas de preditores

Para pequenos valores de  $p$ , é viável analisar 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.

Não é legítimo usar os testes  $t$  à significância de cada coeficiente  $\beta_j$ , no modelo completo, para decidir sobre a exclusão de vários preditores **em simultâneo**.

Os testes  $t$  aos coeficientes  $\beta_j$  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 das variáveis que permanecem no submodelo. **Pode acontecer que um preditor seja dispensável num modelo completo, mas deixe de o ser num submodelo, ou viceversa.**

## Um exemplo

Nos dados relativos ao Exercício 2 (RLM) das aulas práticas, a tabela associada à regressão da variável *Brix* sobre todas 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	*

Mas *não* é legítimo concluir que *Altura*, *Peso* e *pH* são dispensáveis.

```
> 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 *

## Pesquisas completas

Para um número  $p$  de preditores pequeno ou médio, e dispondo de algoritmos e rotinas informáticas adequadas, é possível efectuar uma **pesquisa completa** que assegure que se encontra o subconjunto de  $k$  preditores com o maior valor de  $R^2$  (ou de algum outro critério de qualidade do submodelo).

O algoritmo *leaps and bounds*, de Furnival e Wilson <sup>1</sup> 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`.

---

<sup>1</sup>Furnival, G.W and Wilson, R.W.,Jr. (1974) Regressions by leaps and bounds, *Technometrics*, **16**, 499-511.

# Um exemplo de aplicação da rotina leaps

Apesar do pequeno número de preditores, exemplifiquemos a aplicação da função leaps com os dados brix.

```
> colnames(brix)
[1] "Diametro" "Altura" "Peso" "Brix" "pH" "Acucar"

> library(leaps) <-- carregar o módulo (tem de estar instalado)
> leaps(y=brix$Brix, x=brix[,-4], method="r2", nbest=1) <-- o comando: 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 de preditores: 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.

# Algoritmos de pesquisa sequenciais

Caso não esteja disponível *software* apropriado, ou se o número  $p$  de preditores for demasiado grande, pode recorrer-se a **algoritmos de pesquisa** que simplificam uma regressão linear múltipla **sem analisar todo os possíveis submodelos e sem a garantia de obter os melhores subconjuntos**.

Vamos considerar 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*).

Existem variantes deste algoritmo, não estudadas aqui:

- **algoritmo de inclusão sequencial** (*forward selection*).
- **algoritmos de exclusão/inclusão alternada** (*stepwise selection*).

# O algoritmo de exclusão sequencial

- 1 ajustar o modelo completo, com os  $p$  preditores;
- 2 definir um nível de significância  $\alpha$  para os testes de hipóteses; existem variáveis para as quais **não** se rejeita a hipótese  $\beta_j = 0$ ?
  - ▶ Em caso negativo, não é possível simplificar o modelo (passar ao ponto 3).
  - ▶ Em caso afirmativo, qualquer dessas variáveis é candidata a sair do modelo.
    - ★ se apenas existe uma 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)

**Reajustar o modelo após a exclusão da variável e repetir este ponto 2**
- 3 Quando não existirem variáveis candidatas a sair, ou quando sobrar um único preditor, o algoritmo pára. Tem-se então o submodelo final.

# Um exemplo – Exercício 2 (RLM)

Fixando o nível de significância  $\alpha = 0.05$ :

```
> 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	*

```
> 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 (0.05)
pH	0.33844	0.33322	1.016	0.336316	
Acucar	0.08481	0.03810	2.226	0.053031	. <- Deixou de ser significativo (0.05)


```
> 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 (0.05)

O algoritmo pára aqui. Pode comparar-se o submodelo final com o modelo completo, através dum teste  $F$  parcial.



## Algoritmos sequenciais com base no AIC

O  disponibiliza funções para automatizar pesquisas sequenciais de submodelos, semelhantes à que aqui foi enunciada, mas em que o critério de exclusão dum 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**. No contexto dum **Regressão Linear Múltipla com  $k$  variáveis preditoras**, define-se como

$$AIC = n \cdot \ln \left( \frac{SQRE_k}{n} \right) + 2(k + 1) .$$

**Nota:** O AIC pode tomar valores negativos.

# Interpretando o AIC

$$AIC = n \cdot \ln \left( \frac{SQRE_k}{n} \right) + 2(k+1)$$

- a primeira parcela é função crescente de  $SQRE_k$ , i.e., quanto melhor o ajustamento, mais pequena a primeira parcela;
- a segunda parcela mede a complexidade do modelo ( $k+1$  é o número de parâmetros), pelo que quanto mais parcimonioso o modelo, mais pequena a segunda parcela.

Assim, o AIC mede simultaneamente a qualidade do ajustamento e a simplicidade do modelo.

Um modelo para a variável resposta  $Y$  é considerado **melhor** que outro se tiver um **AIC menor** (quando ajustados com os mesmos dados).

## Algoritmos sequenciais com base no AIC (cont.)

Pode definir-se um algoritmo de exclusão sequencial, com base no critério AIC:

- ajustar o modelo completo e calcular o respectivo AIC.
- ajustar cada submodelo com menos **uma** variável e calcular o respectivo AIC.
- Se nenhum dos AICs dos submodelos considerados for 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 que mais reduz o AIC e regressa-se ao ponto anterior.

## Algoritmos sequenciais com base no AIC (cont.)

Em cada passo de exclusão, o submodelo com menor AIC será aquele que provoca menor aumento no  $SQRE$ , ou seja, que tiver excluído a variável cujo teste a  $\beta_j = 0$  tem maior  $p$ -value.

Assim, o procedimento de exclusão sequencial baseado nos testes  $t$  ou no AIC coincidem na ordem das variáveis a excluir, podendo diferir apenas no critério de paragem.

Em geral, um algoritmo de exclusão sequencial baseado no AIC é mais cauteloso ao excluir do que um algoritmo baseado nos testes  $t$ , sobretudo se o valor de  $\alpha$  usado nos testes for baixo.

Nos algoritmos de exclusão baseados nos testes  $t$ , é aconselhável usar valores relativamente elevados de  $\alpha$ , como  $\alpha = 0.10$ .

## Algoritmos de exclusão sequencial no

A função `step` corre o algoritmo de exclusão sequencial, com base no AIC. Considere ainda o exemplo dos dados `brix` (Exercício 2 RLM):

```
> 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. Neste caso, não se exclui qualquer variável: o AIC do modelo inicial é inferior ao de qualquer submodelo resultante de excluir uma variável. O submodelo final é o modelo inicial.

# Uma palavra final sobre algoritmos de pesquisa

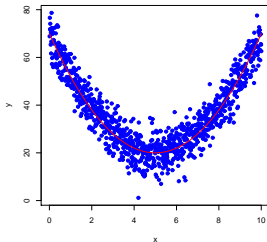
O algoritmo de exclusão sequencial **não** garante a identificação do “melhor submodelo” com um dado número de preditores. Apenas identifica, de forma que não é computacionalmente muito pesada, submodelos que se presume serem “bons”.

**Deve ser usado com bom senso** e o submodelo obtido cruzado 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).

# 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.

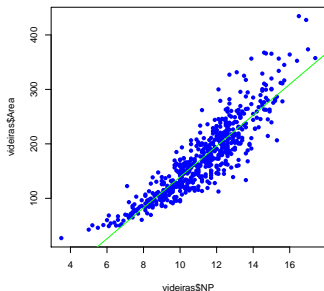
Considere-se, por exemplo, que a relação de fundo entre uma variável resposta  $Y$  e uma única variável preditora  $X$  não é dada por uma recta, mas sim por uma parábola:



## Regressão Polinomial - Exemplo

Considere os dados do Exercício 7 da Regressão Linear Múltipla, relativos a medições sobre  $n = 600$  folhas de videira.

Eis o gráfico das **áreas** vs. **comprimentos de nervuras principais**, com sobreposta a recta de regressão:



Há uma tendência para curvatura. Talvez um polinómio de 2o. grau?



## Regressão Polinomial - Exemplo (cont.)

Qualquer parábola, com equação  $Y = \beta_0 + \beta_1 X + \beta_2 X^2$ , pode ser ajustada e estudada como se se tratasse duma regressão linear entre  $Y$  e as variáveis  $X_1 = X$  e  $X_2 = X^2$ :

```
> 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

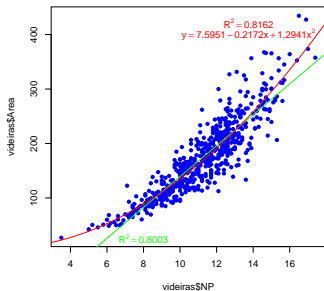
F-statistic: 1325 on 2 and 597 DF, p-value: < 2.2e-16

A equação da parábola ajustada é  $y = 7.5961 - 0.2172x + 1.2941x^2$ .

A rejeição da hipótese  $\beta_2 = 0$  diz que a parábola ajusta-se significativamente melhor do que a recta de regressão.

## Regressão Polinomial - Exemplo (cont.)

Eis a parábola ajustada:



É legítimo afirmar que este modelo de regressão quadrático explica 81.62% da variabilidade nas áreas foliares observadas, uma vez que não houve transformação da variável resposta  $Y$ .

## Regressões Polinomiais (cont.)

O argumento é extensível a qualquer polinómio de qualquer grau, e em qualquer número de variáveis. Dois exemplos:

- Polinómio de grau  $p$  numa variável

$$Y = \beta_0 + \beta_1 \underbrace{x}_{=x_1} + \beta_2 \underbrace{x^2}_{=x_2} + \beta_3 \underbrace{x^3}_{=x_3} + \dots + \beta_p \underbrace{x^p}_{=x_p}$$

- Polinómio de grau 2 em 2 variáveis

$$Y = \beta_0 + \beta_1 \underbrace{x}_{=x_1} + \beta_2 \underbrace{x^2}_{=x_2} + \beta_3 \underbrace{z}_{=x_3} + \beta_4 \underbrace{z^2}_{=x_4} + \beta_5 \underbrace{xz}_{=x_5}$$

## O $R^2$ modificado

O  $R^2$  modificado é uma variante do Coeficiente de Determinação.

- O Coeficiente de Determinação usual:

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQRE}{SQT}$$

- O  $R^2$  modificado (sendo  $QMT = SQT / (n - 1) = s_y^2$ ):

$$R_{mod}^2 = 1 - \frac{QMRE}{QMT} = 1 - \frac{SQRE}{SQT} \cdot \frac{n-1}{n-(p+1)} = 1 - (1 - R^2) \cdot \frac{n-1}{n-(p+1)}.$$

Tem-se sempre  $n-1 > n-(p+1)$ , pelo que  $R_{mod}^2 < R^2$ .

Quando  $n \gg p+1$  (muito mais observações que parâmetros no modelo) tem-se  $R^2 \approx R_{mod}^2$ .

Se  $n$  é pouco maior que o número de variáveis preditoras,  $R_{mod}^2$  é bastante inferior a  $R^2$  (excepto se  $R^2$  fôr muito próximo de 1).

## O $R^2$ modificado (cont.)

Viu-se que o valor de  $R_{mod}^2$  penaliza modelos complexos ajustados com poucas observações. Exemplo: dados brix ( $n=14$  e  $p+1=6$ ).

```
> summary(brix.lm)
[...]  
Multiple R-squared:  0.8483, Adjusted R-squared:  0.7534
```

Um submodelo pode ter  $R_{mod}^2$  maior que um modelo completo.

Exemplo: Exercício 9 (também ilustra o uso do  $R_{mod}^2$  como critério de selecção na função de pesquisa leaps):

```
> 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
```

# Análise de Resíduos e outros diagnósticos

Uma análise de regressão linear não fica completa sem o estudo dos resíduos e de alguns outros diagnósticos.

Grande parte do que se disse sobre resíduos na Regressão Linear Simples mantém-se válido numa Regressão Linear Múltipla.

Relembrar três conceitos relacionados, mas diferentes:

Erros aleatórios

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_{1(i)} + \beta_2 x_{2(i)} + \dots + \beta_p x_{p(i)})$$

Resíduos (variáveis aleatórias - preditores dos erros aleatórios)

$$E_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1(i)} + \hat{\beta}_2 x_{2(i)} + \dots + \hat{\beta}_p x_{p(i)})$$

Resíduos (valores observados)

$$e_i = y_i - (b_0 + b_1 x_{1(i)} + b_2 x_{2(i)} + \dots + b_p x_{p(i)})$$

# Propriedades dos Resíduos sob o Modelo RLM

O modelo de Regressão Linear Múltipla admite que

$$\varepsilon_i \cap \mathcal{N}(0, \sigma^2) \quad \forall i = 1, \dots, n.$$

Sob o modelo RLM, os **resíduos** têm a seguinte distribuição:

$$E_i \cap \mathcal{N}\left(0, \sigma^2(1 - h_{ii})\right) \quad \forall i = 1, \dots, n,$$

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  $\mathcal{C}(\mathbf{X})$ .

Em notação vectorial, o **vector dos  $n$  resíduos**  $E_i$  é dado por:

$$\vec{\mathbf{E}} = \vec{\mathbf{Y}} - \vec{\hat{\mathbf{Y}}} = \vec{\mathbf{Y}} - \mathbf{H}\vec{\mathbf{Y}} = (\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}},$$

## Propriedades dos Resíduos sob o Modelo RLM (cont.)

**Teorema** (Distribuição dos Resíduos no modelo RLM)

*Dado o Modelo de Regressão Linear Múltipla, tem-se:*

$$\vec{\mathbf{E}} \cap \mathcal{N}_n(\vec{\mathbf{0}}, \sigma^2(\mathbf{I}_n - \mathbf{H})) \quad \text{sendo} \quad \vec{\mathbf{E}} = (\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}.$$

O vector dos resíduos  $\vec{\mathbf{E}} = \vec{\mathbf{Y}} - \vec{\hat{\mathbf{Y}}} = \vec{\mathbf{Y}} - \mathbf{H}\vec{\mathbf{Y}} = (\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}$ , tem distribuição **Multinormal** pelo último ponto do Teorema do acetato 240.

O vector esperado de  $\vec{\mathbf{E}}$  resulta das propriedades do acetato 235:

- $E[\vec{\mathbf{E}}] = E[(\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}] = (\mathbf{I}_n - \mathbf{H})E[\vec{\mathbf{Y}}] = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\vec{\beta} = \vec{\mathbf{0}}$ ,  
pois o vector  $\mathbf{X}\vec{\beta} \in \mathcal{C}(\mathbf{X})$ , logo permanece invariante sob a acção da matriz de projecção  $\mathbf{H}$ :  $\mathbf{H}\mathbf{X}\vec{\beta} = \mathbf{X}\vec{\beta}$ .



## Propriedades dos Resíduos sob o Modelo RLM (cont.)

A matriz de covariâncias do vector aleatório dos resíduos,  $\vec{E}$ , calcula-se a partir do facto de a matriz de projecção ortogonal ser (Exercício 4 da RLM) **simétrica** ( $\mathbf{H}^t = \mathbf{H}$ ) e **idempotente** ( $\mathbf{H}^2 = \mathbf{H}\mathbf{H} = \mathbf{H}$ ). Tendo também presentes as propriedades do acetato 236, vem:

- $V[\vec{E}] = V[(\mathbf{I}_n - \mathbf{H})\vec{Y}] = (\mathbf{I}_n - \mathbf{H})V[\vec{Y}](\mathbf{I}_n - \mathbf{H})^t = \sigma^2(\mathbf{I}_n - \mathbf{H})$ .

Embora no modelo RLM os erros aleatórios sejam independentes, **os resíduos não são variáveis aleatórias independentes**, pois as covariâncias entre resíduos diferentes são (em geral), não nulas:

$$\text{cov}[E_i, E_j] = -\sigma^2 h_{ij}, \quad \text{se } i \neq j,$$

onde  $h_{ij}$  indica o **elemento da linha  $i$  e coluna  $j$  da matriz  $\mathbf{H}$** .

## Vários tipos de resíduos

Tal como na RLS, definem-se diferentes tipos de resíduos:

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

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 **excluindo** a  $i$ -ésima observação (associada ao resíduo  $E_i$ ).

# Análise dos resíduos

Tal como para a RLS, também em regressões múltiplas se avalia a validade dos pressupostos do modelo através de **gráficos de resíduos**.

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  $E_j$ s vs.  $\hat{Y}_j$ s**: os pontos devem-se dispor numa banda horizontal, centrada no valor zero, sem outro padrão especial.
- **qq-plot dos resíduos estandardizados vs. distribuição Normal**: a Normalidade dos erros aleatórios corresponde à linearidade.
- **gráfico de resíduos vs. ordem de observação**: para investigar eventuais faltas de independência dos erros aleatórios.

## O efeito alavanca

Outras ferramentas de diagnóstico visam identificar observações individuais que merecem ulterior análise, tal como na RLS. Mas importa adaptar as definições ao contexto de Regressão Múltipla.

Numa RLM o **valor de efeito alavanca** (*leverage*) é o valor  $h_{ii}$  do elemento diagonal da matriz de projecção ortogonal  $\mathbf{H}$ , correspondente à observação  $i$ , que continua balizado entre  $\frac{1}{n}$  e 1.

O **valor médio** das observações alavanca numa RLM é

$$\bar{h} = \frac{p+1}{n},$$

ou seja, a razão entre o número de parâmetros e o número de observações.

# Gráficos de diagnóstico

A **distância de Cook** para avaliar a influência da observação  $i$  define-se agora como:

$$D_i = \frac{\|\vec{\hat{y}} - \vec{\hat{y}}_{(-i)}\|^2}{(p+1) QMRE},$$

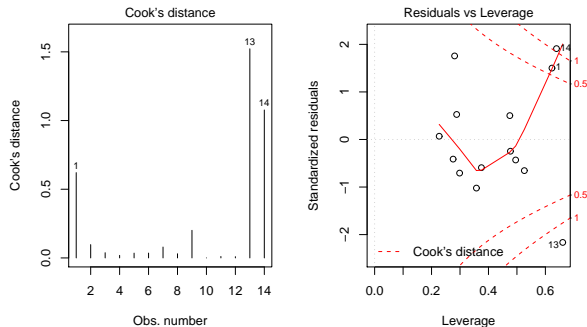
onde  $\vec{\hat{y}}_{(-i)} = \mathbf{X}\vec{\hat{\beta}}_{(-i)}$  é o vector dos  $n$  valores ajustados de  $Y$  obtido estimando os  $\beta$ s sem a observação  $i$ . Expressão equivalente é (sendo  $R_i$  o correspondente resíduo estandardizado):

$$D_i = R_i^2 \left( \frac{h_{ii}}{1 - h_{ii}} \right) \frac{1}{p+1}.$$

Os restantes aspectos da discussão são análogos aos duma RLS.

# Um exemplo de gráficos de diagnóstico

Um exemplo destes gráficos de diagnósticos, para os dados do Exercício 2 da RLM (Brix) é:



Os valores bastante grandes de distância de Cook e efeito alavanca  $h_{ii}$  neste exemplo reflectem o reduzido número de observações ( $n=14$ ) usado para ajustar um modelo com muitos parâmetros ( $p+1=6$ ).

# Advertências finais

1. Podem surgir problemas associados à **multicolinearidade** das variáveis preditoras, ou seja, ao facto das colunas da matriz  $\mathbf{X}$  serem (quase) linearmente dependentes. Nesse caso, podem:

- existir **problemas no cálculo de  $(\mathbf{X}^t\mathbf{X})^{-1}$** , logo no ajustamento do modelo e na estimação dos parâmetros;
- existir **variâncias muito grandes de alguns  $\hat{\beta}_i$ s**, o que significa muita instabilidade na inferência.

Multicolinearidade exacta reflecte redundância de informação nos preditores.

É possível eliminar multicolinearidade exacta ou aproximada, excluindo da análise uma ou várias variáveis preditoras que sejam responsáveis pela (quase) dependência linear dos preditores.

## Advertências finais (cont.)

2. Tal como na Regressão Linear Simples, podem ser usadas transformações da variável resposta, e também de alguma(s) das variáveis preditoras.

Em particular, podem ser úteis transformações que linearizem a relação entre  $Y$  e  $X_1, X_2, \dots, X_p$ . Tais transformações linearizantes podem permitir estudar relações de tipo não-linear através de relações lineares entre as variáveis transformadas.

E.g., (ver Ex.13 RLM) a relação não linear entre  $Y, x_1$  e  $x_2$ ,

$$y = ax_1^b x_2^c$$

torna-se, após logaritmização, numa relação linear entre  $\ln(y)$ ,  $\ln(x_1)$  e  $\ln(x_2)$  (com  $b_0^* = \ln(a)$ ,  $b_1^* = b$  e  $b_2^* = c$ ):

$$\ln(y) = b_0^* + b_1^* \ln(x_1) + b_2^* \ln(x_2) .$$



## Advertências finais (cont.)

3. Não se deve confundir a existência de uma relação linear entre preditores  $X_1, X_2, \dots, X_p$  e variável resposta  $Y$ , com uma relação de causa e efeito.

Pode existir uma relação de causa e efeito.

Mas pode também verificar-se:

- Uma relação de **variação conjunta**, mas não de tipo causal (como por exemplo, em muitos conjuntos de dados morfométricos). Por vezes, preditores e variável resposta são todos efeito de causas comuns subjacentes.
- Uma relação **espúria**, de coincidência numérica.

Uma relação **causal** só pode ser afirmada com base em teoria própria do fenómeno sob estudo, e não com base na relação linear estabelecida estatisticamente.