

Estatística e Delineamento Experimental 2026/2027

Modelo Linear

Elsa Gonçalves
Secção de Matemática, DCEB, ISA-Ulisboa

(Adaptado, Cadima J. (2021). O Modelo Linear, ISA, ULLisboa)

Regressão Linear Múltipla - Abordagem Descritiva

Quando é necessária **mais do que uma variável preditora** para modelar adequadamente a variável resposta de interesse.

Plano em $\mathbb{R}^3$

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

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

No nosso contexto, e colocando:

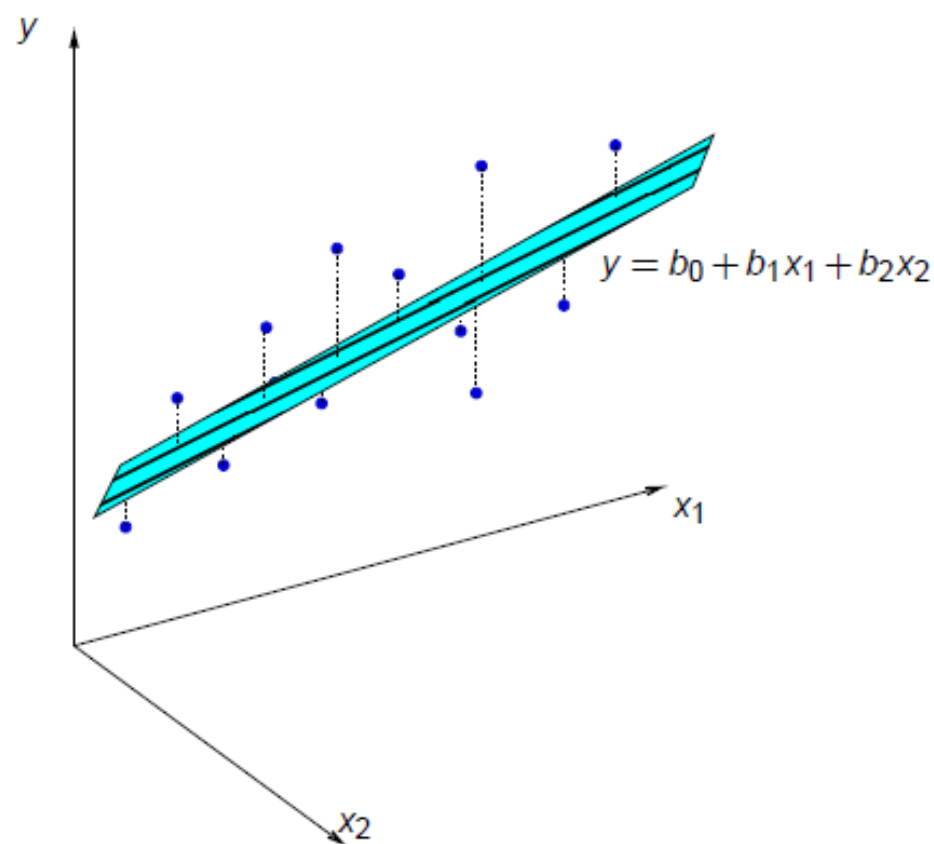
- no eixo vertical (z) a variável resposta Y ;
- noutro eixo (x) um preditor X_1 ;
- no terceiro eixo (y) o outro preditor X_2 ,

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, x_2, y).
Pode ser ajustado pelo mesmo critério que na RLS: minimizar SQRE.

O caso geral: p preditores

Para modelar uma variável resposta Y com base numa regressão linear sobre p variáveis preditoras, $x_1, x_2, \dots, x_p$, admite-se 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:

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

Não é possível visualizar a nuvem de pontos das observações se $p > 2$.

Regressão Múltipla: o hiperplano ajustado

Admite-se que os valores de y oscilam em torno duma combinação linear (afim) das p variáveis preditoras:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p .$$

Trata-se da equação dum **hiperplano** em $\mathbb{R}^{p+1}$.

O **critério** utilizado para ajustar um hiperplano à nuvem de n pontos em $\mathbb{R}^{p+1}$ é o de **minimizar a Soma de Quadrados dos Resíduos**, ou seja, escolher os $p+1$ parâmetros $\{b_j\}_{j=0}^p$ que minimizem:

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

onde os y_i representam 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** pela equação do hiperplano.

Duas abordagens para a estimação dos parâmetros

Para obter os parâmetros que definem o hiperplano que melhor se ajusta às observações pode-se usar uma abordagem:

- analítica; ou
- geométrica.

Nas duas abordagens, a notação vectorial-matricial é vantajosa.

Não existem fórmulas simples, como no caso da RLS, para cada um dos parâmetros b_j isoladamente. Mas é possível indicar uma fórmula única matricial para o conjunto dos $p + 1$ parâmetros do modelo.

Os n valores ajustados $\hat{y}_i$ também definem um vector de $\mathbb{R}^n$, $\vec{\hat{y}}$:

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

O vector $\vec{\hat{y}}$ é uma combinação linear dos vectores $\vec{\mathbf{1}}_n, \vec{\mathbf{x}}_1, \vec{\mathbf{x}}_2, \dots, \vec{\mathbf{x}}_p$

A matriz do modelo $\mathbf{X}$

O vector $\vec{\hat{y}}$ dos valores ajustados pode também escrever-se como um produto envolvendo uma matriz $\mathbf{X}$ cujas colunas sejam os vectores $\vec{1}_n, \vec{x}_1, \dots, \vec{x}_p$.

A matriz $\mathbf{X}$ do modelo

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1(1)} & x_{2(1)} & \cdots & x_{p(1)} \\ 1 & x_{1(2)} & x_{2(2)} & \cdots & x_{p(2)} \\ 1 & x_{1(3)} & x_{2(3)} & \cdots & x_{p(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1(n)} & x_{2(n)} & \cdots & x_{p(n)} \end{bmatrix}$$

$\underbrace{\hspace{1.5cm}}_{=\vec{1}_n} \quad \underbrace{\hspace{1.5cm}}_{=\vec{x}_1} \quad \underbrace{\hspace{1.5cm}}_{=\vec{x}_2} \quad \cdots \quad \underbrace{\hspace{1.5cm}}_{=\vec{x}_p}$

A matriz do modelo $\mathbf{X}$ é de dimensão $n \times (p+1)$.

O vector $\vec{\hat{y}}$ pode ser escrito desta forma: $\vec{\hat{y}} = \mathbf{X}\vec{b}$

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

- O conjunto de **todas** as combinações lineares dum conjunto de vectores chama-se o **subespaço gerado** (**spanned**) por esses vectores.
- O subespaço gerado pelas colunas da matriz do modelo $\mathbf{X}$ chama-se **subespaço das colunas** (**column-space**) da matriz $\mathbf{X}$, $\mathcal{C}(\mathbf{X})$.
- O vector $\vec{\mathbf{y}}$ pertence ao subespaço $\mathcal{C}(\mathbf{X})$
(os vectores $\vec{\mathbf{1}}_n, \vec{\mathbf{x}}_1, \dots, \vec{\mathbf{x}}_p$ são colunas $\mathbf{X}$ e $\vec{\mathbf{y}} = b_0 \vec{\mathbf{1}}_n + b_1 \vec{\mathbf{x}}_1 + b_2 \vec{\mathbf{x}}_2 + \dots + b_p \vec{\mathbf{x}}_p$).
- $\mathcal{C}(\mathbf{X})$ é um subespaço de $\mathbb{R}^n$ ($\mathcal{C}(\mathbf{X}) \subset \mathbb{R}^n$), mas de **dimensão $p+1$**
(se as colunas de $\mathbf{X}$ forem **linearmente independentes**, isto é, se nenhum vector se puder escrever como combinação linear dos restantes).
- Qualquer combinação linear das colunas da matriz $\mathbf{X}$, ou seja, **qualquer elemento de $\mathcal{C}(\mathbf{X})$** se pode escrever como $\mathbf{X}\vec{\mathbf{a}}$, onde $\vec{\mathbf{a}} = (a_0, a_1, a_2, \dots, a_p)$ é o vector dos coeficientes da combinação linear.

Os parâmetros

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

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:

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

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

$$\frac{\partial SQRE}{\partial \vec{b}} = \mathbf{0}$$

Os parâmetros ajustados na RL Múltipla

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

Ou, usando 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 $\vec{y}$.

SQRE na projecção ortogonal

SOLUÇÃO:

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

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

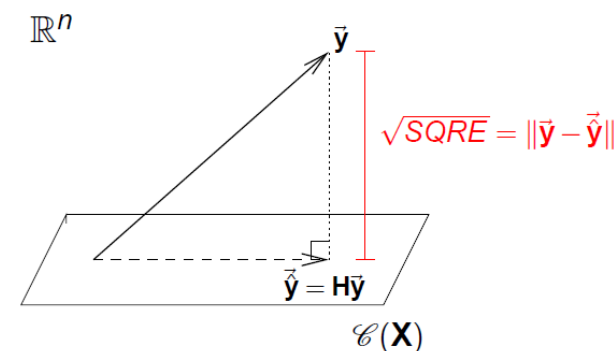
Matriz de projecção ortogonal sobre $\mathcal{C}(\mathbf{X})$

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

Nota: As matrizes de projecção ortogonal $\mathbf{P}$ sobre algum subespaço de $\mathbb{R}^n$ são as matrizes $n \times n$:

- simétricas (isto é, $\mathbf{P}^t = \mathbf{P}$); e
- idempotentes (isto é, $\mathbf{P}\mathbf{P} = \mathbf{P}$).

A matriz $\mathbf{H}$ tem estas propriedades (verifique!).



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

A projecção ortogonal no contexto da RLM

No contexto duma regressão linear múltipla, tem-se:

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

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

O vector de parâmetros ajustado

$$\vec{b} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \vec{y}.$$

Quando $p = 1$ (RLS): $\vec{b} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \vec{y} = \begin{bmatrix} \bar{y} - b_1 \bar{x} \\ \frac{cov_{xy}}{s_x^2} \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$

Os coeficientes b_j

O vector dos parâmetros ajustados pelo método dos mínimos quadrados, $\vec{\mathbf{b}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{y}}$, gera n valores ajustados:

$$\begin{aligned}\vec{\hat{\mathbf{y}}} &= \mathbf{H}\vec{\mathbf{y}} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{y}} = \mathbf{X}\vec{\mathbf{b}} \\ \Leftrightarrow \hat{y}_i &= b_0 + b_1x_{1(i)} + \dots + b_px_{p(i)} \quad , \quad \forall i .\end{aligned}$$

As unidades de medida:

- de b_0 são iguais às de y (e às de $\hat{y}$).
- dos parâmetros b_j das variáveis ($j \neq 0$) são a razão entre as unidades de y e as do preditor x_j correspondente.

Os coeficientes $\{b_j\}_{j=1}^p$ das variáveis preditoras interpretam-se como a diferença (média) em y , associada a aumentar o preditor x_j correspondente em uma unidade, mantendo os restantes preditores constantes.

Resíduos

As unidades de medida dos resíduos $e_i = y_i - \hat{y}_i$ são iguais às de y :

$$\begin{aligned} e_i &= y_i - \hat{y}_i = y_i - (b_0 + b_1 x_{1(i)} + \dots + b_p x_{p(i)}) \quad , \quad \forall i \\ \Leftrightarrow \vec{e} &= \vec{y} - \vec{\hat{y}} = \vec{y} - \mathbf{H}\vec{y} \quad , \end{aligned}$$

O vector de resíduos, $\vec{e}$, também pode ser obtido pré-multiplicando o vector $\vec{y}$ pela matriz $\mathbf{I} - \mathbf{H}$, onde $\mathbf{I}$ é a matriz identidade $n \times n$:

$$\vec{e} = \vec{y} - \mathbf{H}\vec{y} = (\mathbf{I} - \mathbf{H})\vec{y} \quad ,$$

A matriz $\mathbf{I} - \mathbf{H}$ é simétrica e idempotente, logo também é uma matriz de projecção ortogonal. Projecta sobre o subespaço de $\mathbb{R}^n$ constituído pelos vectores ortogonais a qualquer vector de $\mathcal{C}(\mathbf{X})$, chamado o complemento ortogonal de $\mathcal{C}(\mathbf{X})$ e designado por $\mathcal{C}(\mathbf{X})^\perp$.

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:

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

$$SQT = SQR + SQRE$$

Nota: Também na RL Múltipla os y observados (y_i) e os y ajustados ($\hat{y}_i$) têm a mesma média.

O Coeficiente de Determinação na Regressão Linear, $R^2 = \frac{SQR}{SQT}$

Algumas propriedades

$\mathcal{C}(\mathbf{X})$ contém o vector $\vec{\mathbf{1}}_n$ de n uns. Então $\mathbf{H}\vec{\mathbf{1}}_n = \vec{\mathbf{1}}_n$, pois a projecção de qualquer vector num subespaço que já o contém deixa o vector invariante. Logo:

- As médias dos valores observados e ajustados de Y são iguais:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n \hat{y}_i = \frac{1}{n} \vec{\mathbf{1}}_n^t \vec{\mathbf{y}} = \frac{1}{n} \vec{\mathbf{1}}_n^t \mathbf{H} \vec{\mathbf{y}} = \frac{1}{n} \vec{\mathbf{1}}_n^t \mathbf{H}^t \vec{\mathbf{y}} = \frac{1}{n} (\mathbf{H} \vec{\mathbf{1}}_n)^t \vec{\mathbf{y}} = \frac{1}{n} \vec{\mathbf{1}}_n^t \vec{\mathbf{y}} = \bar{y}$$

- A média dos resíduos é zero:

$$\bar{e} = \frac{1}{n} \sum_{i=1}^n e_i = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i) = \bar{y} - \bar{y} = 0.$$

- Em $\mathbb{R}^{p+1}$, o hiperplano ajustado contém o centro de gravidade da nuvem dos n pontos observados: $\bar{y} = b_0 + b_1 \bar{x}_1 + b_2 \bar{x}_2 + \dots + b_p \bar{x}_p$.

Já vimos que $\bar{y} = \bar{\hat{y}} = \frac{1}{n} \sum_{i=1}^n \hat{y}_i$. Mas $\frac{1}{n} \sum_{i=1}^n \hat{y}_i =$

$$\frac{1}{n} \sum_{i=1}^n (b_0 + b_1 x_{1(i)} + b_2 x_{2(i)} + \dots + b_p x_{p(i)}) = b_0 + b_1 \bar{x}_1 + b_2 \bar{x}_2 + \dots + b_p \bar{x}_p$$

A Regressão Linear Múltipla no

O comando `lm` também ajusta uma Regressão Linear Múltipla no . A variável resposta y e as variáveis preditoras $x_1, \dots, x_p$ definem-se mediante uma fórmula semelhante à da RLS.

E.g., sendo y a variável resposta e x_1 , x_2 e x_3 três preditores, a fórmula que especifica a relação será:

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

Comando R para ajustar uma regressão linear múltipla

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

O comando devolve o vector $\vec{b}$ das estimativas dos $p + 1$ parâmetros do modelo, $b_0, b_1, \dots, b_p$.

Um exemplo de RLM no R

Ilustra-se uma Regressão Linear Múltipla no R, com um conjunto de dados famoso: os lírios de Anderson/Fisher, disponíveis na *data frame* `iris`.

```
> help(iris)
```

`iris`

package:datasets

R Documentation

Edgar Anderson's Iris Data

Description:

This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables `sepal length` and `width` and `petal length` and `width`, respectively, for 50 flowers from each of 3 species of iris. The species are `Iris setosa`, `versicolor`, and `virginica`.

Exemplo RLM (cont.)

Uma inspecção inicial dos dados pode ser feita com o comando `head`, que mostra a parte inicial da *data frame*:

```
> head(iris)
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

Uma síntese da informação é dado pelo comando `summary`:

```
> 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	

Note-se que a quinta coluna é um **factor**. Será, para já, ignorado.

A Regressão Múltipla descritiva no (cont.)

Ajuste-se um modelo para prever a variável resposta largura da pétala, a partir do comprimento da pétala e das duas medições das sépalas (largura e comprimento), ignorando as espécies.

RL Múltipla - dados dos lírios

```
> 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 $\mathbb{R}^4$ ($\mathbb{R}^{p+1}$) é:

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

Confirmando as fórmulas (cont.)

Vamos confirmar a fórmula dos parâmetros ajustados pelo método dos mínimos quadrados. O comando `model.matrix` devolve a matriz **X**.

```
> X <- model.matrix(iris2.lm)
> X
```

	(Intercept)	Petal.Length	Sepal.Length	Sepal.Width
1	1	1.4	5.1	3.5
2	1	1.4	4.9	3.0
3	1	1.3	4.7	3.2
4	1	1.5	4.6	3.1
5	1	1.4	5.0	3.6
6	1	1.7	5.4	3.9
7	1	1.4	4.6	3.4
8	1	1.5	5.0	3.4
[...]				
149	1	5.4	6.2	3.4
150	1	5.1	5.9	3.0

Confirmando as fórmulas (cont.)

Os comandos do R para as operações matriciais necessárias para o cálculo de $\vec{b} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{y}$ são:

- `t(A)` indica a transposta da matriz A
- `A %*% B` indica o produto das matrizes A e B.
- `solve(A)` calcula a inversa da matriz A.

```
> y <- iris$Petal.Width  
> b <- solve( t(X) %*% X ) %*% ( t(X) %*% y )  
> b
```

```
              [,1]  
(Intercept) -0.2403074  
Petal.Length  0.5240831  
Sepal.Length -0.2072661  
Sepal.Width   0.2228285
```

Modelos e submodelos

Submodelos

Dado um modelo de regressão linear múltipla, com equação

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

chama-se **submodelo** a uma regressão linear com apenas **alguns** preditores.

Por exemplo, a regressão linear **simples**

$$Petal.Width = b_0 + b_1Petal.Length$$

é um **submodelo** da regressão linear múltipla acabada de ajustar,

$$Petal.Width = b_0 + b_1Petal.Length + b_2Sepal.Length + b_3Sepal.Width$$

Nota: Um submodelo (S) não pode ter preditores que não façam parte do modelo completo (C). A variável resposta tem de ser a mesma.

O R^2 de submodelos

Coeficientes de Determinação de submodelos: $R_s^2 \leq R_c^2$

O R_s^2 dum submodelo não pode exceder o R_c^2 do modelo completo.

O subespaço das colunas do submodelo tem de estar contido no subespaço das colunas do modelo completo: $\mathcal{C}(\mathbf{X}_s) \subseteq \mathcal{C}(\mathbf{X}_c)$. Logo, o ângulo entre $\vec{y}$ e $\vec{y}_s \in \mathcal{C}(\mathbf{X}_s)$ não pode ser menor que o ângulo entre $\vec{y}$ e $\vec{y}_c \in \mathcal{C}(\mathbf{X}_c)$, pois $\vec{y}_s$ também pertence a $\mathcal{C}(\mathbf{X}_c)$.

Ainda o exemplo dos lírios

```
> summary(iris2.lm)$r.sq  
[1] 0.9378503  
> iris.lm <- lm(Petal.Width ~ Petal.Length, data = iris)  
> summary(iris.lm)$r.sq  
[1] 0.9271098
```

- o valor de 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.

Equações de submodelos

Os parâmetros ajustados não são iguais

A equação ajustada num submodelo **não** é a parte correspondente na equação ajustada do modelo.

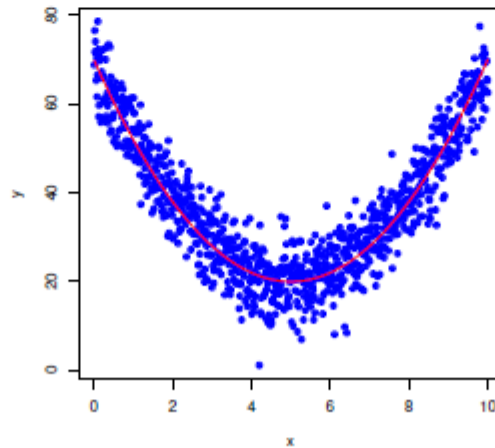
Ainda o exemplo dos lírios

```
> coef(iris.lm)
(Intercept) Petal.Length
-0.3630755    0.4157554
> coef(iris2.lm)
(Intercept) Petal.Length Sepal.Length Sepal.Width
-0.2403074    0.5240831   -0.2072661    0.2228285
```

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.

Imagine-se uma relação de fundo entre uma variável resposta Y e uma única variável preditora X dada por uma parábola:

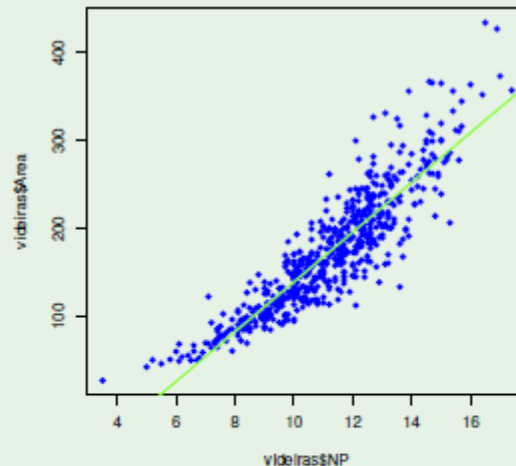


Regressão Polinomial - Exemplo

Exemplo 5 – Folhas de videira

Considere os dados de 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.)

Pode ajustar-se uma qualquer parábola, com equação

$$y = b_0 + b_1x + b_2x^2$$

com uma regressão linear de y sobre os dois preditores $x_1 = x$ e $x_2 = x^2$

```
> videiras.lm2 <- lm( Area ~ NP + I(NP^2) , data=videiras )  
> videiras.lm2
```

Call:

```
lm(formula = Area ~ NP + I(NP^2), data = videiras)
```

Coefficients:

(Intercept)	NP	I(NP^2)
7.5961	-0.2172	1.2941

```
> summary( videiras.lm2 )$r.sq
```

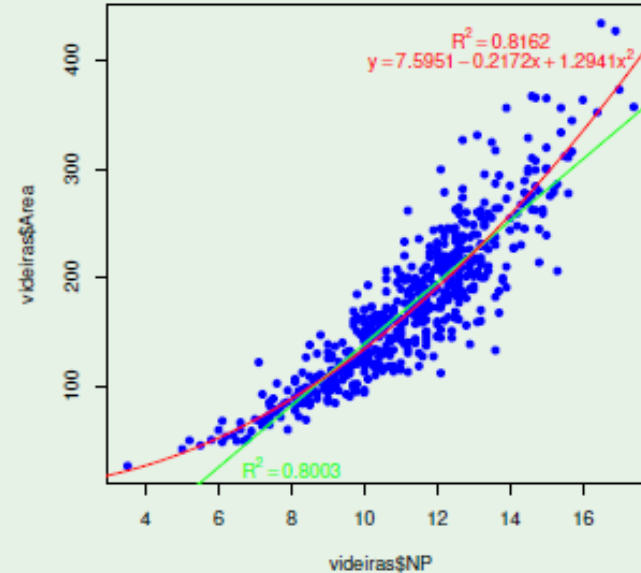
```
[1] 0.8161632
```

A parábola ajustada tem equação $y = 7.5961 - 0.2172x + 1.2941x^2$. O valor $R^2 = 0.8162$ indica que cerca de 82% da variabilidade observada nas áreas foliares é explicada pela regressão quadrática (aqui não houve transformação de y).

Nota: aplicável a qualquer polinómio de qualquer grau e em qualquer número de variáveis.

Regressão Polinomial - Exemplo (cont.)

A parábola ajustada



A equação da recta ajustada é $y = -144.15 + 28.34x$, confirmando que a equação ajustada dum submodelo (neste caso, a recta de regressão) **não** é apenas a parte relevante da equação ajustada dum modelo completo (neste caso, o modelo parabólico).