Modelo Linear

Modelação de relações entre variáveis

Importância central da recolha de informação (dados).

Nas disciplinas introdutórias de Estatística aprende-se a trabalhar com dados relativos a uma variável.

Nesta disciplina: relações (modelos) entre duas ou mais variáveis.

Variáveis podem ser:

- numéricas (medições, rendimentos, contagens, etc.) ou categóricas (factores) (espécies, locais, tratamentos, etc.);
- foco de interesse (variável resposta) ou auxiliares para explicar uma variável resposta (variável preditora ou explicativa).

Modelos determinísticos e modelos estatísticos

Uma relação (modelo) entre duas ou mais variáveis pode ser:

- essencialmente exacta (como na Mecânica: $F = ma$).
  Trata-se de modelos determinísticos.

  Ou
- apenas uma tendência de fundo, sabendo-se que existe variabilidade das observações em torno dessa tendência de fundo. Trata-se de modelos estatísticos ou probabilísticos.

Modelação Estatística

Objectivo (informal): Descrever a relação de fundo entre

- uma variável resposta (ou dependente) y; e
- uma ou mais variáveis preditoras (variáveis explicativas ou independentes), $x_1, x_2, \ldots, x_p$.

Informação: A identificação da relação de fundo é feita com base em n observações do conjunto de variáveis envolvidas na relação.

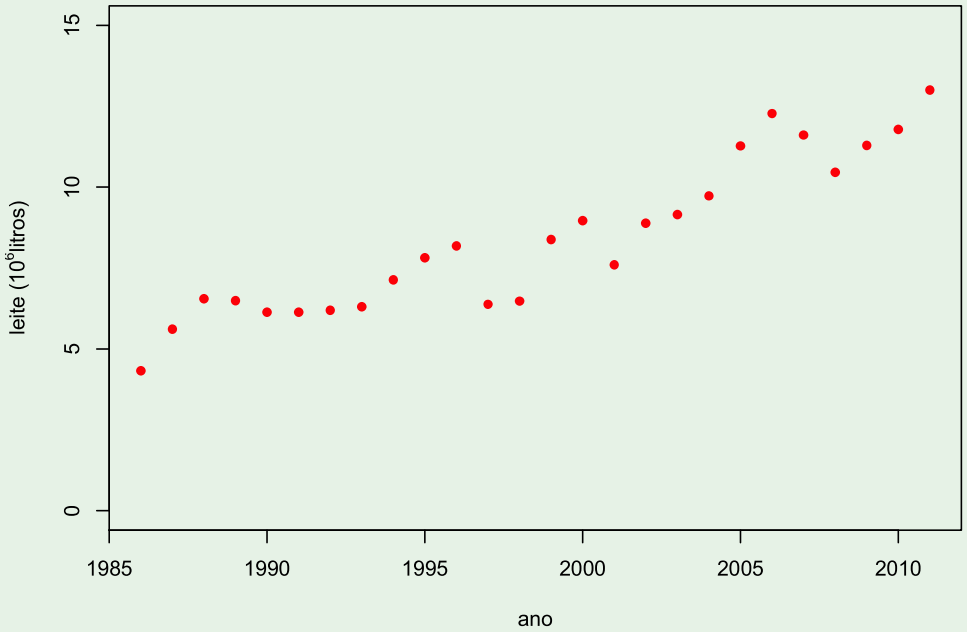
Vamos inicialmente considerar o contexto de um único preditor numérico, para modelar uma única variável resposta numérica.

Motivamos a discussão com dois exemplos.

Exemplo 1

Produção de leite de cabra em Portugal, 1986 a 2011 (INE)

Produção (y) *vs.* Anos (x), $n = 26$ pares de valores, $\{(x_i, y_i)\}_{i=1}^{26}$.

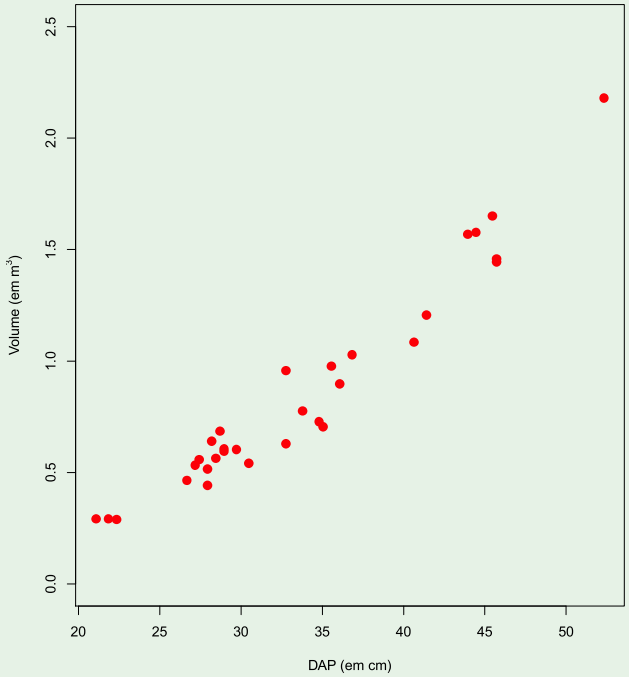


Existe uma tendência de fundo e é aproximadamente linear.
O coeficiente de correlação linear é $r_{xy} = 0.9348$.
Qual a "melhor" equação de recta, $y = b_0 + b_1 x$, para descrever as n observações (e que critério de "melhor")?

Exemplo 2 - relação linear

Volume de tronco *vs.* DAP em cerejeiras

DAP (Diâmetro à altura do peito, variável x) e Volume de troncos (y) de cerejeiras. Existem $n = 31$ pares de medições: $\{(x_i, y_i)\}_{i=1}^{31}$.



A tendência de fundo é aproximadamente linear. O coeficiente de correlação linear é $r_{xy} = 0.9671$. Mas os $n = 31$ pares de observações são apenas uma amostra aleatória duma população mais vasta. Interessa o contexto inferencial: o que se pode dizer sobre a recta populacional $y = \beta_0 + \beta_1 x$?

Regressão Linear Simples - contexto descritivo

Revisão: Estudado nas disciplinas introdutórias de Estatística.

Se n pares de observações $\{(x_i, y_i)\}_{i=1}^{n}$ têm relação linear de fundo, a recta de regressão de y sobre x define-se como:

Recta de Regressão Linear de y sobre x

$$y = b_0 + b_1 x$$

com

$$\begin{array}{rrcl} \text{Declive} & b_1 & = & \dfrac{cov_{xy}}{s_x^2} \\ \text{Ordenada na origem} & b_0 & = & \overline{y} - b_1 \overline{x} \end{array}$$

sendo

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad cov_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})\,.$$

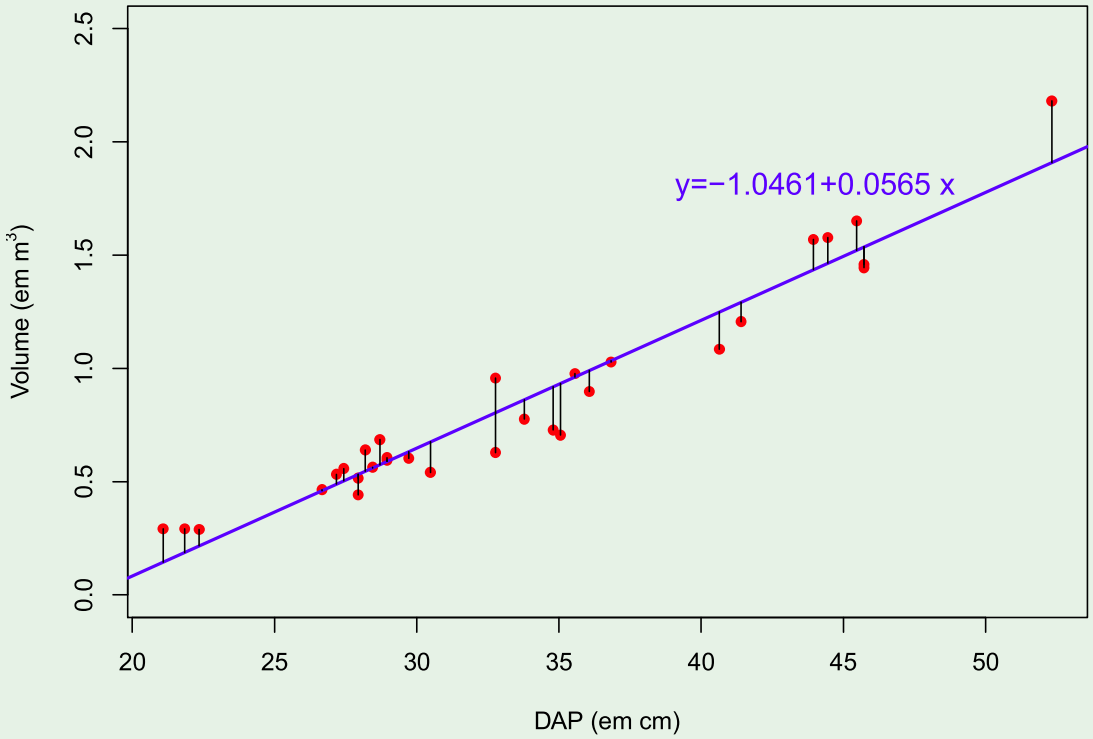
Regressão Linear Simples - contexto descritivo

Exemplo das cerejeiras

$n = 31$ pares de medições, $\{(x_i, y_i)\}_{i=1}^{31}$.
DAP (x) e Volume de troncos (y) de cerejeiras.

$$
\begin{aligned}
cov_{xy} &= 3.5881929 \\
s_x^2 &= 63.5348018 \\
\overline{x} &= 33.6509032 \\
\overline{y} &= 0.8543468
\end{aligned}
$$

$$
\begin{aligned}
b_1 = \frac{cov_{xy}}{s_x^2} &= 0.056476 \\
b_0 = \overline{y} - b_1 \overline{x} &= -1.046122
\end{aligned}
$$



Como se chegou à equação da recta?

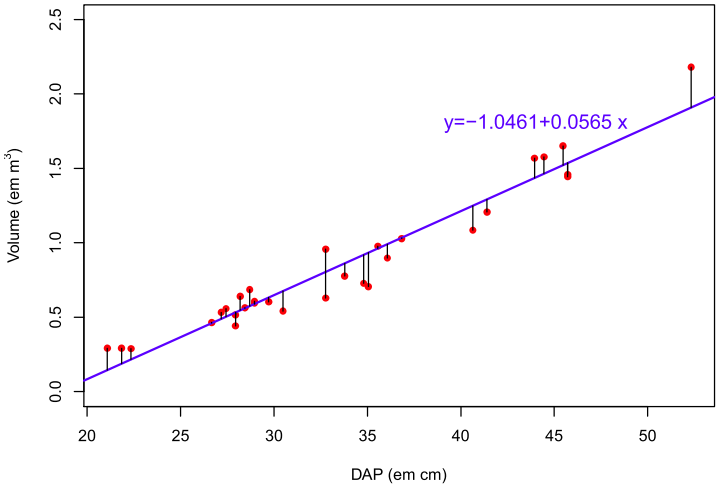
Valores ajustados e Resíduos

Dada uma recta, valores de y podem ser previstos a partir de valores de x, obtendo-se os "valores de y ajustados pela recta", $\hat{y}_i$:

$$\hat{y}_i = b_0 + b_1 x_i \ .$$

Os resíduos são as diferenças entre os valores de y observados e ajustados ou seja, são as diferenças na vertical entre pontos e recta ajustada:

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



O Critério de Mínimos Quadrados

Critério: minimizar a Soma de Quadrados dos Resíduos

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

Determinar b_0 e b_1 que minimizam $SQRE$ é um problema de minimizar uma função ($SQRE$) de duas variáveis (aqui chamadas b_0 e b_1).

Regressão Linear Simples - contexto descritivo

O critério de minimizar Soma de Quadrados dos Resíduos tem, subjacente, um pressuposto:

O papel das 2 variáveis, x e y, não é simétrico.

y – **variável resposta** ("dependente")

- variável que se deseja modelar, prever a partir da variável x.

x – **variável preditora** ("independente")

- variável com base na qual se pretende tirar conclusões sobre y.

Regressão Linear Simples - contexto descritivo

O i-ésimo resíduo é o desvio (com sinal) da observação y_i face à sua previsão a partir da recta:

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

Interpretação do Critério de Mínimos Quadrados

Minimizar a soma de quadrados dos resíduos corresponde a minimizar a soma de quadrados dos "erros de previsão".

O critério tem subjacente a preocupação de prever o melhor possível a variável y, a partir da sua relação com o preditor x.

Regressão Linear Simples - contexto descritivo no R

As regressões lineares são ajustadas no R usando o comando `lm` (as iniciais de linear model).

A função `lm` tem dois argumentos fundamentais:

- `formula` – identifica a variável resposta e as variáveis preditoras; numa RL simples da variável y sobre o preditor x, é da forma: $y \sim x$.
- `data` – indica o nome da *data frame* contendo os dados.

Comando R para a RLS do exemplo das cerejeiras

```
> lm( Volume ~ DAP , data=cerejeiras )

Call: lm(formula = Volume ~ DAP, data = cerejeiras)
Coefficients:
(Intercept)          DAP
   -1.04612      0.05648   <- valores ajustados de b0 e b1
```

Comandos R para o estudo da regressão

Vejamos alguns comandos do R úteis para estudar uma regressão.

Comecemos por guardar a regressão do exemplo das cerejeiras:

```
> cerejeiras.lm  <-  lm(Volume ~ DAP , data=cerejeiras )
```

- `fitted` devolve os valores ajustados $\hat{y}_i = b_0 + b_1 x_i$:

```
> fitted(cerejeiras.lm)
        1         2         3         4         5         6         7         8         9        10
0.1445051 0.1875398 0.2162296 0.4600931 0.4887829 0.5031278 0.5318176 0.5318176 0.5461625 0.5605074
       11        12        13        14        15        16        17        18        19        20
0.5748523 0.5891972 0.5891972 0.6322320 0.6752667 0.8043709 0.8043709 0.8617505 0.9191301 0.9334750
[...]
```

Comandos R (cont.)

- `residuals` devolve os resíduos $e_i = y_i - \hat{y}_i$:

```
> residuals(cerejeiras.lm)
          1           2           3           4           5           6           7           8
0.147158427 0.104123704 0.072602203 0.004303217 0.043573833 0.054714087 -0.090074800 -0.016450998
          9          10          11          12          13          14          15          16
0.093798219 0.002997825 0.110415357 0.005456540 0.016783279 -0.029083129 -0.134414916 -0.175736863
[...]
```

A Soma dos Quadrados dos Resíduos, *SQRE*, pode ser calculada por:

```
> sum(residuals(cerejeiras.lm)^2)

[1] 0.4204087
```

SQRE tem unidades de medida: o quadrado das unidades de y.

Comandos R para a regressão (cont.)

- `predict` – ajusta uma regressão a novas observações, dadas numa *data frame* com nomes de preditores iguais aos do ajustamento.

```
> novos <- data.frame( DAP=c(25, 50) )
> predict( cerejeiras.lm , new=novos )

        1         2
0.3657781 1.7776785
```

O valor $\hat{y}$ ajustado pela recta, para $x = 25$, é (arredondamentos aparte):

$$\begin{aligned} \hat{y} &= b_0 + b_1 x \\ \Leftrightarrow \quad &= -1.04612 + 0.05648 \times 25 \, . \end{aligned}$$

Revisão: Propriedades dos parâmetros da recta

Propriedades dos parâmetros da recta de regressão

- A ordenada na origem b_0:
  - é o valor de y (na recta) associado a $x = 0$;
  - tem unidades de medida iguais às de y.
- O declive b_1:
  - é a variação (média) de y associada a um aumento de uma unidade em x;
  - tem unidades de medida iguais a $\frac{unidades\ de\ y}{unidades\ de\ x}$.

Exemplo das cerejeiras

$$b_1 = 0.05648\,\frac{m^3}{cm}$$

por cada cm a mais no DAP, o volume do tronco aumenta, em média, $0.05648\,m^3$.

Revisão: Propriedades da recta de regressão

Propriedades da recta de regressão

- A recta de regressão passa sempre no centro de gravidade da nuvem de pontos, isto é, no ponto $(\overline{x}, \overline{y})$, como é evidente a partir da fórmula para a ordenada na origem:

$$b_0 = \overline{y} - b_1 \overline{x} \qquad \Leftrightarrow \qquad \overline{y} = b_0 + b_1 \overline{x} \ .$$

- $\overline{y}$ é simultaneamente a média dos y_i observados e dos $\hat{y}_i$ ajustados. (Ver Exercício RLS 5).

- Embora não tenha sido explicitamente exigido, a média dos resíduos e_i é nula, ou seja, $\overline{e} = 0$. (Ver Exercício RLS 5).

Revisão: RLS - As três Somas de Quadrados

Recordar: $s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2$ a variância amostral das observações y_i.

Soma de Quadrados Total (SQT)

SQ Total $\quad SQT = \sum_{i=1}^{n} (y_i - \overline{y})^2 \quad = \quad (n-1)\, s_y^2$

Tem-se: $s_{\hat{y}}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$ a variância amostral dos $\hat{y}_i$ ajustados.

Soma de Quadrados da Regressão (SQR)

SQ Regressão $\quad SQR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 \quad = \quad (n-1)\, s_{\hat{y}}^2$

Soma de Quadrados Residual (SQRE) - já dado

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

Revisão: RLS - Fórmula fundamental e R^2

Fórmula Fundamental da Regressão

Prova-se a seguinte Fórmula Fundamental (ver Exercício RLS 5):

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

Definição: Coeficiente de Determinação

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

R^2 mede a proporção da variabilidade total da variável resposta Y que é explicada pela regressão. Quanto maior, melhor.

Propriedades do Coeficiente de Determinação

Propriedades de $R^2 = \frac{SQR}{SQT}$

- $0 \leq R^2 \leq 1$ (Todas as SQs são não negativas e $SQT = SQR + SQRE$)
- $R^2 = 1$ se, e só se, os n pontos são colineares. ("ideal")

  ($SQT = SQR \Leftrightarrow SQRE = \sum_{i=1}^{n} e_i^2 = 0 \Rightarrow$ $e_i = 0$, para todo o i.

  Logo, todos os resíduos são nulos: os pontos estão todos em cima da recta.)
- $R^2 = 0$ se, e só se, a recta de regressão for horizontal. ("inútil")
  ($SQR = 0 \Leftrightarrow SQRE = SQT$. Toda a variabilidade de y é residual.

  $SQR = 0$ implica $\hat{y}_i = \overline{y}$, para todo o i. A recta é $y = \overline{y} \Leftrightarrow b_1 = 0$)
- Numa regressão linear simples, R^2 é o quadrado do coeficiente de correlação linear entre x e y (Ver Exercício RLS 6):

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

Exemplo das cerejeiras

O coeficiente de determinação R^2 obtem-se aplicando o comando `summary` a uma regressão ajustada. Surge com a designação `Multiple R-Squared`:

```
> summary(cerejeiras.lm)

Call: lm(formula = Volume ~ DAP, data = cerejeiras)
[...]
Residual standard error: 0.1204 on 29 degrees of freedom
Multiple R-squared:  0.9353,  Adjusted R-squared:  0.9331
F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

O valor de R^2 (com maior precisão) pode ser obtido da seguinte forma:

```
> summary(cerejeiras.lm)$r.sq
[1] 0.9353199
```