

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** **catóricas (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 predictoras** (**variáveis explicativas** ou **independentes**), x_1, x_2, \dots, x_p .

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

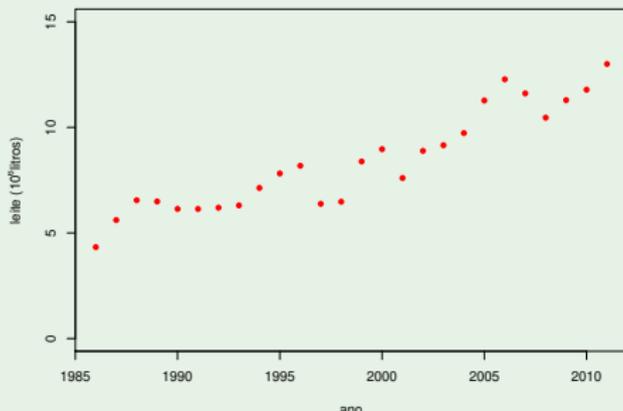
Vamos inicialmente considerar o contexto de **um único preditor numérico**, para modelar **uma única variável resposta numérica**.

Motivamos a discussão com **três 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**.

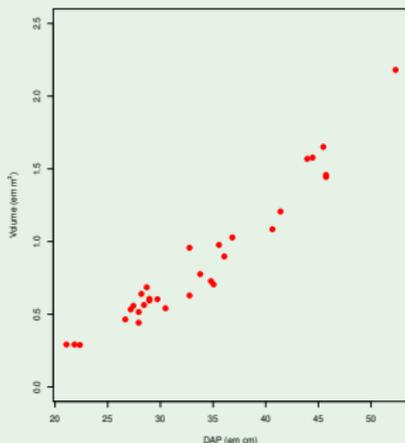
O **coeficiente de correlação linear** é $r_{xy} = 0.9348$.

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

$$\text{Declive} \quad b_1 = \frac{\text{COV}_{xy}}{s_x^2}$$

$$\text{Ordenada na origem} \quad b_0 = \bar{y} - b_1 \bar{x}$$

sendo

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{cov}_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) .$$

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.

$$cov_{xy} = 3.5881929$$

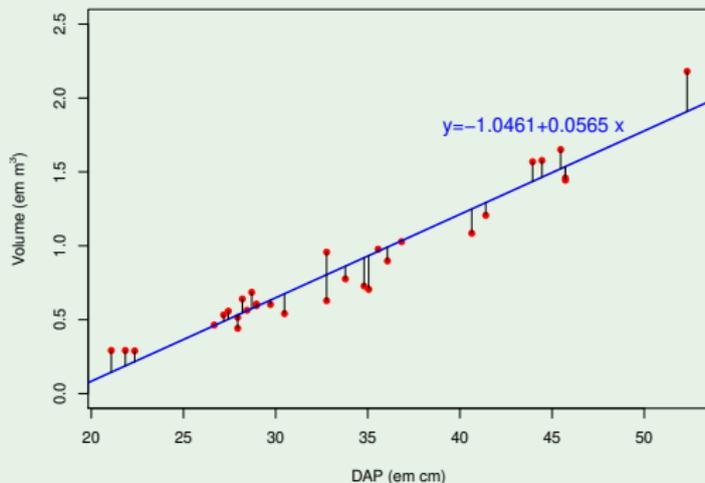
$$s_x^2 = 63.5348018$$

$$\bar{x} = 33.6509032$$

$$\bar{y} = 0.8543468$$

$$b_1 = \frac{cov_{xy}}{s_x^2} = 0.056476$$

$$b_0 = \bar{y} - b_1 \bar{x} = -1.046122$$



Como se chegou à equação da recta?

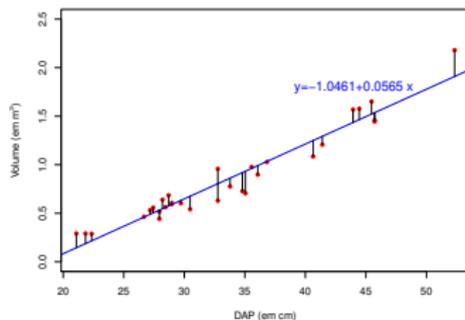
Valores ajustados e Resíduos

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).

Critérios de ajustamento diferentes dariam rectas diferentes. Em vez de distâncias na vertical,

- distâncias na perpendicular?
- distâncias na horizontal?

Em vez de soma de quadrados de distâncias,

- soma das distâncias (valor absoluto dos resíduos)?
- soma de quadrados de alguma percentagem dos (menores) resíduos?

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.

Começemos 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):

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

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 $\frac{\text{unidades de } y}{\text{unidades de } x}$.

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 (\bar{x}, \bar{y}) , como é evidente a partir da fórmula para a ordenada na origem:

$$b_0 = \bar{y} - b_1 \bar{x} \quad \Leftrightarrow \quad \bar{y} = b_0 + b_1 \bar{x} .$$

- \bar{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, $\bar{e} = 0$. (Ver Exercício RLS 5).

Revisão: RLS - As três Somas de Quadrados

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

Soma de Quadrados Total (SQT)

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

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

Soma de Quadrados da Regressão (SQR)

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

Soma de Quadrados Residual (SQRE) - já dado

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

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

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

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”)

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

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 $\hat{y}_i = \bar{y}$, para todo o i . A recta é $y = \bar{y} \Leftrightarrow b_1 = 0$)

- Numa regressão linear **simples**, R^2 é o quadrado do coeficiente de correlação linear entre x e y (Ver Exercício RLS 6):

$$R^2 = r_{xy}^2 = \left(\frac{COV_{xy}}{S_x S_y} \right)^2 \quad \text{se } S_x \neq 0 \text{ e } S_y \neq 0$$

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
```

Regressão - um pouco de história

O **critério de mínimos quadrados** surge no início do Século XIX, associado ao trabalho do francês Legendre, motivado pelo problema de conciliar diferentes observações geodésicas e astronómicas que se sabia estarem afectadas por erros de observação.

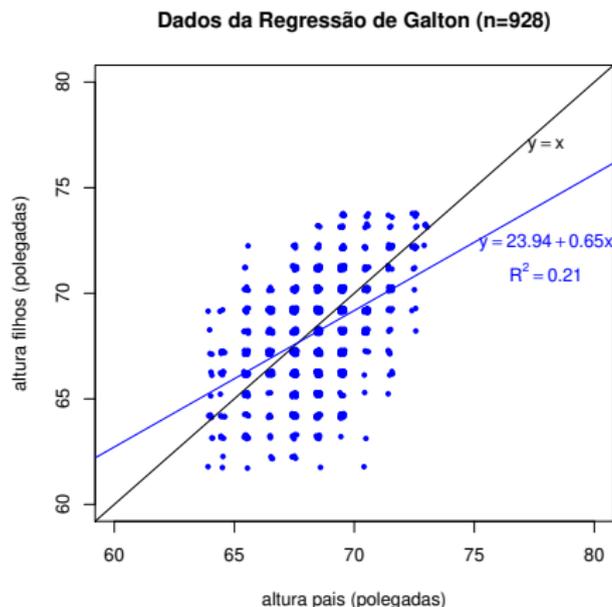
A designação **Regressão** tem origem num estudo de Francis Galton (1886), relacionando a altura de $n = 928$ jovens adultos com a altura (média) dos pais. Galton inventou a designação **eugenia**, conceito que era considerado respeitável até às primeiras décadas do Século XX.

Galton constatou que pais com alturas acima da média tinham tendência a ter filhos com altura acima da média - mas menos que os pais (análogo para os abaixo da média).

Galton chamou ao seu artigo *Regression towards mediocrity in hereditary stature*. A expressão **regressão** ficou associada ao método devido a esta acaso histórico.

Um pouco de história (cont.)

Curiosamente, o exemplo de Galton tem um valor muito baixo do Coeficiente de Determinação.



Algumas ideias prévias sobre modelação

- Todos os modelos são apenas **aproximações** da realidade.
- Pode haver mais do que um modelo adequado a uma relação. Um dado modelo pode ser melhor num aspecto, mas pior noutro.
- O **princípio da parcimónia** na modelação: de entre os modelos considerados **adequados**, é preferível o **mais simples**.
- Os modelos **estatísticos** apenas descrevem **tendência de fundo**: há **variação** das observações em torno da tendência de fundo.
- Num modelo estatístico **não há necessariamente uma relação de causa e efeito entre variável resposta e preditores**. Há apenas **associação**. A eventual existência de uma relação de causa e efeito só pode ser **justificada por argumentos extra-estatísticos**.

Transformações linearizantes

Nalguns casos, a relação de fundo entre x e y é não-linear, mas pode ser linearizada caso se proceda a transformações numa ou em ambas as variáveis.

Tais transformações podem permitir utilizar a Regressão Linear Simples, apesar de a relação original ser não-linear.

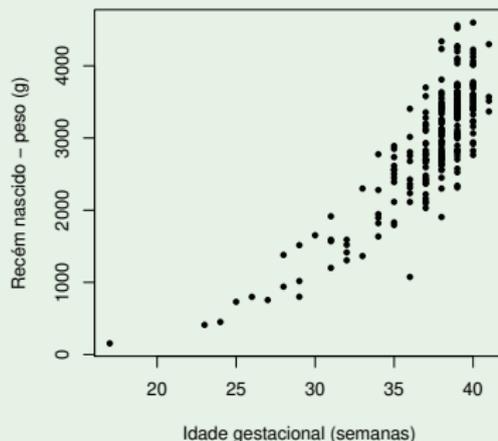
Vamos ver dois exemplos particularmente frequentes de relações não-lineares que são linearizáveis através de transformações da variável resposta e, num caso, também do preditor.

Exemplo 3 - Uma relação não linear

Peso de bebés à nascença

$n = 251$ pares de observações

Idade gestacional (x) e peso de bebé à nascença y , $\{(x_i, y_i)\}_{i=1}^{251}$.



A tendência de fundo é **não-linear**: $y = f(x)$.

Exemplo 3 (cont.)

Neste caso, há uma **questão adicional**:

- Qual a **forma da relação** (qual a natureza da função f)?
 - ▶ f exponencial ($y = ce^{dx}$)?
 - ▶ f função potência ($y = cx^d$)?
 - ▶ outra?

Além das perguntas análogas ao caso linear:

- Como determinar os “melhores” **parâmetros c e d** ?
- E, se os dados forem amostra aleatória, **o que se pode dizer sobre os respectivos parâmetros populacionais?**

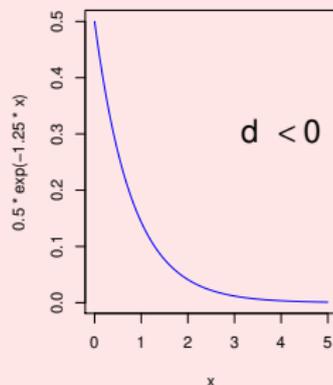
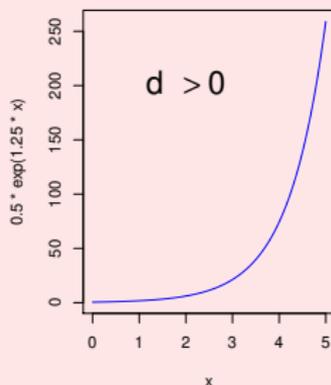
A **Regressão Não Linear** **não** faz parte do programa da disciplina. Mas **transformações linearizantes** de uma ou ambas as variáveis podem criar uma relação linear, que permita usar o Modelo Linear.

Relação exponencial

Relação exponencial

$$y = ce^{dx}$$

$$(y > 0 ; c > 0)$$



Transformação linearizante: $y^* = \ln(y)$ e $x^* = x$

A linearização da relação exponencial

Logaritmizando a equação da exponencial, obtém-se uma **relação linear** entre $y^* = \ln(Y)$ e x :

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

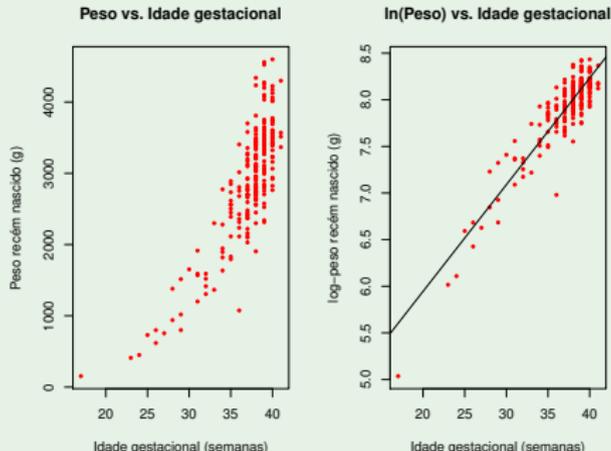
com **declive** $b_1 = d$ e **ordenada na origem** $b_0 = \ln(c)$.

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

Exemplo 3: peso de bebés à nascença

Uma linearização no peso dos bebés

O gráfico de **log-pesos** dos recém-nascidos contra idade gestacional produz uma relação de fundo linear:



Esta linearização significa que a relação original (peso vs. idade gestacional) pode ser considerada exponencial.

Ainda a relação exponencial

Equação Diferencial da exponencial

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

$$y'(x) = d \cdot y(x) ,$$

isto é, que a taxa de variação relativa de y é constante:

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

Primitivando em ordem a x ($P \frac{f'}{f} = \ln|f|$), tem-se (já que $y > 0$):

$$\ln|y(x)| = dx + K \quad \Leftrightarrow \quad y(x) = e^{K+dx} \quad \Leftrightarrow \quad y(x) = e^K e^{dx} = ce^{dx} .$$

O declive b_1 da recta é o valor constante d da taxa de variação relativa de y .

Modelo exponencial de crescimento populacional

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

Mas nenhum crescimento populacional exponencial é sustentável a longo prazo.

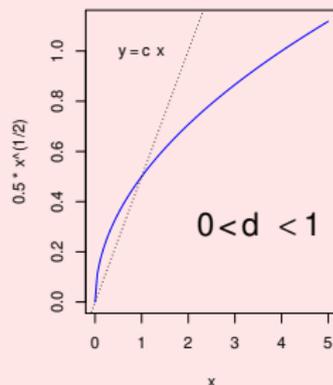
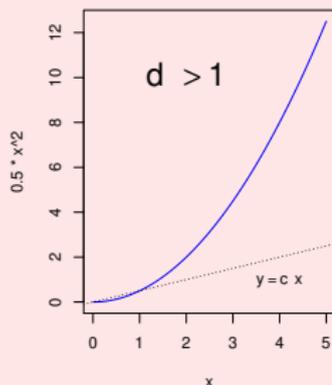
Em **1838 Verhulst** propôs um **modelo de crescimento populacional alternativo**, prevendo os efeitos resultantes da escassez de recursos: o **modelo logístico** (não leccionado).

Relação potência ou alométrica

Relação potência

$$y = cX^d$$

$$(x, y > 0 ; c, d > 0)$$



Transformação linearizante: $y^* = \ln(y)$ e $x^* = \ln(x)$.

A linearização dum relação potência

Logaritmizando, obtém-se:

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

que é uma **relação linear entre $y^* = \ln(y)$ e $x^* = \ln(x)$** .

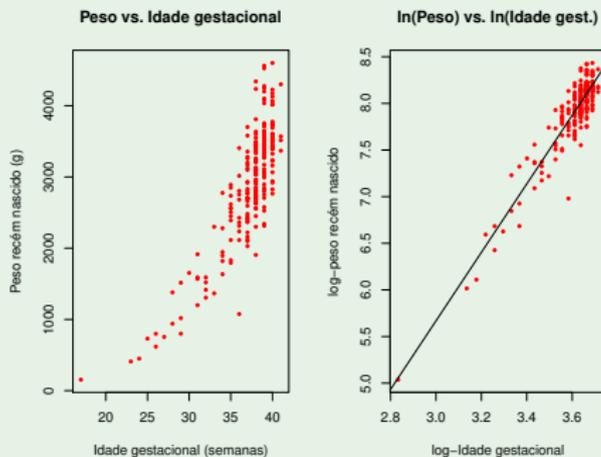
O declive b_1 da recta é o expoente d na relação potência.

A ordenada na origem é $b_0 = \ln(c)$, ou seja, $c = e^{b_0}$.

Outra linearização no Exemplo 3

Outra linearização dos pesos dos bebês

O gráfico de **log-pesos** dos recém-nascidos contra **log-idade gestacional** produz outra **relação de fundo linear**:



Esta linearização significa que a relação original (peso vs. idade gestacional) **também** pode ser considerada uma relação potência.

Ainda a relação potência

Uma Equação Diferencial da potência

Uma relação potência resulta de admitir que y é função de x e a taxa de variação relativa de y , i.e., a razão $\frac{y'(x)}{y(x)}$, é inversamente proporcional a x :

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

Primitivando (em ordem a x), tem-se (pois $y > 0$ e $x > 0$):

$$\underbrace{\ln|y(x)|}_{=y^*} = \underbrace{d}_{=b_1} \underbrace{\ln|x|}_{=x^*} + \underbrace{K}_{=b_0} \quad \Leftrightarrow \quad y(x) = e^{K+\ln(x^d)} \quad \Leftrightarrow \quad y(x) = e^K x^d.$$

O declive b_1 da recta é a constante de proporcionalidade d .

A constante de primitivação K é a ordenada na origem da recta: $K = b_0$.

O contexto alométrico da relação potência

A Equação Diferencial da alometria

Outra forma de obter uma relação potência, muito usada nos estudos alométricos, resulta de admitir que y e x são ambas funções duma terceira variável t (ou seja, $y(t)$ e $x(t)$) e que as taxas de variação relativas de y e x são proporcionais:

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

Primitivando (em ordem a t) tem-se:

$$\ln y = d \ln x + K$$

e exponenciando,

$$y = e^{d \ln x + K} = e^{d \ln x} \cdot e^K = e^{\ln x^d} \cdot \underbrace{e^K}_{=c} \Leftrightarrow y = c x^d.$$

Os estudos de alometria comparam a dimensão de partes diferentes dum organismo. A isometria corresponde ao valor $d=1$.

Advertência sobre transformações linearizantes

A regressão linear simples **não** modela **directamente** relações **não lineares** entre x e y . Pode modelar **uma relação linear** entre as **variáveis transformadas**.

Transformações da **variável-resposta y** têm um impacto grande no ajustamento: **a escala dos resíduos é alterada**.

Conceitos que dependem da escala de Y , como **$SQRE$ e R^2** , **não são directamente comparáveis**, antes e após uma transformação da variável resposta.

Nota: Linearizar, obter os parâmetros b_0 e b_1 da recta e depois desfazer a transformação linearizante **não** produz os mesmos parâmetros ajustados que resultariam de minimizar a soma de quadrados dos resíduos **directamente** na relação não linear, através duma **regressão não linear** (não estudada nesta disciplina).

Regressão Linear Simples - INFERÊNCIA

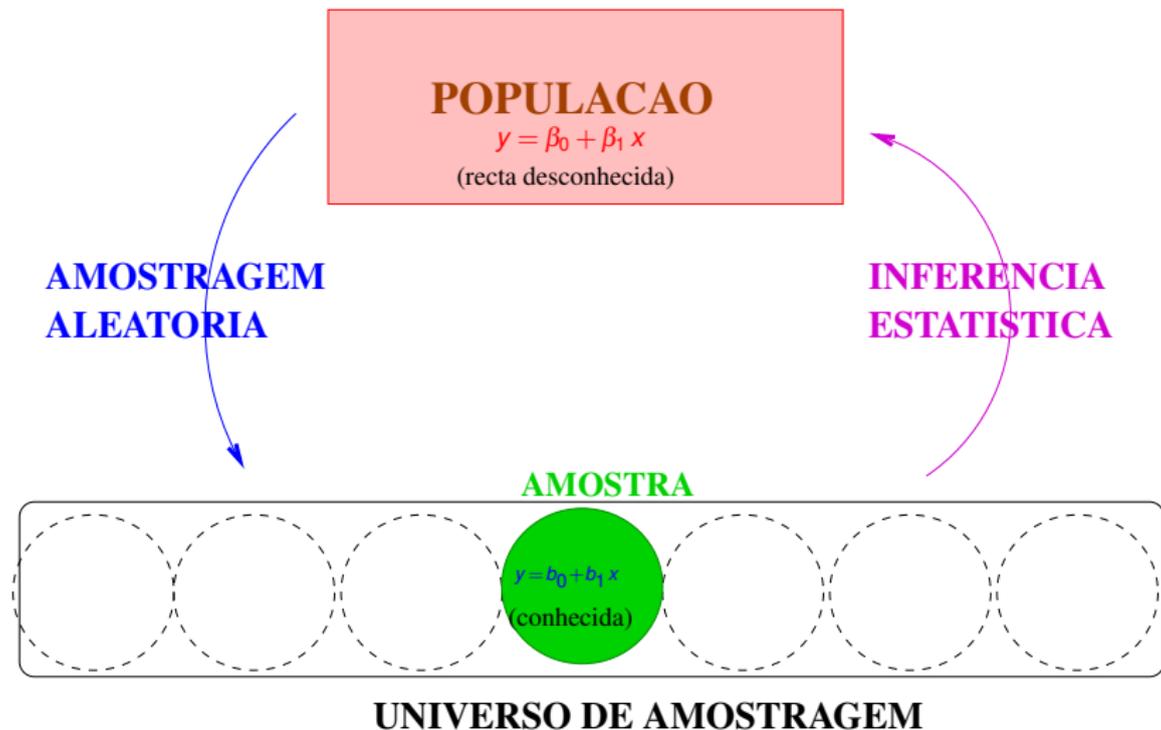
- Até aqui a RLS foi usada apenas como **técnica descritiva**.
Se as n observações fossem a totalidade da população de interesse, pouco mais haveria a dizer.
Mas, com frequência, as n observações são apenas uma **amostra aleatória** de uma população maior.
- A recta de regressão $y = b_0 + b_1 x$ obtida com base na **amostra** é apenas uma **estimativa** de uma **recta populacional**

$$y = \beta_0 + \beta_1 x .$$

Outras amostras dariam outras rectas ajustadas (estimadas).

- Coloca-se o problema da **inferência estatística**.

O problema da Inferência Estatística na RLS



MODELO - Regressão Linear Simples

A fim de se poder fazer inferência sobre a recta populacional, admitem-se **pressupostos adicionais**.

Y – variável resposta **aleatória**.

x – variável preditora **não aleatória** (fixada pelo experimentador ou trabalha-se **condicionalmente** aos valores de x)

Recordar: Uma **variável aleatória** é o conceito que formaliza a realização de experiências aleatórias com resultado numérico.

O modelo será ajustado com base em:

$\{(x_i, Y_i)\}_{i=1}^n$ – n pares de observações de x e Y , sobre n unidades experimentais.

MODELO RLS – Linearidade

Vamos ainda admitir que a relação de fundo entre as variáveis x e Y é linear, com uma variabilidade aleatória em torno dessa relação de fundo, representada por um erro aleatório ε :

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

para todo o $i = 1, \dots, n$.

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

MODELO RLS – Os erros aleatórios

Vamos ainda admitir que os erros aleatórios ε_j :

- Têm valor esperado (valor médio) nulo:

$$E[\varepsilon_j] = 0, \quad \forall j = 1, \dots, n$$

(não é hipótese restritiva).

- Têm distribuição Normal (é restritiva, mas bastante geral).
- Homogeneidade de variâncias: têm sempre a mesma variância

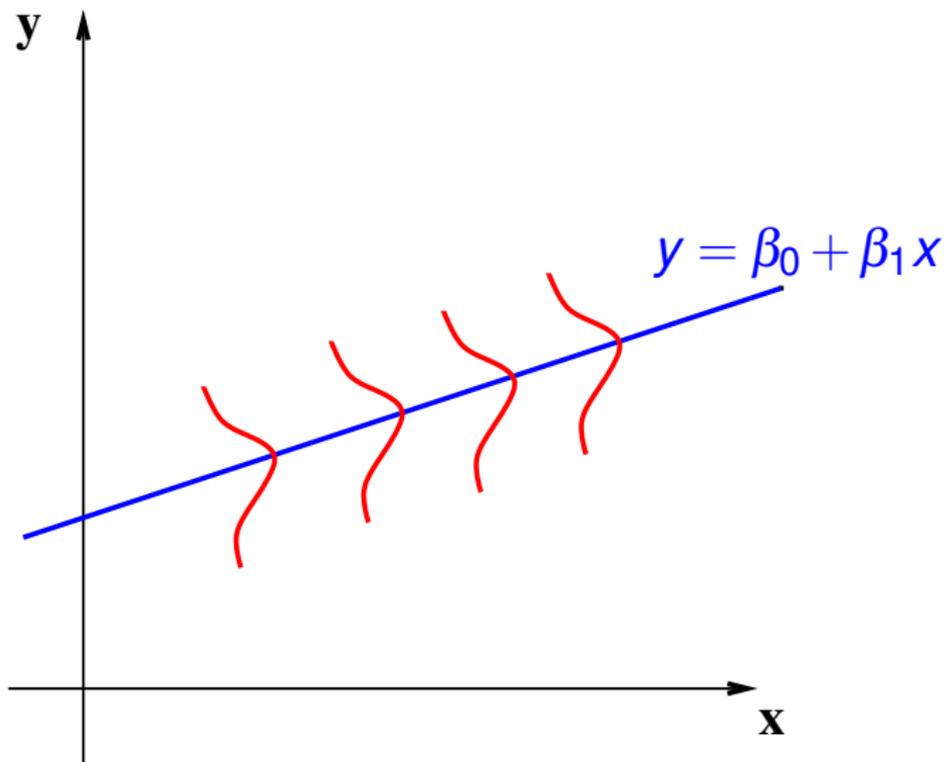
$$V[\varepsilon_j] = \sigma^2, \quad \forall j = 1, \dots, n$$

(é restritiva, mas conveniente).

Ou seja, admite-se que $\varepsilon_j \sim \mathcal{N}(0, \sigma^2)$, para todo o j .

- São variáveis aleatórias independentes (é restritiva, mas conveniente).

MODELO Regressão Linear Simples



MODELO - Regressão Linear Simples

Recapitulando, para efeitos de inferência estatística, admite-se:

O Modelo de Regressão Linear Simples

Temos n pares de observações $\{(x_i, Y_i)\}_{i=1}^n$, tais que:

- 1 $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\forall i = 1, \dots, n$.
- 2 $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, $\forall i = 1, \dots, n$.
- 3 $\{\varepsilon_i\}_{i=1}^n$ v.a. independentes.

NOTA: Nesta disciplina segue-se a convenção que o segundo parâmetro duma Normal é a sua **variância**.

NOTA: Os erros aleatórios são variáveis aleatórias independentes e identicamente distribuídas (i.i.d.).

NOTA: A validade da inferência que se segue depende da validade destes pressupostos do modelo.

Caracterização variáveis aleatórias

Variáveis aleatórias (v.a.) podem ser:

- **Discretas** – tomam um número finito ou infinidade numerável de possíveis valores, x_i (por exemplo, contagens);
- **Contínuas** – tomam valores em intervalos (infinidade não numerável de possíveis valores) (por exemplo, rendimentos).

Ao admitir-se a Normalidade, admite-se que as variáveis resposta que modelamos são contínuas.

Uma v.a. contínua X é caracterizada pela sua **função densidade** $f(x)$, através da qual é possível **calcular probabilidades** de X tomar valores num **dado intervalo** (ver Exercícios Introdutórios 4 e 5):

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

Revisão: valores esperados

O **valor esperado** ou **valor médio** dum a variável aleatória X é o **centro** de gravidade da sua distribuição de probabilidades.

Algumas **propriedades dos valores esperados** de variáveis aleatórias:

Propriedades dos valores esperados

Sejam X e Y variáveis aleatórias e a e b constantes. Então:

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

Revisão: variâncias

A **variância** duma v.a. mede a **dispersão** da sua distribuição. Define-se como:

$$V[X] = E[(X - E[X])^2] = E[X^2] - E^2[X]$$

Propriedades da variância de variáveis aleatórias

Sejam X e Y variáveis aleatórias e a e b constantes. Então:

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

Revisão: covariância

A **covariância** entre duas v.a. mede o grau de **relacionamento linear** entre elas e define-se como:

$$\text{Cov}[X, Y] = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

Propriedades da covariância de variáveis aleatórias

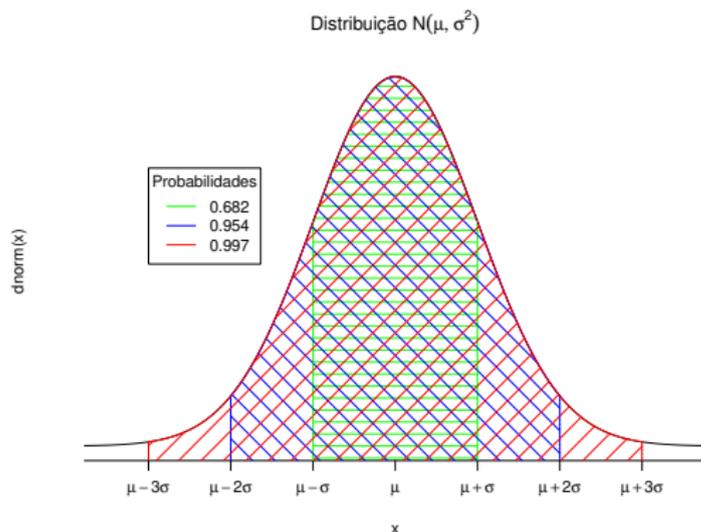
Sejam X e Y variáveis aleatórias e a e b constantes. Então:

- $\text{Cov}[X, Y] = \text{Cov}[Y, X]$.
- $\text{Cov}[X, X] = V[X]$.
- $\text{Cov}[X + a, Y + b] = \text{Cov}[X, Y]$.
- $\text{Cov}[aX, bY] = ab \text{Cov}[X, Y]$.
- $\text{Cov}\left[\sum_{i=1}^n X_i, Y\right] = \sum_{i=1}^n \text{Cov}[X_i, Y]$
- $|\text{Cov}[X, Y]| \leq \sqrt{V[X]V[Y]}$ (Desigualdade de Cauchy-Schwarz).
- **Se X, Y são v.a. independentes**, então $\text{Cov}[X, Y] = 0$.

Revisão: a distribuição Normal

Se a v.a. X tem distribuição Normal, com valor esperado μ e variância σ^2 , escreve-se: $X \sim \mathcal{N}(\mu, \sigma^2)$.

Atenção à convenção nesta UC: o segundo parâmetro é a **variância**.



Propriedades da Normal

Propriedades da distribuição Normal

- Uma **transformação linear** numa Normal tem distribuição Normal. Concretamente: seja $X \sim \mathcal{N}(\mu, \sigma^2)$ e a, b constantes. Então:

$$a + bX \sim \mathcal{N}(a + b\mu, b^2\sigma^2).$$

- Seja $X \sim \mathcal{N}(\mu, \sigma^2)$, então: $\frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$.
- **Combinações lineares de Normais independentes têm distribuição Normal**: se X, Y são Normais independentes e a, b constantes, então $aX + bY$ é Normal (com parâmetros resultantes das propriedades dos acetatos 64 e 65).

Primeiras consequências do MODELO RLS

O modelo RLS obriga a que as observações da variável resposta Y sejam independentes, com distribuição Normal:

Primeiras consequências do Modelo

Dado o Modelo da Regressão Linear Simples, tem-se

- 1 $E[Y_i] = \beta_0 + \beta_1 x_i, \quad \forall i = 1, \dots, n.$
- 2 $V[Y_i] = \sigma^2, \quad \forall i = 1, \dots, n.$
- 3 $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2), \quad \forall i = 1, \dots, n.$
- 4 $\{Y_i\}_{i=1}^n$ v.a. independentes.

NOTA: As observações da variável resposta Y_i não são i.i.d.: embora sejam independentes, normais e de variâncias iguais, as suas médias são diferentes (dependem dos valores de $x = x_i$ associados às observações).

Estimação dos parâmetros do Modelo RLS

A recta do modelo RLS tem dois parâmetros: β_0 e β_1 .

Definem-se **estimadores** desses parâmetros a partir das expressões amostrais obtidas para b_0 e b_1 pelo Método dos Mínimos Quadrados.

$$\text{Recordar: } b_1 = \frac{\text{cov}_{xy}}{s_x^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1) s_x^2} \stackrel{(*)}{=} \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{(n-1) s_x^2} = \sum_{i=1}^n \frac{x_i - \bar{x}}{(n-1) s_x^2} y_i$$

(*) Veja-se o Exercício RLS 3b).

Estimador de β_1

$$\hat{\beta}_1 = \sum_{i=1}^n \frac{x_i - \bar{x}}{(n-1) s_x^2} Y_i = \sum_{i=1}^n c_i Y_i, \quad \text{com } c_i = \frac{x_i - \bar{x}}{(n-1) s_x^2}$$

Nota: O estimador $\hat{\beta}_1$ é combinação linear de Normais independentes, logo tem distribuição Normal.

Estimação dos parâmetros do Modelo RLS (cont.)

Recordar: $b_0 = \bar{y} - b_1 \bar{x}$.

Estimador de β_0

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \frac{1}{n} \sum_{i=1}^n Y_i - \bar{X} \sum_{i=1}^n c_i Y_i = \sum_{i=1}^n \left(\frac{1}{n} - \bar{X} c_i \right) Y_i = \sum_{i=1}^n d_i Y_i,$$

com

$$d_i = \frac{1}{n} - \bar{X} c_i = \frac{1}{n} - \frac{(x_i - \bar{X}) \bar{X}}{(n-1) s_X^2}.$$

Quer $\hat{\beta}_1$, quer $\hat{\beta}_0$, são **combinações lineares** das observações $\{Y_i\}_{i=1}^n$, logo são **combinações lineares de variáveis aleatórias Normais independentes**. Logo, **ambos os estimadores têm distribuição Normal**.

Distribuição dos estimadores RLS

Distribuição dos estimadores dos parâmetros

Dado o Modelo de Regressão Linear Simples,

$$\textcircled{1} \hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{(n-1)s_x^2}\right),$$

$$\textcircled{2} \hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right]\right)$$

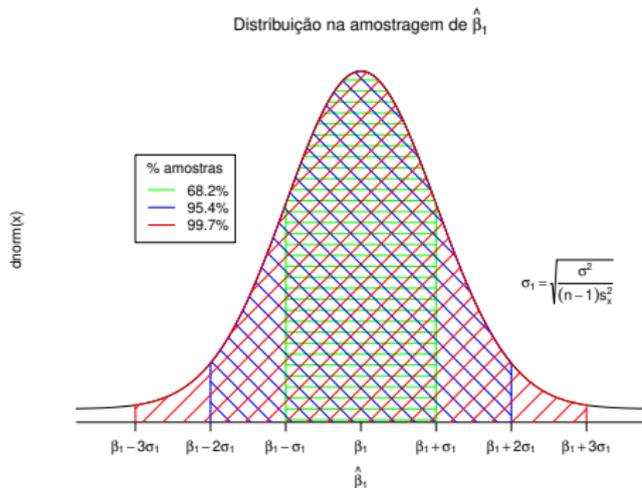
NOTAS:

- 1 Ambos os estimadores são **centrados**: $E[\hat{\beta}_1] = \beta_1$ e $E[\hat{\beta}_0] = \beta_0$.
- 2 Quanto maior $(n-1)s_x^2$, menor a variância dos estimadores.
- 3 A variância de $\hat{\beta}_0$ também diminui com o aumento de n , e com a maior proximidade de \bar{x} de zero.

Significado das distribuições dos estimadores

Interpretação do resultado distribucional do estimador $\hat{\beta}_1$:

se fossem recolhidas todas as possíveis amostras aleatórias de dimensão n (para os valores de x_i fixados), e para cada uma calculado o declive b_1 da recta amostral, a distribuição de frequências desses declives amostrais seria:



Distância da estimativa b_1 a β_1 :

- $< \sigma_{\hat{\beta}_1}$ em $\approx 68\%$ das amostras;
- $< 2\sigma_{\hat{\beta}_1}$ em $\approx 95\%$ das amostras;
- $< 3\sigma_{\hat{\beta}_1}$ em $\approx 99,7\%$ das amostras.

Distribuição dos estimadores RLS

Distribuição dos estimadores (cont.)

Dado o Modelo de Regressão Linear Simples,

$$\textcircled{1} \quad \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \sim \mathcal{N}(0, 1), \quad \text{com } \sigma_{\hat{\beta}_1} = \sqrt{\frac{\sigma^2}{(n-1)S_x^2}} = \frac{\sigma}{\sqrt{(n-1)S_x^2}}$$

$$\textcircled{2} \quad \frac{\hat{\beta}_0 - \beta_0}{\sigma_{\hat{\beta}_0}} \sim \mathcal{N}(0, 1), \quad \text{com } \sigma_{\hat{\beta}_0} = \sqrt{\sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{(n-1)S_x^2} \right]} = \sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)S_x^2}}$$

NOTAS:

- O desvio padrão dum estimador designa-se **erro padrão** (em inglês, *standard error*).
- Não confundir os erros padrão dos estimadores, $\sigma_{\hat{\beta}_1}$ e $\sigma_{\hat{\beta}_0}$, com o desvio padrão σ dos erros aleatórios.

Distribuição dos estimadores RLS

Os resultados do acetato anterior só permitem a inferência sobre os parâmetros β_0 e β_1 (e.g., construir intervalos de confiança ou efectuar testes de hipóteses) caso seja conhecida a **variância dos erros aleatórios**, $\sigma^2 = V[\varepsilon_i]$, que aparece nas expressões de $\sigma_{\hat{\beta}_1}$ e $\sigma_{\hat{\beta}_0}$.

Mas σ^2 é, na prática, desconhecido. **Precisamos de um estimador da variância σ^2 dos erros aleatórios.**

Vamos construí-lo a partir dos **resíduos**.

Erros aleatórios e Resíduos

$$\begin{array}{ll} \text{Erros aleatórios} & \varepsilon_i = Y_i - (\beta_0 + \beta_1 x_i) \quad (\text{desconhecidos}) \\ \text{Resíduos (v.a.)} & E_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \quad (\text{conhecíveis}) \end{array}$$

Os resíduos são **preditores** (conhecíveis) dos erros (desconhecidos).
O numerador da variância amostral dos resíduos é

$$(n-1) s_e^2 = \sum_{i=1}^n E_i^2 = SQRE,$$

porque a média dos resíduos é zero.

É natural que na estimação da variância (comum) dos erros aleatórios ε_i se utilize a variância amostral dos resíduos ou a Soma de Quadrados Residual, *SQRE*.

O Quadrado Médio Residual

Quadrado Médio Residual

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

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

QMRE é estimador centrado de σ^2

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

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

e tem-se $E[QMRE] = \sigma^2$.

Quantidades centrais para a inferência sobre β_0 e β_1

A estimação dos erros padrão com o QMRE transforma as distribuições Normais do slide 74 em distribuições *t-Student*.

Distribuições *t-Student* para a inferência sobre β_0 e β_1

Dado o Modelo de Regressão Linear Simples, prova-se que:

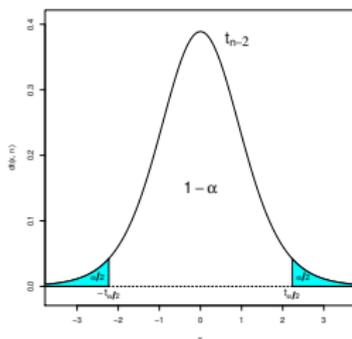
$$1 \quad \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t_{n-2}, \quad \text{com } \hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{QMRE}{(n-1)S_x^2}}$$

$$2 \quad \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \sim t_{n-2}, \quad \text{com } \hat{\sigma}_{\hat{\beta}_0} = \sqrt{QMRE \left[\frac{1}{n} + \frac{\bar{x}^2}{(n-1)S_x^2} \right]}$$

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

Dedução de intervalo de confiança para β_1

Sabemos que $\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t_{n-2}$. Seja $t_{\frac{\alpha}{2}}$ tal que $P[t_{n-2} > t_{\frac{\alpha}{2}}] = \frac{\alpha}{2}$, ou seja, o quantil de ordem $1 - \frac{\alpha}{2}$. Pela simetria da distribuição t-Student, tem-se $P[t_{n-2} < -t_{\frac{\alpha}{2}}] = \frac{\alpha}{2}$.



Tem-se:

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

Aviso: Como em v.a. contínuas a probabilidade de um único valor é sempre nula, é indiferente considerar desigualdades estritas ou não estritas.

Dedução IC para β_1 (cont.)

Trabalhar a dupla desigualdade até isolar β_1 :

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

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

O seguinte **intervalo aleatório** contém β_1 com probabilidade $1 - \alpha$:

$$\left] \hat{\beta}_1 - t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\beta}_1} , \hat{\beta}_1 + t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\beta}_1} \left[$$

Os extremos deste intervalo são variáveis aleatórias.

Interpretação IC para β_1

Como interpretar a conclusão que o **intervalo aleatório**

$$\left] \hat{\beta}_1 - t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\beta}_1}, \hat{\beta}_1 + t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\beta}_1} \left[$$

contém β_1 com probabilidade $1 - \alpha$?

- **A cada amostra concreta corresponde um intervalo concreto,**

$$\left] b_1 - t_{\frac{\alpha}{2} [n-2]} \hat{\sigma}_{\hat{\beta}_1}, b_1 + t_{\frac{\alpha}{2} [n-2]} \hat{\sigma}_{\hat{\beta}_1} \left[$$

- **$(1 - \alpha) \times 100\%$ desses intervalos concretos, para todas as possíveis amostras (de dimensão n e valores de x_i dados) contém o verdadeiro valor de β_1 ; Os restantes $\alpha \times 100\%$ não contém β_1 ;**
- Ao seleccionar **uma** amostra concreta, seleccionamos **um** intervalo concreto, tendo $(1 - \alpha) \times 100\%$ de **confiança** em como contém β_1 .

Intervalo de confiança para β_1

Intervalo de Confiança a $(1 - \alpha) \times 100\%$ para β_1

Dado o Modelo RLS, um intervalo a $(1 - \alpha) \times 100\%$ de confiança para o declive β_1 da recta de regressão populacional é dado por:

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

tendo $t_{\frac{\alpha}{2}[n-2]}$, b_1 e $\hat{\sigma}_{\hat{\beta}_1}$ sido definidos em acetatos anteriores.

NOTAS:

- O intervalo é **centrado em b_1** .
- A **amplitude do intervalo** é $2 \times t_{\frac{\alpha}{2}[n-2]} \hat{\sigma}_{\hat{\beta}_1}$.
- A amplitude **aumenta com QMRE** e **diminui com n e s_x^2** : $\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{QMRE}{(n-1)s_x^2}}$
- A amplitude do IC **aumenta para maiores graus de confiança $1 - \alpha$** .

Intervalo de confiança para β_0

Intervalo de Confiança a $(1 - \alpha) \times 100\%$ para β_0

Dado o Modelo de Regressão Linear Simples, um intervalo a $(1 - \alpha) \times 100\%$ de confiança para a ordenada na origem, β_0 , da recta populacional é:

$$\left] b_0 - t_{\frac{\alpha}{2}[n-2]} \cdot \hat{\sigma}_{\hat{\beta}_0} \quad , \quad b_0 + t_{\frac{\alpha}{2}[n-2]} \cdot \hat{\sigma}_{\hat{\beta}_0} \left[, \right.$$

onde $t_{\frac{\alpha}{2}[n-2]}$, b_0 e $\hat{\sigma}_{\hat{\beta}_0}$ foram definidos em acetatos anteriores.

NOTAS:

- O intervalo é centrado em b_0 .
- A amplitude do intervalo é $2 \times t_{\frac{\alpha}{2}[n-2]} \hat{\sigma}_{\hat{\beta}_0}$.
- A amplitude aumenta com *QMRE* e com \bar{x}^2 e diminui com n e s_x^2 :

$$\hat{\sigma}_{\hat{\beta}_0} = \sqrt{QMRE \cdot \left[\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} \right]}$$

- A amplitude do IC aumenta para maiores graus de confiança $1 - \alpha$.

Intervalos de confiança no R

O exemplo das cerejeiras

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

```
> summary(cerejeiras.lm)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-1.046122	0.095290	-10.98	7.62e-12	***
DAP	0.056476	0.002758	20.48	< 2e-16	***

Na segunda coluna da listagem de saída, são indicados os valores dos **erros padrões estimados**, para cada estimador:

$$\hat{\sigma}_{\hat{\beta}_0} = 0.095290 \qquad \hat{\sigma}_{\hat{\beta}_1} = 0.002758 .$$

Estes valores são usados nos intervalos de confiança para β_0 e β_1 .

Intervalos de confiança de β_0 e β_1 no R

As cerejeiras (cont.)

Para calcular, no R, os intervalos de confiança numa regressão ajustada, usa-se a função `confint`:

```
> confint(cerejeiras.lm)
                2.5 %      97.5 %
(Intercept) -1.24101280 -0.85123173  <- ordenada na origem
DAP          0.05083558  0.06211645  <- declive
```

Por omissão, o IC calculado é a 95% de confiança.

A 95% de confiança, o declive β_1 da recta populacional está no intervalo]0.051, 0.062[e a ordenada na origem β_0 no intervalo] -1.241, -0.851 [.

O nível de confiança pode ser mudado com o argumento `level`:

```
> confint(cerejeiras.lm, level=0.90)
                5 %      95 %
(Intercept) -1.20803258 -0.88421195
DAP          0.05179008  0.06116195
```

Um alerta sobre Intervalos de Confiança

Tal como na construção de qualquer intervalo de confiança, existem dois aspectos contrastantes:

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

Quanto maior o grau de confiança $(1 - \alpha) \times 100\%$ dum intervalo, maior será a sua amplitude, isto é, menor será a sua precisão.

Os mesmos resultados que serviram de base à construção dos intervalos de confiança vão agora ser usados para outro fim: efectuar testes de hipóteses a valores dos parâmetros populacionais β_0 e β_1 .

Revisão de Testes de Hipóteses: Passo 1

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

Passo 1: formular as hipóteses em confronto

- Hipótese Nula H_0 vs.
- Hipótese Alternativa H_1

Exemplo: declive β_1 duma recta de regressão populacional

O objectivo é testar alguma afirmação sobre o valor de β_1 . Por exemplo,

- Hipótese Nula $H_0: \beta_1 = 1$
- Hipótese Alternativa $H_1: \beta_1 \neq 1$

Testes de Hipóteses: Passo 2

Passo 2: estatística de teste

Como optar entre H_0 e H_1 ? Através duma **estatística de teste**, que é:

- quantidade **numérica**, cujo valor **depende apenas** da amostra e de H_0 .
- com **distribuição de probabilidades conhecida**, se H_0 verdade.

Para alguns valores da estatística rejeita-se H_0 , para outros não.

Exemplo: estatística de teste para Hipóteses sobre β_1

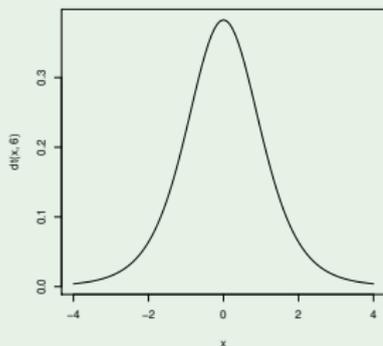
A estatística de teste é a quantidade que serviu de ponto de partida para a construção dos ICs, admitindo H_0 :

$$T = \frac{\hat{\beta}_1 - \overbrace{\beta_1}_{=1}|_{H_0}}{\hat{\sigma}_{\hat{\beta}_1}} \sim t_{n-2}, \text{ se } H_0 \text{ verdade}$$

Testes de Hipóteses (cont.)

Exemplo: como ler os valores calculados da estatística de teste

$$T = \frac{\hat{\beta}_1 - \beta_{1|H_0}}{\hat{\sigma}_{\hat{\beta}_1}} \sim t_{n-2}$$



Numa amostra com $b_1 \approx \beta_{1|H_0}$, não há razões para duvidar de H_0 . Neste caso, a estatística T toma valor $T_{calc} = \frac{b_1 - \beta_{1|H_0}}{\hat{\sigma}_{\hat{\beta}_1}} \approx 0$.

Pelo contrário, se b_1 for muito diferente de $\beta_{1|H_0}$, há razões para duvidar de H_0 . Aí, T_{calc} será (em módulo) grande.

Assim, um hipotético valor $\beta_{1|H_0}$ é plausível se T_{calc} for próximo de zero.

Quanto maior $|T_{calc}|$, menos plausível é H_0 .

Testes de Hipóteses: (cont.)

Como definir a fronteira entre os valores da estatística que levam à rejeição, ou não, de H_0 ?

Há que distinguir entre:

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

Existem **quatro possíveis situações**:

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

Testes de Hipóteses: Passo 3

Não é possível reduzir simultaneamente a probabilidade dos dois erros: diminuir $P[\text{Erro Tipo I}]$ significa reduzir a gama de valores que levam à rejeição de H_0 , aumentando $P[\text{Erro Tipo II}]$.

Procedimento: admitir que o Erro de Tipo I é o mais grave e controlá-lo.

Passo 3: nível de significância do teste

Define-se o nível de significância do teste, α :

$$\alpha = P[\text{Erro de Tipo I}] = P[\text{Rejeitar } H_0 \mid H_0 \text{ verdade}].$$

α define o tamanho da região crítica. Sendo a probabilidade dum erro, queremos α pequeno. Valores usuais são $\alpha = 0.05$, $\alpha = 0.01$.

O papel das duas hipóteses em confronto não é simétrico.

- Hipótese Nula H_0 tem o benefício da dúvida.
- Hipótese Alternativa H_1 tem o ónus da prova.

Testes de Hipóteses: Passo 4

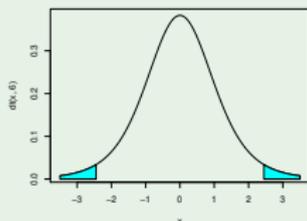
Passo 4: Região Crítica (ou de Rejeição)

É o conjunto de valores possíveis da estatística:

- ao qual associamos a **rejeição de H_0** ;
- é constituída pelos valores “menos plausíveis”, caso seja verdade H_0 (pode ser **bilateral** ou **unilateral**, dependendo de H_1);
- é uma **região de probabilidade α** , se for verdade H_0 .

Exemplo: Região Crítica bilateral (adequada ao exemplo)

Rejeitar $H_0 : \beta_1 = 1$ se $|T_{calc}| > t_{\frac{\alpha}{2}(n-2)}$.



Testes de Hipóteses: Passo 5

Passo 5: Conclusões

- Escolher uma amostra concreta;
- Calcular o valor da estatística para essa amostra;
- Decidir *rejeitar* H_0 ou *não rejeitar* H_0 , consoante o valor da estatística calculado para a amostra escolhida recaia, ou não, na Região Crítica.

É o único passo onde é preciso que existam dados.

Testes de hipóteses para o declive β_1

Sendo válido o Modelo de Regressão Linear Simples, tem-se:

Teste de Hipóteses a β_1 (Bilateral)

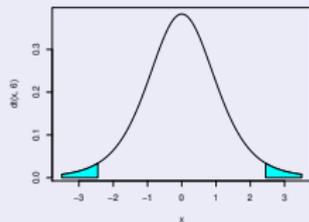
Hipóteses: $H_0 : \beta_1 = c$ vs. $H_1 : \beta_1 \neq c$.

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

Nível de significância do teste: $\alpha = P[\text{Rej. } H_0 | H_0 \text{ verdade}]$

Região Crítica (Região de Rejeição): Bilateral

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



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

Testes de hipóteses sobre o declive β_1

Hipóteses diferentes, que justificam uma RC unilateral direita:

Teste de Hipóteses a β_1 (Unilateral direito)

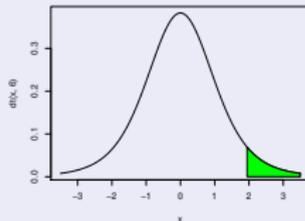
Hipóteses: $H_0 : \beta_1 \leq c$ vs. $H_1 : \beta_1 > c$.

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

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

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



Testes de hipóteses para o declive β_1

Hipóteses diferentes, que justificam uma RC unilateral esquerda:

Teste de Hipóteses a β_1 (Unilateral esquerdo)

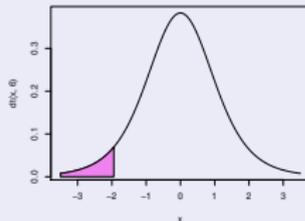
Hipóteses: $H_0 : \beta_1 \geq c$ vs. $H_1 : \beta_1 < c$.

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

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral esquerda

Rejeitar H_0 se $T_{calc} < -t_{\alpha(n-2)}$



Testes de hipóteses para a ordenada na origem β_0

Sendo válido o Modelo de Regressão Linear Simples, tem-se:

Testes de Hipóteses a β_0

$$\text{Hipóteses: } H_0 : \beta_0 \begin{matrix} \geq \\ \leq \end{matrix} c \quad \text{vs.} \quad H_1 : \beta_0 \begin{matrix} < \\ > \end{matrix} c$$

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

Nível de significância do teste: α

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

$$\begin{array}{ll} T_{calc} < -t_{\alpha(n-2)} & \text{(Unilateral esquerdo)} \\ \text{verifica: } |T_{calc}| > t_{\frac{\alpha}{2}(n-2)} & \text{(Bilateral)} \\ T_{calc} > t_{\alpha(n-2)} & \text{(Unilateral direito)} \end{array}$$

Os p -values

Em alternativa, os passos 3 a 5 podem ser substituídos pela indicação do **valor de prova** (p -value) associado ao valor calculado da estatística.

Valores de prova (p -value)

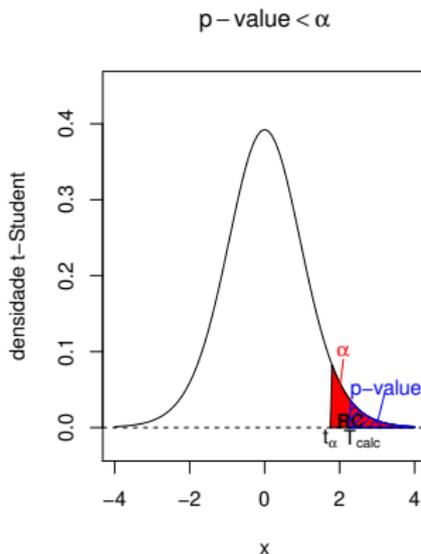
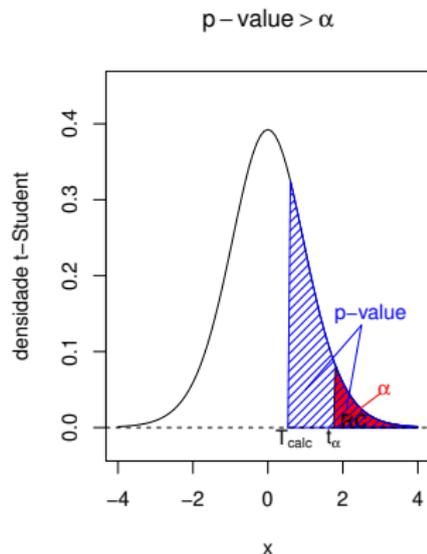
O p -value é a probabilidade da estatística de teste tomar valores mais extremos que o valor calculado a partir da amostra, sob H_0

Para um teste a β_1 , ou β_0 , com Região Crítica **unilateral direita**, o p -value é calculado como:

$$p = P[T > T_{calc}] \quad \text{com} \quad T \sim t_{n-2}$$

A relação de p -values e níveis de significância

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



Em geral: p -value muito pequeno implica rejeição H_0 .

O cálculo dum *p-value*

O cálculo do *p-value* é feito de forma diferente, consoante a natureza da Região Crítica (RC) (unilateral direita ou esquerda, ou bilateral).

Sendo $T \sim t_{n-2}$, tem-se:

RC Unilateral direita: $p = P[T > T_{calc}]$

RC Unilateral esquerda: $p = P[T < T_{calc}]$

RC Bilateral: $p = 2 \times P[T > |T_{calc}|]$.

Testes de hipóteses no

No R, a função `summary`, aplicada ao resultado dum comando `lm` produz a informação essencial para testes de hipóteses a β_0 e β_1 :

Estimate As estimativas b_0 e b_1

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

t value O valor calculado das estatísticas dos testes às hipóteses

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

ou seja,

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

Pr(>|t|) O valor de prova (*p-value*) associado a essa estatística de teste (com região crítica bilateral).

De novo o exemplo dos lírios

Ainda o exemplo dos lírios

```
> summary(cerejeiras.lm)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-1.046122	0.095290	-10.98	7.62e-12	***
DAP	0.056476	0.002758	20.48	< 2e-16	***

Num teste a $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$, a estatística de teste tem valor calculado

$$T_{calc} = \frac{b_1 - \overbrace{\beta_1|_{H_0}}^{=0}}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{0.056476}{0.002758} = 20.48,$$

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

O exemplo dos lírios (cont.)

Ainda o exemplo dos lírios (cont.)

Para testes a valores diferentes de zero dos parâmetros β_j , será preciso completar os cálculos do valor da estatística:

```
> summary(cerejeiras.lm)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-1.046122	0.095290	-10.98	7.62e-12	***
DAP	0.056476	0.002758	20.48	< 2e-16	***

Valor da estatística no teste $H_0 : \beta_1 = 0.05$ vs. $H_1 : \beta_1 \neq 0.05$:

$$T_{calc} = \frac{b_1 - \overbrace{\hat{\beta}_1|_{H_0}}^{=0.05}}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{0.056476 - 0.05}{0.002758} = 2.348078 .$$

Nota: O *p-value* da tabela também não é válido neste caso.

Inferência sobre $\mu_{Y|X} = E[Y|X=x]$

Consideremos agora outro problema inferencial de interesse geral: a inferência sobre o valor esperado da variável resposta Y , dado um valor x da variável preditora, ou seja, sobre o valor de Y na recta populacional, quando $X = x$:

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

Estimador de $\mu_{Y|X} = \beta_0 + \beta_1 x$

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

com c_i e d_i definidos nos acetatos 70 e 71.

Nota: O estimador $\hat{\mu}_{Y|X}$ é combinação linear das observações Y_i (que são Normais e independentes), logo tem distribuição Normal.

A distribuição do estimador de $\mu_{Y|x} = E[Y | X = x]$

Distribuição do estimador $\hat{\mu}_{Y|x}$

Dado o Modelo de Regressão Linear Simples, tem-se

$$\hat{\mu}_{Y|x} = \hat{\beta}_0 + \hat{\beta}_1 x \sim \mathcal{N} \left(\beta_0 + \beta_1 x, \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)S_x^2} \right] \right)$$
$$\Leftrightarrow \frac{\hat{\mu}_{Y|x} - \mu_{Y|x}}{\sigma_{\hat{\mu}_{Y|x}}} \sim \mathcal{N}(0, 1),$$

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

NOTA: Tal como para as distribuições iniciais de $\hat{\beta}_0$ e $\hat{\beta}_1$ (acetato 74), também esta distribuição não é ainda utilizável devido à presença da variância (desconhecida) dos erros aleatórios, σ^2 .

A distribuição para inferência sobre $\mu_{Y|x} = E[Y | X = x]$

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

Dado o Modelo de Regressão Linear Simples, tem-se

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

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

NOTA: A justificação desta distribuição é totalmente análoga à das distribuições de $\hat{\beta}_1$ e $\hat{\beta}_0$ dadas no acetato 78.

Este resultado está na base de intervalos de confiança e/ou testes de hipóteses para $\mu_{Y|x} = E[Y|X = x] = \beta_0 + \beta_1 x$.

Intervalos de confiança para $\mu_{Y|X} = E[Y|X=x]$

Intervalo de confiança para $\mu_{Y|X} = \beta_0 + \beta_1 x$

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

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

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

NOTA: A amplitude do IC aumenta com $QMRE$ e com a distância de x a \bar{x} e diminui com n e s_x^2 . Assim, a estimação de $\mu_{Y|X}$ é melhor para valores de x próximos de \bar{x} .

Inferência sobre $E[Y|X=x]$ no

Valores estimados e intervalos de confiança para $\mu_{Y|X}$ obtêm-se no R com a função `predict`. Os novos valores da variável preditiva são dados, através do argumento `new`, numa `data frame` onde a variável tem o mesmo nome que no ajustamento inicial.

De novo o exemplo das cerejeiras

O volume do tronco esperado de cerejeiras com DAP igual a 30 cm e a 50 cm, é dada por:

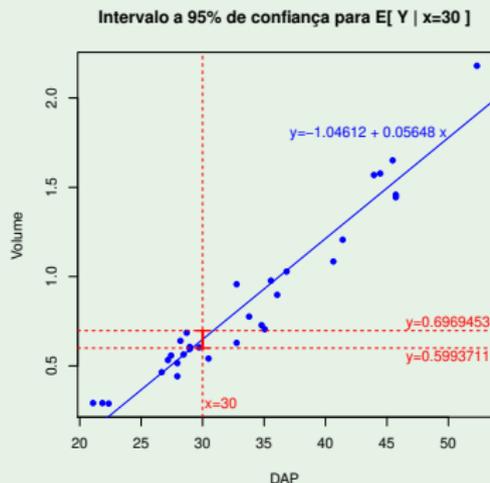
```
> predict(cerejeiras.lm, new=data.frame(DAP=c(30,50)))  
          1          2  
0.6481582 1.7776785
```

Um **intervalo de confiança** obtém-se com o argumento `int="conf"`.

IC para $E[Y|X = x]$ no

IC para $\mu_{Y|X}$ nos dados das cerejeiras

```
> predict(cerejeiras.lm, new=data.frame(DAP=c(30)), int="conf")  
      fit      lwr      upr  
1 0.6481582 0.5993711 0.6969453
```

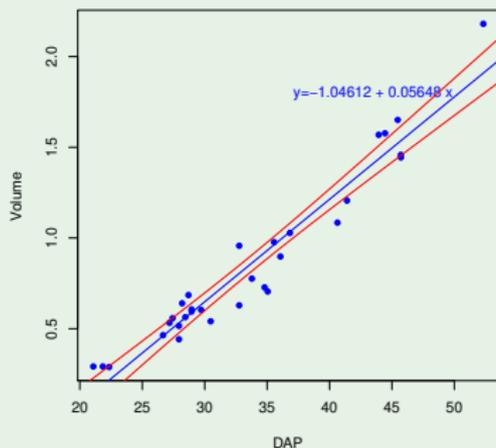


Bandas de confiança para a recta de regressão

Percorrendo valores de x , obtêm-se **bandas de confiança** para a recta de regressão populacional.

Bandas de confiança para a recta populacional das cerejeiras

A 95% de confiança, a recta populacional está contida nas bandas:



Os IC para $\mu_{Y|X}$ dependem do valor de x . Terão maior amplitude quanto mais afastado x estiver da média \bar{x} das observações. As bandas são **encurvadas**.

A variabilidade numa observação individual de Y

Calculámos ICs para o valor esperado de Y , dado um dado valor de x :

$\mu_{Y|x} = \beta_0 + \beta_1 x$. Mas uma observação individual de Y tem uma variabilidade adicional:

$$Y = \beta_0 + \beta_1 x + \varepsilon = \mu_{Y|x} + \varepsilon.$$

O estimador de $\mu_{Y|x}$ tem variância (acetato 105):

$$V[\hat{\mu}_{Y|x}] = \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)S_x^2} \right].$$

A variância do erro aleatório é $V[\varepsilon] = \sigma^2$.

A soma destas variâncias é a variância associada à previsão numa observação individual de Y associada a $X = x$:

$$\sigma_{Indiv}^2 = \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)S_x^2} \right] + \sigma^2 = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)S_x^2} \right].$$

Intervalos de predição para uma observação de Y

Para construir intervalos de predição para uma observação individual de Y , associada ao valor $X = x$, incrementa-se a variância em σ^2 , logo a variância estimada em $QMRE$.

Intervalo de predição para observação individual de Y

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

com $\hat{\mu}_{Y|x} = b_0 + b_1x$ e $\hat{\sigma}_{indiv} = \sqrt{QMRE \left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2} \right]}$.

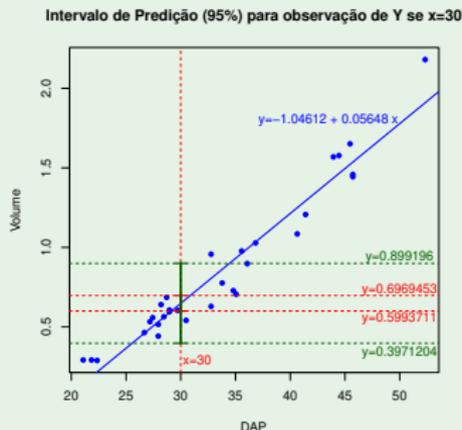
Estes intervalos são (para um mesmo nível $(1-\alpha) \times 100\%$) necessariamente de maior amplitude que os intervalos de confiança para o valor esperado (médio) de Y , $E[Y|X = x]$, vistos antes.

Intervalos de predição para Y no

No R, um **intervalo de predição** para uma observação individual de Y obtém-se através da opção `int="pred"` no comando `predict`.

Intervalos de predição para Y nas cerejeiras

```
> predict(cerejeiras.lm, new=data.frame(DAP=c(30)), int="pred")
      fit      lwr      upr
1 0.6481582 0.3971204 0.899196
```

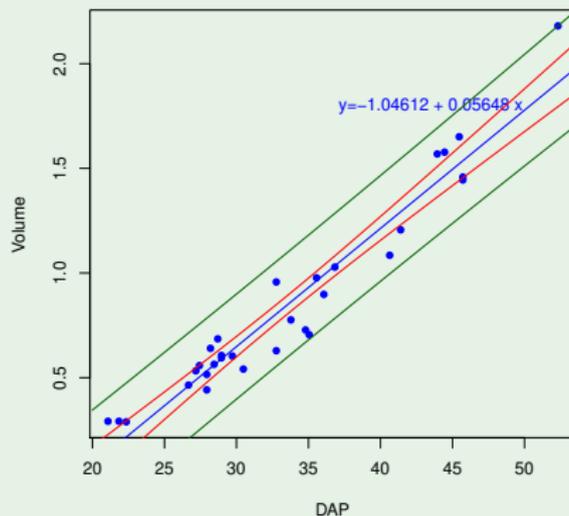


Bandas de predição para uma observação de Y

Variando os valores de x obtêm-se **bandas de predição** para valores individuais de Y .

Bandas de predição para Y nas cerejeiras

No exemplo, 95% dos valores de Y deverão estar contidos entre as seguintes bandas (encurvadas) verdes (a vermelho as bandas de confiança para $\mu_{Y|x}$):



Avaliando a qualidade do ajustamento do Modelo

- Em termos meramente descritivos, a qualidade de ajustamento da recta mede-se com o **Coefficiente de Determinação**, $R^2 = \frac{SQR}{SQT}$.
- Num contexto inferencial, é usual **também** testar a qualidade do ajustamento do Modelo.

A Hipótese Nula do teste de ajustamento global do modelo afirma que o modelo sob estudo **não se distingue do Modelo Nulo**, ou seja, o modelo **sem preditores**, de equação $Y = \beta_0 + \varepsilon$.

Numa **Regressão Linear Simples**, esta Hipótese Nula equivale a $\beta_1 = 0$:

$$Y = \beta_0 + \beta_1 x + \varepsilon \quad \implies \quad Y = \beta_0 + \varepsilon .$$

Em geral, pode escrever-se a **Hipótese Nula** como $H_0 : \mathcal{R}^2 = 0$, onde \mathcal{R}^2 é o **coeficiente de determinação populacional**.

Avaliando o ajustamento do Modelo (cont.)

Numa Regressão Linear Simples, pode testar-se essa hipótese de duas maneiras:

- Testar $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$, usando o teste t de hipóteses a β_1 , considerado no acetato 94.
- Efectuar o teste F ao ajustamento global do modelo.

Apenas esta segunda abordagem se estende ao caso da Regressão Linear Múltipla.

Uma distribuição associada a SQR

Ponto de partida natural para um teste à qualidade de ajustamento do Modelo será saber se SQR (o numerador de R^2) é grande. Ora,

- $SQR = \hat{\beta}_1^2 (n-1) s_x^2$ (ver Exercício RLS 5d).
- No acetato 74 viu-se que: $\frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{(n-1)s_x^2}}} \sim \mathcal{N}(0, 1)$.
- Logo, $\frac{(\hat{\beta}_1 - \beta_1)^2}{\sigma^2 / [(n-1)s_x^2]} \sim \chi_1^2$. [Recordar: $z \sim \mathcal{N}(0, 1) \Rightarrow z^2 \sim \chi_1^2$].
- Se $\beta_1 = 0$, tem-se: $\frac{\hat{\beta}_1^2 (n-1) s_x^2}{\sigma^2} = \frac{SQR}{\sigma^2} \sim \chi_1^2$.

Por outro lado, prova-se que, dado o Modelo RLS, **tem-se sempre:**

$$\frac{SQRE}{\sigma^2} \sim \chi_{n-2}^2.$$

SQR e SQRE

- Sabemos (da disciplina de Estatística) que as distribuições F surgem da seguinte forma:

$$\left. \begin{array}{l} W \sim \chi_{v_1}^2 \\ V \sim \chi_{v_2}^2 \\ W, V \text{ independentes} \end{array} \right\} \Rightarrow \frac{W/v_1}{V/v_2} \sim F_{v_1, v_2} .$$

- É possível mostrar que $SQRE$ e SQR são v.a. independentes.
- Logo, se $\beta_1 = 0$, tem-se $W = \frac{SQR}{\sigma^2} \sim \chi_1^2$ e $V = \frac{SQRE}{\sigma^2} \sim \chi_{n-2}^2$, e

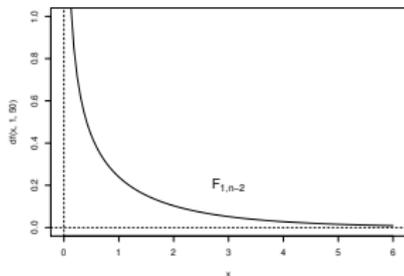
$$\frac{W/1}{V/(n-2)} = \frac{\frac{SQR}{\sigma^2 \cdot 1}}{\frac{SQRE}{\sigma^2 \cdot (n-2)}} = \frac{QMR}{QMRE} \sim F_{(1, n-2)} ,$$

sendo $QMR = \frac{SQR}{1}$ e $QMRE = \frac{SQRE}{n-2}$.

Como usar a estatística F

Vimos que, se $\beta_1 = 0$ tem-se:

$$\frac{QMR}{QMRE} \sim F_{(1, n-2)}$$



Quanto maior $\hat{\beta}_1^2$, mais duvidoso será $H_0 : \beta_1 = 0$. Ao mesmo tempo, maior será $SQR = \hat{\beta}_1^2 (n-1) s_x^2$, pelo que maior será a estatística $F = \frac{QMR}{QMRE}$.

Assim, valores elevados da estatística F sugerem $H_1 : \beta_1 \neq 0$, ou seja, a **Região Crítica** deverá ser **unilateral direita**.

O Teste F de ajustamento global do Modelo

Sendo válido o Modelo de Regressão Linear Simples, tem-se:

Teste F de ajustamento global do modelo

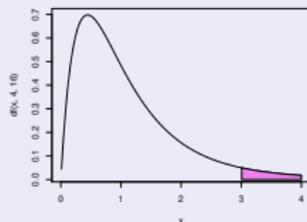
Hipóteses: $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$.

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

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se $F_{calc} > f_{\alpha(1,n-2)}$



O Teste F de ajustamento global do Modelo (cont)

Podem-se re-escrever as hipóteses e estatística do teste usando Coeficientes de Determinação (ver também Exercício RLS 12):

Teste F de ajustamento global do modelo

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

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

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se $F_{calc} > f_{\alpha(1,n-2)}$

A estatística F é uma função crescente do coeficiente de determinação amostral, R^2 .

O teste F no

A informação essencial para efectuar um teste F ao ajustamento global de um modelo de regressão também se obtém através do comando `summary`, aplicado a um objecto `lm`. Em particular:

F-statistic valor calculado da estatística $F = \frac{QMR}{QMRE}$, e os graus de liberdade na distribuição F que lhe está associada.

p-value valor de prova de F_{calc} no teste de ajustamento global do modelo.

Teste F de ajustamento global nos lírios

```
> summary(cerejeiras.lm)
```

```
[...]
```

```
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
```

Outra informação de summary

Na tabela final produzida quando um comando `summary` se aplica a um objecto resultante do comando `lm` são também dados os valores de:

Residual Standard error: Estimativa do desvio padrão σ dos erros aleatórios

ϵ_j :

$$\hat{\sigma} = \sqrt{QMRE} = \sqrt{\frac{SQRE}{n-2}}$$

Multiple R-squared: O Coeficiente de Determinação:

$$R^2 = \frac{SQR}{SQT} = \frac{s_{\hat{y}}^2}{s_y^2} = 1 - \frac{SQRE}{SQT}$$

Adjusted R-squared: O R^2 modificado (mais usado na RL Múltipla):

$$R_{mod}^2 = 1 - \frac{QMRE}{QMT} = 1 - \frac{\hat{\sigma}^2}{s_y^2}, \quad \left(QMT = \frac{SQT}{n-1} \right)$$

A Validação do Modelo (análise dos resíduos)

TODA a inferência feita até aqui admitiu a validade do Modelo Linear, e em particular, dos pressupostos relativos aos erros aleatórios: Normais, de média zero, variância homogénea e independentes.

Uma análise de regressão não fica completa sem que haja uma validação dos pressupostos do modelo.

A validação dos pressupostos relativos aos erros aleatórios (que são desconhecidos) faz-se através dos seus preditores, os resíduos.

O modelo RLS admite erros aleatórios ε_j independentes, com distribuição $\mathcal{N}(0, \sigma^2)$.

Mas os resíduos E_j não são independentes, como se pode verificar a partir do facto de que somam zero (ver Exercício RLS 5). E têm distribuições diferentes.

A distribuição dos Resíduos no Modelo RLS

Distribuição dos Resíduos no Modelo RLS

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

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

(ver também Exercício RLS 19).

É hábito validar os pressupostos do Modelo de Regressão através de gráficos de resíduos.

Não se efectuam testes de Normalidade aos resíduos usuais, uma vez que os resíduos não são independentes.

Diferentes tipos de resíduos

Dado o Modelo RLS, tem-se: $\frac{E_j}{\sqrt{\sigma^2(1-h_{jj})}} \sim \mathcal{N}(0, 1)$.

Duas variantes de resíduos

Resíduos habituais : $E_j = Y_j - \hat{Y}_j$;

Resíduos estandardizados : $R_j = \frac{E_j}{\sqrt{QMRE \cdot (1-h_{jj})}}$.

Resíduos **estandardizados** não têm unidades de medida.

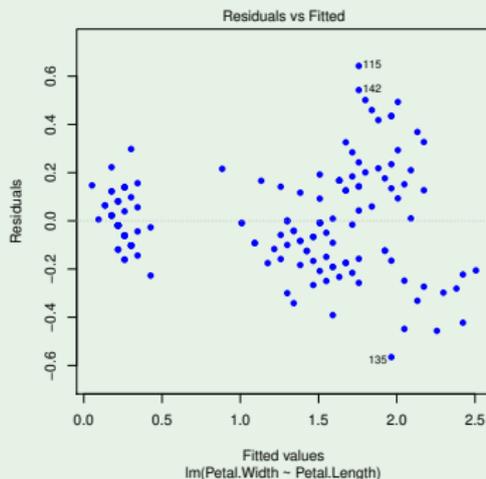
Embora a distribuição dos R_j não seja exactamente $\mathcal{N}(0, 1)$, costumam estar no intervalo $[-3, 3]$.

Gráficos de resíduos vs. \hat{Y}_i

Um gráfico indispensável: Resíduos E_i (usuais) vs. Valores ajustados \hat{Y}_i .

Exemplo dos lírios (Exercício RLS 11)

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



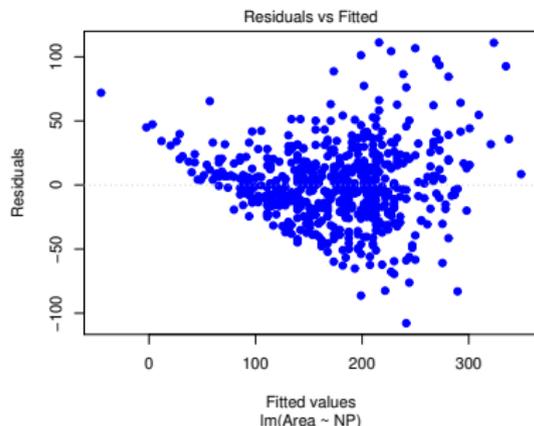
Os resíduos devem dispor-se aproximadamente numa banda horizontal em torno de zero. Dado o Modelo RLS, $cor(E_i, \hat{Y}_i) = 0$ (ver Exercício RLS 19).

Possíveis padrões indicativos de problemas

No gráfico E_i vs. \hat{Y}_i podem surgir padrões indicativos de problemas.

Curvatura na disposição dos resíduos: Indica violação da hipótese de linearidade entre x e y .

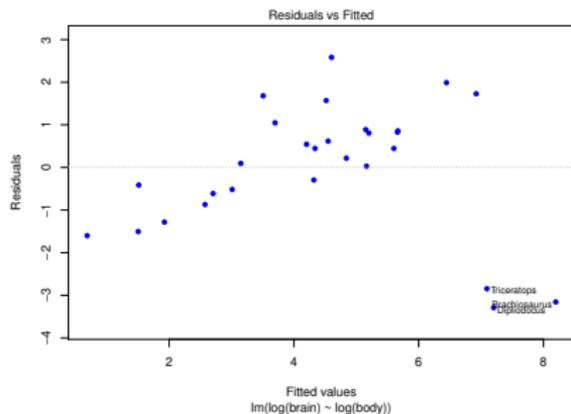
Gráfico em **forma de funil**: Indica violação da hipótese de homogeneidade de variâncias.



Um exemplo de resíduos em **forma de funil** e sugerindo **curvatura** na relação entre as duas variáveis (dados **videiras**, Ex.RLS 13).

Padrões indicativos de problemas (cont.)

Um ou mais **resíduos muito destacados** e/ou **banda oblíqua**: Indica possíveis observações atípicas.



A presença dos dinossáurios nos dados `Animals` (RLS Ex.7) cria uma banda oblíqua de pontos no gráfico E_j vs. \hat{Y}_j .

Gráficos para estudar a hipótese de normalidade

Como foi visto no acetato 125, dado o Modelo, $\frac{E_j}{\sqrt{\sigma^2(1-h_{jj})}} \sim \mathcal{N}(0, 1)$.

Embora os resíduos estandardizados, $R_j = \frac{E_j}{\sqrt{QMRE(1-h_{jj})}}$ não sejam exactamente $\mathcal{N}(0, 1)$, desvios importantes à Normalidade devem fazer duvidar da validade do pressuposto de erros aleatórios Normais.

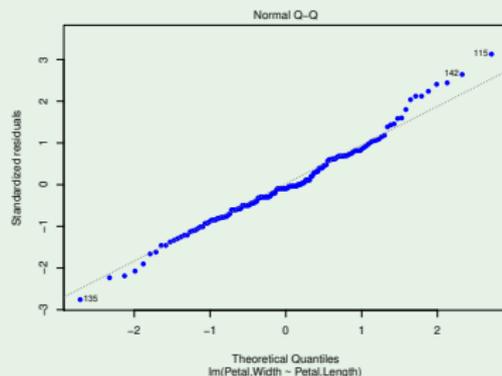
A validade do pressuposto de erros aleatórios Normais através dum **qq-plot** que confronte os **quantis empíricos** dos n resíduos standardizados, com os **quantis teóricos** numa $\mathcal{N}(0, 1)$.

O qq-plot indica concordância com a hipótese de Normalidade dos erros aleatórios se os pontos estiverem aproximadamente em linha recta.

Gráficos para o estudo da Normalidade (cont.)

Exemplo dos lírios (Exercício RLS 11)

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



Este *qq-plot* sugere algum desvio para os resíduos mais extremos, mas não em quantidade ou de forma suficientemente severa para pôr em dúvida o pressuposto da Normalidade dos erros aleatórios.

Gráficos para o estudo de independência

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

Nessas situações serão precisos **modelos específicos para dados com autocorrelação temporal ou espacial**.

Caso haja suspeita de correlação no tempo ou no espaço, será útil inspeccionar **gráficos de resíduos vs. ordem de observação** ou **posição no espaço**, para verificar se existem padrões que sugiram falta de independência.

Observações atípicas

Outras ferramentas de diagnóstico visam identificar observações individuais que merecem ulterior análise.

Observações atípicas (*outliers* em inglês). Conceito sem definição rigorosa, designa observações que se distanciam muito da relação linear de fundo entre Y e a variável preditora.

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

Mas observações distantes da tendência geral podem afectar o próprio ajustamento do modelo, e não serem facilmente identificáveis a partir dos seus resíduos: o maior resíduo estandardizado nos dados `Animals` (Ex. RLS 7) não ultrapassa (em módulo) 2.286.

O efeito alavanca

Efeito alavanca (*leverage*)

Na RLS, o **efeito alavanca** da i -ésima observação é dado por:

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

O valor h_{ii} aparece na variância do i -ésimo resíduo E_i (acetato 125):

$$V[E_i] = \sigma^2 (1 - h_{ii}) .$$

Para h_{ii} grande (próximo de 1), $V[E_i] \approx 0$, logo o resíduo tende a estar próximo do seu valor médio (zero) : a recta de regressão tende a passar próximo do ponto (x_i, y_i) (o ponto “atrai” a recta).

Numa RLS, quanto mais afastado estiver o valor x_i da média \bar{x} , maior será o efeito alavanca da observação.

Efeito alavanca (cont.)

Propriedades do efeito alavanca

Para qualquer observação, verifica-se:

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

O **valor médio** das observações alavanca numa regressão linear simples é a razão entre o no. de parâmetros e o no. de observações:

$$\bar{h} = \frac{2}{n} ,$$

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

Observações com efeito alavanca elevado **podem, ou não, estar dispostas com a mesma tendência de fundo que as restantes observações (isto é, podem, ou não, ser atípicas).**

Observações influentes

Uma observação diz-se **influyente** se, quando retirada da análise, houver variações assinaláveis nos parâmetros estimados b_0 e b_1 , logo, nos \hat{y}_i .

Distância de Cook

Medida frequente para a **influência da observação i** é a **distância de Cook**, que na RLS é:

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

sendo \hat{y}_j o j -ésimo valor ajustado pela recta das n observações e $\hat{y}_{j(-i)}$ o correspondente valor ajustado com a recta estimada sem a observação i .

Expressão equivalente (sendo R_i o resíduo estandardizado):

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

Quanto maior D_i , maior é a influência da i -ésima observação.

Sugere-se $D_i > 0.5$ como **critério de observação influente**.

Uma prevenção

Observações atípicas, influentes ou alavanca **não são o mesmo conceito**, embora possam estar relacionados.

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

R_i^2 grande e h_{ii} grande $\Rightarrow D_i$ grande (observação influente)

R_i^2 pequeno e h_{ii} pequeno $\Rightarrow D_i$ pequeno (observação não influente)

R_i^2 grande e h_{ii} pequeno (ou viceversa) – D_i pode, ou não, ser grande

(Se obs. i é, ou não, influente depende da grandeza relativa de R_i^2 e h_{ii})

Estes diagnósticos servem sobretudo para **identificar observações que merecem maior atenção e consideração**.

Gráficos diagnósticos no

A função `plot`, aplicada a um objecto `lm` produz, além dos gráficos de resíduos antes considerados (com as opções `which=1` e `which=2`), gráficos com os diagnósticos agora definidos.

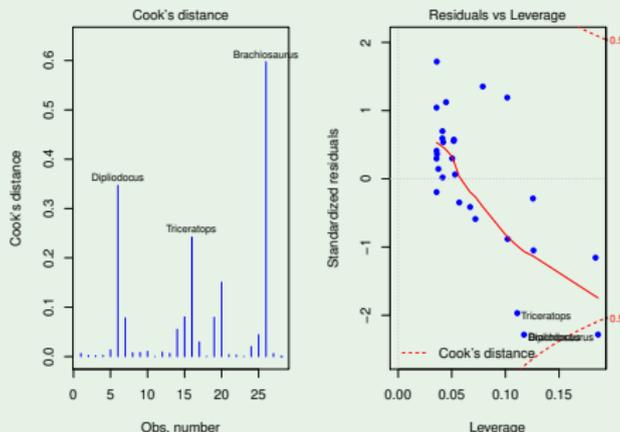
A opção `which=4` produz um diagrama de barras das distâncias de Cook associadas a cada observação.

A opção `which=5` produz um gráfico de resíduos estandardizados (R_i s) no eixo vertical contra valores de h_{ij} (*leverages*) no eixo horizontal, traçando linhas de igual distância de Cook (para os níveis 0.5 e 1, por omissão), que destacam eventuais observações influentes.

Exemplo de gráficos de diagnóstico

Exemplo: dados *Animals* (Ex. RLS 7)

Para os dados dos Exercícios RLS 7 (*Animals*):



Os valores elevados de distância de Cook refletem o distanciamento das espécies de dinossaúros da tendência geral das outras espécies, apesar de haver **três** observações discordantes.

Ainda as transformações de variáveis

Por vezes, violações às hipóteses de Normalidade dos erros aleatórios ou homogeneidade de variâncias podem ser ultrapassadas através de transformações de variáveis.

Mas a utilização de transformações de variáveis com este objectivo deve ser feita com cautela, porque:

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

Nota: Em contexto inferencial, as hipóteses de erros aleatórios aditivos, Normais, de variância homogénea, média zero e independentes terão de ser válidas para a relação linear entre as variáveis transformadas.