Capítulo 1

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 entre duas ou mais variáveis (modelos).

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.

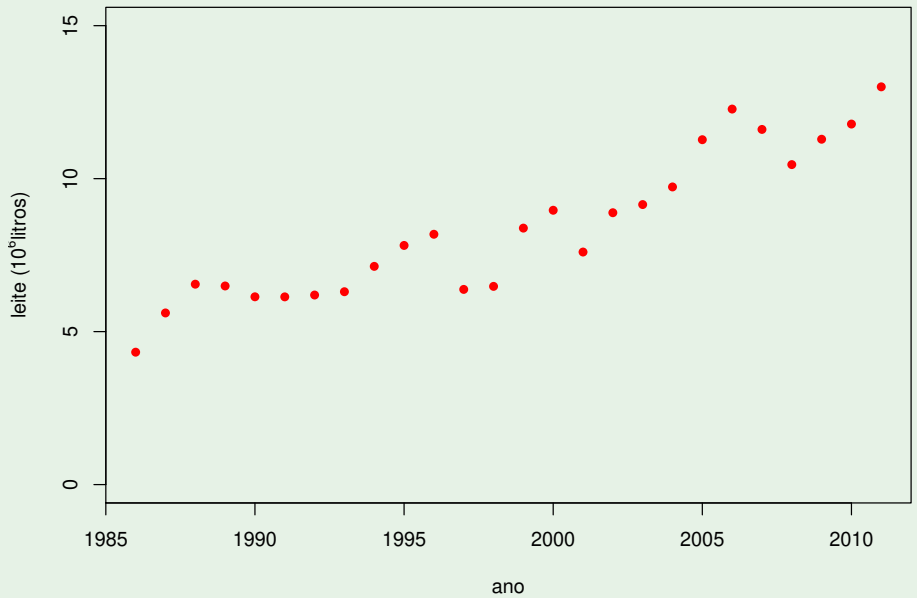
Vamos inicialmente considerar o contexto de um único preditor numérico, para modelar uma única variável resposta numérica.

Motivamos a discussão com 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, $\{(x_i, y_i)\}_{i=1}^{26}$.

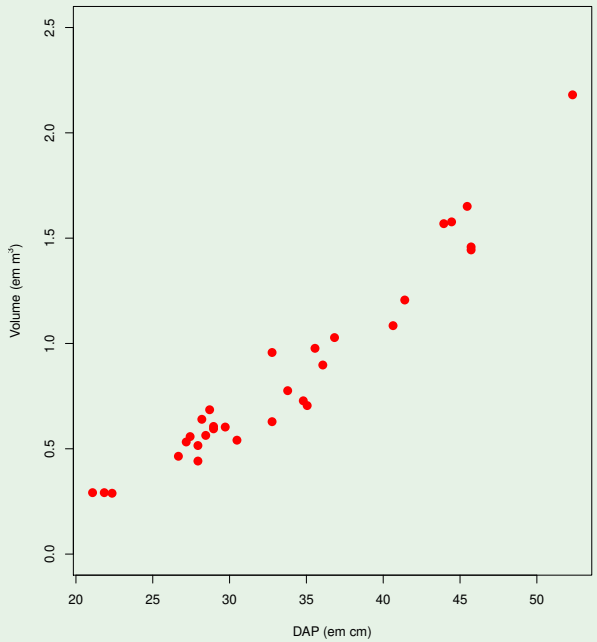


Existe uma tendência de fundo e é aproximadamente linear.
O coeficiente de correlação linear é $r_{xy} = 0.9348$.
Qual a "melhor" equação de recta, $y = b_0 + b_1 x$, para descrever as n observações (e que critério de "melhor")?

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 $y = \beta_0 + \beta_1 x$?

Estatística Descritiva

Duas classes de métodos estatísticos: descritivos e inferenciais.

Estatística Descritiva: Métodos para organizar, apresentar e extrair informação dum conjunto de dados.

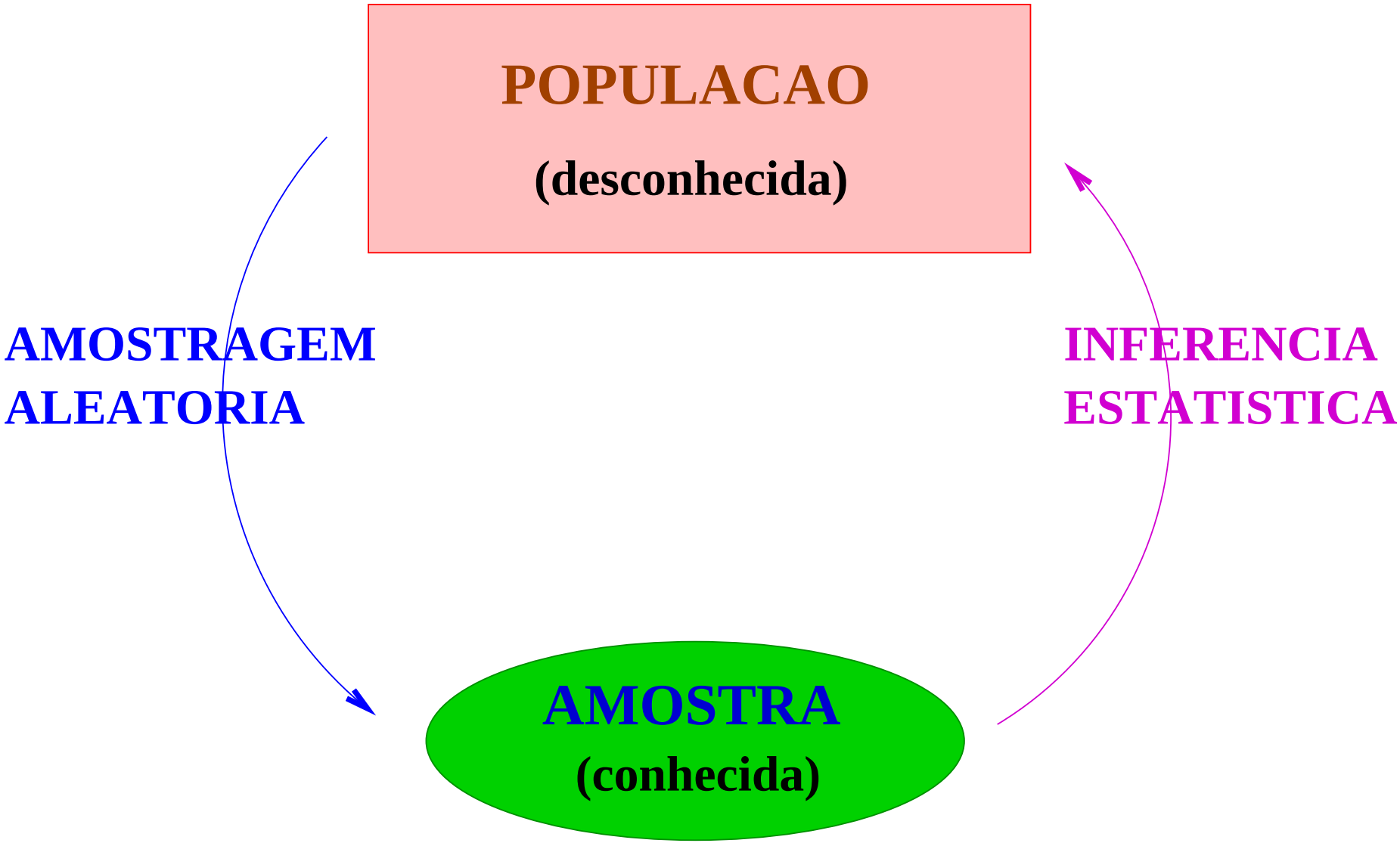
- Os dados podem ser de qualquer tipo: relativos a uma população inteira (censo) ou a uma amostra (aleatória ou não).
- As conclusões apenas dizem respeito às entidades observadas.
- Exemplos de ferramentas descritivas:
  - Para dados de uma só variável
    - ⋆ Cálculo de indicadores (média, variância, quantis, etc.).
    - ⋆ Tabelas de frequências.
    - ⋆ Histogramas, *boxplots* ou outras ferramentas gráficas.
  - Para dados relativos a duas variáveis
    - ⋆ Indicadores (Coeficientes de correlação, covariâncias, etc..)
    - ⋆ Nuvens de pontos (e, se for adequado, rectas de regressão)

Inferência Estatística

Inferência Estatística: procuram-se conclusões relativas a um conjunto vasto de elementos (a população), a partir da observação apenas dum subconjunto dessa população (a amostra).

- Para que se possa falar em inferência estatística, é necessário que a amostra tenha sido escolhida de forma aleatória.
- A inferência estatística baseia-se na Teoria de Probabilidades, que estuda os fenómenos aleatórios.
- Exemplos de ferramentas inferenciais:
  - Estimadores e estudo das suas propriedades.
  - Intervalos de confiança para parâmetros populacionais.
  - Testes de Hipóteses.

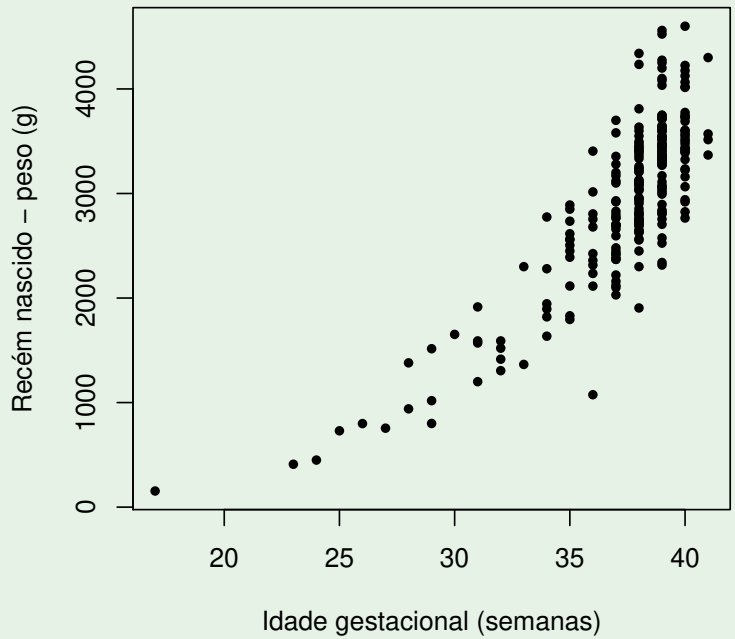
A Inferência Estatística (cont.)



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 ($y = c\,\mathrm{e}^{dx}$)?
  - f função potência ($y = c\,x^d$)?
  - 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.

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.

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

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

sendo

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

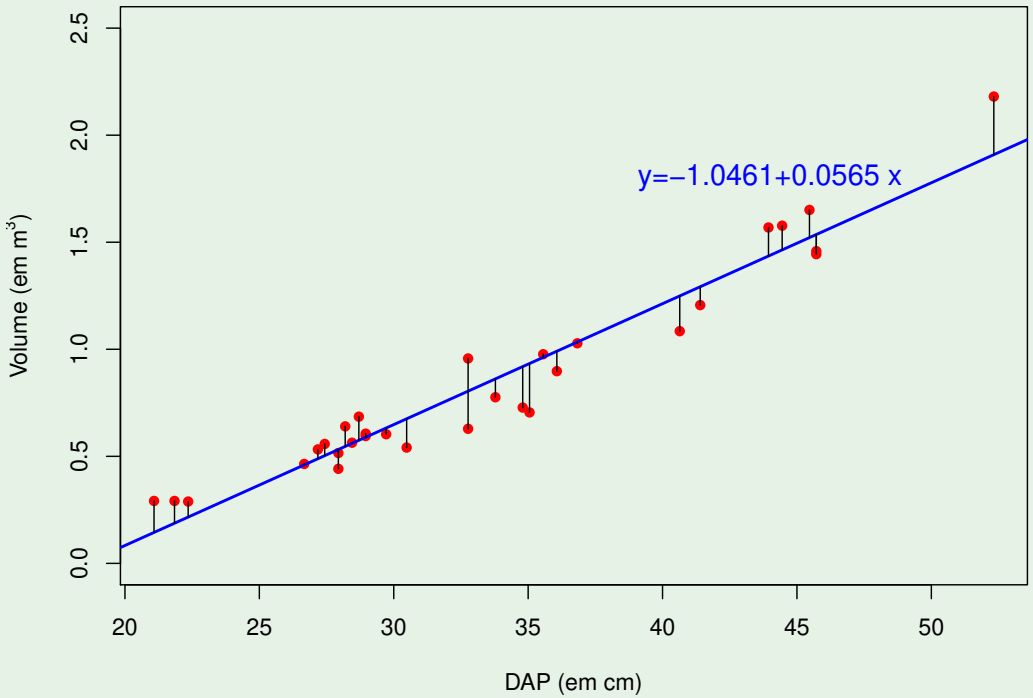
Regressão Linear Simples - contexto descritivo

Exemplo das cerejeiras

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

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

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



Regressão Linear Simples descritiva (cont.)

Como se chegou à equação da recta?

Critério

Minimizar a soma de quadrados residual (isto é, dos resíduos).
Os resíduos são diferenças na vertical entre pontos e recta ajustada:

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

sendo $\hat{y}_i \;=\; b_0 + b_1 x_i$ os "valores de y ajustados pela recta".

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

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)?
- outro critério qualquer?

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")
  - é a variável que se deseja modelar, prever a partir da variável x.
- x – **variável preditora** ("independente")
  - é a variável que se admite conhecida, e 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

O critério de 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.

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

$b_1 = 0.056$: por cada cm a mais no DAP, o volume do tronco aumenta, em média, $0.056 m^3$.

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

Definição: as três Somas de Quadrados

Considere $s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2$ - variância amostral dos y_i observados;

- Define-se SQ Total: $SQT = \sum_{i=1}^{n} (y_i - \overline{y})^2 = {}_{(n-1)}\, s_y^2$.

Considere $s_{\hat{y}}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$ - variância amostral dos $\hat{y}_i$ ajustados;

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

Considere $s_e^2 = \frac{1}{n-1} \sum_{i=1}^{n} (e_i - 0)^2$ - variância amostral dos resíduos e_i;

- Já se tinha definido SQ Residual: $SQRE = \sum_{i=1}^{n} e_i^2 = {}_{(n-1)}\, s_e^2$.

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

$$R^2 = \frac{SQR}{SQT} = \frac{s_{\hat{y}}^2}{s_y^2} \in [0,1] \quad , \quad (s_y^2 \neq 0)$$

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

- $0 \leq R^2 \leq 1$.
- $R^2 = 1$ se, e só se, os n pontos são colineares. ("ideal")
- $R^2 = 0$ se, e só se, a recta de regressão fôr horizontal. ("inútil")
- Numa regressão linear simples, R^2 é o quadrado do coeficiente de correlação linear entre x e y (ver também o Exercício RLS 6):

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

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.

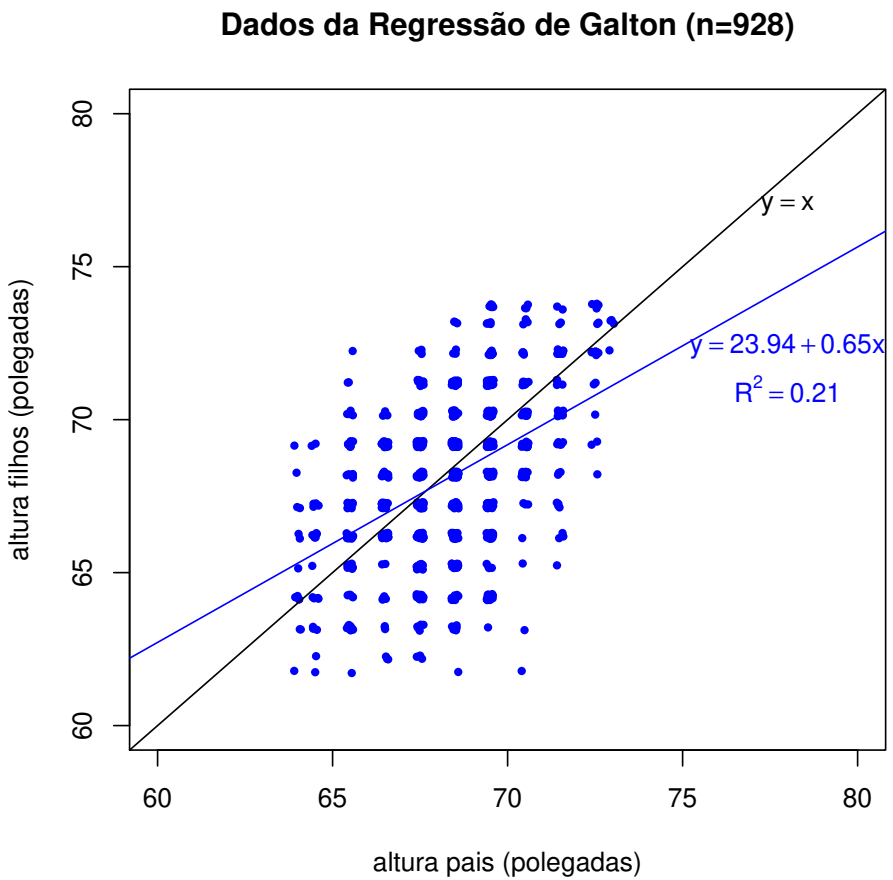
A designação Regressão tem origem num estudo de Francis Galton (1886), relacionando a altura de $n = 928$ jovens adultos com a altura (média) dos pais.

Galton 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.



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.

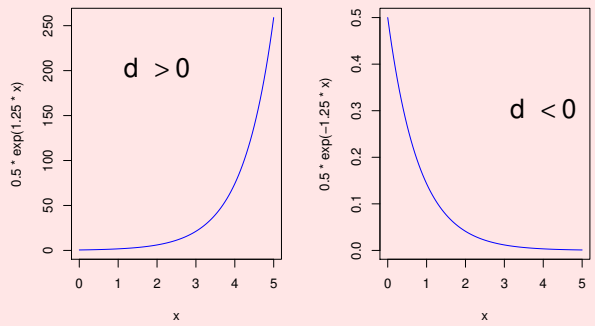
Tais transformações podem permitir utilizar a Regressão Linear Simples, apesar de a relação original ser não-linear.

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

Relação exponencial

Relação exponencial

$$y = c\,e^{d\,x}$$
$$(y>0 \quad ; \quad c>0)$$



Transformação: Logaritmizando, obtém-se:

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

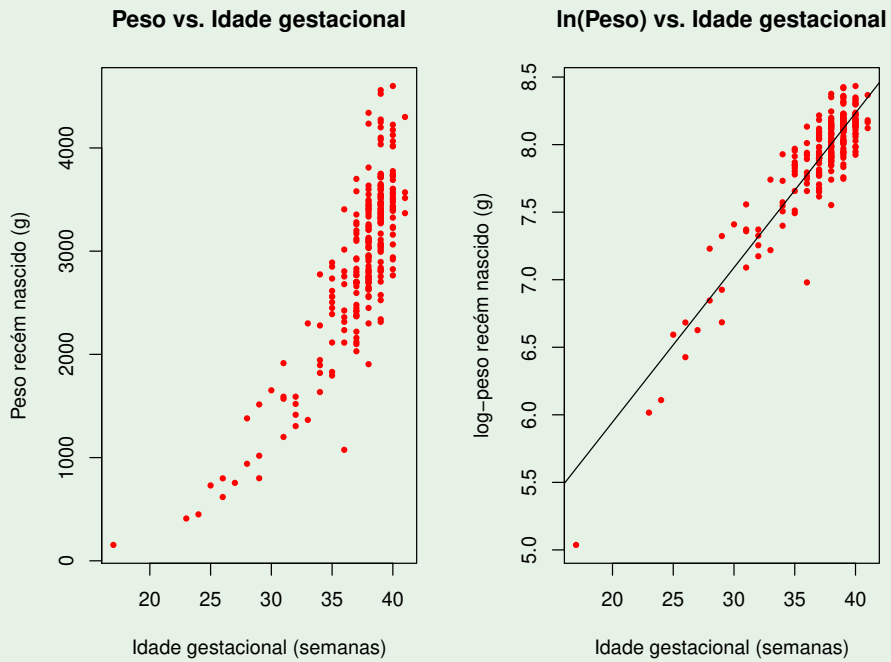
que é uma relação linear entre $y^* = \ln(Y)$ e x, com declive $b_1 = d$ e ordenada na origem $b_0 = \ln(c)$.

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

Um Exemplo

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

Ainda a relação exponencial

A Equação Diferencial duma 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 $\left(P\,\frac{f'}{f} = \ln|f|\right)$, tem-se:

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

O declive b_1 é o valor (constante) d da taxa de variação relativa de y.
A constante de primitivação K é a ordenada na origem da recta: $K = b_0$.

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.

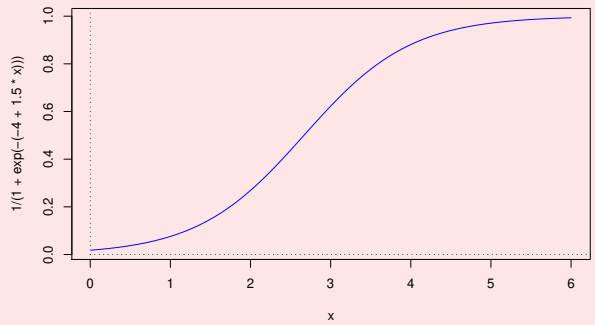
Em 1838 Verhulst propôs uma modelo de crescimento populacional alternativo, prevendo os efeitos resultantes da escassez de recursos: o modelo logístico.

Considera-se aqui uma versão simplificada (com 2 parâmetros) desse modelo, em que a variável y mede a dimensão duma população, relativa a um máximo possível, sendo assim uma proporção.

Relação Logística (com 2 parâmetros)

Relação Logística (2 parâmetros)

$$y = \frac{1}{1+e^{-(c+dx)}}$$



$(d > 0)$

Transformação: Como $y \in]0,1[$, tem-se uma relação linear entre a transformação *logit* de Y, i.e., $y^* = \ln\left(\frac{y}{1-y}\right)$, e x:

$$\begin{aligned}
\Leftrightarrow \quad 1-y &= 1 - \frac{1}{1+e^{-(c+dx)}} = \frac{e^{-(c+dx)}}{1+e^{-(c+dx)}} \\
\Leftrightarrow \quad \frac{y}{1-y} &= \frac{1}{e^{-(c+dx)}} = e^{c+dx} \\
\Leftrightarrow \quad \underbrace{\ln\left(\frac{y}{1-y}\right)}_{=y^*} &= \underbrace{c}_{=b_0} + \underbrace{d}_{=b_1} x
\end{aligned}$$

Ainda a Logística

Equação Diferencial da Logística (2 parâmetros)

A relação logística resulta de admitir que y é função de x e que a taxa de variação relativa de y diminui linearmente com o aumento de y, segundo a expressão:

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

A equação anterior equivale a:

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

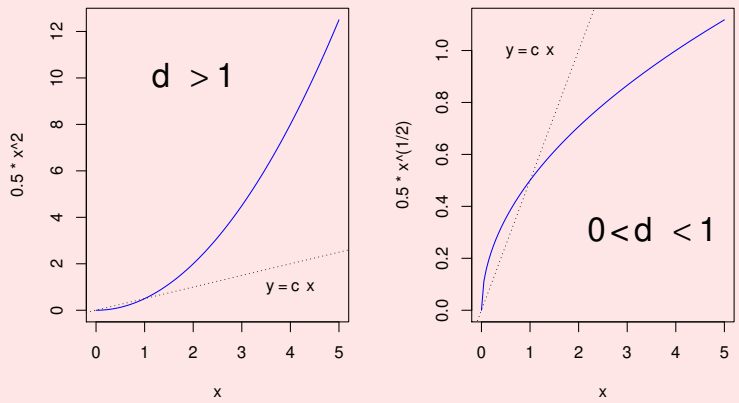
Primitivando (em ordem a x), tem-se:

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

Relação potência ou alométrica

Relação potência

$$y = c\,x^d$$
$$(x,y>0 \quad ; \quad c,d>0)$$



Transformação: Logaritmizando, obtém-se:

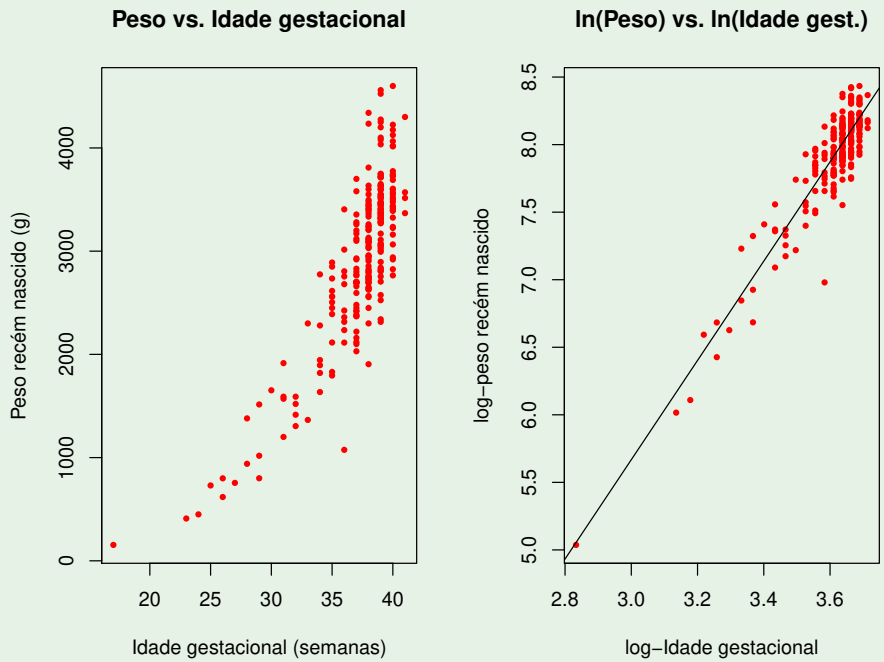
$$\begin{aligned} \ln(y) &= \ln(c) + \ln(x^d) &= \ln(c) + d\ln(x) \\ \Leftrightarrow \quad y^* &= b_0 + b_1 x^* \end{aligned}$$

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. Mas $b_0 = \ln(c)$.

Um exemplo

Outra linearização dos pesos dos bebés

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

Equação diferencial duma relação potência

Uma relação potência resulta de admitir que y e x são funções duma terceira variável t e que a taxa de variação relativa de y é proporcional à taxa de variação relativa de x:

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

Primitivando (em ordem a t) tem-se:

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

e exponenciando,

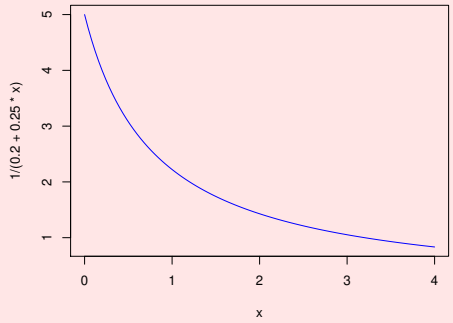
$$y \ = \ \mathrm{e}^{\ln x^d + K} \ = \ \mathrm{e}^{\ln x^d} \cdot \mathrm{e}^K \ = \ x^d \cdot \underbrace{\mathrm{e}^K}_{=c}$$

A relação potência é usada em alometria, que compara o crescimento de partes diferentes dum organismo. A isometria corresponde ao valor $d = 1$.

Relação hiperbólica (ou de proporcionalidade inversa)

Relação de tipo hiperbólico

$$y = \frac{1}{c + d\,x}.$$
$$(x, y > 0 \quad ; \quad c, d > 0)$$



Em Agronomia, tem sido usada para modelar rendimento por planta (y) vs. densidade da cultura ou povoamento (x).

Transformação: Obtém-se uma relação linear entre $y^* = 1/y$ e x:

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

Ainda a relação de tipo hiperbólico

Equação diferencial da relação de tipo hiperbólico

Resulta de admitir que a taxa de variação (diminuição) de y é proporcional ao quadrado de y ou, equivalentemente, que a taxa de variação relativa de y é proporcional a y:

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

Re-escrevendo como $\frac{y'(x)}{y^2(x)} = -d$, e primitivando $\left(P\, f^{\alpha}.f' = \frac{f^{\alpha+1}}{\alpha+1}\right)$:

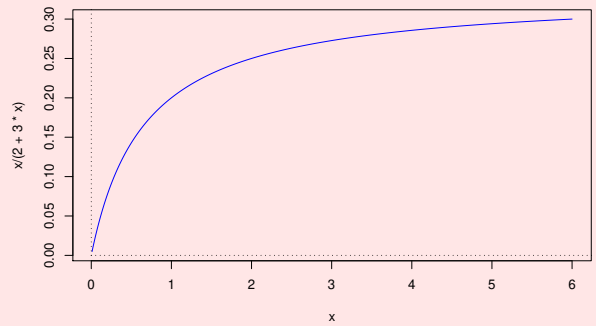
$$-\frac{1}{y(x)} = -d\,x + K \quad \Leftrightarrow \quad y(x) = \frac{1}{d\,x + c}\ ,$$

com $c = -K$.

Relação Michaelis-Menten

Relação Michaelis-Menten

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



Transformação: Tomando recíprocos, obtém-se uma relação linear entre $y^* = \frac{1}{y}$ e $x^* = \frac{1}{x}$:

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

com $b_0 = d$ e $b_1 = c$.

Relação Michaelis-Menten (cont.)

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

Equação Diferencial duma Michaelis-Menten

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

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

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.

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. Esta última abordagem corresponde a efectuar uma regressão não linear, metodologia não englobada nesta disciplina.

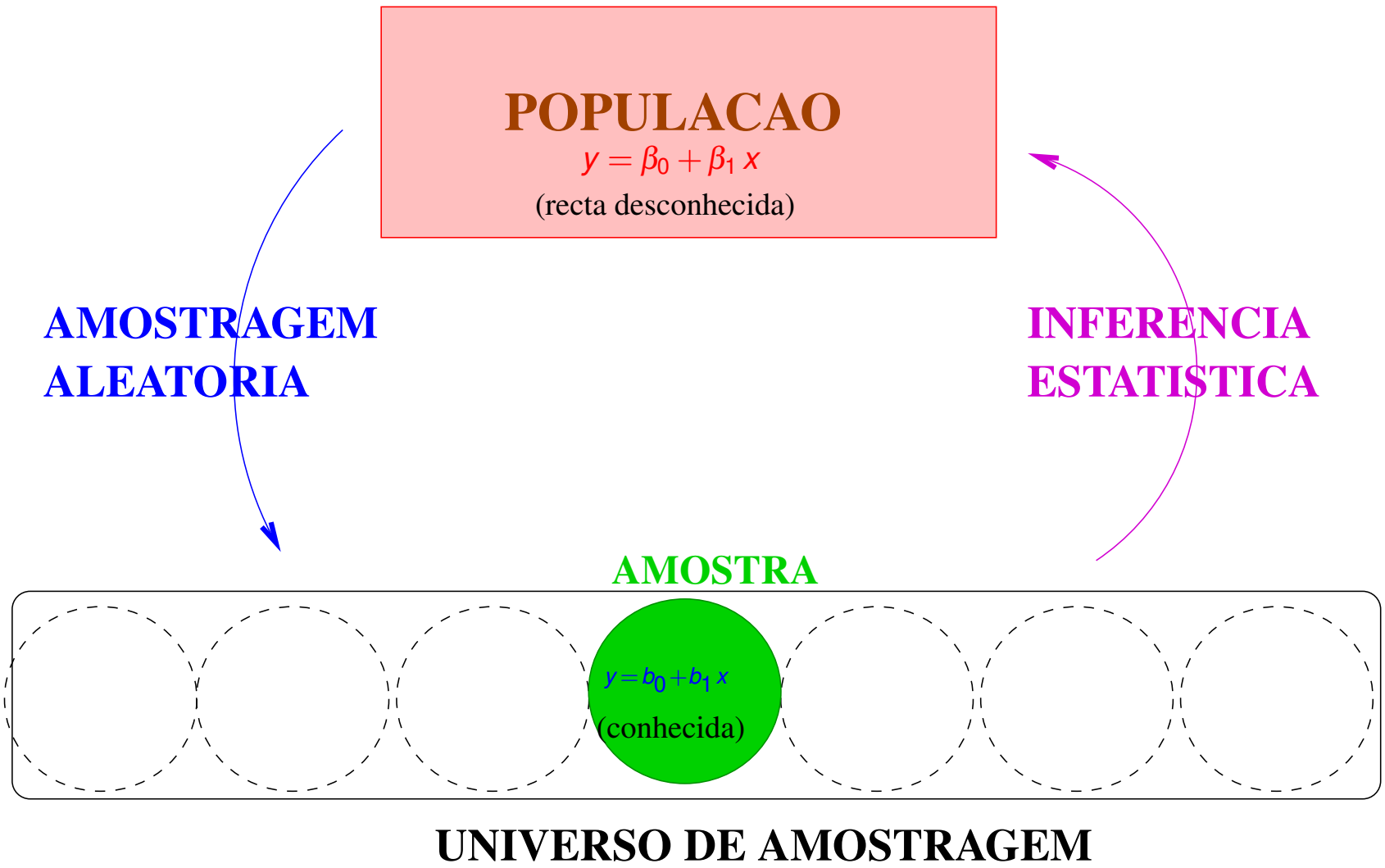
Regressão Linear Simples - INFERÊNCIA

- 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 $i = 1, \ldots, n$.

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_i:

- Têm valor esperado (valor médio) nulo:

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

  (não é hipótese restritiva).

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

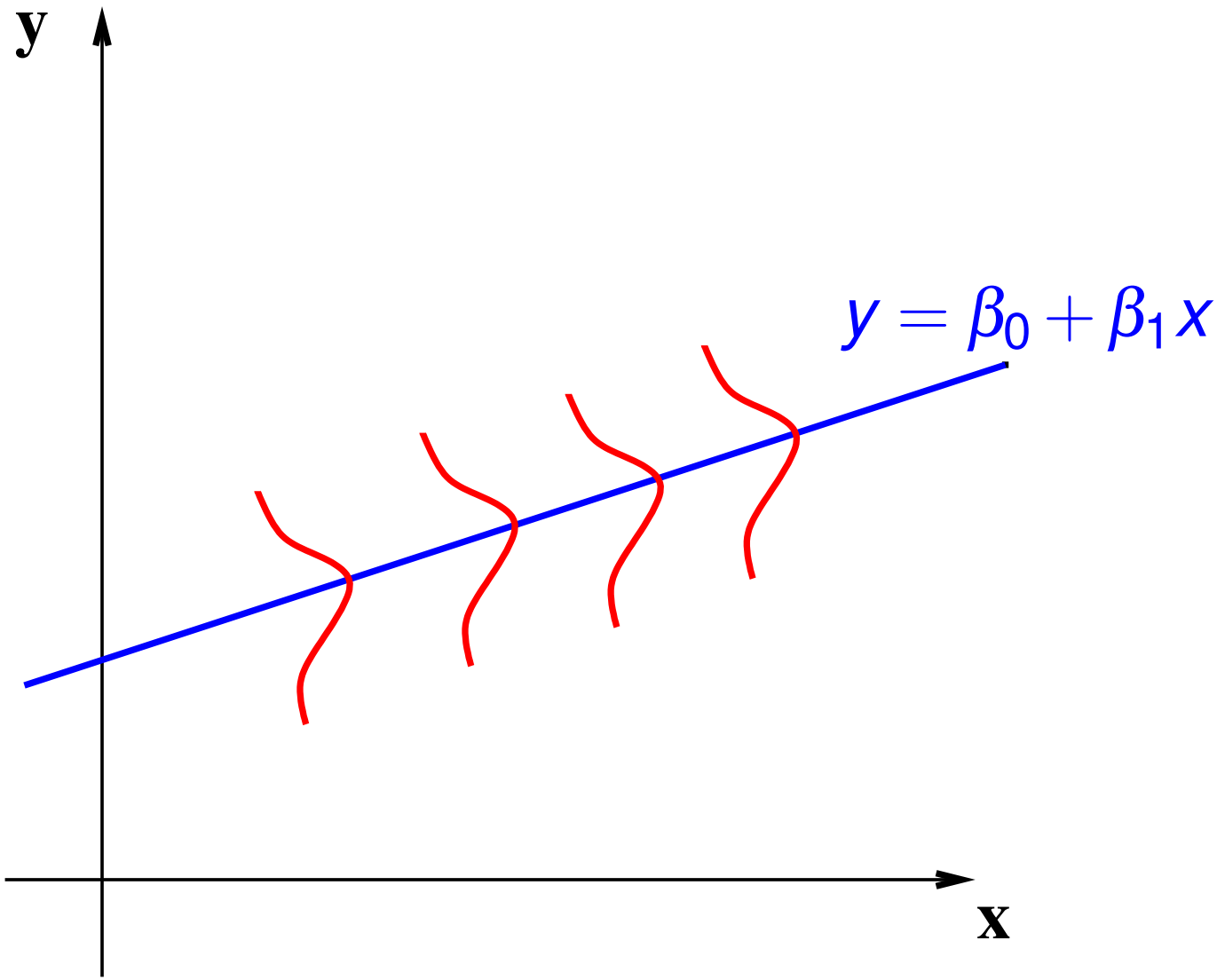
$$V[\varepsilon_i] = \sigma^2\,, \qquad \forall\, i = 1, ..., n$$

  (é restritiva, mas conveniente).

Ou seja, admite-se que $\varepsilon_i \frown \mathcal{N}(0, \sigma^2)$, para todo o i.

- 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. $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ , \quad \forall\, i = 1, ..., n.$
2. $\varepsilon_i \frown \mathcal{N}(0, \sigma^2) \ , \quad \forall\, i = 1, ..., n.$
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).

Cada tipo de v.a. tem as suas ferramentas próprias de caracterização.

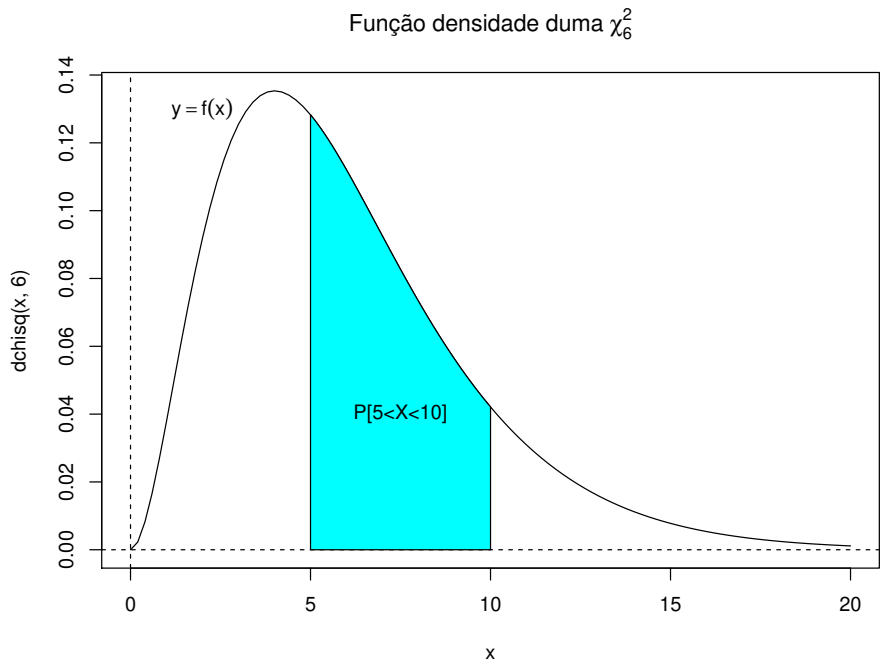
V.a.s discretas são caracterizadas pela sua função de probabilidades, ou função de massa probabilística, (ver Exercícios Introdutórios 4 e 5):

a cada possível valor x_i é associada a respectiva probabilidade p_i

Variáveis aleatórias contínuas

Uma v.a. contínua X é caraterizada pela sua função densidade $f(x)$ (não negativa), através da qual é possível calcular probabilidades de X tomar valores num dado intervalo:

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



Revisão: valores esperados

O valor esperado ou valor médio duma variável aleatória X é o centro de gravidade da sua distribuição de probabilidades:

- Se X discreta: $E[X] = \sum_i x_i \, p_i$;
- Se X contínua: $E[X] = \int_{-\infty}^{+\infty} x \, f(x) \, dx$.

Algumas propriedades dos valores esperados de variáveis aleatórias:

Propriedades dos valores esperados

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

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

Revisão: variâncias

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

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

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:

$$Cov[X,Y] = E[(X-E[X])(Y-E[Y])] = E[XY]-E[X]E[Y]$$

Propriedades da covariância de variáveis aleatórias

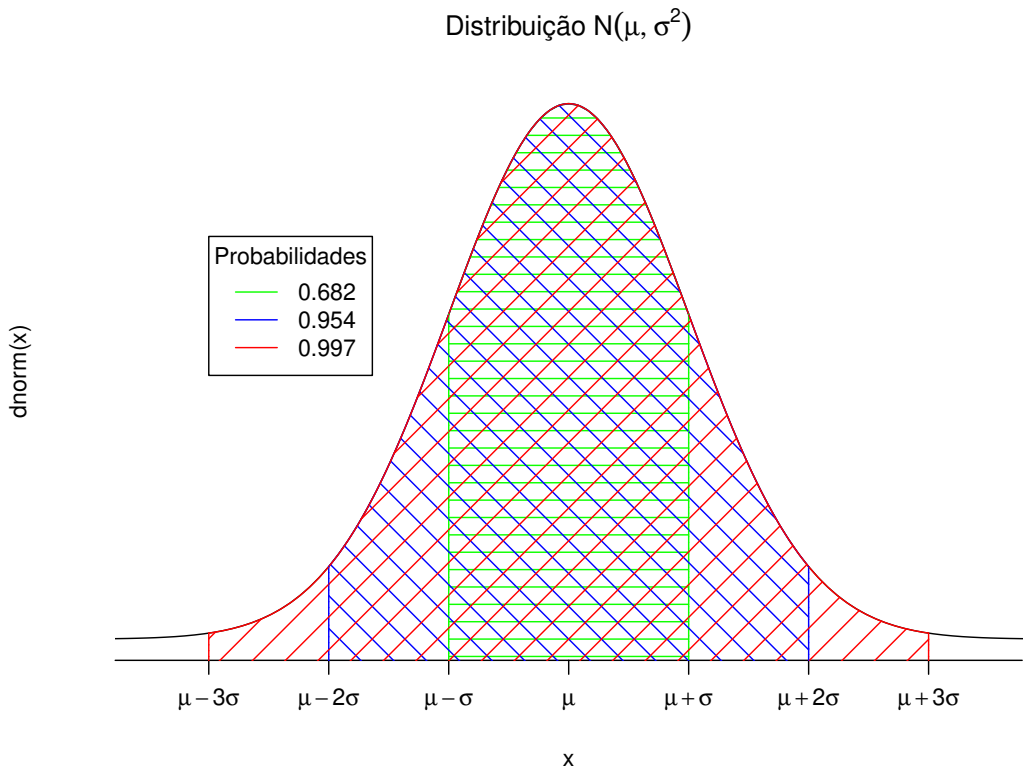
Sejam X, Y e Z variáveis aleatórias e a e b constantes. Então:

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

Revisão: a distribuição Normal

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

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



Propriedades da Normal

Propriedades da distribuição Normal

- Uma transformação linear duma Normal tem distribuição Normal. Mais concretamente, seja $X \frown \mathcal{N}(\mu, \sigma^2)$ e a, b constantes. Então:

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

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

- 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 70 e 71).

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

Recordar: $b_1 = \dfrac{cov_{xy}}{s_x^2} = \dfrac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{(n-1)\,s_x^2} \overset{(\star)}{=} \dfrac{\sum_{i=1}^{n}(x_i-\overline{x})\,y_i}{(n-1)\,s_x^2}$

$(\star)$ Veja-se o Exercício RLS 3b).

Estimador de β_1

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

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

Estimação dos parâmetros do Modelo RLS (cont.)

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

Estimador de β_0

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

com

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

Quer $\hat{\beta}_1$, quer $\hat{\beta}_0$, são combinações lineares das observações $\{Y_i\}_{i=1}^{n}$, logo 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,

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

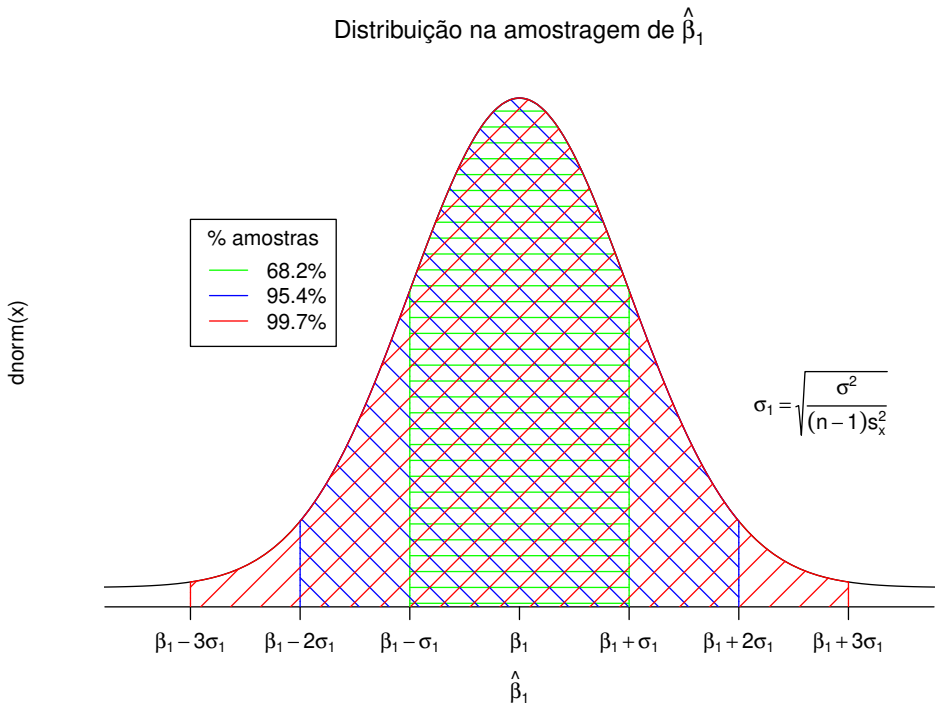
NOTAS:

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 $\overline{x}$ 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 a seguinte:



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,

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

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 construi-lo a partir dos resíduos.

Erros aleatórios e Resíduos

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

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, $SQRE$.

A Soma de Quadrados Residual

Resultados distribucionais de SQRE

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

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

NOTA: Omite-se a demonstração

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

Recordar: Nas distribuições χ^2, $X \frown \chi^2_\nu \Rightarrow E[X] = \nu$. Logo,

$$E\left[\frac{SQRE}{\sigma^2}\right] = n-2 \Leftrightarrow \frac{1}{\sigma^2} E[SQRE] = n-2$$

$$\Leftrightarrow \frac{1}{n-2} E[SQRE] = \sigma^2 \Leftrightarrow E\left[\underbrace{\frac{SQRE}{n-2}}_{=QMRE}\right] = \sigma^2$$

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 de σ^2

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

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

Viu-se no acetato anterior que QMRE é um estimador centrado de σ^2.

Revisão: como surge uma $t-Student$

Veremos agora que a substituição de σ^2 pelo seu estimador *QMRE* no Corolário do acetato 80 transforma a distribuição Normal numa *t-Student*.

Na disciplina de Estatística viu-se como surge uma distribuição $t-Student$:

$$\left.\begin{array}{l} Z \frown \mathcal{N}(0,1) \\ W \frown \chi^2_\nu \\ Z, W \text{ v.a. independentes} \end{array}\right\} \Longrightarrow \frac{Z}{\sqrt{W/\nu}} \frown t_\nu \ .$$

No nosso contexto, tomamos $Z = \frac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}}$, $W = \frac{SQRE}{\sigma^2}$ e $\nu = n-2$.

Quantidades centrais para a inferência sobre β_0 e β_1

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

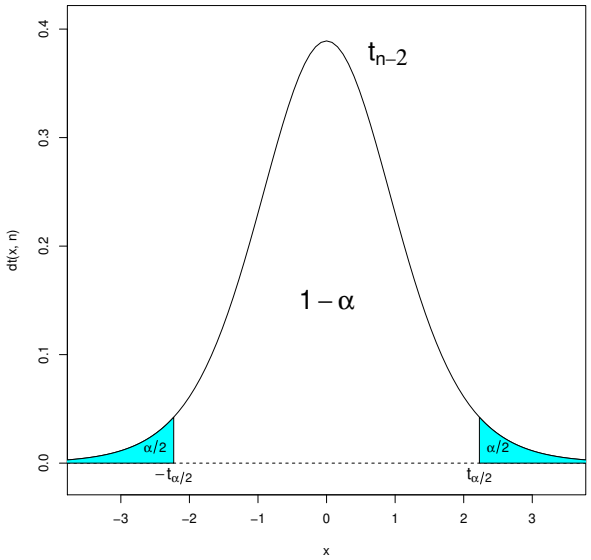
Dado o Modelo de Regressão Linear Simples, tem-se

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

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 $\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \frown t_{n-2}$. 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:

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

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

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

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

Substituindo nos extremos as quantidades amostrais, obtemos um intervalo concreto, chamado intervalo a $(1-\alpha)\times 100\%$ de confiança.

Interpretação IC para β_1

Como interpretar a conclusão que o intervalo aleatório

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

contém β_1 com probabilidade $1-\alpha$?

- A cada amostra concreta corresponde um intervalo concreto,

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

- $(1-\alpha) \times 100\%$ desses intervalos concretos, para todas as possíveis amostras (de dimensão n e valores de x_i dados) contêm o verdadeiro valor de β_1; Os restantes $\alpha \times 100\%$ não contêm 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:

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

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:

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

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 $\overline{x}^2$ e diminui com n e s_x^2:

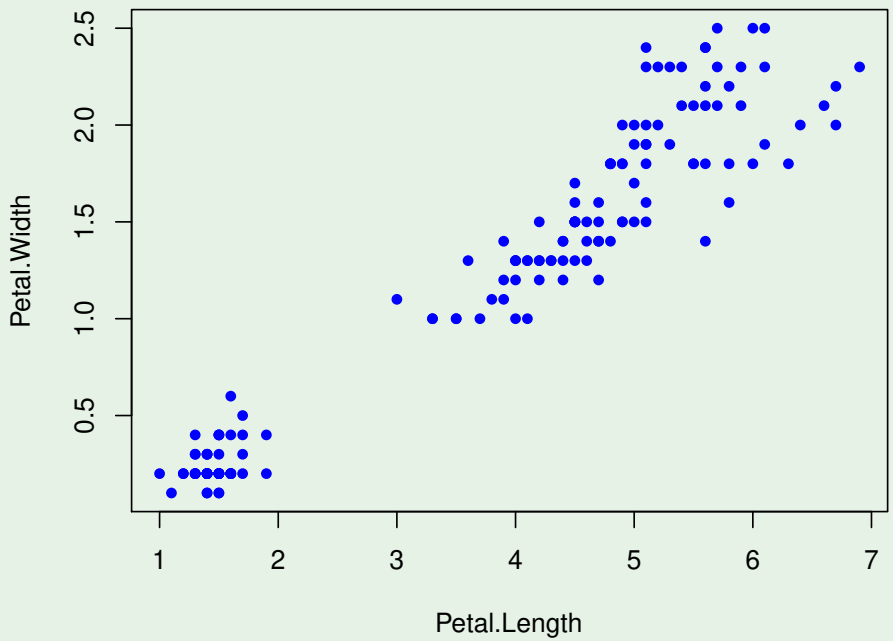
$$\hat{\sigma}_{\hat{\beta}_0} = \sqrt{QMRE\cdot\left[\frac{1}{n}+\frac{\overline{x}^2}{(n-1)\,s_x^2}\right]}$$

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

Um exemplo de RLS

Os lírios de Fisher

A *data frame* `iris`, no `R`, contém medições de 4 variáveis numéricas: comprimento e largura de sépalas e pétalas em $n = 150$ lírios (ver Exercício RLS 8). Eis a nuvem de pontos de largura e comprimento das pétalas (ambas em cm):



Um exemplo de RLS (cont.)

No R, as regressões lineares são ajustadas usando o comando `lm`.

Os lírios de Fisher (cont.)

A regressão de largura sobre comprimento das pétalas é ajustada, e guardada num objecto de nome `iris.lm`, da seguinte forma:

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

A recta estimada é assim:

$$y = -0.3631 + 0.4158\,x$$

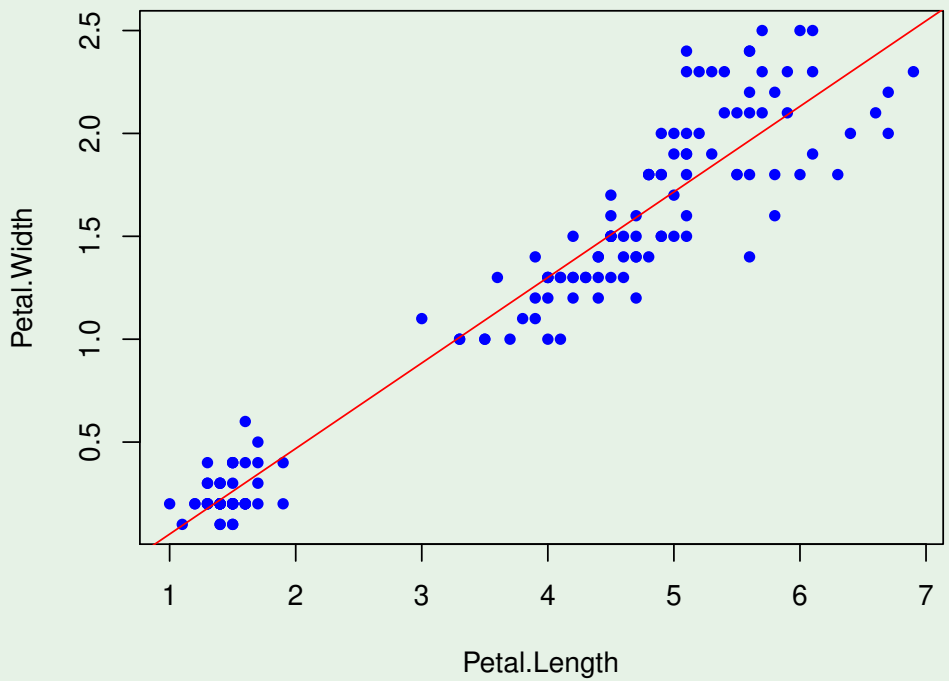
onde y indica a largura da pétala e x o seu comprimento.

Um exemplo de RLS (cont.)

Os lírios de Fisher (cont.)

No R, a recta pode ser sobreposta à nuvem de pontos, após os comandos nos acetatos anteriores, através do comando `abline`:

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



Um exemplo de RLS (cont.)

Os lírios de Fisher (cont.)

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

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

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

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

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

Os lírios de Fisher (cont.)

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

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

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

A 95% de confiança, o declive β_1 da recta populacional está no intervalo $]0.397\,,\,0.435[$ e a ordenada na origem β_0 no intervalo $]-0.442\,,\,-0.285[$.

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

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

Um alerta sobre Intervalos de Confiança

Tal como na construção de intervalos de confiança anteriores (disciplina de Estatística), existem duas facetas contrastantes:

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

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

Nota: 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.

Breve revisão sobre Testes de Hipóteses

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

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

As hipóteses dizem respeito à população. Opta-se entre hipóteses alternativas com base numa amostra aleatória dessa população.

Vamos agora ver Testes de Hipóteses sobre os parâmetros β_0 e β_1 duma recta de regressão populacional.

Revisão de Testes de Hipóteses: Passo 1

Num teste de Hipóteses há cinco passos a seguir.

No primeiro passo, formulam-se hipóteses alternativas em confronto.

Passo 1: hipóteses

Definir 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:

- uma 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.

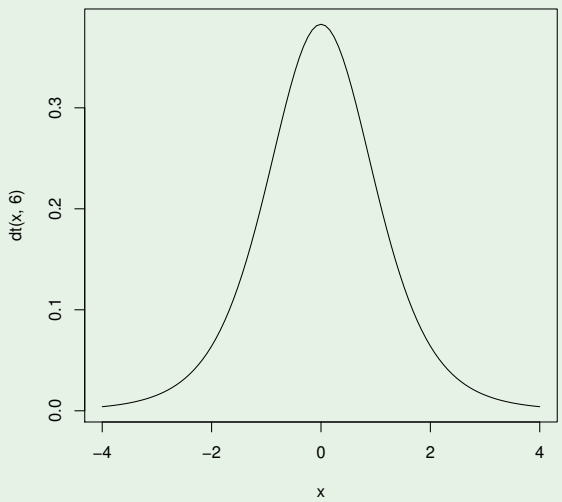
Exemplo: estatística de teste para Hipóteses sobre β_1

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

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

Testes de Hipóteses (cont.)

Exemplo: como ler os valores calculados da estatística de teste



Numa amostra em que o estimador $\hat{\beta}_1$ tome um valor $b_1 \approx \beta_{1|H_0} = 1$, a estatística T toma valor $T_{calc} \approx 0$. Neste caso, não há razões para duvidar de H_0.

Pelo contrário, se b_1 fôr muito diferente de $\beta_{1|H_0} = 1$, o valor calculado da estatística T será (em módulo) grande.

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

Quanto maior seja $|T_{calc}|$, menos plausível será $H_0 : \beta_1 = 1$.

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[Erro Tipo I] significa reduzir a gama de valores que levam à rejeição de H_0, aumentando P[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:

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

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

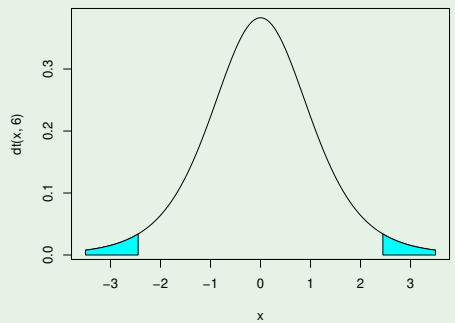
Passo 4: Região Crítica (ou de Rejeição)

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 fôr 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

- Escolhe-se uma amostra concreta;
- Calcula-se o valor da estatística para essa amostra;
- Toma-se a decisão de Rejeitar H_0 ou de 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.

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

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

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: $H_0 : \beta_1 = c \quad \text{vs.} \quad H_1 : \beta_1 \neq c.$

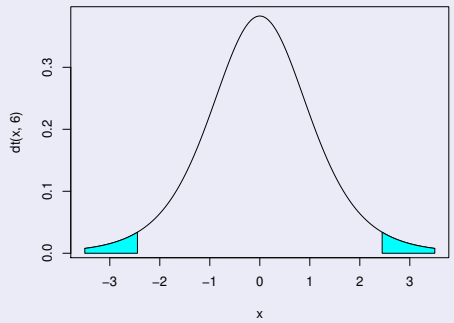
Estatística do Teste: $T = \frac{\hat{\beta}_1 - \overbrace{\beta_1|_{H_0}}^{=c}}{\hat{\sigma}_{\hat{\beta}_1}} \frown t_{n-2} \quad , \quad \text{sob } H_0.$

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

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)

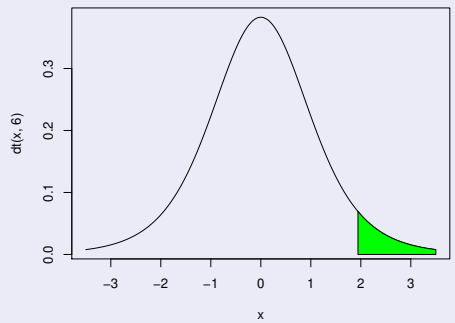
Hipóteses: $H_0 : \beta_1 \leq c$ vs. $H_1 : \beta_1 > c$.

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

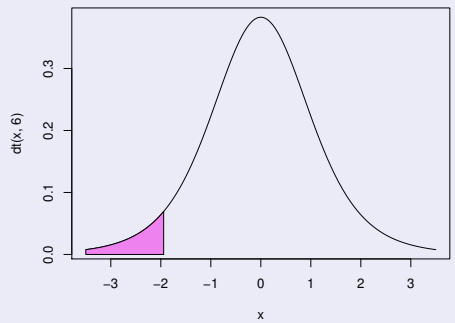
Hipóteses: $H_0 : \beta_1 \geq c$ vs. $H_1 : \beta_1 < c$.

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

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

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

Nível de significância do teste: α

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

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

Testes usando $p-values$

Em alternativa a fixar previamente o nível de significância α, é possível indicar apenas o valor de prova (ou *p*-value) associado ao valor calculado da estatística dum qualquer teste.

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

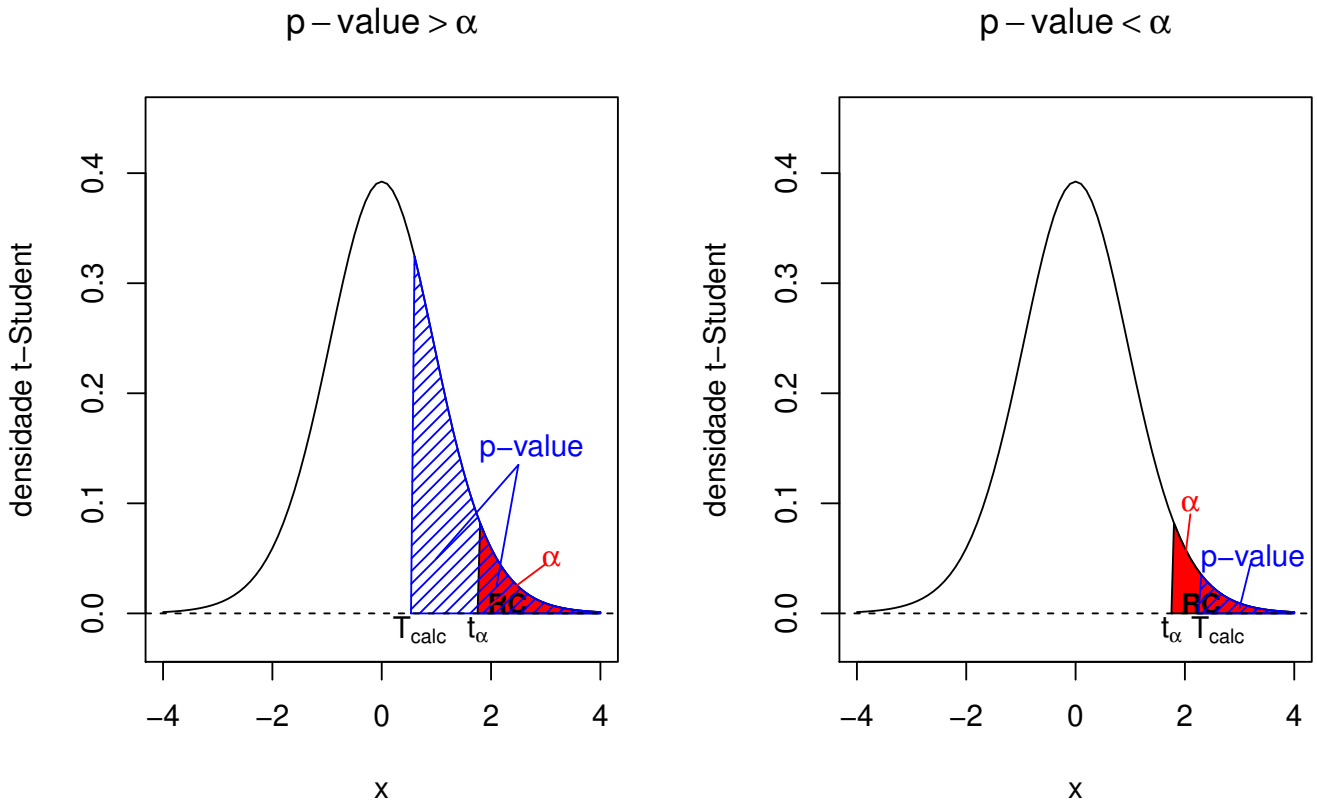
O cálculo do *p-value* é feito de forma diferente, consoante a natureza das hipóteses nula e alternativa conduza a regiões de rejeição unilaterais ou bilaterais.

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

$$p = P[t_{n-2} > T_{calc}] .$$

A relação de *p-values* e níveis de significância

- $p-value > \alpha \Rightarrow$ não rejeição de H_0 ao nível α;
- $p-value < \alpha \Rightarrow$ 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 das hipóteses nula e alternativa:

Teste Unilateral direito $p = P[t_{n-2} > T_{calc}]$

Teste Unilateral esquerdo $p = P[t_{n-2} < T_{calc}]$

Teste Bilateral $p = 2 \times P[t_{n-2} > |T_{calc}|]$.

Testes de hipóteses no R

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

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

ou seja,

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

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

De novo o exemplo dos lírios

Ainda o exemplo dos lírios

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

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

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

cujo valor de prova (*p-value*) é inferior à precisão da máquina ($< 2 \times 10^{-16}$), 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(iris.lm)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.363076   0.039762  -9.131  4.7e-16 ***
Petal.Length  0.415755   0.009582  43.387  < 2e-16 ***
```

Valor da estatística no teste $H_0 : \beta_1 = 0.5$ vs. $H_1 : \beta_1 \neq 0.5$:

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

O exemplo dos lírios (cont.)

Ainda o exemplo dos lírios (cont.)

O valor de prova (bilateral) associado a T_{calc} calcula-se como indicado no acetato 109:

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

Eis o *p-value* calculado no `R`:

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

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

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

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

com c_i e d_i definidos nos acetatos 76 e 77.

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

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

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

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

NOTA: Tal como para as distribuições iniciais de $\hat{\beta}_0$ e $\hat{\beta}_1$ (acetato 80), também esta distribuição não é ainda utilizável devido à presença da variância (desconhecida) dos erros aleatórios, σ^2.

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

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

Dado o Modelo de Regressão Linear Simples, tem-se

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

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

NOTA: A justificação desta distribuição é totalmente análoga à das distribuições de $\hat{\beta}_1$ e $\hat{\beta}_0$ dadas no acetato 86.

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

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:

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

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

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

Inferência sobre $E[Y|X=x]$ no R

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 dos lírios (`iris`, Ex. 8 e 15)

A largura esperada de pétalas de comprimento 1.85 e 4.65, é dada por:

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

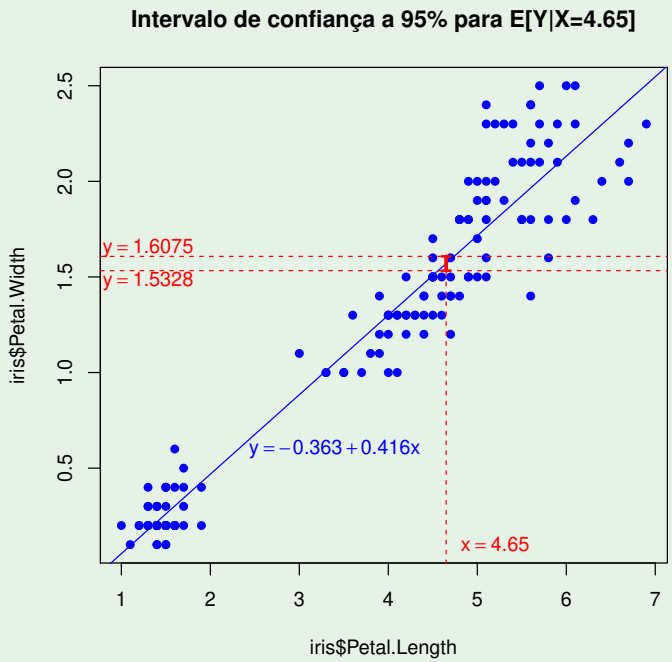
A omissão do argumento `new` produz os valores ajustados de y, os $\hat{y}_i$ associados com os dados usados.

Inferência sobre $E[Y|X = x]$ no R (continuação)

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

IC para $\mu_{Y|X}$ nos dados dos lírios

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

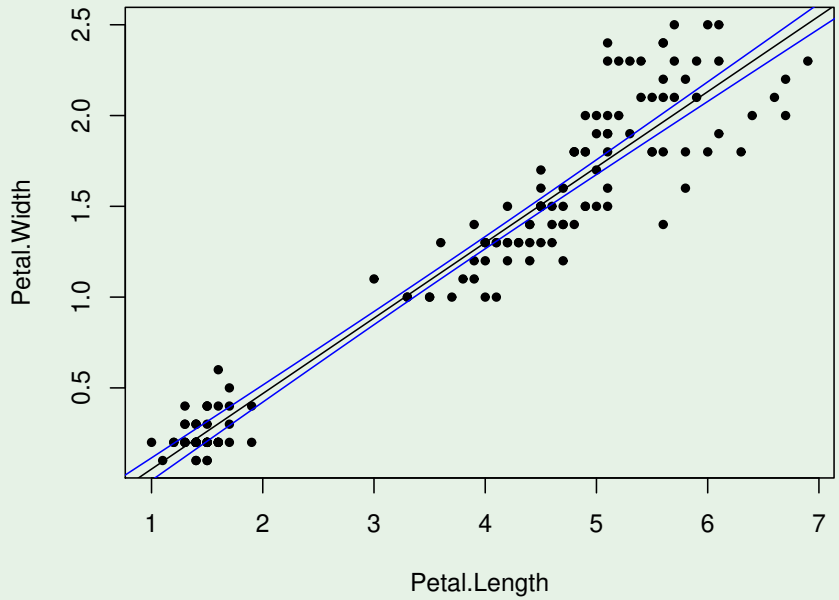


Bandas de confiança para a recta de regressão

Considerando os ICs para uma gama de valores de x, obtêm-se bandas de confiança para a recta de regressão populacional.

Bandas de confiança para a recta populacional dos lírios

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 $\overline{x}$ das observações. As bandas são encurvadas.

A variabilidade duma observação individual de Y

Os ICs acabados de calcular dizem respeito à estimação do valor esperado de Y, para um dado valor de x, $\mu_{Y|x} = \beta_0 + \beta_1 x$. Mas uma observação individual de Y tem uma variabilidade adicional:

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

O estimador de $\mu_{Y|x}$ tem variância (acetato 118):

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

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

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

$$\sigma^2_{Indiv} \; = \; \sigma^2 \left[\frac{1}{n} + \frac{(x - \overline{x})^2}{(n-1)\, s_x^2} \right] \; + \; \sigma^2 \; = \; \sigma^2 \left[1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{(n-1)\, s_x^2} \right] .$$

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*. Assim:

Intervalo de predição para observação individual de Y

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

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

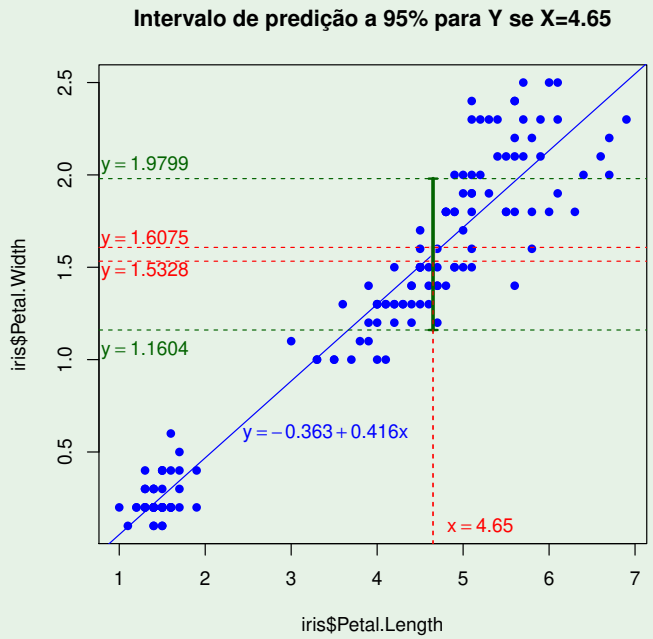
Estes intervalos são (para um mesmo nível $(1-\alpha) \times 100\%$) necessariamente de maior amplitude que os intervalos de confiança para o valor esperado (médio) de Y, $E[Y|X = x]$, vistos antes.

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 nos lírios

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

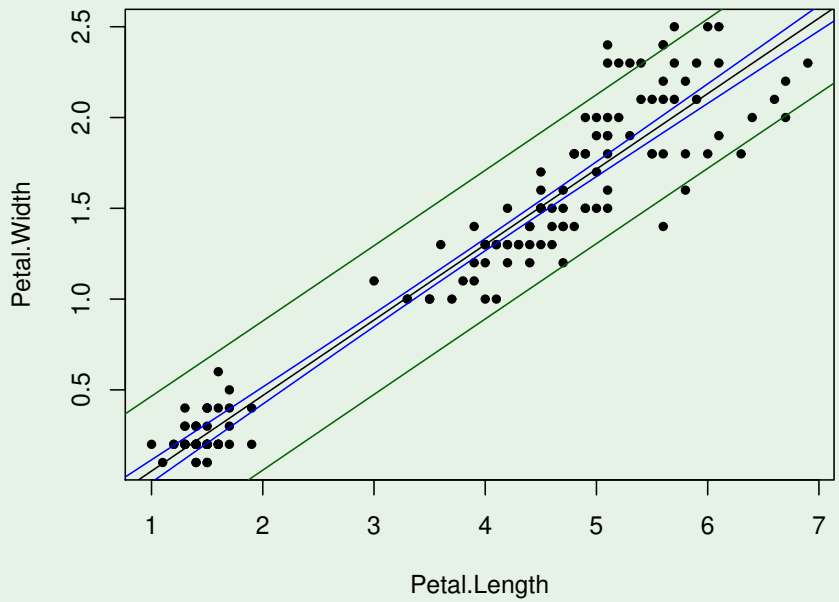


Bandas de predição para uma observação de Y

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

Bandas de predição para Y nos lírios

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



Avaliando a qualidade do ajustamento do Modelo

Como avaliar a qualidade do ajustamento do Modelo?

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

  O teste de ajustamento global do modelo tem a hipótese nula de que o modelo é inútil para prever Y a partir de X:

$$H_0 : \mathscr{R}^2 = 0 \ ,$$

  onde $\mathscr{R}^2$ é o coeficiente de determinação populacional.

  Vamos testar se o R^2 amostral é significativamente diferente de zero.

Avaliando o ajustamento do Modelo (cont.)

O Modelo de Regressão Linear Simples é inútil se $\beta_1 = 0$, isto é, se o Modelo se reduzir ao Modelo Nulo:

$$Y = \beta_0 + \beta_1 x + \varepsilon \qquad \Longrightarrow \qquad Y = \beta_0 + \varepsilon \ .$$

Na RLS pode testar-se essa hipótese de duas maneiras:

- Testar $H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0$, usando o teste t de hipóteses a β_1, considerado no acetato 106.
- Efectuar o teste F ao ajustamento global do modelo. Este teste é descrito seguidamente.

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 80 viu-se que: $\frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{(n-1)\, s_x^2}}} \frown \mathcal{N}(0,1)$.
- Logo, $\frac{(\hat{\beta}_1 - \beta_1)^2}{\sigma^2 / \left[(n-1)\, s_x^2\right]} \frown \chi_1^2$. [Recordar: $Z \frown \mathcal{N}(0,1) \Rightarrow Z^2 \frown \chi_1^2$].
- Se $\beta_1 = 0$, tem-se: $\frac{\hat{\beta}_1^2 \, (n-1) \, s_x^2}{\sigma^2} = \frac{SQR}{\sigma^2} \frown \chi_1^2$.

A quantidade SQR/σ^2 cuja distribuição agora se conhece depende da incógnita σ^2. Esse problema será contornado de forma diferente do que antes.

SQR e *SQRE*

- Sabemos (acetato 83) que $SQRE/\sigma^2 \frown \chi^2_{n-2}$.
- Sabemos (da disciplina de Estatística) que as distribuições F surgem da seguinte forma:

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

- É possível mostrar que *SQRE* e *SQR* são v.a. independentes.
- Logo, se $\beta_1 = 0$, tem-se $W = \frac{SQR}{\sigma^2} \frown \chi^2_1$ e $V = \frac{SQRE}{\sigma^2} \frown \chi^2_{n-2}$, e

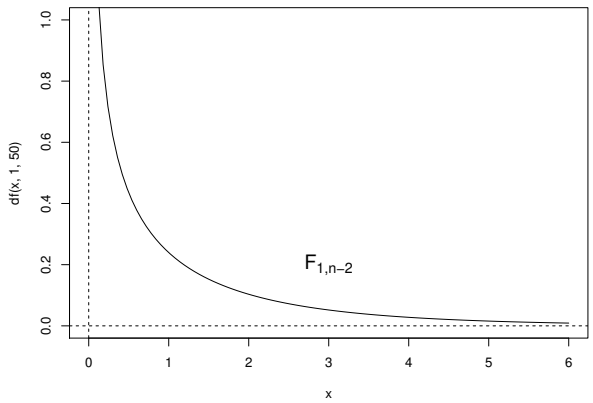
$$\frac{W/1}{V/(n-2)} = \frac{\frac{SQR}{\cancel{\sigma^2}\cdot 1}}{\frac{SQRE}{\cancel{\sigma^2}\cdot(n-2)}} = \frac{QMR}{QMRE} \frown F_{(1,n-2)} ,$$

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

Como usar a estatística F

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

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



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 $F = QMR/QMRE$.

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

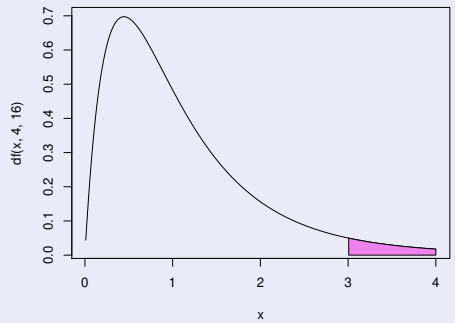
Hipóteses: $H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0$.

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

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

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



O Teste F de ajustamento global do Modelo (cont)

Podem-se re-escrever as hipóteses e estatística do teste usando Coeficientes de Determinação (ver Exercício RLS 16):

Teste F de ajustamento global do modelo

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

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

A informação essencial para efectuar um teste F ao ajustamento global de um modelo de regressão também se obtém através do comando `summary`, aplicado a 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(iris.lm)
(...)
Residual standard error: 0.2065 on 148 degrees of freedom
Multiple R-Squared: 0.9271,     Adjusted R-squared: 0.9266
F-statistic:  1882 on 1 and 148 DF,  p-value: < 2.2e-16
```

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_i:

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

Multiple R-squared: O Coeficiente de Determinação:

$$R^2 = \frac{SQR}{SQT} = \frac{s^2_{\hat{y}}}{s^2_y} = 1 - \frac{SQRE}{SQT}$$

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

$$R^2_{mod} = 1 - \frac{QMRE}{QMT} = 1 - \frac{\hat{\sigma}^2}{s^2_y}, \quad \left(QMT = \frac{SQT}{n-1}\right)$$

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.

A validade dos intervalos de confiança e testes de hipóteses atrás referidos depende da validade desses pressupostos.

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.

Vejamos a distribuição dos resíduos, caso sejam válidos os pressupostos do modelo linear (ver também Exercício RLS 22).

A distribuição dos Resíduos no Modelo RLS

Distribuição dos Resíduos no Modelo RLS

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

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

Um resíduo também é uma combinação linear dos Y_i, logo tem distribuição Normal:

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

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

O modelo RLS admite erros aleatórios com distribuição $\varepsilon_i \frown \mathcal{N}(0, \sigma^2)$.
Mas os resíduos E_i têm variâncias diferentes: $V[E_i] = \sigma^2(1-h_{ii})$.

Dado o Modelo, $\frac{E_i}{\sqrt{\sigma^2(1-h_{ii})}} \frown \mathcal{N}(0,1)$.

Diferentes tipos de resíduos

Três variantes de resíduos

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

Resíduos (internamente) estandardizados : $R_i = \frac{E_i}{\sqrt{QMRE \cdot (1-h_{ii})}}$.

Resíduos Studentizados (ou externamente estandardizados):

$$T_i = \frac{E_i}{\sqrt{QMRE_{[-i]} \cdot (1-h_{ii})}}$$

$QMRE_{[-i]}$ é o valor de $QMRE$ resultante de ajustar a regressão excluíndo a observação i (associada ao resíduo E_i).

Resíduos estandardizados não têm unidades de medida. Costumam estar no intervalo $[-3, 3]$.

É possível mostrar que $T_i = R_i \sqrt{\frac{n-3}{n-2-R_i^2}}$.

Como analisar os resíduos

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

Resíduos usuais (E_i): `residuals`

Resíduos estandardizados (R_i): `rstandard`

Resíduos Studentizados (T_i): `rstudent`

Não se efectuam testes de Normalidade aos resíduos usuais, uma vez que os resíduos não são independentes, como se pode verificar a partir do facto de que somam zero (ver Exercício RLS 5).

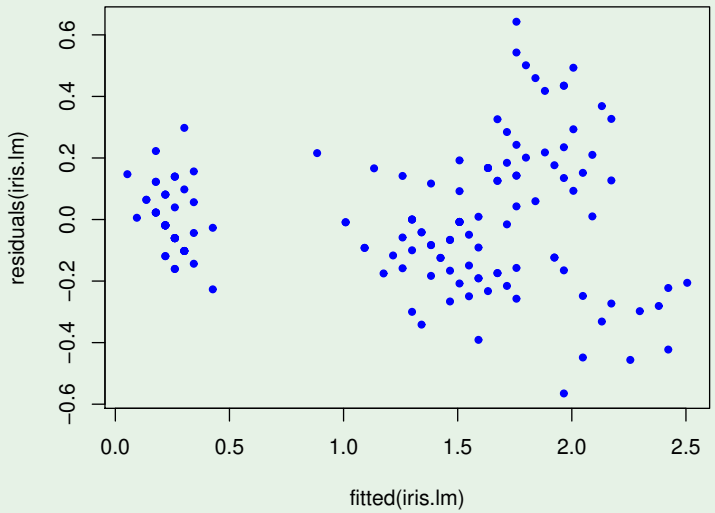
É hábito validar os pressupostos do Modelo de Regressão através de gráficos dos (vários tipos) de resíduos.

Gráficos de resíduos vs. $\hat{Y}_i$

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

Exemplo dos lírios (Ex.8 e 15)

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



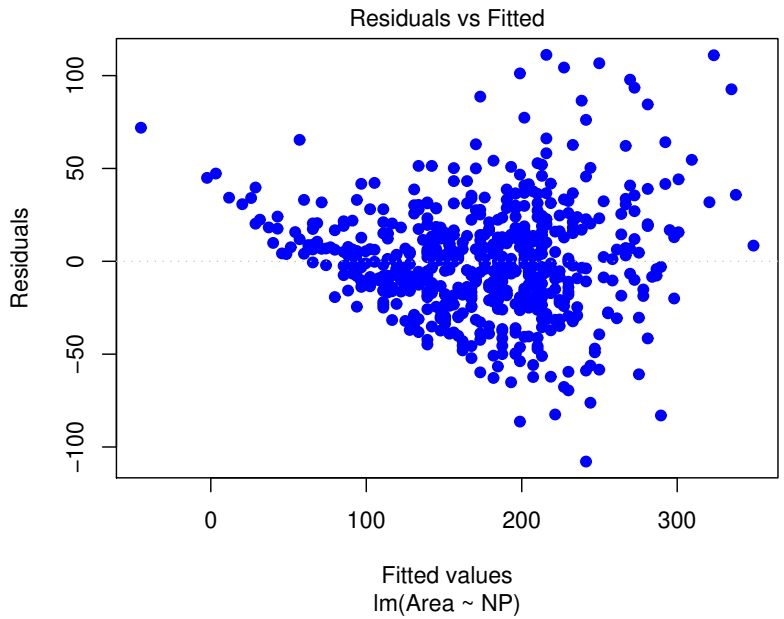
Os resíduos devem dispor-se aproximadamente numa banda horizontal em torno de zero. Sendo válido o Modelo RLS, $cor(E_i, \hat{Y}_i) = 0$ (ver Exercício 22).

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.

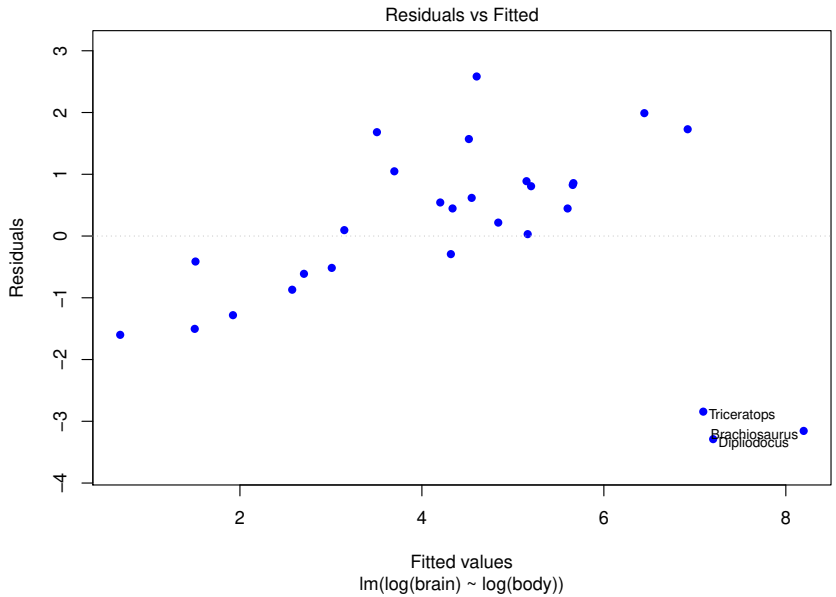
Gráfico em forma de funil: Indica violação da hipótese de homogeneidade de variâncias.



Um exemplo de resíduos em forma de funil e sugerindo alguma curvatura na relação entre as duas variáveis (dados `videiras`).

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` (Ex.9 e 19) cria uma banda oblíqua de pontos no gráfico E_i vs. $\hat{Y}_i$.

Gráficos para estudar a hipótese de normalidade

Como foi visto no acetato 138, dado o Modelo, $\frac{E_i}{\sqrt{\sigma^2\,(1-h_{ii})}} \frown \mathcal{N}(0,1)$.

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

É hábito investigar a validade do pressuposto de erros aleatórios Normais através de:

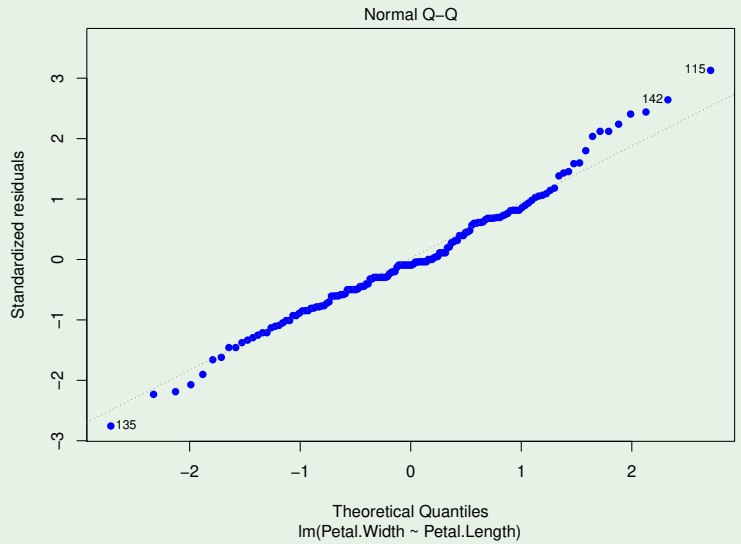
- Um histograma dos resíduos standardizados; ou
- um qq-plot que confronte os quantis empíricos dos n resíduos standardizados, com os quantis teóricos numa $\mathcal{N}(0,1)$.

Gráficos para o estudo da Normalidade (cont.)

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

Exemplo dos lírios (Ex. 8 e 15)

O *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.

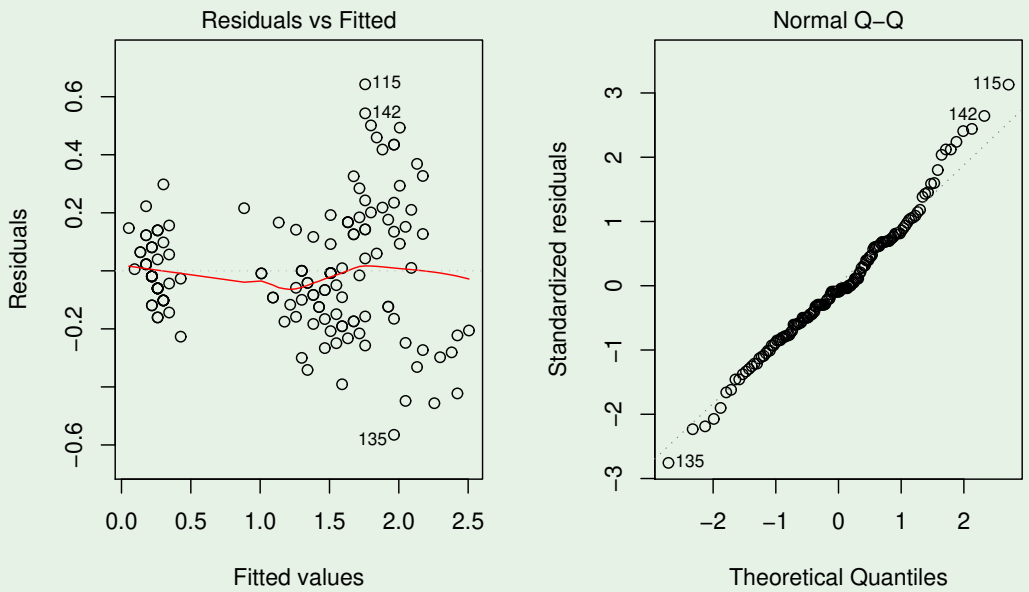
Caso haja suspeita de correlação no tempo ou no espaço, será útil inspeccionar gráficos de resíduos vs. ordem de observação ou posição no espaço, para verificar se existem padrões que sugiram falta de independência.

Estudo de resíduos no R

O comando `plot`, aplicado ao resultado da função `lm` pode produzir seis gráficos, sendo os dois primeiros os que foram vistos antes.

Exemplo dos lírios (Ex. 8 e 15)

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



Observações atípicas

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

Observações atípicas (*outliers* em inglês). Conceito sem definição rigorosa, procura designar 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 ou Studentizados têm distribuição aproximadamente $\mathcal{N}(0,1)$ para n grande, observações para as quais $|R_i| > 3$ ou $|T_i| > 3$ podem ser classificadas como atípicas.

Mas observações distantes da tendência geral podem afectar o próprio ajustamento do modelo, e não serem facilmente identificáveis a partir dos seus resíduos.

O "efeito alavanca"

Efeito alavanca (*leverage*)

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

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

O valor h_{ii} aparece na variância do i-ésimo resíduo E_i (acetato 138):

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

Para h_{ii} grande, $V[E_i] \approx 0$, 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 $\overline{x}$, 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:

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

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 (i.e., podem, ou não, ser atípicas).

Observações influentes

Uma observação diz-se influente 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:

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

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

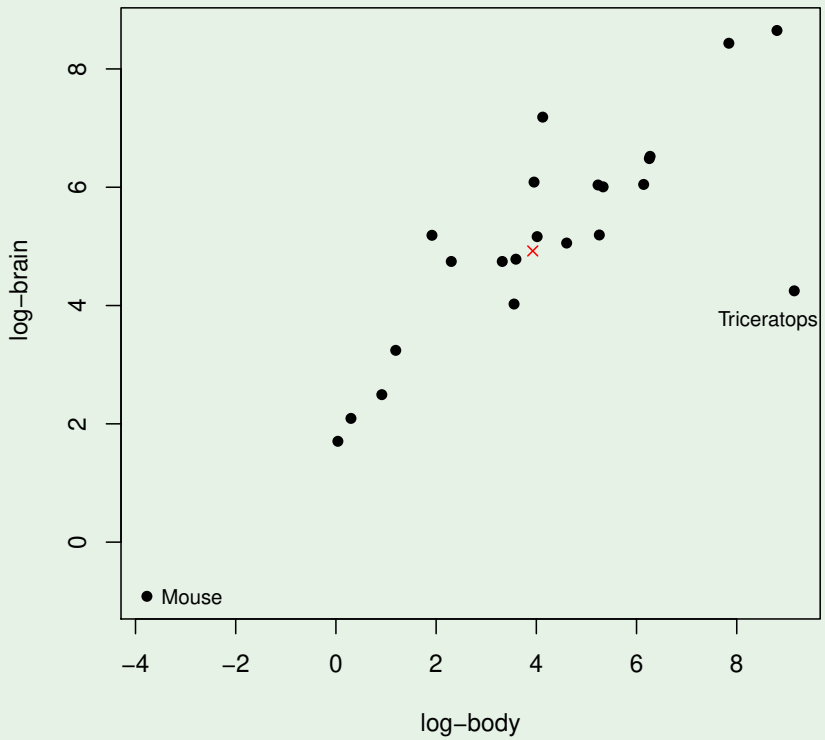
(Se obs. i é, ou não, influente depende da grandeza relativa de R_i^2 e h_{ii})

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

Um exemplo

Exemplo (`Animals`, Ex. 9)

Considerando apenas um subconjunto de 23 das 28 espécies animais do Exercício RLS 9, obtém-se o seguinte gráfico de log-peso do corpo *vs.* log-peso do cérebro:



Há duas espécies mais distantes da nuvem de pontos, mas enquanto o rato se dispõe na mesma tendência de fundo, o *triceratops* não.

A cruz (`x`) indica o centro de gravidade $(\overline{x}, \overline{y})$ da nuvem de pontos.

Exemplo (cont.)

Exemplo (cont.)

Os Resíduos (internamente) estandardizados, distâncias de Cook e valores do efeito alavanca são os seguintes:

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

Gráficos diagnósticos no 

A função `plot`, aplicada a um objecto `lm` produz, além dos gráficos vistos no acetato 147, gráficos com alguns dos diagnósticos agora considerados.

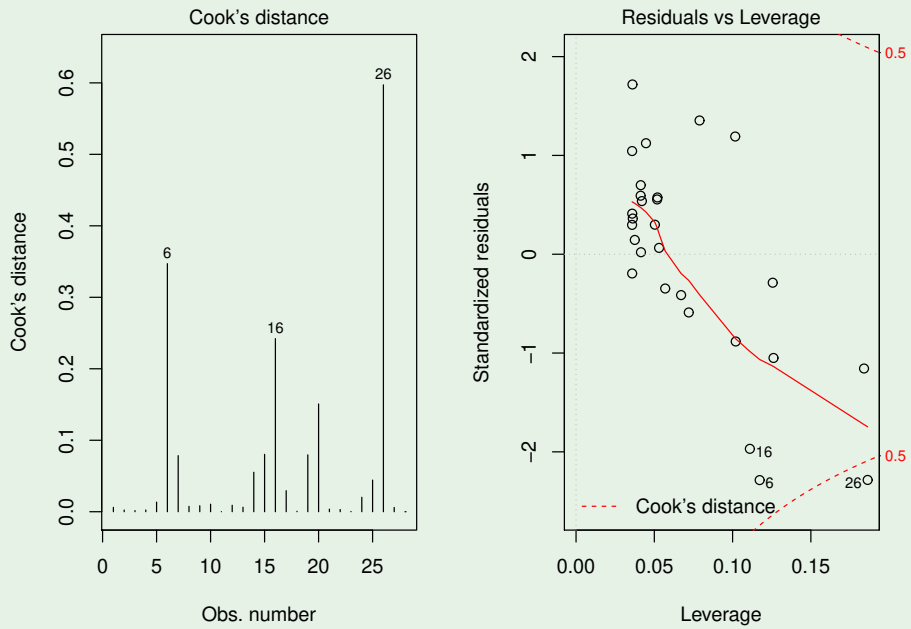
A opção `which=4` produz um diagrama de barras das distâncias de Cook associadas a cada observação.

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

Exemplo de gráficos de diagnóstico

Exemplo: dados `Animals` (Ex. RLS 9 e 19)

Para (a totalidade) dos dados dos Exercícios RLS 9 e 19 (`Animals`):



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

Algumas transformações de variáveis

Por vezes, é possível tornear violações às hipóteses de Normalidade dos erros aleatórios ou homogeneidade de variâncias através de transformações de variáveis. Por exemplo,

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

são propostas usuais para estabilizar as variâncias.

Existe toda uma família Box-Cox de transformações dependentes dum parâmetro (λ):

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

Prevenções sobre transformações

Mas a utilização de transformações da variável resposta Y (e possivelmente também do preditor X) deve ser feita com cautela.

- 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 duma amostra específica, mas não tenham qualquer generalidade.

Transformações linearizantes

Diferente é o problema (já visto mais atrás) de transformações que visam linearizar uma relação original não linear entre x e y.

Prevenções sobre transformações linearizantes:

- Os estimadores que minimizam a soma de quadrados dos resíduos nas relações linearizadas não são os que produzem as soluções óptimas dum problema de minimização de somas de quadrados de resíduos na relação não-linear original.
- As transformações não levaram em conta os erros aleatórios.
- As hipóteses de erros aleatórios aditivos, Normais, de variância homogénea, média zero e independentes terão de ser válidas para as relações lineares entre as variáveis transformadas.