Capítulo 2

Modelo Linear

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), $x_1, x_2, \ldots, x_p$.

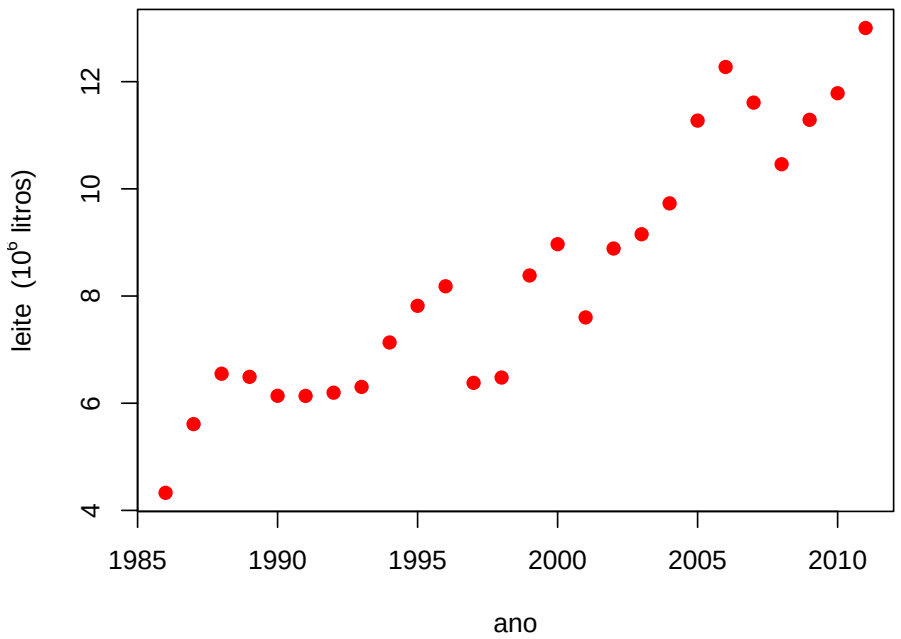
Informação: A identificação da relação de fundo é feita com base em n observações do conjunto de variáveis envolvidas na relação.

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

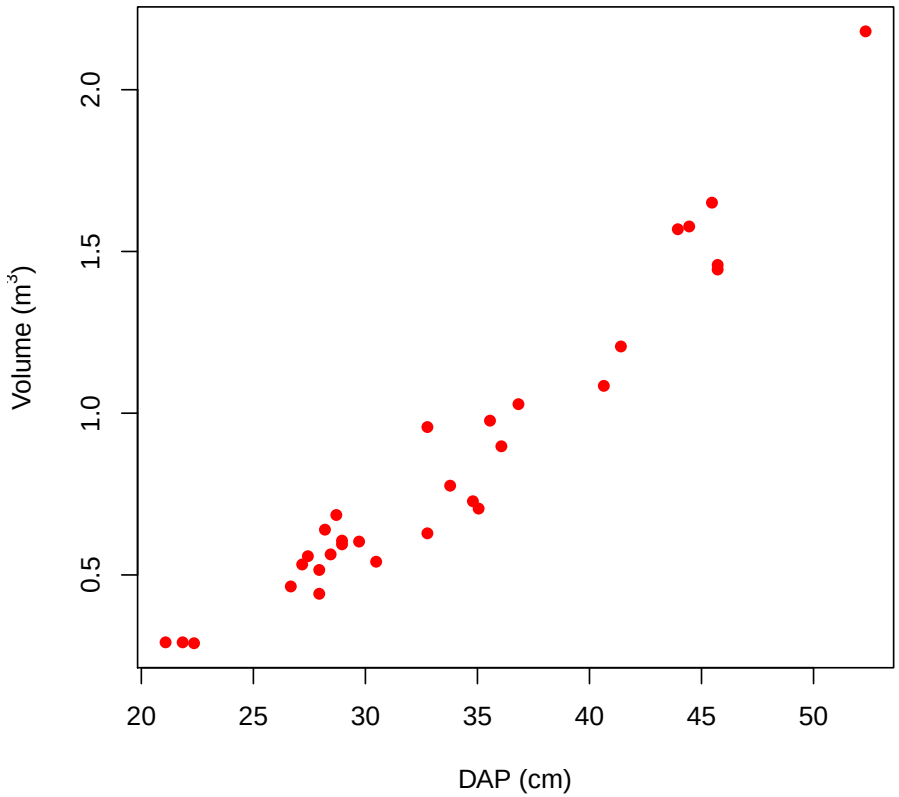
Produção de leite de cabra em Portugal (y, em 10^6 litros) (INE)
vs. Anos (x, de 1986 a 2011) $n = 26$ pares de valores, $\{(x_i, y_i)\}_{i=1}^{26}$



A tendência de fundo é aproximadamente linear.
Interessa o contexto descritivo.
Qual a "melhor" equação de recta, $y = b_0 + b_1\, x$, para descrever as n observações (e qual o critério de "melhor")?

Exemplo 2 - relação linear

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

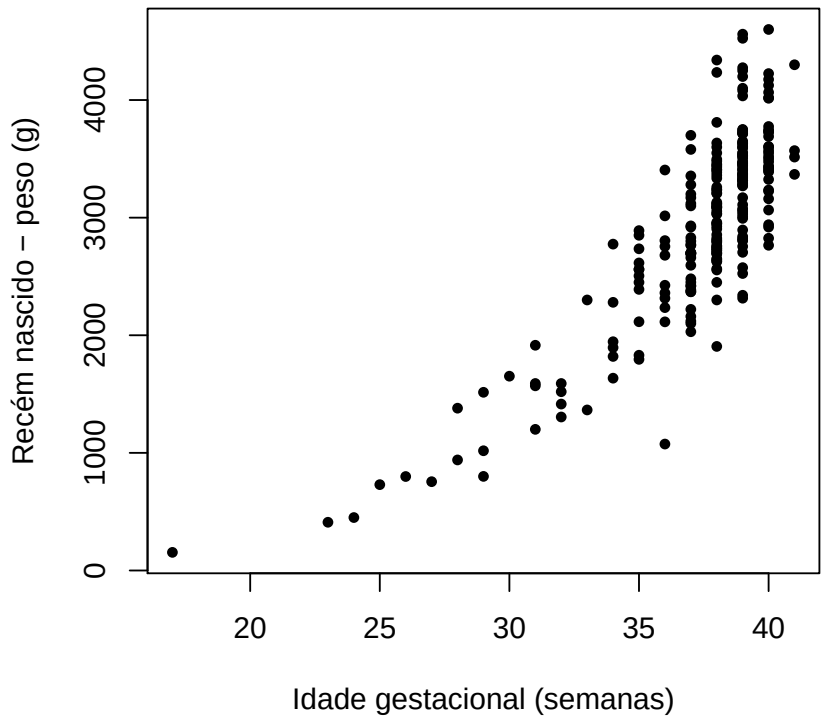


A tendência de fundo é aproximadamente linear. Mas os $n = 31$ pares de observações são apenas uma amostra aleatória duma população mais vasta. Interessa o contexto inferencial: o que se pode dizer sobre a recta populacional $y = \beta_0 + \beta_1 x$?

Exemplo 3 - Uma relação não linear

$n = 251$ 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 ($y = c\,x^d$)?

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?

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.

O Modelo Linear

- O Modelo Linear é um caso particular de modelação estatística;
- engloba um grande número de modelos específicos:
  Regressão Linear (Simples e Múltipla) , Regressão Polinomial, Análise de Variância, Análise de Covariância;
- é o mais completo e bem estudado tipo de modelo;
- serve de base para numerosas extensões
  (Regressão não linear, Modelos Lineares Generalizados, Modelos Lineares Mistos, etc.).

Revisão: Reg. Linear Simples - contexto descritivo

Estudado na disciplina de Estatística (1os. ciclos do ISA),

- apenas como regressão linear **simples**
- apenas no contexto **descritivo**

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

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

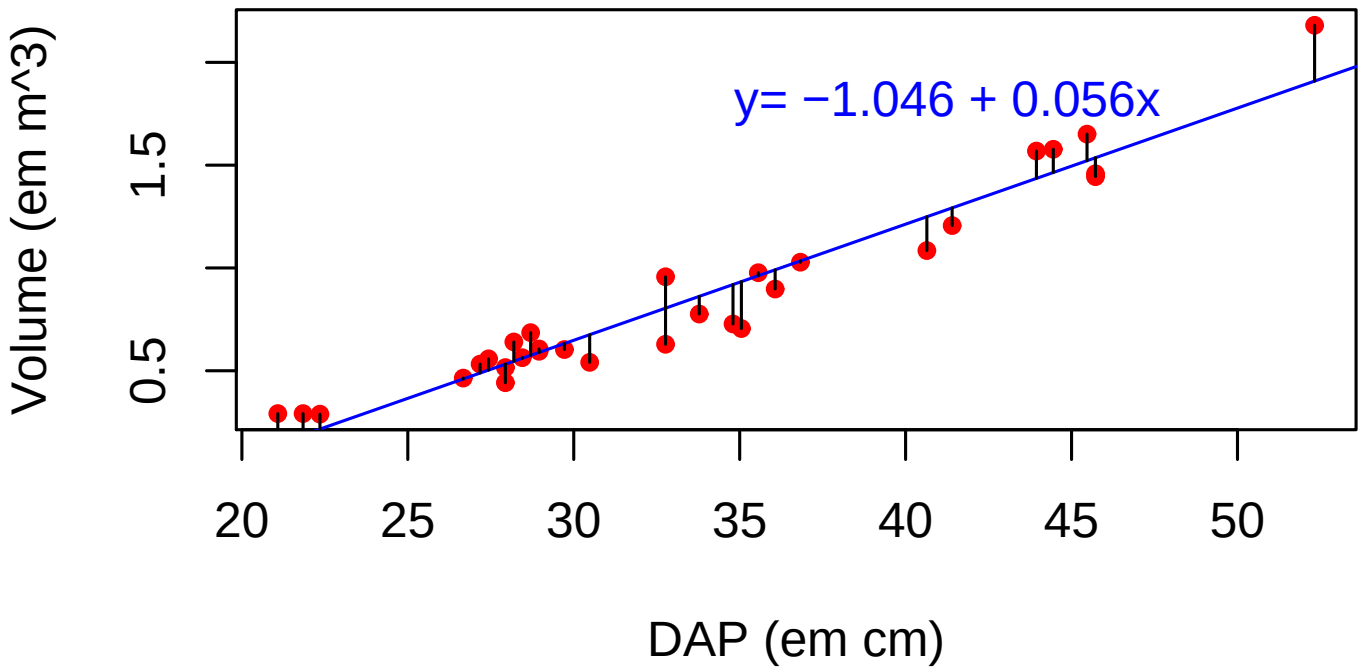
sendo

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

Regressão Linear Simples - contexto descritivo

$n = 31$ pares de medições

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



Revisão: Reg. Linear Simples descritiva (cont.)

Como se chegou à equação da recta?

Critério: Minimizar a soma de quadrados residual (i.e., 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:

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

Critério: 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). (Recordar a matéria de Análise Matemática dos 1os. ciclos).

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

$$e_i = y_i - \hat{y}_i$$

é o desvio (com sinal) da observação y_i face à sua previsão a partir da recta.

> *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

Sabe-se que:

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

No exemplo das cerejeiras: por cada cm a mais no DAP, o volume do tronco aumenta, em média, $0.056 m^3$.

Revisão: 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 5).

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

Revisão: RLS - As três Somas de Quadrados

Algumas quantidades e propriedades importantes na Regressão Linear Simples descritiva.

- s_y^2 - variância amostral dos y_i observados;
- $s_{\hat{y}}^2$ - variância amostral dos $\hat{y}_i$ ajustados;
- s_e^2 - variância amostral dos resíduos e_i;

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

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

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

Revisão: RLS - Fórmula fundamental e R^2

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

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

Papel crucial é desempenhado pelo Coeficiente de Determinação:

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

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

Numa regressão linear simples, tem-se:

- $R^2 = 1$ se, e só se, os n pontos são colineares.
- R^2 é o quadrado do coeficiente de correlação linear entre preditor e variável resposta (ver Exercício 6):

$$R^2 = r_{xy}^2 = \left(\frac{cov_{xy}}{s_x\, s_y}\right)^2$$

Regressão - um pouco de história

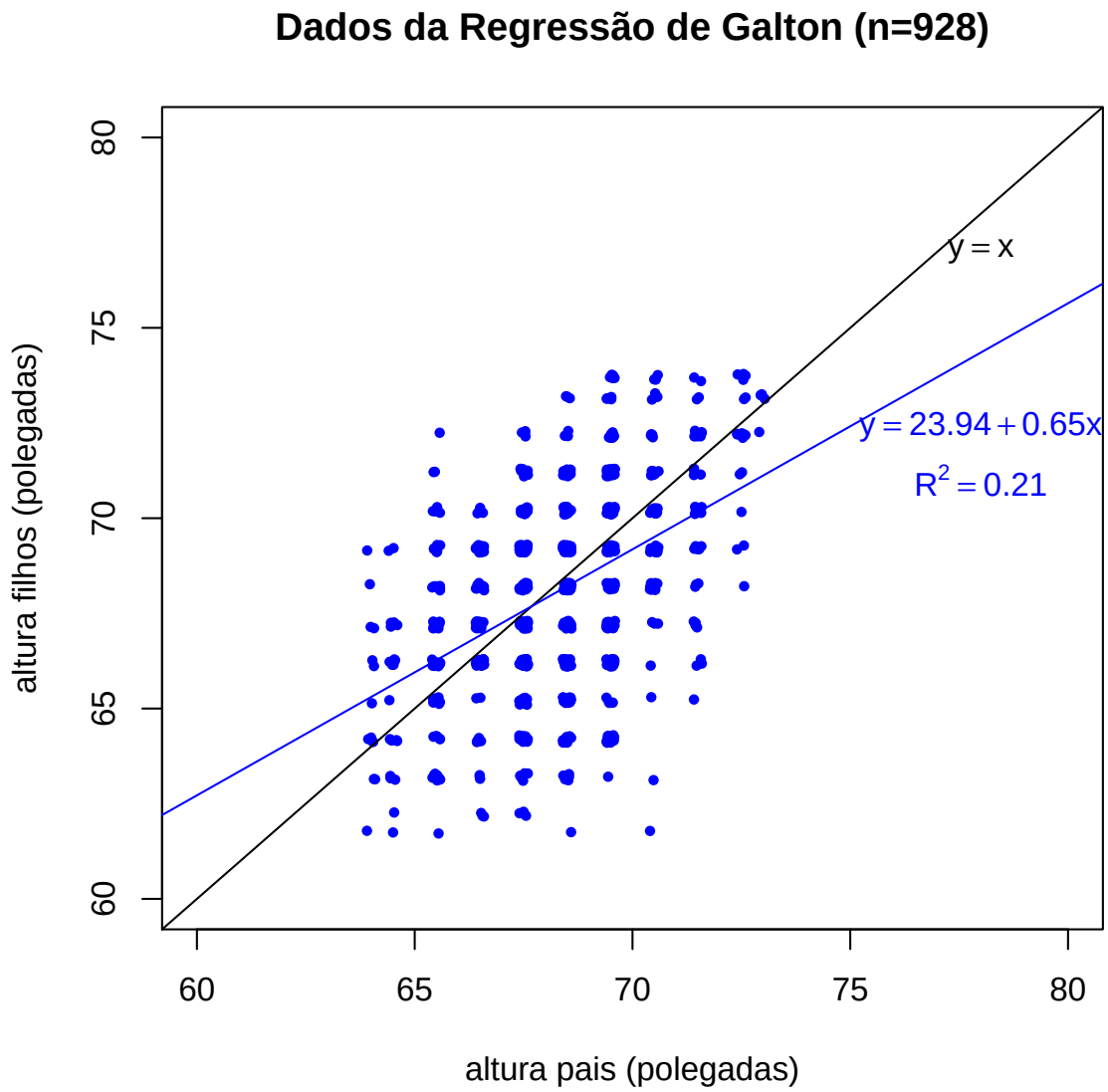
A designação Regressão tem origem num estudo de Francis Galton (1886), relacionando a altura de $n = 928$ jovens adultos com a altura (média) dos pais.

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.

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

Um pouco de história (cont.)



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.

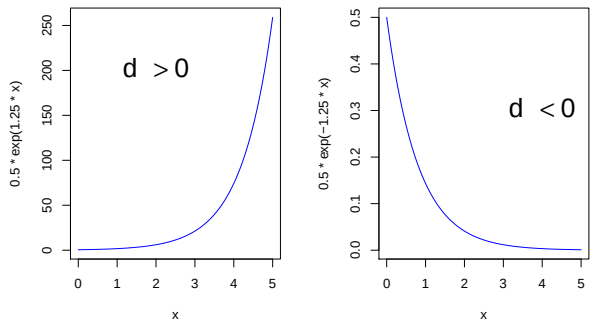
Tais transformações podem permitir utilizar a Regressão Linear Simples, apesar de a relação original ser não-linear.

Vamos ver alguns 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:

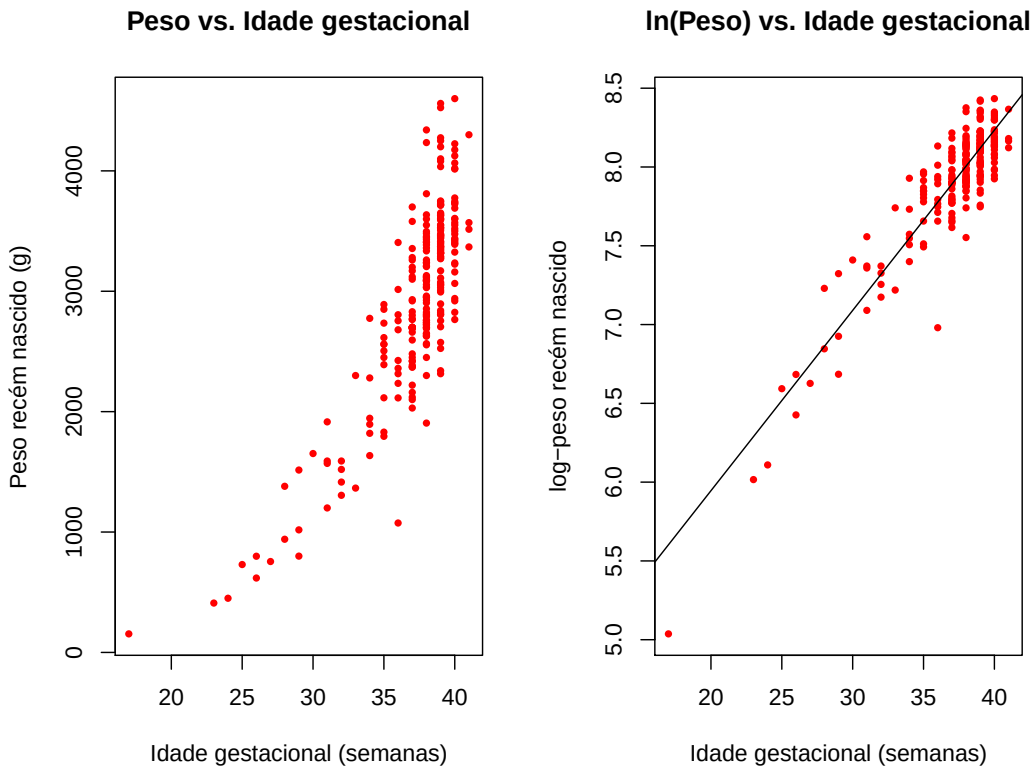
$$\begin{aligned} \ln(y) &= \ln(c) + d\,x \\ \Leftrightarrow \quad y^* &= b_0 + b_1\,x \end{aligned}$$

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 ($b_1 < 0$).

Uma linearização no Exemplo 3

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

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), tem-se:

$$\ln(y(x)) \ = \ \underbrace{d}_{=b_1} x + \underbrace{C}_{=b_0} \qquad \Leftrightarrow \qquad y(x) \ = \ \mathrm{e}^{C}\,\mathrm{e}^{d\,x} .$$

Repare-se que o declive b_1 da recta é o valor (constante) d da taxa de variação relativa de y. A constante de primitivação C é a ordenada na origem da recta: $C = b_0$.

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.

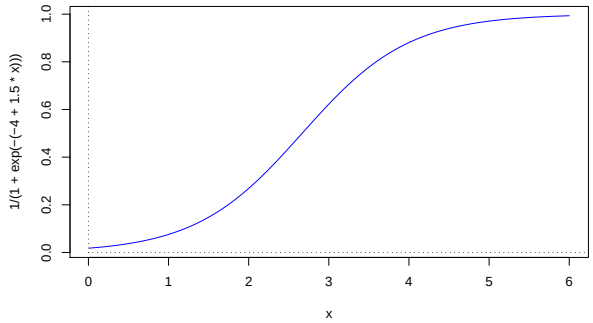
Mas nenhum crescimento populacional exponencial é sustentável a longo prazo.

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. Pode pensar-se 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 : $y = \frac{1}{1+e^{-(c+dx)}}$



$(b_1 > 0)$

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:

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

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

$$\Rightarrow \quad \underbrace{\ln\left(\frac{y}{1-y}\right)}_{=y^*} = \underbrace{c}_{=b_0} + \underbrace{d}_{=b_1} x$$

Ainda a Logística

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

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

De facto, a expressão anterior equivale a:

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

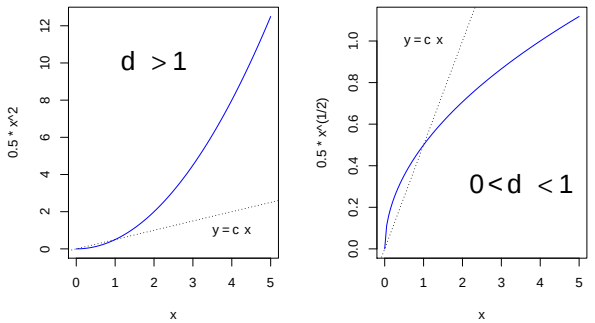
Primitivando (em ordem a x), tem-se:

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

Relação potência ou alométrica

Relação potência : $y = c\,x^d$

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



Transformação : Logaritmizando, obtém-se:

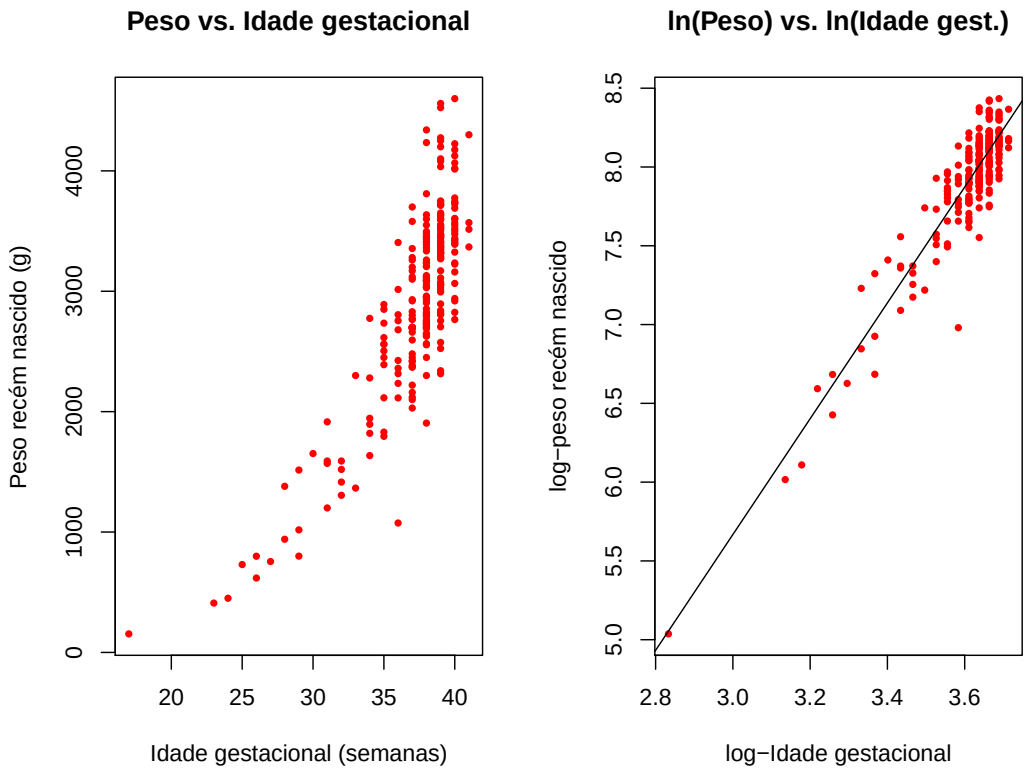
$$\begin{aligned} \ln(y) &= \ln(c) + d\ln(x) \\ \Leftrightarrow \quad y^* &= b_0 + b_1 x^* \end{aligned}$$

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

Outra linearização no Exemplo 3

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 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:

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

De facto, primitivando (em ordem a t), tem-se:

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

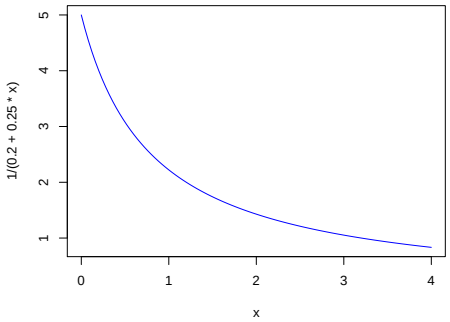
e exponenciando,

$$y \ = \ x^d \cdot \underbrace{\mathrm{e}^C}_{=c}$$

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

Relação hiperbólica (ou de proporcionalidade inversa)

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



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

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

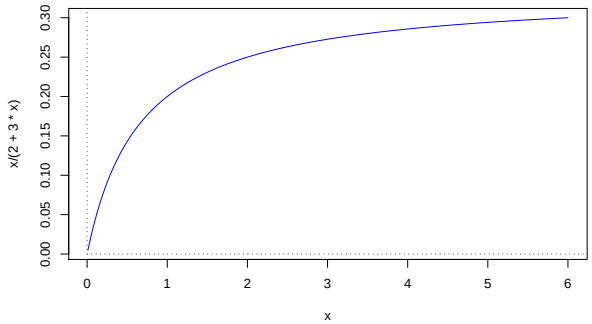
Resulta de admitir que a taxa de variação de y é proporcional ao quadrado de y ou, equivalentemente, que a taxa de variação relativa de y é proporcional a y:

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

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

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 $x^* = \frac{1}{x}$:

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

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

Relação Michaelis-Menten (cont.)

- A relação Michaelis-Menten é muito 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.
- 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.

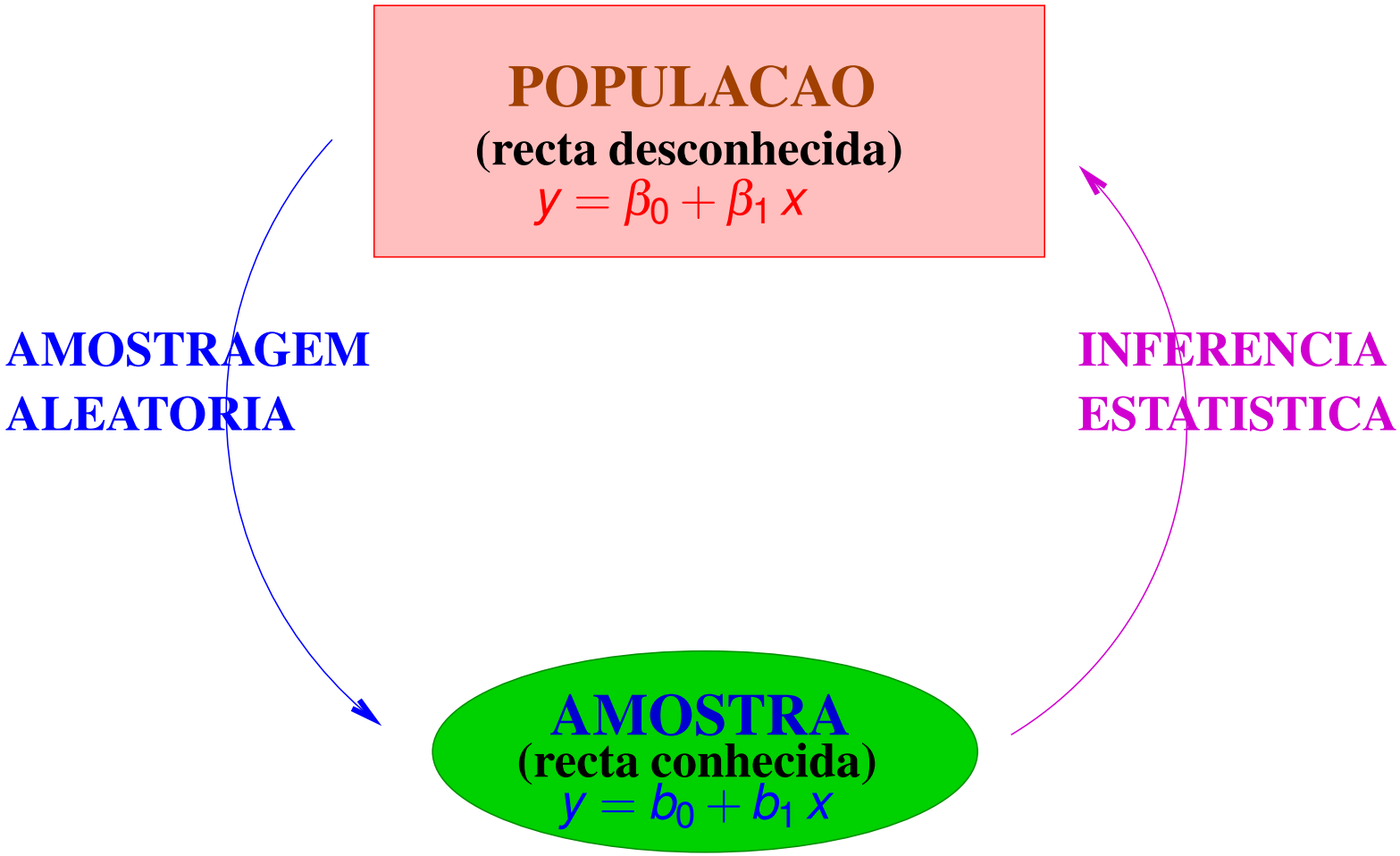
Regressão Linear Simples - INFERÊNCIA

- 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)

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.

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

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).

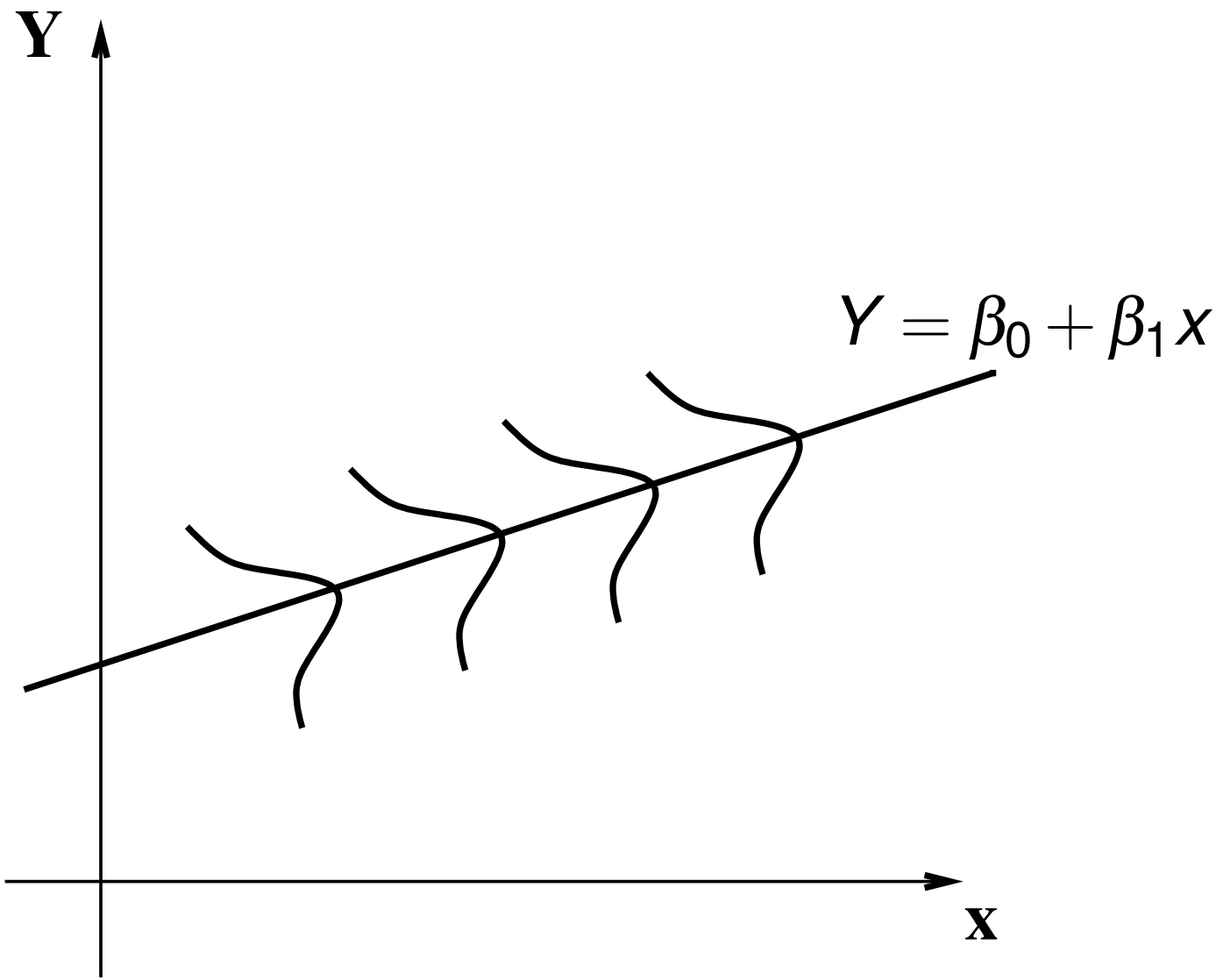
- Têm distribuição Normal (é restritiva, mas bastante geral).
- Homogeneidade de variâncias: têm sempre a mesma variância

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

  (é restritiva, mas conveniente).

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

MODELO Regressão Linear Simples



MODELO - Regressão Linear Simples

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

Definição (O Modelo de Regressão Linear Simples)

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

Revisão: propriedades de valores esperados

O valor esperado ou valor médio duma variável aleatória X é o centro de gravidade da sua distribuição de probabilidades (função de massa probabilística se X discreta, ou função densidade se X contínua).

NOTA: Ver apontamentos de Teoria das Probabilidades (Capítulo II) da Prof. Manuela Neves.

No que se segue, usam-se algumas propriedades dos valores esperados (valores médios) de variáveis aleatórias:

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: propriedades de variâncias

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

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

Algumas propriedades de variâncias 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: propriedades de covariâncias

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

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

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: propriedades da distribuição Normal

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

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

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

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

- Seja $X \cap \mathcal{N}(\mu, \sigma^2)$, então: $\frac{X-\mu}{\sigma} \quad \cap \quad \mathcal{N}(0, 1)$.
- Combinações lineares de Normais independentes têm distribuição Normal. Concretamente: se X, Y são Normais independentes e a, b constantes, então $aX + bY$ é Normal (com parâmetros resultantes das propriedades dos acetatos 122 e 123).

Primeiras consequências do MODELO RLS

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

Teorema (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. $Y_i \cap \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2), \quad \forall i = 1, ..., n.$
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: $b_1 = \frac{cov_{xy}}{s_x^2} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{(n-1)\,s_x^2} \overset{(\star)}{=} \frac{\sum_{i=1}^{n}(x_i-\overline{x})\,y_i}{(n-1)\,s_x^2}$

[$(\star)$ Exercício 3b) de RLS nas aulas práticas]

Definição (Estimador de β_1)

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

Nota: O estimador $\hat{\beta}_1$ é combinação linear de Normais independentes.

Estimação dos parâmetros do Modelo RLS (cont.)

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

Definição (Estimador de β_0)

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x} = \frac{1}{n} \sum_{i=1}^{n} Y_i - \overline{x} \sum_{i=1}^{n} c_i Y_i = \sum_{i=1}^{n} d_i Y_i \,, \tag{2}$$

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

Teorema (Distribuição dos estimadores dos parâmetros)

Dado o Modelo de Regressão Linear Simples,

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

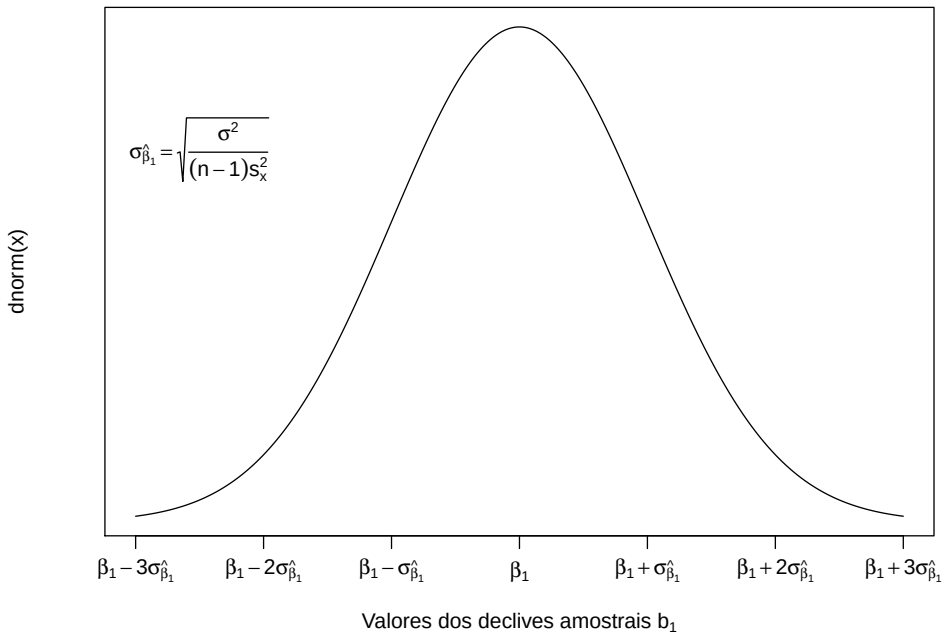
NOTAS:

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 variabilidade dos estimadores.
3. A variabilidade de $\hat{\beta}_0$ também diminui com o aumento de n, e com a maior proximidade de $\overline{x}$ de zero.
4. A demonstração do primeiro resultado está nos Materiais de Apoio (secção Aulas Teóricas) e a do segundo resultado está na resolução do Ex. 13 (secção Aulas Práticas).

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

Corolário

Dado o Modelo de Regressão Linear Simples,

1. $\frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \cap \mathcal{N}(0,1)$, *com* $\sigma_{\hat{\beta}_1} = \sqrt{\frac{\sigma^2}{(n-1)\,s_x^2}} = \sigma / \sqrt{(n-1)\,s_x^2}$
2. $\frac{\hat{\beta}_0 - \beta_0}{\sigma_{\hat{\beta}_0}} \cap \mathcal{N}(0,1)$, *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 Corolário anterior só permitem fazer inferência sobre os parâmetros β_0 e β_1 (e.g., construir intervalos de confiança ou efectuar testes de hipóteses) caso fosse 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

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

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

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

porque a média dos resíduos é zero (ver Exercício 5b de RLS nas aulas práticas).

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

A Soma de Quadrados Residual

Teorema (Resultados distribucionais de SQRE)

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

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

NOTA: Omite-se a demonstração

Corolário

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

Recordar: $X \cap \chi^2_\nu \quad \Rightarrow \quad E[X] = \nu$.

Para propriedades da χ^2, ver apontamentos da Prof. Manuela Neves (Teoria das Probabilidades, p.102 e seguintes).

O Quadrado Médio Residual

Definição (Quadrado Médio Residual)

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

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

- 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 131 transforma a distribuição Normal numa *t-Student*.

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

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

Ver apontamentos Prof. Manuela Neves (Inferência Estatística, Capítulo III, Def. 3.3, p.115).

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

Teorema (Distribuições para a inferência sobre β_0 e β_1)

Dado o Modelo de Regressão Linear Simples, tem-se

1. $\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \cap t_{n-2}$, *com* $\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{QMRE}{(n-1)\, s_x^2}}$
2. $\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \cap t_{n-2}$, *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.

A construção do IC para β_1 (feita na aula Teórica) encontra-se nos Materiais de Apoio (secção Aulas Teóricas).

Intervalo de confiança para β_1

Teorema (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:*

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

sendo $t_{\alpha/2(n-2)}$ *o valor que, numa distribuição* $t_{(n-2)}$*, deixa à direita uma região de probabilidade* $\alpha/2$*. As quantidades* b_1 *e* $\hat{\sigma}_{\hat{\beta}_1}$ *foram definidas em acetatos anteriores.*

NOTA: A amplitude do IC aumenta com *QMRE* e diminui com n e s_x^2:

$$\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{QMRE}{(n-1)\,s_x^2}}$$

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

Intervalo de confiança para β_0

Teorema (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 de regressão populacional é dado por:*

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

onde b_0 *e* $\hat{\sigma}_{\hat{\beta}_0}$ *foram definidos em acetatos anteriores.*

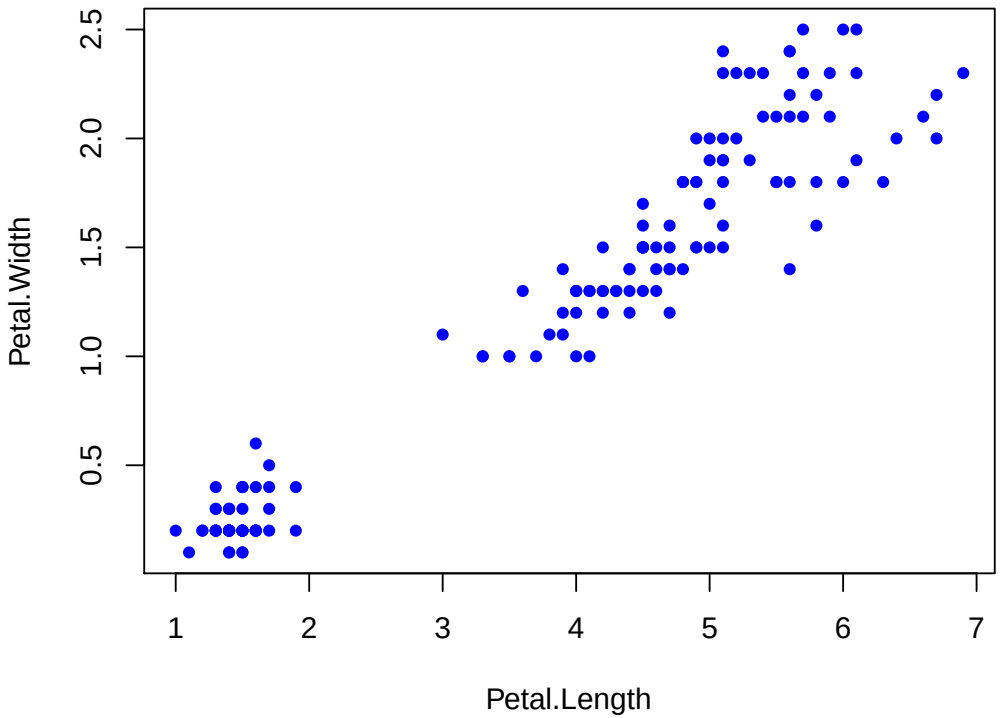
NOTA: A amplitude do IC aumenta com $QMRE$ e com $\overline{x}^2$ e diminui com n e s_x^2:

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

Um exemplo de RLS

A *data frame* `iris`, no `R`, contém medições de 4 variáveis numéricas: comprimento e largura de sépalas e pétalas em $n = 150$ lírios.

Eis a nuvem de pontos relacionando largura e comprimento das pétalas (discutida no Exercício 8 RLS):



Um exemplo de RLS (cont.)

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

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:

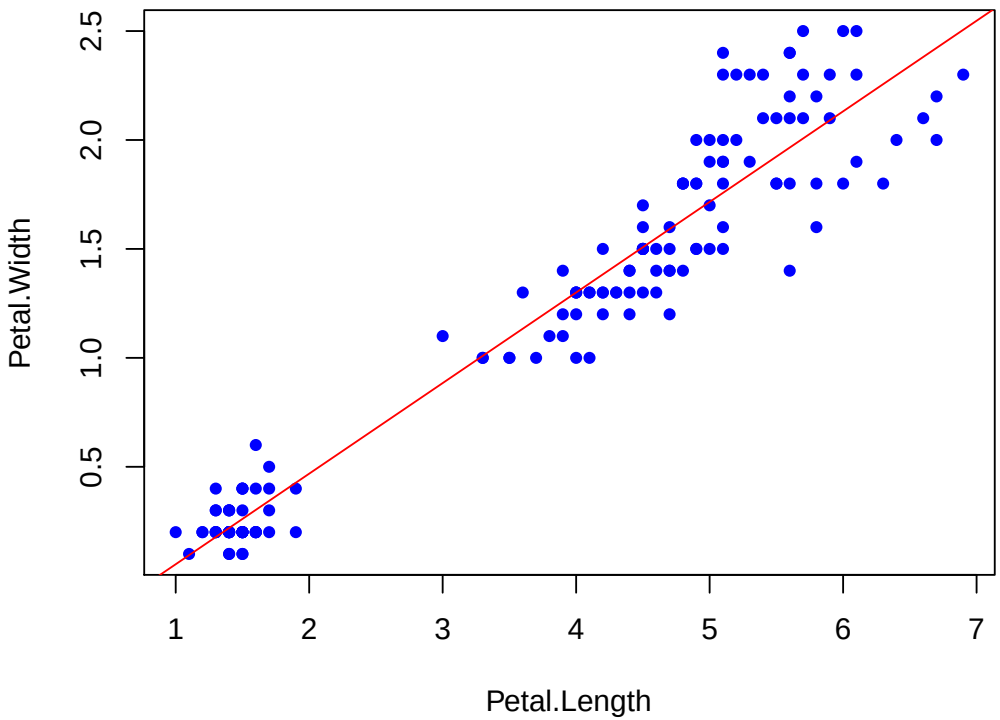
$$y = -0.3631 + 0.4158\,x$$

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

Um exemplo de RLS (cont.)

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.)

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 na construção dos intervalos de confiança para β_0 e β_1.

Intervalos de confiança de β_0 e β_1 no R

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.

Podemos afirmar, a 95% de confiança, que o declive β_1 da recta populacional está no intervalo $]0.397\,,\,0.435[$, e que a respectiva ordenada na origem β_0 pertence ao 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.

Dado um conjunto de observações,

> *quanto maior o grau de confiança* $(1-\alpha)\times 100\%$ *associado a um 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 β_0 e β_1.

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

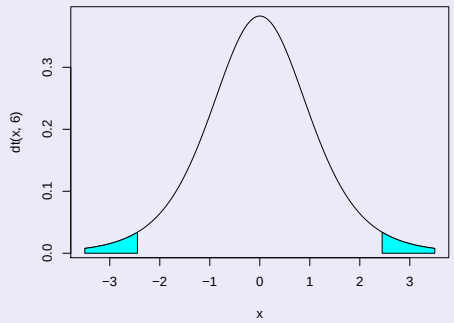
Estatística do Teste: $T = \frac{\hat{\beta}_1 - \overbrace{\beta_1|_{H_0}}^{=c}}{\hat{\sigma}_{\hat{\beta}_1}} \cap t_{n-2} \quad , \quad \text{sob } H_0.$

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 $|T_{calc}| > t_{\alpha/2(n-2)}$



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)

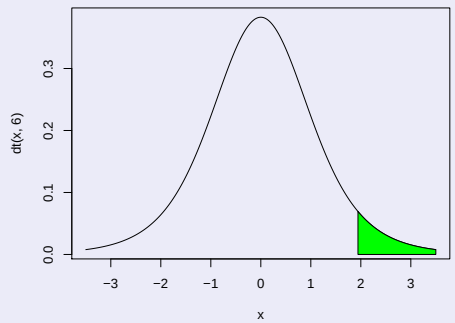
Hipóteses: $H_0 : \beta_1 \leq c$ vs. $H_1 : \beta_1 > c$.

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

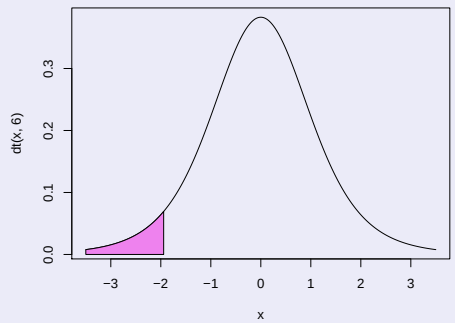
Hipóteses: $H_0 : \beta_1 \geq c$ vs. $H_1 : \beta_1 < c$.

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

Em alternativa a fixar previamente o nível de significância α, é possível indicar apenas o *p*-value associado ao valor calculado da estatística T:

prob. de T tomar valores mais extremos que T_{calc}, sob H_0

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

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: $T = \frac{\hat{\beta}_0 - \overbrace{\beta_0|_{H_0}}^{=c}}{\hat{\sigma}_{\hat{\beta}_0}} \cap t_{n-2}$, 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)
	$\lvert T_{calc} \rvert > t_{\alpha/2(n-2)}$	(Bilateral)
	$T_{calc} > t_{\alpha(n-2)}$	(Unilateral direito)

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.

De novo o exemplo dos lírios

Recordemos os resultados no exemplo dos lírios (acetato 143):

```
> 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.)

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 valor de prova (bilateral) associado a T_{calc} calcula-se como indicado no acetato 150: $p = 2 \times P[t_{n-2} > |-8.792006|]$. 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 $X = x$:

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

O estimador óbvio desta quantidade é

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

usando a notação introduzida nos acetatos 127 e 128.

Nota: O estimador $\hat{\mu}_{Y|x}$ é combinação linear das observações Y_i.

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

Teorema (Distribuição de $\hat{\mu}_{Y|x}$)

Dado o Modelo de Regressão Linear Simples, tem-se

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

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

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

NOTA: Tal como para as distribuições iniciais de $\hat{\beta}_0$ e $\hat{\beta}_1$ (acetato 131), também esta distribuição não é ainda utilizável devido à presença da variância (desconhecida) dos erros aleatórios, σ^2.

A distribuição para inferência sobre $E[Y \mid X = x]$

Teorema (Distrib. de $\hat{\mu}_{Y|X}$, sem quantidades desconhecidas)

Dado o Modelo de Regressão Linear Simples, tem-se

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

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 137.

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

Teorema (IC 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:*

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

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.

Por exemplo, no exemplo dos lírios, 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
```

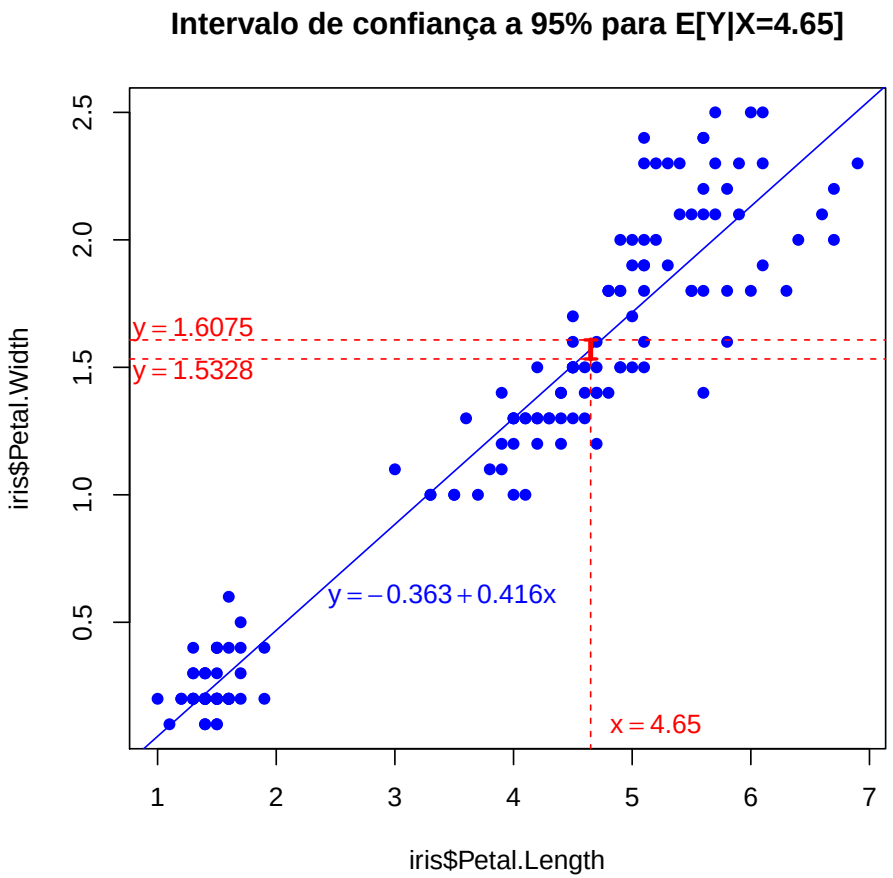
A omissão do argumento `new` produz os valores ajustados de y, os $\hat{y}_i$ associados com os dados usados. Também se pode obter os $\hat{y}_i$ usando o comando `fitted`:

```
> fitted(iris.lm)
```

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

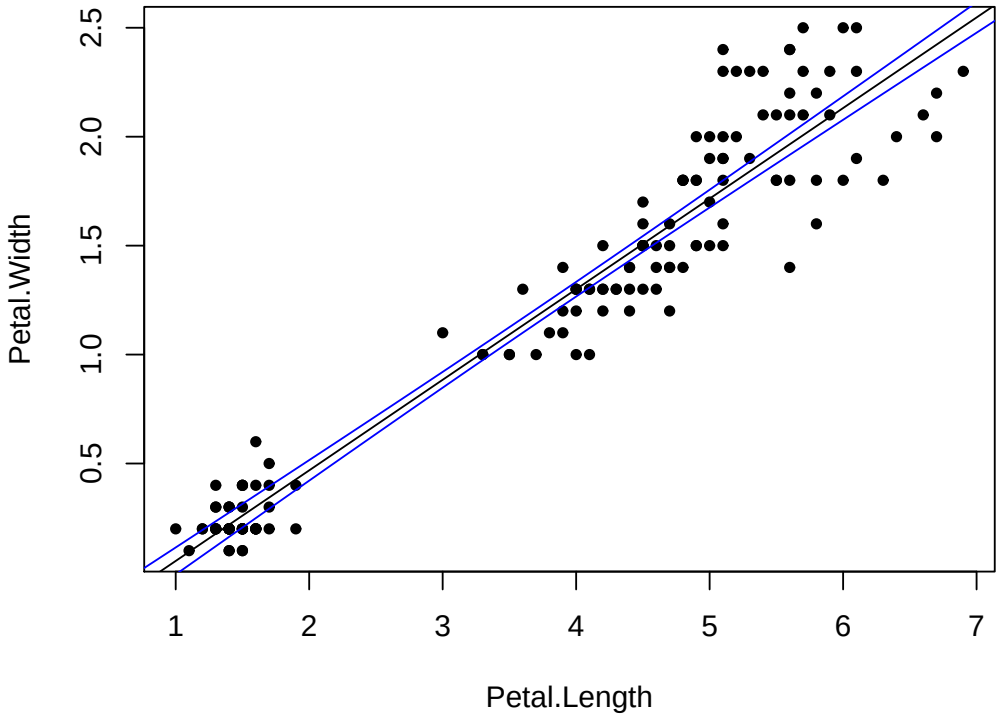
Um intervalo de confiança obtém-se com o argumento `int="conf"`:

```
> 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. No exemplo, e com 95% de confiança, a recta populacional está contida nas seguintes 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 ao valor esperado de Y, para um dado valor de x. Mas uma observação individual de Y tem associada uma variabilidade adicional. De facto,

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

Um predictor dessa observação de Y é dado por:

$$Y_{indiv} \quad = \quad \hat{\beta}_0 + \hat{\beta}_1 x + \varepsilon \quad = \quad \hat{\mu}_{Y|x} + \varepsilon.$$

Como a variabilidade do estimador de $\hat{\mu}_{Y|x}$ é (acetato 155): $V[\hat{\mu}_{Y|x}] = \sigma^2 \left[\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)\,s_x^2}\right]$, e a da flutuação aleatória em torno da recta é $V[\varepsilon] = \sigma^2$, a variância de uma observação individual é:

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

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

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

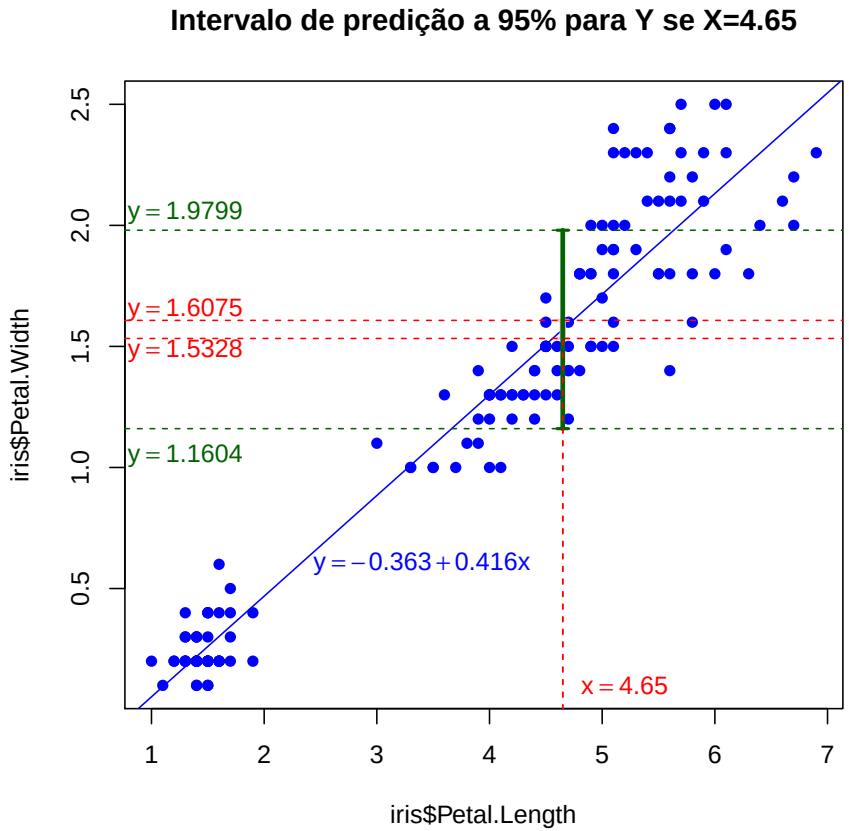
com $\hat{\mu}_{Y|x} = b_0 + b_1 x$ e $\hat{\sigma}_{indiv} = \sqrt{QMRE \left[1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)\, s_x^2}\right]}$.

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 R

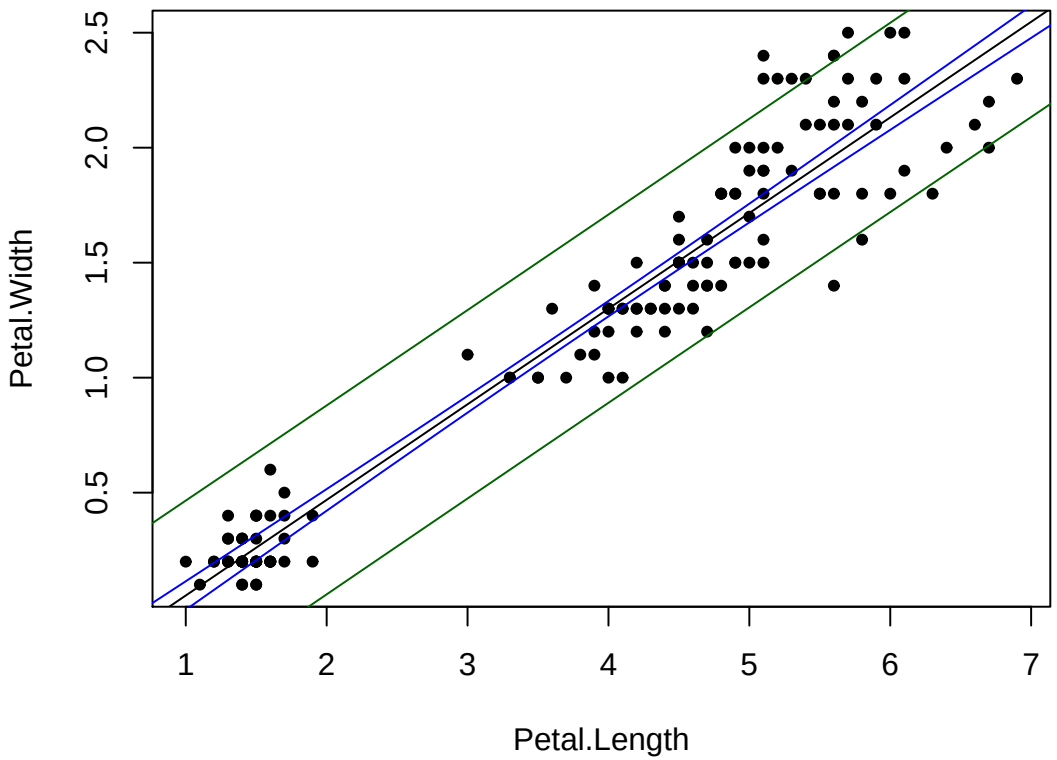
No R, um intervalo de predição para uma observação individual de Y obtém-se através da opção `int="pred"` no comando `predict`:

```
> 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. 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.

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: $Y = \beta_0 + \varepsilon$.

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 146.
- Efectuar o teste F ao ajustamento global do modelo. Este teste é descrito seguidamente.

Apenas a 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 5d das práticas).
- No acetato 131 viu-se que: $\dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{\sigma^2}{(n-1)s_x^2}}} \cap \mathcal{N}(0,1)$.
- Logo, $\dfrac{(\hat{\beta}_1 - \beta_1)^2}{\sigma^2 / \left[(n-1)s_x^2\right]} \cap \chi_1^2$. [Recordar: $Z \cap \mathcal{N}(0,1) \Rightarrow Z^2 \cap \chi_1^2$].
- Se $\beta_1 = 0$, tem-se: $\dfrac{SQR}{\sigma^2} \cap \chi_1^2$.

A quantidade SQR/σ^2 cuja distribuição agora se conhece depende da incógnita σ^2. Mas temos forma de tornear o problema.

SQR e *SQRE*

- Sabemos (acetato 134) que $SQRE/\sigma^2 \cap \chi^2_{n-2}$.
- Sabemos (da disciplina de Estatística) que as distribuições F surgem da seguinte forma:

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

- É possível mostrar que *SQRE* e *SQR* são v.a. independentes.
- Logo, se $\beta_1 = 0$, tem-se, definindo $QMR = SQR/1$ e $QMRE = SQRE/(n-2)$ (onde QM $\equiv$ Quadrados Médios):

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

Como usar a estatística F

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

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

sendo $QMR = SQR/1$ e $QMRE = SQRE/(n-2)$.

E se $\beta_1 \neq 0$?

Quanto maior fôr $\hat{\beta}_1^2$, mais duvidoso será que $\beta_1 = 0$ e, 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 que $\beta_1 \neq 0$.

O Teste F de ajustamento global do Modelo

Sendo válido o Modelo de Regressão Linear Simples, pode efectuar-se o seguinte

Teste F de ajustamento global do modelo

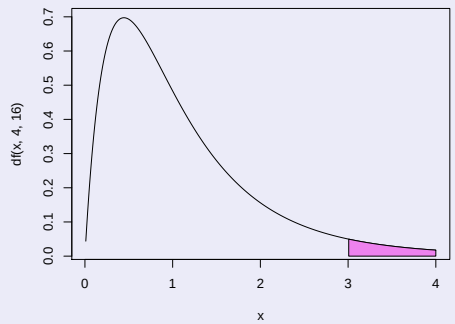
Hipóteses: $H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0.$

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

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)

Pode-se re-escrever as hipóteses e estatística do teste usando Coeficientes de Determinação (ver Exercício 15 de RLS):

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: $F = (n-2)\frac{R^2}{1-R^2} \cap F_{(1,n-2)} \quad \text{se } H_0.$

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

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

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.

```
> 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:

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

A 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 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 21).

A distribuição dos Resíduos no Modelo RLS

Teorema (Distribuição dos Resíduos no Modelo RLS)

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

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

Recordar: O modelo RLS admite que $\varepsilon_i \cap \mathcal{N}(0,\sigma^2)$.
Note que os resíduos E_i têm variâncias diferentes: $V[E_i] = \sigma^2(1-h_{ii})$.

Um resíduo também é uma combinação linear dos Y_i:

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

com

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

Diferentes tipos de resíduos

Definem-se três variantes de resíduos:

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

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

Resíduos Studentizados (ou externamente estandardizados):

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

sendo $QMRE_{[-i]}$ o valor de $QMRE$ resultante de um ajustamento da Regressão excluíndo a i-ésima observação (associada ao resíduo E_i).

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

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`

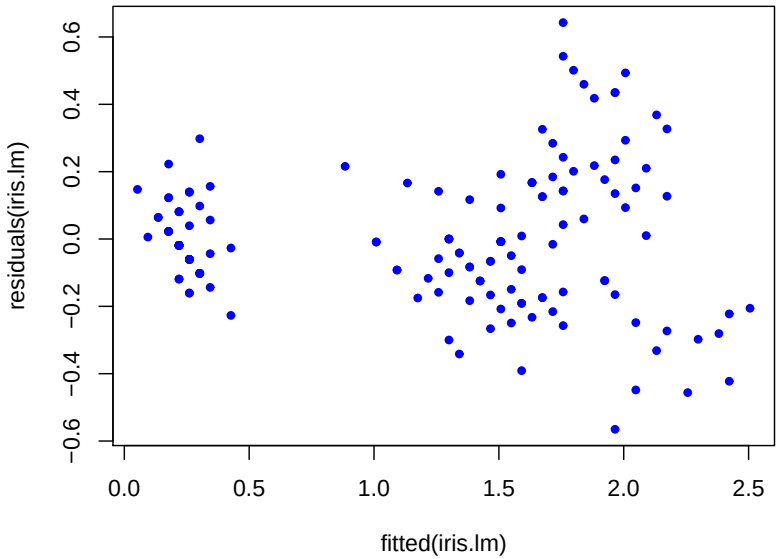
Não se efectuam testes de Normalidade aos resíduos usuais, uma vez que os resíduos não são independentes, como se pode verificar a partir do facto de que somam zero (ver Exercício RLS 5).

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 indispensável é o de Resíduos (usuais) vs. Valores ajustados de Y. No exemplo dos lírios:

```
> 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 21).

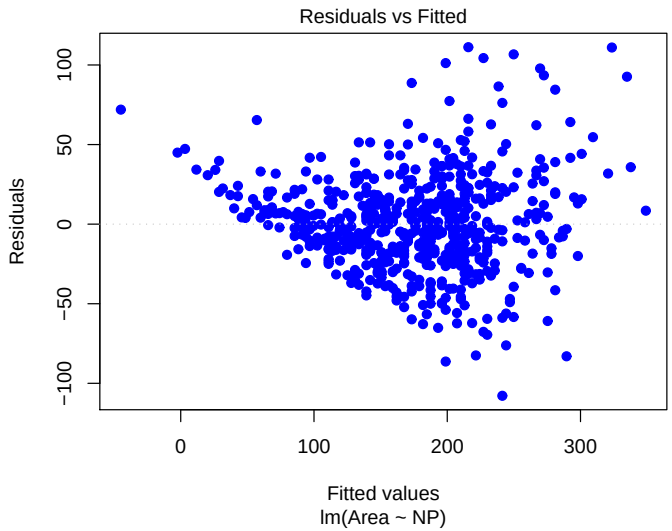
Possíveis padrões indicativos de problemas

Num gráfico de E_i vs. $\hat{Y}_i$ surgem com frequência alguns 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 ou mais resíduos muito destacados, ou banda oblíqua Indica possíveis observações atípicas.



Um exemplo de resíduos em forma de funil, e sugerindo alguma curvatura na relação entre as duas variáveis.

Gráficos para estudar a hipótese de normalidade

Como foi visto no acetato 175, dado o Modelo, $\frac{E_i}{\sqrt{\sigma^2 (1-h_{ii})}} \cap \mathcal{N}(0,1)$.

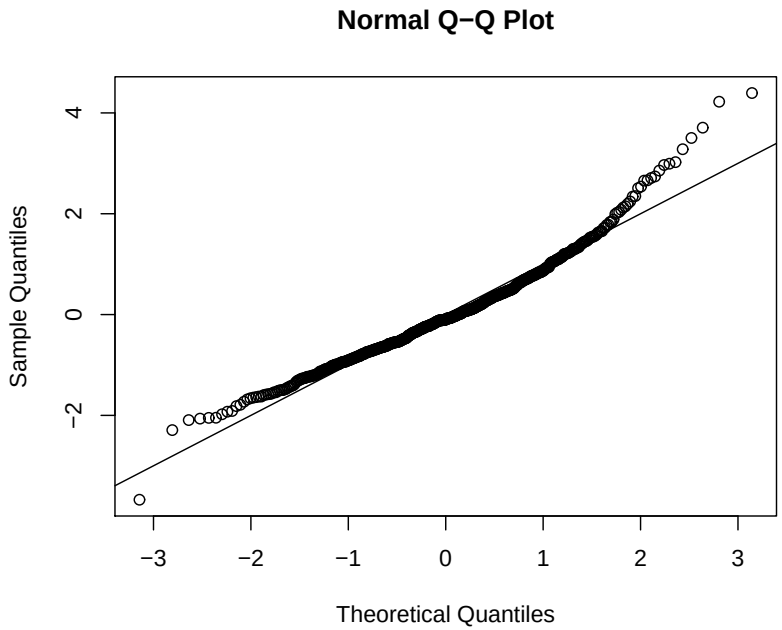
Embora os resíduos estandardizados, $R_i = \frac{E_i}{\sqrt{QMRE\,(1-h_{ii})}}$ não sejam exactamente $\mathcal{N}(0,1)$, desvios importantes à Normalidade devem fazer duvidar da validade do pressuposto de erros aleatórios Normais.

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.)

Um qq-plot indicativo de concordância com a hipótese de Normalidade dos erros aleatórios deverá ter os pontos aproximadamente em cima de uma recta. O exemplo seguinte sugere algum desvio a essa hipótese para os resíduos mais extremos.



Foi criado pelos comandos

```
> qqnorm(rstandard(lm(Area ~ NLdir, data=clopes)))
> abline(0,1)
```

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.

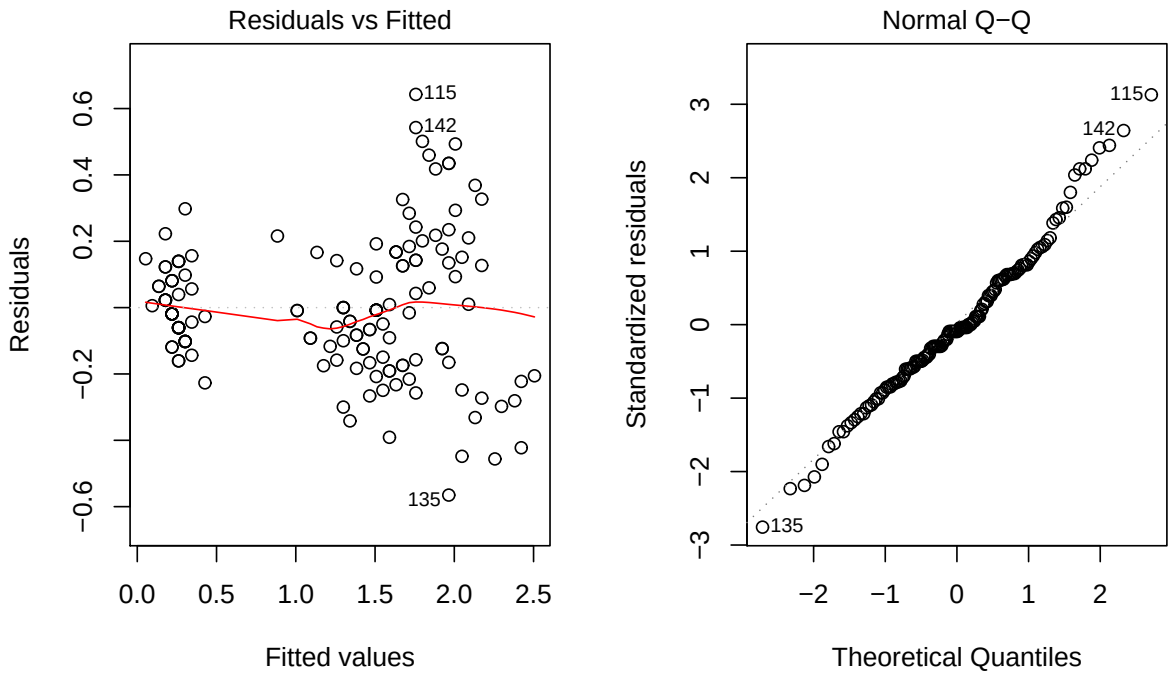
Em casos onde se suspeite de correlação no tempo, ou no espaço, será útil inspeccionar um gráfico de resíduos vs. ordem de observação ou posição no espaço, para verificar se existem padrões que sugiram falta de independência.

Estudo de resíduos no R

O comando `plot`, aplicado a um objecto que resulte de aplicar a função `lm` pode produzir seis gráficos, correspondendo os dois primeiros aos que foram vistos nos acetatos anteriores.

Para o exemplo dos lírios:

```
> 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 da relação linear de fundo entre Y e a variável preditora.

Muitas vezes surgem associadas a resíduos grandes (em módulo). Em particular, e 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 por vezes, observações distantes da tendência geral podem afectar o próprio ajustamento do modelo, e não serem facilmente identificáveis a partir dos seus resíduos.

As chamadas "observações alavanca"

Observações alavanca (*leverage points* em inglês) são observações que tendem a "atrair" a recta de regressão. Na RLS são observações para as quais é elevado o valor

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

também designado valor do efeito alavanca (*leverage*, em inglês).

Assim, numa RLS, quanto mais afastado estiver o valor x_i da média $\overline{x}$, maior será o efeito alavanca.

O papel de h_{ii} resulta da sua presença na expressão da variância do i-ésimo resíduo E_i (ver acetato 175): $V[E_i] = \sigma^2 (1 - h_{ii})$.
Se h_{ii} é elevado, a variância do resíduo E_i é baixa, logo o resíduo tende a estar próximo do seu valor médio (zero), ou seja, a recta de regressão tende a passar próximo desse ponto.

Observações alavanca (cont.)

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 um 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

Observações influentes são observações que, se retiradas da análise, geram variações assinaláveis no conjunto dos valores ajustados de Y e nos parâmetros estimados, b_0 e b_1. Medida frequente para a influência da observação i é a distância de Cook, que na RLS é:

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

sendo $\vec{\hat{\mathbf{y}}}$ o vector dos valores ajustados $\hat{y}_i$ usuais e $\vec{\hat{\mathbf{y}}}_{(-i)}$ o vector dos n valores ajustados de Y obtidos estimando os βs 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

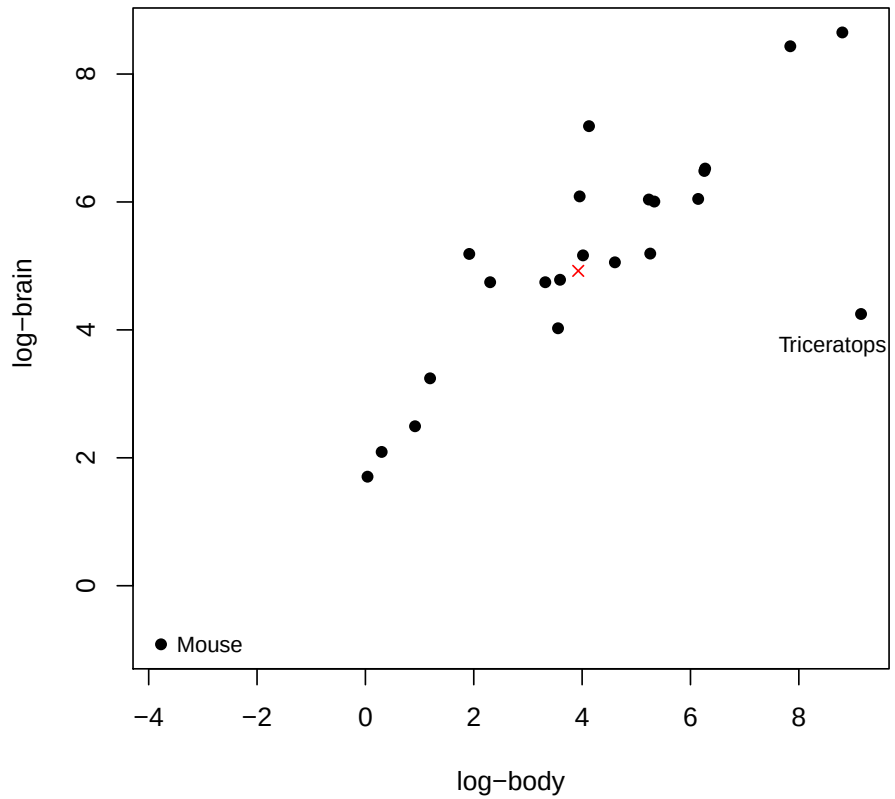
Observações atípicas, influentes ou alavanca, embora podendo estar relacionadas, não são o mesmo conceito.

Por exemplo, uma observação com resíduo (internamente) estandardizado grande e h_{ii} elevado, tem de ter uma distância de Cook grande, logo ser influente. Se tiver R_i^2 grande e h_{ii} pequeno (ou viceversa), pode, ou não, ser influente, consoante a grandeza relativa desses dois valores.

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

Um exemplo

Considerando apenas um subconjunto das espécies animais estudadas no Exercício 9 de Regressão Linear Simples, 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.

Um 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 

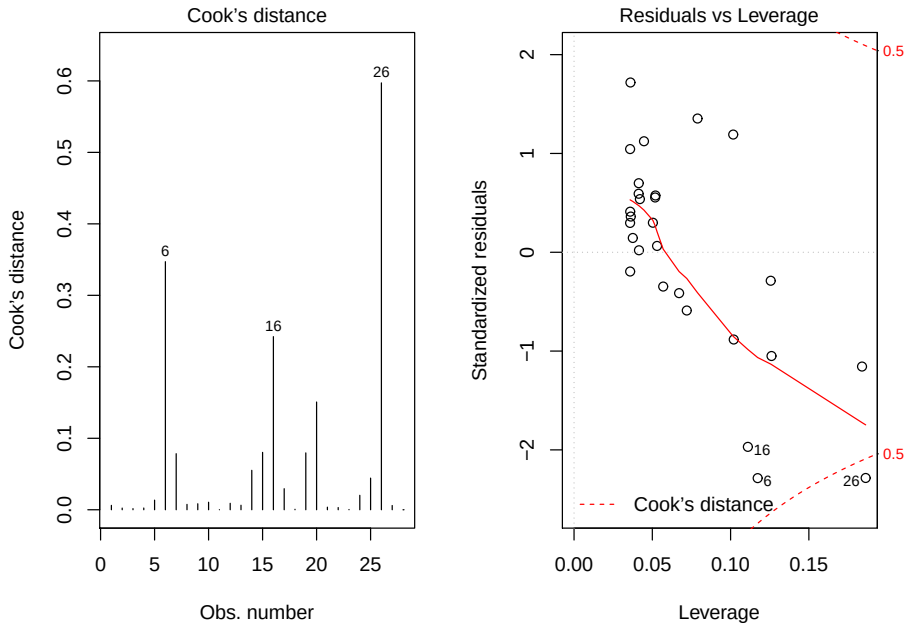
A função `plot`, aplicada a um objecto `lm` produz, além dos gráficos vistos no acetato 183, gráficos com alguns dos diagnósticos agora considerados.

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.

Um exemplo de gráficos de diagnóstico

Um exemplo destes gráficos de diagnósticos, para (a totalidade) dos dados do Exercício RLS 9 (`Animals`) é:



Os valores elevados de distância de Cook reflectem o distanciamento das espécies de dinossáurios da tendência geral das outras espécies, embora o facto de serem três observações discordantes mitiga um pouco o valor destes diagnósticos.

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,

$$\begin{array}{lllll} \text{Se } var(\varepsilon_i) & \propto & E[Y_i] & \text{então} & Y \longrightarrow \sqrt{Y} \\ \text{Se } var(\varepsilon_i) & \propto & (E[Y_i])^2 & \text{então} & Y \longrightarrow \ln Y \\ \text{Se } var(\varepsilon_i) & \propto & (E[Y_i])^4 & \text{então} & Y \longrightarrow 1/Y \end{array}$$

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.