

Capítulo 2

Modelo Linear

Modelação Estatística

Objectivo (informal): Descrever a **relação de fundo** entre

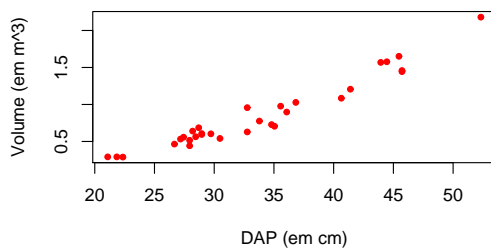
- uma **variável resposta** (ou **dependente**) y ; e
- uma ou mais **variáveis explicativas** (**variáveis predictoras** ou **independentes**), x_1, x_2, \dots, x_p .

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

Exemplo 1

$n = 31$ pares de medições:

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



A **tendência de fundo** é aproximadamente **linear**

Exemplo 1 (cont.)

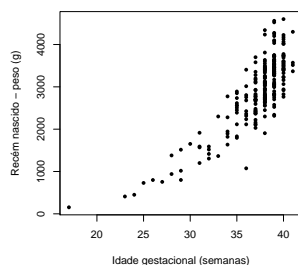
Mas,

- qual a “melhor” equação de recta, $y = b_0 + b_1 x$, para as n observações (e qual o critério de “melhor”)?
- e se os $n = 31$ pares de observações são apenas uma **amostra aleatória** duma população mais vasta, o que se pode dizer da **recta populacional** $y = \beta_0 + \beta_1 x$?

Exemplo 2 - Uma relação não linear

$n = 251$ pares de observações

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



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

Exemplo 2 (cont.)

Neste caso, há uma nova pergunta:

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

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**?

Duas ideias prévias sobre modelação

- Todos os modelos são apenas **aproximações** da realidade. Uns são melhores que outros.
- 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.

Modelo Linear

- um **caso particular** de modelação estatística;
- **engloba um grande número de modelos específicos**: Regressão Linear (Simple 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 generalizações** (Regressão não linear, Modelos Lineares Generalizados, etc.).

Revisão: Reg. Linear Simple - contexto descritivo

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

- apenas como regressão linear **simple**s
- apenas no contexto **descritivo**

Relação linear de fundo? \implies **Recta de Regressão** $y = b_0 + b_1 x$.

$$b_1 = \frac{\text{cov}_{xy}}{s_x^2}$$

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

com

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

Revisão: Reg. Linear Simple descritiva (cont.)

Critério: Minimizar a soma de quadrados dos resíduos

Resíduos:

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

onde $\hat{y}_i = b_0 + b_1 x_i$ são os "y ajustados pela recta"

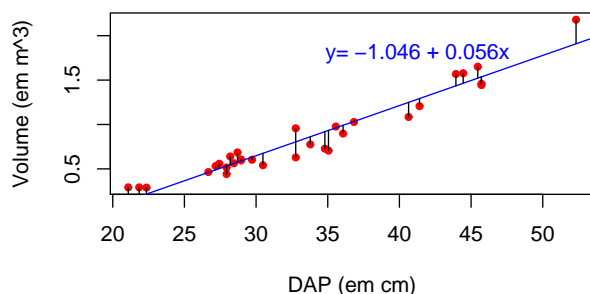
Soma de Quadrados dos Resíduos:

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

Regressão Linear Simple - contexto descritivo

$n = 31$ pares de medições

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



Regressão Linear Simple - contexto descritivo

Crítérios de ajustamento diferentes dão 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 desvios)?
- 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.

Regressão Linear Simples - contexto descritivo

Algumas quantidades importantes na RLS descritiva (ver Exercícios das aulas práticas):

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

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

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

Fórmula Fundamental:

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

Coefficiente de Determinação:

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

Regressão - um pouco de história

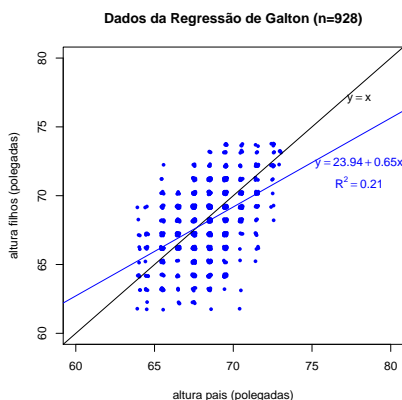
A designação **Regressão** tem origem num estudo de Francis Galton (1886), relacionando a altura de $n = 928$ jovens adultos com a altura (média) dos pais.

Galton constatou que pais com alturas acima da média tinham tendência a ter filhos com altura acima da média - mas menos que os pais (idem 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

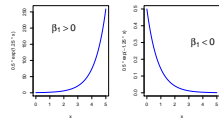
Nalguns casos, a relação de fundo entre x e y é não-linear, mas pode ser linearizada caso se proceda a transformações numa ou em ambas as variáveis.

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.

Relação exponencial

Relação exponencial : $Y = \beta_0 e^{\beta_1 x}$
 ($y > 0$; $\beta_0 > 0$)



Transformação : Logaritimizando, obtém-se:

$$\ln(Y) = \ln(\beta_0) + \beta_1 x$$

$$\Leftrightarrow Y^* = \beta_0^* + \beta_1 x$$

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

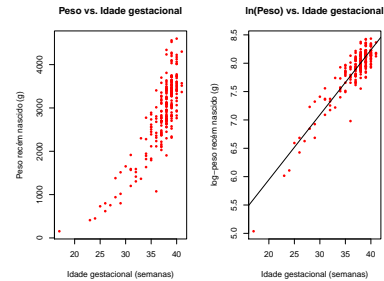
Uma relação exponencial resulta de admitir que y é função de x e que a **taxa de variação relativa de y é constante**:

$$\frac{y'(x)}{y(x)} = \beta_1,$$

i.e., a taxa de variação de y é proporcional a y : $y'(x) = \beta_1 \cdot y(x)$.

Uma linearização no Exemplo 2

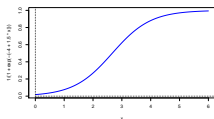
O gráfico de **log-pesos** dos recém-nascidos contra idade gestacional produz uma relação de fundo linear:



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

Relação Logística

Relação Logística : $Y = \frac{1}{1 + \exp^{-(\beta_0 + \beta_1 x)}}$



Transformação : Admitindo que $y \in]0, 1[$, tem-se uma relação linear entre a **função logit** de Y , $\ln(Y/(1 - Y))$, e x :

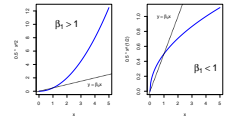
$$\ln\left(\frac{Y}{1 - Y}\right) = \beta_0 + \beta_1 x \quad \Leftrightarrow \quad Y^* = \beta_0 + \beta_1 x.$$

Resulta de admitir que y é função de x e que a **taxa de variação relativa de y diminui com o aumento de y** :

$$\frac{y'(x)}{y(x)} = \beta_1 \cdot [1 - y(x)].$$

Relação potência ou alométrica

Relação potência : $Y = \beta_0 x^{\beta_1}$
 ($x, y > 0$; $\beta_0 > 0$)



Transformação : Logaritimizando, obtém-se:

$$\ln(Y) = \ln(\beta_0) + \beta_1 \ln(x)$$

$$\Leftrightarrow Y^* = \beta_0^* + \beta_1 x^*$$

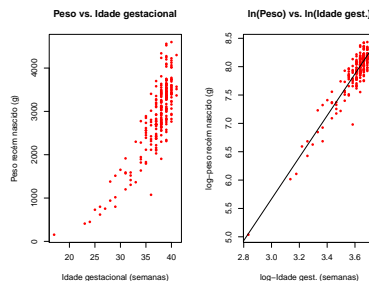
que é uma **relação linear** entre $Y^* = \ln(Y)$ e $x^* = \ln(x)$.

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

$$\frac{y'(t)}{y(t)} = \beta_1 \cdot \frac{x'(t)}{x(t)}.$$

Outra linearização no Exemplo 2

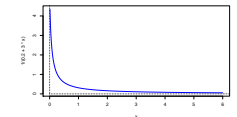
O gráfico de **log-pesos** dos recém-nascidos contra **log-idade gestacional** produz outra relação de fundo linear:



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

Relação hiperbólica ou de proporcionalidade inversa

Relação hiperbólica : $Y = \frac{1}{\beta_0 + \beta_1 x}$
 ($x, y > 0$; $\beta_0 > 0$)



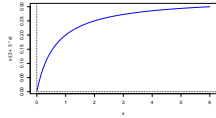
Transformação : Obtém-se uma **relação linear** entre $Y^* = 1/Y$ e x :

$$\frac{1}{Y} = \beta_0 + \beta_1 x \quad \Leftrightarrow \quad Y^* = \beta_0 + \beta_1 x.$$

Usado na modelação de **rendimento por planta (y) vs. densidade da cultura ou povoamento (x)**.

Relação Michaelis-Menten

Relação Michaelis-Menten : $Y = \frac{x}{c+dx}$



Transformação : Tomando recíprocos, obtém-se uma **relação linear** entre $Y^* = \frac{1}{Y}$ e $x^* = \frac{1}{x}$, com $\beta_0^* = d$ e $\beta_1^* = c$:

$$\frac{1}{Y} = \frac{c}{x} + d \Leftrightarrow Y^* = \beta_0^* + \beta_1^* x^*$$

- Em **modelos 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.

Advertência sobre transformações linearizantes

A regressão **linear** simples **não** modela directamente relações **não lineares** entre x e y .

RLS pode modelar relações lineares que se formem **após transformações linearizantes**, ou seja, **a relação linear entre as variáveis transformadas**.

Atenção: Linearizar, obter os parâmetros b_0 e b_1 da recta e depois desfazer a relação não linear **não** produz os mesmos valores dos parâmetros que resultariam de minimizar a soma de quadrados dos resíduos directamente na relação não linear.

Regressão Linear Simples - Inferência

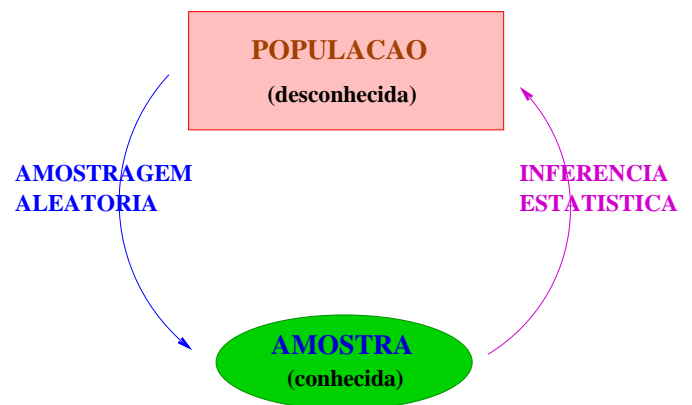
- Até aqui a RLS foi usada apenas como **técnica descritiva**. Se as n observações fossem a totalidade da população de interesse, pouco mais haveria a dizer. Mas, com frequência, as n observações são apenas uma **amostra aleatória** de uma população maior.
- A recta de regressão $y = b_0 + b_1 x$ obtida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



MODELO - Regressão Linear Simples

Admitimos **pressupostos adicionais** para permitir a inferência.

Y – variável resposta **aleatória**.

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

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

MODELO RLS – Linearidade

Vamos ainda admitir que a **relação de fundo entre x e Y é linear**, com uma variabilidade aleatória em torno dessa relação, representada por um **erro aleatório** ε . Para todo o $i = 1, \dots, n$:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

\downarrow \downarrow \downarrow \downarrow \downarrow
 v.a. cte. cte. cte. v.a.

MODELO RLS – Os erros aleatórios

Vamos ainda admitir que os erros aleatórios ε_i :

- Têm valor esperado nulo:

$$E[\varepsilon_i] = 0, \quad \forall i = 1, \dots, n$$

(não é hipótese restritiva).

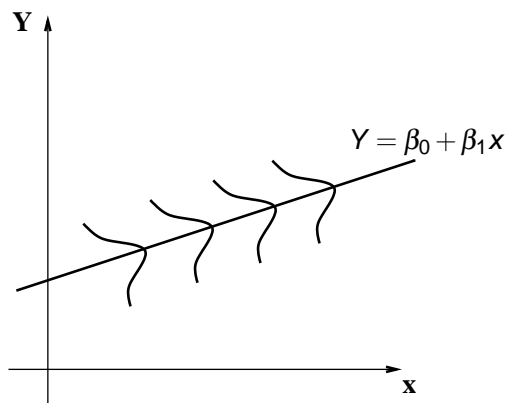
- Têm distribuição Normal (é restritiva, mas bastante geral).
- Homogeneidade de variâncias: têm sempre a mesma variância:

$$V[\varepsilon_i] = \sigma^2, \quad \forall i = 1, \dots, n$$

(é restritiva, mas conveniente).

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

MODELO Regressão Linear Simples



MODELO - Regressão Linear Simples

Definição (O Modelo da Regressão Linear Simples)

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

NOTA: Os erros aleatórios são i.i.d. (independentes e identicamente distribuídos)

Primeiras consequências do MODELO RLS

Teorema (Primeiras consequências do Modelo)

Dado o Modelo da Regressão Linear Simples, tem-se

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

Ver apontamentos da Prof. Manuela Neves (Teoria das Probabilidades, p.95 e seguintes) para propriedades da Normal: <http://www.isa.utl.pt/dm/estat/estat/seb2.pdf>

- **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

O Modelo de Regressão Linear Simples tem dois parâmetros: β_0 e β_1 . Definem-se **estimadores** desses parâmetros a partir das expressões obtidas para b_0 e b_1 pelo Método dos Mínimos Quadrados.

Definição (Estimador de β_1)

$$\hat{\beta}_1 = \frac{\text{cov}_{XY}}{s_x^2} \stackrel{(*)}{=} \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{(n-1) \cdot s_x^2} = \sum_{i=1}^n c_i Y_i, \quad (1)$$

com

$$c_i = \frac{(x_i - \bar{x})}{(n-1) \cdot s_x^2}.$$

(*) Exercício 3b) das aulas práticas.

Estimação dos parâmetros do Modelo RLS

Definição (Estimador de β_0)

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} = \frac{1}{n} \sum_{i=1}^n Y_i - \bar{x} \sum_{i=1}^n c_i Y_i = \sum_{i=1}^n d_i Y_i, \quad (2)$$

com

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

Distribuição dos estimadores RLS

Teorema (Distribuição dos estimadores dos parâmetros)

Dado o Modelo de Regressão Linear Simples,

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

NOTAS:

- 1 Ambos os estimadores são **centrados**: $E[\hat{\beta}_1] = \beta_1$ e $E[\hat{\beta}_0] = \beta_0$
- 2 Quanto maior $(n-1) \cdot S_x^2$, 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 \bar{x} de zero.

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

NOTA: O desvio padrão dum estimador designa-se **erro padrão** (em inglês, *standard error*).

Distribuição dos estimadores RLS

Os resultados do Corolário anterior quase que permitem fazer inferência sobre os parâmetros β_0 e β_1 (e.g., construir intervalos de confiança ou efectuar testes de hipótese). Mas subsiste um problema importante: para além dos próprios parâmetros β_0 e β_1 , há outra **quantidade desconhecida** nos resultados: a variabilidade dos erros aleatórios, $\sigma^2 = V[\varepsilon_i]$.

Precisamos de um estimador da variância σ^2 dos erros aleatórios.

Vamos buscá-lo aos **resíduos**.

Erros aleatórios e Resíduos

$$\begin{aligned} \text{Erros aleatórios} & \varepsilon_i = Y_i - (\beta_0 + \beta_1 x_i). \\ \text{Resíduos} & E_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i). \end{aligned}$$

Os resíduos são os preditores (conhecidos) dos erros (desconhecidos).

O numerador da variância dos resíduos é

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

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

A Soma de Quadrados Residual

Teorema (Resultados distribucionais de SQRE)

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

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

NOTA: Omite-se a demonstração

Corolário

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

Ver apontamentos da Prof. Manuela Neves (Teoria das Probabilidades, p.103 e seguintes) para propriedades da χ^2 :
<http://www.isa.utl.pt/dm/estat/estat/seb2.pdf>

O Quadrado Médio Residual

Definição (Quadrado Médio Residual)

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

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

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

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

- Como se viu no acetato anterior, QMRE é um **estimador centrado**.

Revisão: como surge uma t – Student

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

$$\left. \begin{array}{l} Z \cap \mathcal{N}(0,1) \\ W \cap \chi_v^2 \\ Z, W \text{ v.a. independentes} \end{array} \right\} \Rightarrow \frac{Z}{\sqrt{W/v}} \cap t_v.$$

Ver apontamentos da Prof. Manuela Neves (Introdução à Inferência Estatística, Def. 3.3, p.115):
<http://www.isa.utl.pt/dm/estat/estat/seb3.pdf>.

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

$$\begin{array}{l} \textcircled{1} \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \cap t_{n-2}, \quad \text{com } \hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{QMRE}{(n-1) \cdot S_x^2}}. \\ \textcircled{2} \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \cap t_{n-2}, \quad \text{com } \hat{\sigma}_{\hat{\beta}_0} = \sqrt{QMRE \left[\frac{1}{n} + \frac{\bar{x}^2}{(n-1) \cdot S_x^2} \right]} \end{array}$$

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 .

Intervalo de confiança para β_1

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

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

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

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 SQRE e diminui com n e S_x^2 :

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

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:

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

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

NOTA: A amplitude do IC aumenta com SQRE e com \bar{x}^2 e diminui com n e S_x^2 :

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

Um alerta sobre Intervalos de Confiança

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

- o grau de confiança em como intervalos deste tipo 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, menor será a sua precisão, isto é, maior será a sua amplitude.

Testes de hipóteses para o declive β_1

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

Teste de Hipóteses a β_1 (Bilateral)

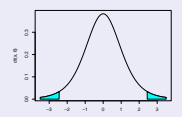
Hipóteses: $H_0 : \beta_1 = c$ vs. $H_1 : \beta_1 \neq c$.

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

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

Região Crítica (Região de Rejeição): Bilateral

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



Testes de hipóteses para o declive β_1

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

Teste de Hipóteses a β_1 (Unilateral direito)

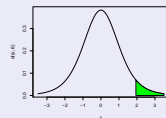
Hipóteses: $H_0 : \beta_1 \leq c$ vs. $H_1 : \beta_1 > c$.

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

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

Teste de Hipóteses a β_1 (Unilateral esquerdo)

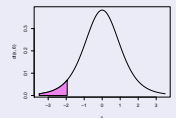
Hipóteses: $H_0 : \beta_1 \geq c$ vs. $H_1 : \beta_1 < c$.

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

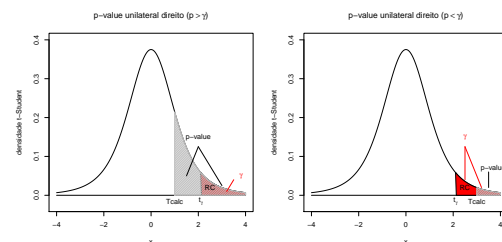
prob. de T tomar valores mais extremos que T_{calc} , sob H_0

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

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 ;



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: $H_0 : \beta_0 = c$ vs. $H_1 : \beta_0 \neq c$

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

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Rejeitar H_0 se
 $T_{calc} < -t_{\alpha(n-2)}$ (Unilateral esquerdo)
 $|T_{calc}| > t_{\alpha/2(n-2)}$ (Bilateral)
 $T_{calc} > t_{\alpha(n-2)}$ (Unilateral direito)

Testes de hipóteses no \mathbb{R}

No software \mathbb{R} , a Regressão Linear Simples de uma variável y sobre uma variável x invoca-se através do comando `lm`:

```
> lm(y ~ x)
```

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, $\hat{\sigma}_{\beta_0}$ e $\hat{\sigma}_{\beta_1}$

t value O valor calculado das estatísticas dos testes às hipóteses $H_0 : \beta_0(\beta_1) = 0$ vs. $H_1 : \beta_0(\beta_1) \neq 0$, ou seja,

$$T_{calc} = b_0 / \hat{\sigma}_{\beta_0} \quad \text{e} \quad T_{calc} = b_1 / \hat{\sigma}_{\beta_1}$$

Pr(>|t|) O valor p (p -value) associado a essa estatística de teste.

Um exemplo

O objecto `iris` do \mathbb{R} contém um conjunto de dados relativos a 150 lírios, nos quais se mediram quatro variáveis numéricas (comprimento e largura de sépalas e pétalas). Para obter uma regressão linear das larguras das pétalas sobre o comprimento das pétalas, podem dar-se os comandos seguintes (sendo o resultado do último comando apenas mostrado parcialmente):

```
> attach(iris)
> iris.lm <- lm(Petal.Width ~ Petal.Length)
> 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 ***
```

Neste caso, devem rejeitar-se as hipóteses $H_0 : \beta_0 = 0$ e $H_0 : \beta_1 = 0$.

Intervalos de confiança de β_0 e β_1 no \mathbb{R}

Para obter intervalos de confiança dos parâmetros da recta no \mathbb{R} , utiliza-se a função `confint` sobre o resultado de um comando `lm`. Por omissão, o IC calculado é a 95% de confiança, mas esse nível pode ser controlado através do argumento `level`:

```
> confint(iris.lm)
                2.5 %      97.5 %
(Intercept) -0.4416501 -0.2845010
Petal.Length  0.3968193  0.4346915

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

Inferência sobre $E[Y|x]$

Problema de interesse geral: fazer inferência sobre o valor esperado da variável aleatória Y , dado um valor da variável preditora x :

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

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 (91) e (92).

A distribuição do estimador de $E[Y|x]$

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

Dado o Modelo de Regressão Linear Simples, tem-se

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

onde $\sigma_{\hat{\mu}_{Y|x}} = \sqrt{\sigma^2 \cdot \left[\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1) \cdot S_x^2}\right]}$ 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 94), também esta distribuição não é ainda utilizável devido à presença da variância (desconhecida) dos erros aleatórios, σ^2 .

A distribuição utilizável do estimador de $E[Y|x]$

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

Dado o Modelo de Regressão Linear Simples, tem-se

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

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

NOTA: A justificação desta distribuição é totalmente análoga à das distribuições de $\hat{\beta}_1$ e $\hat{\beta}_0$ dadas no acetato (100).

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

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

Teorema (IC para $\mu_{Y|x} = \beta_0 + \beta_1 x$)

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

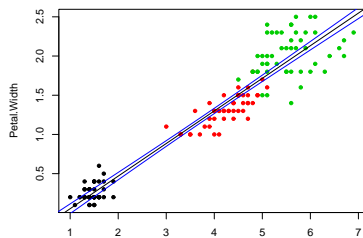
$$\left] \hat{\mu}_{Y|x} - t_{\alpha/2(n-2)} \cdot \hat{\sigma}_{\hat{\mu}_{Y|x}}, \hat{\mu}_{Y|x} + t_{\alpha/2(n-2)} \cdot \hat{\sigma}_{\hat{\mu}_{Y|x}} \right[,$$

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

NOTA: A amplitude do IC aumenta com SQRE e com a distância de x a \bar{x} e diminui com n e $(n-1) \cdot S_x^2$.

Bandas de confiança para a recta de regressão

Os IC para $\mu_{Y|x}$ dependem do valor de x . Terão maior amplitude quanto mais afastado x estiver da média \bar{x} das observações. Considerando os ICs para todos os valores de x nalgum intervalo, obtém-se uma **banda de confiança** em torno da recta de regressão. Exemplo: Dados dos lírios no \mathbb{R} , com `lm(Petal.Width ~ Petal.Length)`:



A variabilidade de uma 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,

$$Y = E[Y|x] + \varepsilon = \beta_0 + \beta_1 x + \varepsilon,$$

pelo que se a variabilidade do estimador de $E[Y|x]$ é (acetato 113)

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

a variância de uma observação individual pode ser aproximada por

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

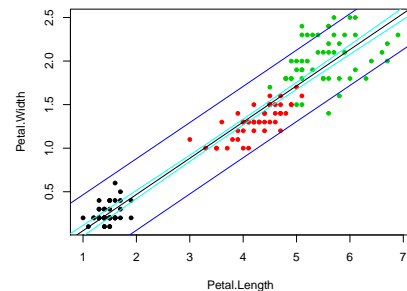
Intervalos de predição para uma observação de Y

Podem-se construir **intervalos de predição para uma observação individual de Y** , associada ao valor $X = x$, incrementando a variância em σ^2 , logo a variância estimada em **QMRE**, ou seja:

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

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

Intervalos de predição para uma observação de Y



Inferência sobre $E[Y|x]$ no \mathbb{R}

Valores estimados e intervalos de confiança para $\mu_{Y|x}$ obtêm-se no \mathbb{R} com a função `predict`. Os valores de x são dados numa `data frame`, com o mesmo nome para x . E.g. (acetato 111), a largura esperada de pétalas de comprimento 1.85 e 4.65, é obtido através de:

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

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 $E[Y|x]$ no \mathbb{R} (continuação)

O **intervalo de confiança** obtém-se através do argumento `int='conf'`:

```
> predict(iris.lm,data.frame(Petal.Length=c(1.85,4.65)),
          int="conf")
           fit      lwr      upr
1 0.406072 0.3569258 0.4552182
2 1.570187 1.5328338 1.6075405
```

Um **intervalo de predição** para uma observação individual de Y obtém-se através de `int='pred'`:

```
> predict(iris.lm,data.frame(Petal.Length=c(1.85,4.65)),
          int="pred")
           fit      lwr      upr
1 0.406072 -0.004915416 0.8170594
2 1.570187 1.160442632 1.9799317
```

Avaliando a qualidade do ajustamento do Modelo

Como avaliar a qualidade do ajustamento do Modelo?

- Em termos meramente descritivos, é frequente usar o **Coefficiente de Determinação**, $R^2 = \frac{SQR}{SQE}$.
- Num contexto inferencial, é usual **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 :

$$H_0: \beta^2 = 0,$$

onde β^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 a $Y = \beta_0 + \varepsilon$.

Podemos testar-se essa hipótese de duas maneiras:

- Testar $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$, usando o teste t de hipóteses a β_1 , considerado no acetato 104.
- Efectuar o **teste F ao ajustamento global do modelo**. O teste é descrito seguidamente.

A segunda abordagem generaliza-se na 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,

- No Exercício 5d (aulas práticas) prova-se: $SQR = \hat{\beta}_1^2 \cdot (n-1) \cdot S_x^2$.
- No acetato 94 vimos que $\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{(n-1)S_x^2}}} \sim \mathcal{N}(0,1)$.
- Logo, $\frac{(\hat{\beta}_1 - \beta_1)^2}{\sigma^2 / [(n-1)S_x^2]} \sim \chi_1^2$.
[Recordar: $Z \sim \mathcal{N}(0,1) \Rightarrow Z^2 \sim \chi_1^2$].
- Se $\beta_1 = 0$, tem-se: $\frac{SQR}{\sigma^2} \sim \chi_1^2$.

SQR e SQRE

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

- Sabemos (acetato 97) que $SQRE/\sigma^2 \sim \chi_{n-2}^2$.
- Sabemos (da disciplina de Estatística) que as distribuições F surgem da seguinte forma:

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

- É possível mostrar que $SQRE$ e SQR são v.a. independentes.
- Logo, se $\beta_1 = 0$, tem-se:

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

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

Como usar a estatística F

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

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

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

E se $\beta_1 \neq 0$? Para valores de β_1 longe de zero, é natural que o estimador $\hat{\beta}_1$ produza valores igualmente longe de zero.

Quanto maior for $\hat{\beta}_1^2$, maior será $SQR = \hat{\beta}_1^2 \cdot (n-1) \cdot S_x^2$, pelo que maior será a estatística $F = QMR/QMRE$.

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

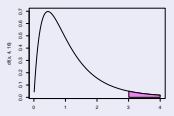
Hipóteses: $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$.

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

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

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



O Teste F de ajustamento global do Modelo (cont)

Pode-se (ver Exercício 14 das aulas práticas) re-escrever as hipóteses e estatística do teste usando **Coefficientes de Determinação**:

Teste F de ajustamento global do modelo

Hipóteses: $H_0 : R^2 = 0$ vs. $H_1 : R^2 > 0$.
 Estatística do Teste: $F = (n-2) \cdot \frac{R^2}{1-R^2} \cap F_{(1,n-2)}$ 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 .
- A hipótese nula $H_0 : R^2 = 0$ afirma que, na população, o coeficiente de correlação (ao quadrado) entre x e y é nulo.

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

p-value O 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 **Coefficiente de Determinação**:

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

Adjusted R-squared O R^2 modificado:

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

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 homogênea e independentes.

A validade dos ICs 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**.

A distribuição dos Resíduos num MRLS

Teorema (Distribuição dos Resíduos no MRLS)

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

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

Note que um resíduo se pode escrever como:

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

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

Note que **os resíduos E_i têm variâncias diferentes**.

Diferentes tipos de resíduos

A procura de um comportamento mais estável para os resíduos conduz à definição de três variantes de resíduos.

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

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

(Não é possível concluir que os R_i tenham distribuição t_{n-2} , uma vez que $QMRE$ e E_i não são independentes)

Resíduos Studentizados (ou externamente standardizados):

$$T_i = \frac{E_i}{\sqrt{QMRE_{[-i]} \cdot (1 - h_{ii})}} \cap t_{n-3}$$

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

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

Os resíduos no R

É hábito fazer representações gráficas dos (vários tipos) de resíduos para validar os pressupostos do Modelo de Regressão Linear.

Não se efectuam testes de Normalidade, uma vez que os resíduos não são independentes, como se pode verificar a partir do facto de que somam zero.

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

Resíduos usuais (E_i): `residuals`

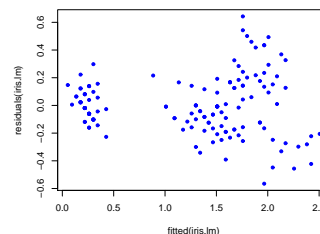
Resíduos estandardizados (R_i): `rstandard`

Resíduos Studentizados (T_i): `rstudent`

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))
```



Não deve existir qualquer padrão aparente. Sendo válido o MRLS, $cor(E_i, \hat{Y}_i) = 0$ (ver exercício 19). Resíduos devem estar aproximadamente numa banda horizontal em torno de zero.

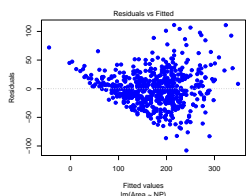
Possíveis padrões indicativos de problemas

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 Indica a possível existência de observações atípicas que podem estar a afectar o ajustamento.



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 (133), dado o MRLS, $\frac{E_i}{\sqrt{\sigma^2 \cdot (1-h_{ii})}} \cap \mathcal{N}(0, 1)$.

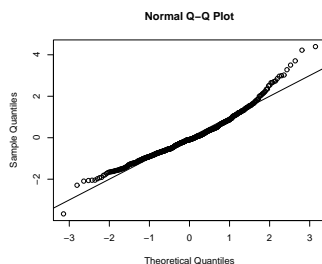
Embora os resíduos standardizados calculáveis, $R_i = \frac{E_i}{\sqrt{QMRE \cdot (1-h_{ii})}}$ não sejam exactamente $\mathcal{N}(0, 1)$, desvios importantes à Normalidade devem fazer duvidar da validade do pressuposto de erros aleatórios Normais. É hábito investigar a validade do pressuposto de erros 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** de n observações 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).

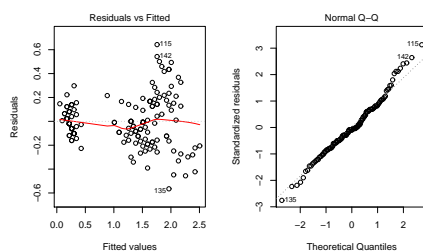
Nesse caso, pode ser útil inspeccionar um **gráfico de resíduos vs. ordem de observação**, para verificar se existem padrões que sugiram falta de independência.

Estudo de resíduos no R

O comando `plot`, aplicado a um objecto que resulte de aplicar a função `lm` pode produzir seis gráficos (quatro por omissão). Os dois primeiros correspondem aos que foram vistos nos acetatos anteriores. Por exemplo, com o exemplo dos lírios,

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

produz os gráficos



Algumas transformações de variáveis

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

Se $var(\varepsilon_i) \propto E[Y_i]$ então $Y \rightarrow \sqrt{Y}$

Se $var(\varepsilon_i) \propto (E[Y_i])^2$ então $Y \rightarrow \ln Y$

Se $var(\varepsilon_i) \propto (E[Y_i])^4$ então $Y \rightarrow 1/Y$

são propostas usuais para estabilizar as variâncias.

Os exemplos acima são casos particulares da **família Box-Cox de transformações**:

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

Prevenções sobre transformações

Mas a utilização de transformações de qualquer (ou ambas) as variáveis 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 heterogêneas) **pode gerar outro** (e.g., não-normalidade);
- Existe o perigo de usar transformações que resolvam o problema numa amostra específica, mas **não tenham qualquer generalidade**.

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:

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