Estatística e Delineamento Experimental 2025/2026

Modelo Linear

Elsa Gonçalves
Secção de Matemática, DCEB, ISA-Ulisboa

(Adaptado, Cadima J. (2021). O Modelo Linear, ISA, UlLisboa)

Regressão Linear – Abordagem Inferencial

Regressão Linear - Inferência

- Até aqui a regressão linear foi usada apenas como técnica descritiva. Se as n observações forem a totalidade da população de interesse, pouco mais há a dizer.

- Mas, com frequência, as n observações são apenas uma amostra aleatória de uma população maior.

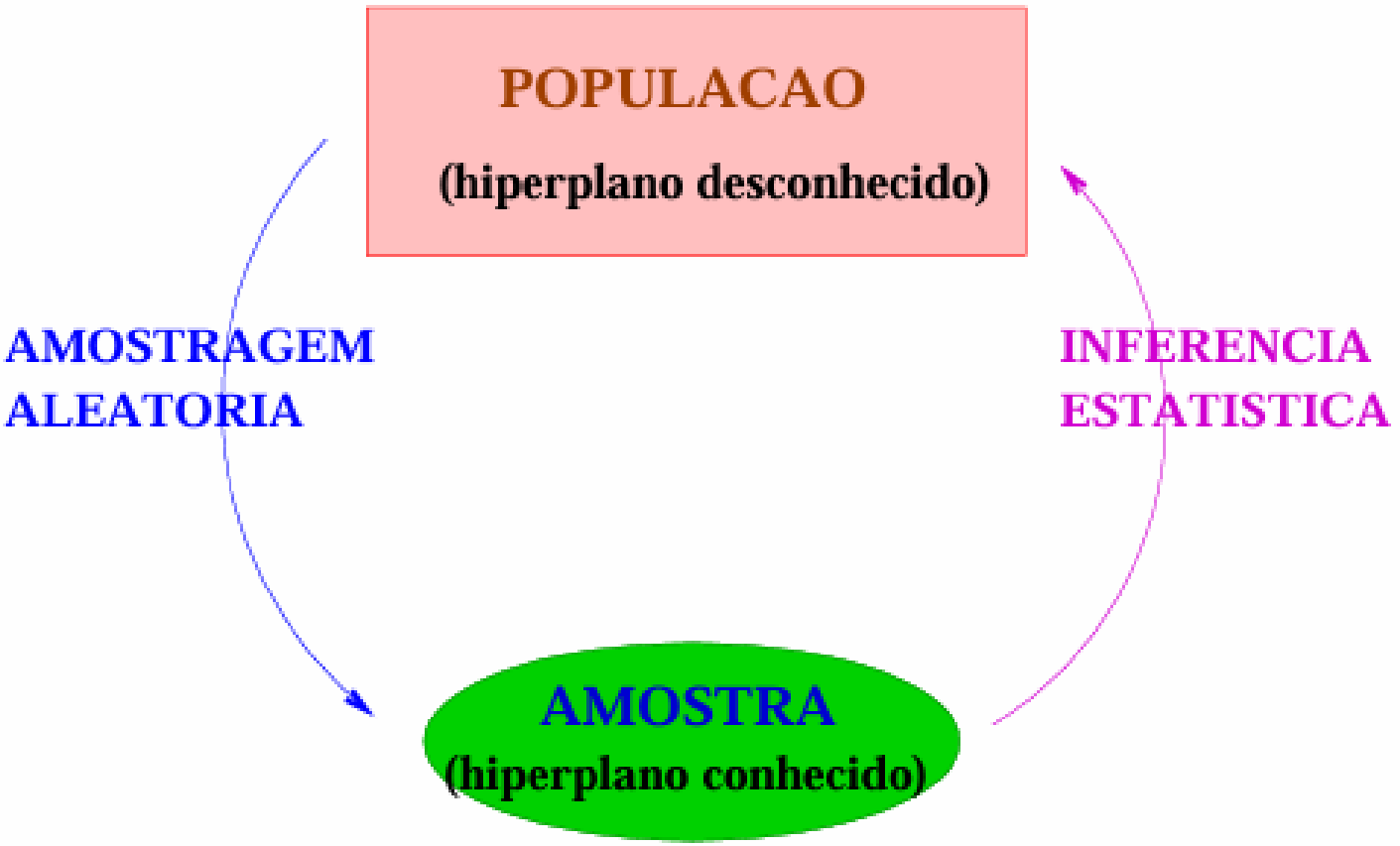
- Um hiperplano ajustado a partir duma dada amostra, $y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$, é apenas uma estimativa de um hiperplano populacional

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \ .$$

  Outras amostras dariam hiperplanos ajustados diferentes.

- Coloca-se o problema da inferência estatística.

O problema da Inferência Estatística na Reg. Linear



MODELO - Regressão Linear

A fim de se poder fazer inferência sobre o hiperplano populacional, vamos admitir pressupostos adicionais.

Y – variável resposta aleatória.

$x_1, ..., x_p$ – variáveis preditoras não aleatórias (fixadas pelo experimentador ou trabalha-se condicionalmente aos valores de $x_1, ..., x_p$)

O modelo será ajustado com base em:

$\{(x_{1(i)}, x_{2(i)}, ..., x_{p(i)}, Y_i)\}_{i=1}^{n}$ – n conjuntos de observações independentes das variáveis $x_1, x_2, ..., x_p$ e Y, sobre n unidades experimentais.

MODELO RL – Linearidade

Vamos ainda admitir que a relação de fundo entre Y e $x_1, x_2, \ldots, x_p$, é linear (afim), com uma variabilidade aleatória em torno dessa relação, representada por um erro aleatório ε. Para todo o $i = 1, \ldots, n$:

$$Y_i = \beta_0 + \beta_1 x_{1(i)} + \ldots + \beta_p x_{p(i)} + \varepsilon_i$$

Y_i	β_0	β_1	$x_{1(i)}$	β_p	$x_{p(i)}$	ε_i
v.a.	cte.	cte.	cte.	cte.	cte.	v.a.

MODELO Regressão Linear – 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).

O Modelo Linear

O modelo para inferência na regressão linear é assim:

O Modelo Linear

1. $Y_i = \beta_0 + \beta_1 x_{1(i)} + \beta_2 x_{2(i)} + \cdots + \beta_p x_{p(i)} + \varepsilon_i, \quad \forall\, i = 1, \ldots, n.$
2. $\varepsilon_i \frown \mathcal{N}(0, \sigma^2), \quad \forall\, i = 1, \ldots, n.$
3. $\{\varepsilon_i\}_{i=1}^n$ v.a. independentes.

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

Dado o modelo, o valor esperado (médio) de Y_i, condicional aos valores $x_1, x_2, \ldots, x_p$ dos preditores, é:

$$\mu_i = E[Y_i \,|\, x_1, x_2, \ldots, x_p] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \,.$$

NOTA: β_j ($j \neq 0$) é a variação média em Y, associada a um aumento de uma unidade em x_j, mantendo os restantes preditores constantes.

O estudo do modelo

Um primeiro objectivo da inferência: os $p+1$ parâmetros do modelo, β_j $(j = 0, 1, \ldots, p)$.

Os parâmetros ajustados $\vec{\mathbf{b}} = (b_0, b_1, b_2, \ldots, b_p)$, são estimativas desses parâmetros.

Para ser possível construir intervalos de confiança e/ou efectuar testes de hipóteses sobre os valores dos parâmetros populacionais β_j, há-que:

- Definir estimadores $\hat{\beta}_j$ dos parâmetros populacionais;
- conhecer as respectivas distribuições de probabilidades (ao abrigo do Modelo);

A validade da inferência depende da validade dos pressupostos do modelo.

A notação matricial/vectorial

$$
\begin{array}{ccccccc}
Y_1 & = & \beta_0 + \beta_1 x_{1(1)} + \beta_2 x_{2(1)} + \cdots + \beta_p x_{p(1)} & + & \varepsilon_1 \\
Y_2 & = & \beta_0 + \beta_1 x_{1(2)} + \beta_2 x_{2(2)} + \cdots + \beta_p x_{p(2)} & + & \varepsilon_2 \\
Y_3 & = & \beta_0 + \beta_1 x_{1(3)} + \beta_2 x_{2(3)} + \cdots + \beta_p x_{p(3)} & + & \varepsilon_3 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\underbrace{Y_n}_{=\vec{\mathbf{Y}}} & = & \underbrace{\beta_0 + \beta_1 x_{1(n)} + \beta_2 x_{2(n)} + \cdots + \beta_p x_{p(n)}}_{=\mathbf{X}\vec{\boldsymbol{\beta}}} & + & \underbrace{\varepsilon_n}_{=\vec{\boldsymbol{\varepsilon}}}
\end{array}
$$

As n equações correspondem a uma única equação vectorial:

$$\vec{\mathbf{Y}} = \mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\varepsilon}}\,,$$

onde:

$$\vec{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{1(1)} & x_{2(1)} & \cdots & x_{p(1)} \\ 1 & x_{1(2)} & x_{2(2)} & \cdots & x_{p(2)} \\ 1 & x_{1(3)} & x_{2(3)} & \cdots & x_{p(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1(n)} & x_{2(n)} & \cdots & x_{p(n)} \end{bmatrix}, \quad \vec{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \vec{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

- $\vec{\mathbf{Y}}$ e $\vec{\boldsymbol{\varepsilon}}$ são vectores aleatórios,
- $\mathbf{X}$ é uma matriz não aleatória e $\vec{\boldsymbol{\beta}}$ um vector não-aleatório.

Modelo Regressão Linear - versão vectorial

O Modelo Linear em notação vectorial

1. $\vec{\mathbf{Y}} = \mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\varepsilon}}$.

2. $\vec{\boldsymbol{\varepsilon}} \frown \mathcal{N}_n(\vec{\mathbf{0}}, \sigma^2 \mathbf{I}_n)$, com $\vec{0} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \cdots \\ 0 \end{bmatrix}$; $\sigma^2 \mathbf{I}_n = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$

- Cada erro aleatório individual ε_i tem distribuição Normal.
- Cada erro aleatório individual tem média zero: $E[\varepsilon_i] = 0$.
- Cada erro aleatório individual tem variância igual: $V[\varepsilon_i] = \sigma^2$.
- Erros aleatórios diferentes são independentes, porque $Cov[\varepsilon_i, \varepsilon_j] = 0$ se $i \neq j$ e, numa Multinormal, isso implica a independência.

A distribuição de $\vec{\mathbf{Y}}$

Teorema (Primeiras Consequências do Modelo)

Dado o Modelo de Regressão Linear, tem-se:

$$\vec{\mathbf{Y}} \frown \mathcal{N}_n(\mathbf{X}\vec{\boldsymbol{\beta}}\,, \sigma^2 \mathbf{I}_n).$$

De facto, $\vec{\mathbf{Y}}$ é soma de vector não aleatório ($\mathbf{X}\vec{\boldsymbol{\beta}}$) e vector aleatório ($\vec{\boldsymbol{\varepsilon}}$):

$$\vec{\mathbf{Y}} = \underbrace{\mathbf{X}\vec{\boldsymbol{\beta}}}_{=\text{“}\vec{\mathbf{a}}\text{”}} + \underbrace{\vec{\boldsymbol{\varepsilon}}}_{=\text{“}\vec{\mathbf{Z}}\text{”}}.$$

- $\vec{\boldsymbol{\varepsilon}} \frown \mathcal{N}(\vec{\mathbf{0}}, \sigma^2 \mathbf{I}_n)$.
- Somar vector constante ($\mathbf{X}\vec{\boldsymbol{\beta}}$) a um vector aleatório Multinormal ($\vec{\boldsymbol{\varepsilon}}$) não destrói a Multinormalidade.
- $E[\vec{\mathbf{Y}}] = E[\mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\varepsilon}}] = \mathbf{X}\vec{\boldsymbol{\beta}} + E[\vec{\boldsymbol{\varepsilon}}] = \mathbf{X}\vec{\boldsymbol{\beta}}$.
- $V[\vec{\mathbf{Y}}] = V[\mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\varepsilon}}] = V[\vec{\boldsymbol{\varepsilon}}] = \sigma^2 \mathbf{I}_n$.

A distribuição de $\vec{\mathbf{Y}}$ (interpretação)

$$\vec{\mathbf{Y}} \frown \mathcal{N}_n(\mathbf{X}\vec{\boldsymbol{\beta}}, \sigma^2 \mathbf{I}_n).$$

Tendo em conta as propriedades da Multinormal:

- Cada observação individual Y_i tem distribuição Normal.
- Cada observação individual Y_i tem média $\mu_i = E[Y_i] = \vec{\mathbf{x}}^t_{[i,]}\vec{\boldsymbol{\beta}} = \beta_0 + \beta_1 x_{1(i)} + \beta_2 x_{2(i)} + \ldots + \beta_p x_{p(i)}$.
- Cada observação individual tem variância igual: $V[Y_i] = \sigma^2$.
- Observações diferentes de Y são independentes, porque $Cov[Y_i, Y_j] = 0$ se $i \neq j$ e, numa Multinormal, isso implica a independência.

O vector de estimadores $\hat{\boldsymbol{\beta}}$

O vector de estimadores $\vec{\hat{\boldsymbol{\beta}}} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)^t$ é definido a partir da equação do vector $\vec{\mathbf{b}}$ de estimativas, mas substituindo o vector $\vec{\mathbf{y}}$ de valores observados de Y pelo vector aleatório $\vec{\mathbf{Y}}$.

Estimadores de Mínimos Quadrados dos parâmetros

$$\vec{\hat{\boldsymbol{\beta}}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{Y}} .$$

O vector $\vec{\hat{\boldsymbol{\beta}}}$ é de dimensão $p+1$. O seu primeiro elemento é o estimador de β_0, o seu segundo elemento é o estimador de β_1, etc...

Em geral, o estimador de β_j está na posição $j+1$ do vector $\vec{\hat{\boldsymbol{\beta}}}$.

Os resultados gerais já referidos permitem facilmente determinar a distribuição de probabilidades do estimador $\vec{\hat{\boldsymbol{\beta}}}$.

A distribuição do vector de estimadores $\vec{\hat{\beta}}$

Teorema (Distribuição do estimador $\vec{\hat{\beta}}$)

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

$$\vec{\hat{\beta}} \frown \mathcal{N}_{p+1}(\vec{\beta}, \sigma^2 (\mathbf{X}^t\mathbf{X})^{-1}) .$$

$\vec{\hat{\beta}}$ é produto de matriz não aleatória, $(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$, e vector aleatório, $\vec{\mathbf{Y}}$:

$$\vec{\hat{\beta}} = \underbrace{(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t}_{\text{"B"}} \underbrace{\vec{\mathbf{Y}}}_{\text{"}\vec{\mathbf{Z}}\text{"}} .$$

- $\vec{\mathbf{Y}} \frown \mathcal{N}_n(\mathbf{X}\vec{\beta}, \sigma^2\mathbf{I}_n)$.
- Multiplicar matriz constante, $(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$, por um vector aleatório Multinormal ($\vec{\mathbf{Y}}$) não destrói a Multinormalidade.
- $E[\vec{\hat{\beta}}] = E[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{Y}}] = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t E[\vec{\mathbf{Y}}] = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}\vec{\beta} = \vec{\beta}$.
- $V[\vec{\hat{\beta}}] = V[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{Y}}] = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t V[\vec{\mathbf{Y}}][(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t]^t =$
  $(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t \cdot \sigma^2\mathbf{I}_n \cdot \mathbf{X}[(\mathbf{X}^t\mathbf{X})^{-1}]^t = \sigma^2 \cdot (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}[(\mathbf{X}^t\mathbf{X})^t]^{-1} = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$.

A distribuição de $\vec{\hat{\boldsymbol{\beta}}}$ (interpretação)

$$\vec{\hat{\boldsymbol{\beta}}} \sim \mathcal{N}_{p+1}(\vec{\boldsymbol{\beta}}\,, \sigma^2 (\mathbf{X}^t\mathbf{X})^{-1}) \,.$$

Tendo em conta as propriedades da Multinormal

- Cada estimador individual $\hat{\beta}_j$ tem distribuição Normal.
- Cada estimador individual tem média $E[\hat{\beta}_j] = \beta_j$, logo é centrado
- Cada estimador individual tem variância $V[\hat{\beta}_j] = \sigma^2 \left(\mathbf{X}^t\mathbf{X}\right)^{-1}_{(j+1,j+1)}$.
  (Note-se o desfasamento nos índices).
- Estimadores individuais diferentes não são (em geral) independentes, porque $(\mathbf{X}^t\mathbf{X})^{-1}$ não é, em geral, uma matriz diagonal: $Cov[\hat{\beta}_i, \hat{\beta}_j] = \sigma^2 \left(\mathbf{X}^t\mathbf{X}\right)^{-1}_{(i+1,j+1)}$.
- Logo, o estimador $\hat{\beta}_j$ de um parâmetro individual β_j tem distribuição $\hat{\beta}_j \sim \mathcal{N}(\beta_j\,, \sigma^2_{\hat{\beta}_j})$, com $\sigma^2_{\hat{\beta}_j} = \sigma^2 \left(\mathbf{X}^t\mathbf{X}\right)^{-1}_{(j+1,j+1)}$.

Se $p = 1$, RLS

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

Estimador de β_1

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

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

Se $p = 1$, RLS

Estimação dos parâmetros do Modelo RLS (cont.)

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

Estimador de β_0

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

com

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

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

Se $p = 1$, RLS

Distribuição dos estimadores RLS

Distribuição dos estimadores dos parâmetros

Dado o Modelo de Regressão Linear Simples,

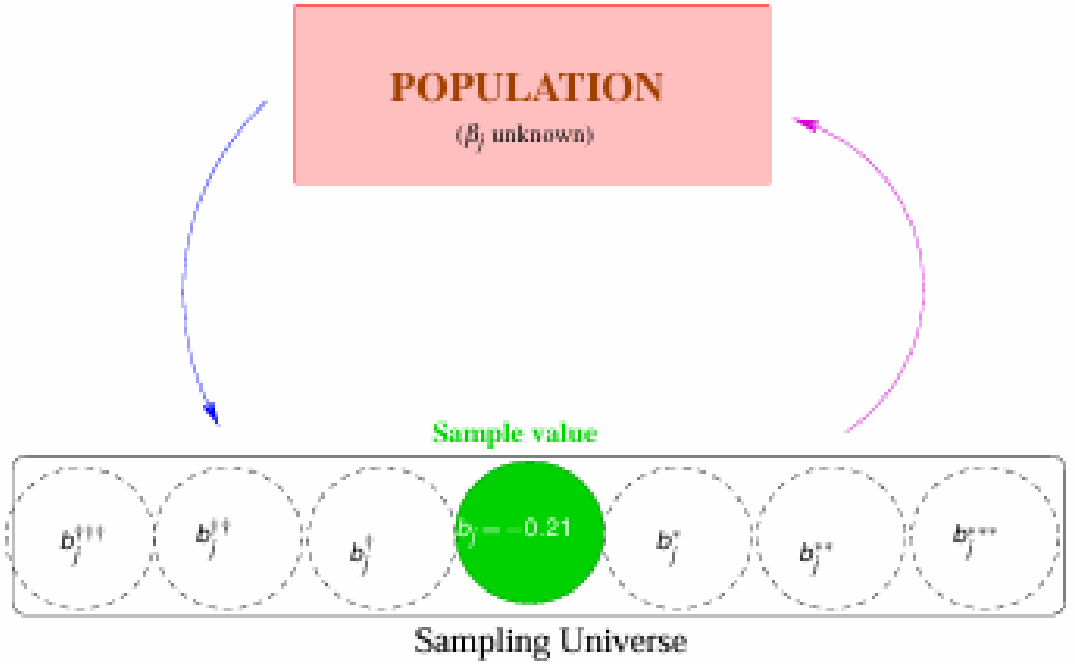
1. $\hat{\beta}_1 \frown \mathcal{N}\left(\beta_1 , \frac{\sigma^2}{(n-1)s_x^2}\right)$,
2. $\hat{\beta}_0 \frown \mathcal{N}\left(\beta_0 , \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}\right]\right)$

NOTAS:

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

A distribuição na amostragem de $\hat{\beta}_j$ (interpretação)

$$\hat{\beta}_j \frown \mathcal{N}(\beta_j\ ,\ \sigma^2_{\hat{\beta}_j}) \qquad \text{com} \qquad \sigma^2_{\hat{\beta}_j} = \sigma^2\,(\mathbf{X}'\mathbf{X})^{-1}_{(j+1,j+1)}\ .$$

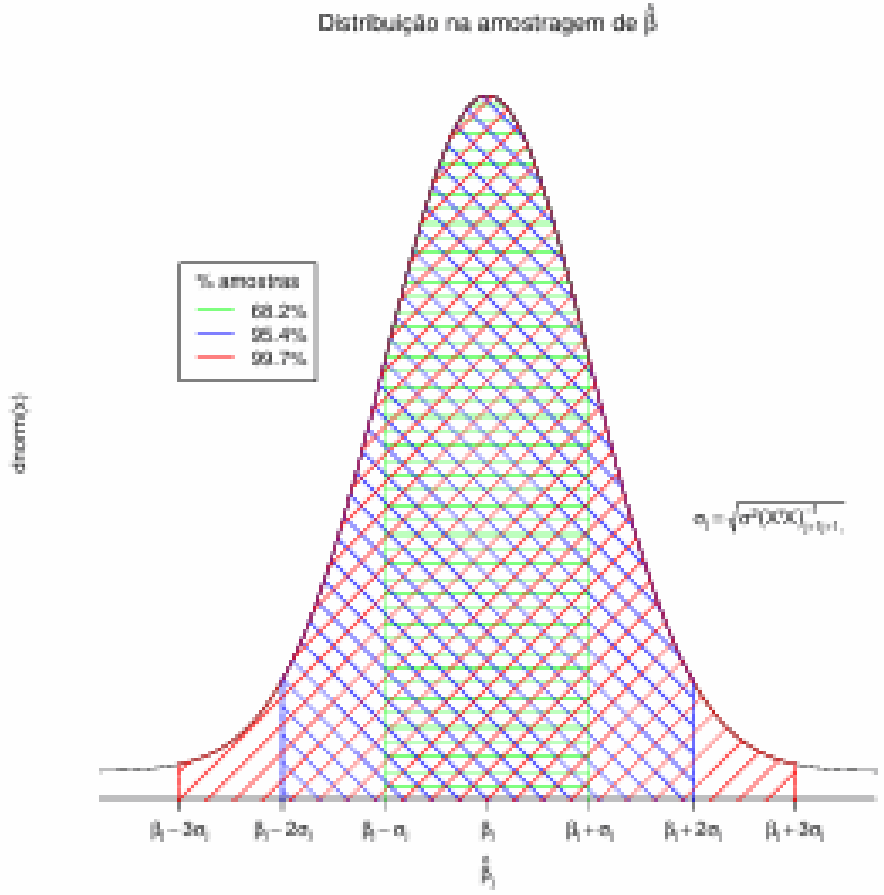


O conjunto de todas as possíveis amostras de dimensão n designa-se o Universo de Amostragem

A distribuição de probabilidades de $\hat{\beta}_j$ pode ser vista como a distribuição dos valores de b_j ao longo do Universo de Amostragem.

A distribuição na amostragem de $\hat{\beta}_j$ (interpretação)

$$\hat{\beta}_j \frown \mathcal{N}(\beta_j\ ,\ \sigma^2_{\hat{\beta}_j}) \qquad \text{com} \qquad \sigma^2_{\hat{\beta}_j} = \sigma^2\,(\mathbf{X}'\mathbf{X})^{-1}_{(j+1,j+1)}\ .$$



A distribuição dum estimador individual

Como se viu, tem-se, $\forall j = 0, 1, ..., p$:

$$\hat{\beta}_j \frown \mathcal{N}\left(\beta_j\ ,\ \sigma^2 (\mathbf{X}^t\mathbf{X})^{-1}_{(j+1,j+1)}\right)$$

$$\Leftrightarrow \quad \frac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}} \frown \mathcal{N}(0,1)\ ,$$

com $\sigma_{\hat{\beta}_j} = \sqrt{\sigma^2 (\mathbf{X}^t\mathbf{X})^{-1}_{(j+1,j+1)}}$.

Este resultado distribucional permitiria construir intervalos de confiança ou fazer testes a hipóteses sobre os parâmetros $\vec{\boldsymbol{\beta}}$, não fosse o desconhecimento da variância σ^2 dos erros aleatórios.

Se $p = 1$, RLS

Distribuição dos estimadores RLS

Distribuição dos estimadores (cont.)

Dado o Modelo de Regressão Linear Simples,

1. $\frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \sim \mathcal{N}(0,1)$, com $\sigma_{\hat{\beta}_1} = \sqrt{\frac{\sigma^2}{(n-1)S_x^2}} = \frac{\sigma}{\sqrt{(n-1)S_x^2}}$
2. $\frac{\hat{\beta}_0 - \beta_0}{\sigma_{\hat{\beta}_0}} \sim \mathcal{N}(0,1)$, 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.

O problema de σ^2 desconhecido

Para poder utilizar um estimador $\hat{\beta}_j$ na inferência, é preciso conhecer a sua distribuição de probabilidades, sem a presença de quantidades não-amostrais desconhecidas, além de β_j.

Para ultrapassar este problema é preciso:

- obter um estimador para σ^2; e
- ver o que acontece à distribuição de $\hat{\beta}_j$ quando σ^2 é substituído pelo seu estimador.

Como $\sigma^2 = V(\varepsilon_i), \forall i$, e como os erros aleatórios ε_i são desconhecidos, é natural procurar um estimador de σ^2 através dos resíduos.

Estimando σ^2

Erros aleatórios (variáveis aleatórias – não observáveis)

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_{1(i)} + \beta_2 x_{2(i)} + \ldots + \beta_p x_{p(i)})$$

Resíduos (variáveis aleatórias – observáveis)

$$E_i = Y_i - \underbrace{(\hat{\beta}_0 + \hat{\beta}_1 x_{1(i)} + \hat{\beta}_2 x_{2(i)} + \ldots + \hat{\beta}_p x_{p(i)})}_{=\hat{Y}_i}$$

Resíduos (observados)

$$e_i = y_i - (b_0 + b_1 x_{1(i)} + b_2 x_{2(i)} + \ldots + b_p x_{p(i)})$$

Quadrado Médio Residual (QMRE)

Define-se o Quadrado Médio Residual como

$$QMRE = \frac{SQRE}{n-(p+1)} = \frac{\sum_{i=1}^{n} E_i^2}{n-(p+1)}$$

Dado o Modelo Linear, $\hat{\sigma}^2 = QMRE$ é um estimador centrado da variância comum dos erros aleatórios, $\sigma^2 = V[\varepsilon_i]$:

$$E[QMRE] = \sigma^2 \ .$$

O Quadrado Médio Residual tem como unidades de medida o quadrado das unidades de Y.

Quantidades fulcrais para a inferência sobre β_j

A estimação dos erros padrão com o QMRE transforma as distribuições normais em distribuições *t-Student*

Teorema (Distribuições para a inferência sobre β_j)

Dado o Modelo de Regressão Linear Múltipla, tem-se

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \frown t_{n-(p+1)}\,, \qquad \forall j = 0, 1, \ldots, p$$

com $\hat{\sigma}_{\hat{\beta}_j} = \sqrt{QMRE \cdot (\mathbf{X}^t\mathbf{X})^{-1}_{(j+1,j+1)}}$.

Este Teorema dá-nos os resultados que servem de base à construção de intervalos de confiança e testes de hipóteses para os parâmetros β_j do modelo populacional.

Se $p = 1$, RLS

Quantidades centrais para a inferência sobre β_0 e β_1

A estimação dos erros padrão com o QMRE transforma as distribuições normais em distribuições *t-Student*

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

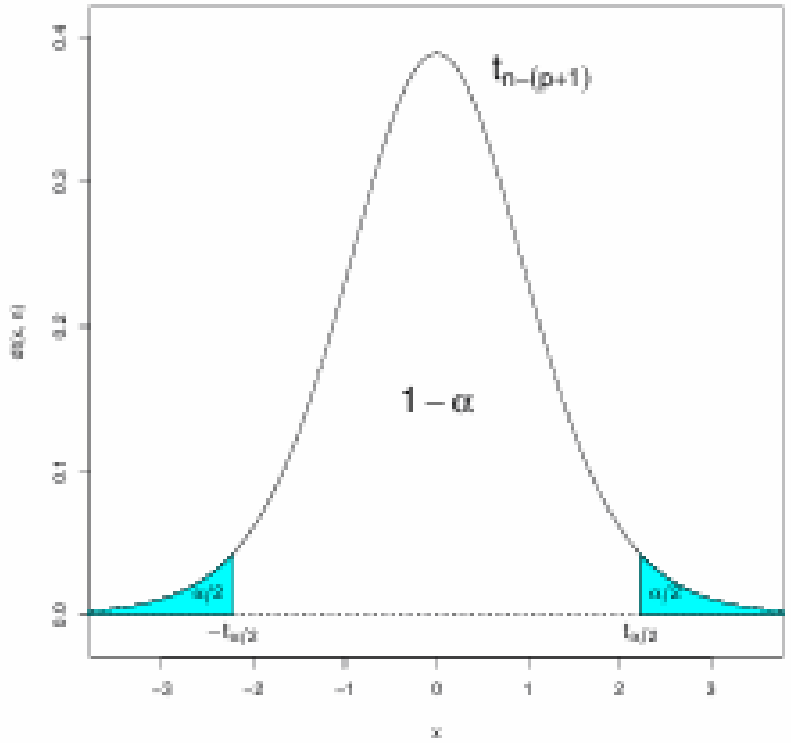
Dado o Modelo de Regressão Linear Simples, prova-se que:

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

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

Dedução de intervalo de confiança para β_j

Sabemos que $\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \frown t_{n-(p+1)}$. Logo,



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

Dedução IC para β_j (cont.)

Trabalhar a dupla desigualdade até isolar β_j:

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

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

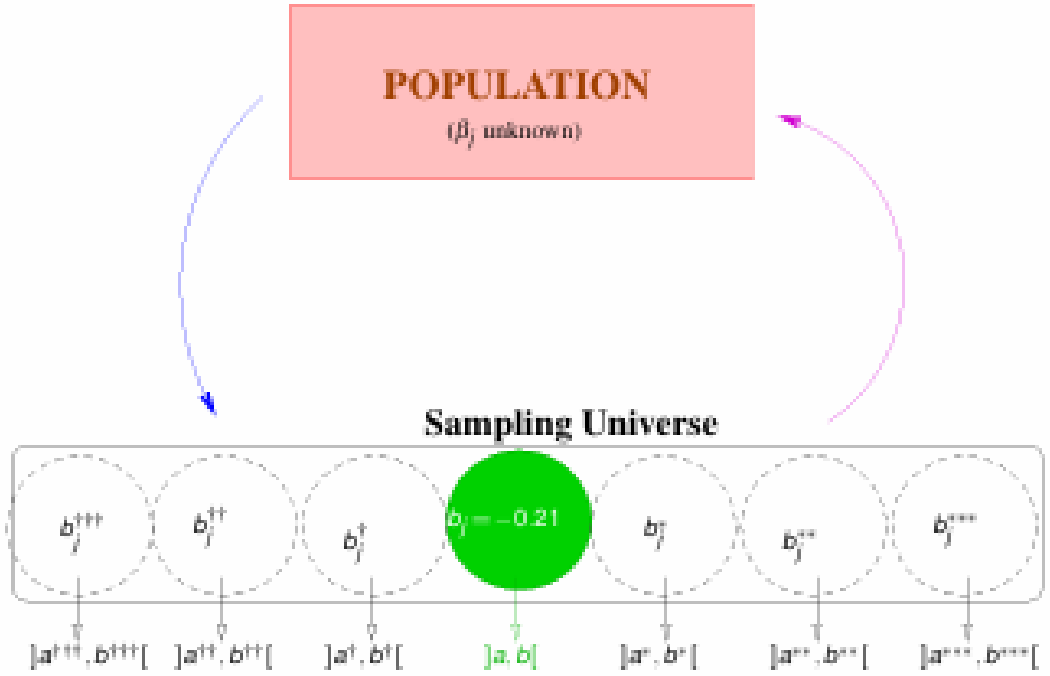
O intervalo aleatório

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

contém β_j com probabilidade $1-\alpha$.

Intervalo aleatório para β_j (interpretação)

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



Cada amostra no Universo de Amostragem gera um intervalo concreto, chamado Intervalo de Confiança

Uma proporção $1-\alpha$ desses intervalos contêm o verdadeiro valor de β_j. Os restantes α não contêm β_j.

Intervalo de confiança para β_j

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

Dado o Modelo de Regressão Linear Múltipla e uma amostra, eis o intervalo a $(1-\alpha)\times 100\%$ de confiança para o parâmetro β_j:

$$\left] \; b_j - t_{\frac{\alpha}{2}\,[n-(p+1)]} \cdot \hat{\sigma}_{\hat{\beta}_j} \;\; , \;\; b_j + t_{\frac{\alpha}{2}\,[n-(p+1)]} \cdot \hat{\sigma}_{\hat{\beta}_j} \; \right[,$$

sendo:

- b_j o elemento $j+1$ do vector das estimativas $\vec{\mathbf{b}}$
- $t_{\frac{\alpha}{2}\,[n-(p+1)]}$ o quantil de ordem $1-\frac{\alpha}{2}$ da distribuição $t_{n-(p+1)}$;
- $\hat{\sigma}_{\hat{\beta}_j} = \sqrt{QMRE \cdot (\mathbf{X}^t\mathbf{X})^{-1}_{(j+1,j+1)}}$ (com o valor de QMRE na nossa amostra).

NOTA: A amplitude do IC aumenta com *QMRE* e o valor diagonal da matriz $(\mathbf{X}^t\mathbf{X})^{-1}$ correspondente ao parâmetro β_j.

Se $p = 1$, RLS

Intervalo de confiança para β_1

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

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

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

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

NOTAS:

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

Se $p = 1$, RLS

Intervalo de confiança para β_0

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

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

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

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

NOTAS:

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

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

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

Exemplo: RLM

Intervalos de confiança para β_j no R

A informação para construir intervalos de confiança para cada β_j obtém-se a partir da função `summary`.

```
> summary(iris2.lm)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.24031    0.17837  -1.347     0.18
Petal.Length  0.52408    0.02449  21.399  < 2e-16 ***
Sepal.Length -0.20727    0.04751  -4.363 2.41e-05 ***
Sepal.Width   0.22283    0.04894   4.553 1.10e-05 ***
```

Estima-se que em média a largura da pétala diminui $0.20727 cm$ por cada aumento de $1 cm$ no comprimento da sépala (mantendo-se as outras medições constantes).

Como $t_{0.025(146)} = 1.976346$, o IC a 95% para β_2 é

$$\left]\ (-0.20727) - (1.976346)(0.04751)\ ,\ (-0.20727) + (1.976346)(0.04751)\ \right[$$

$$\Leftrightarrow \quad \left]\ -0.3012\ ,\ -0.1134\ \right[$$

Exemplo: RLM

Intervalos de confiança para β_j no R (cont.)

Alternativamente, é possível usar a função `confint` para obter os intervalos de confiança para cada β_j individual:

```
> confint(iris2.lm)                    <- IC a 95% confiança (por omissão)
                   2.5 %      97.5 %
(Intercept)   -0.5928277   0.1122129
Petal.Length   0.4756798   0.5724865
Sepal.Length -0.3011547 -0.1133775
Sepal.Width    0.1261101   0.3195470

> confint(iris2.lm , level=0.99)       <- IC a 99% de confiança
                   0.5 %       99.5 %
(Intercept)   -0.70583864   0.22522386
Petal.Length   0.46016260   0.58800363
Sepal.Length -0.33125352 -0.08327863
Sepal.Width    0.09510404   0.35055304
```

Exemplo: RLS

Intervalos de confiança no R

O exemplo das cerejeiras

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

```
> summary(cerejeiras.lm)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.046122   0.095290  -10.98 7.62e-12 ***
DAP          0.056476   0.002758   20.48  < 2e-16 ***
```

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

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

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

Exemplo: RLS

Intervalos de confiança de β_0 e β_1 no R

As cerejeiras (cont.)

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

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

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

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

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

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

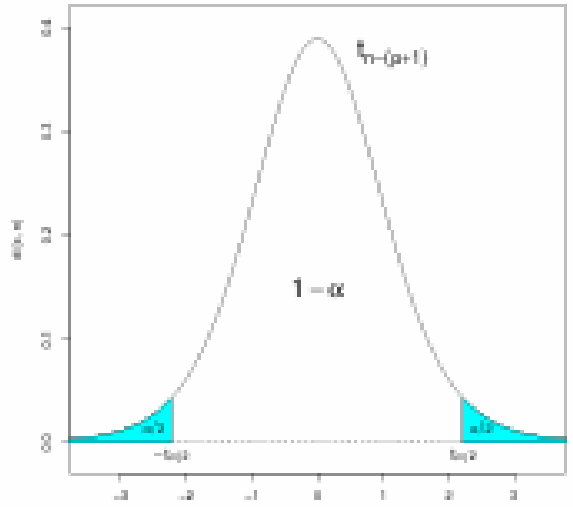
Testes de Hipóteses sobre os parâmetros

O resultado usado para construir ICs também permite Testes a Hipóteses sobre cada β_j. Admitindo a Hipótese Nula $H_0 : \beta_j = c$:

$$T = \frac{\hat{\beta}_j - \overbrace{\beta_{j|H_0}}^{=c}}{\hat{\sigma}_{\hat{\beta}_j}} \frown t_{n-(p+1)}\,, \qquad \forall\, j = 0, 1, \ldots, p$$

Rejeita-se H_0 em favor da Hipótese Alternativa $H_1 : \beta_j \neq c$ se o valor calculado de T na amostra, T_{calc}, recair numa das caudas da distribuição.

Fixando o Nível de Significância α, tem-se a Região Crítica:



Testes de Hipóteses (bilateral) a $\hat{\beta}_j$

Testes de Hipóteses a β_j (Modelo de Regressão Linear Múltipla)

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

Estatística do Teste: $T = \frac{\hat{\beta}_j - \overbrace{\beta_j|_{H_0}}^{=c}}{\hat{\sigma}_{\hat{\beta}_j}} \frown t_{n-(p+1)}$, se H_0 verdade.

Nível de significância do teste: α

Região Crítica (Região de Rejeição bilateral): Rejeitar H_0 se

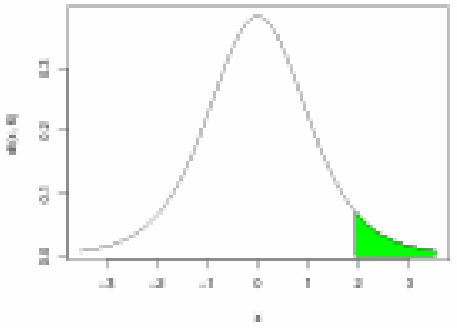
$$T_{calc} > t_{\frac{\alpha}{2}[n-(p+1)]} \quad \text{ou} \quad T_{calc} < -t_{\frac{\alpha}{2}[n-(p+1)]}$$

$$\Longleftrightarrow \quad |T_{calc}| > t_{\frac{\alpha}{2}[n-(p+1)]}$$

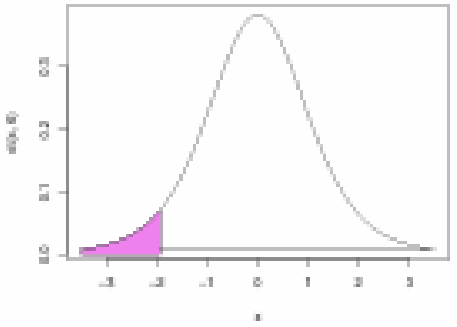
Testes de Hipóteses a $\hat{\beta}_j$ (unilaterais)

$$T = \frac{\hat{\beta}_j - \overbrace{\beta_{j|H_0}}^{=c}}{\hat{\sigma}_{\hat{\beta}_j}} \frown t_{n-(p+1)}$$

Com a Hipótese Alternativa $H_1 : \beta_j > c$, só valores grandes da estatística sugerem a rejeição de $H_0 : \beta_j \leq c$ (ou $H_0 : \beta_j = c$):



Com a Hipótese Alternativa $H_1 : \beta_j < c$, só valores pequenos de T_{calc} sugerem rejeitar $H_0 : \beta_j \geq c$ (ou $H_0 : \beta_j = c$):



Testes de Hipóteses sobre os parâmetros

Dado o Modelo de Regressão Linear Múltipla,

Testes de Hipóteses a β_j (Regressão Linear Múltipla)

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

Estatística do Teste: $T = \frac{\hat{\beta}_j - \overbrace{\beta_j|_{H_0}}^{=c}}{\hat{\sigma}_{\hat{\beta}_j}} \frown t_{n-(p+1)}$, se H_0 verdade.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Rejeitar H_0 se

$T_{calc} < -t_{\alpha[n-(p+1)]}$	(Unilateral esquerdo)
$\lvert T_{calc} \rvert > t_{\alpha/2[n-(p+1)]}$	(Bilateral)
$T_{calc} > t_{\alpha[n-(p+1)]}$	(Unilateral direito)

Os *p-values*

Valores de prova (*p-value*)

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

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

Sendo $T \frown t_{n-(p+1)}$

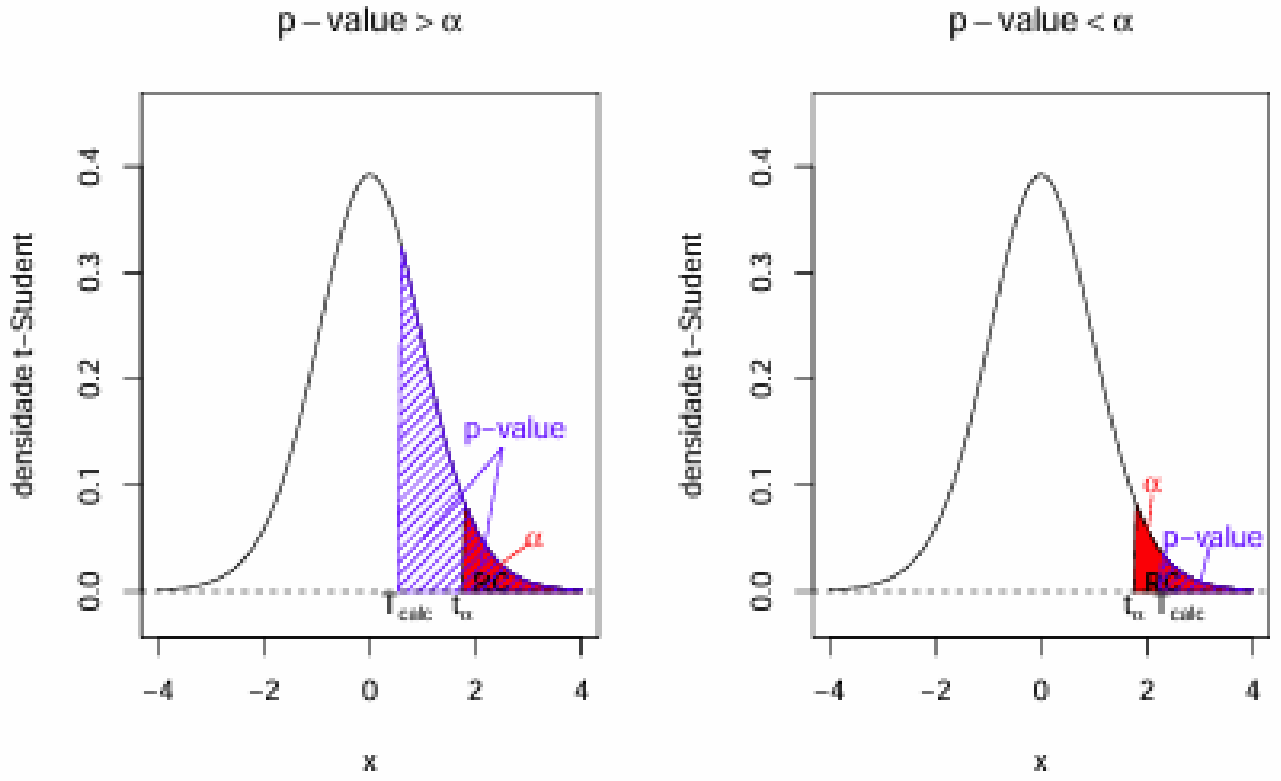
RC Unilateral direita: $p = P[T > T_{calc}]$

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

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

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

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



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

Exemplo: RLS

```
> summary(cerejeiras.lm)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.046122   0.095290  -10.98 7.62e-12 ***
DAP          0.056476   0.002758   20.48  < 2e-16 ***
```

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

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

O valor de prova (*p-value*) indica uma claríssima rejeição da hipótese nula.

Exemplo: RLS

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

```
> summary(cerejeiras.lm)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.046122   0.095290  -10.98 7.62e-12 ***
DAP          0.056476   0.002758   20.48  < 2e-16 ***
```

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

$$T_{calc} = \frac{b_1 - \overbrace{\beta_1|_{H_0}}^{=0.05}}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{0.056476 - 0.05}{0.002758} = 2.348078 \, .$$

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

Combinações lineares dos parâmetros

Seja $\vec{\mathbf{a}} = (a_0, a_1, \ldots, a_p)^t$ um vector não aleatório em $\mathbb{R}^{p+1}$. O produto interno $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$ define uma combinação linear dos parâmetros do modelo:

$$\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} = a_0\beta_0 + a_1\beta_1 + a_2\beta_2 + \ldots + a_p\beta_p \,.$$

Casos particulares importantes são se:

- $\vec{\mathbf{a}}$ tem um único elemento não-nulo, $a_{j+1} = 1$: $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} = \beta_j$.
- $\vec{\mathbf{a}}$ só tem dois elementos não-nulos, $a_{i+1} = 1$ e $a_{j+1} = \pm 1$: $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} = \beta_i \pm \beta_j$.
- $\vec{\mathbf{a}} = (1, x_1, x_2, \ldots, x_p)$: $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$ é o valor esperado de Y associado aos valores indicados das variáveis preditoras:

$$\begin{aligned} \vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \\ &= E[Y \mid X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p] \\ &= \mu_{Y|\vec{\mathbf{x}}} \end{aligned}$$

Inferência sobre combinações lineares dos β_js

Estima-se $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$ com a mesma combinação linear dos estimadores:

$$\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}} = a_0\hat{\beta}_0 + a_1\hat{\beta}_1 + a_2\hat{\beta}_2 + \ldots + a_p\hat{\beta}_p \ .$$

Sabemos determinar a distribuição de probabilidades de $\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}$:

- Sabemos que $\hat{\vec{\boldsymbol{\beta}}} \frown \mathcal{N}_{p+1}\left(\vec{\boldsymbol{\beta}}, \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}\right)$
- Logo, $\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}} \frown \mathcal{N}_1(\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}, \sigma^2\vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}})$
- Ou seja, $\vec{Z} = \frac{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}} - \vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}}{\sqrt{\sigma^2\vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}}} \frown \mathcal{N}(0,1)$;
- Por um raciocínio análogo ao usado nos β_j individuais, tem-se:

$$\frac{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}} - \vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}}{\sqrt{QMRE \cdot \vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}}} \frown t_{n-(p+1)} \ .$$

Quantidades centrais para a inferência sobre $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$

Teorema (Distribuições para combinações lineares dos βs)

Dado o Modelo de Regressão Linear Múltipla, tem-se

$$\frac{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}} - \vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}}{\hat{\sigma}_{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}}} \frown t_{n-(p+1)} \ ,$$

com $\hat{\sigma}_{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}} = \sqrt{QMRE \cdot \vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}}$.

Este Teorema dá-nos os resultados que servem de base à construção de intervalos de confiança e testes de hipóteses para quaisquer combinações lineares dos parâmetros β_j do modelo.

Intervalo de confiança para $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$

Intervalo de Confiança a $(1-\alpha)\times 100\%$ para $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$

Dado o Modelo de Regressão Linear Múltipla e uma amostra, o intervalo a $(1-\alpha)\times 100\%$ de confiança para uma combinação linear dos parâmetros, $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} = a_0\beta_0 + a_1\beta_1 + \ldots + a_p\beta_p$, é:

$$\left] \vec{\mathbf{a}}^t\vec{\mathbf{b}} - t_{\frac{\alpha}{2}[n-(p+1)]} \cdot \hat{\sigma}_{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}} \;,\; \vec{\mathbf{a}}^t\vec{\mathbf{b}} + t_{\frac{\alpha}{2}[n-(p+1)]} \cdot \hat{\sigma}_{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}} \right[,$$

com $\vec{\mathbf{a}}^t\vec{\mathbf{b}} = a_0b_0 + a_1b_1 + \ldots + a_pb_p$ e $\hat{\sigma}_{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}} = \sqrt{QMRE \cdot \vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}}$.

Fórmulas para a estimação de $\beta_i \pm \beta_j$

A fórmula geral $\hat{\sigma}_{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}} = \sqrt{QMRE \cdot \vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}}$ admite uma expressão alternativa no caso particular duma soma ou diferença de dois βs.

Pela fórmula geral da variância duma soma ou diferença de v.a.s,

$$\begin{aligned} V[\hat{\beta}_i \pm \hat{\beta}_j] &= V[\hat{\beta}_i] + V[\hat{\beta}_j] \pm 2\,Cov[\hat{\beta}_i, \hat{\beta}_j]\ . \\ \Leftrightarrow \quad \sigma^2_{\hat{\beta}_i \pm \hat{\beta}_j} &= \sigma^2 \cdot \left[(\mathbf{X}^t\mathbf{X})^{-1}_{[i+1,i+1]} + (\mathbf{X}^t\mathbf{X})^{-1}_{[j+1,j+1]} \pm 2\,(\mathbf{X}^t\mathbf{X})^{-1}_{[i+1,j+1]}\right]\ . \end{aligned}$$

Logo, o erro padrão de $\hat{\beta}_i \pm \hat{\beta}_j$ é:

$$\hat{\sigma}_{\hat{\beta}_i \pm \hat{\beta}_j} = \sqrt{QMRE \cdot \left[(\mathbf{X}^t\mathbf{X})^{-1}_{[i+1,i+1]} + (\mathbf{X}^t\mathbf{X})^{-1}_{[j+1,j+1]} \pm 2\,(\mathbf{X}^t\mathbf{X})^{-1}_{[i+1,j+1]}\right]}\ .$$

ICs para combinações lineares no R

Numa RLM, o IC duma combinação linear genérica $\vec{a}^t\vec{\beta}$, precisa da matriz das (co)variâncias estimadas dos estimadores $\vec{\hat{\beta}}$,

$$\widehat{V[\vec{\hat{\beta}}]} = QMRE \cdot (\mathbf{X}^t\mathbf{X})^{-1} \ ,$$

que é gerada pelo comando `R vcov`.

A matriz das (co)variâncias estimadas no exemplo RLM dos lírios é:

```
> vcov(iris2.lm)
              (Intercept)  Petal.Length Sepal.Length  Sepal.Width
(Intercept)   0.031815766  0.0015144174 -0.005075942 -0.002486105
Petal.Length  0.001514417  0.0005998259 -0.001065046  0.000802941
Sepal.Length -0.005075942 -0.0010650465  0.002256837 -0.001344002
Sepal.Width  -0.002486105  0.0008029410 -0.001344002  0.002394932
```

ICs para combinações lineares no R

O erro padrão estimado de $\hat{\beta}_2 + \hat{\beta}_3$

$$\hat{\sigma}_{\hat{\beta}_2+\hat{\beta}_3} = \sqrt{\hat{V}[\hat{\beta}_2 + \hat{\beta}_3]} = \sqrt{\hat{V}[\hat{\beta}_2] + \hat{V}[\hat{\beta}_3] + 2\hat{Cov}[\hat{\beta}_2, \hat{\beta}_3]}$$

$$\hat{\sigma}_{\hat{\beta}_2+\hat{\beta}_3} = \sqrt{0.002256837 + 0.002394932 + 2(-0.001344002)} = 0.04431439 .$$

```
> vcov(iris2.lm)
              (Intercept)  Petal.Length Sepal.Length  Sepal.Width
(Intercept)   0.031815766  0.0015144174 -0.005075942 -0.002486105
Petal.Length  0.001514417  0.0005998259 -0.001065046  0.000802941
Sepal.Length -0.005075942 -0.0010650465  0.002256837 -0.001344002
Sepal.Width  -0.002486105  0.0008029410 -0.001344002  0.002394932
```

Testes a combinações lineares dos parâmetros

Dado o Modelo de Regressão Linear Múltipla,

Testes de Hipóteses relativos a $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$

Hipóteses: $H_0: \vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} \begin{matrix} \geq \\ = \\ \leq \end{matrix} c$ vs. $H_1: \vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} \begin{matrix} < \\ \neq \\ > \end{matrix} c$

Estatística do Teste: $T = \dfrac{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}} - \overbrace{\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}|_{H_0}}^{=c}}{\hat{\sigma}_{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}}} \frown t_{n-(p+1)}$, se H_0 verdade

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Rejeitar H_0 se

$T_{calc} < -t_{\alpha[n-(p+1)]}$	(Unilateral esquerdo)
$\lvert T_{calc}\rvert > t_{\alpha/2[n-(p+1)]}$	(Bilateral)
$T_{calc} > t_{\alpha[n-(p+1)]}$	(Unilateral direito)

Intervalos de confiança para $E[Y|X_1 = x_1, \ldots, X_p = x_p]$

Como caso particular do resultado anterior, tem-se:

IC para o valor esperado de Y, dados os preditores

Dado o Modelo RLM e uma amostra com os valores $\vec{\mathbf{x}} = (x_1, x_2, \ldots, x_p)^t$ das variáveis preditoras, o valor esperado de Y,

$$\mu_{Y|\vec{\mathbf{x}}} = E[Y|X_1 = x_1, \ldots, X_p = x_p] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \ ,$$

é estimado por $\hat{\mu}_{Y|\vec{\mathbf{x}}} = b_0 + b_1 x_1 + \ldots + b_p x_p$.

Um intervalo a $(1-\alpha) \times 100\%$ de confiança para $\mu_{Y|\vec{\mathbf{x}}}$ é dado por:

$$\left] \ \hat{\mu}_{Y|\vec{\mathbf{x}}} - t_{\frac{\alpha}{2}[n-(p+1)]} \cdot \hat{\sigma}_{\hat{\mu}_{Y|\vec{\mathbf{x}}}} \ , \ \hat{\mu}_{Y|\vec{\mathbf{x}}} + t_{\frac{\alpha}{2}[n-(p+1)]} \cdot \hat{\sigma}_{\hat{\mu}_{Y|\vec{\mathbf{x}}}} \ \right[\ ,$$

com $\hat{\sigma}_{\hat{\mu}_{Y|\vec{\mathbf{x}}}} = \sqrt{QMRE \cdot \vec{\mathbf{a}}^t (\mathbf{X}^t \mathbf{X})^{-1} \vec{\mathbf{a}}}$, onde $\vec{\mathbf{a}} = (1, x_1, x_2, \ldots, x_p)$.

Se $p = 1$, RLS

Fórmulas para uma regressão linear simples

Numa regressão linear simpes, a fórmula da variância de $\hat{\mu}_{Y|x}$ é:

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

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

O intervalo de confiança para $\mu_{Y|x}$ na RLS é:

$$\left]\ (b_0 + b_1\, x) - t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\mu}_{Y|x}} \ ;\ (b_0 + b_1\, x) + t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\mu}_{Y|x}}\ \right[.$$

Exemplo: RLS

Inferência sobre $\mu_{Y|\vec{\mathbf{x}}} = E[Y|\vec{\mathbf{x}}]$ no R

Valores estimados e intervalos de confiança para $\mu_{Y|\vec{\mathbf{x}}}$ obtêm-se com a função `predict`. Os novos valores dos preditores são indicados numa `data frame` (com nomes iguais aos do ajustamento inicial).

No exemplo de Regressão Linear Simples nos lírios, a largura esperada de pétalas de comprimento 1.85 e 4.65, é:

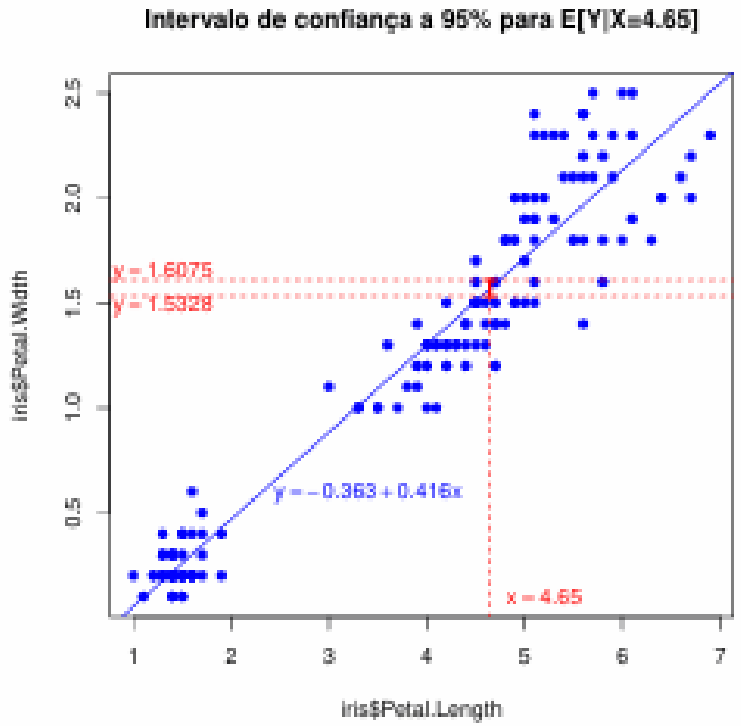
```
> predict(iris.lm, new=data.frame(Petal.Length=c(1.85,4.65)))
       1        2
0.406072 1.570187
```

Exemplo: RLS

Inferência sobre $E[Y|\vec{\mathbf{x}}]$ no R (continuação)

O intervalo de confiança para $\mu_{Y|\vec{\mathbf{x}}}$ obtém-se com o argumento `int="conf"`:

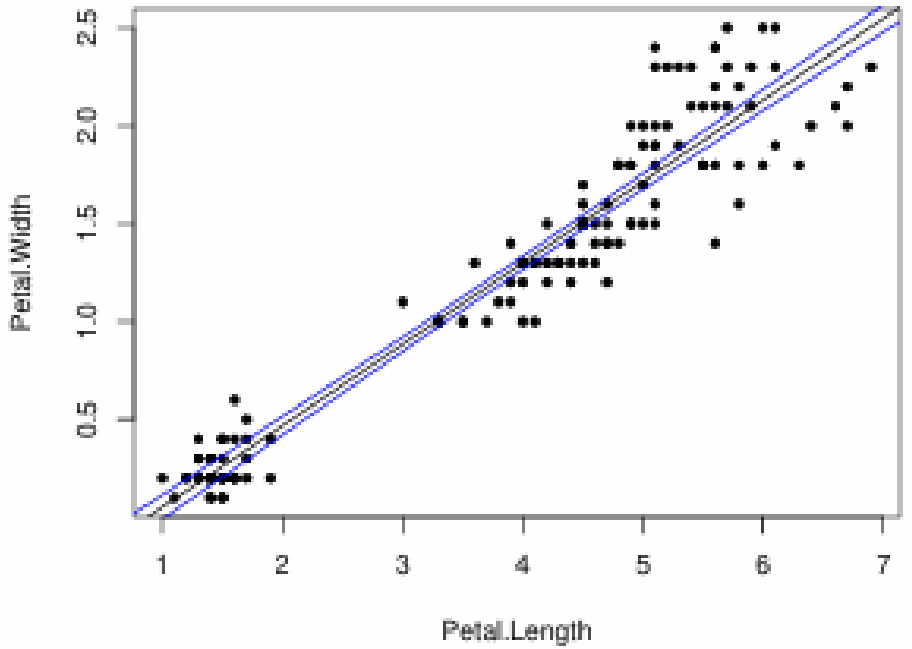
```
> predict(iris.lm,data.frame(Petal.Length=c(4.65)),int="conf")
       fit       lwr       upr
1 1.570187 1.5328338 1.6075405
```



Exemplo: RLS

Bandas de confiança para a recta de regressão

Considerando os ICs para todos os valores de x nalgum intervalo, obtém-se uma banda de confiança que contém a recta de regressão com $(1-\alpha) \times 100\%$ de confiança.



Os IC para $\mu_{Y|x}$ dependem do valor de x

maior amplitude quanto mais afastado x estiver da média $\overline{x}$ das observações. Logo, as bandas são encurvadas.

Exemplo: RLM

RLM: Intervalos de confiança para $E[Y|\vec{\mathbf{x}}]$ no 

O comando `predict` também permite obter ICs para $\mu_{Y|\vec{\mathbf{x}}}$ numa regressão linear múltipla.

Na regressão linear múltipla dos lírios, eis um IC a 95% para a largura esperada de pétalas de flores com:

`Petal.Length=2` `Sepal.Length=5` `Sepal.Width=3.1`

```
> predict(iris2.lm, new=data.frame(Petal.Length=c(2),
+     Sepal.Length=c(5), Sepal.Width=c(3.1)), int="conf")

          fit       lwr       upr
[1,] 0.462297 0.4169203 0.5076736
```

O IC para $E[Y \mid X_1 = 2, X_2 = 5, X_3 = 3.1]$ é: $]\, 0.4169203\ ,\ 0.5076736\, [$.

Não é possível visualizar este intervalo em $\mathbb{R}^4$.

A variabilidade duma observação individual de Y

Consideraram-se intervalos de confiança para o valor esperado de Y,

$$\mu_{Y|\vec{\mathbf{x}}} = E[Y|x_1{=}x_1, x_2{=}x_2, \ldots, x_p{=}x_p] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \ ,$$

usam a variabilidade associada ao estimador $\hat{\mu}_{Y|\vec{\mathbf{x}}}$:

$$\sigma^2_{\hat{\mu}_{Y|\vec{\mathbf{x}}}} = V[\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p] = \sigma^2 \cdot \vec{\mathbf{a}}^t (\mathbf{X}^t\mathbf{X})^{-1} \vec{\mathbf{a}},$$

com $\vec{\mathbf{a}} = (1, x_1, x_2, \ldots, x_p)$.

Uma observação individual de Y tem uma variabilidade adicional, pois:

$$Y \quad = \quad \mu_{Y|\vec{\mathbf{x}}} + \varepsilon \quad = \quad \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon \ .$$

A flutuação aleatória de observações individuais em torno do hiperplano é $V[\varepsilon] = \sigma^2$. Será necessário somar a variância associada à estimação do hiperplano e a variância das observações individuais:

$$\sigma^2_{Indiv} \ = \ V[\hat{\mu}_{Y|\vec{\mathbf{x}}}] + V[\varepsilon] \ = \ \sigma^2 \cdot \vec{\mathbf{a}}^t (\mathbf{X}^t\mathbf{X})^{-1} \vec{\mathbf{a}} + \sigma^2 \ = \ \sigma^2 \cdot \left[\vec{\mathbf{a}}^t (\mathbf{X}^t\mathbf{X})^{-1} \vec{\mathbf{a}} + 1 \right].$$

Intervalos de predição para Y

Podem obter-se intervalos de predição para uma observação individual de Y, associada aos valores $X_1 = x_1, \ldots, X_p = x_p$ das variáveis preditoras.

Nestes intervalos, a estimativa da variância duma observação individual de Y é a estimativa de σ^2_{Indiv}, resultante de substituir σ^2 pelo *QMRE* amostral:

Intervalos de predição para observações individuais

$$\left] \hat{\mu}_{Y|\vec{\mathbf{x}}} - t_{\frac{\alpha}{2}[n-(p+1)]} \cdot \hat{\sigma}_{indiv} \;,\; \hat{\mu}_{Y|\vec{\mathbf{x}}} + t_{\frac{\alpha}{2}[n-(p+1)]} \cdot \hat{\sigma}_{indiv} \right[$$

onde

$$\hat{\mu}_{Y|X} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$$

e

$$\hat{\sigma}_{indiv} = \sqrt{QMRE\left[1 + \vec{\mathbf{a}}^t (\mathbf{X}^t \mathbf{X})^{-1} \vec{\mathbf{a}}\right]} \quad \text{com} \quad \vec{\mathbf{a}} = (1, x_1, x_2, \ldots, x_p).$$

Se $p = 1$, RLS

Fórmulas para a regressão linear simples

Na regressão linear simples usa-se a fórmula

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

Logo,

RLS: Intervalo de predição para observação individual de Y

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

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

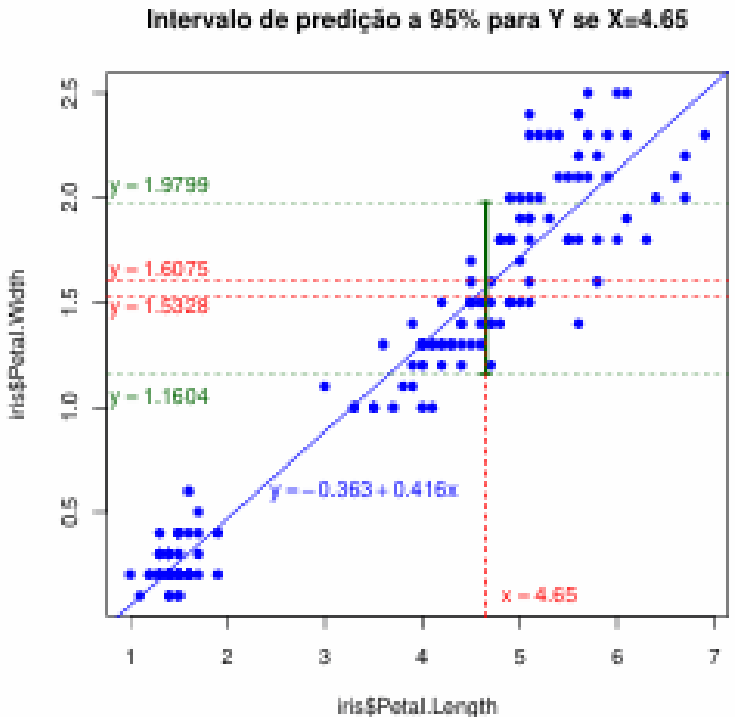
Quer numa regressão linear simples, quer numa múltipla, estes intervalos são necessariamente de maior amplitude que os intervalos de confiança para $\mu_{Y|\vec{x}}$ (para igual nível de confiança $(1-\alpha)\times 100\%$).

Exemplo: RLS

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`:

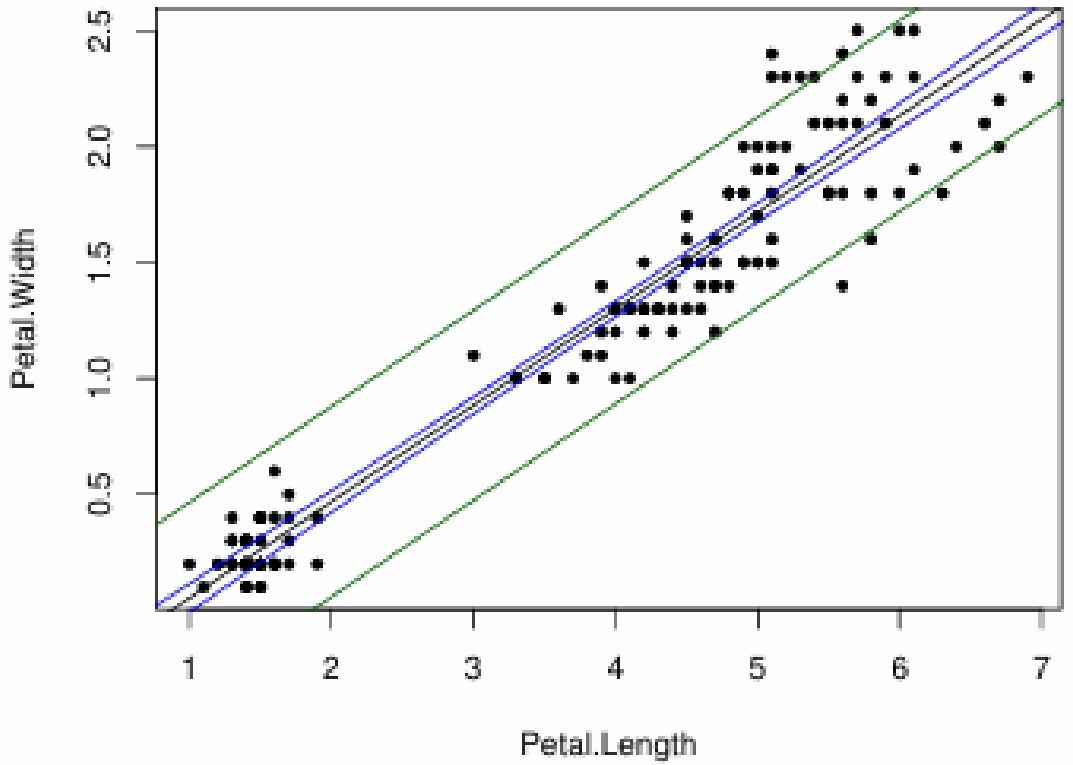
```
> predict(iris.lm,data.frame(Petal.Length=c(4.65)), int="pred")
       fit          lwr      upr
1 1.570187   1.160442632 1.9799317
```



Exemplo: RLS

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.



Exemplo: RLM

Intervalos de predição para Y (cont.)

Eis, na Regressão Linear Múltipla dos lírios, o intervalo de predição para a largura da pétala, num lírio com comprimento de pétala 2, e com sépala de comprimento 5 e largura 3.1:

```
> predict(iris2.lm, data.frame(Petal.Length=c(2),
+     Sepal.Length=c(5), Sepal.Width=c(3.1)), int="pred")

          fit       lwr       upr
[1,] 0.462297 0.08019972 0.8443942
```

O intervalo de predição pedido é: $]\, 0.0802\ ,\ 0.8444\, [$.

O correspondente intervalo de confiança para $\mu_{Y|\vec{\mathbf{x}}}$ era $]\, 0.4169\ ,\ 0.5077\, [$, necessariamente mais curto.

Testando a qualidade do ajustamento global

Numa Regressão Linear, o modelo é inútil se fôr indistinguível do modelo nulo, i.e., do modelo de equação $Y_i = \beta_0 + \varepsilon_i$. O modelo nulo pode ser visto como um submodelo de qualquer modelo linear, em que todas as variáveis preditoras têm coeficiente nulo: $\beta_j = 0, \forall j > 0$.

O teste de ajustamento global visa testar se um dado modelo linear é significativamente diferente do modelo nulo.

As hipóteses em confronto são:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$
[MODELO COMPLETO $\equiv$ MODELO NULO]

vs.

$$H_1 : \exists j = 1, \ldots, p \quad \text{t.q.} \quad \beta_j \neq 0$$
[MODELO COMPLETO $\not\equiv$ MODELO NULO]

NOTA: repare que β_0 não intervém nas hipóteses.

O teste de ajustamento global (cont.)

Definindo:

- O Quadrado Médio da Regressão como $QMR = \frac{SQR}{p}$.
- O Quadrado Médio Residual como $QMRE = \frac{SQRE}{n-(p+1)}$.

Sob a Hipótese Nula do teste de ajustamento global:

$$F = \frac{QMR}{QMRE} \frown F_{[\,p\,,\,n-(p+1)\,]}\,.$$

Esta é a estatística F do teste de ajustamento global.

Expressão alternativa para a estatística do teste F

A estatística do teste F de ajustamento global do modelo numa Regressão Linear Múltipla pode ser escrita na forma alternativa:

$$F = \frac{n-(p+1)}{p} \cdot \frac{R^2}{1-R^2} .$$

A estatística F é uma função crescente do coeficiente de determinação amostral R^2, o que justifica a natureza unilateral direita da região crítica.

As hipóteses do teste também se podem escrever como

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

A hipótese $H_0 : \mathscr{R}^2 = 0$ indica ausência de relação linear entre Y e o conjunto dos preditores. Corresponde a um ajustamento "péssimo" do modelo. A sua rejeição não garante um bom ajustamento.

O Teste F de ajustamento global do Modelo

Teste F de ajustamento global do modelo RLM

Hipóteses: $H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$

vs.

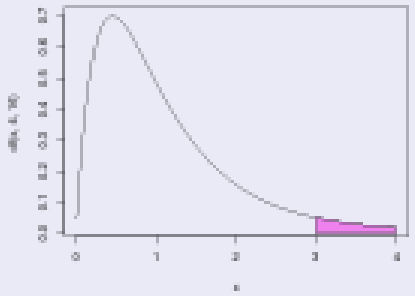
$H_1 : \exists\, j = 1, ..., p \quad \text{tal que} \quad \beta_j \neq 0.$

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

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se $F_{calc} > f_{\alpha[p,n-(p+1)]}$



Outra formulação do teste F de ajustamento global

Teste F de ajustamento global do modelo RLM (alternativa)

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

Estatística do Teste: $F = \frac{n-(p+1)}{p} \cdot \frac{R^2}{1-R^2} \overset{\cdot}{\sim} F_{[p,n-(p+1)]} \quad \text{se } H_0.$

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

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

A hipótese nula $H_0 : \mathscr{R}^2 = 0$ afirma que, na população, o coeficiente de determinação é nulo.

Exemplo inferência RLM: dados Brix

Eis uma RL Múltipla da variável `Brix` sobre todas as restantes:

```
> brix.lm <- lm(Brix ~ . , data=brix)        <- note-se o uso do '.'
> summary(brix.lm)
[...]
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.08878    1.00252   6.073 0.000298 ***
Diametro     1.27093    0.51219   2.481 0.038030 *
Altura      -0.70967    0.41098  -1.727 0.122478
Peso        -0.20453    0.14096  -1.451 0.184841
pH           0.51557    0.33733   1.528 0.164942
Acucar       0.08971    0.03611   2.484 0.037866 *
--
Residual standard error: 0.1366 on 8 degrees of freedom
Multiple R-squared: 0.8483,  Adjusted R-squared: 0.7534
F-statistic: 8.944 on 5 and 8 DF, p-value: 0.003942
```

A linha final contém a informação do teste F de ajustamento global.

As 2 últimas colunas da tabela `Coefficients` contêm a informação para os testes t (bilaterais) a cada $H_0 : \beta_j = 0$.

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:

$$\hat{\sigma} = \sqrt{QMRE} = \sqrt{\frac{SQRE}{n-(p+1)}}$$

Multiple R-squared: O Coeficiente de Determinação:

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

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

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

O R^2 modificado (adjusted R^2)

O Coeficiente de Determinação usual define-se como:

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

O R^2 modificado, sendo $QMT = \frac{SQT}{n-1} = s_y^2$, é:

$$R^2_{mod} = 1 - \frac{QMRE}{QMT} = 1 - \frac{SQRE}{SQT} \cdot \frac{n-1}{n-(p+1)} = 1 - (1-R^2) \cdot \frac{n-1}{n-(p+1)}.$$

Para qualquer modelo linear (com preditores), tem-se: $R^2_{mod} < R^2$.
Se $n \gg p+1$ (muito mais observações que parâmetros), $R^2 \approx R^2_{mod}$.
Se n é pouco maior que p, $R^2_{mod} \ll R^2$ (excepto se $R^2 \approx 1$).

$\frac{QMRE}{QMT} = \frac{\hat{\sigma}^2}{s_y^2}$ é a proporção da variabilidade total de Y que permanece inexplicada após a introdução dos preditores. Logo, R^2_{mod} é o ganho na explicação de s_y^2 associado ao modelo.

O R^2 modificado (cont.)

Viu-se que o valor de R^2_{mod} penaliza modelos complexos ajustados com poucas observações. dados `brix` ($n=14$ e $p+1=6$).

```
> summary(brix.lm)
[...]
Multiple R-squared:  0.8483, Adjusted R-squared:  0.7534
```

Um submodelo pode ter R^2_{mod} maior que um modelo completo.

O princípio da parcimónia na RLM

Recordemos o princípio da parcimónia na modelação: queremos um modelo que descreva adequadamente a relação entre as variáveis, mas que seja o mais simples (parcimonioso) possível.

Caso se disponha de um modelo de Regressão Linear Múltipla com um ajustamento considerado adequado, a aplicação deste princípio traduz-se em saber se será possível obter um modelo com menos variáveis preditoras, sem perder significativamente em termos de qualidade de ajustamento.

Modelo e Submodelos

Se dispomos de um modelo de Regressão Linear Múltipla, com relação de base

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \ ,$$

chamamos submodelo a um modelo de regressão linear múltipla contendo apenas algumas das variáveis preditoras, e.g.,

$$Y = \beta_0 \qquad + \beta_2 x_2 \qquad\qquad + \beta_5 x_5 \ ,$$

Podemos identificar o submodelo pelo conjunto $\mathscr{S}$ das variáveis preditoras que pertencem ao submodelo. No exemplo, $\mathscr{S} = \{2, 5\}$.

O modelo e o submodelo são idênticos se $\beta_j = 0$ para qualquer variável x_j cujo índice não pertença a $\mathscr{S}$.

Comparando modelo e submodelos

Para comparar um modelo e um seu submodelo (identificado pelo conjunto $\mathscr{S}$ dos índices das suas variáveis), precisamos de optar entre as hipóteses:

$$H_0 : \beta_j = 0, \quad \forall j \notin \mathscr{S} \qquad \text{vs.} \qquad H_1 : \exists j \notin \mathscr{S} \quad \text{tal que} \quad \beta_j \neq 0.$$

[SUBMODELO = MODELO] [SUBMODELO $\neq$ MODELO]

NOTA: Esta discussão só envolve coeficientes β_j de variáveis preditoras ($j > 0$). O coeficiente β_0 faz sempre parte dos submodelos e não é relevante do ponto de vista da parcimónia.

Caso não se rejeite H_0, opta-se pelo submodelo (mais parcimonioso).

Caso se rejeite H_0, opta-se pelo modelo completo (ajusta-se significativamente melhor).

Este coeficiente β_0 não é relevante do ponto de vista da parcimónia: a sua presença não implica trabalho adicional de recolha de dados, nem de interpretação do modelo. Apenas permite um melhor ajustamento.

Estatística de teste para comparar modelo/submodelo

A estatística de teste compara as Somas de Quadrados Residuais do:

- modelo completo (referenciado pelo índice C); e do
- submodelo (referenciado pelo índice S)

Seja k o número de preditores do submodelo ($k+1$ parâmetros). Tem-se, sob H_0 ($\beta_j = 0$, para todas as variáveis x_j que não estão no submodelo):

$$F = \frac{\frac{SQRE_S - SQRE_C}{p-k}}{\frac{SQRE_C}{n-(p+1)}} \quad \frown \quad F_{[p-k,\, n-(p+1)]} \, ,$$

Nota: Necessariamente $SQRE_s \geq SQRE_c$.

São os valores grandes da estatística que levantam dúvidas sobre H_0.

O teste a um submodelo (teste F parcial)

Teste F de comparação dum modelo com um seu submodelo

Dado o Modelo de Regressão Linear Múltipla,

Hipóteses:

$$H_0 : \beta_j = 0, \quad \forall j \notin \mathscr{S} \quad \text{vs.} \quad H_1 : \exists j \notin \mathscr{S} \quad \text{tal que} \quad \beta_j \neq 0.$$

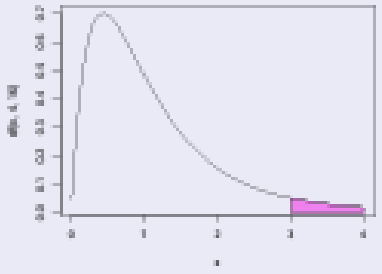
Estatística do Teste:

$$F = \frac{\frac{SQRE_S - SQRE_C}{p-k}}{\frac{SQRE_C}{n-(p+1)}} \quad \frown \quad F_{[p-k,\,n-(p+1)]}, \text{ sob } H_0.$$

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se $F_{calc} > f_{\alpha[p-k,\,n-(p+1)]}$



Expressão alternativa para a estatística do teste

A estatística do teste F parcial pode ser escrita na forma alternativa:

$$F = \frac{n-(p+1)}{p-k} \cdot \frac{R_C^2 - R_S^2}{1-R_C^2} .$$

NOTA: A Soma de Quadrados Total apenas depende dos valores observados da variável resposta Y e não do modelo ajustado. Assim, SQT é igual no modelo completo e no submodelo.

As hipóteses do teste também se podem escrever como

$$H_0 : \mathscr{R}_C^2 = \mathscr{R}_S^2 \qquad \text{vs.} \qquad H_1 : \mathscr{R}_C^2 > \mathscr{R}_S^2 ,$$

A hipótese H_0 indica que o grau de relacionamento linear entre Y e o conjunto dos preditores é idêntico no modelo e no submodelo.

Caso não se rejeite H_0, opta-se pelo submodelo (mais parcimonioso).
Caso se rejeite H_0, opta-se pelo modelo completo (ajusta-se significativamente melhor).

Teste F parcial: formulação alternativa

Teste F de comparação dum modelo com um seu submodelo

Dado o Modelo de Regressão Linear Múltipla,

Hipóteses:

$$H_0 : \mathscr{R}^2_C = \mathscr{R}^2_S \quad \text{vs.} \quad H_1 : \mathscr{R}^2_C > \mathscr{R}^2_S \, .$$

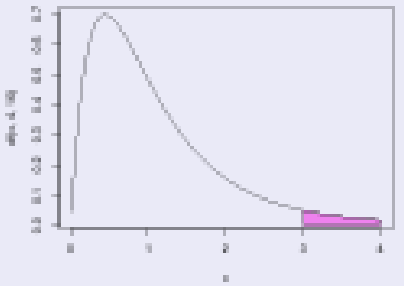
Estatística do Teste:

$$F = \frac{n-(p+1)}{p-k} \cdot \frac{R^2_C - R^2_S}{1-R^2_C} \quad \overset{\cdot}{\sim} \quad F_{[p-k,\, n-(p+1)]}, \text{ sob } H_0.$$

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se $F_{calc} > f_{\alpha[p-k,\, n-(p+1)]}$



O teste a submodelos no R

A informação necessária para um teste F parcial obtem-se através da função `anova`, com dois argumentos: os objectos `lm` resultantes de ajustar o modelo completo e o submodelo sob comparação.

Nos exemplos dos lírios, temos:

```
> anova(iris.lm, iris2.lm)

Analysis of Variance Table
Model 1: Petal.Width ~ Petal.Length
Model 2: Petal.Width ~ Petal.Length + Sepal.Length + Sepal.Width

  Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
1    148 6.3101
2    146 5.3803   2    0.9298 12.616 8.836e-06 ***
```

Os valores $R^2_s = 0.9271$ e $R^2_c = 0.9379$ dos modelos `iris.lm` e `iris2.lm` são significativamente diferentes.

Relações dos testes F parcial

O teste de ajustamento global é equivalente a um teste F parcial comparando um modelo linear e o submodelo nulo (sem preditores).

Caso o modelo e submodelo difiram num único preditor, X_j, o teste F parcial é equivalente ao teste t com as hipóteses $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$.

Nesse caso, não apenas as hipóteses dos dois testes são iguais, como a estatística do teste F parcial é o quadrado da estatística do teste t referido.

- as hipóteses dos dois testes são iguais ($H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$);
- a estatística do teste F parcial é o quadrado da estatística do teste t referido:

$$F_{calc} = T^2_{calc}$$

Tem-se $p - k = 1$, e como é sabido, se uma variável aleatória T tem distribuição t_ν, então o seu quadrado, T^2 tem distribuição $F_{1,\nu}$.

Numa regressão linear simples, o teste t ao declive da recta ser nulo é equivalente ao teste F de ajustamento global. A segunda destas estatística de teste é o quadrado da primeira.

Como escolher um submodelo?

O teste F parcial (teste aos modelos encaixados) permite-nos optar entre um modelo e um seu submodelo. Por vezes, um submodelo pode ser sugerido por:

- razões de índole teórica, sugerindo que determinadas variáveis preditoras não sejam, na realidade, importantes para influenciar os valores de Y.
- razões de índole prática, como a dificuldade, custo ou volume de trabalho associado à recolha de observações para determinadas variáveis preditoras.

Nestes casos, pode ser claro que submodelo(s) se deseja testar.

Como escolher um submodelo? (cont.)

Mas em muitas situações não é evidente qual o subconjunto de variáveis preditoras que se deseja considerar no submodelo. Pretende-se apenas ver se o modelo é simplificável. Nestes casos, a opção por um submodelo não é um problema fácil.

Dadas p variáveis preditoras, o número de subconjuntos, de qualquer cardinalidade, excepto 0 (modelo nulo) e p (o modelo completo) que é possível escolher é dado por $2^p - 2$. A tabela seguinte indica o número desses subconjuntos para $p = 5, 10, 15, 20, 30$.

p	$2^p - 2$
5	30
10	1 022
15	32 766
20	1 048 574
30	1 073 741 822

Para valores de p pequenos, é possível analisar todos os possíveis subconjuntos. Com algoritmos e rotinas informáticas adequadas, a pesquisa completa de todos os possíveis subconjuntos ainda é possível para valores grandes de p (até $p \approx 35$). Mas para p muito grande, uma pesquisa completa é computacionalmente inviável.

Não é legítimo optar pela exclusão de várias variáveis preditoras em simultâneo, com base nos testes t à significância de cada coeficiente β_j no modelo completo.

De facto, os testes t aos coeficientes β_j admitem que todas as restantes variáveis pertencem ao modelo. A exclusão de um qualquer preditor altera o ajustamento: altera os valores estimados b_j e os respectivos erros padrão das variáveis que permanecem no submodelo. Pode acontecer que um preditor seja dispensável num modelo completo, mas deixe de o ser num submodelo, ou viceversa.

Um exemplo:

Há três preditores cuja exclusão individual é admissível (com $\alpha = 0.05$):

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.08878    1.00252   6.073 0.000298 ***
Diametro      1.27093    0.51219   2.481 0.038030 *
Altura       -0.70967    0.41098  -1.727 0.122478
Peso         -0.20453    0.14096  -1.451 0.184841
pH            0.51557    0.33733   1.528 0.164942
Acucar        0.08971    0.03611   2.484 0.037866 *
```

Mas não é legítimo concluir que *Altura*, *Peso* e *pH* são dispensáveis em conjunto.

```
> anova(brix2.lm,brix.lm)
Analysis of Variance Table
Model 1: Brix ~ Diametro + Acucar
Model 2: Brix ~ Diametro + Altura + Peso + pH + Acucar
  Res.Df     RSS Df Sum of Sq    F  Pr(>F)
1     11 0.42743
2      8 0.14925  3   0.27818 4.97 0.03104 *
```

Algoritmos de pesquisa sequenciais

Caso não esteja disponível *software* apropriado, ou se o número *p* de preditores for demasiado grande, pode recorrer-se a algoritmos de pesquisa que simplificam uma regressão linear múltipla sem analisar todo os possíveis submodelos e sem a garantia de obter os melhores subconjuntos.

Vamos considerar um algoritmo que, em cada passo, exclui uma variável preditora, até alcançar uma condição de paragem considerada adequada, ou seja, um algoritmo de exclusão sequencial (*backward elimination*).

Existem variantes deste algoritmo, não estudadas aqui:

- algoritmo de inclusão sequencial (*forward selection*).
- algoritmos de exclusão/inclusão alternada (*stepwise selection*).

O algoritmo de exclusão sequencial com testes aos β_j

1. ajustar o modelo completo, com os p preditores;
2. definir um nível de significância α para os testes de hipóteses a $\beta_j = 0$;
3. para todas as variáveis rejeita-se $H_0 : \beta_j = 0$?
   - Se sim: não é possível simplificar o modelo (passar ao ponto 4).
   - Se não: variáveis em que não se rejeita H_0 são dispensáveis (candidatas à exclusão).
     - ⋆ se apenas existe uma candidata a sair, excluir essa variável;
     - ⋆ se existir mais do que uma variável candidata a sair, excluir a variável associada ao maior *p-value* (isto é, ao valor da estatística t mais próxima de zero)

   Reajustar o modelo após a exclusão da variável e repetir este ponto 3
4. Quando não existirem variáveis candidatas a sair, ou quando sobrar um único preditor, o algoritmo pára. Tem-se então o submodelo final.

Um exemplo – Exercício RLM 2

Dados `brix`: algoritmo de exclusão sequencial

Fixando o nível de significância $\alpha = 0.05$:

```
> summary(lm(Brix ~ Diametro + Altura + Peso + pH + Acucar, data=brix))
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.08878    1.00252   6.073 0.000298 ***
Diametro     1.27093    0.51219   2.481 0.038030 *
Altura      -0.70967    0.41098  -1.727 0.122478
Peso        -0.20453    0.14096  -1.451 0.184841
pH           0.51557    0.33733   1.528 0.164942
Acucar       0.08971    0.03611   2.484 0.037866 *

> summary(lm(Brix ~ Diametro + Altura + pH + Acucar, data=brix))
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.25964    1.05494   5.934 0.000220 ***
Diametro     1.40573    0.53373   2.634 0.027189 *
Altura      -1.06413    0.35021  -3.039 0.014050 *    <-   Passou a ser significativo (0.05)
pH           0.33844    0.33322   1.016 0.336316
Acucar       0.08481    0.03810   2.226 0.053031 .    <-   Deixou de ser significativo (0.05)

> summary(lm(Brix ~ Diametro + Altura + Acucar, data=brix))
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.97183    0.78941   8.832  4.9e-06 ***
Diametro     1.57932    0.50642   3.119  0.01090 *
Altura      -1.11589    0.34702  -3.216  0.00924 **
Acucar       0.09039    0.03776   2.394  0.03771 *    <-   Voltou a ser significativo (0.05)
```

O algoritmo pára aqui. Pode comparar-se o submodelo final com o modelo completo, através dum teste F parcial.

Critério de Informação de Akaike

O R disponibiliza funções para automatizar pesquisas sequenciais de submodelos, semelhantes à que aqui foi enunciada, mas em que o critério de exclusão duma variável em cada passo se baseia no Critério de Informação de Akaike (AIC).

Critério de Informação de Akaike (AIC)

O AIC é uma medida geral da qualidade de ajustamento de modelos. No contexto duma Regressão Linear Múltipla com k variáveis preditoras, define-se como

$$AIC = n \cdot \ln\left(\frac{SQRE_k}{n}\right) + 2(k+1) .$$

Nota: O AIC pode tomar valores negativos.

Interpretando o AIC

$$AIC = n \cdot \ln\left(\frac{SQRE_k}{n}\right) + 2(k+1)$$

- a primeira parcela é função crescente de $SQRE_k$, i.e., quanto melhor o ajustamento, mais pequena a primeira parcela;
- a segunda parcela mede a complexidade do modelo ($k+1$ é o número de parâmetros), pelo que quanto mais parcimonioso o modelo, mais pequena a segunda parcela.

Assim, o AIC depende simultaneamente da qualidade do ajustamento e da simplicidade do modelo.

Um modelo para a variável resposta Y é considerado melhor que outro se tiver um AIC menor (quando ajustados com os mesmos dados).

Algoritmo de exclusão sequencial com base no AIC

Pode definir-se um algoritmo de exclusão sequencial, com base no critério AIC:

- ajustar o modelo completo e calcular o respectivo AIC.
- ajustar cada submodelo com menos uma variável e calcular o respectivo AIC.
- Se nenhum dos AICs dos submodelos considerados for inferior ao AIC do modelo anterior, o algoritmo termina sendo o modelo anterior o modelo final.
  Caso alguma das exclusões reduza o AIC, efectua-se a exclusão que mais reduz o AIC e regressa-se ao ponto anterior.

Algoritmos de exclusão sequencial no R

A função `step` corre o algoritmo de exclusão sequencial, com base no AIC.

```
> brix.lm <- lm(Brix ~ Diametro + Altura + Peso + pH + Acucar, data=brix)
> step(brix.lm, dir="backward")
Start:  AIC=-51.58                    <-- AIC negativo
Brix ~ Diametro + Altura + Peso + pH + Acucar
           Df Sum of Sq     RSS      AIC
<none>                   0.14925 -51.576  <-- modelo original com AIC menor
- Peso      1  0.039279 0.18853 -50.306  <-- modelo sem Peso em 2o. lugar
- pH        1  0.043581 0.19284 -49.990
- Altura    1  0.055631 0.20489 -49.141
- Diametro  1  0.114874 0.26413 -45.585
- Acucar    1  0.115132 0.26439 -45.572
```

Os vários modelos ensaiados são ordenados por ordem crescente de AIC. Neste caso, não se exclui qualquer variável: o AIC do modelo inicial é inferior ao de qualquer submodelo resultante de excluir uma variável. O submodelo final é o modelo inicial.

As duas variantes dos algoritmos

Os algoritmos de exclusão sequencial baseados nos testes *t* ou no AIC coincidem nas variáveis a excluir, podendo diferir apenas no momento de paragem.

Em geral, um algoritmo de exclusão sequencial baseado no AIC é mais cauteloso na exclusão, sobretudo se o valor de α usado nos testes *t* for baixo. Nos algoritmos baseados nos testes *t*, é aconselhável usar valores mais elevados de α, como $\alpha = 0.10$.

Um algoritmo de exclusão sequencial não garante a identificação do "*melhor* submodelo" com um dado número de preditores. Apenas identifica, de forma computacionalmente ligeira, submodelos "bons".

Deve ser usado com bom senso e o submodelo obtido cruzado com outras considerações (e.g., o custo ou dificuldade de obtenção de cada variável, ou o papel que a teoria relativa ao problema em questão reserva a cada preditor).

A Validação do Modelo (análise dos resíduos)

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

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

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

A análise de Resíduos e outros diagnósticos

Uma análise de regressão linear não fica completa sem o estudo dos resíduos e de alguns outros diagnósticos.

O modelo linear admite que $\varepsilon_i \frown \mathcal{N}(0, \sigma^2) \quad \forall i = 1, ..., n.$

Sob o modelo linear, os resíduos têm a seguinte distribuição:

$$E_i \frown \mathcal{N}\left(0, \sigma^2(1 - h_{ii})\right) \qquad \forall i = 1, ..., n\,,$$

sendo h_{ii} o i-ésimo elemento diagonal da matriz $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ de projecção ortogonal sobre o subespaço $\mathscr{C}(\mathbf{X})$.

Este resultado demonstra-se mais facilmente considerando o vector dos resíduos, $\vec{\mathbf{E}} = \vec{\mathbf{Y}} - \hat{\vec{\mathbf{Y}}} = \vec{\mathbf{Y}} - \mathbf{H}\vec{\mathbf{Y}} = (\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}$.

Propriedades dos Resíduos sob o modelo linear

Teorema (Distribuição dos Resíduos no Modelo Linear)

Dado o Modelo Linear, tem-se:

$$\vec{\mathbf{E}} \frown \mathcal{N}_n\left(\vec{\mathbf{0}}, \sigma^2(\mathbf{I}_n - \mathbf{H})\right) \qquad \text{sendo} \quad \vec{\mathbf{E}} = (\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}} .$$

Como no Modelo Linear $\vec{\mathbf{Y}} \frown \mathcal{N}(\mathbf{X}\vec{\boldsymbol{\beta}}, \sigma^2\mathbf{I}_n)$, o vector dos resíduos $\vec{\mathbf{E}} = (\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}$, tem distribuição Multinormal em sentido generalizado

- $E[\vec{\mathbf{E}}] = E[(\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}] = (\mathbf{I}_n - \mathbf{H})E[\vec{\mathbf{Y}}] = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\vec{\boldsymbol{\beta}} = \vec{\mathbf{0}}$,
  pois $\mathbf{X}\vec{\boldsymbol{\beta}} \in \mathscr{C}(\mathbf{X})$, logo permanece invariante sob a projecção: $\mathbf{H}\mathbf{X}\vec{\boldsymbol{\beta}} = \mathbf{X}\vec{\boldsymbol{\beta}}$.
- Pelas propriedades do acetato 126 e o facto de $\mathbf{H}$ ser simétrica ($\mathbf{H}^t = \mathbf{H}$) e idempotente ($\mathbf{H}\mathbf{H} = \mathbf{H}$), tem-se:
  $V[\vec{\mathbf{E}}] = V[(\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}] = (\mathbf{I}_n - \mathbf{H})V[\vec{\mathbf{Y}}](\mathbf{I}_n - \mathbf{H})^t = \sigma^2 \cdot (\mathbf{I}_n - \mathbf{H})$.

Propriedades dos Resíduos no Modelo Linear (cont.)

Nota: Embora no modelo RL os erros aleatórios sejam independentes, os resíduos não são variáveis aleatórias independentes, pois as covariâncias entre resíduos diferentes são (em geral), não nulas:

$$cov(E_i, E_j) = -\sigma^2 \cdot h_{ij} \ , \quad \text{se } i \neq j \ ,$$

onde h_{ij} indica o elemento da linha i e coluna j da matriz $\mathbf{H}$.

Se $\vec{\mathbf{E}} \frown \mathcal{N}_n\left(\vec{\mathbf{0}}, \sigma^2(\mathbf{I}_n - \mathbf{H})\right)$, então cada resíduo tem distribuição:

$$E_i \frown \mathcal{N}\left(0 \ , \ \sigma^2(1 - h_{ii})\right) \ ,$$

onde h_{ii} é o i-ésimo elemento diagonal de $\mathbf{H}$ e

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

Dois tipos de resíduos

Como $\frac{E_i}{\sqrt{\sigma^2(1-h_{ii})}} \frown \mathcal{N}(0,1)$, definem-se resíduos normalizados:

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

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

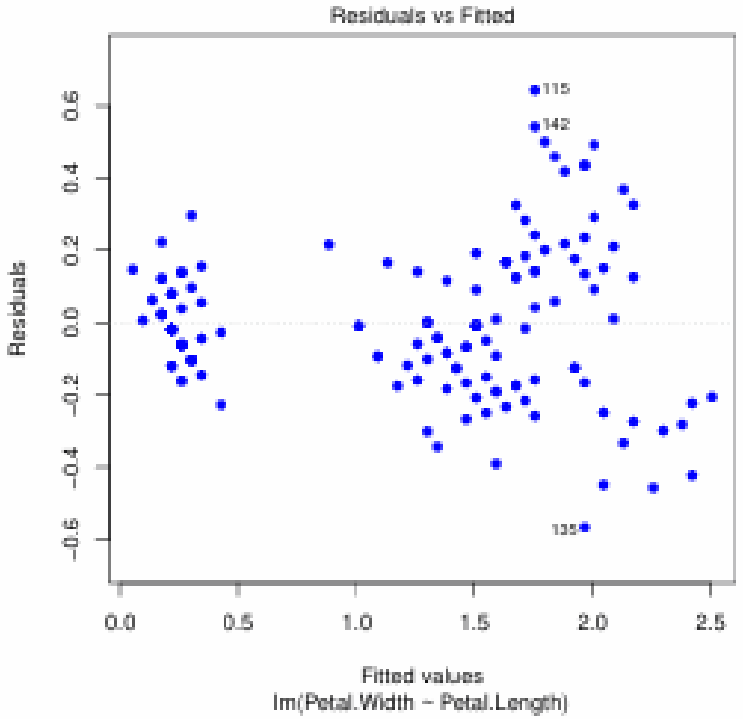
Para grandes amostras, os R_i são aproximadamente $\mathcal{N}(0,1)$.

A função `rstandard` calcula resíduos standardizados (R_i).

Nas regressões lineares, avalia-se a validade dos pressupostos do modelo através de gráficos de resíduos. Não se efectuam testes de Normalidade, já que os resíduos não são (em geral) independentes.

Validação do modelo: (1) Gráficos de resíduos vs. $\hat{Y}_i$

Gráfico indispensável: Resíduos (usuais) vs. Valores ajustados de Y.



- Os resíduos devem estar aproximadamente numa banda horizontal em torno de zero.
- Não deve existir qualquer padrão aparente. Sendo válido o Modelo RL, $cor(E_i, \hat{Y}_i) = 0$.

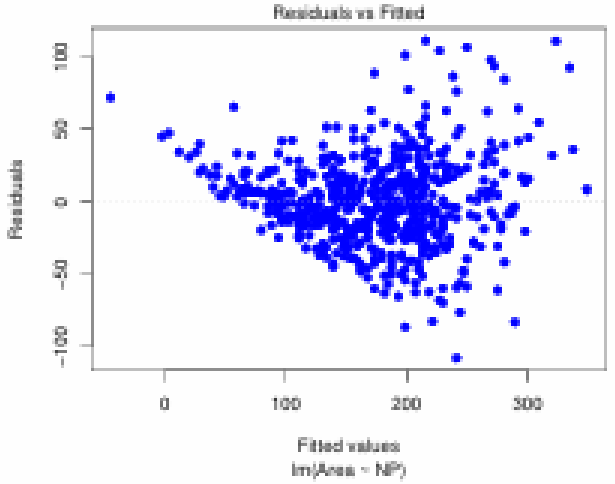
Possíveis padrões indicativos de problemas

Num gráfico de E_i vs. $\hat{Y}_i$ podem surgir padrões problemáticos:

Curvatura na disposição dos resíduos Indica violação da hipótese de linearidade entre *y* e os preditores.

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

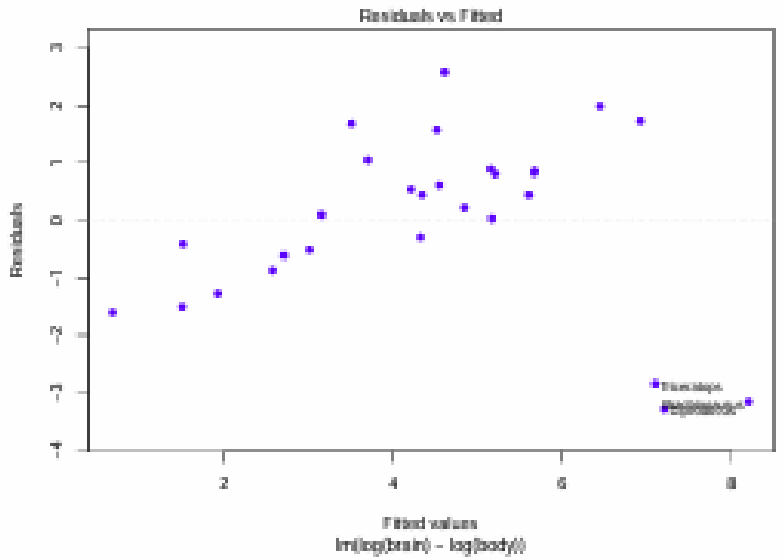
Um ou mais resíduos muito destacados Indica a existência de observações atípicas.



Um exemplo de resíduos em forma de funil, e sugerindo alguma curvatura na relação entre as duas variáveis (dados das folhas de videira `Area` vs. `NP`

Padrões indicativos de problemas (cont.)

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



A presença dos dinossáurios nos dados `Animals` cria uma banda oblíqua de pontos no gráfico E_i vs. $\hat{Y}_i$.

Validação do modelo: (2) Gráficos para avaliar a Normalidade

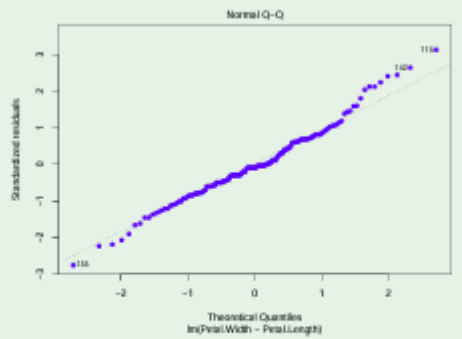
para grandes amostras os resíduos estandardizados $R_i = \frac{E_i}{\sqrt{QMRE \cdot (1-h_{ii})}}$, devem ser aproximadamente $\mathcal{N}(0,1)$

O pressuposto de erros aleatórios Normais pode ser validado com:

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

Um qq-plot concordante com a hipótese de Normalidade dos erros aleatórios deverá apresentar colinearidade aproximada. O exemplo seguinte sugere algum desvio à Normalidade para os resíduos mais extremos.

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



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

Validação do modelo: (3) Gráficos para avaliar independência

Dependência entre erros aleatórios pode surgir como resultado de:

- correlação cronológica;
- correlação espacial.

Nesse caso, pode ser útil inspeccionar gráficos de resíduos vs. ordem de observação ou distribuição espacial dos resíduos, para verificar se existem padrões que sugiram falta de independência. Nesse caso, modelos alternativos para series temporais ou dados espaciais podem ser necessários.

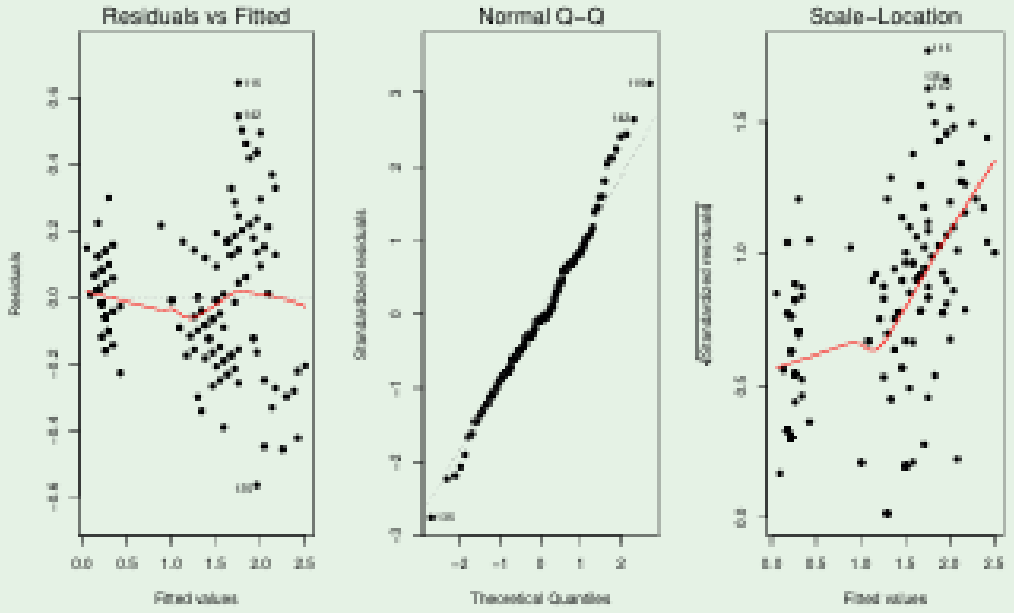
Validação do modelo: (4) Gráficos de resíduos vs. preditores

A presença de não-linearidade em gráficos de resíduos vs. preditores individuais pode sugerir a necessidade de transformações desses preditores.

Estudo de resíduos no R

O comando `plot`, aplicado a um objecto `lm` produz até seis gráficos de resíduos e diagnósticos. Os três primeiros correspondem a gráficos de resíduos. Para o exemplo dos lírios:

```
> plot(iris.lm, which=1:3, pch=16)
```



O terceiro gráfico (argumento `which=3`) é de $\sqrt{|R_i|}$ *vs.* $\hat{Y}_i$.

Observações atípicas

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

Observações atípicas (*outliers* in English). Conceito sem definição rigorosa, procura designar observações que se distanciam da relação linear de fundo entre Y e as variáveis preditoras.

Muitas vezes surgem associadas a resíduos grandes (em módulo). Em particular, e como os resíduos 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"

Define-se o valor do efeito alavanca (*leverage*) da i-ésima observação como sendo o i-ésimo valor diagonal da matriz $\mathbf{H}$: $h_{ii} = \mathbf{H}_{(i,i)}$.

Como $\hat{\vec{\mathbf{Y}}} = \mathbf{H}\vec{\mathbf{Y}}$, tem-se $\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j$ (cada valor ajustado é combinação linear dos valores observados). O efeito alavanca h_{ii} é a ponderação associada a y_i na definição do valor ajustado correspondente, $\hat{y}_i$. Não deveria ser excessivo.

Observações alavanca (*leverage points*) são observações com h_{ii} elevado, que tendem a "atrair" a hipersuperfície ajustada numa regressão.

Como $V[E_i] = \sigma^2(1 - h_{ii})$, se h_{ii} é elevado, a variância do resíduo E_i é baixa e o resíduo tende a estar perto da sua média (zero). Ou seja, a superfície ajustada tende a passar próximo desse ponto.

Observações alavanca (cont.)

Verifica-se, para qualquer observação:

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

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

$$\overline{h} = \frac{p+1}{n} \, ,$$

Logo, quanto mais observações, menor o efeito alavanca médio.

Observações alavanca (cont.)

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 (outliers).

Efeito alavanca numa regressão linear Simples

Numa regressão linear simples, tem-se

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

Assim, numa RLS, o efeito alavanca da observação i depende do valor x_i em relação à média $\overline{x}$: quanto maior $(x_i - \overline{x})^2$, maior h_{ii}. O maior efeito alavanca tem de pertencer a uma das duas observações mais extrema em x.

Numa regressão linear múltipla, os maiores efeitos alavanca correspondem às observações em que os valores dos preditores estão mais afastados do vector das médias dos preditores.

Observações influentes

Observações influentes são observações que, se retiradas da análise, gerariam variações assinaláveis no conjunto dos valores ajustados de Y e nos parâmetros ajustados, b_j.

Medida de influência frequente é a distância de Cook, definida como:

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

sendo $\hat{y}_{[-i]_j}$ o valor ajustado da observação i, obtido estimando os β_js sem a observação i. Expressão equivalente é:

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

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

É frequente considerar $D_i > 0.5$ como limiar de observação influente.

Uma prevenção

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

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

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

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

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

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

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

Diagnósticos no R

`hatvalues` calcula efeitos alavanca (h_{ii}) e `cooks.distance` as D_i.

```
> brix.diagn <- cbind(hatvalues(brix.lm), cooks.distance(brix.lm))
> colnames(brix.diagn) <- c("h_ii", "Di")
> brix.diagn
```

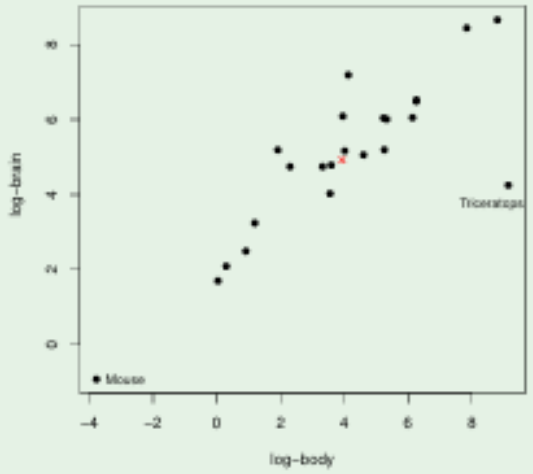
	h_ii	Di
1	0.6231274	0.6209707369
2	0.3576171	0.0969006496
3	0.4750339	0.0380279990
4	0.2881782	0.0186723249
5	0.3751686	0.0351359851
6	0.2985676	0.0354362871
7	0.5260699	0.0793008032
8	0.4955231	0.0304136309
9	0.2809899	0.2009993314
10	0.2268779	0.0002254622
11	0.2757540	0.0108143657
12	0.4771373	0.0092558438
13	0.6609377	1.5222084206
14	0.6390174	1.0769004225

Alguns valores muito elevados reflectem um conjunto de dados pequeno ($n=14$) com um modelo pesado ($p=5$). O efeito alavanca médio é $\overline{h}=\frac{p+1}{n}=0.4286$.

Um exemplo na RLS

Considerando apenas um subconjunto das espécies, obtém-se o seguinte gráfico de log-peso do cérebro *vs.* log-peso do corpo:

```
> library(MASS)
> animaissub <- Animals[-c(6,19,25,26,27),]
> plot(log(brain) ~ log(body) , data=animaissub, pch=16)
```



Eis os correspondentes resíduos (internamente) estandardizados, distâncias de Cook e valores do efeito alavanca:

```
                  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 R

A função `plot`, aplicada a um objecto `lm` produz, além dos gráficos vistos no acetato 209, gráficos com alguns dos diagnósticos agora considerados.

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

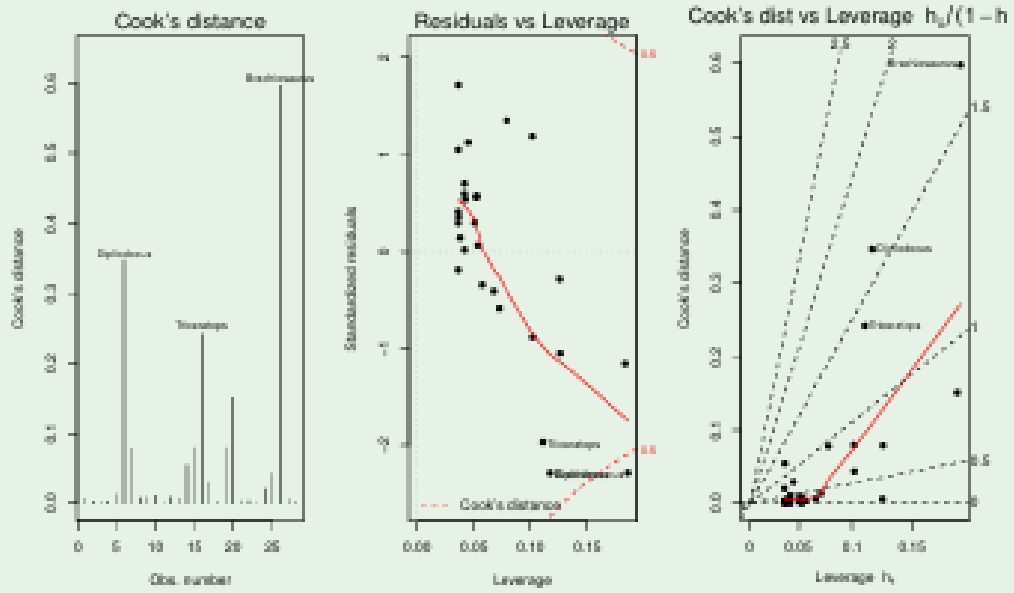
A opção `which=5` produz um gráfico de Resíduos estandardizados (R_is) no eixo vertical contra valores de h_{ii} (*leverages*) no eixo horizontal, traçando linhas de igual distância de Cook (para os níveis 0.5 e 1, por omissão), que destacam eventuais observações influentes.

A opção `which=6` produz um gráfico de distâncias de Cook (eixo vertical) contra valores de $\frac{h_{ii}}{1-h_{ii}}$, com isolinhas de resíduos estandardizados R_i

Um exemplo de gráficos de diagnóstico

Eis estes gráficos de diagnóstico, para os dados `Animals`

```
> plot(Animals.lm, which=4:6)
```



As distâncias de Cook elevadas reflectem o distanciamento das espécies de dinossáurios da tendência geral das outras espécies. O facto de serem três observações discordantes mitiga um pouco o valor destas distâncias.

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.

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

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

Prevenções sobre transformações

Mas a utilização de transformações de variáveis, sobretudo quando afecta a variável resposta, 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.

Advertências finais

1. Podem surgir problemas associados à (quase) multicolinearidade das variáveis preditoras, ou seja, ao facto das colunas da matriz $\mathbf{X}$ serem (quase) linearmente dependentes:

- podem existir problemas numéricos no cálculo de $(\mathbf{X}^t\mathbf{X})^{-1}$, logo no ajustamento do modelo e na estimação dos parâmetros;
- podem existir variâncias muito grandes de alguns $\hat{\beta}_i$s, o que significa muita instabilidade na inferência.

Multicolinearidade reflecte redundância de informação nos preditores. É possível eliminá-la excluíndo da análise uma ou várias variáveis preditoras que sejam responsáveis pela (quase) dependência linear dos preditores.

Advertências finais (cont.)

2. Não se deve confundir a existência de uma relação linear entre preditores X_1, X_2, ..., X_p e uma variável resposta Y, com uma relação de causa e efeito.

Pode existir uma relação de causa e efeito. Mas pode também verificar-se:

- Uma relação de variação conjunta, mas não de tipo causal (como por exemplo, em muitos conjuntos de dados morfométricos). Por vezes, preditores e variável resposta são todos efeito de causas comuns subjacentes.
- Uma relação espúria, de coincidência numérica.

Uma relação causal só pode ser afirmada com base em teoria própria do fenómeno sob estudo, e não com base na relação linear estabelecida estatisticamente.