O contexto inferencial

Até aqui, apenas se considerou o problema descritivo: dados n conjuntos de observações $\{(x_{1(i)}, x_{2(i)}, ..., x_{p(i)}, y_{(i)})\}_{i=1}^{n}$, determinar os $p+1$ coeficientes $\vec{\mathbf{b}} = (b_0, b_1, b_2, ..., b_p)$ que minimizam a soma de quadrados de resíduos

$$SQRE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_{1(i)} + b_2 x_{2(i)} + ... + b_p x_{p(i)})]^2$$

$$\Rightarrow \quad SQRE \text{ minimo se } \quad \vec{\mathbf{b}} = \left(\mathbf{X}^t \mathbf{X}\right)^{-1} \mathbf{X}^t \vec{\mathbf{y}}.$$

Tal como na RL Simples, coloca-se o problema inferencial quando as n observações são uma amostra aleatória duma população mais vasta. É o hiperplano populacional relacionando Y e as p variáveis preditoras que se pretende conhecer:

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

Para esse fim, será necessário admitir alguns pressupostos adicionais.

O Modelo RLM

Na Regressão Linear Múltipla admite-se que as n observações da variável resposta Y são aleatórias e podem ser modeladas como

$$Y_i = \beta_0 + \beta_1 x_{1(i)} + \beta_2 x_{2(i)} + ... + \beta_p x_{p(i)} + \varepsilon_i \ , \qquad i = 1, ..., n$$

Admitem-se válidos pressupostos semelhantes aos do modelo RLS:

O Modelo da Regressão Linear Múltipla - RLM

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

A constante β_j ($j = 1, 2, ..., p$) que multiplica a variável X_j pode ser interpretada como a variação esperada em Y, associada a aumentar X_j em uma unidade, mantendo as restantes variáveis constantes.

A notação matricial/vectorial

As n equações, para as n observações, podem ser escritas como uma única equação, utilizando notação vectorial/matricial:

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

A notação matricial (cont.)

A equação matricial/vectorial do modelo

As n equações correspondem a uma única equação matricial:

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

onde

$$\vec{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x1_{(1)} & x2_{(1)} & \cdots & xp_{(1)} \\ 1 & x1_{(2)} & x2_{(2)} & \cdots & xp_{(2)} \\ 1 & x1_{(3)} & x2_{(3)} & \cdots & xp_{(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x1_{(n)} & x2_{(n)} & \cdots & xp_{(n)} \end{bmatrix}, \vec{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \vec{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

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

A notação matricial (cont.)

Na equação matricial $\vec{\mathbf{Y}} = \mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\varepsilon}}$, tem-se:

- $\vec{\mathbf{Y}}$ – vector aleatório das n variáveis aleatórias resposta;
- $\mathbf{X}$ – matriz do modelo (não aleatória) de dimensões $n \times (p+1)$ cujas colunas são dadas pelas observações de cada variável preditora (e por uma coluna de uns, associada a constante aditiva do modelo);
- $\vec{\boldsymbol{\beta}}$ – vector (não aleatório) dos $p+1$ parâmetros do modelo;
- $\vec{\boldsymbol{\varepsilon}}$ – vector aleatório dos n erros aleatórios.

Representa-se um vector de n observações de Y por $\vec{\mathbf{y}}$.

Com alguns conceitos adicionais podemos escrever também os pressupostos relativos aos erros aleatórios em notação vectorial/matricial.

Ferramentas para vectores aleatórios

O vector de n componentes $\vec{\mathbf{Y}}$, tal como o vector dos n erros aleatórios, $\vec{\boldsymbol{\varepsilon}}$, constituem vectores aleatórios. Vamos definir ferramentas para trabalhar com vectores aleatórios.

Vector esperado

Para qualquer vector aleatório $\vec{\mathbf{W}} = (W_1, W_2, ..., W_k)^t$, define-se o vector esperado de $\vec{\mathbf{W}}$, constituído pelos valores esperados de cada componente:

$$E[\vec{\mathbf{W}}] = \begin{bmatrix} E[W_1] \\ E[W_2] \\ \vdots \\ E[W_k] \end{bmatrix}.$$

Ferramentas para vectores aleatórios (cont.)

Matriz de variâncias-covariâncias de $\vec{\mathbf{W}}$

É a matriz $k \times k$ cujos elementos são as (co-)variâncias de cada par de componentes:

$$V[\vec{\mathbf{W}}] = \begin{bmatrix} V[W_1] & C[W_1, W_2] & C[W_1, W_3] & \dots & C[W_1, W_k] \\ C[W_2, W_1] & V[W_2] & C[W_2, W_3] & \dots & C[W_2, W_k] \\ C[W_3, W_1] & C[W_3, W_2] & V[W_3] & \dots & C[W_3, W_k] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C[W_k, W_1] & C[W_k, W_2] & C[W_k, W_3] & \dots & V[W_k] \end{bmatrix}$$

É necessariamente uma matriz simétrica.

Propriedades do vector esperado

Propriedades de vectores esperados

Tal como para o caso de variáveis aleatórias, também o vector esperado de um vector aleatório $\vec{\mathbf{W}}_{k\times 1}$ tem propriedades simples:

- Se b é um escalar não aleatório, $E[b\vec{\mathbf{W}}] = b\,E[\vec{\mathbf{W}}]$.
- Se $\vec{\mathbf{a}}_{k\times 1}$ é um vector não aleatório, $E[\vec{\mathbf{W}} + \vec{\mathbf{a}}] = E[\vec{\mathbf{W}}] + \vec{\mathbf{a}}$.
- Se $\vec{\mathbf{a}}_{k\times 1}$ é um vector não aleatório, $E[\vec{\mathbf{a}}^t\,\vec{\mathbf{W}}] = \vec{\mathbf{a}}^t\,E[\vec{\mathbf{W}}]$.
- Se $\mathbf{B}_{m\times k}$ é uma matriz não aleatória, $E[\mathbf{B}\vec{\mathbf{W}}] = \mathbf{B}\,E[\vec{\mathbf{W}}]$.

Também o vector esperado da soma de dois vectors aleatórios tem uma propriedade operatória simples:

- Se $\vec{\mathbf{W}}_{k\times 1}$, $\vec{\mathbf{U}}_{k\times 1}$ são vectores aleatórios, $E[\vec{\mathbf{W}} + \vec{\mathbf{U}}] = E[\vec{\mathbf{W}}] + E[\vec{\mathbf{U}}]$.

Propriedades da matriz de (co)variâncias

Propriedades de matrizes de (co-)variâncias

Eis algumas propriedades operatórias das matrizes de variâncias-covariâncias de vectores aleatórios:

- Se b é um escalar não aleatório, $V[b\vec{\mathbf{W}}] = b^2\, V[\vec{\mathbf{W}}]$.
- Se $\vec{\mathbf{a}}_{k\times 1}$ é um vector não aleatório, $V[\vec{\mathbf{W}} + \vec{\mathbf{a}}] = V[\vec{\mathbf{W}}]$.
- Se $\vec{\mathbf{a}}_{k\times 1}$ é um vector não aleatório, $V[\vec{\mathbf{a}}^t\, \vec{\mathbf{W}}] = \vec{\mathbf{a}}^t\, V[\vec{\mathbf{W}}]\, \vec{\mathbf{a}}$.
- Se $\mathbf{B}_{m\times k}$ é uma matriz não aleatória, $V[\mathbf{B}\vec{\mathbf{W}}] = \mathbf{B}\, V[\vec{\mathbf{W}}]\, \mathbf{B}^t$.

A matriz de variâncias-covariâncias da soma de dois vectores aleatórios tem uma propriedade operatória simples se os vectores aleatórios forem independentes:

- Se $\vec{\mathbf{W}}_{k\times 1}$ e $\vec{\mathbf{U}}_{k\times 1}$ forem vectores aleatórios independentes, $V[\vec{\mathbf{W}} + \vec{\mathbf{U}}] = V[\vec{\mathbf{W}}] + V[\vec{\mathbf{U}}]$.

A distribuição Normal Multivariada

Vectores aleatórios têm também distribuições (multivariadas) de probabilidades. Para vectores aleatórios contínuos $\vec{\mathbf{W}}_{k\times 1}$, a distribuição pode ser caracterizada por uma função densidade conjunta $f : \mathbb{R}^k \to \mathbb{R}$. A mais frequente distribuição multivariada para vectores aleatórios é a Multinormal:

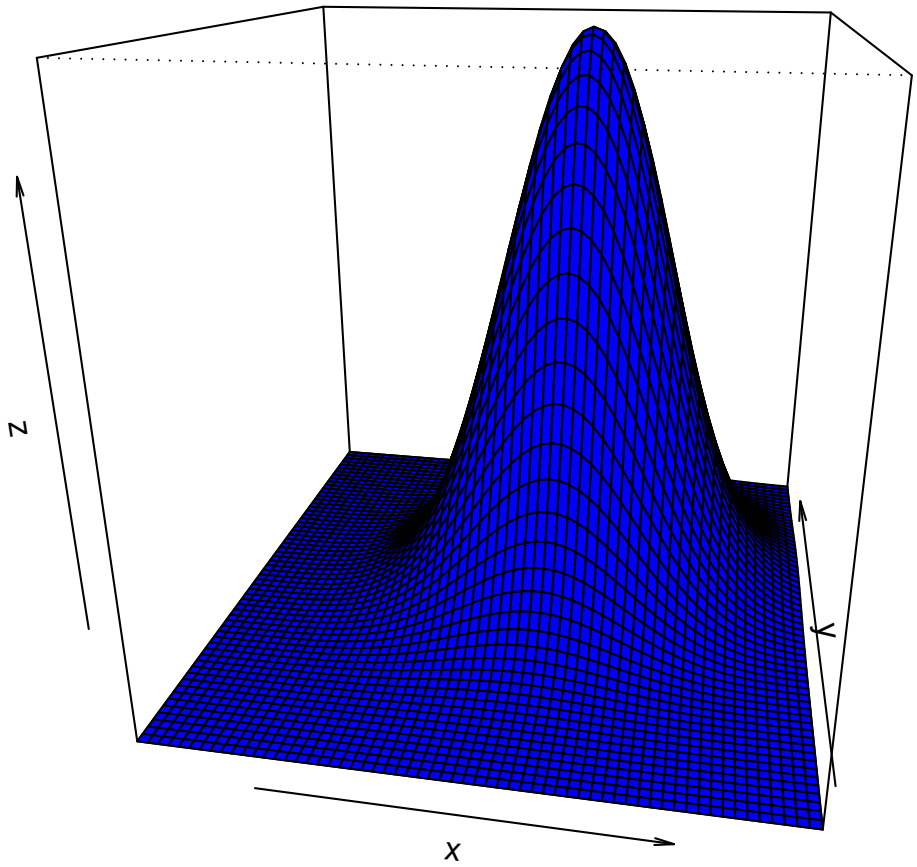
Distribuição Normal Multivariada

O vector aleatório k-dimensional $\vec{\mathbf{W}}$ tem distribuição Multinormal, com parâmetros dados pelo vector $\vec{\boldsymbol{\mu}}$ e a matriz $\boldsymbol{\Sigma}$ se a sua função densidade conjunta for:

$$f(\vec{\mathbf{w}}) = \frac{1}{(2\pi)^{k/2}\sqrt{det(\boldsymbol{\Sigma})}} \mathrm{e}^{-\frac{1}{2}(\vec{\mathbf{w}}-\vec{\boldsymbol{\mu}})^t \boldsymbol{\Sigma}^{-1}(\vec{\mathbf{w}}-\vec{\boldsymbol{\mu}})} \quad , \quad \vec{\mathbf{w}} \in \mathbb{R}^k. \qquad (1)$$

Notação: $\vec{\mathbf{W}} \frown \mathscr{N}_k(\vec{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$.

A densidade Binormal (Multinormal com $k = 2$)



Algumas propriedades da distribuição Multinormal

Propriedades da distribuição Multinormal

Se $\vec{\mathbf{W}} \frown \mathscr{N}_k(\vec{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$:

1. O vector esperado de $\vec{\mathbf{W}}$ é $E[\vec{\mathbf{W}}] = \vec{\boldsymbol{\mu}}$.
2. A matriz de (co)variâncias de $\vec{\mathbf{W}}$ é $V[\vec{\mathbf{W}}] = \boldsymbol{\Sigma}$.
3. Se duas componentes de $\vec{\mathbf{W}}$ têm covariância nula, são independentes: $Cov[W_i, W_j] = 0 \Rightarrow W_i,\ W_j$ independentes.

Nota: Em geral, X, Y *independentes* $\Rightarrow$ $cov[X, Y] = 0$.
Quando a distribuição conjunta de X e Y é Multinormal, tem-se também a implicação contrária.

Nota: Qualquer zero numa matriz de (co)variâncias duma Multinormal indica que as componentes correspondentes são independentes.

Propriedades da Multinormal (cont.)

Propriedades da Multinormal (cont.)

Se $\vec{\mathbf{W}} \frown \mathscr{N}_k(\vec{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$:

4. Qualquer distribuição marginal de $\vec{\mathbf{W}}$ é (multi)normal.
   Em particular, cada componente individual W_i é normal com média μ_i e variância $\boldsymbol{\Sigma}_{(i,i)}$: $W_i \frown \mathscr{N}(\mu_i, \boldsymbol{\Sigma}_{(i,i)})$.
5. Se $\vec{\mathbf{a}}$ um vector não-aleatório $k \times 1$, então $\vec{\mathbf{W}} + \vec{\mathbf{a}} \frown \mathscr{N}_k(\vec{\boldsymbol{\mu}} + \vec{\mathbf{a}}, \boldsymbol{\Sigma})$.
6. Combinações lineares das componentes dum vector multinormal são Normais: $\vec{\mathbf{a}}^t \vec{\mathbf{W}} = a_1 W_1 + a_2 W_2 + \ldots + a_k W_k \frown \mathscr{N}(\vec{\mathbf{a}}^t \vec{\boldsymbol{\mu}}, \vec{\mathbf{a}}^t \boldsymbol{\Sigma} \vec{\mathbf{a}})$.
7. Se $\mathbf{B}$ é matriz $m \times k$ (não aleatória, de característica $m \leq k$), então $\mathbf{B}\vec{\mathbf{W}} \frown \mathscr{N}_m(\mathbf{B}\vec{\boldsymbol{\mu}}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^t)$.

Modelo Regressão Linear Múltipla - versão matricial

O Modelo em notação vectorial/matricial

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 $\vec{\mathbf{0}} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \dots \\ 0 \end{bmatrix}$; $\sigma^2 \mathbf{I}_n = \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ 0 & 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix}$

A condição $\vec{\boldsymbol{\varepsilon}} \frown \mathcal{N}_n(\vec{\mathbf{0}}, \sigma^2 \mathbf{I}_n)$ significa que:

- 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 são independentes, porque $Cov[\varepsilon_i, \varepsilon_j] = 0$ se $i \neq j$, o que, numa Multinormal, implica a independência.

A distribuição do vector $\vec{\mathbf{Y}}$

Primeiras Consequências do Modelo

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

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

Sai directamente, pois $\vec{\mathbf{Y}}$ é da forma $\vec{\mathbf{W}} + \vec{\mathbf{a}}$, soma dum vector Multinormal ($\vec{\mathbf{W}} = \vec{\boldsymbol{\varepsilon}}$) com um vector não aleatório ($\vec{\mathbf{a}} = \mathbf{X}\vec{\boldsymbol{\beta}}$).

Tendo em conta as propriedades da Multinormal:

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

O estimador dos parâmetros do Modelo

Os estimadores $\hat{\beta}_j$ dos parâmetros β_j do modelo obtêm-se adaptando o vector $\vec{\mathbf{b}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{y}}$ (acetato 180) ao contexto inferencial.

Estimador dos parâmetros populacionais

O vector $\vec{\hat{\boldsymbol{\beta}}}$ que estima o vector $\vec{\boldsymbol{\beta}}$ dos parâmetros populacionais é:

$$\vec{\hat{\boldsymbol{\beta}}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\vec{\mathbf{Y}}\,,$$

onde $\mathbf{X}$ é a matriz do modelo e $\vec{\mathbf{Y}}$ o vector das n observações da variável resposta (acetato 196).

Nota: O estimador de β_j está na posição $j+1$ do vector $\vec{\hat{\boldsymbol{\beta}}}$.

Nota: $\vec{\hat{\boldsymbol{\beta}}}$ é da forma $\mathbf{B}\vec{\mathbf{W}}$, com $\mathbf{B} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t$ e $\vec{\mathbf{W}} = \vec{\mathbf{Y}}$ (Multinormal).

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

Distribuição do estimador $\vec{\hat{\boldsymbol{\beta}}}$

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

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

- $E[\vec{\hat{\boldsymbol{\beta}}}] = \vec{\boldsymbol{\beta}}$ e $V[\vec{\hat{\boldsymbol{\beta}}}] = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$.
- 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 $V[\hat{\beta}_j] = \sigma^2\,(\mathbf{X}^t\mathbf{X})^{-1}_{(j+1,j+1)}$. (Note-se o desfasamento nos índices).
- $Cov[\hat{\beta}_i, \hat{\beta}_j] = \sigma^2\,(\mathbf{X}^t\mathbf{X})^{-1}_{(i+1,j+1)}$.
- Estimadores individuais não são (em geral) independentes, porque $(\mathbf{X}^t\mathbf{X})^{-1}$ não é, em geral, uma matriz diagonal.

A distribuição dum estimador individual

Como se viu no acetato anterior, tem-se, $\forall j = 0, 1, ..., p$:

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

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

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

Este resultado generaliza os relativos à Regressão Linear Simples.

O problema de σ^2 desconhecido

O resultado do acetato anterior permitiria construir intervalos de confiança e fazer testes a hipóteses sobre os parâmetros β_j se fosse conhecida a variância σ^2 dos erros aleatórios.

Mas esta é, em geral, desconhecida.

Vai-se, tal como na Regressão Linear Simples:

- obter um estimador para σ^2; e
- ver o que acontece à distribuição do acetato anterior quando σ^2 é substituído pelo seu estimador.

SQRE na Regressão Múltipla

Resultados distribucionais de SQRE

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

- $\frac{SQRE}{\sigma^2} \frown \chi^2_{n-(p+1)}$
- $SQRE$ é independente de $\vec{\hat{\boldsymbol{\beta}}}$.

NOTA: Omite-se a demonstração

Um estimador centrado de σ^2

Dado o Modelo de RLM, $E\left[\frac{SQRE}{n-(p+1)}\right] = \sigma^2$.

NOTA: Os graus de liberdade associados a $SQRE$ são o número de observações (n) menos o número de parâmetros do modelo ($p+1$).

O Quadrado Médio Residual na Regressão Múltipla

Quadrado Médio Residual

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

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

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

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

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

Revisitando o estimador de β_j

Vimos (acetato 210) que para cada estimador $\hat{\beta}_j$ se tem:

$$Z = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 \cdot (\mathbf{X}^t\mathbf{X})^{-1}_{(j+1,j+1)}}} \frown \mathcal{N}(0,1) \ .$$

Temos ainda:

$$W = \frac{SQRE}{\sigma^2} \frown \chi^2_{n-(p+1)} \quad \text{e} \quad Z, W \text{ v.a. independentes} \ .$$

Logo (ver também o acetato 85):

$$\frac{Z}{\sqrt{W/(n-(p+1))}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{QMRE \cdot (\mathbf{X}^t\mathbf{X})^{-1}_{(j+1,j+1)}}} \frown t_{n-(p+1)} \ .$$

Quantidades fulcrais para a inferência sobre β_j

Distribuições para a inferência sobre β_j, $j = 0, 1, ..., p$

Dado o Modelo de Regressão Linear Múltipla, tem-se

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

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

Este resultado serve de base à construção de intervalos de confiança e testes de hipóteses para os parâmetros β_j do modelo populacional.

NOTA: A quantidade fulcral acima é totalmente análoga aos resultados correspondentes na RLS. Assim, os ICs e testes de hipóteses a parâmetros individuais, na RLM, serão análogos aos da RLS.

Intervalo de confiança para β_j

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

Dado o Modelo de Regressão Linear Múltipla, um intervalo a $(1-\alpha)\times 100\%$ de confiança para o parâmetro β_j do modelo é:

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

com $\hat{\sigma}_{\hat{\beta}_j} = \sqrt{QMRE \cdot (\mathbf{X}^t\mathbf{X})^{-1}_{(j+1,j+1)}}$, e sendo $t_{\frac{\alpha}{2}\,[n-(p+1)]}$ o valor que na distribuição $t_{n-(p+1)}$ deixa *à direita* uma região de probabilidade $\frac{\alpha}{2}$.
O valor b_j é o elemento $j+1$ do vector das estimativas $\vec{\mathbf{b}}$ (acetato 180).

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

Intervalos de confiança para β_i no 

A informação básica obtém-se com a função `summary`.

ICs numa regressão múltipla com os lírios

```
> 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 intervalo a 95% de confiança para β_2 é

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

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

Intervalos de confiança para β_j no R (cont.)

Alternativamente, é possível usar a função `confint`.

A função *confint* nos lírios

```
> confint(iris2.lm)
                   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)
                    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
```

Testes de Hipóteses sobre os parâmetros

O mesmo resultado (acetato 215) usado para construir intervalos de confiança serve para construir testes a hipóteses para cada β_j individual.

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)

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

$T_{calc} > t_{\alpha[n-(p+1)]}$ (Unilateral direito)

Combinações lineares dos parâmetros

Seja $\vec{\mathbf{a}}^t = (a_0, a_1, ..., a_p)$ 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:

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

Casos particulares importantes nas aplicações são:

- Se $\vec{\mathbf{a}} = (0, 0, ..., \underbrace{1}_{=a_j}, ..., 0)$ (um único elemento não nulo, de valor 1, na posição $j+1$), então $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} = \beta_j$.
- Se $\vec{\mathbf{a}}$ tem apenas dois elementos não nulos, 1 na posição $i+1$ e ± 1 na posição $j+1$, $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} = \beta_i \pm \beta_j$.
- Se $\vec{\mathbf{a}} = (1, x_1, x_2, ..., x_p)$, onde x_j indica um possível valor da variável preditora X_j, então $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$ representa o valor esperado de Y associado aos valores indicados das variáveis preditoras:

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

Inferência sobre combinações lineares dos β_js

Para estimar $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} = a_0\beta_0 + a_1\beta_1 + a_2\beta_2 + \ldots + a_p\beta_p$, usa-se o estimador:

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

A multinormalidade de $\vec{\hat{\boldsymbol{\beta}}}$ implica a normalidade de qualquer vector combinação linear das suas componentes (acetato 205, ponto 4):

- Sabemos que $\vec{\hat{\boldsymbol{\beta}}} \frown \mathcal{N}_{p+1}\left(\vec{\boldsymbol{\beta}}\,, \sigma^2\,(\mathbf{X}^t\mathbf{X})^{-1}\right)$ (acetato 209);
- Logo, $\vec{\mathbf{a}}^t\vec{\hat{\boldsymbol{\beta}}} \frown \mathcal{N}(\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}\,, \sigma^2\,\vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}\,)$ (acetato 205, ponto 4);
- Ou seja, $\mathbf{Z} = \dfrac{\vec{\mathbf{a}}^t\vec{\hat{\boldsymbol{\beta}}} - \vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}}{\sqrt{\sigma^2\,\vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}}} \frown \mathcal{N}(0,1)$;
- Por um raciocínio análogo ao usado com os βs individuais, tem-se:

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

Quantidade fulcral para a inferência sobre $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$

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 resultado serve de base à construção de intervalos de confiança e testes de hipóteses para quaisquer combinações lineares dos parâmetros β_j do modelo.

NOTA: Repare-se na analogia da estrutura desta quantidade fulcral com os resultados anteriores, relativos a β_js individuais.

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, um intervalo a $(1-\alpha)\times 100\%$ de confiança para a 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:

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

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

Testes de Hipóteses sobre os parâmetros

Dado o Modelo de Regressão Linear Múltipla,

Testes de Hipóteses a $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$ (Regressão Linear Múltipla)

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 $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} = c$ (H_0)

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Rejeitar H_0 se

$T_{calc} < -t_{\alpha[n-(p+1)]}$ (Unilateral esquerdo)

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

$T_{calc} > t_{\alpha[n-(p+1)]}$ (Unilateral direito)

De novo os casos particulares

No acetato 220 viram-se três casos particulares importantes de combinações lineares dos parâmetros.

- No caso de $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} = \beta_j$, os intervalos e testes acabados de ver são idênticos aos dados nos acetatos 216 e 219.
- No caso de $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} = \beta_i \pm \beta_j$, tem-se $\hat{\sigma}_{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}} = \hat{\sigma}_{\hat{\beta}_i \pm \hat{\beta}_j}$, com:

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

As parcelas debaixo da raíz quadrada são elementos da matriz estimada de (co-)variâncias do vector de estimadores $\hat{\vec{\boldsymbol{\beta}}}$:

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

ICs para combinações lineares no R

Um intervalo de confiança para $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$ precisa da matriz das (co)variâncias estimadas dos estimadores $\vec{\hat{\boldsymbol{\beta}}}$, $\widehat{V[\vec{\hat{\boldsymbol{\beta}}}]} = QMRE \cdot (\mathbf{X}^t\mathbf{X})^{-1}$.

No R, esta matriz obtém-se através da função `vcov`.

De novo o exemplo dos lírios (`iris`)

```
> 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
```

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

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

Intervalos de confiança para $\mu_{Y|\vec{\mathbf{x}}}$

Se a combinação linear dos βs que se deseja corresponde ao valor esperado de Y, dado um conjunto de valores das variáveis preditoras, isto é, a $\mu_{Y|\vec{\mathbf{x}}} = E[Y|X_1 = x_1, X_2 = x_2, ..., X_p = x_p]$, então o intervalo de confiança do acetato 223 particulariza-se da seguinte forma:

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

sendo $\vec{\mathbf{x}} = (x_1, x_2, ..., x_p)$ o vector dos valores dos preditores,

$$\hat{\mu}_{Y|\vec{\mathbf{x}}} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_p x_p \ ,$$

e

$$\hat{\sigma}_{\hat{\mu}_{Y|\vec{\mathbf{x}}}} = \sqrt{QMRE \ \cdot \ \vec{\mathbf{a}}^t (\mathbf{X}^t \mathbf{X})^{-1} \vec{\mathbf{a}}},$$

com

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

Intervalos de confiança para $\mu_{Y|\vec{\mathbf{x}}}$ no 

É possível obter o intervalo de confiança referido no acetato anterior através do comando `predict`, tal como na RLS.

O exemplo dos lírios

Um IC a 95% para a largura esperada de pétalas de flores com:

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

usa o vector $\vec{\mathbf{a}} = (\ 1\ ,\ 2\ ,\ 5\ ,\ 3.1\ )^t$. No R:

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

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

O IC para $E[Y|X_1=2, X_2=5, X_3=3.1]$ é: $\left]\ 0.4169\ ,\ 0.5077\ \right[$.

O caso de $\mu_{Y|\vec{\mathbf{x}}_{(i)}}$ (valores do indivíduo i)

Quando $\vec{\mathbf{a}}$ contém valores das variáveis preditoras usados na i-ésima observação então $\vec{\mathbf{a}}^t$ é a linha i da matrix $\mathbf{X}$:

$$\vec{\mathbf{a}}^t = (1, x_{1_{(i)}}, x_{2_{(i)}}, ..., x_{p_{(i)}}) \ ,$$

.

Nesse caso,

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

onde h_{ii} indica o i-ésimo elemento diagonal da matriz de projecção ortogonal $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$. Este valor terá importância no que se segue.

Intervalos de predição para Y

Podem também obter-se, de forma análoga à RLS, intervalos de predição para uma observação individual de Y, associada aos valores $X_1 = x_1, ..., X_p = x_p$ das variáveis preditoras.

Nestes intervalos, a estimativa da variância associada a uma observação individual de Y é acrescida em $QMRE$ unidades:

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

onde $\vec{\mathbf{x}} = (x_1, x_2, ..., x_p)^t$ indica o vector dos valores dos preditores e

$$\hat{\mu}_{Y|\vec{\mathbf{x}}} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_p x_p$$

e

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

Intervalos de predição para Y no R

No  , é possível obter um intervalo de predição através do comando `predict` com o argumento `int="pred"`, tal como na RLS.

O exemplo dos lírios (`iris`)

O intervalo de predição (95%) para a largura da pétala, num lírio com comprimento de pétala 2 e sépala de comprimento 5 e largura 3.1, é:

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

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

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

Avaliando a qualidade do ajustamento global

Numa Regressão Linear Múltipla, com equação

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon \ ,$$

o Modelo Nulo $Y = \beta_0 + \varepsilon$, corresponde a admitir que todas as variáveis preditoras têm simultaneamente coeficiente β_j nulo.

As hipóteses correspondentes à inexistência (H_0), ou existência (H_1), de relacionamento linear são:

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

[MODELO NULO – INÚTIL]

vs.

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

[MODELO NÃO INÚTIL]

NOTA: Optar entre estas Hipóteses não é equivalente a realizar p testes *t-Student* aos β_j individuais.

NOTA: β_0 não intervém nas hipóteses.

Distribuição associada a *SQR*

De novo, o ponto de partida para uma estatística de teste será a Soma de Quadrados associada à Regressão, $SQR = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$.

Tem-se (sem demonstração):

Distribuição associada a SQR, sob H_0

Dado o Modelo de Regressão Linear Múltipla,

- $\frac{SQR}{\sigma^2} \frown \chi^2_p$, se $\beta_1 = \beta_2 = ... = \beta_p = 0$.
- *SQR* e *SQRE* são variáveis aleatórias independentes.

Quadrado Médio da Regressão (QMR)

Defina-se o Quadrado Médio associado à Regressão, $QMR = \frac{SQR}{p}$.

A estatística do teste de ajustamento global

Temos (veja também o acetato 131), se $\beta_j = 0\,, \forall i = 1 : p$

$$\left.\begin{array}{l} W = \frac{SQR}{\sigma^2} \frown \chi^2_p \\ V = \frac{SQRE}{\sigma^2} \frown \chi^2_{n-(p+1)} \\ W, V \text{ independentes} \end{array}\right\} \Rightarrow \quad \frac{W/p}{V/n-(p+1)} = \frac{QMR}{QMRE} \frown F_{p,n-(p+1)} \,.$$

sendo $QMR = \frac{SQR}{p}$ e $QMRE = \frac{SQRE}{n-(p+1)}$.

O Teste F de ajustamento global do Modelo

Sendo válido o Modelo RLM, pode efectuar-se o seguinte

Teste F de ajustamento global do modelo RLM

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

vs.

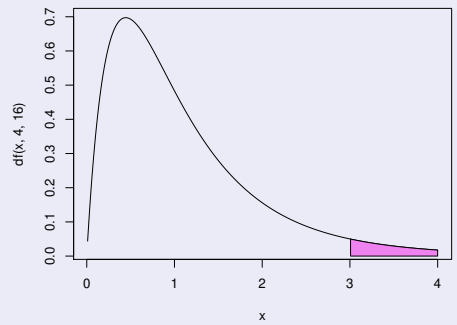
$H_1 : \exists\, j = 1, ..., p \quad \text{tal que} \quad \beta_j \neq 0.$

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



Expressões alternativas no teste F global

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

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

Tal como na Regressão Linear Simples, a estatística F é uma função crescente do Coeficiente de Determinação, R^2.

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 que, na população, o coeficiente de determinação é nulo.

Outra formulação do Teste F de ajustamento global

Teste F de ajustamento global do modelo RLM (alternativa)

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

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

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

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

- A estatística F é uma função crescente do coeficiente de determinação amostral, R^2.
- A hipótese nula $H_0 : \mathscr{R}^2 = 0$ corresponde a ausência de relação linear entre Y e o conjunto dos preditores (Modelo Nulo).

O Quadro-resumo do ajustamento global

Pode sintetizar-se a informação usada num teste de ajustamento global num quadro-resumo da regressão:

Fonte	g.l.	SQ	QM	f_{calc}
Regressão	p	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	$\frac{SQR}{p}$	$\frac{QMR}{QMRE}$
Resíduos	$n-(p+1)$	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$\frac{SQRE}{n-p-1}$	
Total	$n-1$	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	–	–

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 avaliar se um dado modelo difere significativamente dum 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 PIOR]

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: a sua presença não implica trabalho adicional de recolha de dados, nem de interpretação do modelo (ao mesmo tempo que permite um melhor ajustamento do modelo).

Estatística de teste para comparar modelo/submodelo

A estatística de teste envolve a comparação das 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 $R^2_c \geq R^2_s$, logo $SQRE_s \geq SQRE_c$.

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:

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

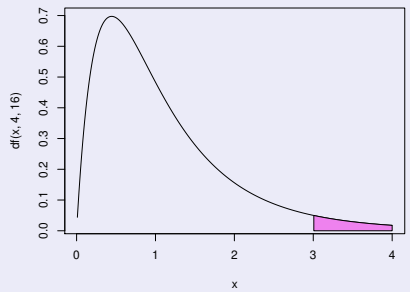
Estatística do Teste:

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

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

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



Expressão alternativa para a estatística do teste

A estatística do teste F de comparação de um modelo completo com p preditores, e um seu submodelo com apenas k preditores 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: Assinale-se que a Soma de Quadrados Total, SQT, apenas depende dos valores observados da variável resposta Y, e não de qual o modelo ajustado. Assim, SQT é igual no modelo completo e no submodelo.

Expressão alternativa para as hipóteses do teste

As hipóteses do teste também se podem escrever como

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

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:

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

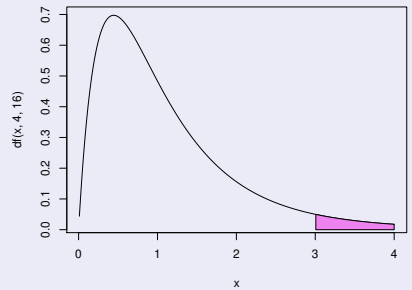
Estatística do Teste:

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

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

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



O teste a submodelos no R

A informação necessária para um teste F parcial obtem-se no R, através da função `anova`, com dois argumentos: os objectos `lm` resultantes de ajustar o modelo completo e o submodelo sob comparação.

O exemplo dos lírios (acetatos 92 e 217)

```
> 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 ***
```

O valor calculado da estatística é $F_{calc} = 12.616$. O respectivo *p-value* é $p = 8.836 \times 10^{-6}$. Rejeita-se a hipótese nula de igualdade de modelo e submodelo.

Relação entre os testes-t e o teste F parcial

Caso o modelo e submodelo difiram num único preditor X_j, o teste F parcial dos acetatos anteriores é equivalente ao teste *t-Student* (acetato 219) com as hipóteses $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$.

Nesse caso:

- 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 $F_{1,\nu}$.

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.

Nota: Veja-sa o Exercício RLM 11 g) para um exemplo.

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

Cuidado com exclusões simultâneas de preditores

Para pequenos valores de p: é viável analisar todos os possíveis subconjuntos de preditores.

Para valores de p até $p \approx 35$: Com algoritmos e rotinas informáticas adequadas, ainda é possível pesquisar todos os subconjuntos.

Mas para p muito grande: uma pesquisa exaustiva é computacionalmente inviável.

Não é legítimo usar testes t à significância de cada β_j no modelo completo para decidir sobre a exclusão de vários preditores em simultâneo.

Um teste t a $\beta_j = 0$ parte do princípio que todas as restantes variáveis pertencem ao modelo. A exclusão de qualquer preditor altera os valores estimados b_j e 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

Dados brix (Exercício RLM 2)

A tabela da regressão da variável *Brix* sobre todas as restantes é:

```
             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 todas dispensáveis.

```
> 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 *
```

Pesquisas completas

Para um número p de preditores não excessivo, e dispondo de algoritmos e rotinas informáticas adequadas, é possível efectuar uma pesquisa completa que assegure que se encontra o subconjunto de k preditores com o maior valor de R^2 (ou de algum outro critério de qualidade do submodelo).

O algoritmo *leaps and bounds*, de Furnival e Wilson [1] é um algoritmo computacionalmente eficiente que permite identificar o melhor subconjunto de preditores, de uma dada cardinalidade k.

Uma rotina implementando o algoritmo encontra-se disponível no `R`, num módulo (*package*) de nome `leaps` (comando com o mesmo nome). Outra rotina análoga encontra-se na função `eleaps` do módulo `subselect`.

[1] Furnival, G.W and Wilson, R.W.,Jr. (1974) Regressions by leaps and bounds, *Technometrics*, **16**, 499-511.

Um exemplo de aplicação da rotina `leaps`

A rotina *leaps* nos dados `brix`

Apesar do pequeno número de preditores, exemplifiquemos a aplicação da função `leaps` com os dados `brix` (Exercício RLM 2).

```
> colnames(brix)
[1] "Diametro" "Altura"   "Peso"     "Brix"     "pH"       "Acucar"

> library(leaps)       <-- carregar o módulo (tem de estar instalado)
> leaps(y=brix$Brix, x=brix[,-4], method="r2", nbest=1) <-- o comando: y resposta, x preditores

$which                 <-- matriz de valores lógicos, indicando resultados (cada coluna um preditor,
      1     2     3     4     5                              cada linha uma cardinalidade de subconjunto)
1 FALSE FALSE FALSE FALSE  TRUE    <-- k=1 ; melhor preditor individual: Acucar
2  TRUE  TRUE FALSE FALSE FALSE    <-- k=2 ; melhor par de preditores: Diametro e Altura
3  TRUE  TRUE FALSE FALSE  TRUE    <-- k=3 ; melhor trio de preditores: Diametro, Altura, Acucar
4  TRUE  TRUE FALSE  TRUE  TRUE
5  TRUE  TRUE  TRUE  TRUE  TRUE
[...]

$r2              <-- Coef. Determinação da melhor solução com o no. k=1,2,3,4,5 de preditores
[1] 0.5091325 0.6639105 0.7863475 0.8083178 0.8482525
```

Repare-se como o melhor submodelo (R^2 mais elevado) com dois preditores não é o submodelo com os preditores `Diametro` e `Acucar`, como sugerido pelos p-values do ajustamento do modelo completo.

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

1. ajustar o modelo completo, com os p preditores;
2. definir um nível de significância α para os testes de hipóteses;
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.

Algoritmos sequenciais com base no AIC

O  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).

Algoritmos sequenciais com base no AIC (cont.)

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 sequenciais com base no AIC (cont.)

Em cada passo de exclusão, o submodelo com menor AIC será aquele que provoca menor aumento no *SQRE*, ou seja, que tiver excluído a variável cujo teste a $\beta_j = 0$ tem maior *p-value*.

Assim, o procedimento de exclusão sequencial baseado nos testes *t* ou no AIC coincidem na ordem das variáveis a excluir, podendo diferir apenas no critério de paragem.

Em geral, um algoritmo de exclusão sequencial baseado no AIC é mais cauteloso ao excluir do que um algoritmo baseado nos testes *t*, sobretudo se o valor de α usado nos testes for baixo.

Nos algoritmos de exclusão baseados nos testes *t*, é aconselhável usar valores relativamente elevados de α, como $\alpha = 0.10$.

Algoritmos de exclusão sequencial no R

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

Dados `brix` (Exercício 2 RLM)

```
> 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
- Peso      1  0.039279 0.18853 -50.306
- 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.

Uma palavra final sobre algoritmos de pesquisa

O algoritmo de exclusão sequencial não garante a identificação do "melhor submodelo" com um dado número de preditores. Apenas identifica, de forma que não é computacionalmente muito pesada, submodelos que se presume serem "bons".

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

Regressão Polinomial

Um caso particular de relação não-linear, mesmo que envolvendo apenas uma variável preditora e a variável resposta, pode ser facilmente tratada no âmbito duma regressão linear múltipla: o caso de relações polinomiais entre Y e um ou mais preditores.

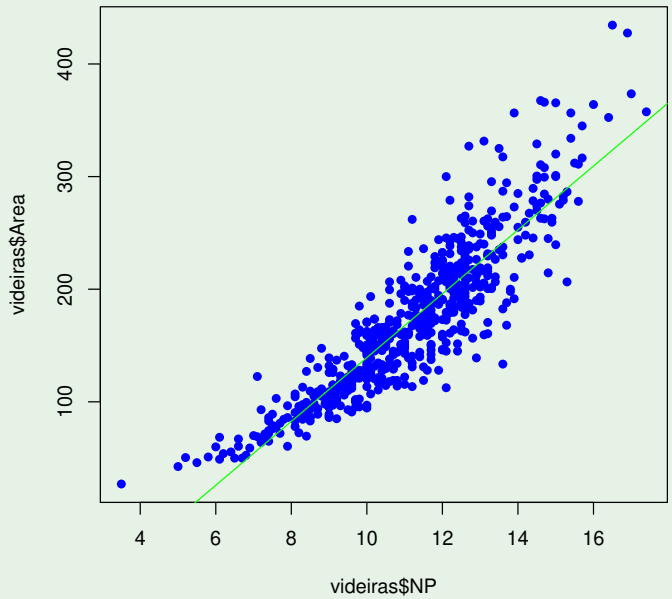
Considere-se, por exemplo, que a relação de fundo entre uma variável resposta Y e uma única variável preditora X é dada por uma parábola, de equação:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 \ .$$

Regressão Polinomial - Exemplo

Exemplo: folhas de videira (Exercício RLM 9)

Eis o gráfico das áreas *vs.* comprimentos de nervuras principais, com sobreposta a recta de regressão:



Há uma tendência para curvatura. Talvez um polinómio de 2º grau?

Regressão Polinomial - Exemplo (cont.)

A parábola de equação $Y = \beta_0 + \beta_1 x + \beta_2 x^2$ pode ser vista como a equação duma regressão linear entre Y e as variáveis $X_1 = X$ e $X_2 = X^2$:

Dados `videiras` (Exercício RLM 9)

```
> summary(lm(Area ~ NP + I(NP^2), data=videiras))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.5961    22.0431   0.345    0.731
NP           -0.2172     4.0125  -0.054    0.957
I(NP^2)       1.2941     0.1801   7.187 1.98e-12 ***
--
Residual standard error: 28.86 on 597 degrees of freedom
Multiple R-squared: 0.8162, Adjusted R-squared: 0.8155
F-statistic:  1325 on 2 and 597 DF,  p-value: < 2.2e-16
```

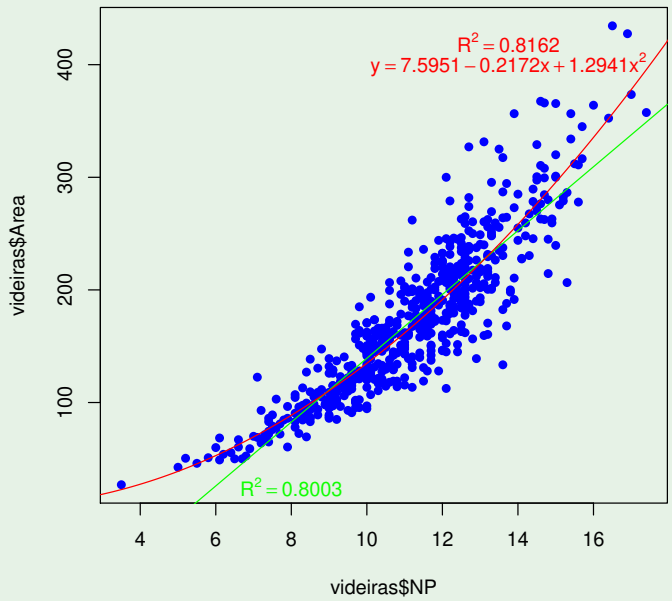
A equação da parábola ajustada é $y = 7.5961 - 0.2172\,x + 1.2941\,x^2$.

A rejeição da hipótese $\beta_2 = 0$ diz que a parábola ajusta-se significativamente melhor do que uma recta de regressão.

Regressão Polinomial - Exemplo (cont.)

Dados `videiras` (Exercício RLM 9)

Eis a parábola ajustada:



É legítimo afirmar que este modelo de regressão quadrático explica 81.62% da variabilidade nas áreas foliares observadas, uma vez que não houve transformação da variável resposta Y.

Regressões Polinomiais (cont.)

O argumento é extensível a qualquer polinómio de qualquer grau, e em qualquer número de variáveis. Dois exemplos:

- Polinómio de grau p numa variável

$$Y = \beta_0 + \beta_1 \underbrace{x}_{=x_1} + \beta_2 \underbrace{x^2}_{=x_2} + \beta_3 \underbrace{x^3}_{=x_3} + \ldots + \beta_p \underbrace{x^p}_{=x_p}$$

- Polinómio de grau 2 em 2 variáveis

$$Y = \beta_0 + \beta_1 \underbrace{x}_{=x_1} + \beta_2 \underbrace{x^2}_{=x_2} + \beta_3 \underbrace{z}_{=x_3} + \beta_4 \underbrace{z^2}_{=x_4} + \beta_5 \underbrace{xz}_{=x_5}$$

O R^2 modificado

O R^2 modificado é uma variante do Coeficiente de Determinação.

- O Coeficiente de Determinação usual:

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

- O R^2 modificado (definindo $QMT = \frac{SQT}{n-1} = s_y^2$) é:

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

Tem-se sempre $n-1 > n-(p+1)$, pelo que $R^2_{mod} < R^2$.

Quando $n \gg p+1$ (muito mais observações que parâmetros no modelo) tem-se $R^2 \approx R^2_{mod}$.

Se n é pouco maior que o número de variáveis preditoras, $R^2_{mod} \ll R^2$ (excepto se R^2 for muito próximo de 1).

O R^2 modificado (cont.)

R^2_{mod} penaliza modelos complexos ajustados com poucas observações.

Exemplo: 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.

Exemplo: dados `milho` (Exercício RLM 11)

(que ilustra o uso do R^2_{mod} como critério de selecção na função de pesquisa `leaps`):

```
> library(leaps)
> leaps(y=milho$y , x=milho[,-10], method="adjr2", nbest=1)
[...]
$adjr2    <-- o maior R2 modificado é no submodelo com k=4 preditores
[1] 0.5493014 0.6337329 0.6544835 0.6807418 0.6798986 0.6779395 0.6745412
[8] 0.6633467 0.6488148
```

Análise de Resíduos e outros diagnósticos

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

Grande parte do que se disse sobre resíduos na Regressão Linear Simples mantém-se válido numa Regressão Linear Múltipla.

Relembrar três conceitos relacionados, mas diferentes:

Erros aleatórios (desconhecidos)

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

Resíduos (variáveis aleatórias - preditores dos erros aleatórios)

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

Resíduos (valores observados)

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

Propriedades dos Resíduos sob o Modelo RLM

O modelo de Regressão Linear Múltipla admite que

$$\varepsilon_i \frown \mathcal{N}(0, \sigma^2) \qquad \forall i = 1, ..., n \, .$$

Distribuição dos Resíduos

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

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

onde h_{ii} é 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 $\mathscr{C}(\mathbf{X})$.

Em notação vectorial, o vector dos n resíduos E_i é dado por:

$$\vec{\mathbf{E}} = \vec{\mathbf{Y}} - \vec{\hat{\mathbf{Y}}} = \vec{\mathbf{Y}} - \mathbf{H}\vec{\mathbf{Y}} = (\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}} \, ,$$

Propriedades dos Resíduos sob o Modelo RLM (cont.)

Sabemos que $\vec{\mathbf{Y}} \frown \mathcal{N}_n(\mathbf{X}\vec{\boldsymbol{\beta}}\,, \sigma^2\mathbf{I}_n)$. Logo, vector dos resíduos $\vec{\mathbf{E}} = (\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}$:

- tem distribuição Multinormal (é da forma $\mathbf{B}\vec{\mathbf{Y}}$, com $\mathbf{B} = (\mathbf{I}_n - \mathbf{H})$ não aleatória);
- tem vector esperado:

$$E[\vec{\mathbf{E}}] \;=\; E[(\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}] \;=\; (\mathbf{I}_n - \mathbf{H})E[\vec{\mathbf{Y}}] \;=\; (\mathbf{I}_n - \mathbf{H})\mathbf{X}\vec{\boldsymbol{\beta}} \;=\; \vec{\mathbf{0}}\,,$$

  pois o vector $\mathbf{X}\vec{\boldsymbol{\beta}} \in \mathscr{C}(\mathbf{X})$, logo permanece invariante sob a acção da matriz de projecção $\mathbf{H}$: $\mathbf{H}\mathbf{X}\vec{\boldsymbol{\beta}} = \mathbf{X}\vec{\boldsymbol{\beta}}$;

- tem matriz de covariâncias:

$$V[\vec{\mathbf{E}}] \;=\; V[(\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}] \;=\; (\mathbf{I}_n - \mathbf{H})\,V[\vec{\mathbf{Y}}](\mathbf{I}_n - \mathbf{H})^t \;=\; \sigma^2\,(\mathbf{I}_n - \mathbf{H})\,,$$

  porque a matriz de projecção ortogonal é (Exercício RLM 4) simétrica ($\mathbf{H}^t = \mathbf{H}$) e idempotente ($\mathbf{H}^2 = \mathbf{H}\,\mathbf{H} = \mathbf{H}$).

Propriedades dos Resíduos sob o Modelo RLM (cont.)

Distribuição do vector dos Resíduos no modelo RLM

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

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

Logo, E_i Normal, com $E[E_i] = 0$ e $V[E_i] = \sigma^2(1 - h_{ii})$.

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

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

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

Vários tipos de resíduos

Tal como na RLS, definem-se diferentes tipos de resíduos:

Três tipos de resíduos (também na RLM)

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

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

Resíduos Studentizados (ou externamente estandardizados):

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

sendo $QMRE_{[-i]}$ o valor de $QMRE$ resultante da hipersuperfície ajustada excluíndo a i-ésima observação.

Análise dos resíduos

Tal como para a RLS, também em regressões múltiplas se avalia a validade dos pressupostos do modelo através de gráficos de resíduos.

Estes gráficos são agora mais importantes do que na RLS, dada a impossibilidade de visualização de nuvens de pontos em espaços de alta dimensionalidade.

Os gráficos mais usuais são os já considerados na RLS e a sua leitura faz-se de forma análoga:

- gráfico de E_is vs. $\hat{Y}_i$s: os pontos devem-se dispor numa banda horizontal, centrada no valor zero, sem outro padrão especial.
- *qq-plot* dos resíduos estandardizados vs. distribuição Normal: a Normalidade dos erros aleatórios corresponde à linearidade.
- gráfico de resíduos vs. ordem de observação: para investigar eventuais faltas de independência dos erros aleatórios.

O efeito alavanca

Outras ferramentas de diagnóstico visam identificar observações individuais que merecem ulterior análise, tal como na RLS. Mas importa adaptar as definições ao contexto de Regressão Múltipla.

Efeito alavanca

Numa RLM o valor de efeito alavanca (*leverage*) é o valor h_{ii} do elemento diagonal da matriz de projecção ortogonal $\mathbf{H}$, correspondente à observação i

- tem-se $\frac{1}{n} \leq h_{ii} \leq 1$;
- o valor médio das observações alavanca numa RLM é a razão entre o número de parâmetros e o número de observações:

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

Gráficos de diagnóstico

Distância de Cook

A distância de Cook para avaliar a influência da observação i define-se agora como:

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

onde $\hat{y}_{j(-i)}$ é o j-ésimo valor ajustado sem a observação i.

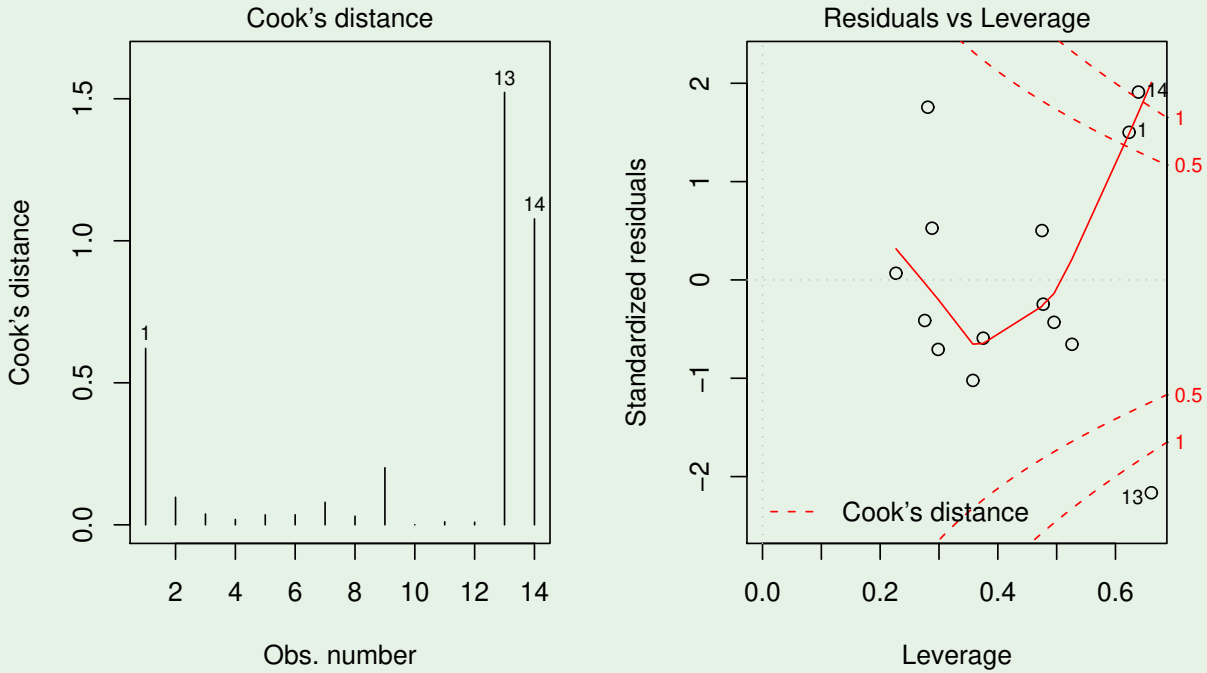
Alternativamente (sendo R_i o correspondente resíduo estandardizado):

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

Os restantes aspectos da discussão são análogos aos duma RLS.

Um exemplo de gráficos de diagnóstico

Dados `brix` (Exercício RLM 2)



Os valores bastante grandes de distância de Cook e efeito alavanca h_{ij} neste exemplo reflectem o reduzido número de observações ($n=14$) usado para ajustar um modelo com muitos parâmetros ($p+1=6$).

Advertências finais

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

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

Multicolinearidade exacta reflecte redundância de informação nos preditores.

É possível eliminar multicolinearidade exacta ou aproximada, excluíndo da análise uma ou várias variáveis preditoras que sejam responsáveis pela (quase) dependência linear dos preditores.

Advertências finais (cont.)

2. Tal como na RLS, podem ser usadas transformações da variável resposta e uma ou mais variáveis preditoras.

São úteis transformações que linearizem uma relação não linear entre Y e $X_1, X_2, ..., X_p$. Tais transformações linearizantes permitem estudar relações não lineares através de relações lineares entre as variáveis transformadas.

E.g., uma relação não linear entre y, x_1 e x_2, da forma:

$$y = a x_1^b x_2^c$$

torna-se, após logaritmização, numa relação linear entre $\ln(y)$, $\ln(x_1)$ e $\ln(x_2)$ (com $b_0 = \ln(a)$, $b_1 = b$ e $b_2 = c$):

$$\ln(y) = \ln(a) + b\ln(x_1) + c\ln(x_2) = b_0 + b_1\ln(x_1) + b_2\ln(x_2) .$$

Nota: Os erros aleatórios aditivos, com os pressupostos usuais, devem ser válidos após as transformações linearizantes.

Advertências finais (cont.)

3. Não se deve confundir a existência de uma relação linear entre preditores X_1, X_2, ..., X_p e 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.