Modelos Matemáticos e Aplicações
Modelo Linear

Jorge Cadima

Secção de Matemática (DCEB) - Instituto Superior de Agronomia (UL)

2018-19

Módulo 2: Modelação Estatística

Introdução aos principais modelos estatísticos.

1. Modelo Linear
2. Modelos Lineares Generalizados
3. Modelos Mistos

Os mais estudados e utilizados modelos estatísticos fazem parte do chamado Modelo Linear.

- Regressão Linear (Simples e Múltipla)
- Regressão Polinomial
- Análise de Variância (ANOVA)
- Análises de Covariância (ANCOVA)

Bibliografia - Modelo Linear

1. **Apontamentos da disciplina Estatística e Delineamento:**
   - Cadima, J. (2018) *O Modelo Linear* (https://fenix-edu.isa.ulisboa.pt/downloadFile/563022967866625/folhas.pdf).
2. **Referências Base:**
   - Draper, N.R. e Smith, H. (1998), *Applied Regression Analysis*, 3a. edição, John Wiley & Sons **[BISA: U10-734] + [SI-78]** (**[BISA: U10-412]** a primeira edição de 1981).
   - Kutner, M.H.; Nachtsheim, C.J.; Neter, J. e Li, W. (2005), *Applied Linear Statistical Models*, Irwin **[BISA: U10-727 e CD-236]**.
   - Montgomery, D.C. e Peck, E.A. (1982), *Introduction to Linear Regression Analysis*, John Wiley & Sons **[BISA: U10-329]**.
   - Seber, G.A.F. (1977), *Linear Regression Analysis*, John Wiley & Sons **[BISA: U10-416]**

Bibliografia (continuação)

3. **Referências de apoio à utilização do R**

- Agresti, Alan (2015) *Foundations of Linear and Generalizd Linear Models*. Wiley Series in Probability and Statistics.
- Fox, John e Weisberg, Harvey Sanford (2011) *An R Companion to Applied Regression*, SAGE publications.
- Maindonald, J. e Brown, W.J. (2003), *Data Analysis and Graphics using R*, Cambridge University Press **[BISA: U10-722]**
- Venables, W.N. e Ripley, B.D. (2002), *Modern Applied Statistics with S (fourth edition)*, Springer-Verlag **[BISA: U10-733]**

Modelação Estatística

Objectivo (informal): Estudar a relação entre

- uma variável resposta (ou dependente) y; e
- uma ou mais variáveis preditoras (variáveis explicativas ou independentes), $x_1, x_2, \ldots, x_p$.

A relação é estudada com base em n observações do conjunto de variáveis envolvidas na relação.

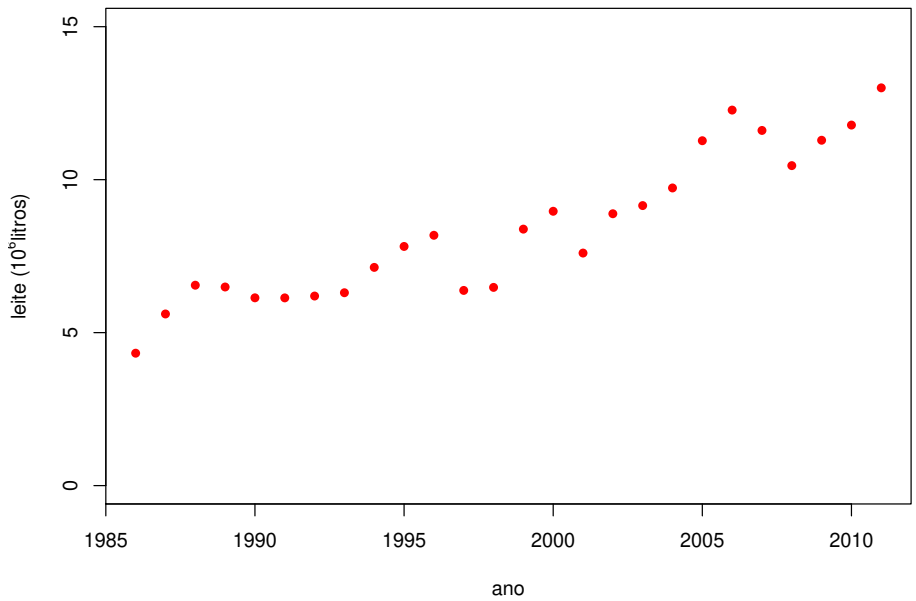
Nesta disciplina apenas se consideram modelos para n observações independentes, e com uma única variável resposta numérica.

Pode ter-se um ou mais preditores, que podem ser numéricos ou categóricos (factores).

Motivamos a discussão com alguns exemplos.

Exemplo 1: regressão linear simples (descritiva)

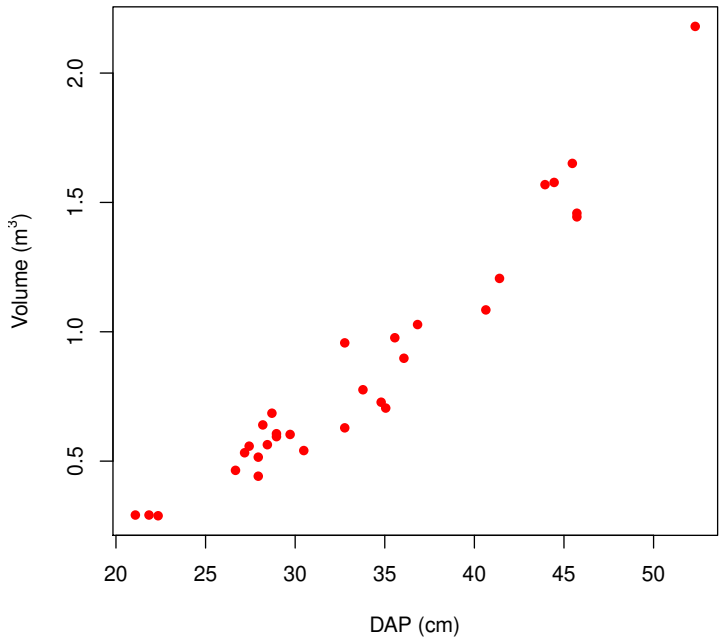
Resposta: Produção de leite de cabra em Portugal (y, `leite`) (10^6 litros).
Preditor: Anos (x, `ano`) (1986 a 2011).
Dados: $n = 26$ pares de valores, $\{(x_i, y_i)\}_{i=1}^{26}$. Na *data frame* `Cabra`.
Fonte: Instituto Nacional de Estatística (INE).



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

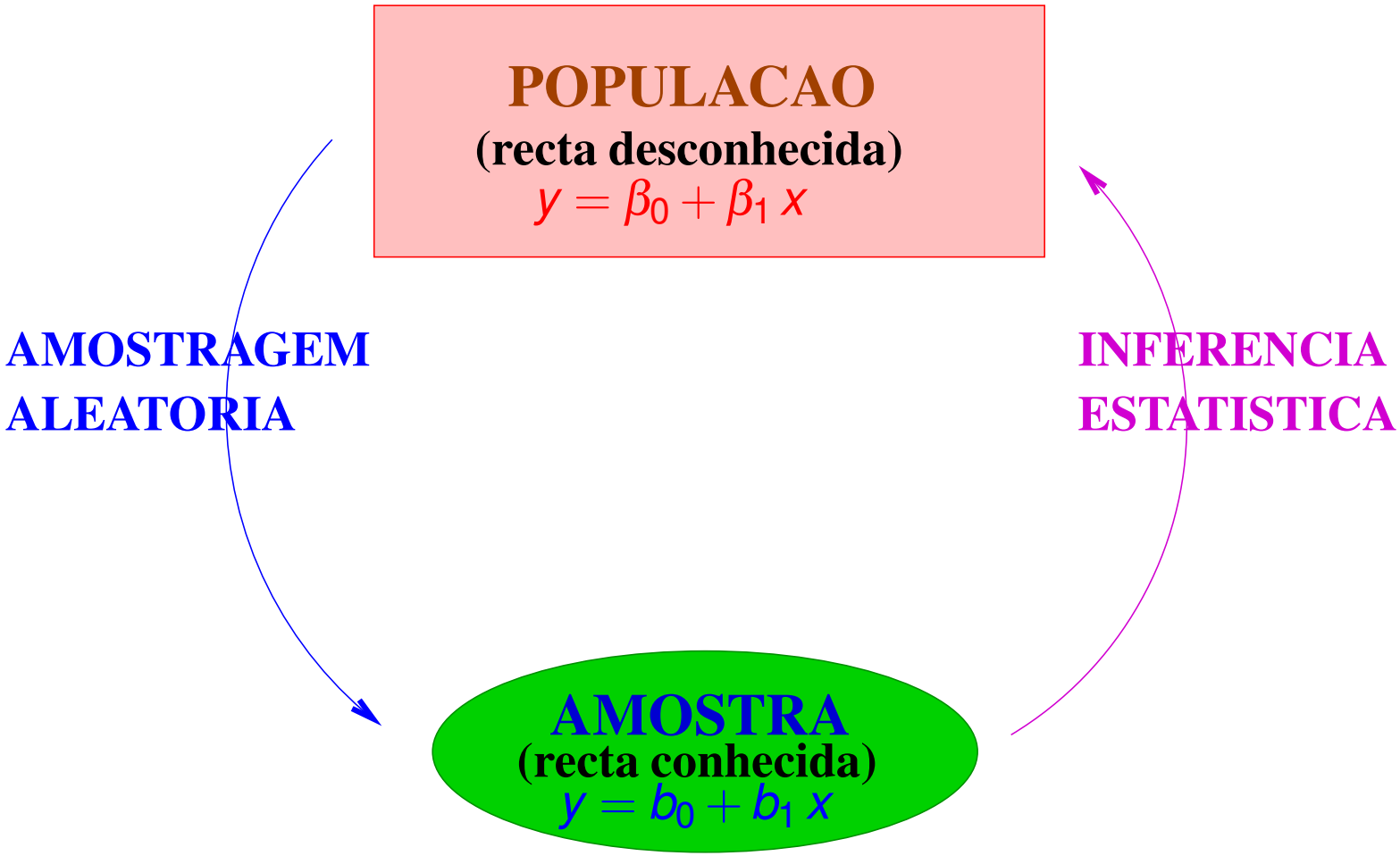
Exemplo 2 - regressão linear simples (inferencial)

Resposta (numérica): Volume de troncos (y) de cerejeiras.
Preditor (numérico): Diâmetro à altura do peito, DAP, (x).
Dados: $n = 31$ pares de medições, $\{(x_i, y_i)\}_{i=1}^{31}$. *Data frame* `trees`.
Fonte: No `R`: ver `help(trees)` para detalhes. Convertido ao sistema métrico.



Tendência de fundo aproximadamente linear.
As observações são uma amostra aleatória duma população maior.
Interessa o contexto inferencial: o que se pode dizer sobre a recta populacional $y = \beta_0 + \beta_1 x$?

O problema da Inferência Estatística na RLS



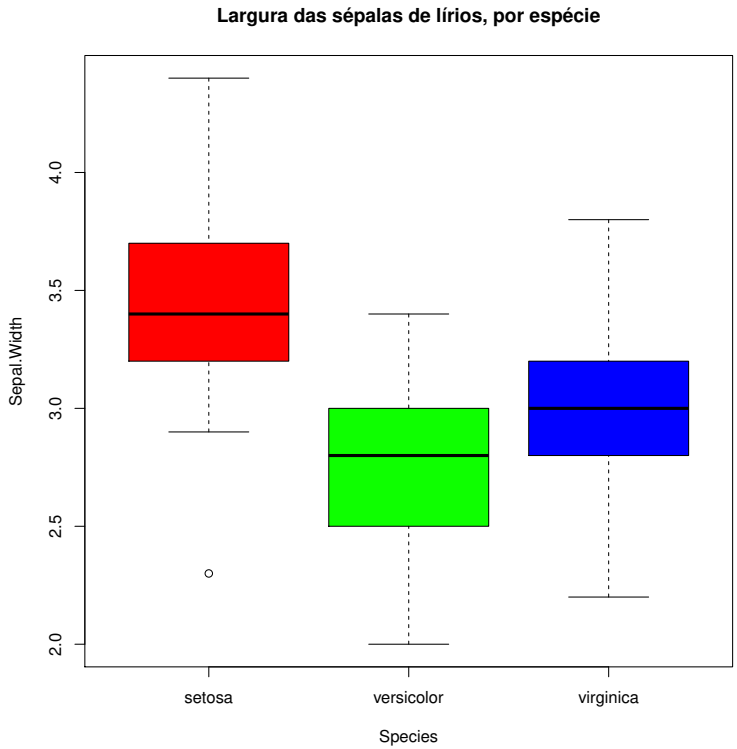
Exemplo 3: ANOVA a um factor

Resposta (numérica): largura de sépalas em lírios.

Preditor (factor): espécie de lírio.

Dados: $n = 150$ medidas, 50 em cada espécie. `Data frame` `iris`.

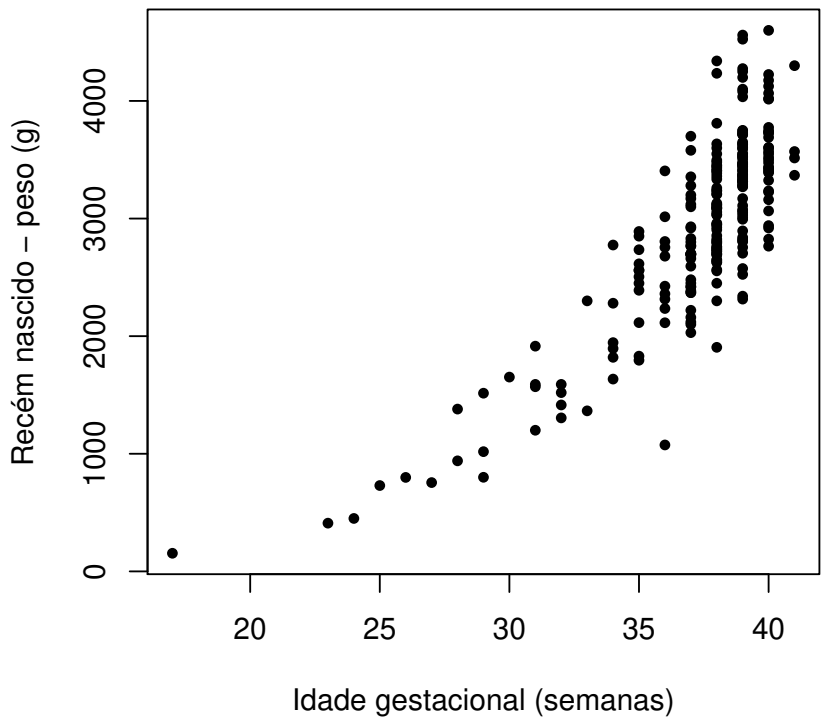
Fonte: No `R`: ver `help(iris)` para detalhes.



Haverá diferenças nos valores médios populacionais de cada espécie?

Exemplo 4 - relação não linear (descritivo)

Resposta (numérica): peso de bebé recém-nascido(y), em g.
Preditor (numérico): Idade gestacional (x), em semanas.
Dados: $n = 251$ pares de observações, $\{(x_i, y_i)\}_{i=1}^{251}$.



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

Exemplo 4 (cont.)

Neste caso, há uma questão adicional:

- Qual a forma da relação $y = f(x)$ (qual a natureza da função f)?
  - f exponencial ($y = c\,\mathrm{e}^{dx}$)?
  - f função potência ($y = c\,x^d$)?

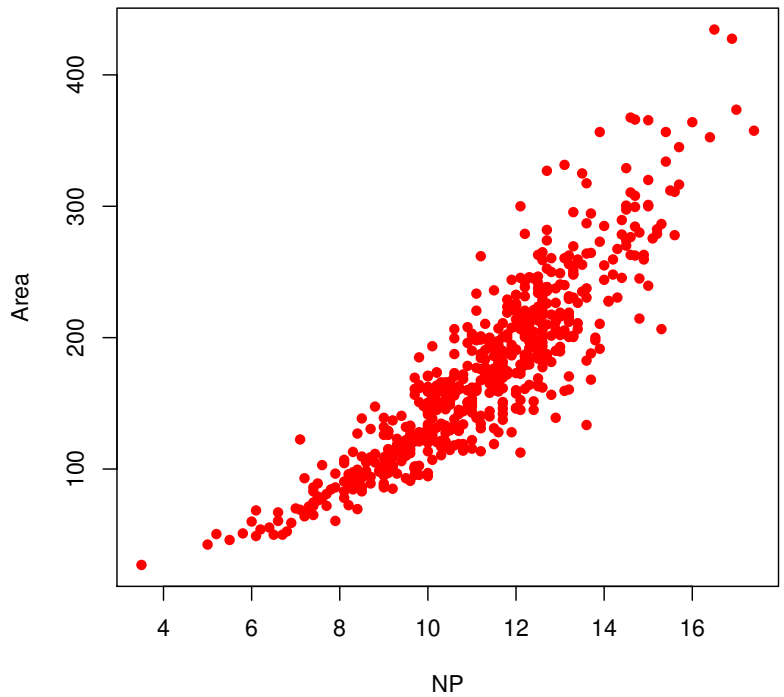
Escolhida a classe de f, há perguntas análogas ao caso linear: como determinar os "melhores" parâmetros c e d?

Relações não lineares estudam-se através da Regressão Não Linear.

Mas muitas relações não lineares podem ser linearizadas através de transformações adequadas das variáveis, e a relação linearizada resultante pode ser estudada com o Modelo Linear.

Exemplo 5 - relação não linear (inferencial)

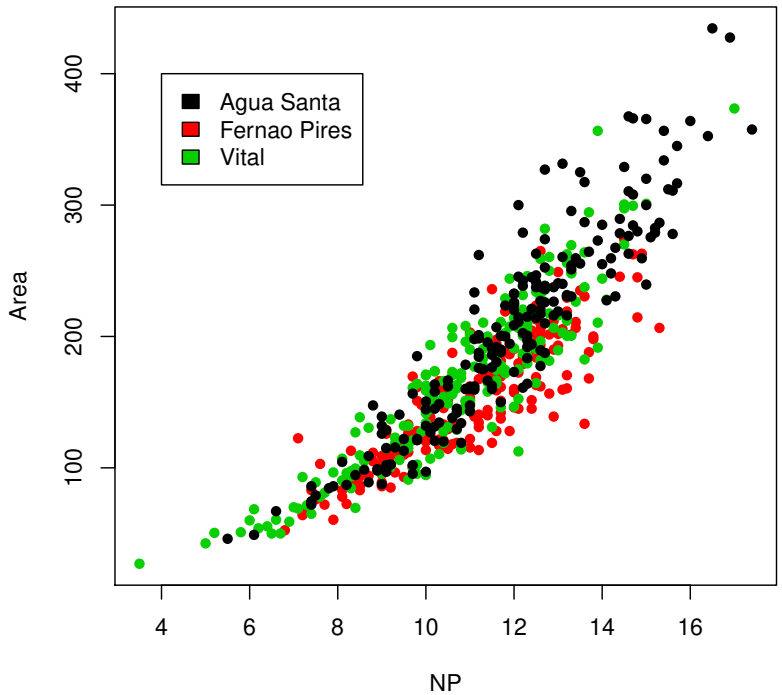
Resposta (numérica): Area de folhas de videira (y, `Area`).
Preditor (numérico): comprimento da nervura principal (x, `NP`).
Dados: $n = 600$ pares de observações, $\{(x_i, y_i)\}_{i=1}^{600}$. *Data frame* `videiras`.
Fonte: Prof. Carlos Lopes, Viticultura, ISA.



Tendência de fundo não-linear, $y = f(x)$. Parábola? Exponencial? Potência?
Dados são amostra aleatória. Que dizer sobre os parâmetros populacionais?

Exemplo 6 - relação de tipo ANCOVA

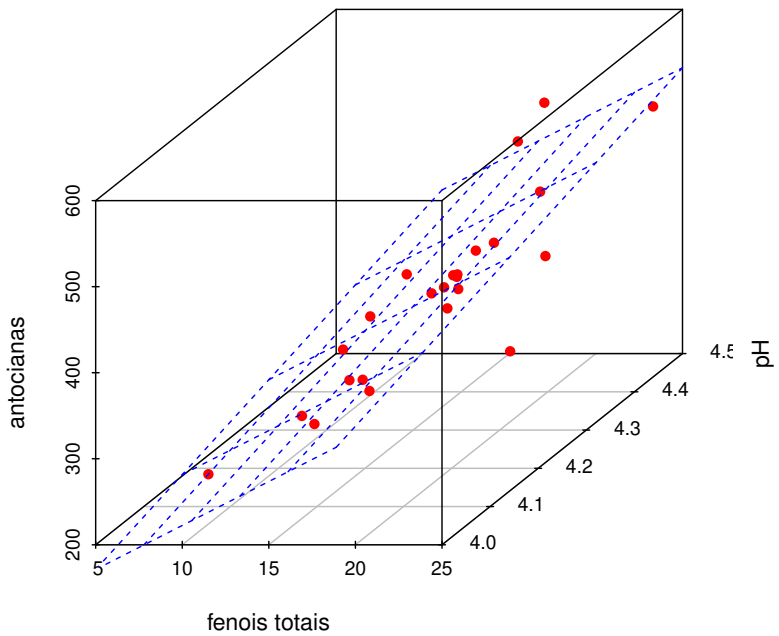
Resposta (numérica): Area de folhas de videira (y, `Area`).
Preditor (numérico): comprimento da nervura principal (x, `NP`).
Preditor (factor): casta (há 3 castas: Água Santa, Fernão Pires e Vital).
Dados: $n = 200$ observações para cada casta. *Data frame* `videiras`.



Uma única curva ajusta-se bem a todas as castas?
Ou haverá curvas diferentes para castas diferentes?

Exemplo 7 - Regressão linear múltipla

Resposta (numérica): Teor de antocianinas (y, `antoci`) (em mg/dm^3).
Preditor (numérico): teor de fenóis totais (x_1, `fentot`).
Preditor (numérico): pH (x_2, `pH`).
Dados: $n = 24$ genótipos casta Tinta Francisca. *Data frame* `Antoci`.
Fonte: Prof. Elsa Gonçalves, Matemática e Genética, ISA (Tabuaço 2003).



Descritivo: qual o "melhor" plano amostral $y = b_0 + b_1 x_1 + b_2 x_2$?
Inferencial: que dizer sobre o plano populacional $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$?

Ideias prévias sobre modelação

- Todos os modelos são apenas aproximações da realidade. Uns são melhores que outros.
- O princípio da parcimónia na modelação: de entre os modelos considerados adequados, é preferível o mais simples.
- Os modelos podem ser:
  - modelos teóricos, baseados em princípios físicos, biológicos, etc.;
  - modelos empíricos, descrevendo a relação observada nos dados.
- Os modelos estatísticos não são determinísticos: descrevem uma relação de fundo, sabendo que há variação das observações em torno dessa relação de fundo. Essa variabilidade é incorporada no modelo.

Ideias prévias sobre modelação (cont.)

- Não há (necessariamente) relação de causa e efeito entre variável resposta e preditores. A Estatística só pode mostrar que há associação. Uma eventual existência de relação causa e efeito é exterior à Estatística.
- No estudo de modelos estatísticos há aspectos diferentes:
  - faceta descritiva: ajustar modelo a dados observados, qualquer que seja a sua origem.
  - faceta inferencial: se os dados são uma amostra aleatória duma população, procurar tirar conclusões sobre a população.

  A inferência exige mais pressupostos e muito mais ferramentaria matemático-estatística.

O Modelo Linear

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

Revisão: Reg. Linear Simples - contexto descritivo

Se n pares de observações $\{(x_i, y_i)\}_{i=1}^{n}$ têm relação linear de fundo, a recta de regressão de y sobre x define-se como:

Recta de regressão linear de y sobre x

$$y = b_0 + b_1 x$$

com

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

sendo

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad ; \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

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

Revisão: Reg. Linear Simples descritiva (cont.)

Como se chegou à equação da recta?

Critério: Minimizar a soma de quadrados residual (i.e., dos resíduos) (Legendre 1805, Gauss 1809).

Os resíduos são distâncias na vertical entre pontos e recta ajustada:

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

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

Soma de Quadrados dos Resíduos:

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

Critério: Determinar b_0 e b_1 que minimizam $SQRE$.
Nota: Unidades de medida de $SQRE$: quadrado das unidades de y.

Revisão: Reg. Linear Simples descritiva (cont.)

Para minimizar *SQRE* tem de se anular as respectivas derivadas parciais em ordem a b_0 e b_1:

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

$$\Rightarrow \begin{cases} \frac{\partial SQRE}{\partial b_0}(b_0, b_1) = 0 \\ \\ \frac{\partial SQRE}{\partial b_1}(b_0, b_1) = 0 \end{cases} \Leftrightarrow \begin{cases} (-2) \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)] = 0 \\ 2 \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)] (x_i) = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \sum_{i=1}^{n} y_i - n b_0 - b_1 \sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} y_i x_i - b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2 = 0 \end{cases} \Leftrightarrow \begin{cases} b_0 = \overline{y} - b_1 \overline{x} \\ \\ b_1 = \frac{cov_{xy}}{s_x^2} \end{cases} .$$

Este ponto crítico tem de ser mínimo, pois a função *SQRE* é positiva e quadrática.

Regressão Linear Simples - contexto descritivo no R

As regressões lineares são ajustadas no R usando o comando `lm` (as iniciais de linear model).

A função `lm` tem dois argumentos fundamentais:

- `formula` – identifica a variável resposta e as variáveis preditoras; numa RL simples da variável y sobre o preditor x, é da forma: $y \sim x$.
- `data` – indica o nome da *data frame* contendo os dados.

Comando R para a RLS do Exemplo 1

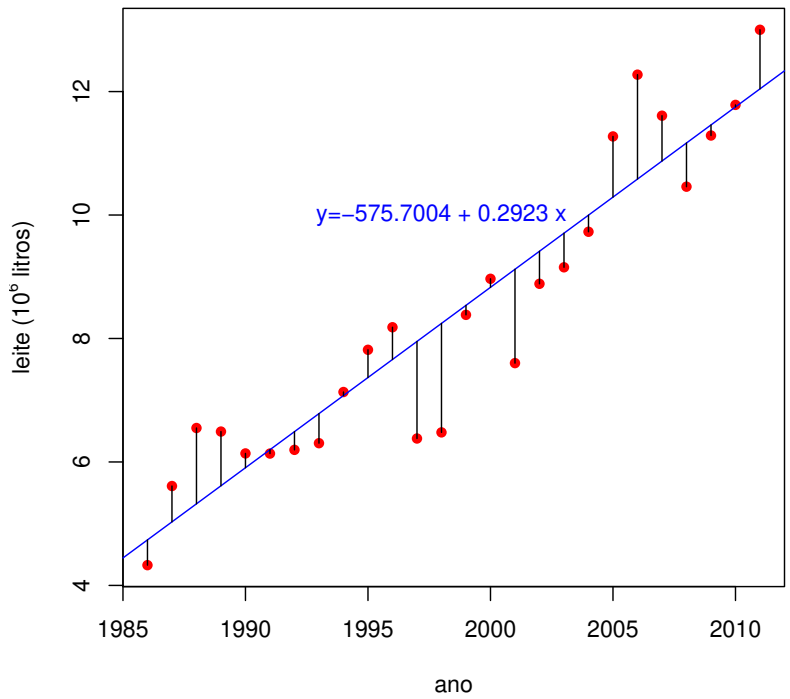
```
> lm( leite ~ ano , data=Cabra )
```

Resultado:

```
Call: lm(formula = leite ~ ano, data = Cabra)
Coefficients:
(Intercept)          ano
  -575.7004       0.2923    <- valores ajustados de b0 e b1
```

Regressão Linear Simples descritiva - Exemplo 1

Exemplo 1: Ano (x) e produção de leite (y) de cabra.
$n = 26$ pares de medições, $\{(x_i, y_i)\}_{i=1}^{26}$.



A recta ajustada minimiza a soma dos quadrados das distâncias, na vertical, entre pontos e recta.

Propriedades da recta de regressão

- A ordenada na origem b_0:
  - é o valor de y (na recta) associado a $x = 0$;
  - tem unidades de medida iguais às de y.
- O declive b_1:
  - é a variação (média) de y associada a um aumento de uma unidade em x;
  - tem unidades de medida iguais a $\frac{\textit{unidades de } y}{\textit{unidades de } x}$.
- A recta de regressão passa sempre no centro de gravidade da nuvem de pontos, isto é, no ponto $(\overline{x}, \overline{y})$, como é evidente a partir da fórmula para a ordenada na origem:

$$b_0 = \overline{y} - b_1 \overline{x} \quad \Leftrightarrow \quad \overline{y} = b_0 + b_1 \overline{x} \ .$$

Comandos R para o estudo da regressão

Alguns comandos para inspeccionar uma regressão linear:

Guardar a regressão do Exemplo 1

```
> Cabra.lm <- lm( leite ~ ano , data=Cabra )
```

- `fitted` devolve os valores ajustados $\hat{y}_i = b_0 + b_1\, x_i$:

```
> fitted(Cabra.lm)
       1        2        3        4        5        6        7        8        9       10
4.737154 5.029418 5.321683 5.613948 5.906212 6.198477 6.490742 6.783006 7.075271 7.367535
      11       12       13       14       15       16       17       18       19       20
7.659800 7.952065 8.244329 8.536594 8.828858 9.121123 9.413388 9.705652 9.997917 10.290182
      21       22       23       24       25       26
10.582446 10.874711 11.166975 11.459240 11.751505 12.043769
```

Comandos R (cont.)

- `residuals` devolve os resíduos $e_i = y_i - \hat{y}_i$:

```
> residuals(Cabra.lm)
         1          2          3          4          5          6          7          8
-0.40915385  0.58058154  1.22831692  0.87805231  0.23178769 -0.06247692 -0.29474154 -0.47900615
         9         10         11         12         13         14         15         16
 0.05772923  0.44946462  0.52220000 -1.57206462 -1.76532923 -0.15359385  0.13814154 -1.52012308
        17         18         19         20         21         22         23         24
-0.52738769 -0.55265231 -0.26891692  0.98281846  1.69155385  0.73428923 -0.70797538 -0.17124000
        25         26
 0.03249538  0.95723077
```

A Soma dos Quadrados dos Resíduos, *SQRE*, pode ser obtida por:

```
> sum(residuals(Cabra.lm)^2)

[1] 18.04768
```

SQRE tem unidades de medida: o quadrado das unidades de y.

Comandos R para a regressão (cont.)

- `predict` – ajusta uma regressão a novas observações, dadas numa *data frame* com nomes de preditores iguais aos do ajustamento.

Usar uma regressão ajustada para prever novos valores

```
> novos <- data.frame( ano=c(1985,2012) )
> predict( Cabra.lm , new=novos )
       1        2
4.444889 12.336034
```

Assim, o valor $\hat{y}$ ajustado pela recta, para $x = 2012$, é (arredondamentos aparte):

$$\begin{aligned} \hat{y} &= b_0 + b_1\, x \\ \Leftrightarrow \quad 12.336034 &= -575.7004 + 0.2923 \times 2012\ . \end{aligned}$$

Comandos R para criar o gráfico do acetato 22

O gráfico do acetato 22 foi criado usando os seguintes comandos:

```
# Repetir a criação da nuvem de pontos:
> plot(Cabra , pch=16 , col="red", ylab=expression(paste("leite (",10^6,"litros)")))

# Ajustar e guardar a regressão linear:
> Cabra.lm <- lm( leite ~ ano , data=Cabra )

# Usar o comando abline para traçar a recta na nuvem de pontos:
> abline( Cabra.lm , col="blue")

# Usar o comando text para escrever a equação da recta:
> text( 1998 , 10 , "y=-575.7004 + 0.2923 x", col="blue")

# Criar um ciclo com o comando for que, para cada um dos 26 pontos, use o
comando lines para traçar um segmento de recta vertical entre cada ponto
observado e o correspondente ponto na recta:
> for (i in 1:26){
+     lines( c(Cabra$ano[i],Cabra$ano[i]) , c(Cabra$leite[i], fitted(Cabra.lm)[i]) )
+ }
```

O critério de mínimos quadrados

O critério de minimizar Soma de Quadrados dos Resíduos, $SQRE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$, tem subjacente, um pressuposto:

Numa RLS, o papel das 2 variáveis, x e y, não é simétrico.

- y – **variável resposta** ("dependente")
  - variável que se quer modelar, prever a partir da variável x.
- x – **variável preditora** ("independente" ou "explicativa")
  - variável que se admite conhecida, e com base na qual se pretende tirar conclusões sobre y.

O critério de mínimos quadrados (cont.)

O *i*-ésimo resíduo

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

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

Minimizar a soma de quadrados dos resíduos corresponde a minimizar a soma de quadrados dos "erros de previsão".

O critério tem subjacente a preocupação de prever o melhor possível a variável y, a partir da sua relação com o preditor x.

Revisão: RLS - contexto descritivo

Quantidades importantes na Regressão Linear (ver Exercícios):

SQT e SQR

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

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

Nota: $\overline{y}$ é a média, quer dos y_i observados, quer dos $\hat{y}_i$ ajustados.

SQRE

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

Nota: A média dos resíduos e_i é nula, ou seja, $\overline{e} = 0$.

Revisão: RLS - contexto descritivo (cont.)

Fórmula Fundamental da Regressão

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

Coeficiente de Determinação

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

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

Recordar que, numa regressão linear simples, R^2 é o quadrado do coeficiente de correlação linear entre preditor e variável resposta:

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

Regressão - um pouco de história

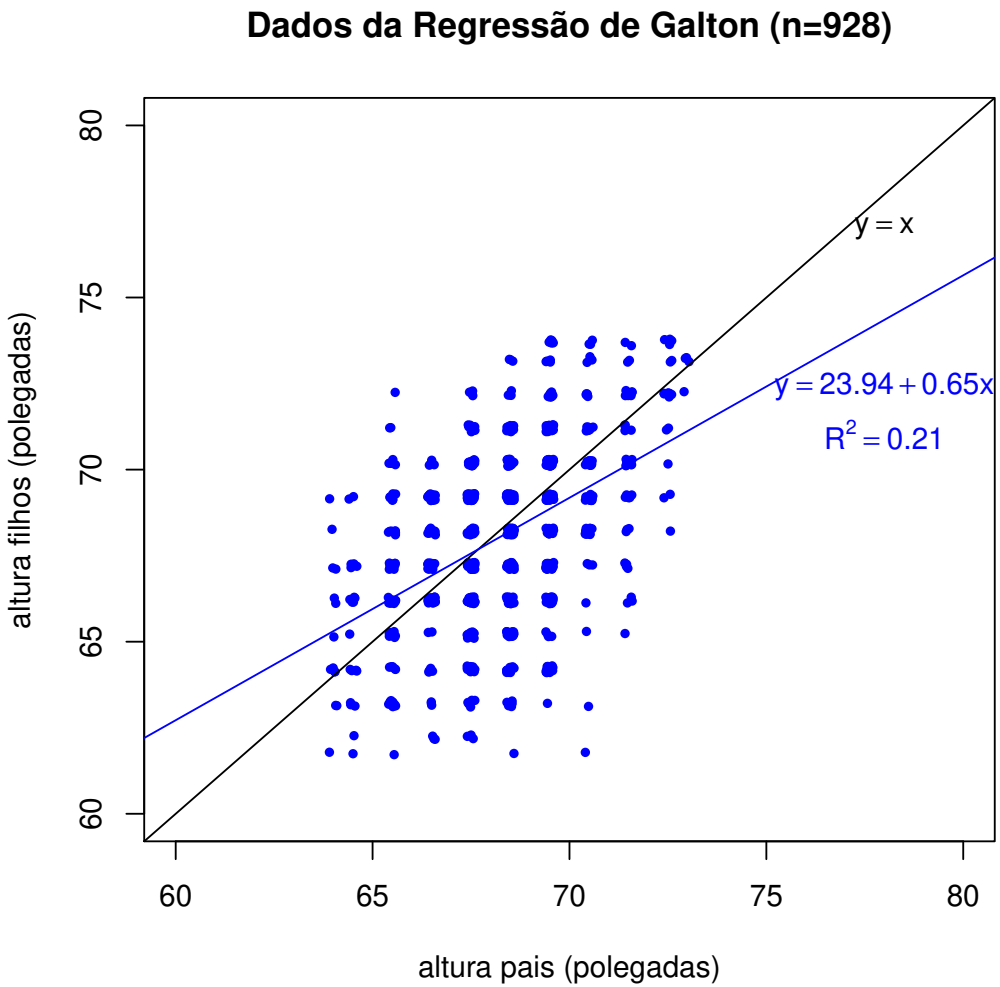
A designação Regressão tem origem num estudo de Francis Galton (1886), relacionando a altura de $n = 928$ jovens adultos com a altura (média) dos pais.

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

Galton chamou ao seu artigo *Regression towards mediocrity in hereditary stature*. A expressão regressão ficou associada ao método devido a esta acaso histórico. Galton foi também o inventor da designação eugenia, conceito do qual era grande defensor.

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

Um pouco de história (cont.)



Exemplo RLS (cont.)

O coeficiente de determinação no R pode ser obtido aplicando o comando `summary` a uma regressão ajustada. Numa extensa listagem de resultados (a maioria dizem respeito ao problema inferencial) surge, com a designação `Multiple R-Squared` o valor de R^2:

```
> summary(Cabra.lm)

Call: lm(formula = leite ~ ano, data = Cabra)
[...]
Residual standard error: 0.8672 on 24 degrees of freedom
Multiple R-squared:  0.8738,    Adjusted R-squared:  0.8685
F-statistic: 166.1 on 1 and 24 DF,  p-value: 2.807e-12
```

O valor de R^2 (com maior precisão) pode ser obtido da seguinte forma:

```
> summary(Cabra.lm)$r.sq
[1] 0.8737681
```

Extrair informação duma regressão ajustada

O comando `lm` produz um objecto de tipo `list`:

```
> is.list(Cabra.lm)  <– pergunta se o objecto Cabra.lm é uma lista

[1] TRUE

> names(Cabra.lm) <– pede os nomes dos objectos que compõem a lista

"coefficients" "residuals" "effects" "rank"  "fitted.values" "assign"
"qr"  "df.residual"  "xlevels"   "call"    "terms" "model"
```

Cada componente da lista pode ser extraído separando o nome da lista e da componente com um cifrão:

```
> Cabra.lm$coef <– nome pode estar incompleto, desde que não ambíguo

 (Intercept)          ano
-575.7003723    0.2922646
```

Para aprofundar o significado de cada componente da lista: `help(lm)`.

Extrair informação duma regressão (cont.)

O comando `summary`, quando aplicado a uma regressão ajustada, produz outro objecto de tipo `list`. Eis as componentes produzidas:

```
> names(summary(Cabra.lm))

[1] "call"          "terms"         "residuals"     "coefficients"
[5] "aliased"       "sigma"         "df"            "r.squared"
[9] "adj.r.squared" "fstatistic"    "cov.unscaled"
```

Componentes individuais podem ser extraídas desta lista de *output*, da forma já indicada.

Uma desvantagem do critério de minimizar SQRE

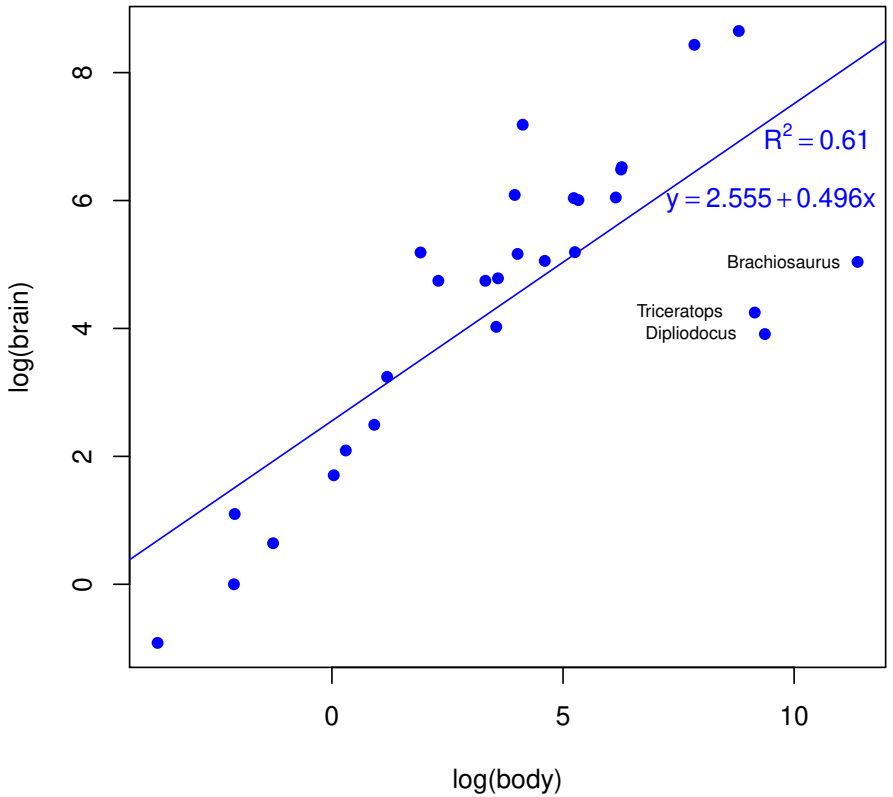
O critério de ajustamento usado (minimizar *SQRE*) tem uma desvantagem: é sensível à presença de observações atípicas.

Ilustremos com um conjunto de dados do módulo `MASS` (iniciais do livro *Modern Applied Statistics with S*, de Venables e Ripley) do `R`.

```
> library(MASS)
> help(Animals)

Animals                    package:MASS                    R Documentation
[...]
Average brain and body weights for 28 species of land animals.
[...]
'body' body weight in kg.
'brain' brain weight in g.
[...]
Source:
P. J. Rousseeuw and A. M. Leroy (1987) _Robust Regression and
Outlier Detection. Wiley, p. 57.
```

RLS e observações atípicas

A generalidade das observações seguem uma relação linear entre os logaritmos do peso do cérebro e do peso do corpo.
Mas três espécies de dinossáurios são observações atípicas, e condicionam o ajustamento da recta.



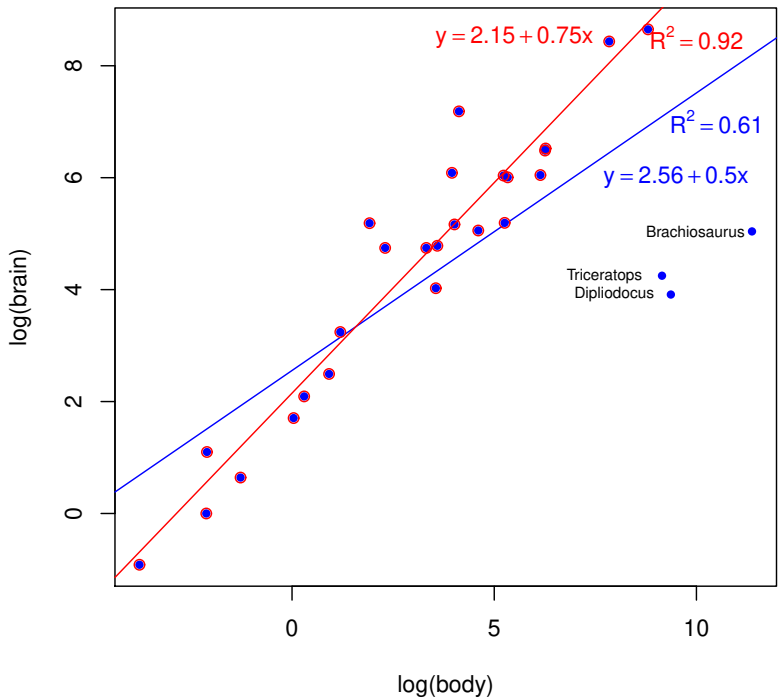
Os comandos no R

O gráfico do acetato anterior foi criado com os seguintes comandos:

```
> plot( log(brain) ~ log(body) , data=Animals, pch=16, col="blue")

> abline( lm(log(brain) ~ log(body) , data=Animals), col="blue")

> text( 10.5 , 7 , expression(R^2==0.61) , col="blue")

> text( 9.5 , 6 , expression(y==2.555+0.496*x) , col="blue")

> text( log(Animals[c(6,16,26),1])-1.6 , log(Animals[c(6,16,26),2]) ,
+       lab=rownames(Animals)[c(6,16,26)] , cex=0.7 )
```

RLS e observações atípicas (cont.)

A exclusão dessas três observações influencia, quer a recta ajustada, quer a sua qualidade.



Embora neste caso seja aceitável a exclusão das 3 observações atípicas (pois pertencem a "outra realidade", de espécies extintas), existem alternativas que usam critérios alternativos de ajustamento robustos.

Relações não lineares e transformações linearizantes

Nalguns casos, uma relação de fundo não linear entre x e y pode ser linearizada através de transformações duma, ou ambas, as variáveis.

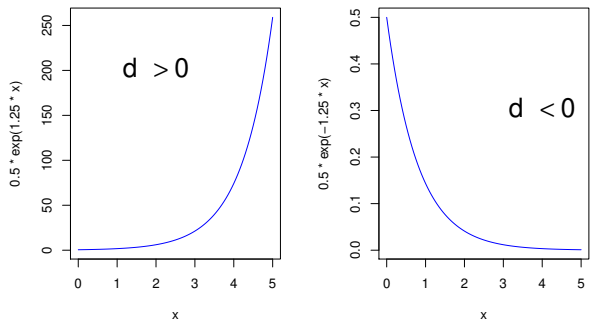
Tais transformações podem permitir utilizar uma regressão linear simples, apesar de a relação original ser não linear.

Estas transformações linearizantes são extensíveis ao caso de haver mais do que um preditor.

Consideremos alguns exemplos particularmente frequentes de relações não lineares que são linearizáveis através de transformações da variável resposta e, nalguns casos, também do preditor.

Relação exponencial

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



Transformação : Logaritmizando, obtém-se:

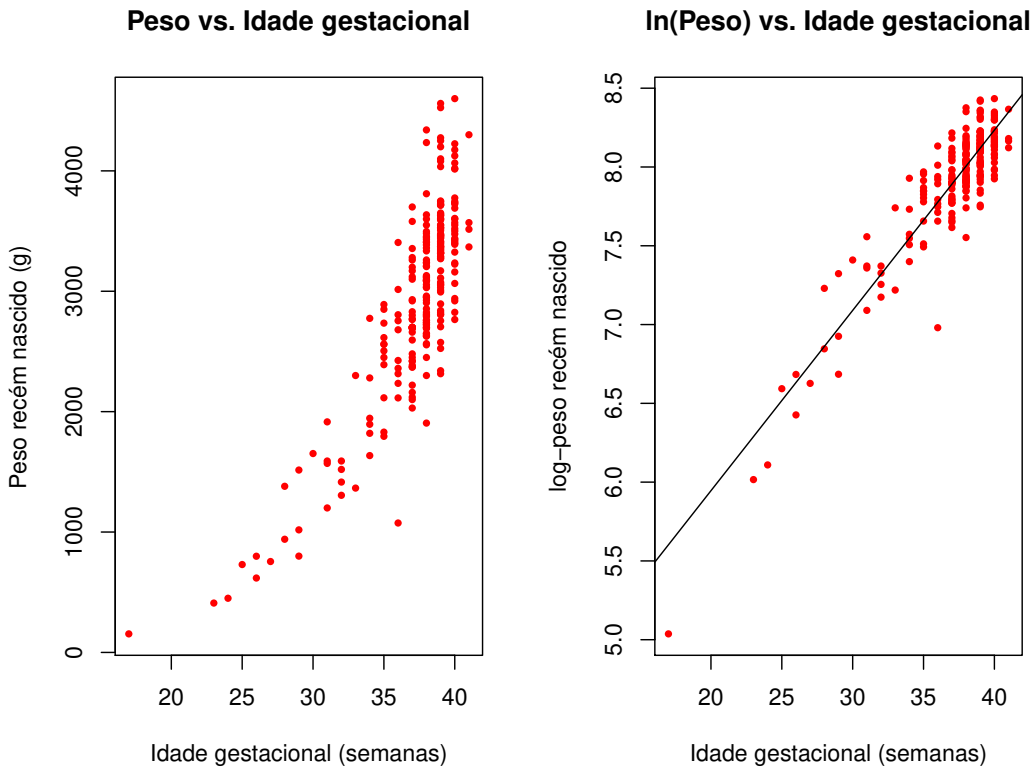
$$\begin{aligned} \ln(y) &= \ln(c) + d\,x \\ \Leftrightarrow \quad y^* &= b_0 + b_1\,x \end{aligned}$$

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

Tem-se $b_0 = \ln(c)$ e $b_1 = d$. O sinal do declive da recta indica se a relação exponencial original é crescente $(b_1 > 0)$ ou decrescente $(b_1 < 0)$.

Uma linearização no Exemplo 4

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



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

Ainda a relação exponencial

Uma relação exponencial resulta de admitir que y é função de x e que a taxa de variação de y, $y'(x)$, é proporcional a y:

$$y'(x) = b_1 \cdot y(x) \ ,$$

i.e., que a taxa de variação relativa de y é constante:

$$\frac{y'(x)}{y(x)} = b_1 \ .$$

Primitivando (em ordem a x), tem-se (admitindo $y(x) > 0$):

$$\ln[y(x)] \ = \ b_1 \, x + K \qquad \Leftrightarrow \qquad y(x) \ = \ \underbrace{\mathrm{e}^{K}}_{=c} \, \mathrm{e}^{b_1 \, x} \ .$$

onde K é a constante de primitivação que, no nosso contexto, corresponde à constante aditiva $b_0 = \ln(c)$.
O declive b_1 da recta é a taxa de variação relativa constante.

Modelo exponencial de crescimento populacional

Um modelo exponencial é frequentemente usado para descrever o crescimento de populações, numa fase inicial onde não se faz ainda sentir a escassez de recursos limitantes.

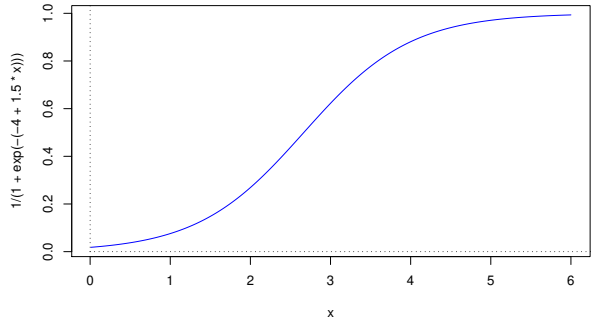
Mas nenhum crescimento populacional exponencial é sustentável a longo prazo.

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

Considera-se aqui uma versão simplificada (com 2 parâmetros) duma curva logística, associada a uma variável resposta que representa a proporção da população em relação ao seu máximo (a capacidade de sustentação do meio).

Relação Logística

Relação Logística : $y = \frac{1}{1+e^{-(b_0+b_1 x)}}$



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

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

Ainda a Logística

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

$$\frac{y'(x)}{y(x)} = b_1 \left[1 - y(x)\right] .$$

De facto, a expressão anterior equivale a:

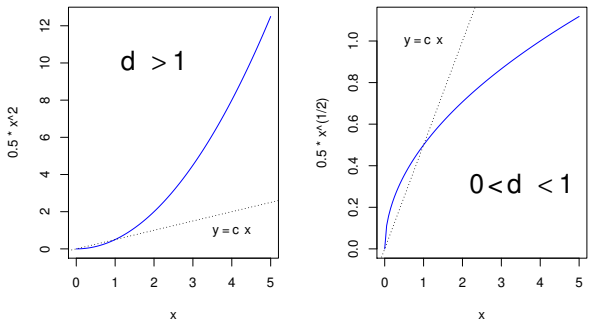
$$\frac{y'(x)}{y(x)\left(1-y(x)\right)} = b_1 \quad \Leftrightarrow \quad \frac{y'(x)}{1-y(x)} + \frac{y'(x)}{y(x)} = b_1$$

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

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

Relação potência (ou alométrica)

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



Transformação : Logaritmizando, obtém-se:

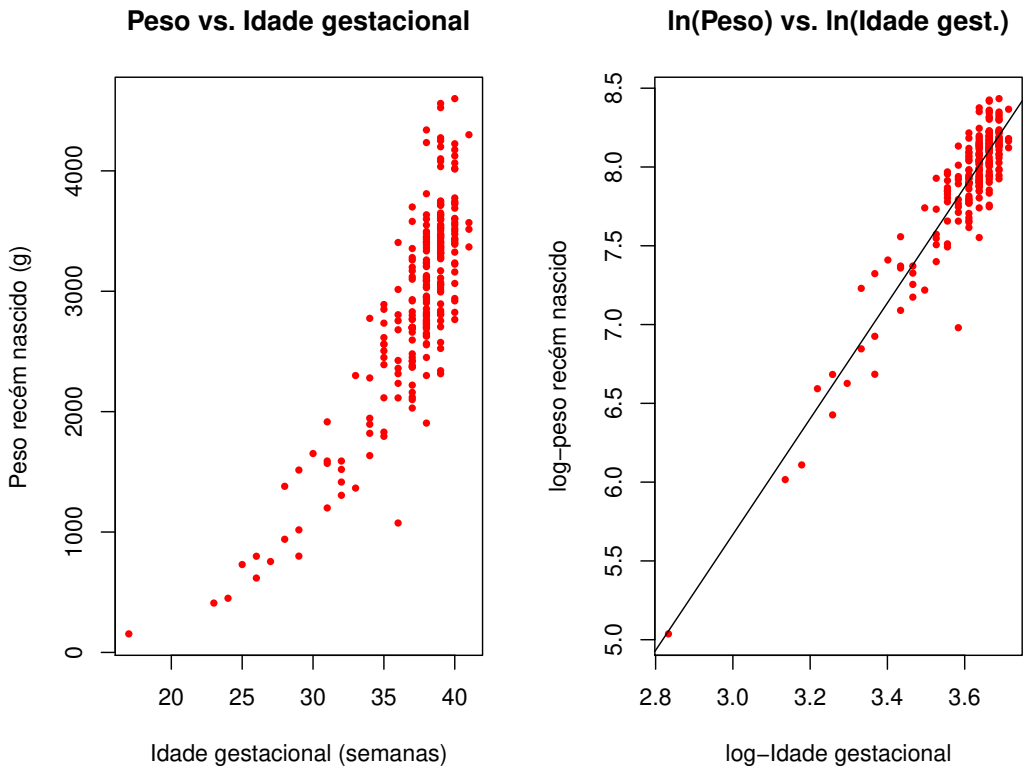
$$\begin{aligned} \ln(y) &= \ln(c) + d\ln(x) \\ \Leftrightarrow \quad y^* &= b_0 + b_1 x^* \end{aligned}$$

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

O declive b_1 da recta é o expoente na relação potência original ($b_1 = d$). Mas $b_0 = \ln(c)$.

Outra linearização no Exemplo 4

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



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

Ainda a relação potência

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

$$\frac{y'(t)}{y(t)} = b_1 \cdot \frac{x'(t)}{x(t)} \ .$$

De facto, primitivando (em ordem a t), tem-se (admitindo $x, y > 0$):

$$\ln y \ = \ b_1 \ln x + K$$

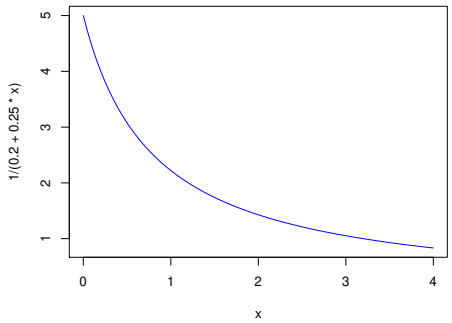
e exponenciando,

$$y \ = \ x^{b_1} \cdot \underbrace{\mathrm{e}^K}_{=b_0}$$

A relação potência é muito usado em estudos de alometria, que comparam o crescimento de partes diferentes dum organismo.
A isometria corresponde ao valor $b_1 = 1$.

Relação de tipo hiperbólico

Relação hiperbólica : $y = \dfrac{1}{b_0 + b_1\, x}$.
$(x,y>0 \quad ; \quad b_0,b_1>0)$



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

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

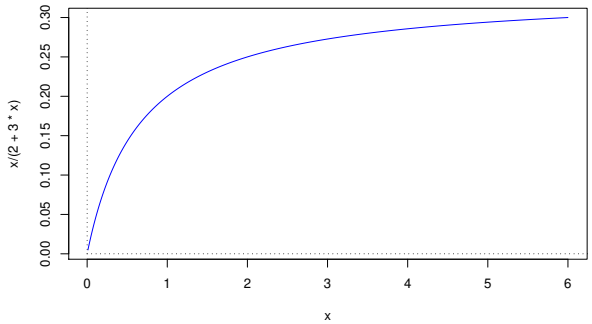
Resulta de admitir que a taxa de variação de y é proporcional ao quadrado de y ou, equivalentemente, que a taxa de variação relativa de y é proporcional a y:

$$y'(x) = -b_1\, y^2(x) \qquad \Leftrightarrow \qquad \frac{y'(x)}{y(x)} = -b_1\, y(x)\,.$$

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

Relação Michaelis-Menten

Relação Michaelis-Menten : $y = \frac{x}{c + d\,x}$



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

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

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

Relação Michaelis-Menten (cont.)

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

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

Advertência sobre transformações linearizantes

A regressão linear simples não modela directamente relações não lineares entre x e y. Pode modelar uma relação linear entre as variáveis transformadas.

Transformações da variável-resposta Y têm um impacto grande no ajustamento: a escala dos resíduos é alterada.

Conceitos que dependem da escala de Y, como $SQRE$ e R^2, não são directamente comparáveis, antes e após uma transformação da variável resposta.

Nota: Linearizar, obter os parâmetros b_0 e b_1 da recta e depois desfazer a transformação linearizante não produz os mesmos valores de parâmetros que resultariam de minimizar a soma de quadrados dos resíduos directamente na relação não linear, através duma Regressão não linear.

A Regressão Linear Múltipla

Por vezes, é necessário mais do que uma variável preditora para modelar a variável resposta de interesse.

Exemplo 7: dados `Antoci`

Num estudo sobre uma população experimental de clones da casta Tinta Francisca, realizado no Tabuaço em 2003, foram medidos os valores das seguintes variáveis para 24 videiras:

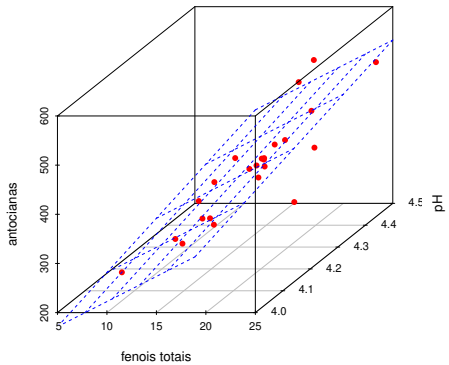
- teor de antocianas (variável `antoci`, em mg/dm^3);
- fenóis totais (variável `fentot`);
- pH (variável `pH`).

Há interesse em estudar a relação entre o teor de antocianas (variável resposta) e o teor de fenóis totais e pH.

As $n = 24$ observações em três variáveis geram uma nuvem de 24 pontos em $\mathbb{R}^3$, que parece dispôr-se em torno de um plano.

O gráfico do Exemplo 7

O gráfico foi feito usando a função `scatterplot3d`, no módulo do `R` com o mesmo nome. Concretamente, foram usados os seguintes comandos:




```
# Carregar o módulo scatterplot3d:
> library(scatterplot3d)

# Criar, e guardar, a nuvem de pontos a 3D:
> s3d <- scatterplot3d( Antoci$fentot, Antoci$pH, Antoci$antoci,
+     pch=16, color="red", xlab="fenois totais", ylab="pH", zlab="antocianas")

# Desenhar um plano ajustado à nuvem:
> s3d$plane3d( lm(antoci ~ fentot + pH , data=Antoci) , col="blue")

# Assinale-se que a list de saída do comando scatterplot3d contém
# um objecto plane3d que é uma função.
```

Plano em $\mathbb{R}^3$

Qualquer plano em $\mathbb{R}^3$, no sistema $x0y0z$, tem equação

$$Ax + By + Cz + D = 0 \ .$$

No nosso contexto, e colocando:

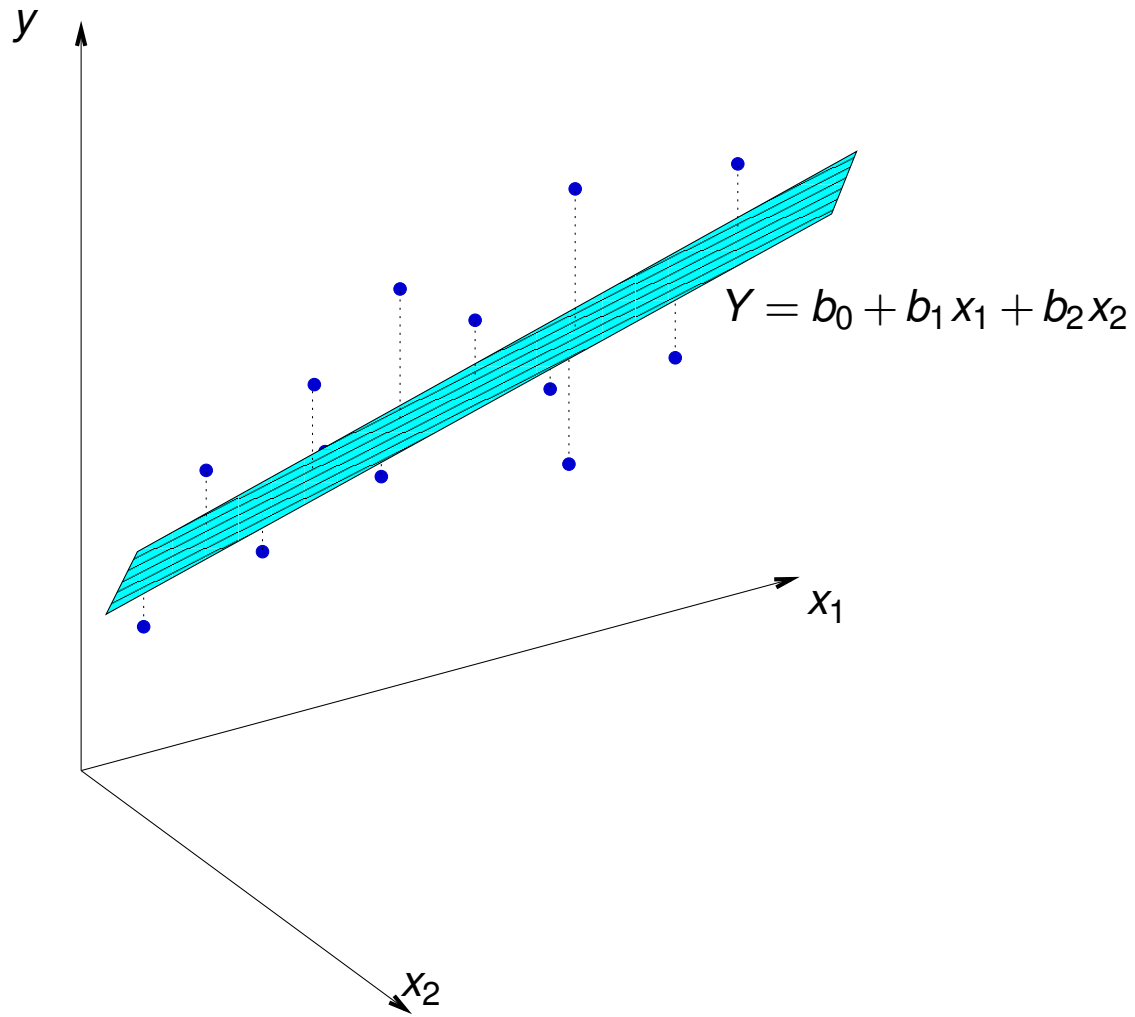
- no eixo vertical (z) a variável resposta Y;
- noutro eixo (x) um preditor X_1;
- no terceiro eixo (y) o outro preditor X_2,

A equação fica (se $C \neq 0$, i.e., para planos não verticais):

$$\begin{aligned} Ax_1 + Bx_2 + Cy + D = 0 \quad &\Leftrightarrow \quad y = -\frac{D}{C} - \frac{A}{C}x_1 - \frac{B}{C}x_2 \\ &\Leftrightarrow \quad y = b_0 + b_1 x_1 + b_2 x_2 \end{aligned}$$

A equação estende a equação da recta, para o caso de 2 preditores.

Regressão Múltipla - representação gráfica ($p = 2$)



$y = b_0 + b_1 x_1 + b_2 x_2$ é a equação dum plano em $\mathbb{R}^3$ $(x_1 0 x_2 0 y)$.
Pode ser ajustado pelo mesmo critério que na RLS: minimizar SQRE.

O caso geral: p preditores

Pretende-se modelar uma variável resposta, Y, com base em p variáveis preditoras, x_1, x_2, ..., x_p. Dispõe-se de n conjuntos de observações:

$$\left\{ \left(x_{1(i)}, x_{2(i)}, \ldots x_{p(i)}, y_i \right) \right\}_{i=1}^{n} .$$

Problema: A representação usual deixa de ser visualizável, uma vez que as observações definem uma nuvem de n pontos no espaço $\mathbb{R}^{p+1}$.

As características fundamentais da representação usual são:

- $p+1$ eixos – um para cada variável em questão.
- n pontos – um para cada indivíduo (unidade experimental) observado.
- Tem-se uma nuvem de n pontos num espaço $(p+1)$-dimensional.

O hiperplano ajustado

Admite-se que os valores de Y oscilam em torno duma combinação linear (afim) das p variáveis preditoras:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p \ .$$

Trata-se da equação dum hiperplano em $\mathbb{R}^{p+1}$.

O critério utilizado para ajustar um hiperplano à nuvem de n pontos em $\mathbb{R}^{p+1}$ é o mesmo: minimizar a Soma de Quadrados dos Resíduos, ou seja, escolher os $p+1$ parâmetros $\{b_j\}_{j=0}^{p}$ que minimizem:

$$SQRE = \sum_{i=}^{n} e_i^2 = \sum_{i=}^{n} (y_i - \hat{y}_i)^2$$

onde os y_i representam os valores observados da variável resposta e

$$\hat{y}_i = b_0 + b_1 x_{1(i)} + b_2 x_{2(i)} + \ldots + b_p x_{p(i)}$$

os valores ajustados pela equação do hiperplano.

Duas abordagens para a estimação dos parâmetros

Para obter os parâmetros que definem o hiperplano que melhor se ajusta às observações pode-se usar uma abordagem:

- analítica; ou
- geométrica.

Nas duas abordagens, a notação vectorial-matricial é vantajosa.

Não existem fórmulas simples, como no caso da RLS, para cada um dos parâmetros b_j isoladamente. Mas é possível indicar uma fórmula única matricial para o conjunto dos $p+1$ parâmetros do modelo.

Vamos seguir uma abordagem geométrica.

Representação alternativa: o espaço das variáveis

A representação gráfica de n observações de Y e das variáveis preditoras em $\mathbb{R}^{p+1}$ não é a única possível.

Há outra representação possível dos dados, que casa conceitos geométricos e conceitos estatísticos.

As n observações de Y definem um vector de n coordenadas em $\mathbb{R}^n$:

$$\vec{\mathbf{y}} = (y_1, y_2, y_3, ..., y_n)^t .$$

Da mesma forma, as n observações duma dada variável preditora definem um vector em $\mathbb{R}^n$:

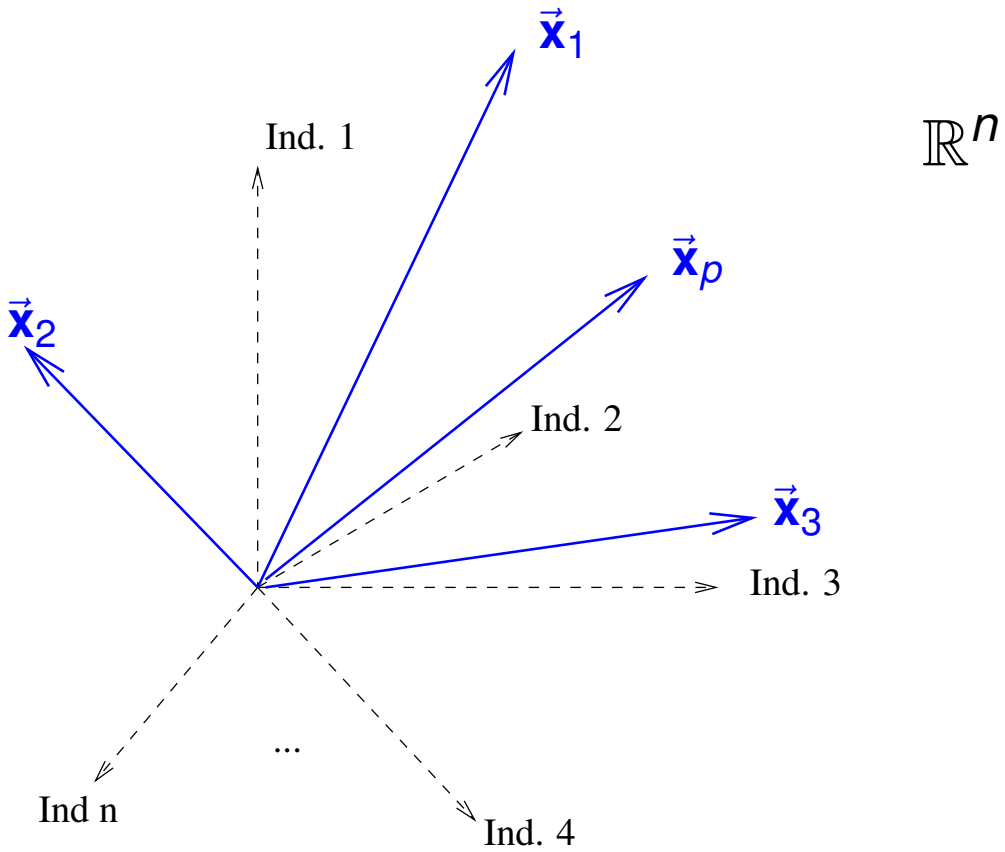
$$\vec{\mathbf{x}}_j = (x_{j_{(1)}}, x_{j_{(2)}}, x_{j_{(3)}}, ..., x_{j_{(n)}})^t \quad (j = 1, 2, ..., p).$$

Logo, podemos representar as variáveis por pontos/vectores em $\mathbb{R}^n$.

A representação em $\mathbb{R}^n$

- cada eixo corresponde a um indivíduo observado;
- cada vector corresponde a uma variável.

O vector de n uns, representado por $\vec{\mathbf{1}}_n$, também é um vector de $\mathbb{R}^n$.



O vector de valores ajustados

Os n valores ajustados $\hat{y}_i$ também definem um vector de $\mathbb{R}^n$, que é uma combinação linear dos vectores $\vec{\mathbf{1}}_n$, $\vec{\mathbf{x}}_1$, $\vec{\mathbf{x}}_2$, ..., $\vec{\mathbf{x}}_p$:

$$\vec{\hat{\mathbf{y}}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \dots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} b_0 + b_1 x1_{(1)} + b_2 x2_{(1)} + \ldots + b_p xp_{(1)} \\ b_0 + b_1 x1_{(2)} + b_2 x2_{(2)} + \ldots + b_p xp_{(2)} \\ b_0 + b_1 x1_{(3)} + b_2 x2_{(3)} + \ldots + b_p xp_{(3)} \\ \dots \\ b_0 + b_1 x1_{(n)} + b_2 x2_{(n)} + \ldots + b_p xp_{(n)} \end{bmatrix}$$

$$= b_0 \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + b_1 \begin{bmatrix} x1_{(1)} \\ x1_{(2)} \\ x1_{(3)} \\ \vdots \\ x1_{(n)} \end{bmatrix} + \ldots + b_p \begin{bmatrix} xp_{(1)} \\ xp_{(2)} \\ xp_{(3)} \\ \vdots \\ xp_{(n)} \end{bmatrix}$$

$$= b_0 \vec{\mathbf{1}}_n + b_1 \vec{\mathbf{x}}_1 + b_2 \vec{\mathbf{x}}_2 + \ldots + b_p \vec{\mathbf{x}}_p$$

O vector $\vec{\hat{\mathbf{y}}}$ dos valores ajustados pode também escrever-se como um produto envolvendo uma matriz $\mathbf{X}$ cujas colunas sejam os vectores $\vec{\mathbf{1}}_n$, $\vec{\mathbf{x}}_1$, ..., $\vec{\mathbf{x}}_p$, ou seja, o vector de n uns e os vectores com as n observações de cada um dos preditores.

A matriz do modelo $\mathbf{X}$

Defina-se a matriz de dimensão $n \times (p+1)$, cuja primeira coluna é a coluna de n uns e cujas restantes colunas são dadas pelas n observações de cada uma das variáveis preditoras:

A matriz $\mathbf{X}$ do modelo

$$\mathbf{X} = \underbrace{\begin{bmatrix} 1 & x_{1_{(1)}} & x_{2_{(1)}} & \cdots & x_{p_{(1)}} \\ 1 & x_{1_{(2)}} & x_{2_{(2)}} & \cdots & x_{p_{(2)}} \\ 1 & x_{1_{(3)}} & x_{2_{(3)}} & \cdots & x_{p_{(3)}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1_{(n)}} & x_{2_{(n)}} & \cdots & x_{p_{(n)}} \end{bmatrix}}_{\begin{matrix} \cdots \\ \end{matrix}}$$

$$\underbrace{}_{=\vec{\mathbf{1}}_n} \quad \underbrace{}_{=\vec{\mathbf{x}}_1} \quad \underbrace{}_{=\vec{\mathbf{x}}_2} \quad \underbrace{}_{=\vec{\mathbf{x}}_p}$$

Um produto matricial $\mathbf{X}\vec{\mathbf{a}}$

O produto duma matriz $\mathbf{X} = \left[\vec{\mathbf{1}}_n | \vec{\mathbf{x}}_1 | \vec{\mathbf{x}}_2 | ... | \vec{\mathbf{x}}_p\right]$ por um qualquer vector $\vec{\mathbf{a}} \in \mathbb{R}^{p+1}$ é sempre uma combinação linear das colunas de $\mathbf{X}$:

$$
\begin{aligned}
\mathbf{X}\vec{\mathbf{a}} &= \begin{bmatrix} 1 & x_{1_{(1)}} & x_{2_{(1)}} & \cdots & x_{p_{(1)}} \\ 1 & x_{1_{(2)}} & x_{2_{(2)}} & \cdots & x_{p_{(2)}} \\ 1 & x_{1_{(3)}} & x_{2_{(3)}} & \cdots & x_{p_{(3)}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1_{(n)}} & x_{2_{(n)}} & \cdots & x_{p_{(n)}} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} \\
&= \begin{bmatrix} a_0 + a_1 x_{1_{(1)}} + a_2 x_{2_{(1)}} + ... + a_p x_{p_{(1)}} \\ a_0 + a_1 x_{1_{(2)}} + a_2 x_{2_{(2)}} + ... + a_p x_{p_{(2)}} \\ a_0 + a_1 x_{1_{(3)}} + a_2 x_{2_{(3)}} + ... + a_p x_{p_{(3)}} \\ \cdots \\ a_0 + a_1 x_{1_{(n)}} + a_2 x_{2_{(n)}} + ... + a_p x_{p_{(n)}} \end{bmatrix} \\
&= a_0 \vec{\mathbf{1}}_n + a_1 \vec{\mathbf{x}}_1 + a_2 \vec{\mathbf{x}}_2 + ... + a_p \vec{\mathbf{x}}_p
\end{aligned}
$$

A matriz $\mathbf{X}$ e o seu subespaço de colunas

- O conjunto de todas as combinações lineares dos $p+1$ vectores $\vec{\mathbf{1}}_n$, $\vec{\mathbf{x}}_1$, ..., $\vec{\mathbf{x}}_p$ chama-se o subespaço gerado por esses vectores.

  Nota: Subespaços são rectas (dimensão 1), planos (dimensão 2) ou hiperplanos (dimensão ≥ 3) que contêm a origem.

- É um subespaço de dimensão $p+1$ (se os $p+1$ vectores forem linearmente independentes).

- Colocando os vectores $\vec{\mathbf{1}}_n$, $\vec{\mathbf{x}}_1$, ..., $\vec{\mathbf{x}}_p$ nas colunas duma matriz $\mathbf{X}$ (de dimensões $n \times (p+1)$) podemos chamar a este subespaço o subespaço das colunas da matriz $\mathbf{X}$, $\mathscr{C}(\mathbf{X}) \subset \mathbb{R}^n$.

- Qualquer combinação linear de colunas da matriz $\mathbf{X}$ é dada por $\mathbf{X}\vec{\mathbf{a}}$, onde $\vec{\mathbf{a}} = (a_0, a_1, a_2, ..., a_p)^t$ é o vector dos coeficientes que define a combinação linear.

Os parâmetros

- Cada escolha possível de coeficientes $\vec{\mathbf{a}} = (a_0, a_1, a_2, ..., a_p)^t$ corresponde a um ponto/vector $\mathbf{X}\vec{\mathbf{a}}$ no subespaço $\mathscr{C}(\mathbf{X})$.
- Essa escolha de coeficientes é única caso as colunas de $\mathbf{X}$ sejam linearmente independentes, isto é, se não houver dependência linear dos vectores $\vec{\mathbf{x}}_1, ..., \vec{\mathbf{x}}_p, \vec{\mathbf{1}}_n$.
- Um dos pontos/vectores do subespaço é a combinação linear dada pelo vector de coeficientes $\vec{\mathbf{b}} = (b_0, b_1, ..., b_p)$ que minimiza *SQRE*. É a combinação linear que desejamos determinar.

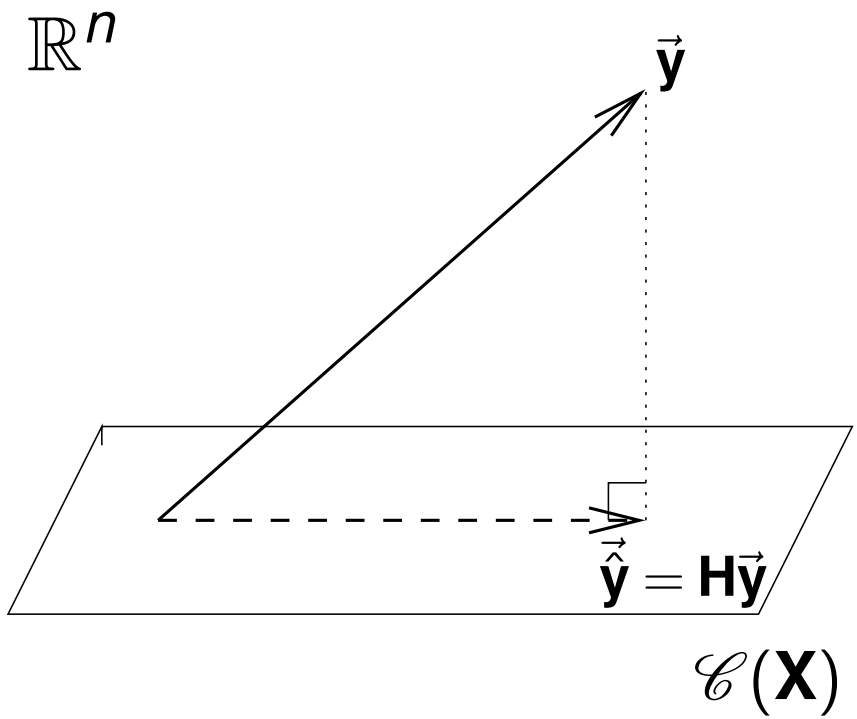
Como identificar esse ponto/vector?

Os parâmetros (cont.)

- Dispomos de um vector de n observações de $\vec{\mathbf{y}}$ que está em $\mathbb{R}^n$ mas, em geral, não está no subespaço $\mathscr{C}(\mathbf{X})$.
- Queremos aproximar esse vector por outro vector, $\vec{\hat{\mathbf{y}}} = b_0\vec{\mathbf{1}}_n + b_1\vec{\mathbf{x}}_1 + \ldots + b_p\vec{\mathbf{x}}_p$, que está no subespaço $\mathscr{C}(\mathbf{X})$.
- Vamos aproximar o vector de observações $\vec{\mathbf{y}}$ pelo vector $\vec{\hat{\mathbf{y}}}$ do subespaço $\mathscr{C}(\mathbf{X})$ que está mais próximo de $\vec{\mathbf{y}}$.

SOLUÇÃO: Tomar a projecção ortogonal de $\vec{\mathbf{y}}$ sobre $\mathscr{C}(\mathbf{X})$: $\vec{\hat{\mathbf{y}}} = \mathbf{H}\vec{\mathbf{y}}$.

O conceito geométrico subjacente à identificação de $\vec{\hat{\mathbf{y}}}$



O vector de $\mathscr{C}(\mathbf{X}) \subset \mathbb{R}^n$ mais próximo dum vector $\vec{\mathbf{y}} \in \mathbb{R}^n$ é o vector $\vec{\hat{\mathbf{y}}}$ que resulta de projectar ortogonalmente $\vec{\mathbf{y}}$ sobre $\mathscr{C}(\mathbf{X})$.

A projecção ortogonal cria um triângulo rectângulo em $\mathbb{R}^n$.

O critério minimiza *SQRE*

O critério de escolher o vector $\vec{\hat{\mathbf{y}}} \in \mathscr{C}(\mathbf{X})$ que minimize a distância ao vector de observações $\vec{\mathbf{y}}$ significa que minimizamos o quadrado dessa distância, que é dado por:

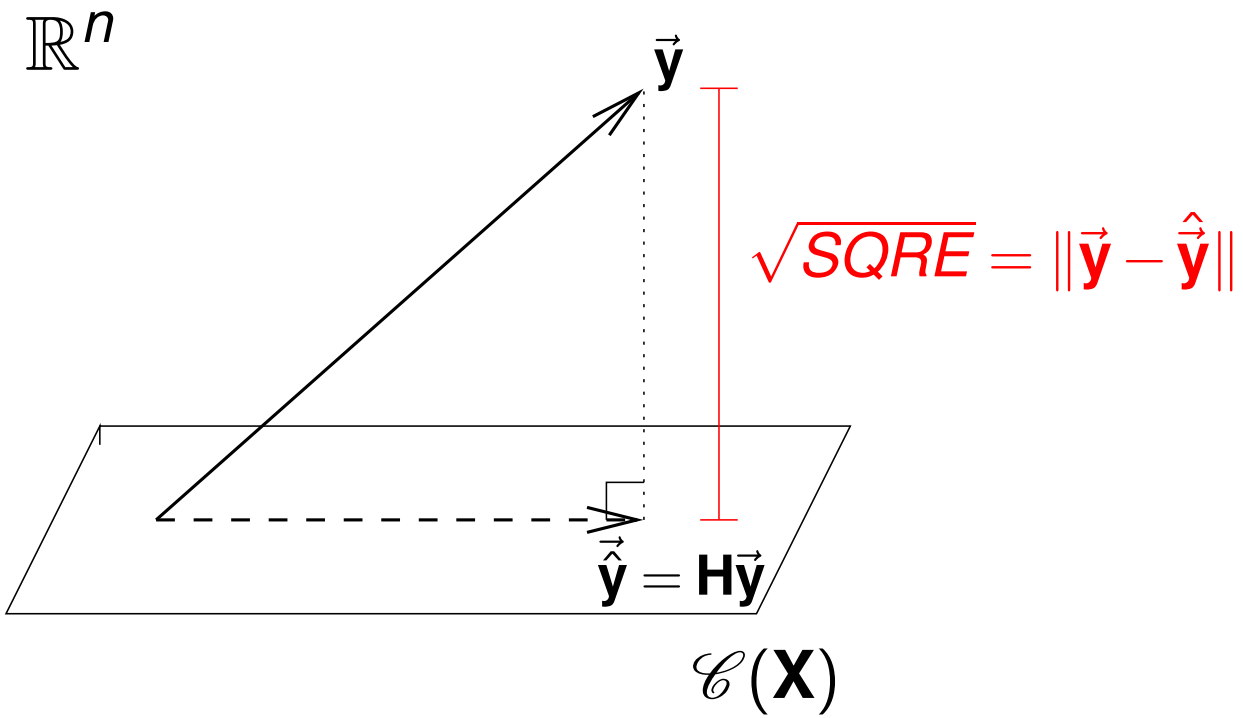
$$dist^2(\vec{\mathbf{y}}, \vec{\hat{\mathbf{y}}}) \;=\; \|\vec{\mathbf{y}} - \vec{\hat{\mathbf{y}}}\|^2 \;=\; \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \;=\; SQRE \;,$$

ou seja, que minimizamos a soma de quadrados dos resíduos.

Recordar: Para qualquer vector $\vec{\mathbf{x}} = (x_1, x_2, ..., x_n)^t$, a norma de $\vec{\mathbf{x}}$ define-se como $\|\vec{\mathbf{x}}\| = \sqrt{\vec{\mathbf{x}}^t \vec{\mathbf{x}}} = \sqrt{\sum_{i=1}^{n} x_i^2}$.

O critério geométrico é equivalente ao critério estatístico usado para ajustar os parâmetros na Regressão Linear.

O conceito geométrico subjacente à obtenção de $\vec{\hat{\mathbf{y}}}$



O quadrado da distância de $\vec{\mathbf{y}}$ a $\vec{\hat{\mathbf{y}}}$ é *SQRE*, a soma dos quadrados dos resíduos.

A projecção ortogonal

A projecção ortogonal de um vector $\vec{\mathbf{y}} \in \mathbb{R}^n$ sobre o subespaço $\mathscr{C}(\mathbf{X})$ gerado pelas colunas (linearmente independentes) de $\mathbf{X}$ faz-se pré-multiplicando $\vec{\mathbf{y}}$ pela matriz de projecção ortogonal sobre $\mathscr{C}(\mathbf{X})$:

Matriz de projecção ortogonal sobre $\mathscr{C}(\mathbf{X})$

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t .$$

As matrizes de projecção ortogonal $\mathbf{P}$ sobre algum subespaço de $\mathbb{R}^n$ são as matrizes $n \times n$:

- simétricas (isto é, $\mathbf{P}^t = \mathbf{P}$); e
- idempotentes (isto é, $\mathbf{PP} = \mathbf{P}$).

A matriz $\mathbf{H}$ tem estas propriedades (Exercício 13: verifique!).

A projecção ortogonal no contexto da RLM

No contexto duma regressão linear múltipla, tem-se:

$$\begin{aligned} & \vec{\hat{\mathbf{y}}} = \mathbf{H}\vec{\mathbf{y}} \\ \Leftrightarrow \quad & \vec{\hat{\mathbf{y}}} = \mathbf{X}\underbrace{(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{y}}}_{=\vec{\mathbf{b}}} \end{aligned}$$

A combinação linear dos vectores $\vec{\mathbf{1}}_n, \vec{\mathbf{x}}_1, ..., \vec{\mathbf{x}}_p$ que gera o vector mais próximo de $\vec{\mathbf{y}}$ tem coeficientes dados pelos elementos do vector $\vec{\mathbf{b}}$:

O vector de parâmetros ajustado

$$\vec{\mathbf{b}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{y}} \, .$$

As três Somas de Quadrados

Recordar as três Somas de Quadrados:

SQRE A Soma de Quadrados dos Resíduos:

$$SQRE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .$$

SQT A Soma de Quadrados Total:

$$SQT = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2 .$$

SQR A Soma de Quadrados associada à Regressão:

$$SQR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 = \sum_{i=1}^{n} \hat{y}_i^2 - n\overline{y}^2 .$$

Pitágoras e a Regressão

O Teorema de Pitágoras aplica-se em qualquer espaço euclideano $\mathbb{R}^n$.
No triângulo rectângulo do acetato 72 produz a seguinte relação:

$$\begin{aligned}
\|\vec{\mathbf{y}}\|^2 &= \|\vec{\hat{\mathbf{y}}}\|^2 + \|\vec{\mathbf{y}} - \vec{\hat{\mathbf{y}}}\|^2 \\
\Leftrightarrow \quad \sum_{i=1}^{n} y_i^2 &= \sum_{i=1}^{n} \hat{y}_i^2 + \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}_{= SQRE} \\
\Leftrightarrow \quad \sum_{i=1}^{n} y_i^2 - n\overline{y}^2 &= \sum_{i=1}^{n} \hat{y}_i^2 - n\overline{y}^2 + SQRE \\
\Leftrightarrow \quad SQT &= SQR + SQRE
\end{aligned}$$

Revisitando Pitágoras

Vimos que a relação fundamental da Regressão Linear ($SQT = SQR + SQRE$) resulta duma aplicação do Teorema de Pitágoras. Mas foi necessário introduzir a subtracção de $n\overline{y}^2$.
Um outro triângulo rectângulo é estatisticamente mais interessante.

Considere-se o vector centrado das observações da variável resposta, isto é, o vector cujo elemento genérico é $y_i - \overline{y}$. Este vector, que será designado $\vec{\mathbf{y}}^c$, obtém-se subtraíndo a $\vec{\mathbf{y}}$ o vector que repete n vezes $\overline{y}$:

$$\vec{\mathbf{y}}^c = \vec{\mathbf{y}} - (\overline{y})\,\vec{\mathbf{1}}_n = (y_1 - \overline{y}, y_2 - \overline{y}, \ldots, y_n - \overline{y})^t.$$

A norma deste vector é $\sqrt{SQT} = \|\vec{\mathbf{y}}^c\| = \sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}$.

Revisitando Pitágoras (cont.)

A projecção ortogonal do vector $\vec{\mathbf{y}}^c$ sobre o subespaço $\mathscr{C}(\mathbf{X})$ gera o vector:

$$\begin{aligned} & \mathbf{H}\vec{\mathbf{y}}^c &=& \mathbf{H}\left(\vec{\mathbf{y}} - (\overline{y}) \cdot \vec{\mathbf{1}}_n\right) \\ \Leftrightarrow \quad & \mathbf{H}\vec{\mathbf{y}}^c &=& \mathbf{H}\vec{\mathbf{y}} - (\overline{y}) \cdot \mathbf{H}\vec{\mathbf{1}}_n \\ \Leftrightarrow \quad & \mathbf{H}\vec{\mathbf{y}}^c &=& \hat{\vec{\mathbf{y}}} - (\overline{y}) \cdot \vec{\mathbf{1}}_n \end{aligned}$$

já que $\mathbf{H}\vec{\mathbf{1}}_n = \vec{\mathbf{1}}_n$, pois o vector $\vec{\mathbf{1}}_n$ já pertence ao subespaço $\mathscr{C}(\mathbf{X})$, logo fica invariante quando projectado nesse mesmo subespaço – ver Exercício 13.

O vector $\mathbf{H}\vec{\mathbf{y}}^c$ tem elemento genérico $\hat{y}_i - \overline{y}$, e a sua norma é

$$\|\mathbf{H}\vec{\mathbf{y}}^c\| = \sqrt{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2} = \sqrt{SQR} \, .$$

Revisitando Pitágoras (cont.)

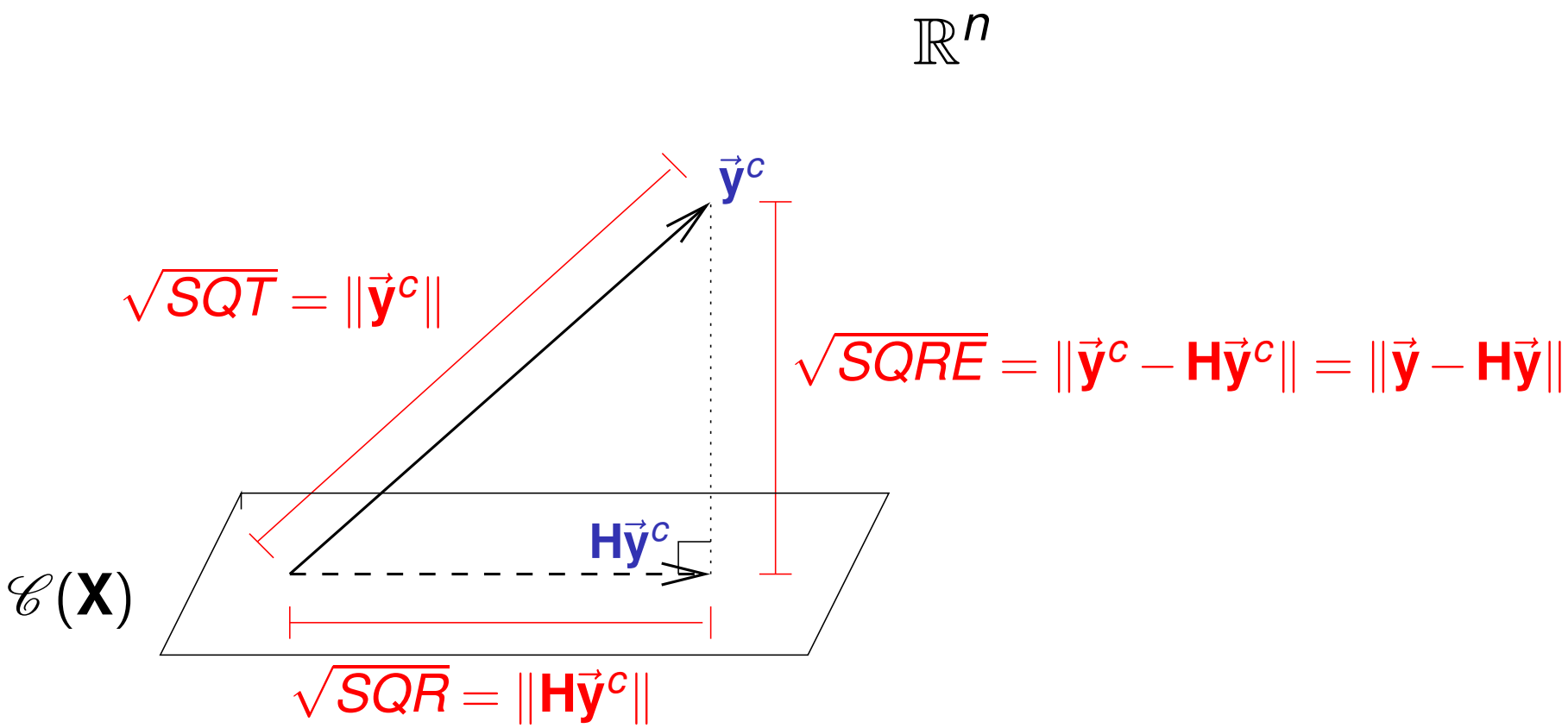
A distância entre o vector $\vec{\mathbf{y}}^c$ e a sua projecção ortogonal sobre $\mathscr{C}(\mathbf{X})$ continua a ser $\sqrt{SQRE}$:

$$\begin{aligned} & \vec{\mathbf{y}}^c - \mathbf{H}\vec{\mathbf{y}}^c &=& (\vec{\mathbf{y}} - \cancel{\overline{y}\vec{\mathbf{1}}_n}) - (\hat{\vec{\mathbf{y}}} - \cancel{\overline{y}\vec{\mathbf{1}}_n}) \\ \Leftrightarrow \quad & \vec{\mathbf{y}}^c - \mathbf{H}\vec{\mathbf{y}}^c &=& \vec{\mathbf{y}} - \hat{\vec{\mathbf{y}}} \end{aligned}$$

pelo que

$$\|\vec{\mathbf{y}}^c - \mathbf{H}\vec{\mathbf{y}}^c\| = \|\vec{\mathbf{y}} - \hat{\vec{\mathbf{y}}}\| = \sqrt{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} = \sqrt{SQRE} \ .$$

Revisitando Pitágoras (cont.)



A fórmula fundamental da Regressão Linear, $SQT = SQR + SQRE$, é uma aplicação directa do Teorema de Pitágoras ao triângulo definido por $\vec{\mathbf{y}}^c$ e a sua projecção ortogonal sobre $\mathscr{C}(\mathbf{X})$.

Pitágoras e o Coeficiente de Determinação

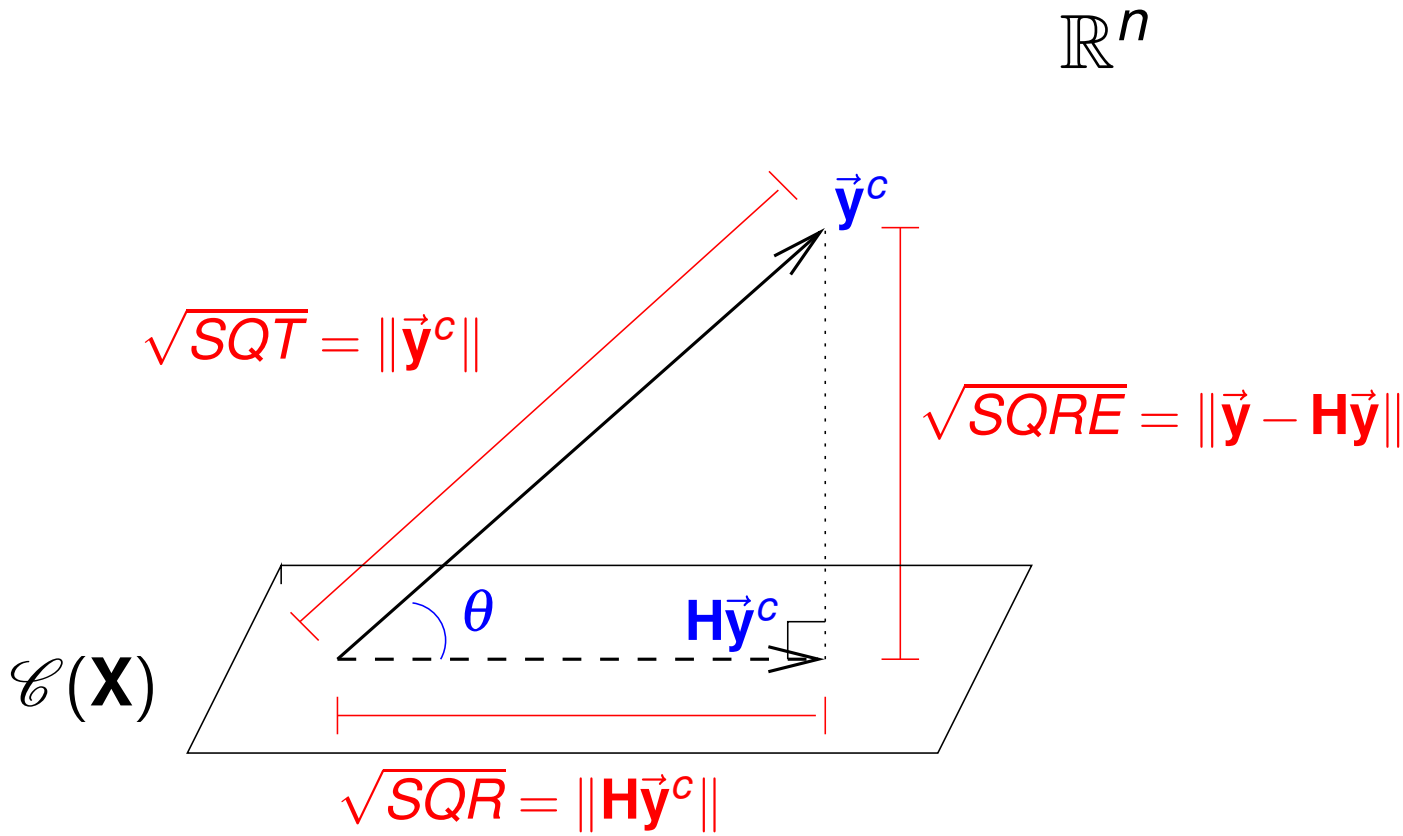
Torna-se evidente outra relação importante entre a geometria no espaço $\mathbb{R}^n$ e a estatística da Regressão Linear:

O coeficiente de determinação $R^2 = \frac{SQR}{SQT}$ é o cosseno ao quadrado do ângulo entre o vector centrado das observações da variável resposta, $\vec{\mathbf{y}}^c$, e a sua projecção ortogonal sobre o subespaço $\mathscr{C}(\mathbf{X})$:

$$\cos^2(\theta) = \frac{SQR}{SQT} = \mathrm{R}^2 ,$$

onde θ é o ângulo entre os vectores $\vec{\mathbf{y}}^c$ e $\mathbf{H}\vec{\mathbf{y}}^c$.

Pitágoras e o Coeficiente de Determinação (cont.)



O Coeficiente de Determinação na Regressão Linear, $R^2 = \frac{SQR}{SQT}$, é o cosseno ao quadrado do ângulo entre $\vec{\mathbf{y}}^c$ e $\mathbf{H}\vec{\mathbf{y}}^c$.

Propriedades do Coeficiente de Determinação

A abordagem geométrica confirma que, também na Regressão Linear Múltipla, são válidas as propriedades (já conhecidas da Regressão Linear Simples) do Coeficiente de Determinação:

- R^2 toma valores entre 0 e 1.
- Quanto mais próximo de 1 estiver R^2, menor o ângulo θ, e portanto melhor será a correspondência entre o vector (centrado) das observações, $\vec{\mathbf{y}}^c$, e o seu ajustamento em $\mathscr{C}(\mathbf{X})$.
- Se $R^2 \approx 0$, o vector $\vec{\mathbf{y}}^c$ é quase perpendicular ao subespaço $\mathscr{C}(\mathbf{X})$ onde se pretende aproximá-lo, e a projecção vai quase anular todas os elementos do vector projectado. O resultado será de má qualidade, uma vez que se perde quase toda a variabilidade nos valores de y.

Propriedades de modelos com constante aditiva

$\mathscr{C}(\mathbf{X})$ contém o vector $\vec{\mathbf{1}}_n$ de n uns (sucede sempre que haja constante aditiva b_0 no modelo). Então $\mathbf{H}\vec{\mathbf{1}}_n = \vec{\mathbf{1}}_n$, pois a projecção de qualquer vector num subespaço que já o contém deixa o vector invariante.

Logo, (ver também o Exercício 13),

- As médias dos valores observados e ajustados de Y são iguais.
- A soma dos resíduos é zero.
- O hiperplano em $\mathbb{R}^{p+1}$ ajustado pela regressão contém o centro de gravidade da nuvem dos n pontos observados, isto é, $\overline{y} = \overline{\mathbf{x}}^t \vec{\mathbf{b}}$, onde $\overline{\mathbf{x}}^t = (1, \overline{x_1}, \overline{x_2}, ..., \overline{x_p}) = \frac{1}{n}\vec{\mathbf{1}}_n^t \mathbf{X}$, ou seja,

$$\overline{y} = b_0 + b_1\,\overline{x_1} + b_2\,\overline{x_2} + ... + b_p\,\overline{x_p}\ .$$

Algumas propriedades dos hiperplanos ajustados

Numa regressão linear múltipla verifica-se:

- os coeficientes $\{b_j\}_{j=1}^{p}$ que multiplicam variáveis preditoras interpretam-se como a variação média em Y, associada a aumentar a variável preditora correspondente em uma unidade, mantendo os restantes preditores constantes.

- o valor do coeficiente de determinação R^2 numa regressão múltipla não pode ser inferior ao valor de R^2 que se obteria excluindo do modelo um qualquer subconjunto de preditores. Em particular, não pode ser inferior ao R^2 das regressões lineares simples de Y sobre cada preditor individual.

Unidades de medida

O vector dos parâmetros ajustados pelo método dos mínimos quadrados, $\vec{\mathbf{b}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{y}}$, gera n valores ajustados:

$$\begin{aligned} \hat{y}_i &= b_0 + b_1 x_{1\,(i)} + \ldots + b_p x_{p\,(i)} \quad , \quad \forall\, i \\ \Leftrightarrow \quad \vec{\hat{\mathbf{y}}} &= \mathbf{X}\vec{\mathbf{b}} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{y}} = \mathbf{H}\vec{\mathbf{y}}. \end{aligned}$$

As unidades de medida de $\hat{y}$ e de y são iguais. As unidades de medida de b_0 são iguais às de y. As unidades dos parâmetros b_j que multiplicam variáveis ($j \neq 0$) são a razão entre as unidades da resposta y e as unidades do preditor x_j correspondente.

O vector de resíduos, $\vec{\mathbf{e}}$, também se obtém pré-multiplicando o vector $\vec{\mathbf{y}}$ por $\mathbf{I} - \mathbf{H}$, onde $\mathbf{I}$ é a matriz identidade $n \times n$:

$$\begin{aligned} e_i &= y_i - \hat{y}_i = y_i - (b_0 + b_1 x_{1\,(i)} + \ldots + b_p x_{p\,(i)}) \quad , \quad \forall\, i \\ \Leftrightarrow \quad \vec{\mathbf{e}} &= \vec{\mathbf{y}} - \vec{\hat{\mathbf{y}}} = \vec{\mathbf{y}} - \mathbf{H}\vec{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\vec{\mathbf{y}}\,, \end{aligned}$$

As unidades de medida dos resíduos $e = y - \hat{y}$ são iguais às de y.

A Regressão Múltipla no R

O comando `lm` também ajusta uma Regressão Múltipla no R.
A variável resposta y e as variáveis preditoras $x_1, ..., x_p$ definem-se mediante uma fórmula semelhante à da RLS.

E.g., sendo `y` a variável resposta e `x1`, `x2` e `x3` três preditores, a fórmula que especifica a relação será:

$$y \sim x1 + x2 + x3$$

Comando `R` para ajustar uma regressão linear múltipla

```
> lm ( y ~ x1 + x2 + x3 + ...  + xp, data=dados)
```

O comando devolve o vector $\vec{\mathbf{b}}$ das estimativas dos $p+1$ parâmetros do modelo, b_0, b_1, ..., b_p.

Um exemplo de RLM no R

Vamos ilustrar uma Regressão Linear Múltipla no R, com um conjunto de dados famoso: os Lírios de Anderson/Fisher, disponíveis na *data frame* `iris`.

```
> help(iris)

iris                    package:datasets                    R Documentation

Edgar Anderson's Iris Data

Description:

This famous (Fisher's or Anderson's) iris data set gives the
measurements in centimeters of the variables sepal length and
width and petal length and width, respectively, for 50 flowers
from each of 3 species of iris.  The species are Iris setosa,
versicolor, and virginica.
```

Exemplo RLM (cont.)

Figura: iris setosa

Figura: iris versicolor

Figura: iris virginica

Exemplo RLM (cont.)

Uma inspecção inicial dos dados pode ser feita com o comando `head`, que mostra a parte inicial do argumento:

```
> head(iris)

  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

Uma síntese da informação é dado pelo comando `summary`:

```
> summary(iris)

 Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa     :50
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

Note-se que a quinta coluna é um factor.

A Regressão Múltipla no R (cont.)

Ajuste-se um modelo para prever a variável resposta largura da pétala, a partir do comprimento da pétala e das duas medições das sépalas (largura e comprimento).

```
> iris2.lm <- lm(Petal.Width ~ Petal.Length + Sepal.Length +
+                                Sepal.Width , data=iris)
> iris2.lm
(...)
Coefficients:
 (Intercept)  Petal.Length  Sepal.Length   Sepal.Width
     -0.2403        0.5241       -0.2073        0.2228
```

O hiperplano ajustado é:

$$PW = -0.2403 + 0.5241\,PL - 0.2073\,SL + 0.2228\,SW$$

Confirmando as fórmulas (cont.)

Confirme-se a fórmula dos parâmetros ajustados pelo método dos mínimos quadrados. O comando `model.matrix` devolve a matriz **X**.

```
> X <- model.matrix(iris2.lm)
> X
```

	(Intercept)	Petal.Length	Sepal.Length	Sepal.Width
1	1	1.4	5.1	3.5
2	1	1.4	4.9	3.0
3	1	1.3	4.7	3.2
4	1	1.5	4.6	3.1
5	1	1.4	5.0	3.6
6	1	1.7	5.4	3.9
7	1	1.4	4.6	3.4
8	1	1.5	5.0	3.4
[...]				
149	1	5.4	6.2	3.4
150	1	5.1	5.9	3.0

Confirmando as fórmulas (cont.)

Os comandos do R para as operações matriciais necessárias para o cálculo de $\vec{\mathbf{b}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{y}}$ são:

- `t(A)` indica a transposta da matriz A
- `A %*% B` indica o produto das matrizes A e B.
- `solve(A)` calcula a inversa da matriz A.

```
> y <- iris$Petal.Width
> b <- solve( t(X) %*% X ) %*% ( t(X) %*% y )
> b

                    [,1]
(Intercept)   -0.2403074
Petal.Length   0.5240831
Sepal.Length  -0.2072661
Sepal.Width    0.2228285
```

Confirmam-se os valores do acetato 91.

Modelos e submodelos

Dado um modelo de regressão linear múltipla, com equação $y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_p x_p$, chama-se submodelo a uma regressão linear com apenas alguns preditores. Por exemplo, a regressão linear simples

$$Petal.Width = b_0 + b_1 Petal.Length$$

é um submodelo da regressão linear múltipla acabada de ajustar,

$$Petal.Width = b_0 + b_1 Petal.Length + b_2 Sepal.Length + b_3 Sepal.Width$$

O R^2 dum submodelo não pode exceder o R^2 do modelo completo.

No modelo do acetato anterior $R^2 = 0.9379$. Na RLS com *Petal.Length* $R^2 = 0.9271$.

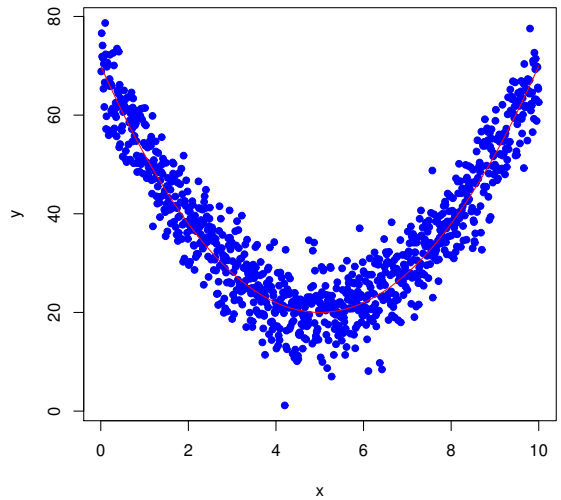
A equação ajustada num submodelo não é a parte correspondente na equação ajustada do modelo.

Vimos que $PW = -0.2403 + 0.5241\,PL - 0.2073\,SL + 0.2228\,SW$. Mas na RLS só com o preditor *PL*, a equação ajustada é $PW = -0.3631 + 0.4158\,PL$.

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.

Imagine-se uma relação de fundo entre uma variável resposta Y e uma única variável preditora X dada por uma parábola:

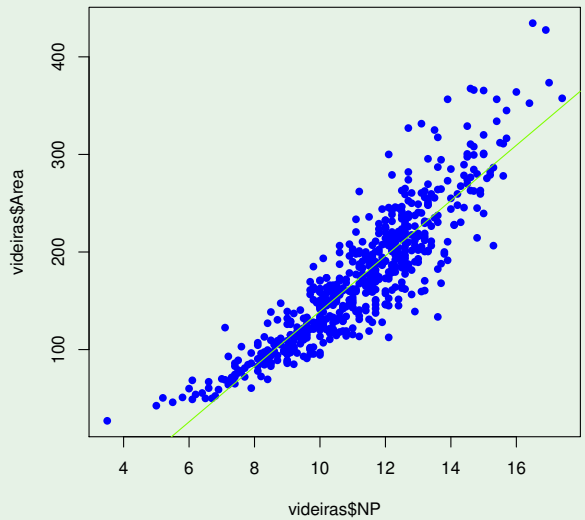


Regressão Polinomial - Exemplo

Exemplo 5 – Folhas de videira

Considere os dados de medições sobre $n = 600$ folhas de videira.

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.)

Pode ajustar-se uma qualquer parábola, com equação

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 ,$$

com uma regressão linear de Y sobre os dois preditores $X_1 = X$ e $X_2 = X^2$:

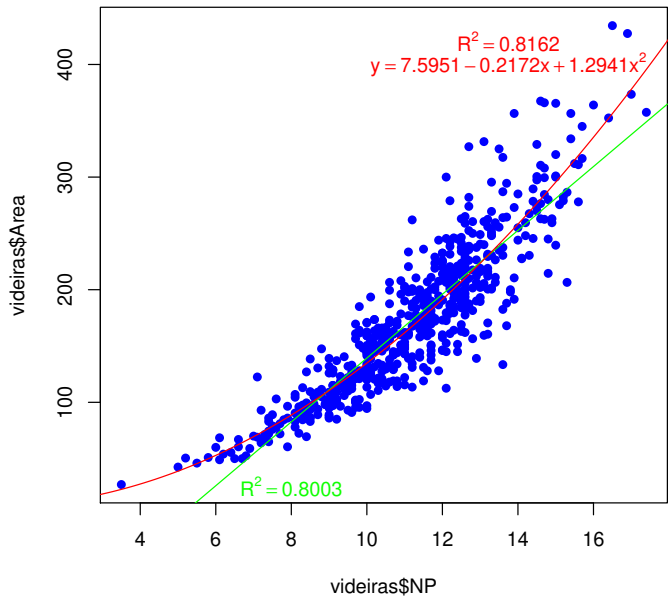
```
> lm( Area ~ NP + I(NP^2) , data=videiras )

 Call:
lm(formula = Area ~ NP + I(NP^2), data = videiras)
 Coefficients:
(Intercept)            NP       I(NP^2)
     7.5961       -0.2172        1.2941
```

```
> summary(lm( Area ~ NP + I(NP^2) , data=videiras ))$r.sq

[1] 0.8161632
```

A parábola ajustada tem equação $y = 7.5961 - 0.2172x + 1.2941x^2$. O valor $R^2 = 0.816$ indica que cerca de 82% da variabilidade observada nas áreas foliares é explicada pela regressão quadrática - aqui não houve transformação de y.

Regressão Polinomial - Exemplo (cont.)

Eis a parábola ajustada:



A equação da recta ajustada é $y = -144.15 + 28.34\,x$, o que confirma que a equação ajustada dum submodelo (neste caso, a recta de regressão) não é apenas a parte relevante da equação ajustada dum modelo (neste caso, o modelo parábolico).

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

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

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.

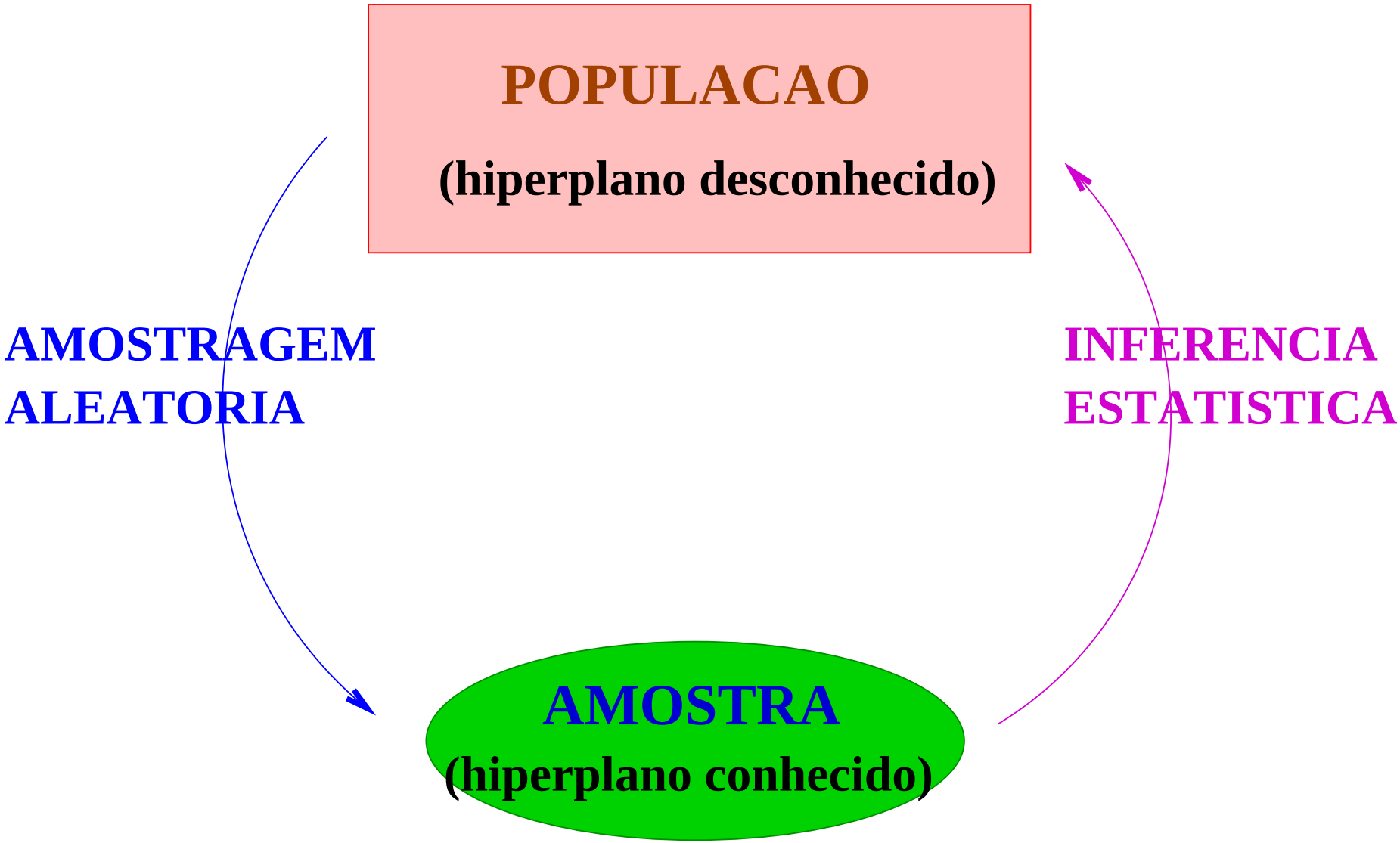
- Um hiperplano ajustado, $y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$, é apenas uma estimativa de um "hiperplano populacional"

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

  Outras amostras dariam hiperplanos ajustados diferentes.

- Coloca-se o problema da inferência estatística.

O problema da Inferência Estatística na RL



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 de x_1, x_2, ..., x_p e Y, sobre n unidades experimentais. Admite-se que as n observações de Y são independentes.

MODELO RL – Linearidade

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

$$\begin{array}{ccccccccccc} Y_i & = & \beta_0 & + & \beta_1 & x_{1(i)} & + \ldots + & \beta_1 & x_{p(i)} & + & \varepsilon_i \\ \downarrow & & \downarrow & & \downarrow & \downarrow & & \downarrow & \downarrow & & \downarrow \\ \text{v.a.} & & \text{cte.} & & \text{cte.} & \text{cte.} & & \text{cte.} & \text{cte.} & & \text{v.a.} \end{array}$$

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

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

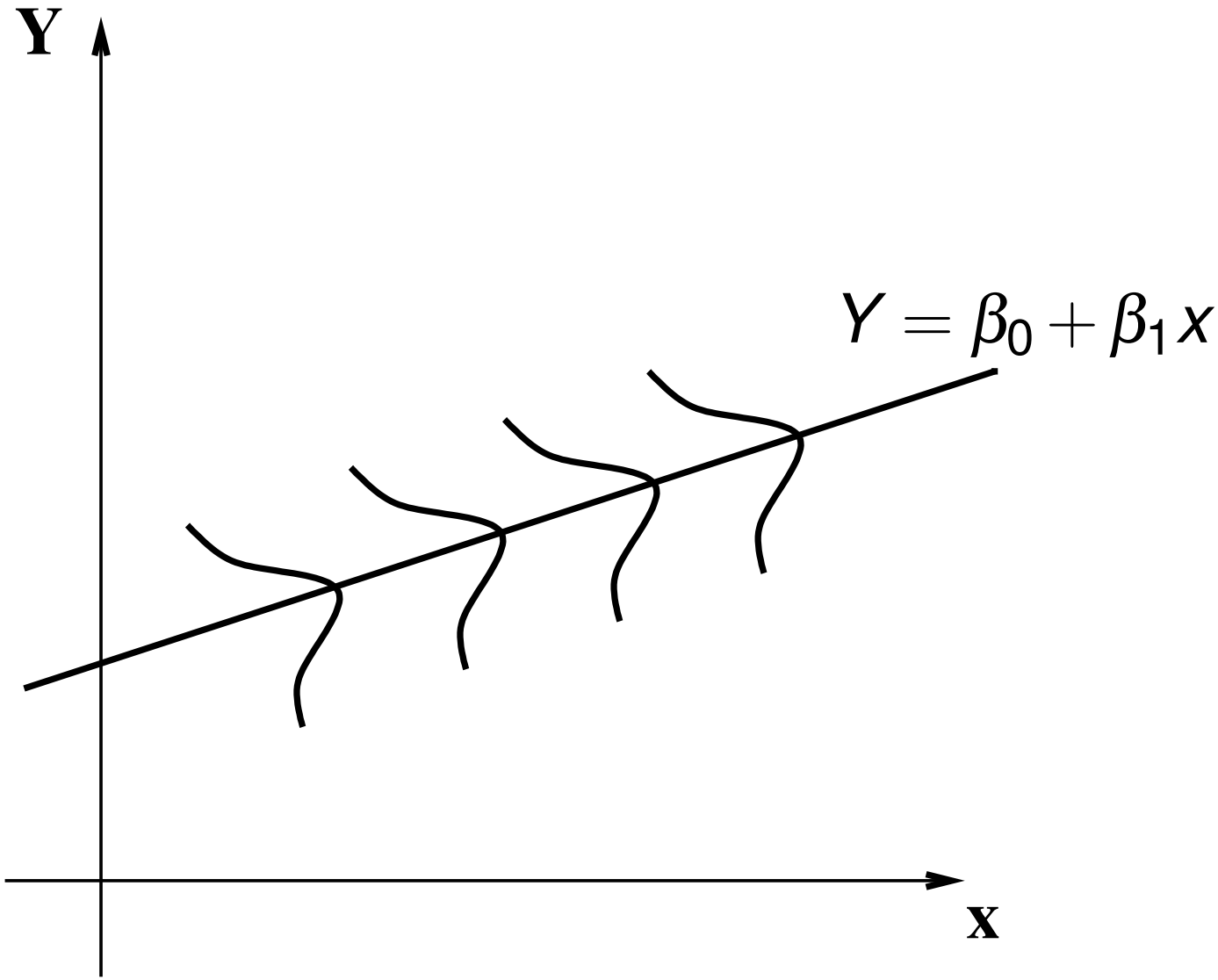
NOTA: Pelo modelo, o valor esperado (médio) de Y_i, condicional aos valores $x_1, x_2, ..., x_p$ dos preditores, é:

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

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.

MODELO Regressão Linear Simples

Ilustrando, no caso duma regressão linear simples:



O estudo do modelo

Para cada caso de estudo haverá que analisar os parâmetros do modelo: as $p+1$ constantes β_j $(j = 0, 1, ..., p)$.

Admite-se que se dispõe de uma amostra de n observações $\{(x_{1(i)}, x_{2(i)}, ..., x_{p(i)}, y_i\}_{i=1}^{n}$, com base nas quais se ajusta o modelo. Os parâmetros ajustados $\vec{\mathbf{b}} = (b_0, b_1, b_2, ..., b_p)$, obtidos pela fórmula do acetato 74, são estimativas desses parâmetros.

Objectivo: obter estimadores $\hat{\beta}_j$ dos parâmetros populacionais, com distribuição de probabilidades conhecida, de forma a poder-se construir intervalos de confiança e/ou efectuar testes de hipóteses sobre os valores dos parâmetros populacionais β_j.

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

A notação matricial/vectorial

O estudo do modelo (nomeadamente com mais de um preditor) exigirá ferramentas para o estudo de vectores aleatórios.

As equações do modelo para as n observações (acetato 105) podem ser escritas como uma única equação, utilizando notação vectorial/matricial:

$$\begin{array}{lcl} Y_1 & = & \beta_0 + \beta_1 x_{1(1)} + \beta_2 x_{2(1)} + \cdots + \beta_p x_{p(1)} + \varepsilon_1 \\ Y_2 & = & \beta_0 + \beta_1 x_{1(2)} + \beta_2 x_{2(2)} + \cdots + \beta_p x_{p(2)} + \varepsilon_2 \\ Y_3 & = & \beta_0 + \beta_1 x_{1(3)} + \beta_2 x_{2(3)} + \cdots + \beta_p x_{p(3)} + \varepsilon_3 \\ \vdots & \vdots & \vdots \\ Y_n & = & \beta_0 + \beta_1 x_{1(n)} + \beta_2 x_{2(n)} + \cdots + \beta_p x_{p(n)} + \varepsilon_n \end{array}$$

A notação vectorial (cont.)

As n equações correspondem a uma única equação vectorial:

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

onde

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

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 vectorial (cont.)

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

- $\vec{\mathbf{Y}}$ é o vector aleatório das n variáveis aleatórias resposta;
- $\mathbf{X}$ é a 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);
- $\vec{\boldsymbol{\beta}}$ é o vector (não aleatório) de $p+1$ parâmetros do modelo;
- $\vec{\boldsymbol{\varepsilon}}$ é o vector aleatório dos n erros aleatórios.

Representa-se um vector de n observações de Y (correspondentes a uma amostra concreta) por $\vec{\mathbf{y}}$.

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

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 (acetato 74), mas substituindo o vector $\vec{\mathbf{y}}$ de valores observados de Y pelo vector aleatório $\vec{\mathbf{Y}}$.

Estimadores de MQ dos parâmetros numa RL

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

Os estimadores assim obtidos são estimadores de mínimos quadrados.

Veremos que, dado o Modelo Linear, são também estimadores de máxima verosimilhança.

Distribuição de Y_i no Modelo Linear

Distribuição de Y_i

Dado o Modelo Linear, tem-se, para qualquer $i = 1, 2, ..., n$,

$$Y_i \frown \mathcal{N}(\mu_i , \sigma^2) ,$$

com $\mu_i = E[Y_i \,|\, x_{1(i)}, x_{2(i)}, ..., x_{p(i)}] = \beta_0 + \beta_1 x_{1(i)} + \beta_2 x_{2(i)} + ... + \beta_p x_{p(i)}$.

Assim, o valor da função densidade para cada observação y_i é:

$$f(y_i \mid \beta_0, \beta_1, ..., \beta_p) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(y_i-\mu_i)^2}{2\sigma^2}} .$$

Como as observações de Y são independentes, a densidade conjunta das n observações é:

$$f(y_1, y_2, ..., y_n \mid \beta_0, \beta_1, ..., \beta_p) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(y_i-\mu_i)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-\frac{n}{2}} \mathrm{e}^{-\sum_{i=1}^{n} \frac{(y_i-\mu_i)^2}{2\sigma^2}} .$$

A estimação por Máxima Verosimilhança

A função verosimilhança

No Modelo Linear, a função verosimilhança de n observações Y_i é:

$$\mathscr{L}(\beta_0, \beta_1, ..., \beta_p \mid y_1, y_2, ..., y_n) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\sum_{i=1}^{n} \frac{(y_i-\mu_i)^2}{2\sigma^2}},$$

com $\mu_i = \beta_0 + \beta_1 x_{1(i)} + \beta_2 x_{2(i)} + ... + \beta_p x_{p(i)}$.

Os estimadores de máxima verosimilhança dos parâmetros β_0, β_1, ..., β_p são os valores que maximizam esta função verosimilhança, ou seja, são os $\hat{\beta}_j$ que minimizam $\sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 = SQRE$, (sendo $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1(i)} + \hat{\beta}_2 x_{2(i)} + ... + \hat{\beta}_p x_{p(i)}$).

No Modelo Linear, os estimadores dos β_j de Mínimos Quadrados (acetato 111) são também estimadores de Máxima Verosimilhança.

Ferramentas para vectores aleatórios

Já se introduziram 3 vectores aleatórios:

- $\vec{\mathbf{Y}}$ (das n observações da variável resposta);
- $\vec{\boldsymbol{\varepsilon}}$ (dos n erros aleatórios); e
- $\vec{\hat{\boldsymbol{\beta}}}$ (dos $p+1$ estimadores $\hat{\beta}_j$).

São necessárias ferramentas para trabalhar com vectores aleatórios.

Para qualquer vector aleatório $\vec{\mathbf{Z}} = (Z_1, Z_2, ..., Z_k)^t$, define-se:

- O vector esperado de $\vec{\mathbf{Z}}$, constituído pelos valores esperados de cada componente:

$$\vec{\boldsymbol{\mu}}_Z = E[\vec{\mathbf{Z}}] = \begin{bmatrix} E[Z_1] \\ E[Z_2] \\ \vdots \\ E[Z_k] \end{bmatrix}.$$

Se $\mathbf{W}$ fôr uma matriz aleatória, também se define $E[\mathbf{W}]$ como a matriz do valor esperado de cada elemento.

Ferramentas para vectores aleatórios (cont.)

- a matriz de variâncias-covariâncias de $\vec{\mathbf{Z}}$ é constituída pelas (co)variâncias de cada par de componentes:

$$V[\vec{\mathbf{Z}}] = \begin{bmatrix} V[Z_1] & C[Z_1,Z_2] & C[Z_1,Z_3] & \dots & C[Z_1,Z_k] \\ C[Z_2,Z_1] & V[Z_2] & C[Z_2,Z_3] & \dots & C[Z_2,Z_k] \\ C[Z_3,Z_1] & C[Z_3,Z_2] & V[Z_3] & \dots & C[Z_3,Z_k] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C[Z_k,Z_1] & C[Z_k,Z_2] & C[Z_k,Z_3] & \dots & V[Z_k] \end{bmatrix}$$

Propriedades do vector esperado

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

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

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

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

Propriedades da matriz de (co)variâncias

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

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

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

A distribuição Normal Multivariada

Vectores aleatórios têm distribuições multivariadas de probabilidades. A mais frequente distribuição multivariada é a Multinormal:

Distribuição Normal Multivariada

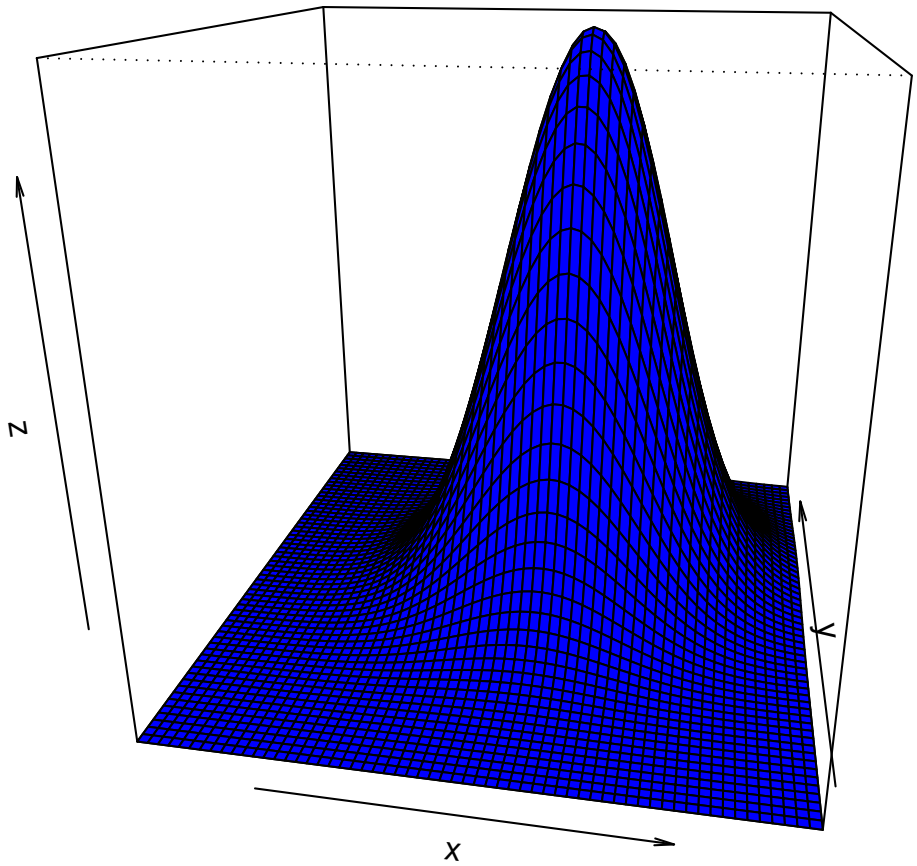
O vector aleatório k-dimensional $\vec{\mathbf{Z}}$ tem distribuição Multinormal, com parâmetros dados pelo vector $\vec{\boldsymbol{\mu}}$ e a matriz invertível $\boldsymbol{\Sigma}$ se a sua função densidade conjunta fôr:

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

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

Nota: Define-se uma Multinormal em sentido generalizado, quando $\boldsymbol{\Sigma}$ é apenas semi-definida positiva, usando a inversa generalizada $\boldsymbol{\Sigma}^{-}$.

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



Algumas propriedades da distribuição Multinormal

Teorema (Propriedades da Multinormal)

Se $\vec{\mathbf{Z}} \frown \mathcal{N}_k(\vec{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$*:*

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

Nota: Nas disciplinas introdutórias de Estatística mostra-se que 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 elemento nulo numa matriz de (co)variâncias duma Multinormal indica que as componentes correspondentes são independentes.

Propriedades da Multinormal (cont.)

Teorema (Propriedades da Multinormal)

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

4. *Todas as distribuições marginais de* $\vec{\mathbf{Z}}$ *são (multi)normais. Em particular, cada componente* Z_i *é normal com média* μ_i *e variância* $\boldsymbol{\Sigma}_{(i,i)}$*:* $Z_i \frown \mathscr{N}(\mu_i, \boldsymbol{\Sigma}_{(i,i)})$.
5. *Se* $\vec{\mathbf{a}}$ *um vector (não-aleatório)* $k \times 1$*, então* $\vec{\mathbf{Z}} + \vec{\mathbf{a}} \frown \mathscr{N}_k(\vec{\boldsymbol{\mu}} + \vec{\mathbf{a}}, \boldsymbol{\Sigma})$.
6. *Combinações lineares das componentes dum vector multinormal são Normais:* $\vec{\mathbf{a}}^t \vec{\mathbf{Z}} = a_1 Z_1 + a_2 Z_2 + \ldots + a_k Z_k \frown \mathscr{N}(\vec{\mathbf{a}}^t \vec{\boldsymbol{\mu}}, \vec{\mathbf{a}}^t \boldsymbol{\Sigma} \vec{\mathbf{a}})$.
7. *Se* $\mathbf{B}$ *é matriz* $m \times k$ *(não aleatória, de característica* $m \leq k$*), então* $\mathbf{B}\vec{\mathbf{Z}} \frown \mathscr{N}_m(\mathbf{B}\vec{\boldsymbol{\mu}}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^t)$.

Nota: No último resultado, se $\mathbf{B}$ é matriz não aleatória de característica $m > k$, a distribuição de $\mathbf{B}\vec{\mathbf{Z}}$ é Multinormal em sentido generalizado.

Modelo Regressão Linear - versão vectorial

O Modelo Linear em notação vectorial

1. $\vec{\mathbf{Y}} = \mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\varepsilon}}$.
2. $\vec{\boldsymbol{\varepsilon}} \frown \mathcal{N}_n(\vec{\mathbf{0}}, \sigma^2 \mathbf{I}_n)$.

Na segunda destas hipóteses são feitas quatro afirmações (tendo em conta as propriedades da Multinormal, referidas atrás):

- 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 $\vec{\mathbf{Y}}$

O seguinte Teorema é consequência directa dos acetatos 120 e 121.

Teorema (Primeiras Consequências do Modelo)

Dado o Modelo de Regressão Linear, tem-se:

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

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 estimador dos parâmetros do Modelo

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

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

onde $\mathbf{X}$ e $\vec{\mathbf{Y}}$ são a matriz e o vector definidos no acetato 109.

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 $\vec{\hat{\boldsymbol{\beta}}}$

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

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

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

Tendo em conta as propriedades da Multinormal (acetatos 120 e 121):

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

A distribuição dum estimador individual

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

$$\begin{aligned} \hat{\beta}_j &\frown \mathcal{N}\left(\beta_j\ ,\ \sigma^2 \left(\mathbf{X}^t\mathbf{X}\right)^{-1}_{(j+1,j+1)}\right) \\ \Leftrightarrow \qquad \frac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}} &\frown \mathcal{N}(0,1)\ , \end{aligned}$$

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

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.

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 mais quantidades não-amostrais.

Para ultrapassar este problema vai-se:

- obter um estimador para σ^2; e
- ver o que acontece à distribuição do acetato anterior 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)

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

Resíduos (observados)

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

O estimador de máxima verosimilhança de σ^2, a variância dos erros aleatórios, é dado por

$$\hat{\sigma}^2_{MV} = \frac{SQRE}{n} .$$

Mas este estimador tem um problema: não é centrado, já que $E\left[\frac{SQRE}{n}\right] = \frac{n-(p+1)}{n}\sigma^2$.

O Quadrado Médio Residual

Uma simples modificação do estimador de máxima verosimilhança gera um estimador centrado.

Quadrado Médio Residual

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, σ^2:

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

Quantidades fulcrais para a inferência sobre β_j

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

Teorema (Distribuições para a inferência sobre β_j)

Dado o Modelo de Regressão Linear Múltipla, tem-se

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

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.

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:

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

com o valor amostral de $QMRE$; $\hat{\sigma}_{\hat{\beta}_j} = \sqrt{QMRE \cdot (\mathbf{X}^t\mathbf{X})^{-1}_{(j+1,j+1)}}$; e sendo $t_{\alpha/2[n-(p+1)]}$ o valor que na distribuição $t_{n-(p+1)}$ deixa *à direita* uma região de probabilidade $\alpha/2$. O valor b_j é o elemento $j+1$ do vector das estimativas $\vec{\mathbf{b}}$ (acetato 73).

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

Intervalos de confiança para β_j no R

A informação básica para a construção de intervalos de confiança para cada parâmetro β_j obtém-se, no R, a partir das tabelas produzidas pela função `summary`. No exemplo do acetato 91:

```
> 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 ***
```

Assim, estima-se que em média a largura da pétala diminui $0.20727cm$ por cada aumento de $1cm$ 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

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

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

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
```

Testes de Hipóteses sobre os parâmetros

O mesmo resultado (acetato 130) usado para construir intervalos de confiança serve para construir testes a hipóteses para cada β_j individual. 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: $T = \dfrac{\hat{\beta}_j - \overbrace{\beta_j|_{H_0}}^{=c}}{\hat{\sigma}_{\hat{\beta}_j}} \frown t_{n-(p+1)}$

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)

Combinações lineares dos parâmetros

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

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

Casos particulares importantes nas aplicações são:

- Se $\vec{\mathbf{a}}$ tem um único elemento não-nulo, na posição $j+1$, $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} = \beta_j$.
- 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 uma qualquer observação 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:

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

Inferência sobre combinações lineares dos β_js

A multinormalidade do vector de estimadores $\vec{\hat{\boldsymbol{\beta}}}$ implica a normalidade de qualquer vector que seja combinação linear das suas componentes (acetato 121). Mais concretamente,

- 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 125);
- 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 121);
- Ou seja, $\vec{\mathbf{Z}} = \frac{\vec{\mathbf{a}}^t\vec{\hat{\boldsymbol{\beta}}} - \vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}}{\sqrt{\sigma^2 \vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}}} \frown \mathcal{N}(0,1)$;
- Por um raciocínio análogo ao usado aquando dos βs individuais, tem-se então

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

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

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

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

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, 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:

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

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

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

Para quaisquer valores das variáveis preditoras, tem-se:

IC para $\mu_{Y|\vec{\mathbf{x}}} = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$

Dado o Modelo RLS e dados os valores $\vec{\mathbf{x}} = (x_1, x_2, ..., x_p)^t$ das variáveis preditoras, um intervalo a $(1-\alpha)\times 100\%$ de confiança para o valor esperado de Y,

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

é dado por:

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

com $\hat{\mu}_{Y|\vec{\mathbf{x}}} = b_0 + b_1 x_1 + ... + b_p x_p$ e $\hat{\sigma}_{\hat{\mu}_{Y|\vec{\mathbf{x}}}} = \sqrt{QMRE \cdot \vec{\mathbf{a}}^t (\mathbf{X}^t \mathbf{X})^{-1} \vec{\mathbf{a}}}$, onde $\vec{\mathbf{a}} = (1, x_1, x_2, ..., x_p)$.

Testes de Hipóteses sobre os parâmetros

Dado o Modelo de Regressão Linear Múltipla,

Testes de Hipóteses 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: $T = \dfrac{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}} - \overbrace{\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}|_{H_0}}^{=c}}{\hat{\sigma}_{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}}} \frown t_{n-(p+1)}$.

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)

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

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
```

Numa regressão linear simples, a variabilidade do estimador $\hat{\mu}_{Y|\vec{\mathbf{x}}}$ de $\mu_{Y|\vec{\mathbf{x}}} = E[Y|X = x]$ tem uma fórmula específica:

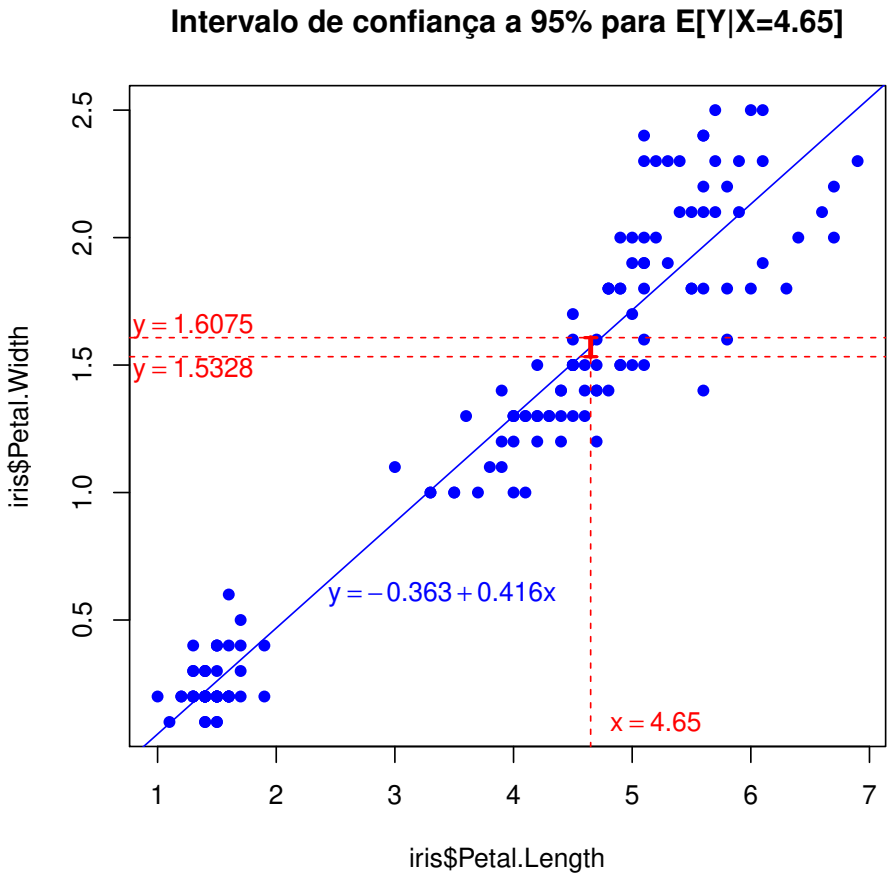
$$\sigma^2_{\hat{\mu}_{Y|\vec{\mathbf{x}}}} = V[\hat{\mu}_{Y|x}] = \sigma^2 \cdot \left[\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)\cdot s_x^2}\right] .$$

Substituindo σ^2 por *QMRE* tem-se um estimador da variância, $\hat{\sigma}^2_{\hat{\mu}_{Y|\vec{\mathbf{x}}}}$.

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

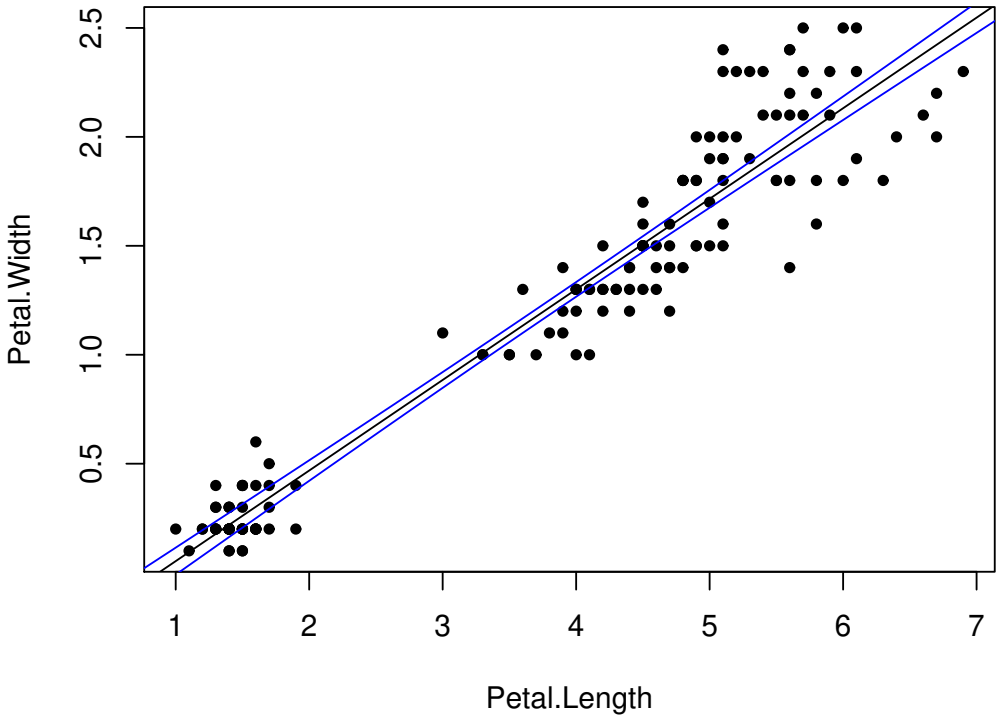
O intervalo de confiança obtém-se através do argumento `int="conf"`:

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



Bandas de confiança para a recta de regressão

Considerando os ICs para todos os valores de x nalgum intervalo, obtém-se uma banda de confiança para a recta de regressão.



Os IC para $\mu_{Y|x}$ dependem do valor de x, como se verifica na fórmula do acetato 141: terão maior amplitude quanto mais afastado x estiver da média $\overline{x}$ das observações. Logo, as bandas são encurvadas.

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

Numa regressão linear múltipla, também se obtém um intervalo de confiança (referido no acetato 138) para o valor esperado de Y , dado um conjunto de valores $X_1 = x_1, \ldots, X_p = x_p$ das variáveis preditoras, através do comando `predict`.

No exemplo da regressão linear múltipla dos lírios, um IC a 95% para a largura esperada de pétalas de flores com `Petal.Length=2`, `Sepal.Length=5` e `Sepal.Width=3.1` é pedido assim:

```
> 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 $E[Y|X_1 = 2, X_2 = 5, X_3 = 3.1]$ é: $]\, 0.4169\ ,\ 0.5077\, [$.

ICs para combinações lineares no R

O intervalo de confiança para uma combinação linear genérica $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$, numa RLM, necessita da matriz das (co)variâncias estimadas dos estimadores $\hat{\vec{\boldsymbol{\beta}}}$, $\hat{V}[\hat{\vec{\boldsymbol{\beta}}}] = QMRE \cdot (\mathbf{X}^t\mathbf{X})^{-1}$, dada pela função `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
```

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

$$\begin{aligned} \hat{\sigma}_{\hat{\beta}_2+\hat{\beta}_3} &= \sqrt{\hat{V}[\hat{\beta}_2+\hat{\beta}_3]} = \sqrt{\hat{V}[\hat{\beta}_2]+\hat{V}[\hat{\beta}_3]+2\hat{Cov}[\hat{\beta}_2,\hat{\beta}_3]} \\ \hat{\sigma}_{\hat{\beta}_2+\hat{\beta}_3} &= \sqrt{0.002256837+0.002394932+2(-0.001344002)} = 0.04431439\,. \end{aligned}$$

A variabilidade duma observação individual de Y

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

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

Esses intervalos de confiança para $\mu_{Y|\vec{\mathbf{x}}}$, basearam-se na variabilidade associada ao estimador $\hat{\mu}_{Y|\vec{\mathbf{x}}}$:

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

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

Mas uma observação individual de Y tem associada uma variabilidade adicional, pois:

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

A flutuação aleatória de observações individuais em torno do hiperplano é $V[\varepsilon] = \sigma^2$. Ao tentar prever-se valores de observações individuais 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 $X_1 = x_1, ..., X_p = x_p$ 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

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

onde

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

e

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

RLS: a variabilidade duma observação individual de Y

Numa regressão linear simples existe uma fórmula específica:

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

Logo,

RLS: Intervalo de predição para observação individual de Y

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

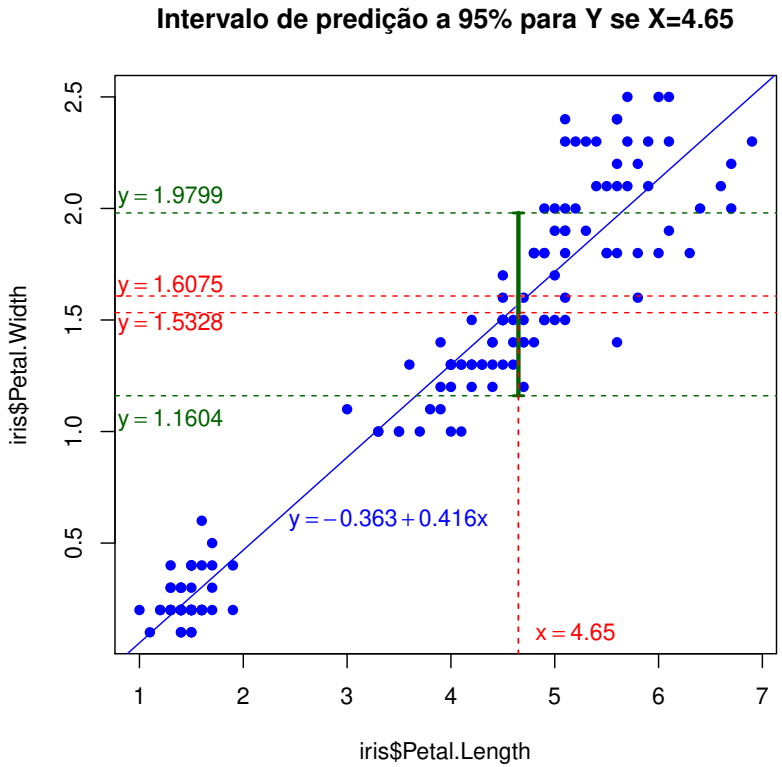
com $\hat{\mu}_{Y|x} = b_0 + b_1 x$ e $\hat{\sigma}_{Indiv} = \sqrt{QMRE \cdot \left[1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)\cdot s_x^2}\right]}$.

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

Intervalos de predição para Y no 

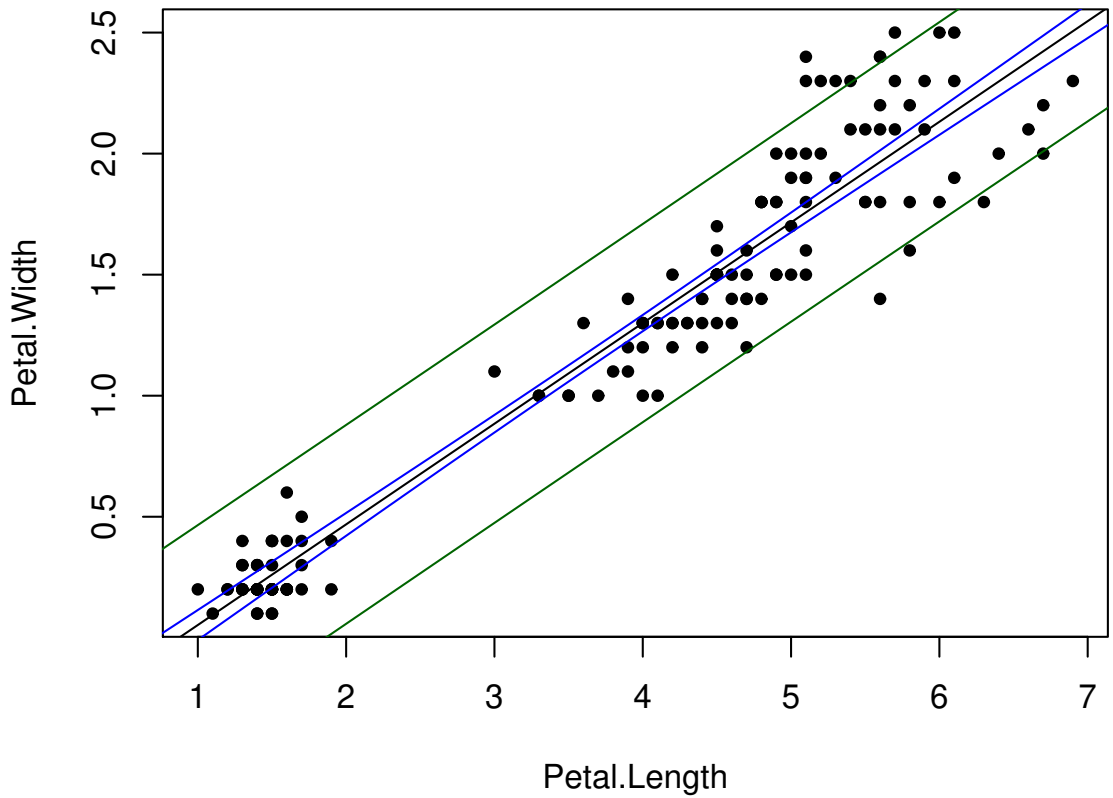
No R, um intervalo de predição para uma observação individual de Y obtém-se através da opção int="pred" no comando predict:

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



Bandas de predição para uma observação de Y

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



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.

Avaliando 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.

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

As hipóteses em confronto são:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$
$$[\text{MODELO COMPLETO} \equiv \text{MODELO NULO}]$$
vs.
$$H_1 : \exists j = 1, ..., p \quad \text{t.q.} \quad \beta_j \neq 0$$
$$[\text{MODELO COMPLETO} \not\equiv \text{MODELO NULO}]$$

NOTA: repare que β_0 não intervém nas hipóteses.

O Teste F de ajustamento global do Modelo

Sendo válido o Modelo RLM, define-se o Quadrado Médio associado à Regressão com sendo $QMR = \frac{SQR}{p}$. Pode efectuar-se o seguinte

Teste F de ajustamento global do modelo RLM

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

vs.

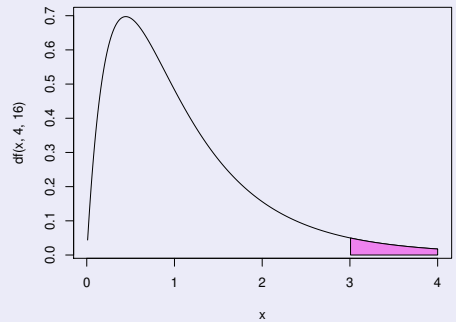
$H_1 : \exists\, j = 1, ..., p \quad \text{tal que} \quad \beta_j \neq 0.$

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



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:

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

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.

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: $F = \frac{n-(p+1)}{p} \cdot \frac{R^2}{1-R^2} \underset{H_0}{\sim} F_{(p,n-(p+1))}$

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

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:

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

[SUBMODELO OK] [SUBMODELO PIOR]

NOTA: Esta discussão só envolve coeficientes β_j de variáveis preditoras. O coeficiente β_0 faz sempre parte dos submodelos. 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.

Uma estatística de teste para a comparação 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)

Vamos admitir que o submodelo tem k preditores ($k+1$ parâmetros):

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

caso $\beta_j = 0$, para todas as variáveis x_j que não pertençam ao submodelo.

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:

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

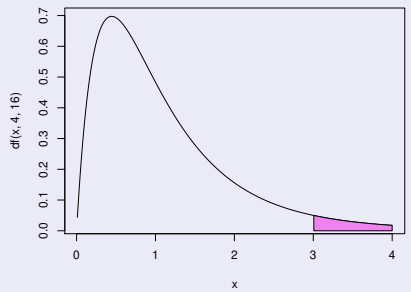
Estatística do Teste:

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

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:

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

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:

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

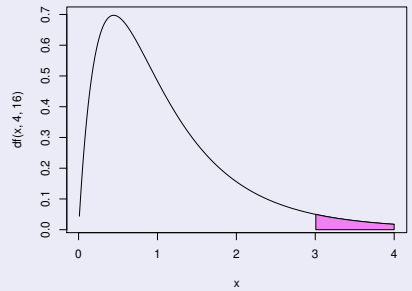
Estatística do Teste:

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

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

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



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:

```
> iris.lm <- lm(Petal.Width ~ Petal.Length , data=iris)
> 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 $R^2 = 0.9271$ e $R^2 = 0.9379$ 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 com o seu submodelo nulo (sem preditores).

Caso o modelo e submodelo difiram num único preditor, X_j, o teste F parcial é equivalente ao teste t (acetato 134) 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.

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, à partida, 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 (conjunto vazio) e p (o modelo completo) que é possível escolher é dado por $2^p - 2$. A tabela seguinte indica o número desses subconjuntos para $p = 5, 10, 15, 20$.

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

Cuidado com exclusões simultâneas de preditores

Para valores de p pequenos, é possível analisar todos os possíveis subconjuntos. Com o apoio de 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: dados Brix

Nos dados relativos ao Exercício 10 de Regressão Linear, a tabela associada à 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 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 pequeno ou médio, existem algoritmos e rotinas informáticas que efectuam uma pesquisa completa e determinam 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 identifica o melhor subconjunto de preditores, para 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`

Apesar do pequeno número de preditores, exemplifiquemos a aplicação da função `leaps` com os dados `brix`.

```
> colnames(brix)       <-- para ver nomes das variáveis
[1] "Diametro" "Altura"   "Peso"     "Brix"     "pH"       "Acucar"

> library(leaps)       <-- para 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 preditores escolhidos
      1     2     3     4     5    <-- colunas <-> variáveis preditoras;
                                       linhas <-> dimensão k de subconjunto
1 FALSE FALSE FALSE FALSE  TRUE    <-- k=1 ; melhor preditor individual é Acucar
2  TRUE  TRUE FALSE FALSE FALSE    <-- k=2 ; melhor par de preditores é Diametro e Altura
3  TRUE  TRUE FALSE FALSE  TRUE    <-- k=3 ; melhor trio de preditores: Diametro, Altura e 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

Alternativamente, podem usar-se algoritmos de pesquisa mais ligeiros computacionalmente, mas que não analisam todo os possíveis submodelos e não garantem a obtenção dos melhores subconjuntos.

Algoritmos simples deste tipo são sequenciais, alterando uma variável preditora em cada passo do algoritmo, até se alcançar uma condição de paragem. Em particular, os algoritmos sequenciais podem ser:

- de exclusão sequencial (*backward elimination*) quando, partindo do modelo completo, consideram a possível exclusão duma variável em cada passo do algoritmo.
- de inclusão sequencial (*forward selection*) quando, partindo do modelo nulo, consideram a possível inclusão duma variável em cada passo do algoritmo
- de exclusão/inclusão alternada (*stepwise selection*) quando, para uma dada "direcção de marcha" pré-fixada, admitem alternar exclusões/inclusões.

Algoritmos sequenciais com base no AIC

O R disponibiliza funções para automatizar pesquisas sequenciais de submodelos em que o critério de exclusão/inclusão duma variável em cada passo se baseia no 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, pode definir-se como

Critério de Informação de Akaike no Modelo Linear

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

Um modelo para a variável resposta Y é considerado melhor que outro se tiver um AIC menor (o que favorece modelos com $SQRE$ menor, mas também com menos parâmetros).

Algoritmos sequenciais com base no AIC (cont.)

Num algoritmo de exclusão sequencial, com base no critério AIC:

- ajusta-se o modelo completo e calcula-se o respectivo AIC.
- ajustam-se todos os submodelos com menos uma variável, e calculam-se os respectivos AICs.
- Se nenhum dos AICs obtidos excluindo uma variável fôr inferior ao AIC do modelo anterior, o algoritmo termina sendo o modelo anterior o modelo final.
  Caso alguma das exclusões reduza o AIC, exclui-se o preditor associado à maior redução de AIC e regressa-se ao ponto anterior.

Algoritmos de selecção sequencial no R

A função `step` corre algoritmos de selecção sequencial, com base no AIC.

Considere-se de novo o exemplo dos dados `brix`:

```
> brix.lm <- lm(Brix ~ Diametro+Altura+Peso+pH+Acucar, data = brix)
> step(brix.lm, dir="backward")

Start:  AIC=-51.58
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
```

Neste caso, não se exclui qualquer variável: o AIC do modelo inicial é menor que o de qualquer submodelo com menos um preditor. O modelo final é o modelo inicial.

Uma palavra final sobre algoritmos de pesquisa

Os algoritmos de selecção sequencial não garantem a identificação do "melhor submodelo" com um dado número de preditores. Apenas identificam, de forma que não é computacionalmente muito pesada, submodelos que se presume serem "bons".

Devem ser usados com bom senso e os submodelos obtidos cruzados 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).

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

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

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

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

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:

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

O vector dos resíduos $\vec{\mathbf{E}} = (\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}$, tem distribuição Multinormal (em sentido generalizado) pelo último ponto do Teorema do acetato 121.

O vector esperado de $\vec{\mathbf{E}}$ resulta das propriedades do acetato 116:

- $E[\vec{\mathbf{E}}] = E[(\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}] = (\mathbf{I}_n - \mathbf{H})E[\vec{\mathbf{Y}}] = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\vec{\boldsymbol{\beta}} = \vec{\mathbf{0}}$,
  pois o vector $\mathbf{X}\vec{\boldsymbol{\beta}} \in \mathscr{C}(\mathbf{X})$, logo permanece invariante sob a acção da matriz de projecções $\mathbf{H}$: $\mathbf{H}\mathbf{X}\vec{\boldsymbol{\beta}} = \mathbf{X}\vec{\boldsymbol{\beta}}$.

Propriedades dos Resíduos no Modelo Linear (cont.)

A matriz de covariâncias de $\vec{\mathbf{E}}$ calcula-se tendo em conta as propriedades do acetato 117 e o facto da matriz de projecção ortogonal $\mathbf{H}$ ser simétrica ($\mathbf{H}^t = \mathbf{H}$) e idempotente ($\mathbf{HH} = \mathbf{H}$):

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

O resultado anterior diz-nos que 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 $\mathbf{H}$.

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 $\mathbf{H}$.

Vários tipos de resíduos

Definem-se diferentes tipos de resíduos:

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

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

Resíduos Studentizados (ou externamente estandardizados):

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

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

Análise dos resíduos

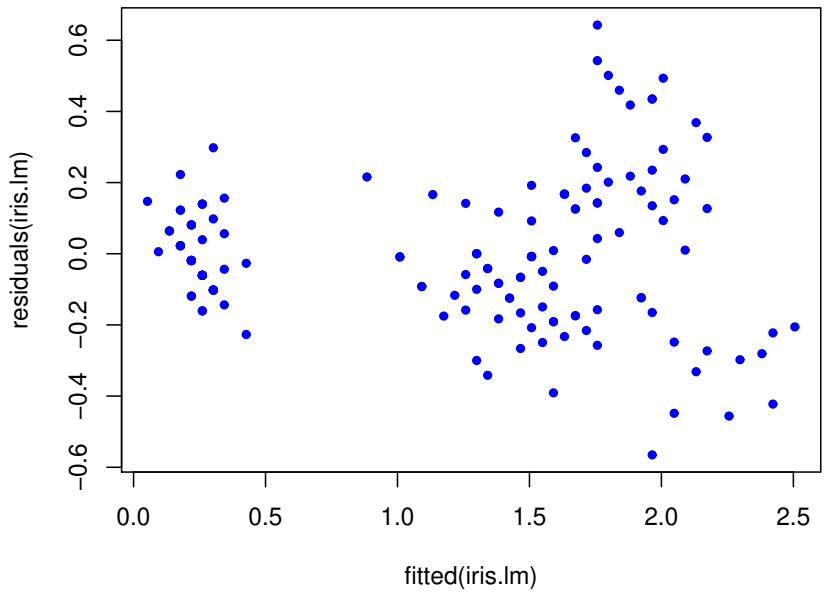
Nas regressões lineares, avalia-se a validade dos pressupostos do modelo através de gráficos de resíduos. Não se efectuam testes de Normalidade, uma vez que os resíduos não são (em geral) independentes.

Os gráficos mais usuais são os seguintes:

- gráfico dos E_i vs. $\hat{Y}_i$: 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 neste gráfico.
- gráfico de resíduos vs. ordem de observação: para investigar eventuais faltas de independência dos erros aleatórios.

Gráficos de resíduos vs. $\hat{Y}_i$

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



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

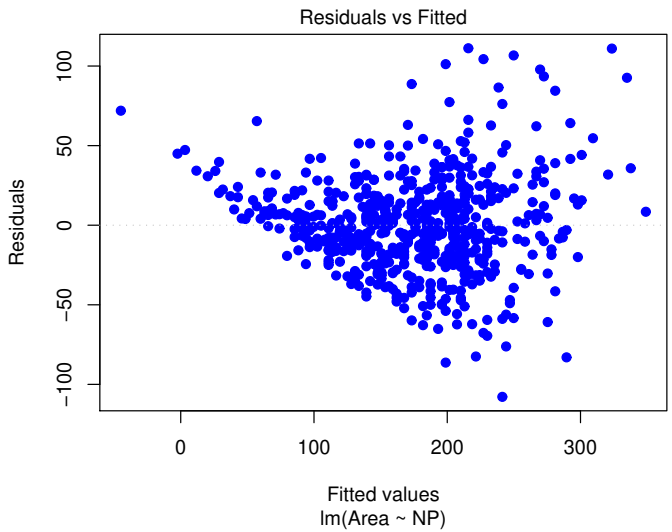
Possíveis padrões indicativos de problemas

Num gráfico de E_i vs. $\hat{Y}_i$ surgem com frequência alguns padrões indicativos de problemas.

Curvatura na disposição dos resíduos Indica violação da hipótese de linearidade entre x e y.

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

Um ou mais resíduos muito destacados Indica a possível existência de observações atípicas que podem estar a afectar o ajustamento.



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

Gráficos para estudar a hipótese de normalidade

Como foi visto no acetato 176, dado o ML, $\frac{E_i}{\sqrt{\sigma^2 \cdot (1-h_{ii})}} \frown \mathcal{N}(0,1)$.

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

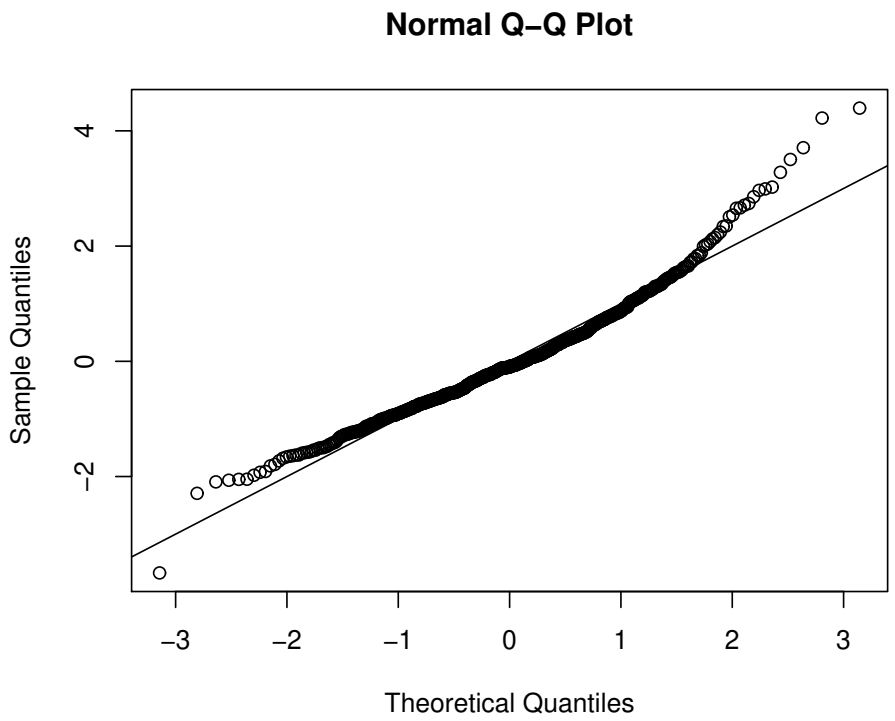
É hábito investigar a validade do pressuposto de erros aleatórios Normais através de:

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

Gráficos para o estudo da Normalidade (cont.)

Um qq-plot indicativo de concordância com a hipótese de Normalidade dos erros aleatórios deverá apresentar colinearidade aproximada.

O exemplo seguinte sugere algum desvio à Normalidade para os resíduos mais extremos.



Gráficos para o estudo de independência

Dependência entre erros aleatórios pode surgir com observações que sejam sequenciais no tempo (como resultado, por exemplo, de um "tempo de retorno" de um aparelho de medição).

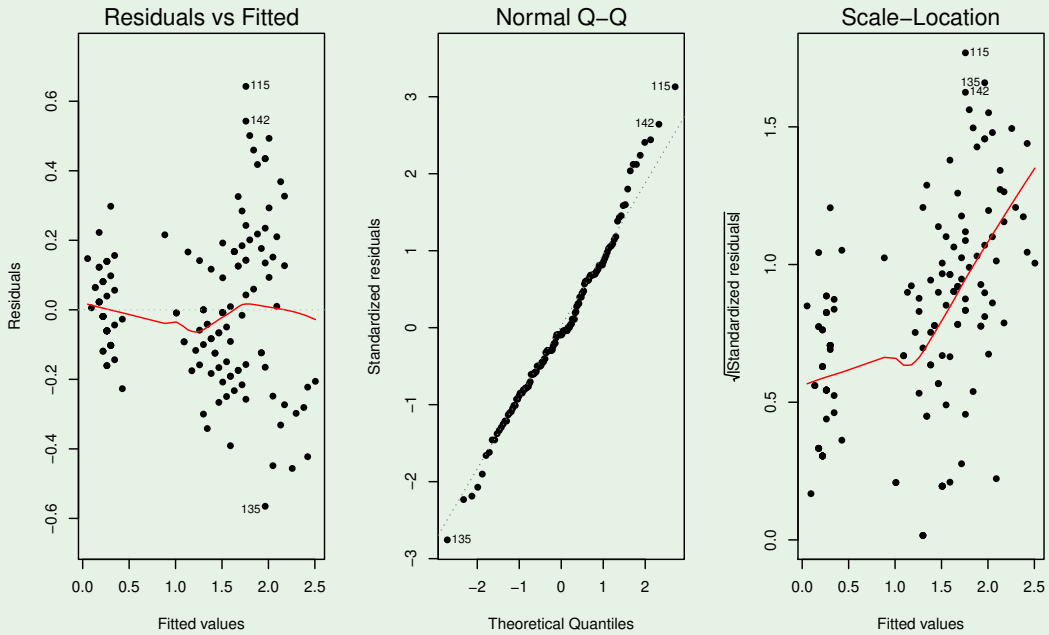
Nesse caso, pode ser útil inspeccionar um gráfico de resíduos vs. ordem de observação, para verificar se existem padrões que sugiram falta de independência.

Falta de independência ocorre igualmente com dados espaciais. No caso destes efeitos serem importantes, serão necessários modelos alternativos específicos para esse tipo de dados.

Estudo de resíduos no 

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

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



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* em inglês). Conceito sem definição rigorosa, procura designar observações que se distanciam da relação linear de fundo entre Y e a variável preditora.

Muitas vezes surgem associadas a resíduos grandes (em módulo). Em particular, e como os resíduos Studentizados têm distribuição aproximadamente $\mathcal{N}(0,1)$ para n grande, observações para as quais $|T_i| > 3$ 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 $\mathbf{H}$: $h_{ii} = \mathbf{H}_{(i,i)}$.

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

De facto (ver acetato 176), $V[E_i] = \sigma^2 (1 - h_{ii})$. Se h_{ii} é elevado, a variância do resíduo E_i é baixa, logo o resíduo tende a estar próximo do seu valor médio (zero), ou seja, a superfície ajustada tende a passar próximo desse ponto.

Observações alavanca (cont.)

Verifica-se, para qualquer observação:

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

Se os valores dos preditores da i-ésima observação forem repetidos num total de r observações, o efeito alavanca não pode exceder $\frac{1}{r}$. Assim, repetir observações de Y para os mesmos valores da variável preditora é uma forma de impedir efeitos alavancas excessivos.

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

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

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).

Numa regressão linear simples, tem-se

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

pelo que, numa RLS, quanto mais afastado estiver o valor x_i em relação à média $\overline{x}$, maior será o efeito alavanca.

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:

$$D_i = \frac{\|\vec{\hat{\mathbf{y}}} - \vec{\hat{\mathbf{y}}}_{(-i)}\|^2}{(p+1) \cdot QMRE},$$

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

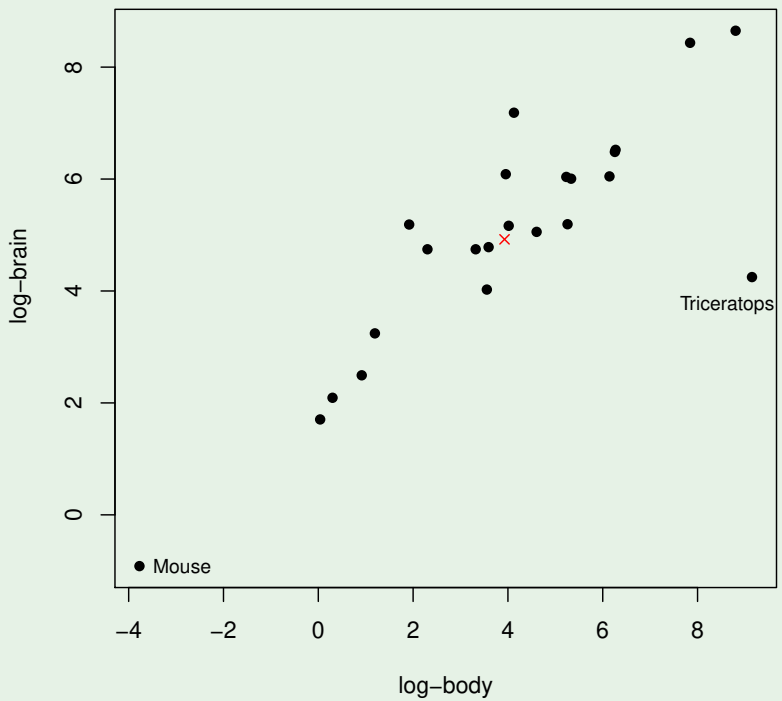
Observações atípicas, influentes ou alavanca, embora podendo estar relacionadas, não são o mesmo conceito.

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

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

Um exemplo na RLS

Considerando apenas um subconjunto das espécies animais do Exercício 4 de Regressão Linear, obtém-se o seguinte gráfico de log-peso do cérebro *vs.* log-peso do corpo:



Um exemplo na RLS (cont.)

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

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

Gráficos diagnósticos no 

A função `plot`, aplicada a um objecto `lm` produz, além dos gráficos vistos no acetato 186, gráficos com alguns dos diagnósticos agora considerados.

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

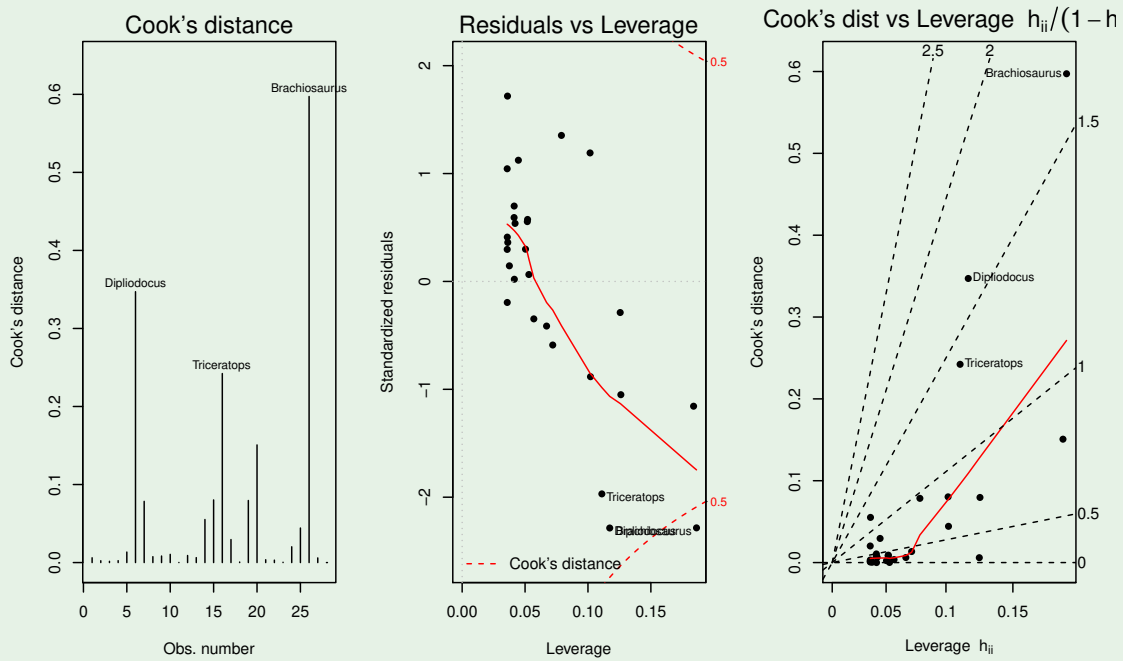
A opção `which=5` produz um gráfico de Resíduos estandardizados (R_is) no eixo vertical contra valores de h_{ii} (*leverages*) no eixo horizontal, traçando linhas de igual distância de Cook (para os níveis 0.5 e 1, por omissão), que destacam eventuais observações influentes.

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 (resultantes da última fórmula do acetato 191).

Um exemplo de gráficos de diagnóstico

Eis estes gráficos de diagnóstico, para os dados `Animals` (Ex. 6 RL):

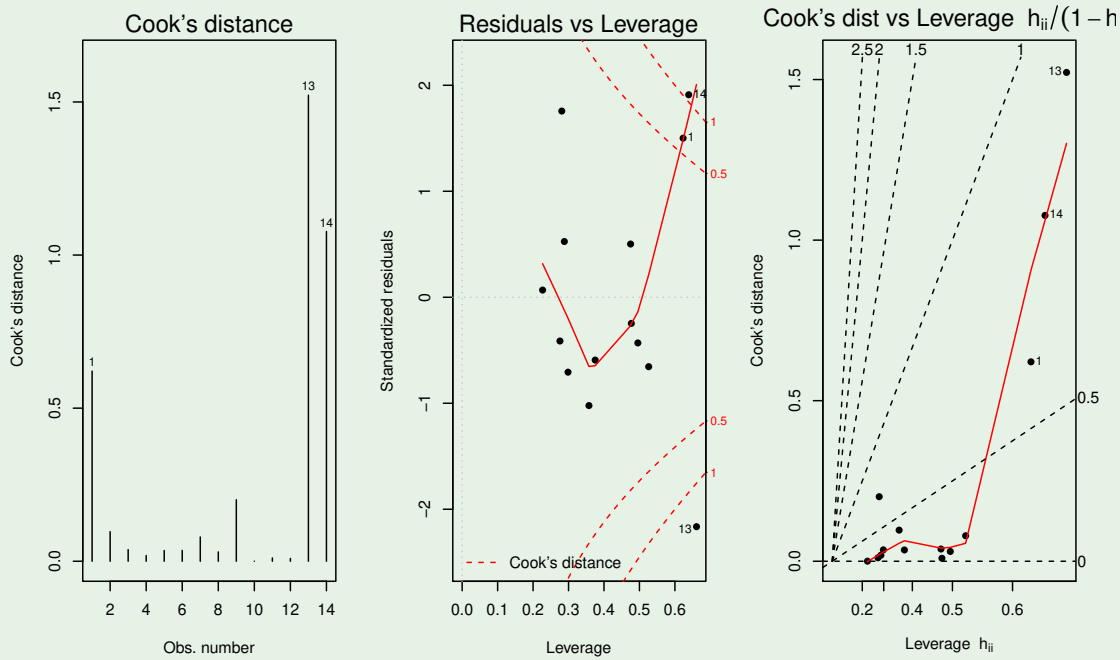
```
> plot(Animals.lm, which=4:6)
```



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.

Outro exemplo de gráficos de diagnóstico

Outro exemplo destes gráficos de diagnósticos, para os dados Brix:



Os valores muito elevados de distância de Cook e h_{ii} reflectem o reduzido número de observações ($n=14$) no ajustamento dum modelo com muitos parâmetros ($p+1=6$).

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.

Transformações linearizantes

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

Prevenções sobre transformações linearizantes:

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

O R^2 modificado (*adjusted* R^2)

Uma variante do Coeficiente de Determinação é o R^2 modificado.

O Coeficiente de Determinação usual:

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

O R^2 modificado (sendo $QMT = SQT/(n-1)$):

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

Para qualquer modelo linear (com preditores), verifica-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 o número de variáveis preditoras, $R^2_{mod} \ll R^2$ (excepto se $R^2 \approx 1$).

O R^2 modificado (cont.)

Viu-se que o valor de R^2_{mod} penaliza modelos complexos ajustados com poucas observações. Exercício 10: 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: Exercício 19

(também 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
```

Três advertências

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 numéricos 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 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.

Três advertências (cont.)

2. Tal como na Regressão Linear Simples, podem ser encaradas transformações, quer da variável resposta, quer de uma ou várias das variáveis preditoras.

Em particular, podem ser úteis transformações que linearizem a relação entre Y e X_1, X_2, ..., X_p. Tais transformações linearizantes podem permitir estudar relações de tipo não-linear através de relações lineares entre as variáveis transformadas.

E.g., a relação não linear entre Y, x_1 e x_2,

$$Y = \beta_0 x_1^{\beta_1} x_2^{\beta_2}$$

torna-se, após uma logaritmização, numa relação linear entre $\ln(Y)$, $\ln(x_1)$ e $\ln(x_2)$ (com $\beta_0^* = \ln(\beta_0)$):

$$\ln(Y) = \beta_0^* + \beta_1 \ln(x_1) + \beta_2 \ln(x_2) \ .$$

Três advertências (cont.)

3. 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.

II.3. Análise de Variância (ANOVA)

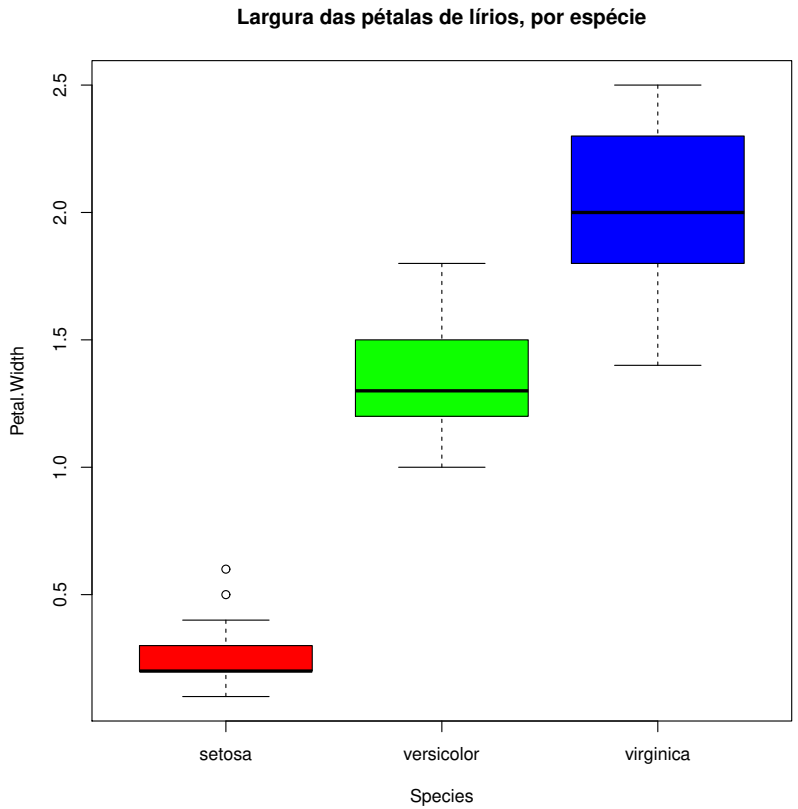
A Regressão Linear visa modelar uma variável resposta numérica (quantitativa), à custa de uma ou mais variáveis preditoras, igualmente numéricas.

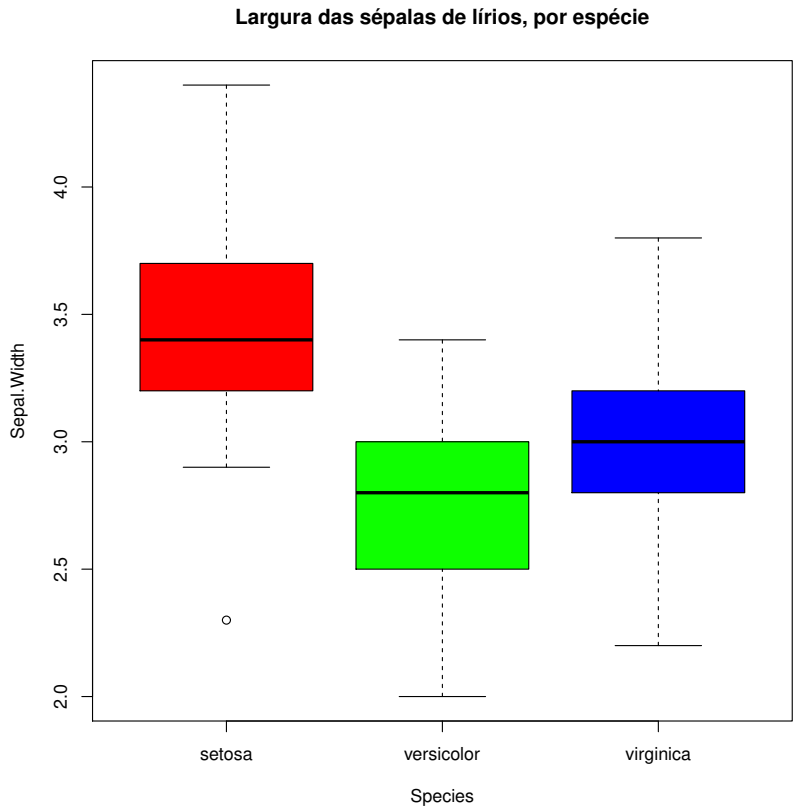
Mas uma variável resposta numérica pode depender de variáveis qualitativas (categóricas), ou seja, de um ou mais factores.

A Análise de Variância (ANOVA) é uma metodologia estatística para lidar com este tipo de situações.

A ANOVA foi desenvolvida nos anos 30 do Século XX, na Estação Experimental Agrícola de Rothamstead (Inglaterra), por R.A. Fisher.

Dois exemplos: os lírios por espécie





As larguras das pétalas parecem diferir entre as espécies dos lírios.

As larguras das sépalas diferem menos.

Pode afirmar-se que as diferenças observadas reflectem verdadeiras diferenças nos valores médios populacionais de cada espécie?

A ANOVA como caso particular do Modelo Linear

Embora a Análise de Variância tenha surgido como método autónomo, quer a Análise de Variância, quer a Regressão Linear, são particularizações do Modelo Linear.

Introduzir a ANOVA através das suas semelhanças com a Regressão Linear permite aproveitar boa parte da teoria estudada até aqui.

Terminologia:

Variável resposta Y: uma variável numérica (quantitativa), que se pretende estudar e modelar.

Factor : uma variável preditora categórica (qualitativa);

Níveis do factor : as diferentes categorias ("valores") do factor, ou seja, diferentes situações experimentais onde se efectuam observações de Y.

A ANOVA a um Factor

O mais simples modelo ANOVA é a ANOVA a um Factor (totalmente casualizado), ou seja, um modelo para situações onde a modelação da variável resposta (numérica) se baseia numa única variável preditora categórica.

Para estudar os efeitos dum factor, com k níveis, sobre uma variável resposta Y, admitimos que temos n observações independentes de Y, sendo n_i $(i = 1, ..., k)$ correspondentes ao nível i do factor. Logo,

$$n_1 + n_2 + \cdots + n_k = n\,.$$

No caso de igual número de observações em cada nível,

$$n_1 = n_2 = n_3 = \cdots = n_k \qquad (= n_c)\,,$$

diz-se que estamos perante um delineamento equilibrado. Por múltiplas razões, delineamentos equilibrados são aconselháveis.

A dupla indexação de Y

Na regressão indexam-se as n observações de Y com um único índice, variando de 1 a n.

Neste novo contexto, é preferível utilizar dois índices para indexar as observações de Y:

- um (i) indica o nível do factor a que a observação corresponde;
- outro (j) permite distinguir as observações num mesmo nível.

Assim, a j-ésima observação de Y, no i-ésimo nível do factor, é representada por Y_{ij} , (com $i = 1, ..., k$ e $j = 1, ..., n_i$) .

Um modelo para Y_{ij}

Admite-se que os valores de Y poderão variar por:

- corresponderem a níveis diferentes do factor; ou
- devido a flutuação aleatória.

A natureza mais pobre da nossa variável preditora estará associada a um modelo mais simples do que na regressão.

Em geral, admitimos que o valor esperado (médio) de Y pode diferir nas k situações (níveis do factor) em que é observado.

Uma primeira formulação do modelo é dada pela equação de base:

$$Y_{ij} = \mu_i + \varepsilon_{ij} \qquad \text{com} \quad E[\varepsilon_{ij}] = 0 \ .$$

Aqui, μ_i representa o valor esperado das observações Y_{ij} efectuadas no nível i do factor.

Um modelo para Y_{ij} (cont.)

Para poder enquadrar a ANOVA na teoria do Modelo Linear já estudada, é conveniente re-escrever as médias de nível na forma:

$$E[Y_{ij}] \;=\; \mu_i \;=\; \mu + \alpha_i \;.$$

O parâmetro μ é comum a todas as observações, enquanto os parâmetros α_i são específicos para cada nível (i) do factor.
Cada α_i é designado o efeito do nível i.

Admite-se que Y_{ij} oscila aleatoriamente em torno do seu valor médio:

$$Y_{ij} \;=\; \mu + \alpha_i + \varepsilon_{ij} \;,$$

com $E[\varepsilon_{ij}] = 0$.

O modelo ANOVA como um Modelo Linear

A equação geral

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} ,$$

significa que as n_1 observações efectuadas no nível $i = 1$ ficam $Y_{1j} = \mu + \alpha_1 + \varepsilon_{1j}$; as n_2 observações efectuadas no nível $i = 2$ ficam $Y_{2j} = \mu + \alpha_2 + \varepsilon_{2j}$; e assim por diante.

Para encaixar este conjunto de equações no contexto do modelo linear, a equação geral pode ser vista como sendo da forma:

$$Y_{ij} = \mu + \alpha_1 \mathscr{I}_{1_{ij}} + \alpha_2 \mathscr{I}_{2_{ij}} + \ldots + \alpha_k \mathscr{I}_{k_{ij}} + \varepsilon_{ij} ,$$

onde as variáveis indicatrizes de nível do factor se definem como:

$$\mathscr{I}_{m_{ij}} = \begin{cases} 1 & \text{se} \quad i = m , \\ 0 & \text{se} \quad i \neq m . \end{cases}$$

O modelo ANOVA como um Modelo Linear (cont.)

A equação de base do modelo ANOVA a um factor pode ser escrito na forma vectorial/matricial, como no modelo de regressão linear. Seja

- $\vec{\mathbf{Y}}$ o vector n-dimensional com a totalidade das observações da variável resposta. Admite-se que as n_1 primeiras correspondem ao nível 1 do factor, as n_2 seguintes ao nível 2, e assim de seguida.
- $\vec{\mathbf{1}}_n$ o vector de n uns, já considerado na regressão.
- $\vec{\mathscr{I}}_i$ o vector da variável indicatriz do nível i do factor. Para cada observação, esta variável toma o valor 1 se a observação corresponde ao nível i do factor, e o valor 0 caso contrário ($i = 1, ..., k$). Numa ANOVA, as variáveis indicatrizes desempenham o papel dos preditores.
- $\vec{\boldsymbol{\varepsilon}}$ o vector dos n erros aleatórios.

Os vectores das variáveis indicatrizes

Por exemplo, se se fizerem $n = 9$ observações, com $n_1 = 3$ observações no primeiro nível do factor, $n_2 = 4$ no segundo nível e $n_3 = 2$ observações no terceiro nível, os vectores $\vec{\mathscr{I}}_2$ e $\vec{\mathscr{I}}_3$ serão:

$$\vec{\mathscr{I}}_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad , \quad \vec{\mathscr{I}}_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

A relação de base em notação vectorial

Em notação matricial/vectorial, a equação de base que descreve as n observações de Y pode escrever-se como no Modelo Linear:

$$\begin{aligned} \vec{\mathbf{Y}} &= \mu\,\vec{\mathbf{1}}_n + \alpha_1\,\vec{\mathscr{I}}_1 + \alpha_2\,\vec{\mathscr{I}}_2 + \alpha_3\,\vec{\mathscr{I}}_3 + \vec{\boldsymbol{\varepsilon}} \\ \Leftrightarrow \quad \vec{\mathbf{Y}} &= \mathbf{X}\vec{\boldsymbol{\beta}} \quad + \quad \vec{\boldsymbol{\varepsilon}}\,, \end{aligned}$$

sendo as colunas da matriz $\mathbf{X}$ constituidas pelo vector dos n uns e pelas variáveis indicatrizes; e o vector dos parâmetros $\vec{\boldsymbol{\beta}}$ constituido por μ e os efeitos α_i.

No exemplo com as $n_1 = 3$, $n_2 = 4$ e $n_3 = 2$ observações:

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \\ Y_{31} \\ Y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix}$$

O problema do excesso de parâmetros

Existe um problema "técnico": as colunas desta matriz $\mathbf{X}$ são linearmente dependentes, pelo que a matriz $\mathbf{X}^t\mathbf{X}$ não é invertível.

Existe um excesso de parâmetros no modelo. Soluções possíveis:

1. retirar o parâmetro μ do modelo.
   - corresponde a retirar a coluna de uns da matriz $\mathbf{X}$;
   - cada α_i equivalerá a μ_i, a média do nível;
   - não se pode generalizar a situações mais complexas;
   - mais difícil de encaixar na teoria já dada do Modelo Linear.
2. tomar $\alpha_1 = 0$: será a solução utilizada.
   - corresponde a excluir a 1a. variável indicatriz do modelo (e de $\mathbf{X}$);
   - permite aproveitar a teoria do Modelo Linear e é generalizável.
3. impor restrições aos parâmetros: e.g., $\sum_{i=1}^{k} \alpha_i = 0$.
   - Foi a solução clássica, ainda hoje frequente em livros de ANOVA;
   - mais difícil de encaixar na teoria geral do Modelo Linear.

Cada solução tem implicações na forma de interpretar os parâmetros.

A relação de base para o nosso exemplo (cont.)

Admitindo $\alpha_1 = 0$, re-escrevemos o modelo como:

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \\ Y_{31} \\ Y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix}$$

Agora $\mu = \mu_1$ é o valor médio das observações do nível $i = 1$:

$$\begin{aligned} E[Y_{1j}] &= \mu_1 && , \forall j = 1, ..., n_1 \\ E[Y_{2j}] = \mu_2 &= \mu_1 + \alpha_2 && , \forall j = 1, ..., n_2 \\ E[Y_{3j}] = \mu_3 &= \mu_1 + \alpha_3 && , \forall j = 1, ..., n_3 \end{aligned}$$

Os efeitos de nível α_i

No modelo para uma ANOVA a um factor (acetato 212), cada α_i ($i > 1$) representa o acréscimo que transforma a média do primeiro nível na média do nível i:

$$\begin{aligned} \alpha_1 &= 0 \\ \alpha_2 &= \mu_2 - \mu_1 \\ \alpha_3 &= \mu_3 - \mu_1 \\ &\vdots \quad \vdots \quad \vdots \\ \alpha_k &= \mu_k - \mu_1 \end{aligned}$$

A igualdade de todas as médias populacionais de nível μ_i equivale a que todos os efeitos de nível sejam nulos: $\alpha_i = 0\ , \quad \forall\ i$.

Consideremos agora os estimadores destes parâmetros.

A matriz $\mathbf{X}$ numa ANOVA a um factor

Na ANOVA a um factor, a matriz $\mathbf{X}$ tem nas suas k colunas os vectores $\vec{\mathbf{1}}_n, \vec{\mathscr{I}}_2, \vec{\mathscr{I}}_3, \ldots, \vec{\mathscr{I}}_k$ e indica quais as observações correspondentes a cada nível do factor.

Como a equação do modelo ANOVA é um caso particular da equação do Modelo Linear, a fórmula dos parâmetros ajustados pelo método dos mínimos quadrados é igualmente

$$\vec{\mathbf{b}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{y}} \, .$$

A natureza especial da matriz $\mathbf{X}$ na ANOVA (os seus elementos só tomam valores 0 e 1) faz com que os resultados gerais, válidos para qualquer Modelo Linear, produzam expressões específicas no contexto da ANOVA, existindo fórmulas fáceis para cada estimador dum parâmetro individual.

Os parâmetros ajustados

Com $\overline{Y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$ a média das n_i observações de Y no nível i,

tem-se que os parâmetros populacionais são estimados pelas quantidades amostrais correspondentes:

Parâmetros estimados numa ANOVA a um factor

$$\begin{aligned}
\hat{\mu}_1 &= \overline{Y}_{1.} \\
\hat{\alpha}_2 &= \overline{Y}_{2.} - \overline{Y}_{1.} \\
\hat{\alpha}_3 &= \overline{Y}_{3.} - \overline{Y}_{1.} \\
\vdots \quad &\vdots \qquad \vdots \\
\hat{\alpha}_k &= \overline{Y}_{k.} - \overline{Y}_{1.}
\end{aligned}$$

Os estimadores das médias de nível

Dados os estimadores referidos no acetato anterior, e uma vez que as médias de cada nível (além do primeiro) são dadas por $\mu_i = \mu_1 + \alpha_i$, temos que os estimadores de cada média de nível são

$$\begin{aligned}
\hat{\mu}_1 &= \overline{Y}_{1\cdot} \\
\hat{\mu}_2 = \hat{\mu}_1 + \hat{\alpha}_2 &= \overline{Y}_{2\cdot} \\
\hat{\mu}_3 = \hat{\mu}_1 + \hat{\alpha}_3 &= \overline{Y}_{3\cdot} \\
&\vdots \\
\hat{\mu}_k = \hat{\mu}_1 + \hat{\alpha}_k &= \overline{Y}_{k\cdot}
\end{aligned}$$

sendo $\overline{Y}_{i\cdot}$ a média das n_i observações de Y no nível i do factor.

Os valores ajustados $\hat{Y}_{ij}$

Do que foi visto, decorre que qualquer observação tem valor ajustado:

$$\hat{Y}_{ij} = \hat{\mu}_i = \hat{\mu}_1 + \hat{\alpha}_i = \overline{Y}_{i\cdot} \ .$$

Ou seja, os valores ajustados $\hat{Y}_{ij}$ são iguais para todas as observações num mesmo nível i do factor, e são dadas pela média amostral das observações nesse nível.

Tal como na Regressão, os valores ajustados $\hat{Y}$ resultam de projectar ortogonalmente o vector $\vec{\mathbf{Y}}$ dos valores observados da variável resposta, sobre o subespaço $\mathscr{C}(\mathbf{X})$ gerado pelas colunas da matriz $\mathbf{X}$.

Numa ANOVA a um factor, o subespaço $\mathscr{C}(\mathbf{X})$ tem natureza especial.

O subespaço $\mathscr{C}(\mathbf{X})$ numa ANOVA a um factor

Qualquer vector no subespaço $\mathscr{C}(\mathbf{X})$ tem de ter valores iguais para todas as observações dum mesmo nível do factor:

$$a_1 \vec{\mathbf{1}}_n + a_2 \vec{\mathscr{I}}_2 + a_3 \vec{\mathscr{I}}_3 + \ldots + a_k \vec{\mathscr{I}}_k = \begin{bmatrix} a_1 \\ \ldots \\ a_1 \\ \hline a_1 + a_2 \\ \ldots \\ a_1 + a_2 \\ \hline a_1 + a_3 \\ \ldots \\ a_1 + a_3 \\ \hline (\ldots) \\ \hline a_1 + a_k \\ \ldots \\ a_1 + a_k \end{bmatrix}$$

O vector $\vec{\hat{\mathbf{Y}}}$ pertence a $\mathscr{C}(\mathbf{X})$, logo tem esta natureza.

O modelo ANOVA a 1 factor para efeitos inferenciais

Modelo ANOVA a um factor, com k níveis

Existem n observações, Y_{ij}, n_i das quais associadas ao nível i ($i = 1, ..., k$) do factor. Tem-se:

1. $Y_{ij} = \mu_1 + \alpha_i + \varepsilon_{ij}$, $\forall i=1,...,k$, $\forall j=1,...,n_i$ $(\alpha_1 = 0)$.
2. $\varepsilon_{ij} \frown \mathcal{N}(0, \sigma^2)$, $\forall i, j$
3. $\{\varepsilon_{ij}\}_{i,j}$ v.a.s independentes.

O modelo tem k parâmetros desconhecidos: a média de Y no primeiro nível do factor, μ_1, e os acréscimos α_i ($i > 1$) que geram as médias de cada um dos $k-1$ restantes níveis do factor. Ou seja,

$$\vec{\boldsymbol{\beta}} = (\mu_1, \alpha_2, \alpha_3, \cdots, \alpha_k)^t .$$

O modelo ANOVA a um factor - notação vectorial

De forma equivalente, em notação vectorial,

Modelo ANOVA a um factor - notação vectorial

O vector $\vec{\mathbf{Y}}$ das n observações verifica:

1. $\vec{\mathbf{Y}} = \mu_1 \vec{\mathbf{1}}_n + \alpha_2 \vec{\mathscr{I}}_2 + \alpha_3 \vec{\mathscr{I}}_3 + ... + \alpha_k \vec{\mathscr{I}}_k + \vec{\boldsymbol{\varepsilon}} = \mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\varepsilon}}$, sendo $\vec{\mathbf{1}}_n$ o vector de n uns; $\vec{\mathscr{I}}_2$, $\vec{\mathscr{I}}_3$, ..., $\vec{\mathscr{I}}_k$ as variáveis indicatrizes dos níveis indicados; $\mathbf{X} = \left[\vec{\mathbf{1}}_n \mid \vec{\mathscr{I}}_2 \mid \vec{\mathscr{I}}_3 \mid \cdots \mid \vec{\mathscr{I}}_k \right]$ a matriz do modelo e $\vec{\boldsymbol{\beta}} = (\mu_1, \alpha_2, \alpha_3, \cdots, \alpha_k)^t$.
2. $\vec{\boldsymbol{\varepsilon}} \frown \mathscr{N}_n(\vec{\mathbf{0}}, \sigma^2 \mathbf{I}_n)$, sendo $\mathbf{I}_n$ a matriz identidade $n \times n$.

Trata-se de um modelo análogo a um modelo de Regressão Linear Múltipla, diferindo apenas na natureza das variáveis preditoras, que são aqui variáveis indicatrizes dos níveis 2 a k do factor.

O teste aos efeitos do factor

A hipótese de que nenhum dos níveis do factor afecte a média da variável resposta corresponde à hipótese

$$\begin{aligned} & \alpha_2 = \alpha_3 = \ldots = \alpha_k = 0 \\ \Leftrightarrow \quad \mu_1 = & \mu_2 = \mu_3 = \cdots = \mu_k \end{aligned}$$

Dado o paralelismo com os modelos de Regressão Linear, esta hipótese corresponde a dizer que todos os coeficientes das "variáveis preditoras" (na ANOVA, as variáveis indicatrizes $\vec{\mathscr{I}}_i$) são nulos. Logo, é possível testar esta hipótese, através dum teste F de ajustamento global do modelo (ver acetato 153).

Tratando-se dum caso particular do modelo linear, neste contexto há fórmulas específicas.

Os graus de liberdade

No contexto da ANOVA a um factor, chama-se SQF (de Factor), em vez de SQR, à Soma de Quadrados associada ao ajustamento do modelo (i.e., o numerador da variância dos valores ajustados de Y).

Numa ANOVA a um factor, o número de preditores do modelo (as variáveis indicatrizes dos níveis $2, 3, ..., k$) é $p = k - 1$ e o número de parâmetros do modelo é $p + 1 = k$. Logo, os graus de liberdade associados a cada Soma de Quadrados são:

SQxx	g.l.
SQF	$k - 1$
SQRE	$n - k$

Os Quadrados Médios continuam a ser os quocientes das Somas de Quadrados a dividir pelos respectivos graus de liberdade.

O Teste F aos efeitos do factor numa ANOVA

Sendo válido o Modelo de ANOVA a um factor, tem-se então:

Teste F aos efeitos do factor

Hipóteses: $H_0 : \alpha_i = 0 \quad \forall i=2,\ldots,k$ vs. $H_1 : \exists i=2,..,k \text{ t.q. } \alpha_i \neq 0$.

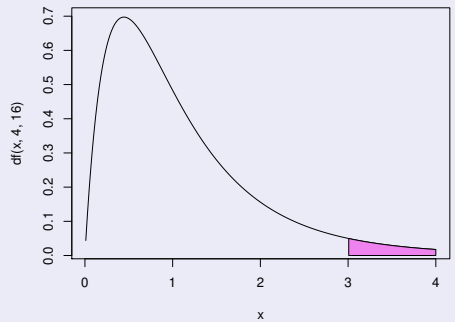
[FACTOR NÃO AFECTA] vs. [FACTOR AFECTA Y]

Estatística do Teste: $F = \frac{QMF}{QMRE} \frown F_{(k-1,n-k)}$ se H_0.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rej. H_0 se $F_{calc} > f_{\alpha(k-1,n-k)}$



Também as Somas de Quadrados e Quadrados Médios têm fórmulas específicas neste contexto.

Os resíduos e *SQRE*

Viu-se antes (acetato 222) que $\hat{Y}_{ij} = \hat{\mu}_i = \overline{Y}_{i\cdot}$, pelo que o resíduo da observação Y_{ij} é dado por:

$$E_{ij} \;=\; Y_{ij} - \hat{Y}_{ij} \;=\; Y_{ij} - \overline{Y}_{i\cdot}\;,$$

Logo, a Soma de Quadrados dos Resíduos é dada por:

$$SQRE \;=\; \sum_{i=1}^{k}\sum_{j=1}^{n_i} E_{ij}^2 \;=\; \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i\cdot}\right)^2 \;=\; \sum_{i=1}^{k} (n_i - 1)\, S_i^2\;,$$

onde $S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot})^2$ é a variância amostral das n_i observações de Y no i-ésimo nível do factor.

SQRE mede variabilidade no seio dos k níveis.

Fórmulas para delineamentos equilibrados

No caso de um delineamento equilibrado, i.e., $n_1 = n_2 = ... = n_k (= n_c)$ tem-se:

$$SQRE = (n_c - 1) \sum_{i=1}^{k} S_i^2$$

$$QMRE = \frac{n_c - 1}{n - k} \sum_{i=1}^{k} S_i^2 = \frac{1}{k} \sum_{i=1}^{k} S_i^2 \ ,$$

já que $n = n_c \cdot k$.

Assim, em delineamentos equilibrados, o Quadrado Médio Residual *QMRE* é a média das k variâncias de nível, nos valores da variável resposta Y.

Em delineamentos não equilibrados, o QMRE é uma média ponderada dos S_i^2.

A Soma de Quadrados associada ao Factor

A Soma de Quadrados associada ao Factor, *SQF*, é dada por:

$$SQF = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\hat{Y}_{ij} - \overline{Y}_{..} \right)^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\overline{Y}_{i.} - \overline{Y}_{..} \right)^2$$

$$\Leftrightarrow \quad SQF = \sum_{i=1}^{k} n_i \left(\overline{Y}_{i.} - \overline{Y}_{..} \right)^2$$

sendo $\overline{Y}_{..} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}$ a média da totalidade das n observações.

SQF mede variabilidade entre as médias amostrais de cada nível.

Fórmulas para delineamentos equilibrados

No caso de um delineamento equilibrado $n_1 = n_2 = ... = n_k (= n_c)$,

$$SQF \quad = \quad n_c \sum_{i=1}^{k} (\overline{Y}_{i.} - \overline{Y}_{..})^2 \; = \; n_c (k-1) \cdot S^2_{\overline{Y}_{i..}} \; ,$$

onde $S^2_{\overline{Y}_{i..}} = \frac{1}{k-1} \sum_{i=1}^{k} (\overline{Y}_{i.} - \overline{Y}_{..})^2$ indica a variância amostral das k médias de nível amostrais.

$$QMF \quad = \quad \frac{SQF}{k-1} \quad = \quad n_c \cdot S^2_{\overline{Y}_{i..}} \; .$$

Assim, em delineamentos equilibrados, o Quadrado Médio associado aos efeitos do Factor, QMF, é proporcional à variância das k médias de nível da variável Y.

A relação entre Somas de Quadrados

A relação fundamental entre as três Somas de Quadrados (mesmo com delineamentos não equilibrados) tem um significado particular:

$$\begin{array}{ccccc} SQT & = & SQF & + & SQRE \\ \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{..})^2 & = & \sum_{i=1}^{k} n_i(\overline{Y}_{i.}-\overline{Y}_{..})^2 & + & \sum_{i=1}^{k}(n_i-1)S_i^2 \ . \end{array}$$

onde:

SQT $= (n-1)s_y^2$ mede a variabilidade total das n observações de Y;

SQF mede a variabilidade entre diferentes níveis do factor (variabilidade inter-níveis);

$SQRE$ mede a variabilidade no seio de cada nível - e que portanto não é explicada pelo factor (variabilidade intra-níveis).

Esta é a origem histórica do nome "Análise da Variância": a variância de Y é decomposta ("analisada") em parcelas, associadas a diferentes causas. Neste modelo, as causas podem ser o efeito do factor ou outras não explicadas pelo modelo (residuais).

O quadro-resumo da ANOVA a 1 Factor

Pode-se coleccionar esta informação numa tabela-resumo da ANOVA.

Fonte	g.l.	SQ	QM	f_{calc}
Factor	$k-1$	$SQF = \sum_{i=1}^{k} n_i \cdot (\overline{y}_{i.} - \overline{y}_{..})^2$	$QMF = \frac{SQF}{k-1}$	$\frac{QMF}{QMRE}$
Resíduos	$n-k$	$SQRE = \sum_{i=1}^{k} (n_i - 1)\, s_i^2$	$QMRE = \frac{SQRE}{n-k}$	
Total	$n-1$	$SQT = (n-1)\, s_y^2$	–	–

ANOVAs a um Factor no R

Para efectuar uma ANOVA a um Factor no R, organizam-se os dados numa `data.frame` com duas colunas:

1. uma para os valores (numéricos) da variável resposta;
2. outra para o factor (com a indicação dos seus níveis).

A fórmula usada no `R` para especificar uma ANOVA a um factor é semelhante à duma regressão linear, indicando o factor como preditor.

Por exemplo, para efectuar uma ANOVA de larguras das pétalas sobre espécies, nos dados dos $n = 150$ lírios, a fórmula é:

$$\texttt{Petal.Width} \quad \sim \quad \texttt{Species}$$

uma vez que a *data frame* `iris` contém uma coluna de nome `Species` que foi definida como factor.

ANOVAs a um factor no R (cont.)

Embora seja possível usar o comando `lm` para efectuar uma ANOVA (a ANOVA é caso particular do Modelo Linear), existe outro comando que organiza a informação da forma mais tradicional numa ANOVA: `aov`.

E.g., a ANOVA da largura de pétalas sobre espécies para os lírios invoca-se da seguinte forma:

```
> aov(Petal.Width ~ Species, data=iris)

Call:  aov(formula = Petal.Width ~ Species, data = iris)
Terms:
                  Species Residuals
Sum of Squares   80.41333   6.15660
Deg. of Freedom        2       147

Residual standard error: 0.20465
```

O resultado produzido é diferente do obtido com o comando `lm`.

ANOVAs a um factor no R (cont.)

A função `summary` também pode ser aplicada ao resultado de uma ANOVA, produzindo o quadro-resumo completo da ANOVA.
Vejamos a ANOVA do primeiro dos dois exemplos que motivou esta discussão (acetato 207):

```
> iris.aov <- aov(Petal.Width ~ Species , data=iris)
> summary(iris.aov)
             Df Sum Sq Mean Sq F value    Pr(>F)
Species       2 80.413  40.207  960.01 < 2.2e-16 ***
Residuals   147  6.157   0.042
```

Neste caso, rejeita-se claramente a hipótese de que os acréscimos de nível, α_i, sejam todos nulos, pelo que se rejeita a hipótese de larguras médias de pétalas iguais em todas as espécies.

Conclusão: o factor (espécie) afecta a variável resposta (largura da pétala).

Os parâmetros estimados, no 

Para obter as estimativas dos parâmetros μ_1, α_2, α_3, ..., α_k, pode aplicar-se a função `coef` ao resultado da ANOVA.

No exemplo dos lírios, temos:

```
> coef(iris.aov)
(Intercept) Speciesversicolor  Speciesvirginica
      0.246             1.080              1.780
```

Estes são os valores estimados dos parâmetros

- $\hat{\mu}_1 = 0.246$: média amostral de larguras de pétalas *setosa*;
- $\hat{\alpha}_2 = 1.080$: acréscimo que, somado à média amostral das *setosa*, dá a média amostral das larguras de pétalas *versicolor*;
- $\hat{\alpha}_3 = 1.780$: acréscimo que, somado à média amostral das *setosa*, dá a média amostral das larguras de pétalas *virginica*.

Parâmetros estimados no R (cont.)

Para melhor interpretar os resultados, vejamos as médias por nível do factor da variável resposta, através da função `model.tables`, com o argumento `type=“means”`:

```
> model.tables(iris.aov , type="mean")
Tables of means
 Grand mean
    1.199333

 Species
Species
setosa versicolor  virginica
 0.246      1.326      2.026
```

O R ordena os níveis de um factor por ordem alfabética.

ANOVAs como modelo Linear no 

Também é possível estudar uma ANOVA através do comando `lm`, nomeadamente para fazer inferência sobre os parâmetros do modelo:

```
> summary(lm(Petal.Width ~ Species , data=iris))
Call: lm(formula = Petal.Width ~ Species, data = iris)
(...)
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.24600    0.02894    8.50 1.96e-14 ***
Speciesversicolor  1.08000    0.04093   26.39  < 2e-16 ***
Speciesvirginica   1.78000    0.04093   43.49  < 2e-16 ***
--
Residual standard error: 0.2047 on 147 degrees of freedom
Multiple R-squared: 0.9289, Adjusted R-squared: 0.9279
F-statistic:   960 on 2 and 147 DF,  p-value: < 2.2e-16
```

Material Complementar: A exploração ulterior de H_1

A Hipótese Nula, no teste F numa ANOVA a 1 Factor, afirma que todos os níveis do factor têm efeito nulo, isto é, que a média da variável resposta Y é igual nos k níveis do Factor:

$$\begin{aligned} & \alpha_2 = \alpha_3 = \ldots = \alpha_k = 0 \\ \Leftrightarrow \quad & \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k \end{aligned}$$

A Hipótese Alternativa diz que pelo menos um dos níveis do factor tem uma média de Y diferente do primeiro nível:

$$\begin{aligned} & \exists\, i \quad \text{tal que} \quad \alpha_i \neq 0 \qquad (i > 1) \\ \Leftrightarrow \quad & \exists\, i \quad \text{tal que} \quad \mu_1 \neq \mu_i \qquad (i > 1) \end{aligned}$$

Ou seja, nem todas as médias de nível de Y são iguais

MC: A exploração ulterior de H_1 (cont.)

Caso se opte pela Hipótese Alternativa, fica em aberto (excepto quando $k = 2$) a questão de saber quais os níveis do factor cujas médias diferem entre si.

Uma possibilidade consiste em efectuar testes aos α_is, com base na teoria já estudada anteriormente. Mas quanto maior fôr k, mais sub-hipóteses alternativas existem, mais testes haverá para fazer. A multiplicação do número de testes faz perder o controlo do nivel de significância α global para o conjunto de todos os testes.

É possível construir testes de hipóteses relativos a todas as diferenças $\mu_i - \mu_j$, definidas pelas médias populacionais de Y nos níveis i, j de um factor ($i, j = 1, ..., k$, com $i \neq j$), controlando o nível de significância global α do conjunto dos testes. Tais testes chamam-se testes de comparações múltiplas de médias.

MC: As comparações múltiplas

O nível de significância α nos testes de comparação múltipla é a probabilidade de rejeitar qualquer das hipóteses $\mu_i = \mu_j$, caso ela seja verdade, ou seja, é um nível de significância global.

Alternativamente, podem-se construir intervalos de confiança para cada diferença $\mu_i - \mu_j$, com um nível $(1-\alpha) \times 100\%$ de confiança de que os verdadeiros valores de $\mu_i - \mu_j$ pertencem a todos os intervalos.

A teoria mais usada para comparações múltiplas é a de Tukey.
A distribuição de Tukey é uma distribuição com dois parâmetros: k e ν associada à amplitude Studentizada duma amostra aleatória Normal.
No nosso contexto, k indica o número total de médias sob comparação e ν os g.l. do QMRE.

MC: Intervalos de Confiança de Tukey para $\mu_i - \mu_j$

Seja $q_{\alpha(k,n-k)}$ o valor que numa distribuição de Tukey com parâmetros k e $n-k$, deixa à direita uma região de probabilidade α.

Com $(1-\alpha)\times 100\%$ de confiança, todas as diferenças de médias de nível $\mu_i - \mu_j$ estão em intervalos da forma:

$$\left] \quad (\overline{y}_{i\cdot} - \overline{y}_{j\cdot}) - q_{\alpha(k,n-k)}\sqrt{\frac{QMRE}{n_c}} \quad , \quad (\overline{y}_{i\cdot} - \overline{y}_{j\cdot}) + q_{\alpha(k,n-k)}\sqrt{\frac{QMRE}{n_c}} \quad \right[$$

Qualquer intervalo deste tipo que não contenha o valor zero corresponde a afirmar que $\mu_i = \mu_j$ não é admissível.

MC: Testes de Tukey para $\mu_i - \mu_j = 0\ , \quad \forall\, i,j$

Alternativamente, é possível testar a Hipótese Nula de que todas as diferenças de pares de médias de nível, $\mu_i - \mu_j$, sejam nulas, em cujo caso

$$\left|\overline{Y}_{i\cdot} - \overline{Y}_{j\cdot}\right| \quad < \quad q_{\alpha\,(k,n-k)} \cdot \sqrt{\frac{QMRE}{n_c}}\ ,$$

com probabilidade $(1-\alpha)$. Qualquer diferença de médias amostrais de nível, $\overline{Y}_{i\cdot} - \overline{Y}_{j\cdot}$, cujo módulo exceda o limiar

$$q_{\alpha\,(k,n-k)} \cdot \sqrt{\frac{QMRE}{n_c}}$$

indica que, para esse par de níveis i,j, se deve considerar $\mu_i \neq \mu_j$.

O nível (global) de significância de todas estas comparações é α, ou seja, a probabilidade de se concluir que $\mu_i \neq \mu_j$ (para algum par i,j), se em todos os casos $\mu_i = \mu_j$, é α.

MC: Comparações Múltiplas de Médias no R

As comparações múltiplas de médias de nível, com base no resultado de Tukey, podem ser facilmente efectuadas no R.

Os valores da função distribuição cumulativa e os quantis $q_{\alpha\,(k,n-k)}$ duma distribuição de Tukey são calculados no R, através das funções `ptukey` e `qtukey`, respectivamente.

Para se obter o termo de comparação nos testes de hipóteses a que $\mu_i - \mu_j = 0$, o quantil de ordem $1-\alpha$ na distribuição de Tukey é obtido a partir do comando

```
> qtukey(1-α, k, n-k)
```

O valor de $\sqrt{QMRE}$ é dado pelo comando `aov`, sob a designação "*Residual standard error*".

MC: Comparações Múltiplas de Médias no R (cont.)

Os intervalos de Confiança a $(1-\alpha)\times 100\%$ para as diferenças de médias são obtidos através do comando `TukeyHSD`. Por exemplo, para o segundo exemplo relativo aos dados dos lírios (acetato 207):

```
> TukeyHSD(aov(Sepal.Width ~ Species, data=iris))
Tukey multiple comparisons of means
95% family-wise confidence level
$Species
                       diff         lwr        upr     p adj
versicolor-setosa    -0.658 -0.81885528 -0.4971447 0.0000000
virginica-setosa     -0.454 -0.61485528 -0.2931447 0.0000000
virginica-versicolor  0.204  0.04314472  0.3648553 0.0087802
```

O intervalo a 95% de confiança para $\mu_2 - \mu_1$ (`versicolor-setosa`) é

$$]\ -0.8189\ ,\ -0.4971\ [\ .$$

Neste exemplo, nenhum dos intervalos inclui o valor 0, concluindo-se que $\mu_i \neq \mu_j,\ \forall\ i \neq j$ (i.e., todas as médias de espécie são diferentes).

MC: Comparações Múltiplas de Médias no R (cont.)

O valor de prova indicado (`p adj`) deve ser interpretado como o valor de α para o qual cada diferença de médias, $\overline{y}_{i.} - \overline{y}_{j.}$, seria, pela primeira vez, considerado não significativo.

```
> TukeyHSD(aov(Sepal.Width ~ Species, data=iris))
Tukey multiple comparisons of means
95% family-wise confidence level
$Species
                           diff         lwr        upr     p adj
versicolor-setosa    -0.658 -0.81885528 -0.4971447 0.0000000
virginica-setosa     -0.454 -0.61485528 -0.2931447 0.0000000
virginica-versicolor  0.204  0.04314472  0.3648553 0.0087802
```

Assim, para $\alpha = 0.00878$, a diferença de médias amostrais para as espécies *virginica* e *versicolor* já seria considerada não significativa. Ou seja, um intervalo com mais de $(1 - \alpha) \times 100\% = 99.122\%$ de confiança para essa diferença de médias conteria o valor zero.

MC: Delineamentos não equilibrados

Nota: O resultado de Tukey é válido para delineamentos equilibrados.

Quando o delineamento da ANOVA a um Factor não é equilibrado (isto é, existe diferente número de observações nos vários níveis do factor), os teste/ICs de Tukey agora enunciados não são, em rigor, válidos.

Mas, para delineamentos em que o desequilíbrio no número de observações não seja muito acentuado, é possível um resultado aproximado, que a função `TukeyHSD` do  incorpora.

Análise de Resíduos na ANOVA a 1 Factor

A validade dos pressupostos do modelo estuda-se de forma idêntica ao que foi visto na Regressão Linear, tal como os diagnósticos para observações especiais. Mas há algumas particularidades.

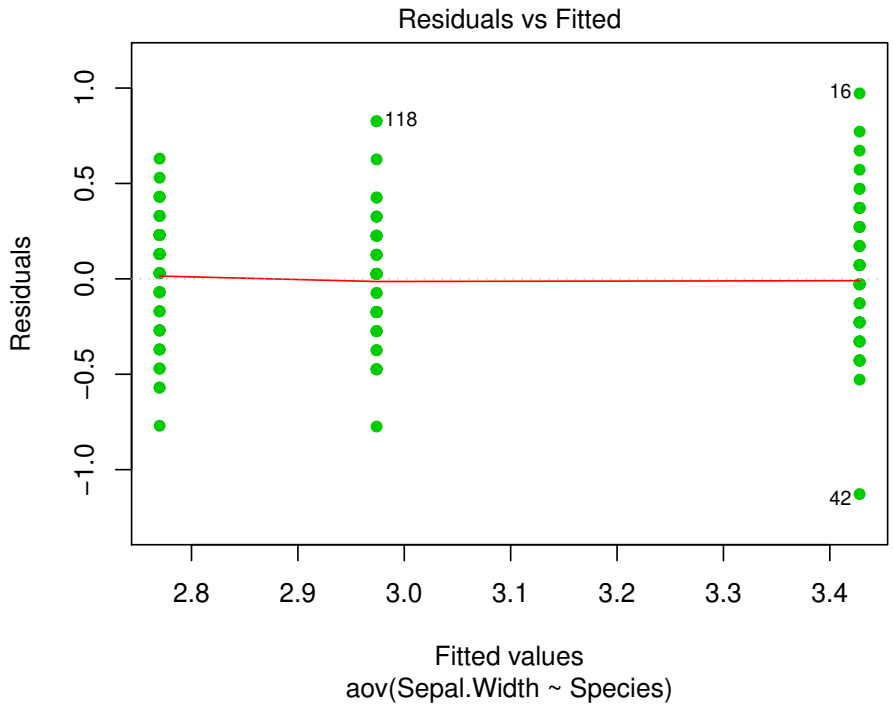
Numa ANOVA a um factor, os resíduos aparecem empilhados em k colunas nos gráficos de $\hat{y}_{ij}$ vs. e_{ij}, porque qualquer valor ajustado $\hat{y}_{ij} = \overline{y}_{i.}$ é igual para observações num mesmo nível do factor.

Este padrão não corresponde a qualquer violação dos pressupostos do modelo.

Analogamente, todas as observações dum mesmo nível do factor terão idêntico efeito alavanca, igual a $h_{ii} = \frac{1}{n_i}$. Sobretudo no caso de delineamentos equilibrados, isto torna os efeitos alavanca pouco úteis neste contexto.

Análise de Resíduos na ANOVA a 1 Factor (cont.)

Padrão de resíduos numa ANOVA a 1 Factor
(o exemplo considerado é `Sepal.Width` $\sim$ `Species`, nos lírios)



Violações aos pressupostos da ANOVA

Violações aos pressupostos do modelo não têm sempre igual gravidade. Alguns comentários gerais:

- O teste F da ANOVA (e as comparações múltiplas de Tukey) são relativamente robustos a desvios à hipótese de normalidade.
- As violações ao pressuposto de variâncias homogéneas são em geral menos graves no caso de delineamentos equilibrados, mas podem ser graves em delineamentos não equilibrados.
- A falta de independência entre erros aleatórios é a violação mais grave dos pressupostos e deve ser evitada, o que é em geral possível com um delineamento experimental adequado.

Uma advertência

Na formulação clássica do modelo ANOVA a um Factor, e a partir da equação-base

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} , \qquad \forall\, i,j$$

em vez de impor a condição $\alpha_1 = 0$, impõe-se a condição $\sum_i \alpha_i = 0$.

Esta condição alternativa:

- Muda a forma de interpretar os parâmetros (μ é agora uma espécie de média geral de Y e α_i o desvio da média do nível i em relação a essa média geral);
- Muda os estimadores dos parâmetros.
- Não muda o resultado do teste F à existência de efeitos do factor, nem a qualidade global do ajustamento.

A nossa abordagem (a restrição $\alpha_1 = 0$), além de generalizável a modelos com mais factores, permite aproveitar directamente os resultados do Modelo Linear estudados na RLM.

Delineamentos e Unidades experimentais

No delineamento das experiências para posterior análise através duma ANOVA (ou regressão linear), as n observações da variável resposta correspondem a n diferentes unidades experimentais (indivíduos, parcelas de terreno, locais, etc.).

Princípios gerais, já conhecidos, na selecção destas unidades experimentais são:

- casualização, ou seja aleatoriedade na escolha das unidades experimentais (por exemplo, na associação que lhes é feita de um dado nível do factor, caso seja controlável). É importante para se poder trabalhar com a Teoria de Probabilidades e também para evitar enviesamentos.
- repetição de observações (por exemplo, associadas a um mesmo nível do factor). É importante para se poder determinar a variabilidade associada à estimação (erros padrões).

Há outros conceitos importantes de delineamento.

Heterogeneidade nas unidades experimentais

Variabilidade nas unidades experimentais não atribuível aos preditores é considerada variação aleatória e contemplada nos erros aleatórios. Assim, heterogeneidade não controlada nas unidades experimentais contribui para aumentar o valor de *SQRE* e de *QMRE*.

Aumentar *QMRE* significa, no teste aos efeitos do factor, diminuir o valor calculado da estatística *F*, afastando-a da região crítica. Assim,

numa ANOVA

heterogeneidade não controlada nas unidades experimentais contribui para esconder a presença de eventuais efeitos do(s) factor(es).

numa Regressão Linear

heterogeneidade não controlada nas unidades experimentais contribui para piorar a qualidade de ajustamento do modelo, diminuindo o seu Coeficiente de Determinação.

Controlar a heterogeneidade

Na prática, é impossível tornar as unidades experimentais totalmente homogéneas: a natural variabilidade de plantas, animais, terrenos, localidades geográficas, células, etc. significa que existe variabilidade não controlável entre unidades experimentais.

Mesmo que seja possível ter unidades experimentais (quase) homogéneas, isso tem uma consequência que pode ser indesejável: restringir a validade dos resultados ao tipo de unidades experimentais com as características utilizadas na experiência.

Caso se saiba que existe um factor de variabilidade importante nas unidades experimentais, a melhor forma de controlar os seus efeitos consiste em contemplar a existência desse factor de variabilidade no delineamento e no modelo, de forma a filtrar os seus efeitos.

Um exemplo

Pretende-se analisar o rendimento de 5 diferentes variedades de trigo. Os rendimentos são também afectados pelos tipo de solos usados.

Nem sempre é possível ter terrenos homogéneos numa experiência. Mesmo que seja possível, pode não ser desejável, por se limitar a validade dos resultados a um único tipo de solos.

Admita-se que estamos interessados em quatro terrenos com diferentes tipos de solos. Cada terreno pode ser dividido em cinco parcelas viáveis para o trigo.

Em vez de repartir aleatoriamente as 5 variedades pelas 20 parcelas, é preferível forçar cada tipo de terreno a conter uma parcela com cada variedade. Apenas dentro dos terrenos haverá casualização.

Um exemplo (cont.)

A situação descrita no acetato anterior é a seguinte:

Terreno 1	Var.1	Var.3	Var.4	Var.5	Var.2
Terreno 2	Var.4	Var.3	Var.5	Var.1	Var.2
Terreno 3	Var.2	Var.4	Var.1	Var.3	Var.5
Terreno 4	Var.5	Var.2	Var.4	Var.1	Var.3

Houve uma restrição à casualização total: dentro de cada terreno há casualização, mas obriga-se cada terreno a ter uma parcela associada a cada nível do factor variedade.

Delineamentos factoriais a dois factores

O delineamento agora exemplificado é um caso particular de um delineamento factorial a dois factores (*two-way ANOVA*), sendo um dos factores a variedade de trigo e o outro o tipo de solos.

Um delineamento factorial é um delineamento em que há observações para todas as possíveis combinações de níveis de cada factor.

Assim, existência de mais do que um factor pode resultar de:

- pretender-se realmente estudar eventuais efeitos de mais do que um factor sobre a variável resposta;
- a tentativa de controlar a variabilidade experimental.

Historicamente, a segunda situação ficou associada à designação blocos, e na primeira fala-se apenas em factores. Mas são situações análogas.

Modelo ANOVA a 2 Factores (sem interacção)

Estudaremos dois diferentes modelos ANOVA para um delineamento factorial com 2 factores.

Admita-se a existência de:

- Uma variável resposta Y, da qual se efectuam n observações.
- Um Factor A, com a níveis.
- Um Factor B, com b níveis.

Um primeiro modelo prevê a existência de dois diferentes tipos de efeitos condicionando os valores de Y: os efeitos associados aos níveis de cada um dos factores.

Modelo ANOVA a 2 Factores (sem interacção)

Notação: Cada observação da variável resposta será agora identificada com três índices, Y_{ijk}, onde:

- i indica o nível i do Factor A.
- j indica o nível j do Factor B.
- k indica a repetição k na célula (i,j).

Célula: cruzamento dum nível dum Factor com um nível do outro Factor. Corresponde a uma dada situação experimental.

O número de observações na célula (i,j) é representado por n_{ij}.

Tem-se $\sum_{i=1}^{a} \sum_{j=1}^{b} n_{ij} = n$.

Se o número de observações fôr igual em todas as células ($n_{ij} = n_c\,, \quad \forall\, i,j$), estamos perante um delineamento equilibrado.

A modelação de Y

Neste primeiro modelo, admite-se que o valor esperado de cada observação é da forma:

$$E[Y_{ijk}] \quad = \quad \mu_{ij} \quad = \quad \mu + \alpha_i + \beta_j \quad , \quad \forall\, i,j,k.$$

O parâmetro μ é comum a todas as observações.

Cada parâmetro α_i funciona como um acréscimo que pode diferir entre níveis do Factor A, e é designado o efeito do nível i do factor A.

Cada parâmetro β_j funciona como um acréscimo que pode diferir entre níveis do Factor B, e é designado o efeito do nível j do factor B.

A variação de Y_{ijk} em torno do seu valor médio é representada por um erro aleatório aditivo, ε_{ijk}, de média zero:

$$Y_{ijk} \;=\; \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \;.$$

A equação-base em notação vectorial

A equação de base do modelo ANOVA a dois factores (sem interacção) também pode ser escrita na forma vectorial.

Seja

- $\vec{\mathbf{Y}}$ o vector aleatório n-dimensional com a totalidade das observações da variável resposta.
- $\vec{\mathbf{1}}_n$ o vector de n uns.
- $\vec{\mathscr{I}}_{A_i}$ a variável indicatriz de pertença ao nível i do Factor A.
- $\vec{\mathscr{I}}_{B_j}$ a variável indicatriz de pertença ao nível j do Factor B.
- $\vec{\boldsymbol{\varepsilon}}$ o vector aleatório dos n erros aleatórios.

A equação-base em notação vectorial: primeira tentativa

Se se admitem efeitos para todos os níveis de ambos os factores, temos a equação-base:

$$\vec{\mathbf{Y}} = \mu\vec{\mathbf{1}}_n + \alpha_1\vec{\mathscr{I}}_{A_1} + \alpha_2\vec{\mathscr{I}}_{A_2} + \ldots + \alpha_a\vec{\mathscr{I}}_{A_a} + \beta_1\vec{\mathscr{I}}_{B_1} + \beta_2\vec{\mathscr{I}}_{B_2} + \ldots + \beta_b\vec{\mathscr{I}}_{B_b} + \vec{\boldsymbol{\varepsilon}}$$

A matriz $\mathbf{X}$ definida com base neste modelo teria dependências lineares por duas diferentes razões:

- a soma das indicatrizes do Factor A daria a coluna dos uns, $\vec{\mathbf{1}}_n$;
- a soma das indicatrizes do Factor B daria a coluna dos uns, $\vec{\mathbf{1}}_n$.

A matriz $\mathbf{X}$ na primeiro tentativa

$$
\mathbf{X} = \left[\begin{array}{c|cccc|cccc}
1 & 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\
1 & 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\
1 & 1 & 0 & \dots & 0 & 0 & 1 & \dots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 1 & 0 & \dots & 0 & 0 & 0 & \dots & 1 \\
1 & 1 & 0 & \dots & 0 & 0 & 0 & \dots & 1 \\
\hline
1 & 0 & 1 & \dots & 0 & 1 & 0 & \dots & 0 \\
1 & 0 & 1 & \dots & 0 & 1 & 0 & \dots & 0 \\
1 & 0 & 1 & \dots & 0 & 1 & 0 & \dots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & 1 & \dots & 0 & 0 & 0 & \dots & 1 \\
1 & 0 & 1 & \dots & 0 & 0 & 0 & \dots & 1 \\
\hline
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\hline
1 & 0 & 0 & \dots & 1 & 1 & 0 & \dots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & 0 & \dots & 1 & 0 & 0 & \dots & 1 \\
1 & 0 & 0 & \dots & 1 & 0 & 0 & \dots & 1
\end{array}\right]
\begin{array}{c}
\\ \\
\uparrow \\ \vec{\mathbf{1}}_n
\end{array}
\begin{array}{cccc}
\\ \\
\uparrow & \uparrow & & \uparrow \\ \vec{\mathscr{I}}_{A_1} & \vec{\mathscr{I}}_{A_2} & \dots & \vec{\mathscr{I}}_{A_a}
\end{array}
\begin{array}{cccc}
\\ \\
\uparrow & \uparrow & & \uparrow \\ \vec{\mathscr{I}}_{B_1} & \vec{\mathscr{I}}_{B_2} & \dots & \vec{\mathscr{I}}_{B_b}
\end{array}
$$

Nem mesmo a exclusão da coluna $\vec{\mathbf{1}}_n$ resolve o problema.

Equação-base em notação vectorial: 2a. tentativa

Doravante, admitimos que foram excluídas do modelo as parcelas associadas ao primeiro nível de cada Factor, isto é:

$$\alpha_1 = 0 \qquad \text{e} \qquad \beta_1 = 0 \ ,$$

o que corresponde a excluir as colunas $\vec{\mathscr{I}}_{A_1}$ e $\vec{\mathscr{I}}_{B_1}$ da matriz $\mathbf{X}$.

A equação-base do modelo ANOVA a 2 Factores, sem interacção, fica:

$$\vec{\mathbf{Y}} = \mu \vec{\mathbf{1}}_n + \alpha_2 \vec{\mathscr{I}}_{A_2} + \ldots + \alpha_a \vec{\mathscr{I}}_{A_a} + \beta_2 \vec{\mathscr{I}}_{B_2} + \ldots + \beta_b \vec{\mathscr{I}}_{B_b} + \vec{\boldsymbol{\varepsilon}}$$

A matriz do delineamento na ANOVA a 2 Factores (sem interacção)

$$
\mathbf{X} = \left[\begin{array}{c|ccc|ccc}
1 & 0 & \dots & 0 & 0 & \dots & 0 \\
1 & 0 & \dots & 0 & 0 & \dots & 0 \\
1 & 0 & \dots & 0 & 1 & \dots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & \dots & 0 & 0 & \dots & 1 \\
1 & 0 & \dots & 0 & 0 & \dots & 1 \\
\hline
1 & 1 & \dots & 0 & 0 & \dots & 0 \\
1 & 1 & \dots & 0 & 0 & \dots & 0 \\
1 & 1 & \dots & 0 & 0 & \dots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
1 & 1 & \dots & 0 & 0 & \dots & 1 \\
1 & 1 & \dots & 0 & 0 & \dots & 1 \\
\hline
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\hline
1 & 0 & \dots & 1 & 0 & \dots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & \dots & 1 & 0 & \dots & 1 \\
1 & 0 & \dots & 1 & 0 & \dots & 1
\end{array}\right]
$$

$$
\begin{array}{ccccccc}
\uparrow & \uparrow & & \uparrow & \uparrow & & \uparrow \\
\vec{\mathbf{1}}_n & \vec{\mathscr{I}}_{A_2} & \dots & \vec{\mathscr{I}}_{A_a} & \vec{\mathscr{I}}_{B_2} & \dots & \vec{\mathscr{I}}_{B_b}
\end{array}
$$

O modelo ANOVA a dois factores, sem interacção

Juntando os pressupostos necessários à inferência,

Modelo ANOVA a dois factores, sem interacção

Existem n observações, Y_{ijk}, n_{ij} das quais associadas à célula (i,j) $(i=1,...,a;\ j=1,...,b)$. Tem-se:

1. $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \varepsilon_{ijk}$, $\forall\, i{=}1,...,a;\ j{=}1,...,b;\ k{=}1,...,n_{ij}$ $\quad (\alpha_1 = 0; \beta_1 = 0)$.
2. $\varepsilon_{ijk} \frown \mathcal{N}(0, \sigma^2), \quad \forall\, i,j,k$
3. $\{\varepsilon_{ijk}\}_{i,j,k}$ v.a.s independentes.

O modelo tem $a+b-1$ parâmetros desconhecidos:

- o parâmetro μ_{11};
- os $a-1$ acréscimos α_i $(i>1)$; e
- os $b-1$ acréscimos β_j $(j>1)$.

Testando a existência de efeitos

Um teste de ajustamento global do modelo tem como hipótese nula que todos os efeitos, quer do factor A, quer do factor B são simultaneamente nulos. Não distingue os efeitos de cada factor.

Mais útil é testar separadamente a existência de efeitos de cada factor:

- Teste I: $H_0 : \alpha_i = 0\,, \quad \forall i = 2, ..., a$;
- Teste II: $H_0 : \beta_j = 0\,, \quad \forall j = 2, ..., b$.

Teste aos efeitos do Factor B

O modelo do Acetato ANOVA a 2 Factores, sem interacção (Acetato 269) tem equação de base, em notação vectorial,

$$\vec{\mathbf{Y}} = \mu\vec{\mathbf{1}}_n + \alpha_2\vec{\mathscr{I}}_{A_2} + \ldots + \alpha_a\vec{\mathscr{I}}_{A_a} + \beta_2\vec{\mathscr{I}}_{B_2} + \ldots + \beta_b\vec{\mathscr{I}}_{B_b} + \vec{\boldsymbol{\varepsilon}}$$

O facto de ser um Modelo Linear permite aplicar a teoria já conhecida para este tipo de modelos, para testar as hipóteses

$$H_0 : \beta_j = 0\,, \quad \forall j = 2, \ldots, b \qquad \text{vs.} \qquad H_1 : \exists j \quad \text{tal que } \beta_j \neq 0\ .$$

Trata-se dum teste F parcial comparando o modelo completo

$$(\text{Modelo } M_{A+B}) \qquad Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \varepsilon_{ijk}\ ,$$

com o submodelo de equação de base

$$(\text{Modelo } M_A) \qquad Y_{ijk} = \mu_{11} + \alpha_i + \varepsilon_{ijk}\ ,$$

que é um modelo ANOVA a 1 Factor (factor A).

A construção do teste aos efeitos do Factor B

Pode-se:

- construir as matrizes $\mathbf{X}$ do delineamento para o modelo (M_{A+B}) e o submodelo (M_A).
- Obter os estimadores de parâmetros $\hat{\vec{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{Y}}$, para a matriz $\mathbf{X}$ correspondente a cada modelo.
- Obter as respectivas Somas de Quadrados Residuais, que designaremos $SQRE_{A+B}$ e $SQRE_A$.
- Efectuar o teste F parcial indicado, com a estatística de teste:

$$\text{(Efeitos Factor B)} \qquad F = \frac{\overbrace{SQRE_A - SQRE_{A+B}}^{=SQB}}{\frac{SQRE_{A+B}}{n-(a+b-1)}} = \frac{QMB}{QMRE}$$

definindo $QMB = \frac{SQB}{b-1} = \frac{SQRE_A - SQRE_{A+B}}{b-1}$

A construção do teste aos efeitos do Factor A

Consideremos também um teste aos efeitos do Factor A. Defina-se:

- $SQA = SQF_A$, a Soma de Quadrados do Factor no Modelo M_A;
- $QMA = \frac{SQA}{a-1}$, o Quadrado Médio do Factor no Modelo M_A;
- $SQRE_{A+B}$ e $QMRE = \frac{SQRE_{A+B}}{n-(a+b-1)}$, como antes.

É possível provar que a estatística

$$F = \frac{QMA}{QMRE} = \frac{\frac{SQA}{a-1}}{\frac{SQRE_{A+B}}{n-(a+b-1)}}$$

tem distribuição $F_{(a-1,n-(a+b-1))}$, caso $\alpha_i = 0$, para qualquer $i=2,\ldots,a$.

O Teste F aos efeitos do factor A

Sendo válido o Modelo de ANOVA a dois factores, sem interacção:

Teste F aos efeitos do factor A

Hipóteses: $H_0 : \alpha_i = 0 \quad \forall i=2,...,a$ vs. $H_1 : \exists i=2,..,a \text{ t.q. } \alpha_i \neq 0.$

[A NÃO AFECTA Y] vs. [A AFECTA Y]

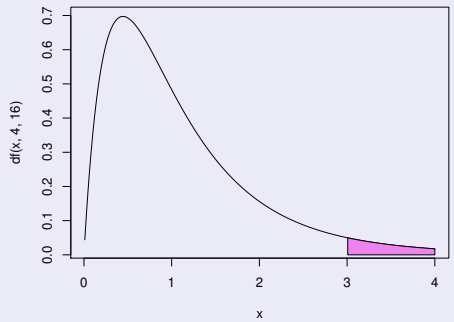
Estatística do Teste: $F = \frac{QMA}{QMRE} \frown F_{(a-1,n-(a+b-1))}$ se H_0.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se

$F_{calc} > f_{\alpha(a-1,n-(a+b-1))}$



O Teste F aos efeitos do factor B

Sendo válido o Modelo de ANOVA a dois factores, sem interacção:

Teste F aos efeitos do factor B

Hipóteses: $H_0 : \beta_j = 0 \quad \forall j=2,\ldots,b$ vs. $H_1 : \exists j=2,..,b \text{ t.q. } \beta_j \neq 0.$

[B NÃO AFECTA Y] vs. [B AFECTA Y]

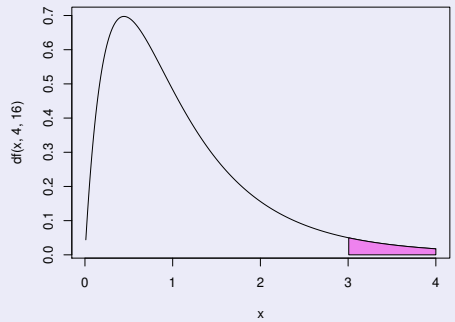
Estatística do Teste: $F = \frac{QMB}{QMRE} \frown F_{(b-1,\,n-(a+b-1))}$ se H_0.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se

$F_{calc} > f_{\alpha(b-1,n-(a+b-1))}$



A nova decomposição de SQT

Tendo em conta as Somas de Quadrados antes definidas, tem-se:

$$\begin{aligned} SQB &= SQRE_A - SQRE_{A+B} \\ SQA &= SQF_A = SQT - SQRE_A \end{aligned}$$

Somando estas SQs a $SQRE_{A+B}$, obtém-se:

$$SQRE_{A+B} + SQA + SQB = SQT$$

que é uma nova decomposição de SQT, em três parcelas, associadas ao facto de haver agora dois factores com efeitos previstos no modelo, mais a variabilidade residual.

Trocando a ordem dos factores

A troca do papel dos factores A e B levaria a definir as Somas de Quadrados de cada factor de forma diferente.

Designando por M_B o modelo ANOVA a um factor, mas apenas com o factor que temos chamado B, ter-se-ia agora:

$$\begin{aligned} SQB &= SQF_B = SQT - SQRE_B \\ SQA &= SQRE_B - SQRE_{A+B} \, . \end{aligned}$$

Continua a ser verdade que SQT se pode decompor na forma

$$SQT = SQA + SQB + SQRE_{A+B} \, .$$

Justificam-se testes análogos aos dos acetatos 274 e 275.
Mas as duas formas alternativas de definir SQA e SQB apenas produzem resultados iguais no caso de delineamentos equilibrados, pelo que só nesse caso a ordem dos factores é arbitrária.

SQA e *SQB* em delineamentos equilibrados

A expressão para SQA obtida no acetato 273 é a Soma de Quadrados do Factor (SQF_A) do Modelo M_A, apenas com o Factor A.

Nesse modelo, os valores ajustados são $\hat{Y}_{ijk} = \overline{Y}_{i..}$ (acetato 223), onde $\overline{Y}_{i...}$ indica a média de todas as observações de Y associadas ao nível i do factor A. Logo, e indicando por $\overline{Y}_{...}$ a média global das n observações de Y, tem-se:

$$SQF_A = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_c}(\hat{Y}_{ijk} - \overline{Y}_{...})^2 = b\,n_c \cdot \sum_{i=1}^{a}(\overline{Y}_{i..} - \overline{Y}_{...})^2 = SQA\,.$$

Da mesma forma, num delineamento equilibrado, *SQB* é a Soma de Quadrados do Factor (SQF_B) do Modelo M_B, apenas com o Factor B:

Nesse modelo, os valores ajustados são $\hat{Y}_{ijk} = \overline{Y}_{.j.}$ (acetato 223), logo:

$$SQF_B = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_c}(\hat{Y}_{ijk} - \overline{Y}_{...})^2 = a\,n_c \cdot \sum_{j=1}^{b}(\overline{Y}_{.j.} - \overline{Y}_{...})^2 = SQB\,.$$

O quadro-resumo da ANOVA a 2 Factores (sem interacção; delineamento equilibrado)

Fonte	g.l.	SQ	QM	f_{calc}
Factor A	$a-1$	$SQA = b\,n_c \cdot \sum_{i=1}^{a} (\overline{y}_{i..} - \overline{y}_{...})^2$	$QMA = \frac{SQA}{a-1}$	$\frac{QMA}{QMRE}$
Factor B	$b-1$	$SQB = a\,n_c \cdot \sum_{j=1}^{b} (\overline{y}_{.j.} - \overline{y}_{...})^2$	$QMB = \frac{SQB}{b-1}$	$\frac{QMB}{QMRE}$
Resíduos	$n-(a+b-1)$	$SQRE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_c} (y_{ijk} - \hat{y}_{ijk})^2$	$QMRE = \frac{SQRE}{n-(a+b-1)}$	
Total	$n-1$	$SQT = (n-1)\,s_y^2$	–	–

ANOVA a dois Factores, sem interacção no R

Para efectuar uma ANOVA a dois Factores (sem interacção) no R, convém organizar os dados numa `data.frame` com três colunas:

1. uma para os valores (numéricos) da variável resposta;
2. outra para o factor A (com a indicação dos seus níveis);
3. outra para o factor B (com a indicação dos seus níveis).

As fórmulas utilizadas no R para indicar uma ANOVA a dois Factores, sem interacção, são semelhantes às usadas na Regressão Linear com dois preditores, devendo o nome dos dois factores ser separado pelo símbolo $+$:

```
y  ~  fA + fB
```

Um exemplo

O rendimento de cinco variedades de aveia (*manchuria, svansota,velvet, trebi* e *peatland*) foi registado em seis diferentes localidades [2]. Em cada localidade foi semeada uma e uma só parcela com cada variedade (havendo casualização em cada localidade).

```
> summary(aov(Y1 ~ Var + Loc, data=immer))
            Df  Sum Sq Mean Sq F value    Pr(>F)
Var          4  2756.6   689.2  4.2309   0.01214 *
Loc          5 17829.8  3566.0 21.8923 1.751e-07 ***
Residuals   20  3257.7   162.9
```

Há alguma indicação de efeitos significativos entre variedades, e muita entre localidades. E num modelo sem efeito de localidades (blocos)?

```
> summary(aov(Y1 ~ Var, data=immer))
            Df  Sum Sq Mean Sq F value Pr(>F)
Var          4  2756.6   689.2   0.817 0.5264
Residuals   25 21087.6   843.5
```

[2] Dados em Immer, Hayes e LeRoy Powers, Statistical adaptation of barley varietal adaptation, Journal of the American Society for Agronomy, 26, 403-419, 1934.

Significado dos parâmetros

- Se $i = j = 1$: $\vec{\mathbf{Y}}_{11k} = \mu + \vec{\boldsymbol{\varepsilon}}_{11k} \Rightarrow E[\vec{\mathbf{Y}}_{11k}] = \mu_{11}$.
- Se $i > 1$ e $j = 1$: $\vec{\mathbf{Y}}_{i1k} = \mu + \alpha_i + \vec{\boldsymbol{\varepsilon}}_{i1k} \Rightarrow E[\vec{\mathbf{Y}}_{i1k}] = \mu_{11} + \alpha_i$.
- Se $i = 1$ e $j > 1$: $\vec{\mathbf{Y}}_{1jk} = \mu + \beta_j + \vec{\boldsymbol{\varepsilon}}_{1jk} \Rightarrow E[\vec{\mathbf{Y}}_{1jk}] = \mu_{11} + \beta_j$.
- Se $i > 1$ e $j > 1$: $\vec{\mathbf{Y}}_{ijk} = \mu + \alpha_i + \beta_j + \vec{\boldsymbol{\varepsilon}}_{ijk} \Rightarrow E[\vec{\mathbf{Y}}_{ijk}] = \mu_{11} + \alpha_i + \beta_j$.

Há $a + b - 1$ parâmetros:

- uma média da primeira célula, μ_{11};
- $a - 1$ efeitos de nível do factor A, α_i ($i > 1$);
- $b - 1$ efeitos de nível do factor B, β_j ($j > 1$).

Mas há ab situações experimentais (as ab células resultantes de cruzar cada nível de cada factor). Este modelo é demasiado rígido: os valores esperados nas células $i > 1$ e $j > 1$ não são livres, devido à falta de parâmetros.

Fórmulas para delineamentos equilibrados

Sejam:

$\overline{Y}_{i..}$ a média amostral das $b\,n_c$ observações do nível i do Factor A, $\overline{Y}_{i..} = \frac{1}{b\,n_c} \sum_{j=1}^{b} \sum_{k=1}^{n_c} Y_{ijk}$

$\overline{Y}_{.j.}$ a média amostral das $a\,n_c$ observações do nível j do Factor B, $\overline{Y}_{.j.} = \frac{1}{a\,n_c} \sum_{i=1}^{a} \sum_{k=1}^{n_c} Y_{ijk}$

$\overline{Y}_{...}$ a média amostral da totalidade das $n = a\,b\,n_c$ observações, $\overline{Y}_{...} = \frac{1}{n} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_c} Y_{ijk}$.

Se o delineamento é equilibrado, ou seja, $n_{ij} = n_c \ , \ \forall\, i,j \ ,$ tem-se:

- $\hat{\mu}_{11} = \overline{Y}_{1..} + \overline{Y}_{.1.} - \overline{Y}_{...}$
- $\hat{\alpha}_i = \overline{Y}_{i..} - \overline{Y}_{1..}$
- $\hat{\beta}_j = \overline{Y}_{.j.} - \overline{Y}_{.1.}$

Fórmulas para delineamentos equilibrados (cont.)

Tendo em conta estas fórmulas e a equação base do Modelo, tem-se que os valores ajustados de cada observação dependem apenas das médias dos respectivos níveis em cada factor e da média geral de todas as observações:

$$\hat{Y}_{ijk} \quad = \quad \hat{\mu}_{11} + \hat{\alpha}_i + \hat{\beta}_j \quad = \quad \overline{Y}_{i..} + \overline{Y}_{.j.} - \overline{Y}_{...} \qquad , \quad \forall\, i,j,k$$

Aviso: Ao contrário do que sucede na ANOVA a um factor, os valores ajustados $\hat{Y}_{ijk}$ não são a média das observações de Y na célula (i,j).

Modelos com interacção

Um modelo ANOVA a 2 Factores, sem interacção, foi considerado para um delineamento factorial, isto é, em que se cruzam todos os níveis de um e outro factor. Mas trata-se dum modelo pouco flexível.

Um modelo sem efeitos de interacção é utilizado sobretudo quando existe uma única observação em cada célula, i.e., $n_{ij} = 1, \forall\ i, j$.

Na presença de repetições nas células, a forma mais natural de modelar um delineamento com dois factores é a de prever a existência de um terceiro tipo de efeitos: os efeitos de interacção.

A ideia é incorporar na equação base do modelo para Y_{ijk} uma parcela $(\alpha\beta)_{ij}$ que permita que em cada célula haja um efeito específico associado à combinação dos níveis i do Factor A e j do Factor B:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \ .$$

Os valores esperados de Y_{ijk} (modelo com interacção)

Vamos admitir as seguintes restrições aos parâmetros:

$$\alpha_1 = 0 \quad ; \quad \beta_1 = 0 \quad ; \quad (\alpha\beta)_{1j} = 0 \ , \ \forall j \quad ; \quad (\alpha\beta)_{i1} = 0 \ , \ \forall i.$$

Tem-se:

- Para a primeira célula ($i = j = 1$): $\mu_{11} = E[Y_{11k}] = \mu$.
- Nas restantes células $(1, j)$ do primeiro nível do Factor A: $\mu_{1j} = E[Y_{1jk}] = \mu_{11} + \beta_j$.
- Nas restantes células $(i, 1)$ do primeiro nível do Factor B: $\mu_{i1} = E[Y_{i1k}] = \mu_{11} + \alpha_i$.
- Nas células genéricas (i, j), com $i > 1$ e $j > 1$, $\mu_{ij} = E[Y_{ijk}] = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij}$.

Os efeitos α_i e β_j designam-se efeitos principais de cada Factor.

Variáveis indicatrizes de célula

A versão vectorial do modelo com interacção associa os novos efeitos $(\alpha\beta)_{ij}$ a variáveis indicatrizes de cada célula, excluíndo as células associadas ao primeiro nível de qualquer dos factores.
A equação-base do modelo ANOVA a 2 Factores, com interacção, é:

$$\begin{aligned}\vec{\mathbf{Y}} &= \mu\vec{\mathbf{1}}_n + \alpha_2\vec{\mathscr{I}}_{A_2} + \ldots + \alpha_a\vec{\mathscr{I}}_{A_a} + \beta_2\vec{\mathscr{I}}_{B_2} + \ldots + \beta_b\vec{\mathscr{I}}_{B_b} + \\ &\quad + (\alpha\beta)_{22}\vec{\mathscr{I}}_{A_2:B_2} + (\alpha\beta)_{23}\vec{\mathscr{I}}_{A_2:B_3} + \ldots + (\alpha\beta)_{ab}\vec{\mathscr{I}}_{A_a:B_b} + \vec{\boldsymbol{\varepsilon}}\end{aligned}$$

onde $\vec{\mathscr{I}}_{A_i:B_j}$ representa a variável indicatriz da célula correspondente ao nível *i* do Factor A e nível *j* do factor B.

Existem neste modelo *ab* parâmetros.

Modelo ANOVA a 2 factores, com interacção (cont.)

O ajustamento deste modelo faz-se de forma análoga ao ajustamento de modelos anteriores.

A matriz **X** do delineamento é agora constituída por ab colunas:

- uma coluna de uns, $\vec{\mathbf{1}}_n$, associada ao parâmetro μ_{11}.
- $a-1$ colunas de indicatrizes de nível do factor A, $\vec{\mathscr{I}}_{A_i}$, $(i>1)$, associadas aos parâmetros α_i.
- $b-1$ colunas de indicatrizes de nível do factor B, $\vec{\mathscr{I}}_{B_j}$, $(j>1)$, associadas aos parâmetros β_j.
- $(a-1)(b-1)$ colunas de indicatrizes de célula, $\vec{\mathscr{I}}_{A_i:B_j}$, $(i,j>1)$, associadas aos efeitos de interacção $(\alpha\beta)_{ij}$.

Como em modelos anteriores, $\vec{\hat{\mathbf{Y}}} = \mathbf{H}\vec{\mathbf{Y}}$, sendo **H** a matriz que projecta ortogonalmente sobre o espaço $\mathscr{C}(\mathbf{X})$ gerado pelas colunas desta matriz **X**. E também, $SQRE_{A*B} = \|\vec{\mathbf{Y}} - \vec{\hat{\mathbf{Y}}}\|^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(Y_{ijk} - \hat{Y}_{ijk})^2$.

Os três testes ANOVA

Neste delineamento, desejamos fazer um teste à existência de cada um dos três tipos de efeitos:

- $H_0 : (\alpha\beta)_{ij} = 0\,, \quad \forall\, i = 2, ..., a\,,\ \forall j = 2, ..., b$;
- $H_0 : \alpha_i = 0\,, \quad \forall i = 2, ..., a$; e
- $H_0 : \beta_j = 0\,, \quad \forall j = 2, ..., b$.

As estatísticas de teste para cada um destes testes obtêm-se a partir da decomposição da Soma de Quadrados Total em parcelas convenientes.

O modelo ANOVA a dois factores, com interacção

Juntando os pressupostos necessários à inferência,

Modelo ANOVA a dois factores, com interacção (Modelo M_{A*B})

Existem n observações, Y_{ijk}, n_{ij} das quais associadas à célula (i,j) $(i=1,...,a;\ j=1,...,b)$. Tem-se:

1. $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$, $\quad \forall i{=}1,...,a\ ;\ j{=}1,...,b\ ;\ k{=}1,...,n_{ij}$
   $(\alpha_1{=}0\ ;\ \beta_1{=}0\ ;\ (\alpha\beta)_{1j}{=}0\ ,\ \forall j;\ (\alpha\beta)_{i1}{=}0\ ,\ \forall i)$.
2. $\varepsilon_{ijk} \frown \mathcal{N}(0, \sigma^2)$
3. $\{\varepsilon_{ijk}\}_{i,j,k}$ v.a.s independentes.

O modelo tem *ab* parâmetros desconhecidos:

- μ_{11};
- os $a-1$ acréscimos α_i $(i > 1)$;
- os $b-1$ acréscimos β_j; e
- os $(a-1)(b-1)$ efeitos de interacção $(\alpha\beta)_{ij}$, para $i > 1$, $j > 1$.

Testando efeitos de interacção

Para testar a existência de efeitos de interacção,

$$H_0 : (\alpha\beta)_{ij} = 0\,, \quad \forall\, i = 2,...,a\ ,\ \forall j = 2,...,b\ ,$$

pode efectuar-se um teste F parcial comparando o modelo

$$(\text{Modelo } M_{A*B}) \qquad Y_{ijk} \;=\; \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}\ ,$$

com o submodelo

$$(\text{Modelo } M_{A+B}) \qquad Y_{ijk} \;=\; \mu_{11} + \alpha_i + \beta_j + \varepsilon_{ijk}\ ,$$

Designa-se Soma de Quadrados associada à interacção à diferença

$$SQAB \;=\; SQRE_{A+B} - SQRE_{A*B}$$

Testando os efeitos principais de cada Factor

Para testar os efeitos principais do Factor B,
$H_0 : \beta_j = 0\,, \quad \forall j = 2, ..., b$, pode partir-se dos modelos

$$\begin{aligned} (\text{Modelo } M_{A+B}) \qquad Y_{ijk} &= \mu_{11} + \alpha_i + \beta_j + \varepsilon_{ijk} \\ (\text{Modelo } M_A) \qquad Y_{ijk} &= \mu_{11} + \alpha_i + \varepsilon_{ijk}\ , \end{aligned}$$

e tomar (como no modelo sem efeitos de interacção):

$$\begin{aligned} SQB &= SQRE_A - SQRE_{A+B} \\ SQA &= SQF_A = SQT - SQRE_A \end{aligned}$$

Nota: Estas duas Somas de Quadrados definem-se de forma idêntica à usada no modelo sem efeitos de interacção.

A decomposição de SQT

Definimos :

$$\begin{aligned} SQAB &= SQRE_{A+B} - SQRE_{A*B} \\ SQB &= SQRE_{A} - SQRE_{A+B} \\ SQA &= SQF_A = SQT - SQRE_A \end{aligned}$$

Somando estas Somas de Quadrados a $SQRE_{A*B}$, obtém-se:

$$SQRE_{A*B} + SQAB + SQA + SQB = SQT$$

Esta decomposição de SQT gera as quantidades nas quais se baseiam as estatísticas dos três testes associados ao Modelo M_{A*B}.

O quadro-resumo

Com base na decomposição do acetato 293 podemos construir o quadro resumo da ANOVA a 2 Factores, com interacção.

Fonte	g.l.	SQ	QM	f_{calc}
Factor A	$a-1$	SQA	$QMA = \frac{SQA}{a-1}$	$\frac{QMA}{QMRE}$
Factor B	$b-1$	SQB	$QMB = \frac{SQB}{b-1}$	$\frac{QMB}{QMRE}$
Interacção	$(a-1)(b-1)$	$SQAB$	$QMAB = \frac{SQAB}{(a-1)(b-1)}$	$\frac{QMAB}{QMRE}$
Resíduos	$n-ab$	$SQRE$	$QMRE = \frac{SQRE}{n-ab}$	
Total	$n-1$	$SQT = (n-1)\,s_y^2$	–	–

Os g.l. de cada tipo de efeitos correspondem ao número de parâmetros desse tipo que sobram após a imposição das restrições. Como em qualquer modelo linear, os g.l. residuais são o número de observações (n) menos o número de parâmetros do modelo (ab).

O Teste F aos efeitos de interacção

Sendo válido o Modelo ANOVA a dois factores, com interacção:

Teste F aos efeitos de interacção

Hipóteses: $H_0 : (\alpha\beta)_{ij} = 0 \quad \forall i,j$ vs. $H_1 : \exists i,j \text{ t.q. } (\alpha\beta)_{ij} \neq 0$.

[NÃO HÁ INTERACÇÃO] vs. [HÁ INTERACÇÃO]

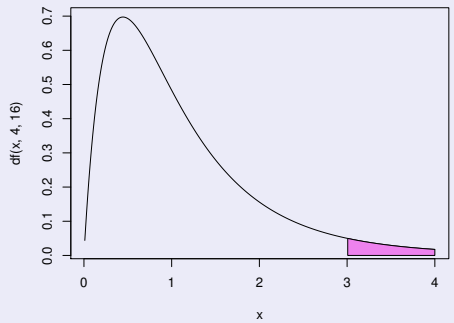
Estatística do Teste: $F = \frac{QMAB}{QMRE} \frown F_{((a-1)(b-1),n-ab)}$ se H_0.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se

$F_{calc} > f_{\alpha((a-1)(b-1),n-ab)}$



O Teste F aos efeitos principais do factor A

Sendo válido o Modelo ANOVA a 2 factores com interacção tem-se:

Teste F aos efeitos principais do factor A

Hipóteses: $H_0 : \alpha_i = 0 \quad \forall i=2,...,a$ vs. $H_1 : \exists i=2,..,a \text{ t.q. } \alpha_i \neq 0.$

[$\nexists$ EFEITOS DE A] vs. [$\exists$ EFEITOS DE A]

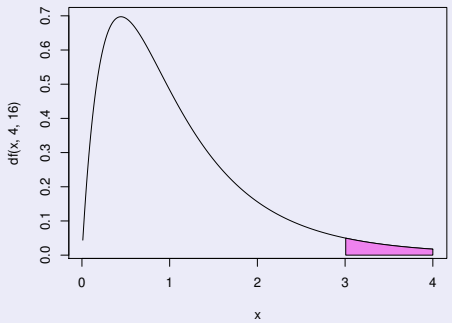
Estatística do Teste: $F = \frac{QMA}{QMRE} \frown F_{(a-1,n-ab)}$ se H_0.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se

$F_{calc} > f_{\alpha(a-1,n-ab)}$



O Teste F aos efeitos principais do factor B

Sendo válido o Modelo ANOVA a 2 factores com interacção tem-se:

Teste F aos efeitos principais do factor B

Hipóteses: $H_0 : \beta_j = 0 \quad \forall j=2,\ldots,b$ vs. $H_1 : \exists j=2,..,b \text{ t.q. } \beta_j \neq 0.$

[$\nexists$ EFEITOS DE B] vs. [$\exists$ EFEITOS DE B]

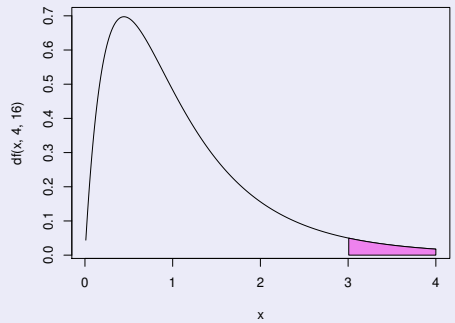
Estatística do Teste: $F = \frac{QMB}{QMRE} \frown F_{(b-1,n-ab)}$ se H_0.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se

$F_{calc} > f_{\alpha(b-1,n-ab)}$



ANOVA a dois Factores, com interacção no R

Para efectuar uma ANOVA a dois Factores, com interacção, no R, organizam-se os dados de forma igual à usada para o modelo sem interacção: uma `data.frame` com três colunas:

1. uma para a variável resposta;
2. outra para o factor A;
3. outra para o factor B.

As fórmulas utilizadas no R para indicar uma ANOVA a dois Factores, com interacção, recorrem ao símbolo `*`:

```
y  ~  fA * fB
```

sendo `y` o nome da variável resposta e `fA` e `fB` os nomes dos factores.

Estimação da interacção necessita de repetições

Para se poder estudar efeitos de interacção, é necessário que haja repetições nas células.

Os graus de liberdade do *SQRE* neste modelo são $n - ab$. Se houver uma única observação em cada célula, tem-se $n = ab$, ou seja, tantos parâmetros quantas as observações existentes. Nesse caso, nem sequer será possível definir o Quadrado Médio Residual, *QMRE*.

Num delineamento com uma única observação por célula é obrigatório optar por um modelo sem interacção. Havendo repetições, é mais natural considerar um modelo com interacção.

Valores ajustados de Y no modelo com interacção

Sejam

- $\overline{Y}_{ij.}$ a média amostral das n_{ij} observações da célula (i,j),
- $\overline{Y}_{i..}$ a média amostral das $\sum_j n_{ij}$ observações do nível i do Factor A,
- $\overline{Y}_{.j.}$ a média amostral das $\sum_i n_{ij}$ observações do nível j do Factor B,
- $\overline{Y}_{...}$ a média amostral da totalidade das $n = \sum_i \sum_j n_{ij}$ observações.

Os valores ajustados $\hat{Y}_{ijk}$ são iguais para todas as observações numa mesma célula, e são dados pela média amostral da célula:

$$\hat{Y}_{ijk} = \overline{Y}_{ij.} .$$

Estimadores de parâmetros

Os estimadores dos parâmetros num modelo ANOVA a 2 Factores, com interacção, são:

- $\hat{\mu}_{11} = \overline{Y}_{11\cdot}$
- $\hat{\alpha}_i = \overline{Y}_{i1\cdot} - \overline{Y}_{11\cdot} \qquad (i > 1)$
- $\hat{\beta}_j = \overline{Y}_{1j\cdot} - \overline{Y}_{11\cdot} \qquad (j > 1)$
- $\widehat{(\alpha\beta)}_{ij} = (\overline{Y}_{ij\cdot} + \overline{Y}_{11\cdot}) - (\overline{Y}_{i1\cdot} + \overline{Y}_{1j\cdot}) \qquad (i,j > 1).$

Intervalos de confiança ou testes de hipóteses para qualquer dos parâmetros individuais, ou combinações lineares desses parâmetros, podem ser efectuados utilizando a teoria geral do Modelo Linear.

Soma de Quadrados Residual

Como os valores ajustados correspondem às medias amostrais da célula onde se efectuaram as observações, $\hat{Y}_{ijk} = \overline{Y}_{ij.}$, tem-se:

$$SQRE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} (Y_{ijk} - \hat{Y}_{ijk})^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} (Y_{ijk} - \overline{Y}_{ij.})^2$$

$$\Leftrightarrow \quad SQRE = \sum_{i=1}^{a}\sum_{j=1}^{b} (n_{ij} - 1) S_{ij}^2 \ ,$$

sendo S_{ij}^2 a variância amostral das observações da célula (i, j).

Num delineamento equilibrado, tem-se $n = n_c ab$, e o Quadrado Médio Residual será a média simples das variâncias amostrais de célula, S_{ij}^2:

$$QMRE = \frac{SQRE}{n - ab} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b} S_{ij}^2 \ .$$

Outras SQs para delineamentos equilibrados

Para delineamentos equilibrados (com n_c observações por célula) é possível obter igualmente fórmulas simples para as Somas de Quadrados associadas aos efeitos principais de cada factor.

Estas fórmulas correspondem (tal como no modelo sem efeitos de interacção) às Somas de Quadrados associadas a cada factor, caso se ajustasse (aos mesmos dados) um modelo ANOVA apenas com esse factor:

$$SQA = bn_c \sum_{i=1}^{a} (\overline{Y}_{i..} - \overline{Y}_{...})^2$$

$$SQB = an_c \sum_{j=1}^{b} (\overline{Y}_{.j.} - \overline{Y}_{...})^2$$

Uma advertência

Na formulação clássica do modelo ANOVA a dois Factores, com interacção, e a partir da equação-base $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$, em vez de impor as condições $\alpha_1 = \beta_1 = (\alpha\beta)_{i1} = (\alpha\beta)_{1j} = 0$ $(\forall\, i, j)$, admite-se a existência de acréscimos de todos os tipos para qualquer valor de i e j e impõe-se as condições:

- $\sum_i \alpha_i = 0$;
- $\sum_j \beta_j = 0$;
- $\sum_i (\alpha\beta)_{ij} = 0\ , \qquad \forall\, j$;
- $\sum_j (\alpha\beta)_{ij} = 0\ , \qquad \forall\, i$.

Estas condições alternativas:

- mudam a forma de interpretar os parâmetros;
- mudam os estimadores dos parâmetros;
- não mudam o resultado dos testes F à existência de efeitos.

Material complementar: Comparações múltiplas de médias de células

O número potencialmente grande de comparações possíveis entre médias de célula aconselha a utilização de métodos de comparação múltipla, que permitam controlar globalmente o nível de significância do conjunto de testes de hipóteses (ou grau de confiança do conjunto de intervalos de confiança).

O mais utilizado dos métodos de comparação múltipla está associado ao nome de Tukey. Foi já introduzido no material complementar ao estudo de delineamentos a 1 Factor. Adapta-se facilmente à comparação múltipla de médias de células.

MC: O Teste de Tukey

Teste de Tukey para médias de células

Admite-se que o delineamento é equilibrado, com $n_c > 1$ repetiçoes em todas as ab células.

Rejeita-se a igualdade das médias das células (i,j) e (i',j'), a favor da hipótese $\mu_{ij} \neq \mu_{i'j'}$, se

$$|\overline{Y}_{ij\cdot} - \overline{Y}_{i'j'\cdot}| \quad > \quad q_{\alpha(ab,n-ab)} \cdot \sqrt{\frac{QMRE}{n_c}} \ ,$$

sendo $q_{\alpha(ab,n-ab)}$ o valor que deixa à direita uma região de probabilidade α numa distribuição de Tukey com parâmetros $k = ab$ (o número total de médias de célula) e $\nu = n - ab$ (os graus de liberdade associados ao $QMRE$).

MC: Intervalos de Confiança para $\mu_{ij} - \mu_{i'j'}$

Com grau de confiança global $(1-\alpha)\times 100\%$, todas as diferenças de médias de pares de células, $\mu_{ij} - \mu_{i'j'}$, estão em intervalos da forma:

$$\left] \left(\overline{y}_{ij\cdot} - \overline{y}_{i'j'\cdot}\right) - q_{\alpha(ab,n-ab)}\sqrt{\frac{QMRE}{n_c}} \quad , \quad \left(\overline{y}_{ij\cdot} - \overline{y}_{i'j'\cdot}\right) + q_{\alpha(ab,n-ab)}\sqrt{\frac{QMRE}{n_c}} \right[$$

Conclui-se que $\mu_{ij} \neq \mu_{i'j'}$ se o intervalo correspondente a este par de células não contém o valor zero.

MC: Tukey no R

A obtenção dos Intervalos de Confiança de Tukey no R, para a diferença da média de células, no caso de um delineamento a dois Factores, é análogo ao caso de um único factor:

```
> TukeyHSD(aov(y ~ fA * fB, data=dados))
```

O R produz também intervalos de confiança para as médias de nível de cada Factor isoladamente.

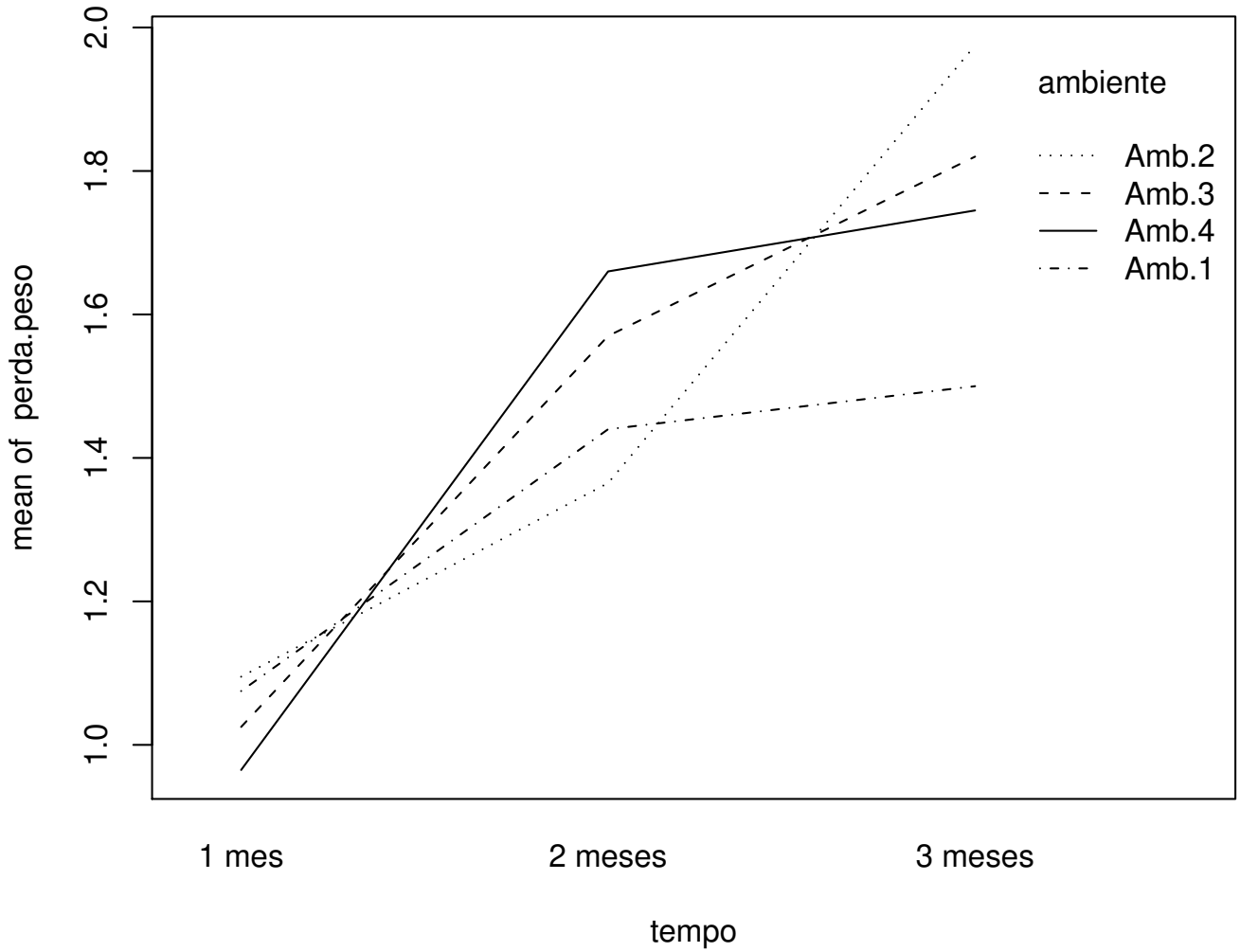
É possível representar graficamente estes Intervalos de Confiança encaixando o comando anterior na função `plot`.

MC: Visualização gráfica de efeitos de interacção

A existência de efeitos de interacção transparece em gráficos onde:

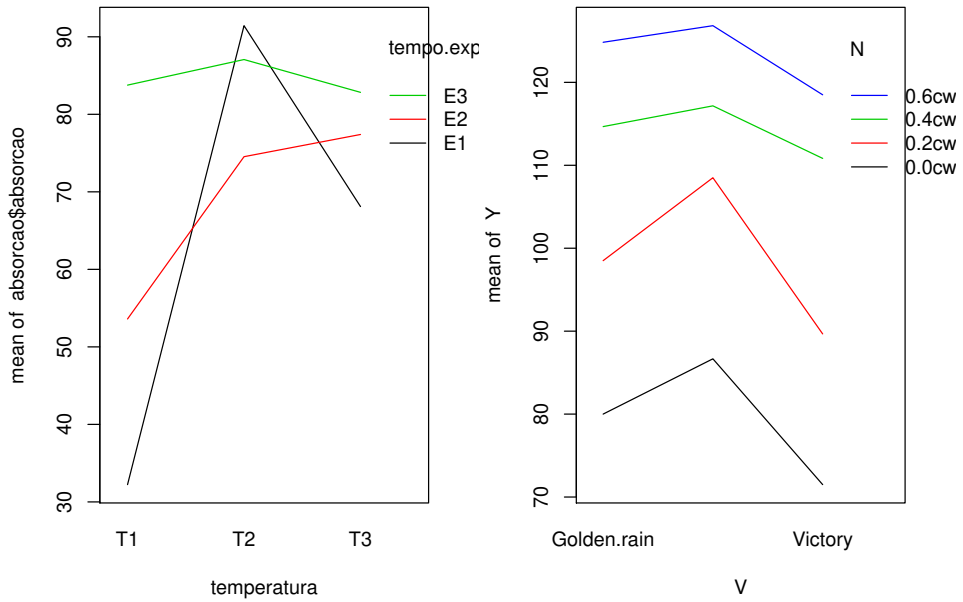
- O eixo horizontal é associado aos níveis de um factor (e.g., fA);
- no eixo vertical serão indicados os valores médios da variável resposta Y em cada célula;
- para cada célula, indica-se um ponto cujas coordenadas são determinadas pelo nível do primeiro factor e respectiva média de célula da variável resposta;
- unem-se com segmentos de recta os pontos correspondentes a um mesmo nível do segundo factor (e.g., fB).

MC: Exemplo


mean of perda.peso
2.0
1.8
1.6
1.4
1.2
1.0
1 mes
2 meses
3 meses
tempo
ambiente
Amb.2
Amb.3
Amb.4
Amb.1

MC: Como ler os gráficos de interacção

A inexistência de interacção significativa produz linhas aproximadamente "paralelas" (ver exemplo da direita).
Havendo interacção, as linhas estarão longe de qualquer paralelismo (ver exemplo da esquerda).



A confirmação da significância dos efeitos de interacção exige que se efectue o respectivo teste F.

Comentários finais sobre ANOVA

1. Delineamentos factoriais com vários factores

Um delineamento factorial (isto é, com observações para todas as combinações de níveis de cada factor) pode ser definido com qualquer número de factores.

Num delineamento factorial a três factores – A, B e C – cada observação da variável resposta indexa-se com quatro índices: Y_{ijkl} indica a observação l no nível i do Factor A, nível j do Factor B e nível k do Factor C. A equação de base para Y_{ijkl} prevê a existência de sete tipos de efeitos:

- três efeitos principais de cada factor, α_i, β_j e γ_k.
- três efeitos de interacção dupla associados a cada combinação de níveis de dois Factores diferentes: $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$ e $(\beta\gamma)_{jk}$.
- um efeito de tripla interacção para as células onde se cruzam níveis dos três factores: $(\alpha\beta\gamma)_{ijk}$

1. O modelo factorial a três factores

A equação de base do modelo é agora da forma:

$$Y_{ijkl} = \mu_{111} + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkl} \ ,$$

excluíndo-se efeitos sempre que um dos índices fôr 1.
O modelo tem *abc* parâmetros.

A Soma de Quadrados Total vai ser agora decomposta em oito parcelas: *SQA*, *SQB*, *SQC*, *SQAB*, *SQAC*, *SQBC*, *SQABC* e *SQRE*.
As sete *SQ*s associadas a efeitos são definidas pela diferença das Somas de Quadrados Residuais de modelos onde se vão sucessivamente omitindo os efeitos correspondentes.

1. O modelo factorial a três factores (cont.)

Os graus de liberdade associados a cada tipo de efeito generalizam conceitos anteriores:

- Para as *SQ*s de efeitos principais de factor, são os números de níveis, menos um: $a-1$, $b-1$ e $c-1$.
- para as interacções duplas, são o produto dos graus de liberdade de cada factor: $(a-1)(b-1)$, $(a-1)(c-1)$ e $(b-1)(c-1)$.
- para as interacções triplas, são o produto dos graus de liberdade dos três efeitos principais: $(a-1)(b-1)(c-1)$.
- para o residual, o número de observações menos o número de parâmetros, $n-abc$.

Haverá sete testes: um para cada tipo de efeitos. As estatísticas desses sete testes são todas do tipo $\frac{QMx}{QMRE}$, onde x designa o tipo de efeitos em questão. As estatísticas desses testes terão, sob H_0, distribuição F com graus de liberdade dados pelos g.l. do numerador e do denominador, respectivamente.

2. Outros tipos de delineamentos experimentais

Existem numerosos outros delineamentos mais complexos.

Nos delineamentos hierarquizados, os níveis dum factor (dito subordinado) dependem dos níveis considerados para outro factor (dito dominante). Exemplo: pretende-se saber se, em média, uma variável resposta varia entre espécies (um factor) e, dentro das espécies, entre variedades (outro factor). O delineamento não é factorial, mas sim hierarquizado. Variedade é o factor dominante e espécie o factor subordinado.

Alguns delineamentos visam reduzir o número de situações experimentais que é necessário estudar. Exemplo: quadrados latinos.

Outros delineamentos visam ultrapassar dificuldades práticas na execução de uma experiência, como é o caso dos delineamentos em parcelas divididas (*split plots*).

3. Métodos não paramétricos de tipo ANOVA

Uma forma alternativa de estudar problemas análogos aos objectivos de ANOVAs resulta da utilização de métodos não paramétricos.

Métodos não paramétricos são métodos em que não se exigem hipóteses tão fortes como os métodos clássicos, (e.g., a hipótese de normalidade). A sua maior generalidade tem como contrapartida uma menor capacidade de rejeitar as hipóteses nulas caso elas sejam falsas (i.e., têm menor potência), quando os pressupostos adicionais dos métodos clássicos são válidos.

Com grande frequência, embora nem sempre, os métodos não paramétricos substituem os valores observados da variável resposta pelas ordens (*ranks*) dessas observações. As estatísticas de teste são então funções dessas ordens.

3. Métodos não paramétricos de tipo ANOVA (cont.)

O teste de Kruskal-Wallis é uma alternativa não paramétrica à ANOVA a 1 Factor, em que:

- Cada observação é substituída pela sua ordem;
- A estatística de teste compara as ordens médias em cada nível do factor com a ordem média global.
- A hipótese nula é que nos vários níveis do factor as observações seguem a mesma distribuição.
- A hipótese alternativa é que a distribuição dos vários níveis difere apenas nas suas localizações (medianas).

II.3. Análise de Covariância: um exemplo

A Regressão Linear e as Análises de Variância estudadas até aqui, são casos particulares do Modelo Linear, que inclui também as Análises de Covariância.

Em qualquer destas três situações se procura modelar uma variável resposta quantitativa (numérica) Y. O que distingue as três situações é a natureza das variáveis preditoras.

- Numa Regressão Linear, as variáveis preditoras são variáveis igualmente quantitativas (numéricas).
- Numa Análise de Variância, as variáves preditoras são factores (variáveis qualitativas, ou categóricas).
- Numa Análise de Covariância, entre as variáveis preditoras encontramos quer variáveis numéricas, quer factores.

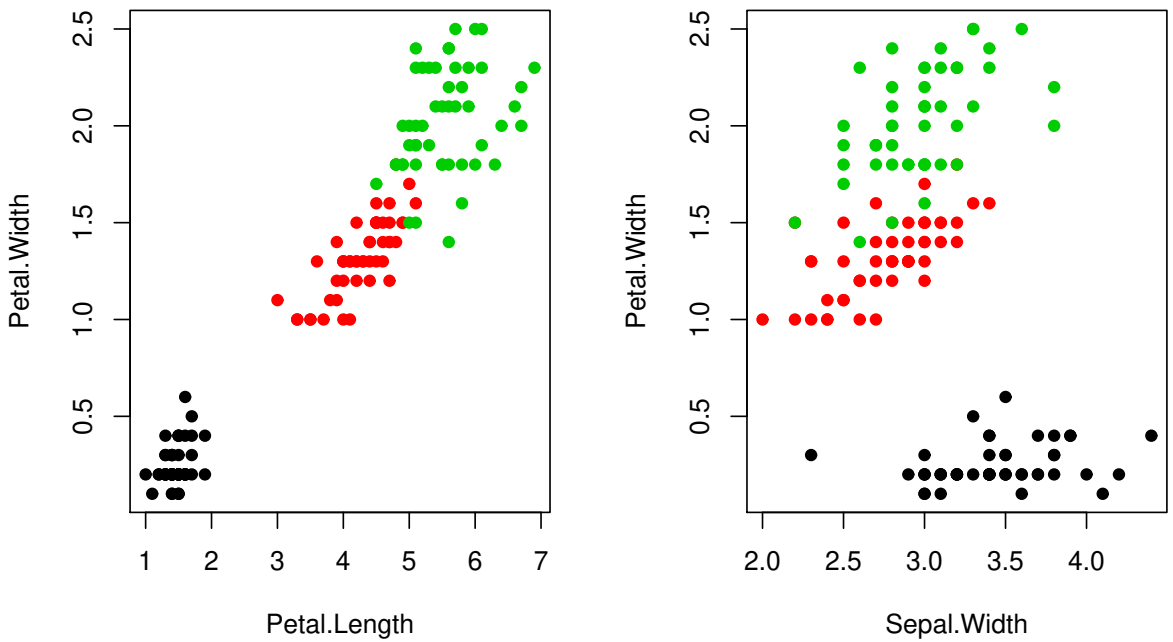
Um exemplo de Análise de Covariância (cont.)

A Análise de Covariância será apenas discutida no contexto dum problema específico de interesse prático, associado à Regressão Linear.

Admita que se verificou ser válida uma regressão linear simples entre uma variável Y e um preditor x, num dado contexto. Surge de forma natural a questão de saber se a recta de regressão teórica é, ou não, idêntica, noutros contextos aparentados, ou seja, noutros níveis de um dado factor.

Um exemplo de Análise de Covariância (cont.)

No exemplo dos lírios, a relação entre Largura de Pétala e Comprimento de Pétala (gráfico à esquerda) talvez seja comum para as três espécies de lírios (*setosa*, *versicolor* e *virginica*).
Já a relação entre Largura de Pétala e Largura de Sépala é claramente diferente para cada espécie (e quase inexistente, enquanto relação linear, para o conjunto das três espécies - gráfico à direita):



Um exemplo de Análise de Covariância (cont.)

O problema em questão pode ser formulado como um problema de Análise de Covariância pois consiste no estudo duma relação linear entre y e x, mas influenciada também por uma variável qualitativa: o factor espécie, que tem três níveis, ou seja, três diferentes espécies.

O problema será formulado de tal forma que admitir a existência de uma única relação nas três espécies seja admitir a igualdade entre um modelo de regressão linear completo e um seu submodelo - permitindo assim usar a teoria de que já dispomos para esse efeito.

Um exemplo de Análise de Covariância (cont.)

Considere-se o exemplo de três contextos aparentados (e.g. espécies, localidades, anos, etc.), nas quais a relação entre uma variável resposta Y e uma preditora X seja dada, respectivamente, por:

- Contexto 1: $Y = \beta_0 + \beta_1 x + \varepsilon$
- Contexto 2: $Y = \beta_0^* + \beta_1^* x + \varepsilon$
- Contexto 3: $Y = \beta_0^{**} + \beta_1^{**} x + \varepsilon$

Vamos considerar que o primeiro contexto é o nível de referência e escrever os parâmetros dos contextos restantes à custa dos primeiros:

$$\beta_0^* = \beta_0 + \alpha_{0:2} \quad ; \quad \beta_1^* = \beta_1 + \alpha_{1:2}$$
$$\beta_0^{**} = \beta_0 + \alpha_{0:3} \quad ; \quad \beta_1^{**} = \beta_1 + \alpha_{1:3}$$

Com os parâmetros de cada recta escritos desta forma, a hipótese de que as três rectas de regressão sejam iguais é a hipótese

$$\alpha_{0:2} = \alpha_{0:3} = \alpha_{1:2} = \alpha_{1:3} = 0 \ .$$

As variáveis associadas aos acréscimos

Considere que se fazem n observações para ajustar o modelo, sendo

- n_1 correspondentes ao primeiro contexto;
- n_2 correspondentes ao segundo contexto;
- n_3 correspondentes ao terceiro contexto.

Definam-se as variáveis indicatrizes $\vec{\mathscr{I}}_j$ de pertença aos níveis (como na Análise de Variância). Definam-se também vectores com os valores da variável X num dado contexto i ($i > 1$) e zero noutras posições, que serão representados por $x \star \vec{\mathscr{I}}_j$:

$$\vec{\mathscr{I}}_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad x \star \vec{\mathscr{I}}_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ 0 \\ 0 \end{bmatrix}, \quad \vec{\mathscr{I}}_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}, \quad x \star \vec{\mathscr{I}}_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ x_8 \\ x_9 \end{bmatrix}$$

A equação de base para as 3 rectas

Podemos agora escrever a relação de base entre o vector $\vec{\mathbf{Y}}$ das n observações da variável resposta, e o preditor X, da seguinte forma:

$$\vec{\mathbf{Y}} = \beta_0 \cdot \vec{\mathbf{1}}_n + \beta_1 \cdot \vec{\mathbf{x}} + \alpha_{0:2} \cdot \vec{\mathscr{I}}_2 + \alpha_{0:3} \cdot \vec{\mathscr{I}}_3 + \alpha_{1:2} \cdot \vec{\mathbf{x}} \star \vec{\mathscr{I}}_2 + \alpha_{1:3} \cdot \vec{\mathbf{x}} \star \vec{\mathscr{I}}_3 \ .$$

No exemplo com as $n_1 = 3$, $n_2 = 4$ e $n_3 = 2$ observações:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \\ Y_9 \end{bmatrix} = \begin{bmatrix} 1 & x_1 & 0 & 0 & 0 & 0 \\ 1 & x_2 & 0 & 0 & 0 & 0 \\ 1 & x_3 & 0 & 0 & 0 & 0 \\ 1 & x_4 & 1 & 0 & x_4 & 0 \\ 1 & x_5 & 1 & 0 & x_5 & 0 \\ 1 & x_6 & 1 & 0 & x_6 & 0 \\ 1 & x_7 & 1 & 0 & x_7 & 0 \\ 1 & x_8 & 0 & 1 & 0 & x_8 \\ 1 & x_9 & 0 & 1 & 0 & x_9 \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \alpha_{0:2} \\ \alpha_{0:3} \\ \alpha_{1:2} \\ \alpha_{1:3} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \end{bmatrix}$$

A equação de base (cont.)

Isto é,

$$Y_i = \begin{cases} \beta_0 + \beta_1 x_i + \varepsilon_i, & \text{se } i = 1,2,3 \\ (\beta_0 + \alpha_{0:2}) + (\beta_1 + \alpha_{1:2}) x_i + \varepsilon_i, & \text{se } i = 4,5,6,7 \\ (\beta_0 + \alpha_{0:3}) + (\beta_1 + \alpha_{1:3}) x_i + \varepsilon_i, & \text{se } i = 8,9\ . \end{cases} \tag{1}$$

O modelo do acetato 324 ajusta, às observações de cada um dos três contextos, uma recta de regressão distinta.

Caso os parâmetros de acréscimo $\alpha_{i:j}$ sejam *todos* iguais a zero, a recta de regressão é a mesma, para os três contextos.

A relação de base para comparar 3 rectas

Temos assim uma equação do tipo modelo linear com $3 \times 2 = 6$ parâmetros (e variáveis preditoras $\vec{\mathbf{x}}$, $\vec{\mathscr{I}}_2$, $\vec{\mathscr{I}}_3$, $\vec{\mathbf{x}} \star \vec{\mathscr{I}}_2$, $\vec{\mathbf{x}} \star \vec{\mathscr{I}}_3$), que ajusta rectas de regressão diferentes para as observações de cada um dos 3 contextos. Caso $\alpha_{0:2} = \alpha_{0:3} = \alpha_{1:2} = \alpha_{1:3} = 0$, obtém-se o submodelo correspondente a ajustar uma única recta aos 3 contextos:

$$\begin{aligned} \vec{\mathbf{Y}} &= \beta_0 \vec{\mathbf{1}}_n + \beta_1 \vec{\mathbf{x}} + \alpha_{0:2} \vec{\mathscr{I}}_2 + \alpha_{0:3} \vec{\mathscr{I}}_3 + \alpha_{1:2} \vec{\mathbf{x}} \star \vec{\mathscr{I}}_2 + \alpha_{1:3} \vec{\mathbf{x}} \star \vec{\mathscr{I}}_3 + \vec{\boldsymbol{\varepsilon}} \\ \vec{\mathbf{Y}} &= \beta_0 \cdot \vec{\mathbf{1}}_n + \beta_1 \cdot \vec{\mathbf{x}} + \vec{\boldsymbol{\varepsilon}} \end{aligned}$$

Um teste F parcial comparando estes dois modelos permite testar a admissibilidade duma recta única para os três contextos considerados.

Outras comparações no exemplo

É possível fazer outras comparações, com base no modelo

$$\vec{\mathbf{Y}} = \beta_0 \vec{\mathbf{1}}_n + \beta_1 \vec{\mathbf{x}} + \alpha_{0:2} \vec{\mathscr{I}}_2 + \alpha_{0:3} \vec{\mathscr{I}}_3 + \alpha_{1:2} \vec{\mathbf{x}} \star \vec{\mathscr{I}}_2 + \alpha_{1:3} \vec{\mathbf{x}} \star \vec{\mathscr{I}}_3 + \vec{\boldsymbol{\varepsilon}}$$

- A hipótese de três rectas paralelas (i.e., com o mesmo declive), mas podendo ter diferentes ordenadas na origem, é a hipótese $\alpha_{1:2} = \alpha_{1:3} = 0$.
- A hipótese de três rectas com igual ordenada na origem, mas declives diferentes, é a hipótese $\alpha_{0:2} = \alpha_{0:3} = 0$.
- A hipótese de a primeira e segunda recta terem o mesmo declive, é a hipótese $\alpha_{1:2} = 0$.
- A hipótese de a segunda e terceira recta terem o mesmo declive, é a hipótese $\alpha_{1:2} = \alpha_{1:3}$, ou seja, $\alpha_{1:2} - \alpha_{1:3} = 0$.

Estas hipóteses (ou outras análogas) podem ser testadas através de testes já vistos no estudo geral do modelo linear.

Cruzando factores com variáveis numéricas no R

No R, um modelo de regressão de y sobre x, admitindo rectas diferentes para cada nível do factor f, é dado pela fórmula: $y \sim x * f$

No exemplo dos $n = 150$ lírios,

```
> modespecie.lm <- lm(Petal.Length ~ Sepal.Length * Species)
> summary(modespecie.lm)
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                    0.8031     0.5310   1.512    0.133
Sepal.Length                   0.1316     0.1058   1.244    0.216
Speciesversicolor             -0.6179     0.6837  -0.904    0.368
Speciesvirginica              -0.1926     0.6578  -0.293    0.770
Sepal.Length:Speciesversicolor  0.5548     0.1281   4.330 2.78e-05 ***
Sepal.Length:Speciesvirginica   0.6184     0.1210   5.111 1.00e-06 ***
--
Residual standard error: 0.2611 on 144 degrees of freedom
Multiple R-squared: 0.9789,     Adjusted R-squared: 0.9781
F-statistic:  1333 on 5 and 144 DF,  p-value: < 2.2e-16
```

Um exemplo no R. As 3 rectas.

As três rectas ajustadas pelo modelo conjunto:

Para a espécie *Setosa* (referência):

$$PL = 0.8031 + 0.1316\,SL$$

Para a espécie *Versicolor*:

$$PL = (0.8031 - 0.6179) + (0.1316 + 0.5548)\,SL = 0.1851 + 0.6865\,SL$$

Para a espécie *Virginica*:

$$PL = (0.8031 - 0.1926) + (0.1316 + 0.6184)\,SL = 0.6105 + 0.7501\,SL$$

São as mesmas rectas que resultam de ajustar apenas as $n_i = 50$ observações de cada espécie.

Um exemplo no R. Recta única?

De novo o exemplo dos 150 lírios. Pretende-se modelar Comprimento das Pétalas, à custa de Comprimento das Sépalas.
Recta única ou rectas diferenciadas por espécie?

```
> modunico.lm <- lm(Petal.Length ~ Sepal.Length)
> modespecie.lm <- lm(Petal.Length ~ Sepal.Length*Species)
> anova(modunico.lm, modespecie.lm)
Analysis of Variance Table
Model 1: Petal.Length ~ Sepal.Length
Model 2: Petal.Length ~ Sepal.Length * Species
  Res.Df     RSS  Df Sum of Sq     F    Pr(>F)
1    148 111.459
2    144   9.818   4   101.641 372.7 < 2.2e-16 ***
```

Rejeita-se a hipótese de uma recta única, em favor de rectas diferentes.

Um exemplo no R. Rectas paralelas?

No R, um modelo de regressão de y sobre x, que admite rectas paralelas, mas com diferentes ordenadas na origem para cada nível de um factor f, pode ser indicado pela fórmula: $y \sim x + f$

```
> modparalelas.lm <- lm(Petal.Length ~ Sepal.Length + Species)
> summary(modparalelas.lm)
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.70234    0.23013  -7.397 1.01e-11 ***
Sepal.Length       0.63211    0.04527  13.962  < 2e-16 ***
Speciesversicolor  2.21014    0.07047  31.362  < 2e-16 ***
Speciesvirginica   3.09000    0.09123  33.870  < 2e-16 ***
--
Residual standard error: 0.2826 on 146 degrees of freedom
Multiple R-squared: 0.9749,     Adjusted R-squared: 0.9744
F-statistic:  1890 on 3 and 146 DF,  p-value: < 2.2e-16
```

Um exemplo no R: as 3 rectas paralelas

As três rectas ajustadas pelo modelo de rectas paralelas:

Para a espécie *Setosa* (referência):

$$PL = -1.70234 + 0.63211\, SL$$

Para a espécie *Versicolor*:

$$PL = (-1.70234 + 2.21014) + 0.63211\, SL = 0.50780 + 0.63211\, SL$$

Para a espécie *Virginica*:

$$PL = (-1.70234 + 3.09000) + 0.63211\, SL = 1.38766 + 0.63211\, SL$$

Um exemplo no R. Rectas paralelas? (cont.)

Mas é admissível que as três rectas sejam paralelas?
Vamos fazer um teste aos modelos encaixados que admitem rectas paralelas e rectas diferentes.

```
> modparalelas.lm <- lm(Petal.Length ~ Sepal.Length + Species)
> modespecie.lm <- lm(Petal.Length ~ Sepal.Length * Species)
> anova(modparalelas.lm,modespecie.lm)
Analysis of Variance Table
Model 1: Petal.Length ~ Sepal.Length + Species
Model 2: Petal.Length ~ Sepal.Length * Species
  Res.Df     RSS  Df Sum of Sq      F    Pr(>F)
1    146 11.6571
2    144  9.8179   2    1.8393 13.489 4.272e-06 ***
```

Rejeita-se a hipótese de rectas paralelas.

Os pressupostos

Os testes anteriormente referidos são válidos caso se verifiquem os pressupostos já admitidos nos Modelos Lineares, i.e., que os erros aleatórios da equação do modelo verificam:

- $\varepsilon_{ij} \frown \mathcal{N}(0, \sigma^2), \ \forall i, j$;
- erros aleatórios independentes.

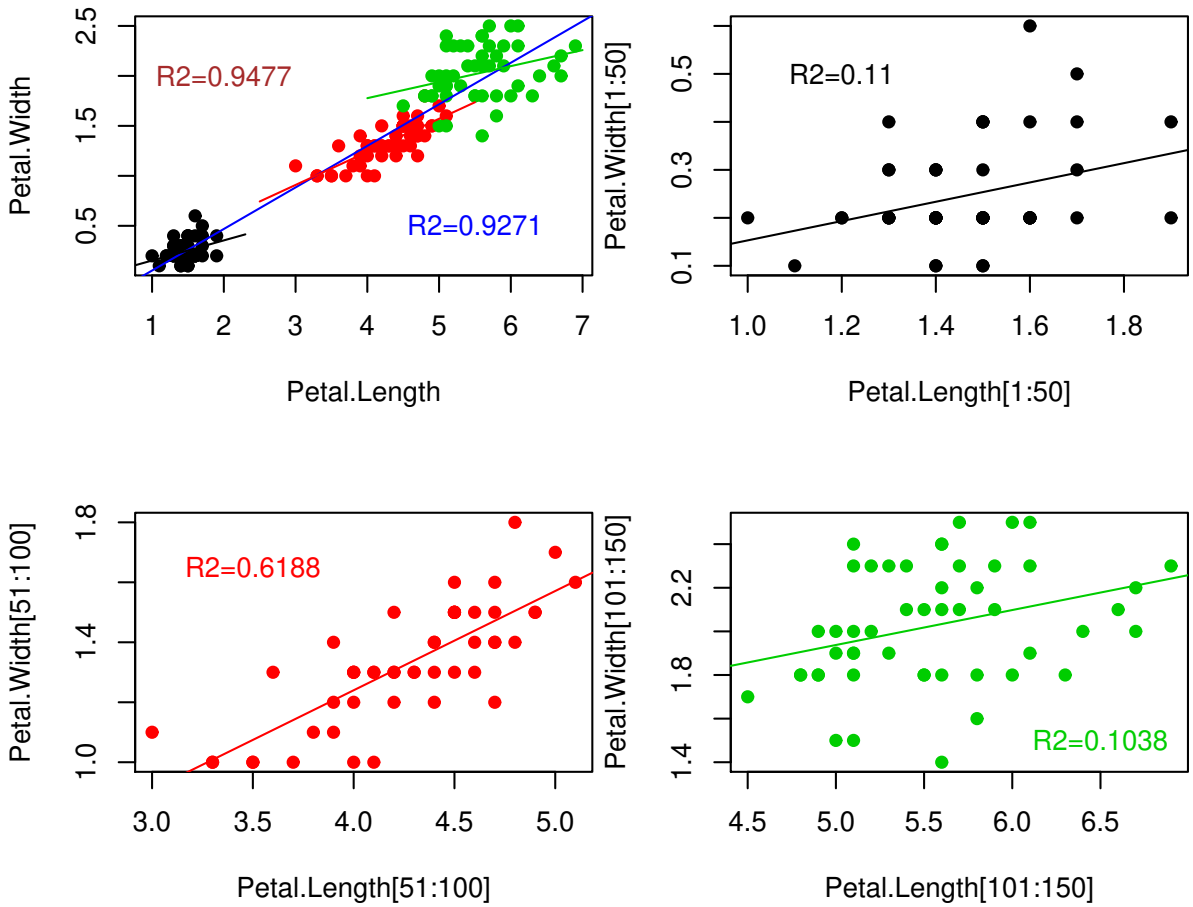
Trata-se (quase) dos mesmos pressupostos que seria necessário supor para ajustar cada recta, de forma separada, usando apenas as n_i observações correspondentes ao seu contexto.

Mas há pressupostos adicionais em relação ao ajustamento em separado: a independência e a homogeneidade das variâncias dos erros aleatórios têm de ser válidas no conjunto dos 3 contextos comparados.

Estes resultados generalizam-se para qualquer número de rectas, ou para modelos lineares com qualquer número de preditores.

Um exemplo

Exemplo do acetato 320, com as relações entre Largura de Pétala e Comprimento de Pétala única, diferenciada e separada, para as três espécies de lírios (*setosa*, *versicolor* e *virginica*). Atenção aos R^2!



Comparando os Coeficientes de Determinação

É possível relacionar os Coeficiente de Determinação do modelo conjunto, R^2, e dos s modelos individuais, $R^2_{[i]}$. Tem-se:

$$R^2 = \frac{\sum_{i=1}^{s} R^2_{[i]}\, SQT_i + SQF}{\sum_{i=1}^{s} SQT_i + SQF}.$$

sendo SQR_i e SQT_i as SQs das observações do nível i, e SQF a SQ do Factor na ANOVA de todas as observações, sobre o factor que determina os s casos comparados (sem a variável preditora numérica).

- se $SQF \approx 0$ (i.e., o Factor não tem efeitos significativos sobre Y), R^2 é aproximadamente uma média ponderada dos $R^2_{[i]}$ (com pesos SQT_i). Neste caso, $R^2 \approx 1$ só se a generalidade dos $R^2_{[i]} \approx 1$.
- para SQF grande (i.e., efeitos significativos do Factor sobre Y), a separação das médias de Y em cada grupo predomina na expressão. $SQF \gg \sum_{i=1}^{s} SQT_i \Rightarrow R^2 \approx 1$.

Ainda o exemplo do Acetato 335

Os valores de cada Soma de Quadrados, bem como do Coeficiente de Determinação, para cada um dos modelos referidos no exemplo do Acetato 335, são:

	SQT	SQR	SQRE	QMRE	R2
setosa	0.54420	0.05985029	0.4843497	0.01009062	0.1099785
versicolor	1.91620	1.18583401	0.7303660	0.01521596	0.6188467
virginica	3.69620	0.38349444	3.3127056	0.06901470	0.1037537
conjunto	86.56993	82.04251207	4.5274213	0.03144043	0.9477022

Resultados ANOVA a 1 Factor: Petal.Width ~ Species
SQF=80.41333 SQRE=6.15660

É o valor elevado de *SQF* que gera um valor elevado do R^2 conjunto.

NOTA: o modelo único não surge nesta comparação.

Generalizando para qualquer número de preditores

A ideia de fundo usada para comparar rectas de regressão linear em s contextos diferentes pode ser generalizada para estudar qualquer regressão linear múltipla em s contextos diferentes.

Para cada preditor, admite-se a possibilidade de haver acréscimos no respectivo coeficiente (em relação ao coeficiente do primeiro contexto), diferentes em cada um dos restantes contextos.