

Modelos Matemáticos e Aplicações

Modelo Linear

Jorge Cadima

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

2023-24

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. (2020) [O Modelo Linear](#)

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**]

3 Referências de apoio à utilização do R

- ▶ Agresti, Alan (2015) *Foundations of Linear and Generalized 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, \dots, x_p .

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

Os nossos modelos

No Modelo Linear apenas se consideram modelos:

- com uma única variável resposta **numérica**.
- ajustadas com n observações **independentes**.

Quanto aos **preditores**:

- pode ter-se **um ou mais preditores**;
- os preditores podem ser **numéricos ou categóricos (factores)**.

Motivamos a discussão com alguns **exemplos**.

Exemplo 1: regressão linear simples (descritiva)

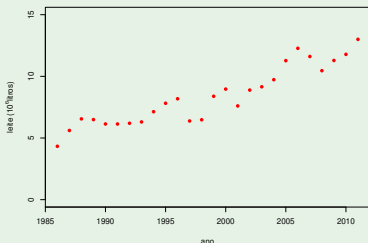
Exemplo 1: Leite de Cabra

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)

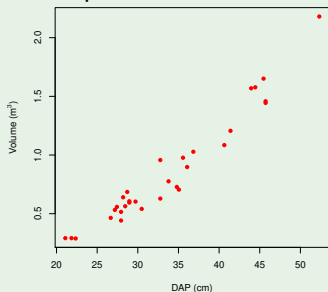
Exemplo 2: Volume de troncos de cerejeiras

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.

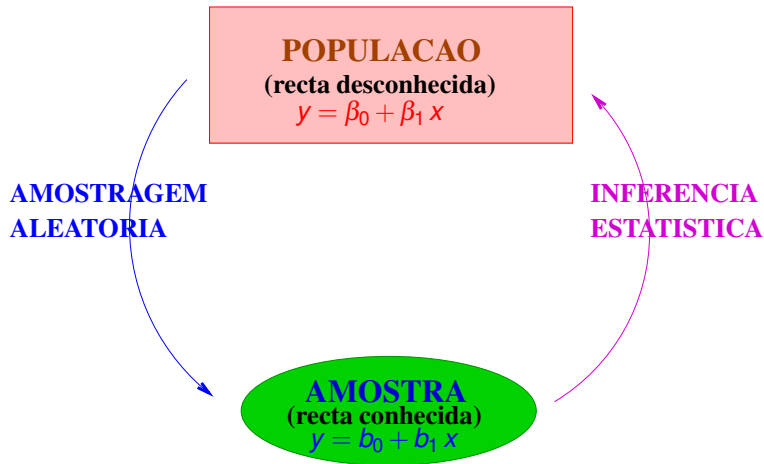


Tendência de fundo aproximadamente linear.

Temos uma **amostra aleatória** duma população maior. Interessa o **contexto**

inferencial: o que sabemos sobre a **recta populacional** $y = \beta_0 + \beta_1 x$?

O problema da Inferência Estatística na RLS



Exemplo 3: ANOVA a um factor

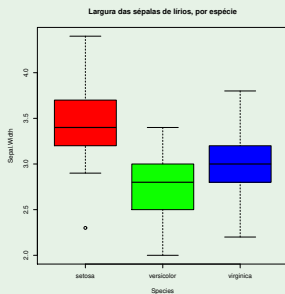
Exemplo 3: Largura de Sépalas em lírios

Resposta (numérica): largura de sépalas de 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?

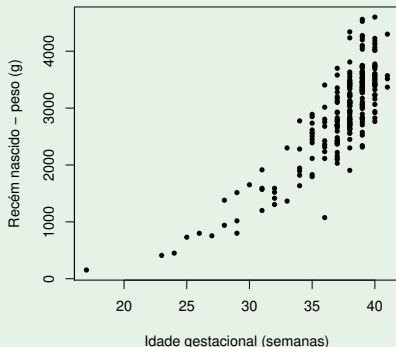
Exemplo 4 - relação não linear (descritivo)

Exemplo 4: Pesos de bebês à nascença

Resposta (numérica): peso de bebê 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 ($y = ce^{dx}$)?
 - ▶ f função potência ($y = cx^d$)?

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** (não faz parte do programa de MMA).

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)

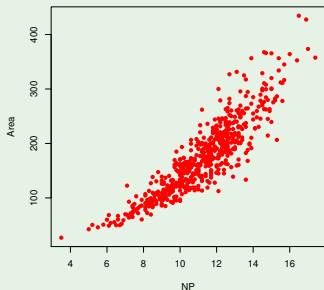
Exemplo 5: Áreas de folhas de videira

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

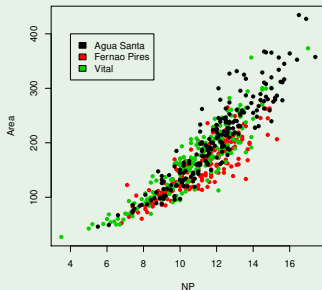
Exemplo 6: Áreas de folhas de videira

Resposta (numérica): Área 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 são preferíveis curvas diferentes para castas diferentes?

Exemplo 7 - Regressão linear múltipla

Exemplo 7: Teor de antocianas na casta Tinta Francisca

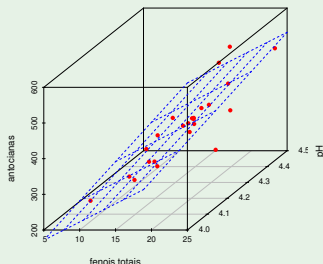
Resposta (numérica): Teor de antocianas (y , $antoci$) (em mg/dm^3).

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, ISA (Tabuaço 2003).



Descritivo: qual o “melhor” plano amostral $y = b_0 + b_1x_1 + b_2x_2$?

Inferencial: que dizer sobre o plano populacional $y = \beta_0 + \beta_1x_1 + \beta_2x_2$?

Ideias prévias sobre modelação

- Todos os modelos são apenas **aproximações** da realidade.
- Podem existir **diferentes modelos adequados**.
- 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-se que há **variação** das observações em torno dessa relação de fundo. Essa variabilidade tem de ser **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

Dados n pares de observações $\{(x_i, y_i)\}_{i=1}^n$, tem-se:

Recta de regressão linear de y sobre x

$$y = b_0 + b_1 x$$

com

$$\text{Declive } b_1 = \frac{\text{COV}_{xy}}{s_x^2} \quad \left(\frac{\text{unidades de } y}{\text{unidades de } x} \right)$$

$$\text{Ordenada na origem } b_0 = \bar{y} - b_1 \bar{x} \quad (\text{unidades de } y)$$

sendo

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

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 1795-1809).

Resíduos e Soma de Quadrados Residual

Os **resíduos** são distâncias **na vertical** (afectadas de sinal) entre pontos e recta ajustada:

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

sendo $\hat{y}_i = b_0 + b_1 x_i$ os “valores de **y ajustados pela recta**”.

Soma de Quadrados dos Resíduos:

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

Critério: Determinar b_0 e b_1 que minimizam $SQRE$.

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 :

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

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

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

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

Regressão Linear Simples - contexto descritivo no R

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

O comando `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 )
```

```
Call: lm(formula = leite ~ ano, data = Cabra)
```

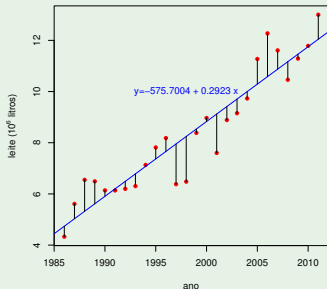
```
Coefficients:
```

```
(Intercept)      ano  
-575.7004      0.2923 <- valores ajustados de b0 e b1
```

Regressão Linear Simples descritiva - Exemplo 1

Exemplo 1: Leite de Cabra

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



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

Os parâmetros da recta de regressão

Propriedades dos parâmetros

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

Exemplo 1: Leite de Cabra

O declive ajustado $b_1 = 0.2923$ significa que, em média, a produção de leite de cabra aumentou 0.2923×10^6 litros por ano.

Mais propriedades da recta de regressão

Propriedades da recta de regressão

- A recta de regressão passa sempre no centro de gravidade da nuvem de pontos, isto é, no ponto (\bar{x}, \bar{y}) .

Pela fórmula da ordenada na origem: $b_0 = \bar{y} - b_1 \bar{x} \Leftrightarrow \bar{y} = b_0 + b_1 \bar{x}$.

- A média dos y_i observados é à média dos \hat{y}_i ajustados: $\bar{y} = \bar{\hat{y}}$.

$$\bar{\hat{y}} = \frac{1}{n} \sum_{i=1}^n \hat{y}_i = \frac{1}{n} \sum_{i=1}^n (b_0 + b_1 x_i) = \frac{1}{n} \underbrace{\sum_{i=1}^n b_0}_{=nb_0} + b_1 \underbrace{\frac{1}{n} \sum_{i=1}^n x_i}_{=\bar{x}} = b_0 + b_1 \bar{x} = \bar{y}.$$

- A média (e soma) dos resíduos é nula: $\bar{e} = 0$.

$$\bar{e} = \frac{1}{n} \sum_{i=1}^n e_i = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i) = \frac{1}{n} \underbrace{\sum_{i=1}^n y_i}_{=\bar{y}} - \frac{1}{n} \underbrace{\sum_{i=1}^n \hat{y}_i}_{=\bar{\hat{y}}} = \bar{y} - \bar{\hat{y}} = 0.$$

Comandos R para o estudo da regressão

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 calculada 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.

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

```
      1      2  
4.444889 12.336034
```

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

$$\hat{y} = b_0 + b_1 x$$
$$\Leftrightarrow 12.336034 = -575.7004 + 0.2923 \times 2012 .$$

O critério de mínimos quadrados

O critério de minimizar a Soma de Quadrados dos Resíduos,

$$SQRE = \sum_{i=1}^n (y_i - \hat{y}_i)^2, \text{ tem subjacente um pressuposto:}$$

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

variável resposta y (“dependente”) é a **variável que se quer modelar**, a partir da variável x .

variável preditora x (“independente” ou “explicativa”) é a **variável que se admite conhecida**, e **com base na qual se pretende tirar conclusões sobre y** .

A recta de regressão de y sobre x é diferente da recta de regressão de x 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: As três Somas de Quadrados

Recordar: $s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$ a variância amostral das observações y_i .

Soma de Quadrados Total (SQT)

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

Tem-se: $s_{\hat{y}}^2 = \frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ a variância amostral dos \hat{y}_i ajustados.

Soma de Quadrados da Regressão (SQR)

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

Soma de Quadrados Residual (SQRE) - já dado

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

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.

Propriedades do Coeficiente de Determinação

Propriedades de $R^2 = \frac{SQR}{SQT}$

- $0 \leq R^2 \leq 1$ (Todas as SQs são não negativas e $SQT = SQR + SQRE$)
- $R^2 = 1$ se, e só se, os n pontos são colineares. (“ideal”)
($SQT = SQR \Leftrightarrow SQRE = 0$. Logo, todos os resíduos são nulos: os pontos estão todos em cima da recta.)
- $R^2 = 0$ se, e só se, a recta de regressão for horizontal. (“inútil”)
($SQR = 0 \Leftrightarrow SQRE = SQT$. Toda a variabilidade de y é residual, não há variabilidade nos \hat{y}_i . Recta é $y = \bar{y} \Leftrightarrow b_1 = 0$)
- Numa regressão linear **simples**, R^2 é o quadrado do coeficiente de correlação linear entre x e y (Ver Exercícios):

$$R^2 = r_{xy}^2 = \left(\frac{COV_{xy}}{s_x s_y} \right)^2 \quad \text{se } s_x \neq 0 \text{ e } s_y \neq 0$$

Exemplo 1: leite da cabra

O coeficiente de determinação R^2 obtem-se aplicando o comando `summary` a uma `regressão ajustada`. Surge com a designação `Multiple R-Squared`.

```
> 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 suas componentes:

```
> 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: `summary(Cabra.lm)$r.sq` extrai o valor de R^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 inventou a designação **eugenia**, conceito que era considerado respeitável até ao início do Século XX.

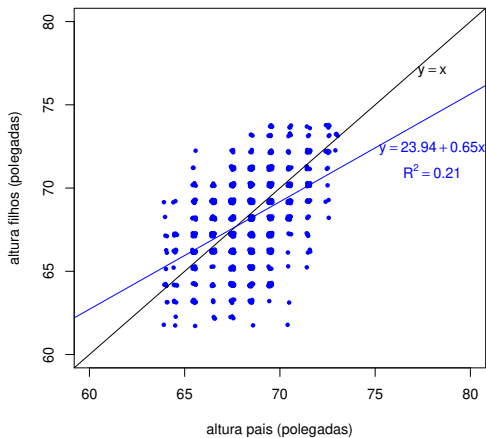
Galton constatou que pais com estaturas acima da média tinham tendência a ter filhos com estatura acima da média - mas menos que os pais (análogo para os abaixo da média).

Galton chamou ao seu artigo *Regression towards mediocrity in hereditary stature*. A designação **regressão** ficou associada ao método devido a esta acaso histórico.

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

Um pouco de história (cont.)

Dados da Regressão de Galton (n=928)



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.

Animals - no módulo MASS

```
> library(MASS)    <- carregar o módulo 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.
```

Exemplo: os dados Animals

> Animals

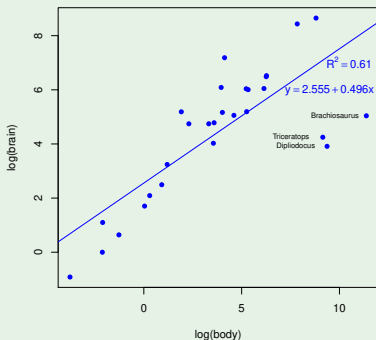
| | body | brain |
|------------------|-----------|--------|
| Mountain beaver | 1.350 | 8.1 |
| Cow | 465.000 | 423.0 |
| Grey wolf | 36.330 | 119.5 |
| Goat | 27.660 | 115.0 |
| Guinea pig | 1.040 | 5.5 |
| Dipliodocus | 11700.000 | 50.0 |
| Asian elephant | 2547.000 | 4603.0 |
| Donkey | 187.100 | 419.0 |
| Horse | 521.000 | 655.0 |
| Potar monkey | 10.000 | 115.0 |
| Cat | 3.300 | 25.6 |
| Giraffe | 529.000 | 680.0 |
| Gorilla | 207.000 | 406.0 |
| Human | 62.000 | 1320.0 |
| African elephant | 6654.000 | 5712.0 |
| Triceratops | 9400.000 | 70.0 |
| Rhesus monkey | 6.800 | 179.0 |
| Kangaroo | 35.000 | 56.0 |
| Golden hamster | 0.120 | 1.0 |
| Mouse | 0.023 | 0.4 |
| Rabbit | 2.500 | 12.1 |
| Sheep | 55.500 | 175.0 |
| Jaguar | 100.000 | 157.0 |
| Chimpanzee | 52.160 | 440.0 |
| Rat | 0.280 | 1.9 |
| Brachiosaurus | 87000.000 | 154.5 |
| Mole | 0.122 | 3.0 |
| Pig | 192.000 | 180.0 |

RLS e observações atípicas

Exemplo: Animals

A generalidade das observações seguem uma **relação linear** entre os **logaritmos** do peso do cérebro e do peso do corpo.

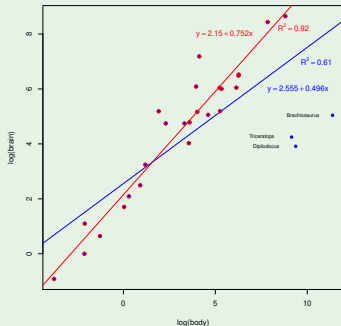
Mas três espécies de dinossaúros são **observações atípicas**, e condicionam o ajustamento da recta.



RLS e observações atípicas (cont.)

Exemplo: Animals

Excluindo essas observações muda a recta ajustada e a sua qualidade.



Neste caso é aceitável excluir as 3 observações atípicas por pertencerem a “outra realidade” (espécies extintas). Há **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.

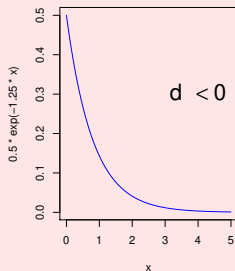
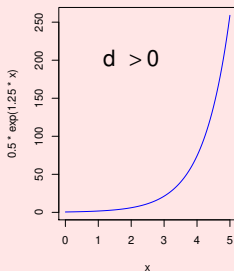
Consideremos alguns exemplos particularmente frequentes de relações não lineares que são linearizáveis através de transformações da variável resposta e, nalguns casos, também do preditor.

Relação exponencial

Relação exponencial

$$y = ce^{dx}$$

$$(y > 0 ; c > 0)$$



Transformação linearizante: $y^* = \ln(y)$ e $x^* = x$

A linearização da relação exponencial

Logaritmizando a equação da exponencial, obtém-se uma **relação linear** entre $y^* = \ln(Y)$ e x :

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

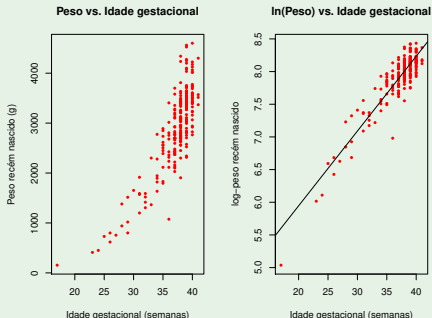
com **declive** $b_1 = d$ e **ordenada na origem** $b_0 = \ln(c) \Leftrightarrow c = e^{b_0}$.

O **sinal do declive da recta** indica se a relação exponencial original é **crescente** ($b_1 > 0$) ou **decrecente** ($b_1 < 0$).

Exemplo 4: peso de bebés à nascença

Uma linearização no peso dos bebés

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



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

Ainda a relação exponencial

Equação Diferencial da exponencial

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

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

isto é, que a **taxa de variação relativa de y** é constante:

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

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

$$\ln|y(x)| = dx + K \quad \Leftrightarrow \quad y(x) = e^{K+dx} \quad \Leftrightarrow \quad y(x) = e^K e^{dx} .$$

O declive b_1 da recta é o valor constante d da taxa de variação relativa de y .

A constante de primitivação K é a ordenada na origem da recta: $K = b_0$.

Modelo logístico 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 Verhulst¹ propôs um 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 (capacidade de sustentação do meio, *carrying capacity of the environment*).

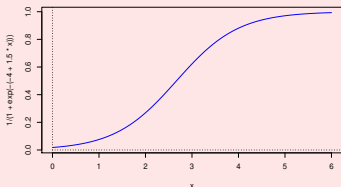
¹Verhulst, P.-F. (1838), Notice sur la loi que la population poursuit dans son accroissement. *Corresp. Math. Phys.* **10**, 113-121

Relação Logística (com 2 parâmetros)

Relação Logística (2 parâmetros)

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

$$(y \in]0, 1])$$



($d > 0$)

Transformação linearizante: transformação *logit* de y , i.e.,

$$y^* = \ln\left(\frac{y}{1-y}\right) \quad \text{e} \quad x^* = x$$

A linearização da relação logística

Como $y \in]0, 1[$, a **transformação logit**, $y^* = \ln\left(\frac{y}{1-y}\right)$, está bem definida.

A relação logística entre y e x corresponde a uma relação **linear** entre $y^* = \ln\left(\frac{y}{1-y}\right)$ e $x^* = x$:

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

Ainda a Logística

Equação Diferencial da Logística (2 parâmetros)

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

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

A equação anterior equivale a:

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

Primitivando (em ordem a x), tem-se (pois $\int \frac{f'}{f} = \ln(|f|)$):

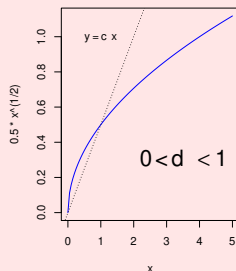
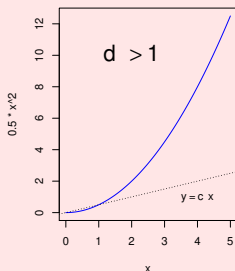
$$\begin{aligned} \ln y(x) - \ln(1 - y(x)) &= dx + K \\ \Leftrightarrow \ln\left(\frac{y}{1 - y}\right) &= b_1 x + b_0. \end{aligned}$$

Relação potência ou alométrica

Relação potência

$$y = cX^d$$

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



Transformação linearizante: $y^* = \ln(y)$ e $x^* = \ln(x)$.

A linearização dum relação potência

Logaritmizando, obtém-se:

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

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

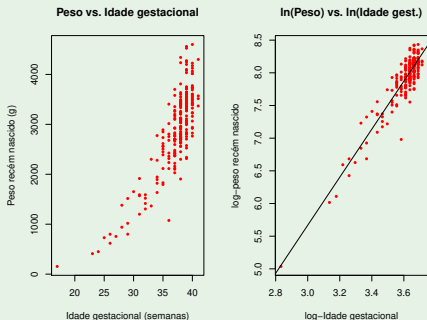
O declive b_1 da recta é o expoente d na relação potência.

A ordenada na origem é $b_0 = \ln(c)$, ou seja, $c = e^{b_0}$.

Outra linearização no Exemplo 4

Outra linearização dos pesos dos bebês

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



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

Ainda a relação potência

Uma Equação Diferencial da potência

Uma relação potência resulta de admitir que y é função de x e a taxa de variação relativa de y , i.e., a razão $\frac{y'(x)}{y(x)}$, é inversamente proporcional a x :

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

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

$$\underbrace{\ln|y(x)|}_{=y^*} = \underbrace{d}_{=b_1} \underbrace{\ln|x|}_{=x^*} + \underbrace{K}_{=b_0} \quad \Leftrightarrow \quad y(x) = e^{K+\ln(x^d)} \quad \Leftrightarrow \quad y(x) = e^K x^d.$$

O declive b_1 da recta é a constante de proporcionalidade d .

A constante de primitivação K é a ordenada na origem da recta: $K = b_0$.

Outra Eq. Diferencial para a relação potência

A Equação Diferencial da alometria

Outra forma de obter uma relação potência, muito usada nos estudos alométricos, resulta de admitir que y e x são ambas funções duma terceira variável t (ou seja, $y(t)$ e $x(t)$) e que as taxas de variação relativas de y e x são proporcionais:

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

Primitivando (em ordem a t) tem-se:

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

e exponenciando,

$$y = e^{d \ln x + K} = e^{d \ln x} \cdot e^K = x^d \cdot \underbrace{e^K}_{=c} \Leftrightarrow y = cx^d.$$

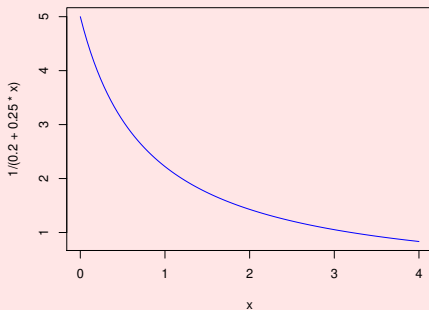
Os estudos de **alometria** comparam a dimensão de partes diferentes dum organismo. A **isometria** corresponde ao valor $d=1$.

Relação hiperbólica (ou de proporcionalidade inversa)

Relação de tipo hiperbólico

$$y = \frac{1}{c + dx}$$

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



Transformação linearizante: $y^* = 1/y$ e $x^* = x$

A linearização duma relação hiperbólica

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

$$\begin{aligned}y &= \frac{1}{c + dx} && \Leftrightarrow && \frac{1}{y} = c + dx \\ & && \Leftrightarrow && y^* = b_0 + b_1 x .\end{aligned}$$

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

Relações de tipo hiperbólico têm sido usadas, em Agronomia, para modelar a relação entre **rendimento (yield) por planta (y)** vs. **densidade da cultura ou povoamento (crop density, x)**, nalgumas culturas.

Atenção: Para valores de y próximos de zero, o recíproco pode ficar muito grande. Observações com $y_i \approx 0$ tendem a dominar o ajustamento da relação linearizada.

Ainda a relação de tipo hiperbólico

Equação diferencial da relação de tipo hiperbólico

Resulta de admitir que a taxa de variação (diminuição) de y é proporcional ao quadrado de y :

$$y'(x) = -d y^2(x)$$

ou equivalentemente, que a taxa de variação relativa de y é proporcional a y :

$$\Leftrightarrow \frac{y'(x)}{y(x)} = -d y(x).$$

Re-escrevendo a equação como $\frac{y'(x)}{y^2(x)} = -d$, e primitivando $\left(\int f^\alpha \cdot f' = \frac{f^{\alpha+1}}{\alpha+1}\right)$, tem-se:

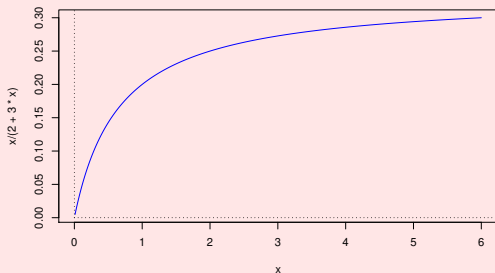
$$-\underbrace{\frac{1}{y(x)}}_{=y^*} = -\underbrace{d}_{=b_1} x + \underbrace{K}_{=b_0} \Leftrightarrow y(x) = \frac{1}{dx + c},$$

com $c = -K$. A constante de proporcionalidade $(-d)$ é o simétrico do declive da recta (b_1) .

Relação Michaelis-Menten

Relação Michaelis-Menten

$$y = \frac{x}{c+dx}$$



A recta horizontal $y = \frac{1}{d}$ é uma assintota à direita.

Transformação linearizante: $y^* = \frac{1}{y}$ e $x^* = \frac{1}{x}$

Linearizando a relação Michaelis-Menten

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

$$\begin{aligned}y = \frac{x}{c + dx} &\Leftrightarrow \frac{1}{y} = \frac{c + dx}{x} \\&\Leftrightarrow \frac{1}{y} = \frac{c}{x} + d = c \cdot \frac{1}{x} + d \\&\Leftrightarrow y^* = b_0 + b_1 x^*,\end{aligned}$$

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

Atenção: Para valores de y ou x próximos de zero, o recíproco pode ficar muito grande. **Observações** com $y_i \approx 0$ e/ou $x_i \approx 0$ tendem a dominar o ajustamento da relação linearizada.

Relação Michaelis-Menten (cont.)

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

Relação Michaelis-Menten (cont.)

Equação Diferencial duma Michaelis-Menten

Uma relação Michaelis-Menten resulta de admitir que y é função de x e a taxa de variação de y é proporcional ao quadrado da razão entre y e x :

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

Re-escrevendo a equação como $\frac{y'(x)}{y^2(x)} = c \frac{1}{x^2}$, e primitivando $\left(\int f^\alpha \cdot f' = \frac{f^{\alpha+1}}{\alpha+1} \right)$, tem-se:

$$\begin{aligned} -\frac{1}{y(x)} &= -c \frac{1}{x} + K \Leftrightarrow \frac{1}{y(x)} = \underbrace{c \frac{1}{x} - K}_{\Leftrightarrow y^* = b_1 x^* + b_0} = \frac{c - Kx}{x} \\ \Leftrightarrow y(x) &= \frac{x}{dx + c} , \end{aligned}$$

com $d = -K = b_0$ e $c = b_1$, a constante de proporcionalidade.

Advertência sobre transformações linearizantes

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

Pode ser necessário ter **mais do que uma variável preditora** para modelar a variável resposta de interesse.

Exemplo 7: dados *Antoci*

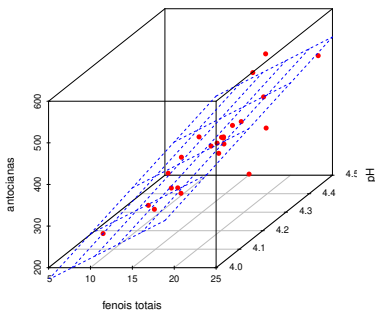
Num estudo sobre uma população experimental de clones da casta Tinta Francisca, realizado no Tabuaço em 2003, foram medidos os valores das seguintes variáveis para 24 videiras:

- 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.

O gráfico do Exemplo 7

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 foi feito usando a função `scatterplot3d`, no módulo (*package*) do R com o mesmo nome.



O módulo `rggobi`, acoplado ao programa `Ggobi`, é uma ferramenta mais poderosa para a visualização de gráficos tri-dimensionais.

Plano em \mathbb{R}^3

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

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

No nosso contexto, e colocando:

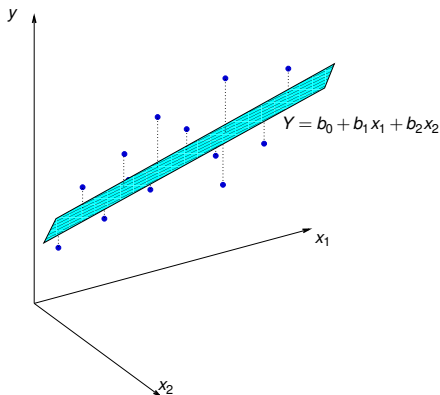
- no eixo vertical (z) a variável resposta Y ;
- noutro eixo (x) um preditor X_1 ;
- no terceiro eixo (y) o outro preditor X_2 ,

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

$$\begin{aligned} Ax_1 + Bx_2 + Cy + D = 0 &\Leftrightarrow Cy = -D - Ax_1 - Bx_2 \\ &\Leftrightarrow y = -\frac{D}{C} - \frac{A}{C}x_1 - \frac{B}{C}x_2 \\ &\Leftrightarrow y = b_0 + b_1x_1 + b_2x_2 \end{aligned}$$

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

Regressão linear múltipla ($p=2$)



$y = b_0 + b_1 x_1 + b_2 x_2$ é a equação dum plano em \mathbb{R}^3 (x_1, x_2, y).

A equação tem 3 parâmetros: b_0 , b_1 e b_2 .

Podem ser ajustados 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, \dots, x_p . Dispõe-se de n conjuntos de observações:

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

Problema: A representação usual deixa de ser visualizável se $p > 2$, 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.

Regressão Múltipla: o hiperplano ajustado

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

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

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 de **minimizar a Soma de Quadrados dos Resíduos**, ou seja, escolher os $p+1$ parâmetros $\{b_j\}_{j=0}^p$ que minimizem:

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

onde os y_i representam os valores observados da variável resposta e

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

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 :

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

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

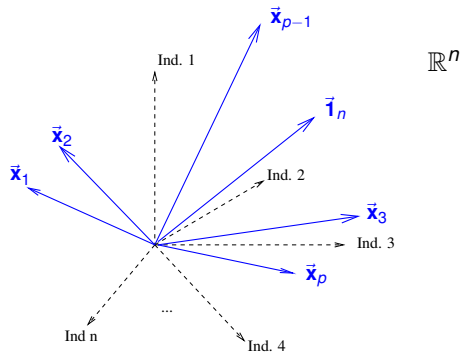
$$\vec{x}_j = (x_{j(1)}, x_{j(2)}, x_{j(3)}, \dots, x_{j(n)})^t \quad (j = 1, 2, \dots, p).$$

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

A representação em \mathbb{R}^n

Nesta **representação alternativa**,

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



O **vector de n uns**, representado por \vec{i}_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 , $\vec{\hat{y}}$:

$$\begin{aligned}\vec{\hat{y}} &= \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \dots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} b_0 + b_1 x_{1(1)} + b_2 x_{2(1)} + \dots + b_p x_{p(1)} \\ b_0 + b_1 x_{1(2)} + b_2 x_{2(2)} + \dots + b_p x_{p(2)} \\ b_0 + b_1 x_{1(3)} + b_2 x_{2(3)} + \dots + b_p x_{p(3)} \\ \dots \\ b_0 + b_1 x_{1(n)} + b_2 x_{2(n)} + \dots + b_p x_{p(n)} \end{bmatrix} \\ &= b_0 \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + b_1 \begin{bmatrix} x_{1(1)} \\ x_{1(2)} \\ x_{1(3)} \\ \vdots \\ x_{1(n)} \end{bmatrix} + \dots + b_p \begin{bmatrix} x_{p(1)} \\ x_{p(2)} \\ x_{p(3)} \\ \vdots \\ x_{p(n)} \end{bmatrix} \\ &= b_0 \vec{\mathbf{1}}_n + b_1 \vec{\mathbf{x}}_1 + b_2 \vec{\mathbf{x}}_2 + \dots + b_p \vec{\mathbf{x}}_p\end{aligned}$$

O vector $\vec{\hat{y}}$ é uma **combinação linear** dos vectores $\vec{\mathbf{1}}_n$, $\vec{\mathbf{x}}_1$, $\vec{\mathbf{x}}_2$, ..., $\vec{\mathbf{x}}_p$

A matriz do modelo \mathbf{X}

O vector $\vec{\hat{y}}$ dos valores ajustados pode também escrever-se como um produto envolvendo uma matriz \mathbf{X} cujas colunas sejam os vectores $\vec{\mathbf{1}}_n, \vec{\mathbf{x}}_1, \dots, \vec{\mathbf{x}}_p$.

A matriz \mathbf{X} do modelo

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

$\underbrace{\hspace{1.5cm}}_{=\vec{\mathbf{1}}_n} \quad \underbrace{\hspace{1.5cm}}_{=\vec{\mathbf{x}}_1} \quad \underbrace{\hspace{1.5cm}}_{=\vec{\mathbf{x}}_2} \quad \cdots \quad \underbrace{\hspace{1.5cm}}_{=\vec{\mathbf{x}}_p}$

A matriz do modelo \mathbf{X} é de dimensão $n \times (p + 1)$.

Os produtos matriciais $\mathbf{X}\vec{a}$

Os produtos da forma $\mathbf{X}\vec{a}$ são **combinações lineares das colunas da matriz \mathbf{X}** :

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

O **vector $\vec{\mathbf{y}}$** pode ser escrito desta forma: $\vec{\mathbf{y}} = \mathbf{X}\vec{\mathbf{b}}$, para algum vector de coeficientes (ainda desconhecido) $\vec{\mathbf{b}} \in \mathbb{R}^{p+1}$.

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

- O conjunto de **todas** as combinações lineares dum conjunto de vectores chama-se o **subespaço gerado** (spanned) por esses vectores.
- O subespaço gerado pelas colunas da matriz do modelo \mathbf{X} chama-se **subespaço das colunas** (column-space) da matriz \mathbf{X} , $\mathcal{C}(\mathbf{X})$.
- O vector \vec{y} pertence ao subespaço $\mathcal{C}(\mathbf{X})$
(os vectores $\vec{1}_n, \vec{x}_1, \dots, \vec{x}_p$ são colunas \mathbf{X} e $\vec{y} = b_0 \vec{1}_n + b_1 \vec{x}_1 + b_2 \vec{x}_2 + \dots + b_p \vec{x}_p$).
- $\mathcal{C}(\mathbf{X})$ é um subespaço de \mathbb{R}^n ($\mathcal{C}(\mathbf{X}) \subset \mathbb{R}^n$), mas de **dimensão $p+1$**
(se as colunas de \mathbf{X} forem **linearmente independentes**, isto é, se nenhum vector se puder escrever como combinação linear dos restantes).
- Qualquer combinação linear das colunas da matriz \mathbf{X} , ou seja, **qualquer elemento de $\mathcal{C}(\mathbf{X})$** se pode escrever como $\mathbf{X}\vec{a}$, onde $\vec{a} = (a_0, a_1, a_2, \dots, a_p)$ é o vector dos coeficientes da combinação linear.

Os parâmetros

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

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

onde os y_i são os valores observados da variável resposta e $\hat{y}_i = b_0 + b_1 x_{1(i)} + b_2 x_{2(i)} + \dots + b_p x_{p(i)}$ os **valores ajustados**. É a combinação linear que desejamos determinar.

Como identificar esse ponto/vector?

Geometria

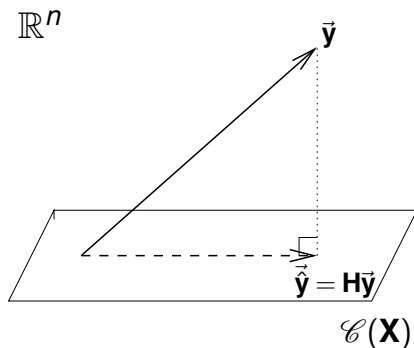
Vamos usar argumentos geométricos.

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

SOLUÇÃO:

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

A projecção ortogonal de \vec{y} sobre $\mathcal{C}(\mathbf{X})$



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

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

O critério minimiza *SQRE*

Recordar definições relativas a vectores:

- A **norma** dum vector $\vec{\mathbf{x}} = (x_1, x_2, \dots, x_n)^t$ é $\|\vec{\mathbf{x}}\| = \sqrt{\vec{\mathbf{x}}^t \vec{\mathbf{x}}} = \sqrt{\sum_{i=1}^n x_i^2}$.
- A **distância** entre dois vectores $\vec{\mathbf{x}}$ e $\vec{\mathbf{y}}$ é a norma da sua diferença:
$$\text{dist}(\vec{\mathbf{x}}, \vec{\mathbf{y}}) = \|\vec{\mathbf{x}} - \vec{\mathbf{y}}\| = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}.$$

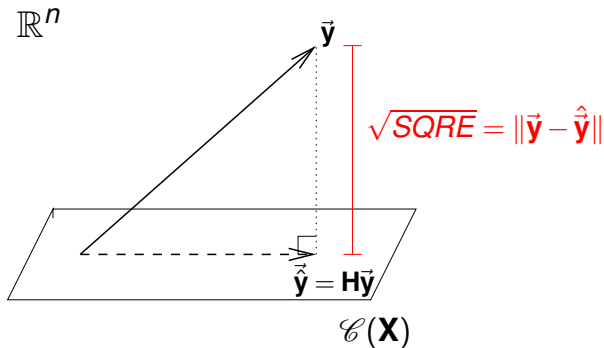
Escolher o vector $\vec{\hat{\mathbf{y}}} \in \mathcal{L}(\mathbf{X})$ que minimiza a distância ao vector de observações $\vec{\mathbf{y}}$ significa minimizar o quadrado dessa distância:

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

ou seja, **minimizar a soma de quadrados dos resíduos**.

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

SQRE na projecção ortogonal



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

A projecção ortogonal

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

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

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

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, $\mathbf{P}\mathbf{P} = \mathbf{P}$).

A matriz \mathbf{H} tem estas propriedades (verifique!).

A projecção ortogonal no contexto da RLM

No contexto duma regressão linear múltipla, tem-se:

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

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

O vector de parâmetros ajustado

$$\vec{b} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{y}.$$

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:

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

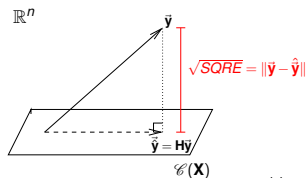
SQR A Soma de Quadrados associada à Regressão:

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

Pitágoras e a Regressão

O **Teorema de Pitágoras** aplica-se em qualquer espaço euclidiano \mathbb{R}^n . No triângulo rectângulo do acetato 82 produz a seguinte relação:

$$\|\vec{y}\|^2 = \|\vec{\hat{y}}\|^2 + \|\vec{y} - \vec{\hat{y}}\|^2$$



$$\Leftrightarrow \sum_{i=1}^n y_i^2 = \sum_{i=1}^n \hat{y}_i^2 + \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{= SQRE}$$

$$\Leftrightarrow \sum_{i=1}^n y_i^2 - n\bar{y}^2 = \sum_{i=1}^n \hat{y}_i^2 - n\bar{y}^2 + SQRE$$

$$\Leftrightarrow \mathbf{SQT} = \mathbf{SQR} + \mathbf{SQRE}$$

Revisitando Pitágoras

A relação fundamental da Regressão Linear ($SQT = SQR + SQRE$) resultou de aplicar o Teorema de Pitágoras. Mas foi necessário subtrair $n\bar{y}^2$. Outro triângulo rectângulo é estatisticamente mais interessante.

Defina-se o **vector centrado**, \vec{y}^c , cujo elemento genérico é o desvio de cada y_i em relação à média: $y_i - \bar{y}$.

$$\vec{y}^c = \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} \bar{y} \\ \bar{y} \\ \vdots \\ \bar{y} \end{bmatrix} = \vec{y} - (\bar{y})\vec{1}_n.$$

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

Revisitando Pitágoras (cont.)

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

$$\begin{aligned}\mathbf{H}\vec{y}^c &= \mathbf{H}[\vec{y} - (\bar{y}) \cdot \vec{\mathbf{1}}_n] \\ \Leftrightarrow \mathbf{H}\vec{y}^c &= \mathbf{H}\vec{y} - (\bar{y}) \cdot \mathbf{H}\vec{\mathbf{1}}_n \\ \Leftrightarrow \mathbf{H}\vec{y}^c &= \hat{\vec{y}} - (\bar{y}) \cdot \vec{\mathbf{1}}_n\end{aligned}$$

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

O vector $\mathbf{H}\vec{y}^c$ tem elemento genérico $\hat{y}_i - \bar{y}$. A sua norma é:

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

Revisitando Pitágoras (cont.)

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

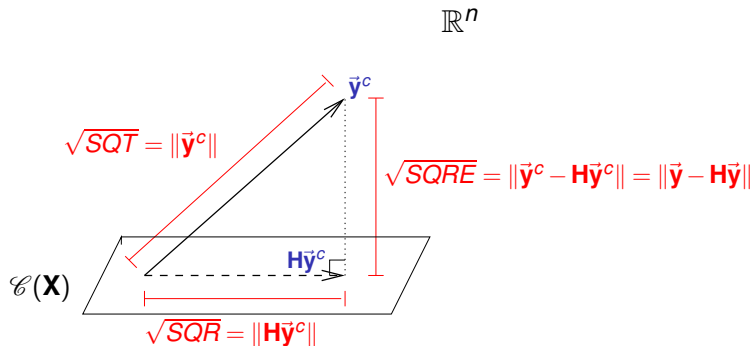
$$\begin{aligned}\vec{\mathbf{y}}^c - \mathbf{H}\vec{\mathbf{y}}^c &= [\vec{\mathbf{y}} - \cancel{\bar{y}\vec{\mathbf{1}}_n}] - [\vec{\hat{\mathbf{y}}} - \cancel{\bar{y}\vec{\mathbf{1}}_n}] \\ \Leftrightarrow \vec{\mathbf{y}}^c - \mathbf{H}\vec{\mathbf{y}}^c &= \vec{\mathbf{y}} - \vec{\hat{\mathbf{y}}}\end{aligned}$$

pelo que

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

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{y}^c e a sua projecção ortogonal sobre $\mathcal{C}(\mathbf{X})$.



Pitágoras e o Coeficiente de Determinação

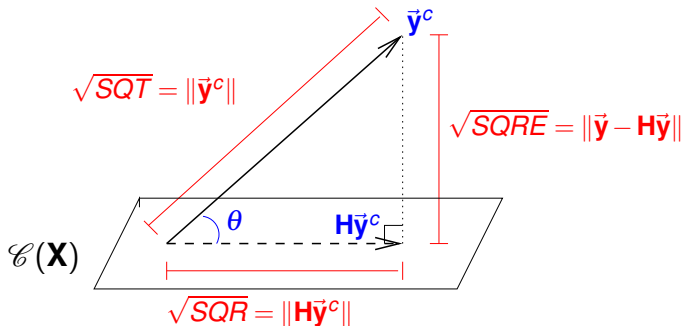
Torna-se evidente outra relação importante entre a geometria no espaço \mathbb{R}^n e a estatística da Regressão Linear:

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{y}^c , e a sua projecção ortogonal sobre o subespaço $\mathcal{C}(\mathbf{X})$:

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

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

Pitágoras e o Coeficiente de Determinação (cont.)

 \mathbb{R}^n 

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

Propriedades do Coeficiente de Determinação

A abordagem geométrica confirma que também na Regressão Linear Múltipla se verificam as propriedades 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 a correspondência entre o vector (centrado) das observações, \vec{y}^c , e o seu ajustamento em $\mathcal{C}(\mathbf{X})$.
- Se $R^2 \approx 0$, o vector \vec{y}^c é quase perpendicular ao subespaço $\mathcal{C}(\mathbf{X})$ onde se pretende aproximá-lo, e a projecção vai quase anular todas os elementos do vector projectado, ou seja, $\hat{y}_i - \bar{y} \approx 0$. **O resultado é de má qualidade**: perde-se quase toda a variabilidade nos valores de $\hat{y}_i \approx \bar{y}$.

Propriedades de modelos com constante aditiva

$\mathcal{C}(\mathbf{X})$ contém o vector $\vec{\mathbf{1}}_n$ de n uns. 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:

- As médias dos valores observados e ajustados de Y são iguais:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n \hat{y}_i = \frac{1}{n} \vec{\mathbf{1}}_n^t \vec{\hat{y}} = \frac{1}{n} \vec{\mathbf{1}}_n^t \mathbf{H}\vec{y} = \frac{1}{n} \vec{\mathbf{1}}_n^t \mathbf{H}^t \vec{y} = \frac{1}{n} (\mathbf{H}\vec{\mathbf{1}}_n)^t \vec{y} = \frac{1}{n} \vec{\mathbf{1}}_n^t \vec{y} = \bar{y}$$

- A soma dos resíduos é zero:

$$\bar{e} = \frac{1}{n} \sum_{i=1}^n e_i = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i) = \bar{y} - \bar{y} = 0.$$

- Em \mathbb{R}^{p+1} , o hiperplano ajustado contém o centro de gravidade da nuvem dos n pontos observados: $\bar{y} = b_0 + b_1 \bar{x}_1 + b_2 \bar{x}_2 + \dots + b_p \bar{x}_p$.

Já vimos que $\bar{y} = \bar{\hat{y}} = \frac{1}{n} \sum_{i=1}^n \hat{y}_i$. Mas $\frac{1}{n} \sum_{i=1}^n \hat{y}_i =$

$$\frac{1}{n} \sum_{i=1}^n (b_0 + b_1 x_{1(i)} + b_2 x_{2(i)} + \dots + b_p x_{p(i)}) = b_0 + b_1 \bar{x}_1 + b_2 \bar{x}_2 + \dots + b_p \bar{x}_p$$

Os coeficientes b_j

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:

$$\begin{aligned}\vec{\hat{\mathbf{y}}} &= \mathbf{H}\vec{\mathbf{y}} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{y}} = \mathbf{X}\vec{\mathbf{b}} \\ \Leftrightarrow \hat{y}_i &= b_0 + b_1x_{1(i)} + \dots + b_px_{p(i)}, \quad \forall i.\end{aligned}$$

As unidades de medida:

- de b_0 são iguais às de y (e às de \hat{y}).
- dos parâmetros b_j das variáveis ($j \neq 0$) são a razão entre as unidades de y e as do preditor x_j correspondente.

Os coeficientes $\{b_j\}_{j=1}^p$ das variáveis predictoras interpretam-se como a diferença (média) em y , associada a aumentar o preditor x_j correspondente em uma unidade, mantendo os restantes preditores constantes.

Resíduos

As **unidades de medida** dos **resíduos** $e_i = y_i - \hat{y}_i$ são iguais às de y :

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


O vector de **resíduos**, \vec{e} , também pode ser obtido **pré-multiplicando** o vector \vec{y} pela matriz $\mathbf{I} - \mathbf{H}$, onde \mathbf{I} é a matriz identidade $n \times n$:

$$\vec{e} = \vec{y} - \mathbf{H}\vec{y} = (\mathbf{I} - \mathbf{H})\vec{y} \quad ,$$

A matriz $\mathbf{I} - \mathbf{H}$ é **simétrica** e **idempotente**, logo também é uma matriz de **projectão ortogonal**. Projecta sobre o subespaço de \mathbb{R}^n constituído pelos **vectores ortogonais** a qualquer vector de $\mathcal{C}(\mathbf{X})$, chamado o **complemento ortogonal** de $\mathcal{C}(\mathbf{X})$ e designado por $\mathcal{C}(\mathbf{X})^\perp$.

O vector \vec{e} é a **projectão ortogonal** de \vec{y} sobre $\mathcal{C}(\mathbf{X})^\perp$.

A Regressão Linear Múltipla no

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

E.g., sendo y a variável resposta e x_1, x_2 e x_3 três preditores, a fórmula que especifica a relação será:

$$y \sim x_1 + x_2 + x_3$$

Comando R para ajustar uma regressão linear múltipla

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

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

Um exemplo de RLM no R

Ilustra-se 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 inspeção inicial dos dados pode ser feita com o comando `head`, que mostra a parte inicial da *data frame*:

```
> 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**. Será, para já, ignorado.

A Regressão Múltipla descritiva no (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), **ignorando as espécies**.

RL Múltipla - dados dos lírios

```
> 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 em \mathbb{R}^4 (\mathbb{R}^{p+1}) é:

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

Confirmando as fórmulas (cont.)

Vamos confirmar a fórmula dos parâmetros ajustados pelo método dos mínimos quadrados. O comando `model.matrix` devolve a matriz \mathbf{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 $\vec{b} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \vec{y}$ 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 101.

Modelos e submodelos

Submodelos

Dado um modelo de regressão linear múltipla, com equação

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_px_p ,$$

chama-se **submodelo** a uma regressão linear com apenas **alguns** preditores.

Por exemplo, a regressão linear **simples**

$$Petal.Width = b_0 + b_1Petal.Length$$

é um **submodelo** da regressão linear múltipla acabada de ajustar,

$$Petal.Width = b_0 + b_1Petal.Length + b_2Sepal.Length + b_3Sepal.Width$$

Nota: Um submodelo (S) não pode ter preditores que não façam parte do modelo completo (C). A variável resposta tem de ser a mesma.

O R^2 de submodelos

Coefficientes de Determinação de submodelos: $R_S^2 \leq R_C^2$

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

O subespaço das colunas do submodelo tem de estar contido no subespaço das colunas do modelo completo: $\mathcal{C}(\mathbf{X}_S) \subseteq \mathcal{C}(\mathbf{X}_C)$. Logo, o ângulo entre \vec{y} e $\vec{y}_S \in \mathcal{C}(\mathbf{X}_S)$ não pode ser menor que o ângulo entre \vec{y} e $\vec{y}_C \in \mathcal{C}(\mathbf{X}_C)$, pois \vec{y}_S também pertence a $\mathcal{C}(\mathbf{X}_C)$.

No modelo do acetato 101: $R^2 = 0.9379$.

Na RL Simples só com o preditor `Petal.Length`: $R^2 = 0.9271$.

Ainda o exemplo dos lírios

```
> summary(iris2.lm)$r.sq
```

```
[1] 0.9378503
```

```
> iris.lm <- lm(Petal.Width ~ Petal.Length, data = iris)
```

```
> summary(iris.lm)$r.sq
```

```
[1] 0.9271098
```

Equações de submodelos

Os parâmetros ajustados não são iguais

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

Ainda o exemplo dos lírios

```
> coef(iris.lm)
```

```
(Intercept) Petal.Length  
-0.3630755    0.4157554
```

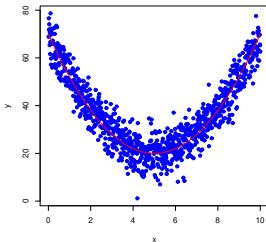
```
> coef(iris2.lm)
```

```
(Intercept) Petal.Length Sepal.Length Sepal.Width  
-0.2403074    0.5240831   -0.2072661    0.2228285
```

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:

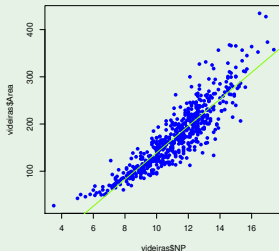


Regressão Polinomial - Exemplo

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 $X_2 = X^2$:

```
> videiras.lm2 <- lm( Area ~ NP + I(NP^2) , data=videiras )  
> videiras.lm2
```

Call:

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

Coefficients:

| (Intercept) | NP | I(NP^2) |
|-------------|---------|---------|
| 7.5961 | -0.2172 | 1.2941 |

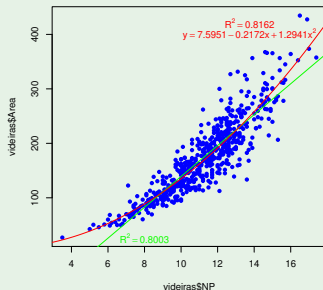
```
> summary( videiras.lm2 )$r.sq
```

```
[1] 0.8161632
```

A parábola ajustada tem equação $y = 7.5961 - 0.2172x + 1.2941x^2$. O valor $R^2 = 0.8162$ 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.)

A parábola ajustada



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

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

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

- 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

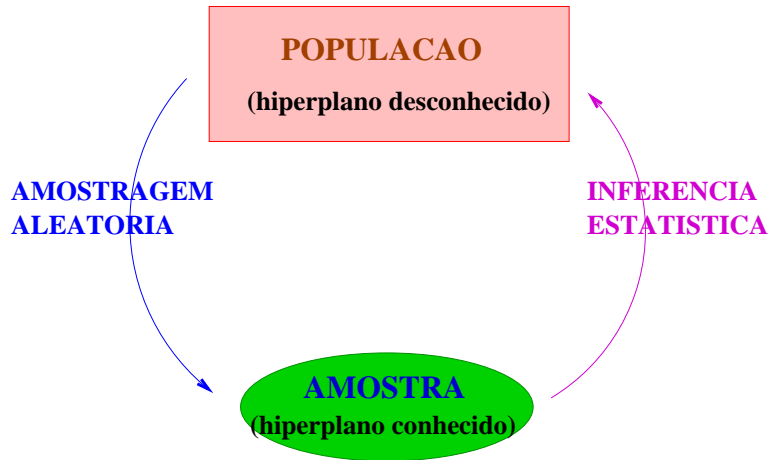
- Até aqui a regressão linear foi usada apenas como **técnica descritiva**. Se as n observações forem a totalidade da população de interesse, pouco mais há a dizer.
- Mas, com frequência, as n observações são apenas uma **amostra aleatória** de uma população maior.
- Um hiperplano ajustado a partir duma dada amostra, $y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p$, é apenas uma **estimativa** de um **hiperplano populacional**

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

Outras amostras dariam hiperplanos ajustados diferentes.

- Coloca-se o problema da **inferência estatística**.

O problema da Inferência Estatística na Reg. Linear



MODELO - Regressão Linear

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

Y – variável resposta **aleatória**.

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

O modelo será ajustado com base em:

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

MODELO RL – Linearidade

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

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

MODELO Regressão Linear – Os erros aleatórios

Vamos ainda admitir que os erros aleatórios ε_j :

- Têm **valor esperado** (valor médio) **nulo**:

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

(não é hipótese restritiva).

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

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

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

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

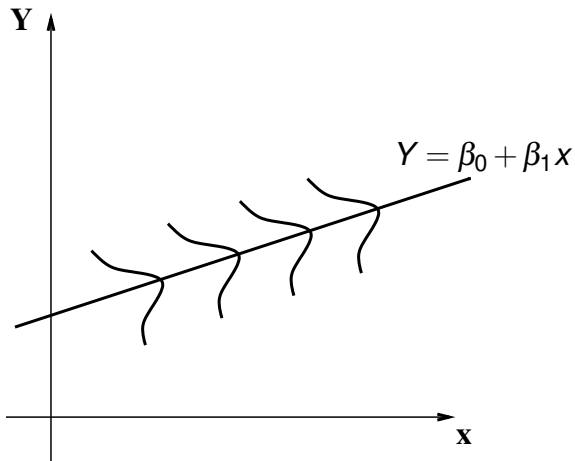
Dado o modelo, o valor esperado (médio) de Y_i , condicional aos valores x_1, x_2, \dots, x_p dos preditores, é:

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

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

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

Os **parâmetros ajustados** $\vec{b} = (b_0, b_1, b_2, \dots, b_p)$, obtidos pela fórmula do acetato 84, são **estimativas** desses parâmetros.

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

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

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

A notação matricial/vectorial

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 117) podem ser escritas como **uma única equação**, utilizando notação vectorial/matricial:

$$\begin{array}{rccccccc} Y_1 & = & \beta_0 + \beta_1 x_{1(1)} + \beta_2 x_{2(1)} + \cdots + \beta_p x_{p(1)} & + & \varepsilon_1 \\ Y_2 & = & \beta_0 + \beta_1 x_{1(2)} + \beta_2 x_{2(2)} + \cdots + \beta_p x_{p(2)} & + & \varepsilon_2 \\ Y_3 & = & \beta_0 + \beta_1 x_{1(3)} + \beta_2 x_{2(3)} + \cdots + \beta_p x_{p(3)} & + & \varepsilon_3 \\ \vdots & & \vdots & & \vdots \\ Y_n & = & \beta_0 + \beta_1 x_{1(n)} + \beta_2 x_{2(n)} + \cdots + \beta_p x_{p(n)} & + & \varepsilon_n \\ \underbrace{\hspace{1.5cm}}_{=\vec{Y}} & & \underbrace{\hspace{4.5cm}}_{=X\vec{\beta}} & & \underbrace{\hspace{1.5cm}}_{=\vec{\varepsilon}} \end{array}$$

A notação vectorial (cont.)

As n equações correspondem a **uma única equação vectorial**:

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

onde:

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

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

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

O **vector de estimadores** $\vec{\hat{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)^t$ é definido a partir da equação do vector $\vec{\mathbf{b}}$ de estimativas (acetato 84), mas substituindo o vector $\vec{\mathbf{y}}$ de valores observados de Y pelo **vector aleatório** $\vec{\mathbf{Y}}$.

Estimadores de Mínimos Quadrados dos parâmetros

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

Os estimadores assim obtidos são **estimadores de mínimos quadrados**.

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

Ferramentas para vectores aleatórios

Já se introduziram 3 **vectores aleatórios**:

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

São necessárias **ferramentas** para trabalhar com vectores aleatórios.

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

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

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

Se \mathbf{W} for 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{Z} é constituída pelas (co)variâncias de cada par de componentes:

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

É uma **matriz simétrica**.

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

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

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

A distribuição Normal Multivariada

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

Distribuição Normal Multivariada

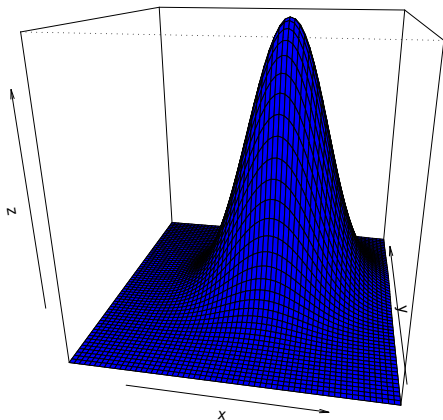
O vector aleatório k -dimensional $\vec{\mathbf{Z}}$ tem **distribuição Multinormal**, com **parâmetros** dados pelo vector $\vec{\boldsymbol{\mu}}$ e a matriz invertível $\boldsymbol{\Sigma}$ se a sua função densidade conjunta for:

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

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

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 $\vec{Z} \sim \mathcal{N}_k(\vec{\mu}, \Sigma)$:

- 1 O vector esperado de \vec{Z} é $E[\vec{Z}] = \vec{\mu}$.
- 2 A matriz de (co)variâncias de \vec{Z} é $V[\vec{Z}] = \Sigma$.
- 3 Se duas componentes de \vec{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 $\vec{Z} \sim \mathcal{N}_k(\vec{\mu}, \Sigma)$:

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

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

Modelo Regressão Linear - versão vectorial

O Modelo Linear em notação vectorial

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

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

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{Y}

O seguinte Teorema é consequência directa dos acetatos 129 e 130.

Teorema (Primeiras Consequências do Modelo)

Dado o Modelo de Regressão Linear, tem-se:

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

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

$$\vec{Y} = \underbrace{\mathbf{X}\vec{\beta}}_{= "a"} + \underbrace{\vec{\epsilon}}_{= "z"}.$$

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

A distribuição de \vec{Y} (interpretação)

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

Tendo em conta as propriedades da Multinormal:

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

O estimador dos parâmetros do Modelo

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

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

onde \mathbf{X} e $\vec{\mathbf{Y}}$ são a matriz e o vector definidos no acetato 121.

O vector $\vec{\hat{\beta}}$ 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 $\vec{\hat{\beta}}$.

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

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

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

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

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

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

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

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

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

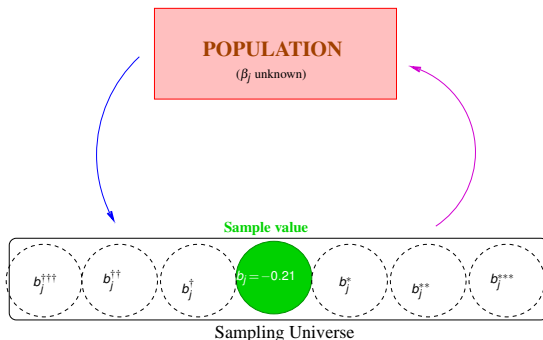
$$\vec{\hat{\beta}} \sim \mathcal{N}_{p+1}(\vec{\beta}, \sigma^2 (\mathbf{X}^t \mathbf{X})^{-1}).$$

Tendo em conta as propriedades da Multinormal (acetatos 129 e 130):

- Cada estimador individual $\hat{\beta}_j$ tem distribuição **Normal**.
- Cada estimador individual tem média $E[\hat{\beta}_j] = \beta_j$, logo é **centrado (unbiased)**.
- 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: $Cov[\hat{\beta}_i, \hat{\beta}_j] = \sigma^2 (\mathbf{X}^t \mathbf{X})^{-1}_{(i+1,j+1)}$.
- Logo, o estimador $\hat{\beta}_j$ de um parâmetro individual β_j tem distribuição $\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma_{\hat{\beta}_j}^2)$, com $\sigma_{\hat{\beta}_j}^2 = \sigma^2 (\mathbf{X}^t \mathbf{X})^{-1}_{(j+1,j+1)}$.

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

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma_{\hat{\beta}_j}^2) \quad \text{com} \quad \sigma_{\hat{\beta}_j}^2 = \sigma^2 (\mathbf{X}^t \mathbf{X})_{(j+1, j+1)}^{-1} .$$

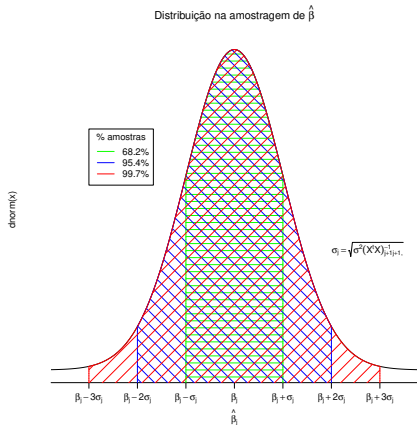


O conjunto de todas as possíveis amostras de dimensão n designa-se o **Universo de Amostragem** (Sampling Universe).

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

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

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma_{\hat{\beta}_j}^2) \quad \text{com} \quad \sigma_{\hat{\beta}_j}^2 = \sigma^2 (\mathbf{X}^t \mathbf{X})_{(j+1,j+1)}^{-1}.$$



A distribuição dum estimador individual

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

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 (\mathbf{X}^t \mathbf{X})_{(j+1, j+1)}^{-1})$$
$$\Leftrightarrow \frac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}} \sim \mathcal{N}(0, 1),$$

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

Este resultado distribucional permitiria construir intervalos de confiança ou fazer testes a hipóteses sobre os parâmetros $\vec{\beta}$, 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 quantidades não-amostrais desconhecidas, além de β_j .

Para ultrapassar este problema é preciso:

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

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

Estimando σ^2

Erros aleatórios (variáveis aleatórias – não observáveis)

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

Resíduos (variáveis aleatórias – observáveis)

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

Resíduos (observados)

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

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

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

Mas o estimador $\hat{\sigma}_{MV}^2$ não é centrado: $E \left[\hat{\sigma}_{MV}^2 \right] = E \left[\frac{SQRE}{n} \right] = \frac{n-(p+1)}{n} \sigma^2.$

O Quadrado Médio Residual

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

Quadrado Médio Residual (QMRE)

Define-se o **Quadrado Médio Residual** como

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

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

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

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

Quantidades fulcrais para a inferência sobre β_j

Teorema (Distribuições para a inferência sobre β_j)

Dado o Modelo de Regressão Linear Múltipla, tem-se

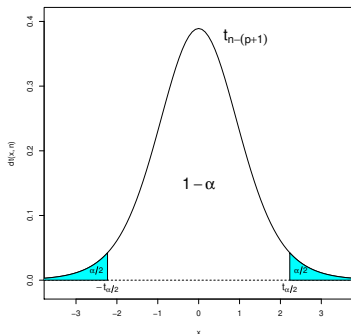
$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-(p+1)}, \quad \forall j=0, 1, \dots, p$$

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

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.

Dedução de intervalo de confiança para β_j

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



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

Dedução IC para β_j (cont.)

Trabalhar a dupla desigualdade até isolar β_j :

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

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

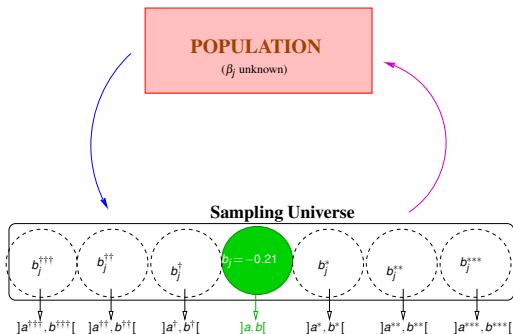
○ intervalo aleatório

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

contém β_j com probabilidade $1 - \alpha$.

Intervalo aleatório para β_j (interpretação)

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



Cada amostra no **Universo de Amostragem** gera um **intervalo concreto**, chamado **Intervalo de Confiança** (**Confidence Interval**).

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

Intervalo de confiança para β_j

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

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

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

sendo:

- b_j o elemento $j+1$ do vector das estimativas $\vec{\mathbf{b}}$ (acetato 83);
- $t_{\frac{\alpha}{2}[n-(p+1)]}$ o quantil de ordem $1 - \frac{\alpha}{2}$ da distribuição $t_{n-(p+1)}$;
- $\hat{\sigma}_{\hat{\beta}_j} = \sqrt{QMRE \cdot (\mathbf{X}^t \mathbf{X})_{(j+1,j+1)}^{-1}}$ (com o valor de QMRE na nossa amostra).

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

Intervalos de confiança para β_j no

A informação para construir intervalos de confiança para cada β_j obtém-se a partir da função `summary`. No exemplo do slide 101:

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

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

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

$$\begin{aligned} &] (-0.20727) - (1.976346)(0.04751) , (-0.20727) + (1.976346)(0.04751) [\\ & \Leftrightarrow] -0.3012 , -0.1134 [\end{aligned}$$

Intervalos de confiança para β_j no \mathbb{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
```

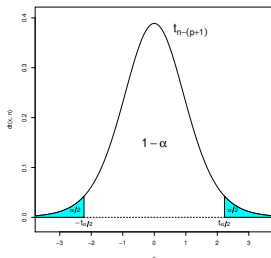
Testes de Hipóteses sobre os parâmetros

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

$$T = \frac{\hat{\beta}_j - \overbrace{\beta_j}^{=c}}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-(p+1)}, \quad \forall j=0, 1, \dots, p$$

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

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



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

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

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

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

Nível de significância do teste: α

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

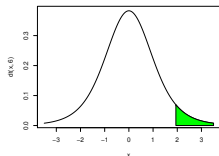
$$T_{calc} > t_{\frac{\alpha}{2}}[n-(p+1)] \quad \text{ou} \quad T_{calc} < -t_{\frac{\alpha}{2}}[n-(p+1)]$$

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

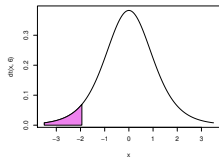
Testes de Hipóteses a $\hat{\beta}_j$ (unilaterais)

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

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



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



Testes de Hipóteses sobre os parâmetros

Dado o Modelo de Regressão Linear Múltipla,

Testes de Hipóteses a β_j (Regressão Linear Múltipla)

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

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

Nível de significância do teste: α

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

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

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

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

Combinações lineares dos parâmetros

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

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

Casos particulares importantes são se:

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

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

Inferência sobre combinações lineares dos β_j s

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

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

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

- Sabemos que $\vec{\hat{\beta}} \sim \mathcal{N}_{p+1}(\vec{\beta}, \sigma^2(\mathbf{X}^t \mathbf{X})^{-1})$ (slide 135);
- Logo, $\vec{a}^t \vec{\hat{\beta}} \sim \mathcal{N}_1(\vec{a}^t \vec{\beta}, \sigma^2 \vec{a}^t (\mathbf{X}^t \mathbf{X})^{-1} \vec{a})$ (slide 130);
- Ou seja, $\vec{Z} = \frac{\vec{a}^t \vec{\hat{\beta}} - \vec{a}^t \vec{\beta}}{\sqrt{\sigma^2 \vec{a}^t (\mathbf{X}^t \mathbf{X})^{-1} \vec{a}}} \sim \mathcal{N}(0, 1)$;
- Por um raciocínio análogo ao usado nos β_j individuais, tem-se:

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

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

Teorema (Distribuições para combinações lineares dos β s)

Dado o Modelo de Regressão Linear Múltipla, tem-se

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

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

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 $\vec{a}^t \vec{\beta}$

A estrutura análoga da quantidade *pivot* (slide 156) gera intervalos de confiança com a mesma estrutura dos IC para cada β_j .

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

Dado o Modelo de Regressão Linear Múltipla e uma amostra, o intervalo a $(1 - \alpha) \times 100\%$ de confiança para uma combinação linear dos parâmetros,

$\vec{a}^t \vec{\beta} = a_0 \beta_0 + a_1 \beta_1 + \dots + a_p \beta_p$, é:

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

com $\vec{a}^t \vec{b} = a_0 b_0 + a_1 b_1 + \dots + a_p b_p$ e $\hat{\sigma}_{\vec{a}^t \vec{\beta}} = \sqrt{QMRE \cdot \vec{a}^t (\mathbf{X}^t \mathbf{X})^{-1} \vec{a}}$.

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

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

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

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

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

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

ICs para combinações lineares no

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

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

que é gerada pelo comando R `vcov`.

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

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

ICs para combinações lineares no

O erro padrão estimado de $\hat{\beta}_2 + \hat{\beta}_3$ (fórmula do acetato 158) é:

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

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

```
> vcov(iris2.lm)
```

| | (Intercept) | Petal.Length | Sepal.Length | Sepal.Width |
|--------------|--------------|---------------|--------------|--------------|
| (Intercept) | 0.031815766 | 0.0015144174 | -0.005075942 | -0.002486105 |
| Petal.Length | 0.001514417 | 0.0005998259 | -0.001065046 | 0.000802941 |
| Sepal.Length | -0.005075942 | -0.0010650465 | 0.002256837 | -0.001344002 |
| Sepal.Width | -0.002486105 | 0.0008029410 | -0.001344002 | 0.002394932 |

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

Como caso particular do resultado anterior, tem-se:

IC para o valor esperado de Y , dados os preditores

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

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

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

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

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

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

Testes a combinações lineares dos parâmetros

Dado o Modelo de Regressão Linear Múltipla,

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

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

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

Nível de significância do teste: α

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

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

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

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

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

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

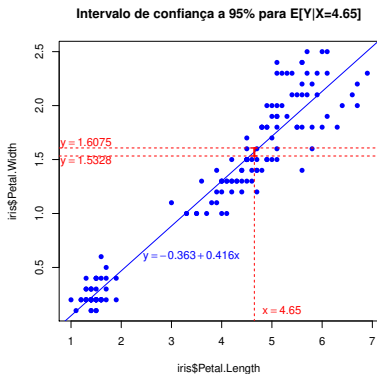
No exemplo de [Regressão Linear Simples](#) nos lírios, a largura esperada de pétalas de comprimento 1.85 e 4.65, é:

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

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

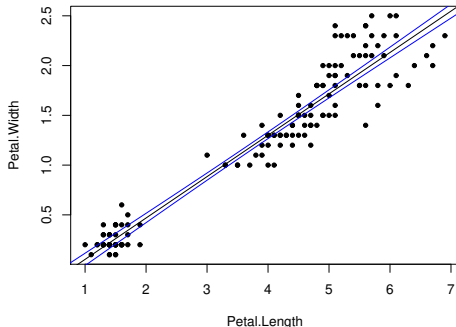
O intervalo de confiança para $\mu_{Y|\bar{x}}$ obtém-se com o argumento `int="conf"`:

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



Bandas de confiança para a recta de regressão

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



Os IC para $\mu_{Y|x}$ dependem do valor de x (fórmula do acetato 166). Terão maior amplitude quanto mais afastado x estiver da média \bar{x} das observações. Logo, as bandas são encurvadas.

Fórmulas para uma regressão linear simples

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

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

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

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

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

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

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

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

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

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

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

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

A variabilidade numa observação individual de Y

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

$$\mu_{Y|\vec{x}} = E[Y|X_1=x_1, X_2=x_2, \dots, X_p=x_p] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p,$$

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

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

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

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

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

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

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

Intervalos de predição para Y

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

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

Intervalos de **predição** para observações individuais

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

onde

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

e

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

Fórmulas para a regressão linear simples

Na regressão linear simples usa-se a fórmula do acetato 166:

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

Logo,

RLS: Intervalo de predição para observação individual de Y

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

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

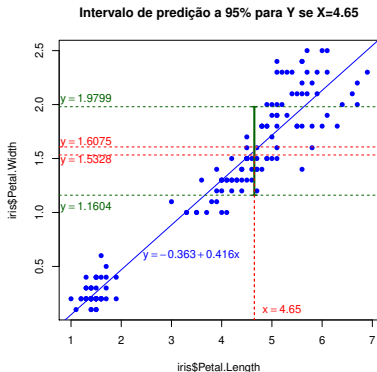
Quer numa regressão linear simples, quer numa múltipla, estes intervalos são necessariamente **de maior amplitude** que os intervalos de confiança para $\mu_{Y|\bar{x}}$ (para igual nível de confiança $(1 - \alpha) \times 100\%$).

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`:

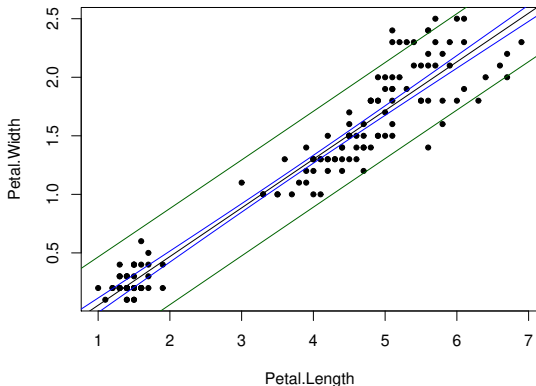
```
> predict(iris.lm,data.frame(Petal.Length=c(4.65)), int="pred")
```

| | 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 $\mu_{Y|\bar{x}}$ era] 0.4169 , 0.5077 [, necessariamente mais curto.

Testando a qualidade do ajustamento global

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

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

As hipóteses em confronto são:

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

[MODELO COMPLETO \equiv MODELO NULO]

vs.

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

[MODELO COMPLETO \neq MODELO NULO]

NOTA: repare que β_0 não intervém nas hipóteses.

O teste de ajustamento global (cont.)

Definindo:

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

Sob a Hipótese Nula do teste de ajustamento global:

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

Esta é a **estatística F** do teste de ajustamento global.

Expressão alternativa para a estatística do teste F

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

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

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

As hipóteses do teste também se podem escrever como

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

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

O Teste F de ajustamento global do Modelo

Teste F de ajustamento global do modelo RLM

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

vs.

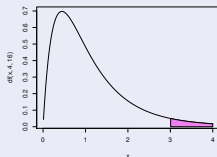
$H_1 : \exists j = 1, \dots, p$ tal que $\beta_j \neq 0$.

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

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

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



Outra formulação do teste F de ajustamento global

Teste F de ajustamento global do modelo RLM (alternativa)

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

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

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

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

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

Exemplo inferência RLM: dados Brix (Exercício 9)

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

```
> brix.lm <- lm(Brix ~ . , data=brix)           <- note-se o uso do '.'
> summary(brix.lm)
[...]
```

Coefficients:

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

--

Residual standard error: 0.1366 on 8 degrees of freedom
Multiple R-squared: 0.8483, Adjusted R-squared: 0.7534
F-statistic: 8.944 on 5 and 8 DF, p-value: 0.003942

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

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

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.,

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

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

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

Comparando modelo e submodelos

Para avaliar se um dado modelo difere significativamente dum seu submodelo (identificado pelo conjunto \mathcal{S} dos índices das suas variáveis), precisamos de optar entre as hipóteses:

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

[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. Apenas permite um melhor ajustamento.

Estatística de teste para comparar modelo/submodelo

Considere-se:

- modelo completo com p preditores e Soma de Quadrados Residual $SQRE_C$;
- submodelo com k preditores e Soma de Quadrados Residual $SQRE_S$

Dada a Hipótese Nula:

$\beta_j = 0$ para todos os preditores x_j que não pertencem ao submodelo,

tem-se:

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

Nota: O denominador $\frac{SQRE_C}{n-(p+1)}$ é o Quadrado Médio Residual do modelo completo, $QMRE_C$.

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:

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

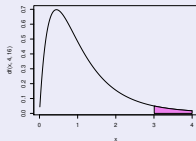
Estatística do Teste:

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

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

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

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:

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

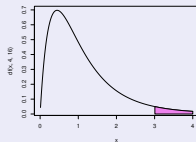
Estatística do Teste:

$$F = \frac{n-(p+1)}{p-k} \cdot \frac{R_C^2 - R_S^2}{1 - R_C^2} \sim F_{[p-k, n-(p+1)]}, \text{ sob } H_0 .$$

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

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



O teste a submodelos no

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

Nos exemplos dos lírios, temos:

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

```
Analysis of Variance Table
```

```
Model 1: Petal.Width ~ Petal.Length
```

```
Model 2: Petal.Width ~ Petal.Length + Sepal.Length + Sepal.Width
```

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

Os valores $R_s^2 = 0.9271$ e $R_c^2 = 0.9379$ 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 e o submodelo nulo (sem preditores).

Caso o modelo e submodelo difiram num único preditor, X_j , o teste F parcial é equivalente ao teste t (acetato 153) 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.

(Não) 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.

(Não) 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 |

(Não) Cuidado com exclusões simultâneas de preditores

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

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

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

(Não) Um exemplo: dados Brix (Exercício 9)

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

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

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

```
> anova(brix2.lm, brix.lm)
```

Analysis of Variance Table

Model 1: Brix ~ Diametro + Acucar

Model 2: Brix ~ Diametro + Altura + Peso + pH + Acucar

| | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) | |
|---|--------|---------|----|-----------|------|---------|---|
| 1 | 11 | 0.42743 | | | | | |
| 2 | 8 | 0.14925 | 3 | 0.27818 | 4.97 | 0.03104 | * |

(Não) 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 ² 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 **e leaps** do módulo **subselect**.

²Furnival, G.W and Wilson, R.W.,Jr. (1974) Regressions by leaps and bounds, *Technometrics*, **16**, 499-511.

(Não) 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: preditores; linhas: dimensão k de subconjunto
1 FALSE FALSE FALSE FALSE TRUE <-- k=1 ; melhor preditor individual: Acucar
2 TRUE  TRUE FALSE FALSE FALSE <-- k=2 ; melhor par de preditores: Diametro e Altura
3 TRUE  TRUE FALSE FALSE TRUE  <-- k=3 ; melhor trio de preditores: Diametro, Altura 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_S^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.


(Não) 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.

(Não) Algoritmos sequenciais com base no AIC

O  disponibiliza funções para automatizar pesquisas sequenciais de submodelos em que o critério de exclusão/inclusão dum 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 baseada na Verosimilhança**. No contexto dum **Regressão Linear Múltipla com k variáveis preditoras**, pode definir-se como

Critério de Informação de Akaike no Modelo Linear

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

É legítimo comparar os valores do AIC de diferentes modelos, **desde que ajustados com os mesmos dados e admitindo a mesma distribuição para a variável resposta Y** .

(Não) Interpretação do AIC

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

- A primeira parcela mede a qualidade do ajustamento do modelo aos dados. Quanto menor, melhor.
- A segunda parcela mede a complexidade do modelo, através do número de preditores. Quanto menor, melhor.

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).

O AIC pode ser usado para optar entre um modelo e um qualquer seu submodelo.

(Não) Algoritmos sequenciais com 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 for 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.

(Não) Algoritmos de selecção sequencial no

A função `step` corre algoritmos de selecção sequencial, com base no AIC. Considere-se de novo o exemplo dos dados `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.**

(Não) Nota 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 $\varepsilon_i \sim \mathcal{N}(0, \sigma^2) \quad \forall i = 1, \dots, n$.

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

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

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 $\mathcal{C}(\mathbf{X})$.

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

Propriedades dos Resíduos sob o modelo linear

Teorema (Distribuição dos Resíduos no Modelo Linear)

Dado o Modelo Linear, tem-se:

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

Como no Modelo Linear $\vec{Y} \sim \mathcal{N}(\mathbf{X}\vec{\beta}, \sigma^2\mathbf{I}_n)$, o vector dos resíduos $\vec{E} = (\mathbf{I}_n - \mathbf{H})\vec{Y}$, tem distribuição **Multinormal** em sentido generalizado (acetato 130).

O vector esperado de \vec{E} resulta das propriedades do acetato 125:

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

Propriedades dos Resíduos no Modelo Linear (cont.)

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

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

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

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

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

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

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

Dois tipos de resíduos

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

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

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

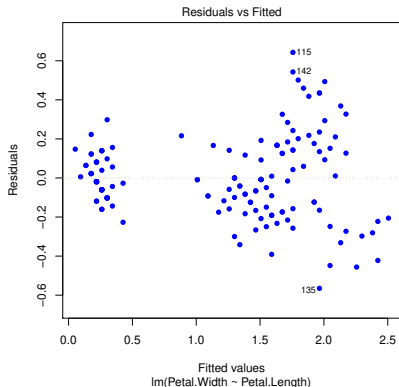
Para grandes amostras, os R_i são aproximadamente $\mathcal{N}(0, 1)$.

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

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

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

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



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

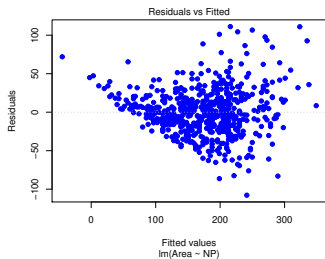
Possíveis padrões indicativos de problemas

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

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

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

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



Um exemplo de resíduos em **forma de funil**, e sugerindo alguma **curvatura** na relação entre as duas variáveis (dados das folhas de videira, Exercício 18, Area vs. NP).

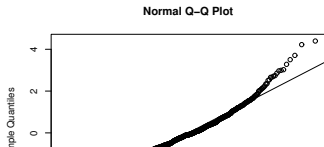
Validação do modelo: (2) Gráficos para avaliar a Normalidade

Como foi visto no acetato 204, para grandes amostras os resíduos estandardizados $R_j = \frac{E_j}{\sqrt{QMRE \cdot (1-h_{jj})}}$, devem ser aproximadamente $\mathcal{N}(0,1)$.

O pressuposto de erros aleatórios Normais pode ser validado com:

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

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



Validação do modelo: (3) Gráficos para avaliar independência

Dependência entre erros aleatórios pode surgir como resultado de:

- **correlação cronológica;**
- **correlação espacial.**

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

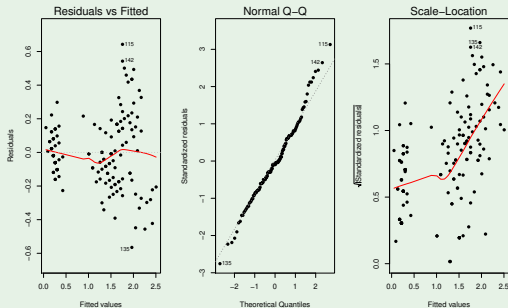
Validação do modelo: (4) Gráficos de resíduos vs. preditores

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

Estudo de resíduos no

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

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



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

Observações atípicas

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

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

Muitas vezes surgem associadas a resíduos grandes (em módulo). Em particular, e como os resíduos Studentizados têm distribuição aproximadamente $\mathcal{N}(0, 1)$ para n grande, observações para as quais $|R_i| > 3$ (ou $|T_i| > 3$) podem ser classificadas como atípicas.

Mas por vezes, observações distantes da tendência geral **podem afectar o próprio ajustamento do modelo**, e não serem facilmente identificáveis a partir dos seus resíduos.

As chamadas “observações alavanca”

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

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

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

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

Observações alavanca (cont.)

Verifica-se, para **qualquer** observação:

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

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

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

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

Observações alavanca (cont.)

Observações com um efeito alavanca elevado podem, ou não, estar dispostas com a mesma tendência de fundo que as restantes observações, i.e., podem, ou não, ser atípicas (outliers).

Efeito alavanca numa regressão linear Simples

Numa regressão linear simples, tem-se

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

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

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

Observações influentes

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

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

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

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

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

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

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

Diagnósticos no

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

Dados Brix (Exercício 9)

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

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

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

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.

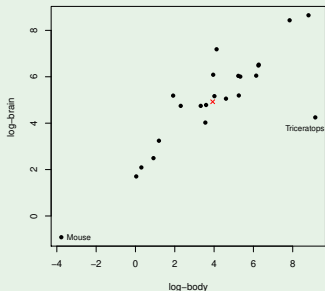
Estes diagnósticos servem sobretudo para **identificar observações que merecem maior atenção e consideração.**

Um exemplo na RLS

Dados Animals (Exercício 6)

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

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



Um exemplo na RLS (cont.)

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

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

Gráficos diagnósticos no

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

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

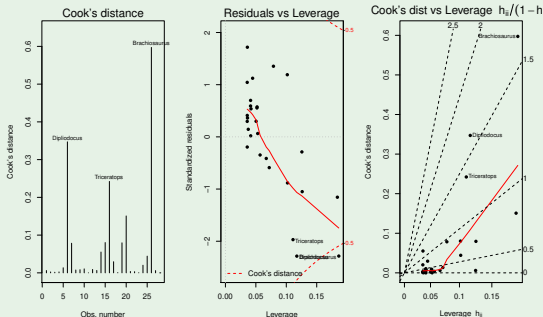
A opção `which=5` produz um gráfico de Resíduos estandardizados (R_i s) no eixo vertical contra valores de h_{ij} (*leverages*) no eixo horizontal, traçando linhas de igual distância de Cook (para os níveis 0.5 e 1, por omissão), que destacam eventuais observações influentes.

A opção `which=6` produz um gráfico de distâncias de Cook (eixo vertical) contra valores de $\frac{h_{ij}}{1-h_{ij}}$, com isolinhas de resíduos estandardizados R_i (resultantes da última fórmula do acetato 214).

Um exemplo de gráficos de diagnóstico

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

```
> 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.

O R^2 modificado (adjusted R^2)

O Coeficiente de Determinação usual define-se como:

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

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

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

Para qualquer modelo linear (com preditores), tem-se: $R_{mod}^2 < R^2$.

Se $n \gg p+1$ (muito mais observações que parâmetros), $R^2 \approx R_{mod}^2$.

Se n é pouco maior que p , $R_{mod}^2 \ll R^2$ (excepto se $R^2 \approx 1$).

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

O R^2 modificado (cont.)

Viu-se que o valor de R_{mod}^2 penaliza modelos complexos ajustados com poucas observações. Exercício 9: 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_{mod}^2 maior que um modelo completo.

Exemplo: Exercício 19

((Não) também ilustra o uso do R_{mod}^2 como critério de selecção na função 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
```

Algumas transformações de variáveis

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

Se $\text{var}(\varepsilon_j) \propto E[Y_j]$ então $Y \rightarrow \sqrt{Y}$

Se $\text{var}(\varepsilon_j) \propto (E[Y_j])^2$ então $Y \rightarrow \ln Y$

Se $\text{var}(\varepsilon_j) \propto (E[Y_j])^4$ então $Y \rightarrow 1/Y$

são propostas usuais para estabilizar as variâncias.

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

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

Prevenções sobre transformações

Mas a utilização de transformações de 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 numa amostra específica, mas **não tenham qualquer generalidade**.

Transformações linearizantes

Diferente é o problema de transformações que visam linearizar uma relação original não linear entre variável resposta e preditores. Tais transformações linearizantes também podem ser úteis em regressões lineares múltiplas.

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 $\beta_0^* = \ln(\beta_0)$):

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

Prevenções sobre transformações linearizantes

- As estimativas que minimizam a soma de quadrados dos resíduos nas relações linearizadas **não são** os que produzem **as soluções ótimas** dum problema de minimização de somas de quadrados de resíduos na relação não-linear original.
- 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.**

Advertências finais

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

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

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

Advertências finais (cont.)

2. Não se deve confundir a existência de uma relação linear entre preditores X_1, X_2, \dots, 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.

Análise de Variância (ANOVA)

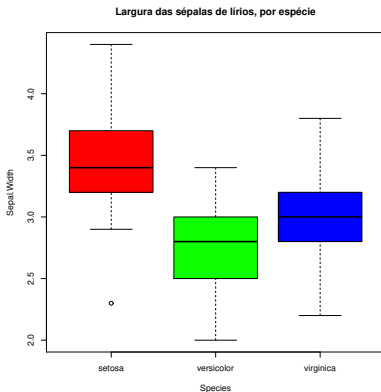
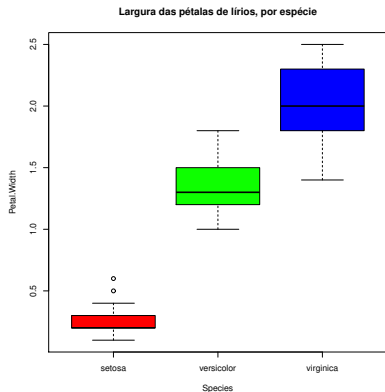
A Regressão Linear visa modelar uma variável resposta numérica (quantitativa), à custa de uma ou mais variáveis preditoras, igualmente numéricas.

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

Na ANOVA a **um factor** (*one-way ANOVA*), a modelação da variável resposta (numérica) baseia-se numa única variável preditora categórica.

Admitimos que há n observações independentes da variável resposta Y , sendo n_i ($i = 1, \dots, k$) correspondentes ao nível i do factor. Logo,

$$n_1 + n_2 + \dots + n_k = n.$$

Delineamento equilibrados e 1 Factor

No caso de igual número de observações em cada nível,

$$n_1 = n_2 = n_3 = \dots = n_k \quad (= n_c),$$

diz-se que estamos perante um **delineamento equilibrado** (*balanced design*).

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 $i=1, \dots, k$ e $j=1, \dots, n_i$).

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 da equação de base do modelo é:

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

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 a equação com uma constante aditiva comum:

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

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

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

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

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:

$$Y_{ij} = \mu + \alpha_1 \mathcal{I}_{1ij} + \alpha_2 \mathcal{I}_{2ij} + \dots + \alpha_k \mathcal{I}_{kij} + \varepsilon_{ij} ,$$

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

$$\mathcal{I}_{mij} = \begin{cases} 1 & \text{se } i = m , \\ 0 & \text{se } i \neq m . \end{cases}$$

A equação em notação vectorial

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{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.

$\vec{1}_n$ o vector de n uns, já considerado na regressão.

\vec{J}_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 ($i = 1, \dots, k$). Numa ANOVA, as variáveis indicatrizes desempenham o papel dos preditores.

$\vec{\epsilon}$ 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{J}_2 e \vec{J}_3 serão:

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

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:

$$\begin{aligned}\vec{Y} &= \mu \vec{\mathbf{1}}_n + \alpha_1 \vec{\mathcal{J}}_1 + \alpha_2 \vec{\mathcal{J}}_2 + \alpha_3 \vec{\mathcal{J}}_3 + \vec{\epsilon} \\ \Leftrightarrow \vec{Y} &= \mathbf{X}\vec{\beta} + \vec{\epsilon}.\end{aligned}$$

As colunas da matriz \mathbf{X} são o vector dos n uns e as variáveis indicatrizes. O vector dos parâmetros $\vec{\beta}$ é constituído por μ e os efeitos α_j .

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

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

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_j equivalerá a μ_j , a média do nível;
 - ▶ não se pode generalizar a situações mais complexas;
 - ▶ mais difícil de encaixar na teoria já dada do Modelo Linear.
- 2 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.
- 3 **impor a restrição $\alpha_1 = 0$: 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.

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 $i = 1$:

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

Os efeitos de nível α_i

Na equação do modelo ANOVA a um factor (acetato 235), cada α_i ($i > 1$) representa o **acréscimo** que transforma a média do primeiro nível na média do nível i :

$$\alpha_1 = 0$$

$$\alpha_2 = \mu_2 - \mu_1$$

$$\alpha_3 = \mu_3 - \mu_1$$

$$\vdots \quad \vdots \quad \vdots$$

$$\alpha_k = \mu_k - \mu_1$$

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

Esta é a **Hipótese Nula** num teste à existência, ou não, de efeitos do factor.

O modelo ANOVA a 1 factor para efeitos inferenciais

Acrescentando os restantes pressupostos do Modelo Linear:

Modelo ANOVA a um factor, com k níveis

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

- 1 $Y_{ij} = \mu_1 + \alpha_i + \varepsilon_{ij}$, $\forall i=1, \dots, k$, $\forall j=1, \dots, n_i$ (com $\alpha_1 = 0$).
- 2 $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, $\forall i, j$
- 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 efeitos α_i ($i > 1$) de cada um dos $k-1$ restantes níveis do factor. Ou seja, tem o vector de parâmetros:

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

O modelo ANOVA a um factor - notação vectorial

De forma equivalente, em notação vectorial,

Modelo ANOVA a um factor - notação vectorial

1 $\vec{Y} = \mu_1 \vec{1}_n + \alpha_2 \vec{\mathcal{I}}_2 + \alpha_3 \vec{\mathcal{I}}_3 + \dots + \alpha_k \vec{\mathcal{I}}_k + \vec{\epsilon} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$, sendo

- ▶ \vec{Y} o vector aleatório das n observações da variável resposta;
- ▶ $\vec{1}_n$ o vector de n uns;
- ▶ $\vec{\mathcal{I}}_2, \vec{\mathcal{I}}_3, \dots, \vec{\mathcal{I}}_k$ as variáveis indicatrizes dos níveis indicados;
- ▶ $\mathbf{X} = \left[\vec{1}_n \mid \vec{\mathcal{I}}_2 \mid \vec{\mathcal{I}}_3 \mid \dots \mid \vec{\mathcal{I}}_k \right]$ a matriz do modelo; e
- ▶ $\vec{\beta} = (\mu_1, \alpha_2, \alpha_3, \dots, \alpha_k)^t$.

2 $\vec{\epsilon} \sim \mathcal{N}_n(\vec{0}, \sigma^2 \mathbf{I}_n)$, sendo \mathbf{I}_n a matriz identidade $n \times n$.

Trata-se de um **Modelo Linear**, 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

$$\alpha_2 = \alpha_3 = \dots = \alpha_k = 0$$
$$\Leftrightarrow \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

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 $\vec{\mathcal{I}}_i$) são nulos.

Logo, **é possível testar esta hipótese, através dum teste F de ajustamento global do modelo** (slide 177).

Neste contexto há fórmulas específicas.

Os graus de liberdade

Numa ANOVA a um factor, o número de preditores do modelo (as variáveis indicatrizes dos níveis $j > 1$) é $p = k - 1$ e o número de parâmetros do modelo é $p + 1 = k$.

Chama-se **SQF** (de **F**actor), em vez de **SQR**, à Soma de Quadrados associada ao ajustamento do modelo.

Os **graus de liberdade** associados a cada Soma de Quadrados são:

| <u>SQxx</u> | <u>g.l.</u> |
|-------------|-------------|
| 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

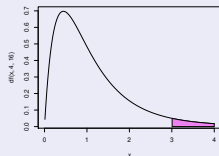
Hipóteses: $H_0 : \alpha_j = 0 \quad \forall j=2,\dots,k$ vs. $H_1 : \exists i=2,\dots,k$ t.q. $\alpha_j \neq 0$.
[FACTOR NÃO AFECTA] vs. [FACTOR AFECTA Y]

Estatística do Teste: $F = \frac{QMF}{QMRE} \sim F_{[k-1, n-k]}$ se H_0 .

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

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



As Somas de Quadrados e Quadrados Médios têm fórmulas específicas do contexto.

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

Como o modelo ANOVA é um caso particular do Modelo Linear, a fórmula dos estimadores de mínimos quadrados dos parâmetros é igualmente:

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

e o vector dos valores ajustados $\vec{\mathbf{Y}}$ resulta de projectar ortogonalmente $\vec{\mathbf{Y}}$ sobre o subespaço $\mathcal{L}(\mathbf{X})$ das colunas da matriz \mathbf{X} : $\vec{\mathbf{Y}} = \mathbf{H} \vec{\mathbf{Y}}$.

Mas a matriz do modelo \mathbf{X} tem uma natureza especial: como as suas k colunas são os vectores $\vec{\mathbf{1}}_n, \vec{\mathcal{I}}_2, \vec{\mathcal{I}}_3, \dots, \vec{\mathcal{I}}_k$, os elementos da matriz \mathbf{X} na ANOVA só tomam valores 0 e 1.

Como consequência, quer a matriz de projecções \mathbf{H} , quer o vector $\vec{\mathbf{Y}}$, têm uma natureza específica neste contexto.

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

Numa ANOVA a um factor, qualquer vector no subespaço $\mathcal{C}(\mathbf{X})$ tem de ter valores iguais para todas as observações dum mesmo nível do factor:

$$a_1 \vec{\mathbf{1}}_n + a_2 \vec{\mathcal{J}}_2 + a_3 \vec{\mathcal{J}}_3 + \dots + a_k \vec{\mathcal{J}}_k = \begin{bmatrix} a_1 \\ \dots \\ a_1 \\ \hline a_1 + a_2 \\ \dots \\ a_1 + a_2 \\ \hline a_1 + a_3 \\ \dots \\ a_1 + a_3 \\ \hline (\dots) \\ \hline a_1 + a_k \\ \dots \\ a_1 + a_k \end{bmatrix}$$

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

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

Concretamente, no vector $\vec{\hat{Y}} = \mathbf{H}\vec{Y}$, todos os valores \hat{Y}_{ij} num mesmo nível i do factor são dados pela média amostral das n_i observações Y_{ij} nesse nível:

$$\hat{Y}_{ij} = \bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij},$$

Repare-se que para minimizar a Soma de Quadrados dos Resíduos,

$$SQRE = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \hat{Y}_{ij})^2,$$

e como os valores ajustados \hat{Y}_{ij} são iguais para todas as observações num mesmo nível i do factor, minimiza-se a soma relativa a cada nível tomando: $\hat{Y}_{ij} = \bar{Y}_i$.

Os parâmetros ajustados

Os parâmetros populacionais são μ_1 e $\alpha_j = \mu_j - \mu_1$.

Os parâmetros populacionais são estimados pelas quantidades amostrais correspondentes:

Parâmetros estimados numa ANOVA a um factor

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

Os resíduos, *SQRE* e *QMRE*

Viu-se (acetato 250) que $\hat{Y}_{ij} = \hat{\mu}_i = \bar{Y}_{i.}$, pelo que o resíduo da observação Y_{ij} é dado por:

$$E_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \bar{Y}_{i.},$$

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

$$SQRE = \sum_{i=1}^k \sum_{j=1}^{n_i} E_{ij}^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 = \sum_{i=1}^k (n_i - 1) S_i^2,$$

onde $S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2$ é a variância amostral das n_i observações de Y no i -ésimo nível do factor. *SQRE* mede variabilidade **no seio** dos k níveis.

O **Quadrado Médio Residual** é uma média ponderada das variâncias de nível S_i^2 , com pesos $n_i - 1$ ($n - k = \sum_i (n_i - 1)$):

$$QMRE = \frac{SQRE}{n - k} = \frac{1}{n - k} \sum_{i=1}^k (n_i - 1) S_i^2.$$

Fórmulas para delineamentos equilibrados

No caso de um **delineamento equilibrado**, i.e., $n_1 = n_2 = \dots = n_k (= n_c)$ e $n = n_c \cdot k$, logo:

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

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

Assim, em delineamentos equilibrados, o Quadrado Médio Residual $QMRE$ é a média (simples) das k variâncias de nível, s_i^2 .

A Soma de Quadrados associada ao Factor

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

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

$$\begin{aligned} SQF &= \sum_{i=1}^k \sum_{j=1}^{n_i} (\hat{Y}_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{Y}_{i.} - \bar{Y}_{..})^2 \\ \Leftrightarrow SQF &= \sum_{i=1}^k n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2 \end{aligned}$$

SQF mede variabilidade entre as médias amostrais de cada nível.

Fórmulas para delineamentos equilibrados

No caso de um **delineamento equilibrado**,

$$SQF = n_c \sum_{i=1}^k (\bar{Y}_{i.} - \bar{Y}_{..})^2 = n_c(k-1) \cdot S_{Y_{i.}}^2,$$

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

$$QMF = \frac{SQF}{k-1} = n_c \cdot S_{Y_{i.}}^2.$$

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:

$$\sum_{i=1}^k \sum_{j=1}^{n_j} (Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^k n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2 + \sum_{i=1}^k (n_i - 1) S_i^2 .$$

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 | $k - 1$ | $SQF = \sum_{i=1}^k n_i \cdot (\bar{y}_i - \bar{y}_{..})^2$ | $QMF = \frac{SQF}{k-1}$ | $\frac{QMF}{QMRE}$ |
| 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

Para efectuar uma ANOVA a um Factor no , 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:

$$\text{Petal.Width} \sim \text{Species}$$

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

ANOVAs a um factor no (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`.

ANOVA 1 Factor (Írrios, acetato 230)

```
> 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 (cont.)

A função `summary` também pode ser aplicada ao resultado de uma ANOVA, produzindo o quadro-resumo completo da ANOVA.

ANOVA 1 Factor (lírios, acetato 230)

```
> 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, o teste F rejeita claramente a hipótese de os acréscimos de nível, α_j , serem 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 $\mu_1, \alpha_2, \alpha_3, \dots, \alpha_k$, pode aplicar-se a função `coef` ao resultado da ANOVA.

ANOVA 1 Factor (Írios, acetato 230)

```
> 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 (cont.)

As médias por nível do factor da variável resposta, podem ser obtidas com a função `model.tables` e o argumento `type="means"`:

ANOVA 1 Factor (Írios, acetato 230)

```
> model.tables(iris.aov , type="means")
```

```
Tables of means
```

```
Grand mean
```

```
1.199333
```

```
Species
```

```
Species
```

```
setosa versicolor virginica
```

```
0.246
```

```
1.326
```

```
2.026
```

O  ordena os níveis dum factor por ordem alfabética.

ANOVAs como modelo Linear no

Também é possível usar o comando `lm`, nomeadamente para fazer inferência sobre os parâmetros do modelo:

ANOVA 1 Factor (Írrios, acetato 230)

```
> 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
```

A exploração ulterior de H_1

Uma rejeição da Hipótese Nula $\alpha_2 = \alpha_3 = \dots = \alpha_k = 0 \Leftrightarrow \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$ deixa em aberto a questão de quais os pares de médias que devem ser considerados significativamente diferentes.

O estudo de quais os pares de níveis i, j para os quais se deve concluir que $\mu_i \neq \mu_j$ exige outros testes. Aconselha-se a utilização de **testes de comparações múltiplas**, a fim de controlar o **nível de significância global** de todas as $\binom{k}{2}$ comparações de pares de médias.

Entre estes, destacam-se os **testes de Tukey** e os **testes de Scheffé**.

Esta matéria não é leccionada nesta disciplina.

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 e_{ij} vs. \hat{y}_{ij} , porque qualquer valor ajustado $\hat{y}_{ij} = \bar{y}_i$ é igual para observações num mesmo nível do factor.

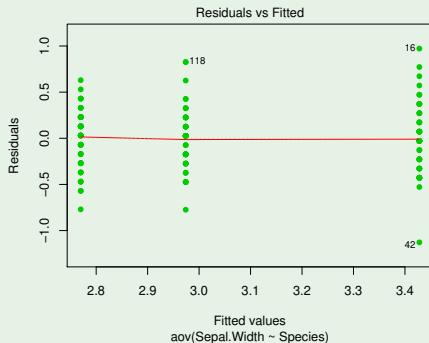
Este padrão **não** corresponde a qualquer violação dos pressupostos do modelo.

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

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

Gráfico de resíduos numa ANOVA a 1 factor

```
> plot(aov(Sepal.Width ~ Species, data=iris), which=1)
```



Violações aos pressupostos da ANOVA

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

- O teste F da ANOVA³ são relativamente robustos a desvios à hipótese de normalidade.
- As violações ao pressuposto de variâncias homogêneas 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.

³E as comparações múltiplas de Tukey.

Uma advertência

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

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

em vez de impor a condição $\alpha_1 = 0$, impõe-se a condição $\sum_{i=1}^k \alpha_i = 0$.

Esta restrição alternativa:

- Muda a forma de **interpretar** os parâmetros (μ é agora uma **média geral de Y** e α_i o desvio da média do nível i em relação a μ);
- 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 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 por ANOVA (ou regressão linear), as observações da variável resposta correspondem a n diferentes **unidades experimentais** (indivíduos, parcelas de terreno, locais, etc.). **Princípios gerais** da selecção destas unidades experimentais:

Casualização

A **casualização**, ou seja **aleatoriedade** na escolha das unidades experimentais e 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
- se **evitar enviesamentos** (mesmo inconscientes).

Repetição

A **repetição** de observações **independentes** é necessária para se **estimar a variabilidade associada à estimação** (erros padrões) e **minorar o impacte de observações atípicas**.

Repetições e pseudo-repetições

Repetições e pseudo-repetições

Há que distinguir **repetições** e **pseudo-repetições**.

Por exemplo, num estudo sobre frutos do tomateiro, é diferente:

- seleccionar frutos **dum mesmo tomateiro**; ou
- seleccionar frutos de **tomateiros diferentes**.

As características genóticas, fenotípicas e ambientais, são idênticas ou muito semelhantes para frutos **duma mesma planta**. Trata-se de **pseudo-repetições** (com valores correlacionados), que **não são repetições independentes**.

Pseudo-repetições **podem ser úteis**: substituindo cada grupo de pseudo-repetições por **uma única observação média** pode-se **diminuir a variabilidade entre diferentes observações independentes**, tornando a inferência mais precisa.

Heterogeneidade nas unidades experimentais

Variabilidade nas unidades experimentais não atribuível aos preditores é considerada variação inexplicada 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 R^2 .

Controlar a heterogeneidade

Fora de ambientes laboratoriais, é impossível tornar as unidades experimentais totalmente homogéneas: a natural variabilidade de plantas, animais, terrenos, localidades geográficas, células, etc. significa que existe variabilidade não controlada de 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 factorial design*), sendo um dos factores a **variedade de trigo** e o outro o **tipo de solos**.

Um **delineamento factorial** (*factorial design*) 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 (*two-way ANOVAs*).

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.

Representação delinearmento factorial (2 factores)

| | | Factor B | | | | |
|----------|--------|----------|-------|-------|-------|-------|
| | | B_1 | B_2 | B_3 | ... | B_b |
| FACTOR A | Níveis | | | | | |
| | A_1 | × × × | × × × | × × × | ... | × × × |
| | A_2 | × × × | × × × | × × × | ... | × × × |
| | A_3 | × × × | × × × | × × × | ... | × × × |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| A_a | × × × | × × × | × × × | ... | × × × | |

Atenção: Esta esquematização **não** corresponde a qualquer organização **espacial**.

Célula: cruzamento dum nível dum Factor com um nível do outro Factor. Corresponde a uma dada **situação experimental**.

Neste delinearmento, há **ab** situações experimentais (células), cada uma com **n_{ij}** observações.

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** ($i = 1, 2, \dots, a$).
- j indica o **nível j do Factor B** ($j = 1, 2, \dots, b$).
- k indica a **repetição k na célula (i, j)** ($k = 1, 2, \dots, n_{ij}$).

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 for igual em todas as células ($n_{ij} = n_c, \forall i, j$), estamos perante um **delineamento equilibrado**.

A modelação de Y

Num primeiro modelo, admite-se que o valor esperado de cada observação é da forma:

$$E[Y_{ijk}] = \mu_{ij} = \mu + \alpha_i + \beta_j, \quad \forall i, j, k.$$

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{Y} o vector aleatório n -dimensional com a totalidade das observações da variável resposta.
- $\vec{1}_n$ o vector de n uns.
- \mathcal{I}_{A_i} a variável indicatriz de pertença ao nível i do Factor A.
- \mathcal{I}_{B_j} a variável indicatriz de pertença ao nível j do Factor B.
- $\vec{\varepsilon}$ o vector aleatório dos n erros aleatórios.

Uma primeira equação-base em notação vectorial

Se se admitem efeitos para **todos** os níveis de ambos os factores, temos a equação-base:

$$\vec{Y} = \mu \vec{1}_n + \alpha_1 \vec{J}_{A_1} + \alpha_2 \vec{J}_{A_2} + \dots + \alpha_a \vec{J}_{A_a} + \beta_1 \vec{J}_{B_1} + \beta_2 \vec{J}_{B_2} + \dots + \beta_b \vec{J}_{B_b} + \vec{\epsilon}$$

A matriz **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, $\vec{1}_n$;
- a soma das indicatrizes do Factor B daria a coluna dos uns, $\vec{1}_n$.

A matriz do modelo X na primeira tentativa

$$\mathbf{X} = \left[\begin{array}{c|ccc|ccc|ccc}
 1 & 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\
 1 & 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\
 1 & 1 & 0 & \dots & 0 & 0 & 1 & \dots & 0 \\
 \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
 1 & 1 & 0 & \dots & 0 & 0 & 0 & \dots & 1 \\
 1 & 1 & 0 & \dots & 0 & 0 & 0 & \dots & 1 \\
 \hline
 1 & 0 & 1 & \dots & 0 & 1 & 0 & \dots & 0 \\
 1 & 0 & 1 & \dots & 0 & 1 & 0 & \dots & 0 \\
 1 & 0 & 1 & \dots & 0 & 1 & 0 & \dots & 0 \\
 \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
 1 & 0 & 1 & \dots & 0 & 0 & 0 & \dots & 1 \\
 1 & 0 & 1 & \dots & 0 & 0 & 0 & \dots & 1 \\
 \hline
 \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
 \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
 \hline
 1 & 0 & 0 & \dots & 1 & 1 & 0 & \dots & 0 \\
 \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
 \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
 1 & 0 & 0 & \dots & 1 & 0 & 0 & \dots & 1 \\
 1 & 0 & 0 & \dots & 1 & 0 & 0 & \dots & 1
 \end{array} \right]$$

$\uparrow \vec{\mathbf{1}}_n$ $\uparrow \vec{\mathbf{A}}_1$ $\uparrow \vec{\mathbf{A}}_2$ \dots $\uparrow \vec{\mathbf{A}}_a$ $\uparrow \vec{\mathbf{B}}_1$ $\uparrow \vec{\mathbf{B}}_2$ \dots $\uparrow \vec{\mathbf{B}}_b$

Nem mesmo a exclusão da coluna $\vec{\mathbf{1}}_n$ resolve o problema.

Restrições à equação do modelo

Doravante, admitimos que foram **excluídas do modelo as parcelas associadas ao primeiro nível de cada Factor**, isto é:

$$\alpha_1 = 0 \quad \text{e} \quad \beta_1 = 0 ,$$

o que corresponde a **excluir as colunas** $\vec{\mathcal{J}}_{A_1}$ e $\vec{\mathcal{J}}_{B_1}$ da matriz \mathbf{X} .

A equação-base do modelo ANOVA a 2 Factores, sem interação, fica:

$$\vec{Y} = \mu \vec{1}_n + \alpha_2 \vec{\mathcal{J}}_{A_2} + \dots + \alpha_a \vec{\mathcal{J}}_{A_a} + \beta_2 \vec{\mathcal{J}}_{B_2} + \dots + \beta_b \vec{\mathcal{J}}_{B_b} + \vec{\varepsilon}$$

O parâmetro μ é agora o **valor esperado de Y para observações da célula** ($i=1, j=1$), que passamos a representar por μ_{11} :

$$Y_{11k} = \mu + \varepsilon_{11k} \quad \Rightarrow \quad E[Y_{11k}] = \mu = \mu_{11} .$$

A matriz do delineamento na ANOVA a 2 Factores (sem interacção)

$$\mathbf{X} = \begin{bmatrix}
 1 & 0 & \dots & 0 & 0 & \dots & 0 \\
 1 & 0 & \dots & 0 & 0 & \dots & 0 \\
 1 & 0 & \dots & 0 & 1 & \dots & 0 \\
 \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
 1 & 0 & \dots & 0 & 0 & \dots & 1 \\
 1 & 0 & \dots & 0 & 0 & \dots & 1 \\
 \hline
 1 & 1 & \dots & 0 & 0 & \dots & 0 \\
 1 & 1 & \dots & 0 & 0 & \dots & 0 \\
 1 & 1 & \dots & 0 & 0 & \dots & 0 \\
 \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
 1 & 1 & \dots & 0 & 0 & \dots & 1 \\
 1 & 1 & \dots & 0 & 0 & \dots & 1 \\
 \hline
 \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
 \hline
 1 & 0 & \dots & 1 & 0 & \dots & 0 \\
 \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
 1 & 0 & \dots & 1 & 0 & \dots & 1 \\
 1 & 0 & \dots & 1 & 0 & \dots & 1
 \end{bmatrix}$$

\uparrow \uparrow \dots \uparrow \uparrow \dots \uparrow
 $\mathbf{1}_n$ \mathcal{A}_2 \dots \mathcal{A}_a \mathcal{B}_2 \dots \mathcal{B}_b

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) ($i = 1, \dots, a; j = 1, \dots, b$). Tem-se:

1 $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \varepsilon_{ijk}$, $\forall i=1, \dots, a; j=1, \dots, b; k=1, \dots, n_{ij}$ ($\alpha_1 = 0$; $\beta_1 = 0$).

2 $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$, $\forall i, j, k$

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: $H_0 : \alpha_i = 0, \quad \forall i = 2, \dots, a ;$
- Teste II: $H_0 : \beta_j = 0, \quad \forall j = 2, \dots, b.$

Teste aos efeitos do Factor B

O modelo ANOVA a 2 Factores, sem interacção (slide 285) tem equação (vectorial) de base:

$$\vec{Y} = \mu \vec{1}_n + \alpha_2 \vec{J}_{A_2} + \dots + \alpha_a \vec{J}_{A_a} + \beta_2 \vec{J}_{B_2} + \dots + \beta_b \vec{J}_{B_b} + \vec{\epsilon}$$

Sendo um Modelo Linear, há teoria para testar as hipóteses:

$$H_0 : \beta_j = 0, \quad \forall j = 2, \dots, b \quad \text{vs.} \quad H_1 : \exists j \text{ tal que } \beta_j \neq 0 .$$

Trata-se dum teste F parcial comparando o modelo completo de $a + b - 1$ parâmetros:

$$(\text{Modelo } M_{A+B}) \quad Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \varepsilon_{ijk} ,$$

com o submodelo, de a parâmetros e equação de base:

$$(\text{Modelo } M_A) \quad Y_{ijk} = \mu_{11} + \alpha_i + \varepsilon_{ijk} ,$$

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 modelo (M_{A+B}) e submodelo (M_A).
- Obter as matrizes de projecção $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ de cada modelo.
- Obter os vectores projectado $\vec{\mathbf{Y}} = \mathbf{H}\vec{\mathbf{Y}}$ e de resíduos $\vec{\mathbf{E}} = (\mathbf{I} - \mathbf{H})\vec{\mathbf{Y}}$ de cada modelo.
- Obter as Somas de Quadrados Residuais, $SQRE_{A+B}$ e $SQRE_A$.
- Efectuar o teste F parcial indicado, com a estatística de teste:

$$\text{(Efeitos Factor B)} \quad F = \frac{\overbrace{SQRE_A - SQRE_{A+B}}^{=SQB}}{b-1}}{\frac{SQRE_{A+B}}{n-(a+b-1)}} = \frac{QMB}{QMRE}$$

$$\text{definindo } QMB = \frac{SQB}{b-1} = \frac{SQRE_A - SQRE_{A+B}}{b-1}$$

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: $H_0 : \beta_j = 0 \quad \forall j=2,\dots,b$ vs. $H_1 : \exists j=2,\dots,b$ t.q. $\beta_j \neq 0$.

[B NÃO AFECTA Y] vs. [B AFECTA Y]

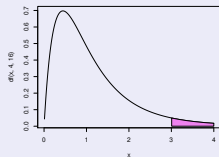
Estatística do Teste: $F = \frac{QMB}{QMRE} \sim F_{(b-1, n-(a+b-1))}$ se H_0 .

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 estatística do teste aos efeitos do Factor A

De forma análoga,

- $SQA = SQF_A$, 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 $QMRE = \frac{SQRE_{A+B}}{n-(a+b-1)}$.

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 $i=2, \dots, a$.

O Teste F aos efeitos do factor A

Sendo válido o Modelo ANOVA a dois factores, sem interacção:

Teste F aos efeitos do factor A

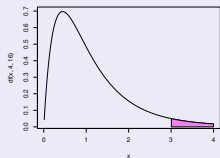
Hipóteses: $H_0 : \alpha_j = 0 \quad \forall j=2,\dots,a$ vs. $H_1 : \exists j=2,\dots,a \text{ t.q. } \alpha_j \neq 0.$
[A NÃO AFECTA Y] vs. [A AFECTA Y]

Estatística do Teste: $F = \frac{QMA}{QMRE} \sim F_{[a-1, n-(a+b-1)]}$ se H_0 .

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

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



A nova decomposição de SQT

Tendo em conta as Somas de Quadrados antes definidas, tem-se:

$$SQB = SQRE_A - SQRE_{A+B}$$

$$SQA = SQF_A = SQT - SQRE_A$$

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.

Aviso: Trocando a ordem dos factores

A troca do papel dos factores A e B define Somas de Quadrados de forma diferente. Designando por M_B o modelo ANOVA a um factor, com o factor B, resulta:

$$\begin{aligned}SQB &= SQF_B = SQT - SQRE_B \\SQA &= SQRE_B - SQRE_{A+B}.\end{aligned}$$

Continua a ser verdade que SQT se pode decompor na forma

$$SQT = SQA + SQB + SQRE_{A+B}.$$

Justificam-se testes análogos aos dos slides 291 e 289.

As duas definições alternativas de SQA e SQB só são iguais com delineamentos equilibrados. Só nesse caso a ordem dos factores é arbitrária.

SQA e SQB em delineamentos equilibrados

Num delineamento equilibrado, SQA e SQB são ambas Somas de Quadrados do Factor de Modelos só com um Factor (A ou B, slide 290).

Logo, na fórmula para $SQA = SQF_A$, (slide 254), tem-se $\hat{Y}_{ijk} = \bar{Y}_{i..}$ onde $\bar{Y}_{i..}$ indica a média de Y no nível i do factor A. Sendo $\bar{Y}_{...}$ a média global das n observações de Y , tem-se:

$$SQF_A = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_c} (\hat{Y}_{ijk} - \bar{Y}_{...})^2 = bn_c \cdot \sum_{i=1}^a (\bar{Y}_{i..} - \bar{Y}_{...})^2 = SQA.$$


Da mesma forma, $SQB = SQF_B$ define-se com base nos valores ajustados pelo Modelo M_B , apenas com o Factor B, sendo $\hat{Y}_{ijk} = \bar{Y}_{.j.}$. Logo:

$$SQF_B = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_c} (\hat{Y}_{ijk} - \bar{Y}_{...})^2 = an_c \cdot \sum_{j=1}^b (\bar{Y}_{.j.} - \bar{Y}_{...})^2 = SQB.$$


O quadro-resumo da ANOVA a 2 Factores (sem interacção; delineamento equilibrado)

| Fonte | g.l. | SQ | QM | f_{calc} |
|----------|-------------------|---|---------------------------------------|--------------------|
| Factor A | $a - 1$ | $SQA = b n_c \cdot \sum_{i=1}^a (\bar{y}_{i..} - \bar{y}_{...})^2$ | $QMA = \frac{SQA}{a-1}$ | $\frac{QMA}{QMRE}$ |
| Factor B | $b - 1$ | $SQB = a n_c \cdot \sum_{j=1}^b (\bar{y}_{.j.} - \bar{y}_{...})^2$ | $QMB = \frac{SQB}{b-1}$ | $\frac{QMB}{QMRE}$ |
| 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

Para efectuar uma ANOVA a dois Factores (sem interacção) no , 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  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 **+**:

$$y \sim fA + fB$$

Um exemplo

Dados immer de cevada (*package* MASS)

O rendimento de cinco variedades (*manchuria*, *svansota*, *velvet*, *trebi* e *peatland*) foi registado em seis localidades ^a. Em cada localidade foi semeada uma parcela com cada variedade (com casualização).

```
> 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 | | |

^aDados em Immer, Hayes & LeRoy Powers, Statistical adaptation of barley varietal adaptation, Journal of the American Society for Agronomy, 26, 403-419, 1934.

A interpretação dos parâmetros

A interpretação do significado dos parâmetros do modelo depende da convenção usada para resolver o problema da multicolinearidade das colunas da matriz \mathbf{X} .

Vejamos a interpretação dos parâmetros resultante da convenção $\alpha_1 = \beta_1 = 0$.

Uma observação de Y efectuada na célula $(1, 1)$, correspondente ao cruzamento do primeiro nível de cada factor, será da forma:

$$Y_{11k} = \mu_{11} + \underbrace{\alpha_1}_{=0} + \underbrace{\beta_1}_{=0} + \varepsilon_{11k} \quad \implies \quad E[Y_{11k}] = \mu_{11}$$

O parâmetro μ_{11} corresponde ao valor esperado da variável resposta Y na célula cujas indicatrizes foram excluídas da matriz do delineamento.

A interpretação dos parâmetros α_j

Uma observação de Y efectuada na célula $(i, 1)$, com $i > 1$ (cruzamento dum nível do factor A diferente do primeiro, com o primeiro nível do Factor B) é da forma:

$$Y_{i1k} = \mu_{11} + \alpha_i + \underbrace{\beta_1}_{=0} + \varepsilon_{i1k} \quad \implies \quad \mu_{i1} = E[Y_{i1k}] = \mu_{11} + \alpha_i$$

O parâmetro $\alpha_j = \mu_{j1} - \mu_{11}$ corresponde ao **acréscimo** no valor esperado da variável resposta Y associado a observações do nível $i > 1$ do Factor A (relativamente às observações do primeiro nível do Factor A), quando $j=1$. Designa-se o **efeito do nível i do factor A**.

Interpretação dos parâmetros α_j

Tabela com médias populacionais de célula (situação experimental):

| | | Factor B | | | | |
|----------|----------------------------------|----------------------------------|------------|------------|------------|------------|
| | | B_1 | B_2 | B_3 | ... | B_b |
| FACTOR A | Níveis | | | | | |
| | A_1 | μ_{11} | μ_{12} | μ_{13} | ... | μ_{1b} |
| | A_2 | $\mu_{21} = \mu_{11} + \alpha_2$ | μ_{22} | μ_{23} | ... | μ_{2b} |
| | A_3 | $\mu_{31} = \mu_{11} + \alpha_3$ | μ_{32} | μ_{33} | ... | μ_{3b} |
| | \vdots | \vdots | \vdots | \vdots | \ddots | \vdots |
| A_a | $\mu_{a1} = \mu_{11} + \alpha_a$ | μ_{a2} | μ_{a3} | ... | μ_{ab} | |

A interpretação dos parâmetros β_j

Uma observação de Y efectuada na célula $(1, j)$, com $j > 1$ (cruzamento do primeiro nível do factor A com um nível do Factor B diferente do primeiro) é da forma:

$$Y_{1jk} = \mu_{11} + \underbrace{\alpha_1}_{=0} + \beta_j + \varepsilon_{1jk} \quad \implies \quad \mu_{1j} = E[Y_{1jk}] = \mu_{11} + \beta_j$$

O parâmetro $\beta_j = \mu_{1j} - \mu_{11}$ corresponde ao **acréscimo** no valor esperado da variável resposta Y associado a observações do nível j do Factor B (relativamente às observações do primeiro nível do Factor B), quando $i=1$. Designa-se o **efeito do nível j do factor B**.

Interpretação dos parâmetros β_j

Tabela com médias populacionais de célula (situação experimental):

| | | Factor B | | | | |
|----------|----------|------------|---------------------------------|---------------------------------|----------|---------------------------------|
| | | B_1 | B_2 | B_3 | ... | B_b |
| FACTOR A | A_1 | μ_{11} | $\mu_{12} = \mu_{11} + \beta_2$ | $\mu_{13} = \mu_{11} + \beta_3$ | ... | $\mu_{1b} = \mu_{11} + \beta_b$ |
| | A_2 | μ_{21} | μ_{22} | μ_{23} | ... | μ_{2b} |
| | A_3 | μ_{31} | μ_{32} | μ_{33} | ... | μ_{3b} |
| | \vdots | \vdots | \vdots | \vdots | \ddots | \vdots |
| | A_a | μ_{a1} | μ_{a2} | μ_{a3} | ... | μ_{ab} |

Observações de Y no caso geral

Mas este modelo é pouco flexível: não existem mais parâmetros e os valores esperados nas restantes células já estão fixados.

Para observações de Y efectuadas numa célula genérica (i, j) , com $i > 1$ e $j > 1$, tem-se:

$$Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \varepsilon_{ijk} \quad \implies \quad \mu_{ij} = E[Y_{ijk}] = \mu_{11} + \alpha_i + \beta_j.$$

Todas as parcelas destes valores esperados de Y já foram usados. Não há flexibilidade para descrever situações específicas de células com $i > 1$ e $j > 1$.

Um modelo sem efeitos de interacção é utilizado sobretudo quando existe uma única observação em cada célula, i.e., $n_{ij} = 1, \forall i, j$.

Fórmulas para delineamentos equilibrados

Sejam:

$\bar{Y}_{i..}$ a média amostral das $b n_c$ observações do nível i do Factor A,

$$\bar{Y}_{i..} = \frac{1}{b n_c} \sum_{j=1}^b \sum_{k=1}^{n_c} Y_{ijk}$$

$\bar{Y}_{.j.}$ a média amostral das $a n_c$ observações do nível j do Factor B,

$$\bar{Y}_{.j.} = \frac{1}{a n_c} \sum_{i=1}^a \sum_{k=1}^{n_c} Y_{ijk}$$

$\bar{Y}_{...}$ a média amostral da totalidade das $n = a b n_c$ observações,

$$\bar{Y}_{...} = \frac{1}{n} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_c} Y_{ijk}.$$

Se o delineamento é equilibrado, ou seja, $n_{ij} = n_c, \forall i, j$, tem-se:

- $\hat{\mu}_{11} = \bar{Y}_{1..} + \bar{Y}_{.1.} - \bar{Y}_{...}$
- $\hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{1..}$
- $\hat{\beta}_j = \bar{Y}_{.j.} - \bar{Y}_{.1.}$

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:

$$\hat{Y}_{ijk} = \hat{\mu}_{11} + \hat{\alpha}_i + \hat{\beta}_j = \bar{Y}_{i..} + \bar{Y}_{.j.} - \bar{Y}_{...}, \quad \forall i, j, k$$

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 mesma situação experimental (célula (i, j)).

Modelos com efeitos de interacção

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 efeitos α_i e β_j designam-se agora efeitos principais de cada Factor.

Os valores esperados de Y_{ijk} (modelo com interacção)

Vamos admitir as seguintes restrições aos parâmetros:

$$\alpha_1 = 0 \quad ; \quad \beta_1 = 0 \quad ; \quad (\alpha\beta)_{1j} = 0, \forall j \quad ; \quad (\alpha\beta)_{i1} = 0, \forall i.$$

| Níveis | | Factor B | | | | |
|----------|-------|----------|-------|-------|-----|-------|
| | | B_1 | B_2 | B_3 | ... | B_b |
| FACTOR A | A_1 | × × × | × × × | × × × | ... | × × × |
| | A_2 | × × × | × × × | × × × | ... | × × × |
| | A_3 | × × × | × × × | × × × | ... | × × × |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | A_a | × × × | × × × | × × × | ... | × × × |

Apenas as observações que **não** são da primeira coluna e/ou primeira linha têm **parcelas correspondentes aos efeitos de interacção**.

Apenas observações que **não** estão associadas a A_1 têm **efeitos** α_j .

Apenas observações que **não** estão associadas a B_1 têm **efeitos** β_j .

Os valores esperados de Y_{ijk} (modelo com interacção)

Com as restrições, 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 $j > 1$,
 $\mu_{ij} = E[Y_{ijk}] = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij}$.

Existem neste modelo **ab** parâmetros:

- uma média da primeira célula, μ_{11} ;
- $a-1$ efeitos principais de nível do factor A, α_i ($i > 1$);
- $b-1$ efeitos principais de nível do factor B, β_j ($j > 1$);
- $(a-1)(b-1)$ efeitos de interacção, $(\alpha\beta)_{ij}$ ($i > 1, j > 1$).

Variáveis indicatrizes de célula

A equação-base do modelo ANOVA a 2 Factores, com interacção, define-se recorrendo a **variáveis indicatrizes de células** com $i > 1$ e $j > 1$, $\vec{\mathcal{I}}_{A_i:B_j}$:

$$\vec{Y} = \mu \vec{\mathbf{1}}_n + \alpha_2 \vec{\mathcal{I}}_{A_2} + \dots + \alpha_a \vec{\mathcal{I}}_{A_a} + \beta_2 \vec{\mathcal{I}}_{B_2} + \dots + \beta_b \vec{\mathcal{I}}_{B_b} + \\ + (\alpha\beta)_{22} \vec{\mathcal{I}}_{A_2:B_2} + (\alpha\beta)_{23} \vec{\mathcal{I}}_{A_2:B_3} + \dots + (\alpha\beta)_{ab} \vec{\mathcal{I}}_{A_a:B_b} + \vec{\epsilon}$$

A matriz do modelo \mathbf{X} tem agora 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, $\vec{\mathcal{I}}_{A_i}$, ($i > 1$), associadas aos parâmetros α_i .
- $b-1$ colunas de indicatrizes de nível do factor B, $\vec{\mathcal{I}}_{B_j}$, ($j > 1$), associadas aos parâmetros β_j .
- $(a-1)(b-1)$ colunas de indicatrizes de célula, $\vec{\mathcal{I}}_{A_i:B_j}$, ($i, j > 1$), associadas aos efeitos de interacção $(\alpha\beta)_{ij}$.

Os três testes ANOVA

Neste delineamento, desejamos fazer um teste à existência de cada um dos três tipos de efeitos:

- $H_0 : (\alpha\beta)_{ij} = 0, \quad \forall i = 2, \dots, a, \forall j = 2, \dots, b ;$
- $H_0 : \alpha_j = 0, \quad \forall j = 2, \dots, a ;$ e
- $H_0 : \beta_j = 0, \quad \forall j = 2, \dots, b .$

As estatísticas de teste para cada um destes testes obtêm-se a partir da decomposição da Soma de Quadrados Total em parcelas convenientes.

Como em modelos anteriores, $\vec{Y} = \mathbf{H}\vec{Y}$, sendo \mathbf{H} a matriz que projecta ortogonalmente sobre o espaço $\mathcal{L}(\mathbf{X})$ gerado pelas colunas da matriz \mathbf{X} .

E também: $SQRE = \|\vec{Y} - \hat{Y}\|^2 = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (Y_{ijk} - \hat{Y}_{ijk})^2.$

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) ($i = 1, \dots, a$; $j = 1, \dots, b$). Tem-se:

- 1 $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$, $\forall i=1, \dots, a$; $j=1, \dots, b$; $k=1, \dots, n_{ij}$
com $\alpha_1 = 0$; $\beta_1 = 0$; $(\alpha\beta)_{ij} = 0$ se $i = 1$ e/ou $j = 1$.
- 2 $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$
- 3 $\{\varepsilon_{ijk}\}_{i,j,k}$ v.a.s independentes.

O modelo tem ab parâmetros desconhecidos.

Testando efeitos de interacção

Para testar a existência de efeitos de interacção,

$$H_0 : (\alpha\beta)_{ij} = 0, \quad \forall i = 2, \dots, a, \quad \forall j = 2, \dots, b,$$

pode efectuar-se um teste F parcial comparando o modelo

$$\text{(Modelo } M_{A*B}) \quad Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk},$$

com o submodelo (2 factores, sem efeitos de interacção):

$$\text{(Modelo } M_{A+B}) \quad Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \varepsilon_{ijk},$$

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, $H_0 : \beta_j = 0, \quad \forall j = 2, \dots, b$, pode partir-se dos modelos

$$\begin{array}{ll} \text{(Modelo } M_{A+B}) & Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \varepsilon_{ijk} \\ \text{(Modelo } M_A) & Y_{ijk} = \mu_{11} + \alpha_i + \varepsilon_{ijk}, \end{array}$$

e tomar:

$$\begin{aligned} SQB &= SQRE_A - SQRE_{A+B} \\ SQA &= SQF_A = SQT - SQRE_A \end{aligned}$$

Nota: Estas duas Somas de Quadrados definem-se de forma idêntica à usada no modelo sem efeitos de interacção.

A decomposição de SQT

Até aqui definimos :

$$SQAB = SQRE_{A+B} - SQRE_{A*B}$$

$$SQB = SQRE_A - SQRE_{A+B}$$

$$SQA = SQF_A = SQT - SQRE_A$$

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} .

A tabela de síntese

Com base na decomposição do acetato 314 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 **graus de liberdade de cada tipo de efeitos** são o número de parâmetros desse tipo que sobram após a imposição das restrições.

Os **graus de liberdade 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]

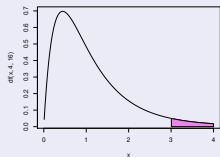
Estatística do Teste: $F = \frac{QMAB}{QMRE} \sim F_{[(a-1)(b-1), n-ab]}$ se H_0 .

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

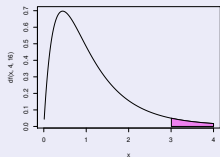
Hipóteses: $H_0 : \alpha_j = 0 \quad \forall j=2,\dots,a$ vs. $H_1 : \exists i=2,\dots,a$ t.q. $\alpha_j \neq 0$.
[\nexists EFEITOS DE A] vs. [\exists EFEITOS DE A]

Estatística do Teste: $F = \frac{QMA}{QMRE} \sim F_{[a-1, n-ab]}$ 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

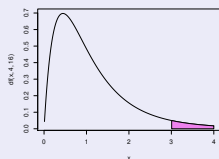
Hipóteses: $H_0 : \beta_j = 0 \quad \forall j=2,\dots,b$ vs. $H_1 : \exists j=2,\dots,b$ t.q. $\beta_j \neq 0$.
[\nexists EFEITOS DE B] vs. [\exists EFEITOS DE B]

Estatística do Teste: $F = \frac{QMB}{QMRE} \sim F_{[b-1, n-ab]}$ 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

Para efectuar uma ANOVA a dois Factores, com interacção, no , 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  para indicar uma ANOVA a dois Factores, com interacção, recorrem ao símbolo `*`:

$$y \sim fA * fB$$

sendo `y` o nome da variável resposta e `fA` e `fB` os nomes dos factores.

Um exemplo dum modelo 2 factores com interacção

Dados rendimento casta Negra Mole

Estudo de **selecção de génotipos** da casta Negra Mole (factor **clone**) com bom **rendimento** (variável resposta **rend**) ao longo dos **anos** (factor **ano**).

```
> NegraMole.aov <- aov(rend ~ ano*clone, data=NegraMole)
```

```
> summary(NegraMole.aov)
```

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------|-----|--------|---------|---------|--------------|
| ano | 4 | 203.61 | 50.90 | 77.460 | < 2e-16 *** |
| clone | 6 | 26.39 | 4.40 | 6.694 | 1.41e-06 *** |
| ano:clone | 24 | 18.08 | 0.75 | 1.146 | 0.294 |
| Residuals | 245 | 161.00 | 0.66 | | |

Há claros efeitos de **ano** e de **clone**. Não há efeitos significativos de **interacção** o que, para a selecção de génotipos é bom (comportamento previsível).

Ainda o exemplo Negra Mole

Dados rendimento casta Negra Mole

As **médias** geral, por ano, por genótipo e por célula (cruzamento ano \times genótipo) obtêm-se com o comando `model.tables`.

```
> model.tables(NegraMole.aov, type="means")
```

```
Tables of means
```

```
Grand mean
```

```
2.2237          <-- rendimento médio global
```

```
ano
```

```
LOU94 LOU95 LOU96 LOU97 LOU98
```

```
1.033 2.786 3.378 2.425 1.496          <-- 96 foi bom ano, 94 e 98 maus
```

```
clone
```

```
NM0307 NM0507 NM0703 NM1006 NM2001 NM2015 NM2102
```

```
2.4410 1.7295 2.2294 1.8306 2.2362 2.5246 2.5747          <-- há diferenças significativas
```

```
ano:clone
```

```
clone
```

```
ano      NM0307 NM0507 NM0703 NM1006 NM2001 NM2015 NM2102
```

```
LOU94 1.465 0.710 0.675 0.814 1.409 0.949 1.209          <-- Ano mau é mau para todos os genótipos.
```

```
LOU95 2.994 2.290 2.783 2.310 2.557 3.619 2.947
```

```
LOU96 3.786 2.784 3.472 2.653 3.205 3.587 4.160          <-- Ano bom é bom para todos os genótipos.
```

```
LOU97 2.728 1.728 2.667 2.272 2.547 2.205 2.831          0 que se passa em cada célula é bastante
```

```
LOU98 1.233 1.135 1.550 1.105 1.463 2.263 1.727          previsível, dado não haver interação.
```

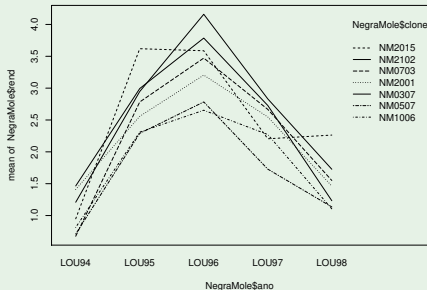
(No) Visualização gráfica de efeitos de interacção

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).

Gráfico de interação para Negra Mole

- > `attach(NegraMole)`
- > `interaction.plot(x.factor=ano, trace.factor=clone, response=rend)`
- > `detach(NegraMole)`



A ausência de interação significativa reflecte-se em “curvas aproximadamente paralelas”.

Dados do Exercício ANOVA 7 (sapotis)

Variável resposta: Conteúdo de taninos na polpa

Factor: temperatura de conservação (alta/baixa)

Factor: tempo de armazenamento (0/3/6/9 dias)

Dados Sapoti (Exercício ANOVA 7)

```
> sapoti.aov <- aov(taninos ~ temperatura * tempo , data=sapoti)
> summary(sapoti.aov)
```

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) | |
|-------------------|----|--------|---------|---------|----------|-----|
| temperatura | 1 | 206.0 | 206.0 | 238.6 | 5.72e-14 | *** |
| tempo | 3 | 288.0 | 96.0 | 111.2 | 3.27e-14 | *** |
| temperatura:tempo | 3 | 968.0 | 322.7 | 373.7 | < 2e-16 | *** |
| Residuals | 24 | 20.7 | 0.9 | | | |

Todos os tipos de efeitos são claramente significativos.

(No) Havendo interacção significativa, as linhas dum gráfico de interacção estarão longe de qualquer paralelismo

Exercise ANOVA 7 (cont.)

Conteúdo taninos nos sapotis

As **médias** global, por temperatura, por tempo de armazenamento e por célula podem obter-se com o comando `model.tables`.

```
> model.tables(sapoti.aov , type="means")
```

```
Tables of means
```

```
Grand mean
```

```
22.14375 <-- conteúdo médio global
```

```
temperatura
```

```
alta baixa
```

```
24.681 19.606 <-- conteúdo médio por temperatura
```

```
tempo
```

```
0 3 6 9
```

```
25.862 23.825 20.987 17.900 <-- conteúdo médio por tempo de armazenamento
```

```
temperatura:tempo
```

```
tempo
```

```
temperatura 0 3 6 9
```

```
alta 19.50 26.85 25.97 26.40
```

```
<-- em geral maior, excepto sem armazenamento
```

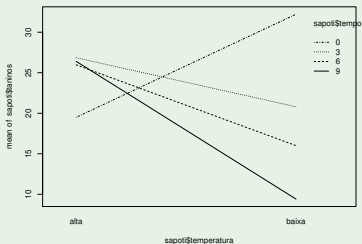
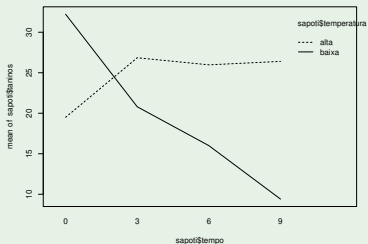
```
baixa 32.22 20.80 16.00 9.40
```

```
<-- em geral menor, excepto sem armazenamento
```

(No) Gráficos com interação

Dados Sapoti (Exercício ANOVA 7)

```
> attach(sapoti)
> interaction.plot(response=taninos,x.factor=tempo,trace.factor=temperatura)
> interaction.plot(response=taninos,x.factor=temperatura,trace.factor=tempo)
> detach(sapoti)
```



A significância da interação tem de ser **garantida através do respectivo teste F** .

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 $n - ab$. Com 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, $QMRE = \frac{SQRE}{n - ab}$.

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

\bar{Y}_{ij} a média amostral das n_{ij} observações da célula (i, j) ,

$\bar{Y}_{i..}$ a média amostral das $\sum_j n_{ij}$ observações do nível i do Factor A,

$\bar{Y}_{.j}$ a média amostral das $\sum_i n_{ij}$ observações do nível j do Factor B,

$\bar{Y}_{...}$ 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**:

$$\hat{Y}_{ijk} = \bar{Y}_{ij.}$$

Estimadores de parâmetros

Os estimadores dos parâmetros num modelo ANOVA a 2 Factores, com interacção, são:

- $\hat{\mu}_{11} = \bar{Y}_{11}.$
- $\hat{\alpha}_i = \bar{Y}_{i1.} - \bar{Y}_{11.} \quad (i > 1)$
- $\hat{\beta}_j = \bar{Y}_{1j.} - \bar{Y}_{11.} \quad (j > 1)$
- $(\hat{\alpha}\hat{\beta})_{ij} = (\bar{Y}_{ij.} + \bar{Y}_{11.}) - (\bar{Y}_{i1.} + \bar{Y}_{1j.}) \quad (i, j > 1).$

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 médias amostrais da célula onde se efectuaram as observações, $\hat{Y}_{ijk} = \bar{Y}_{ij.}$, tem-se:

$$SQRE = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (Y_{ijk} - \hat{Y}_{ijk})^2 = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y}_{ij.})^2$$

$$\Leftrightarrow SQRE = \sum_{i=1}^a \sum_{j=1}^b (n_{ij} - 1) S_{ij}^2,$$

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 :

$$QMRE = \frac{SQRE}{n - ab} = \frac{\cancel{n_c} \uparrow}{ab(\cancel{n_c} \downarrow)} \sum_{i=1}^a \sum_{j=1}^b S_{ij}^2 = \frac{1}{ab} \sum_{i=1}^a \sum_{j=1}^b S_{ij}^2.$$

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 são iguais às das Somas de Quadrados correspondentes num modelo sem efeitos de interacção:

$$SQA = bn_c \sum_{i=1}^a (\bar{Y}_{i..} - \bar{Y}_{...})^2$$

$$SQB = an_c \sum_{j=1}^b (\bar{Y}_{.j.} - \bar{Y}_{...})^2$$

Uma advertência

Na formulação clássica do modelo ANOVA a dois Factores, com interação, e a partir da equação-base $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$, em vez de impor as condições $\alpha_1 = \beta_1 = (\alpha\beta)_{11} = (\alpha\beta)_{1j} = 0$ ($\forall i, j$), admite-se a existência de acréscimos de todos os tipos para qualquer valor de i e j e impõe-se as condições:

- $\sum_i \alpha_i = 0$;
- $\sum_j \beta_j = 0$;
- $\sum_i (\alpha\beta)_{ij} = 0$, $\forall j$;
- $\sum_j (\alpha\beta)_{ij} = 0$, $\forall i$.

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.

(No) Comentários finais sobre ANOVA

1. Um delineamento factorial pode ser definido com qualquer número de factores.

Num delineamento factorial a três factores (Factores A, B e C, com a, b e c níveis) há abc situações experimentais, todas com observações.

Cada observação indexa-se com quatro índices: Y_{ijkl} indica a observação l na célula (i, j, k) . Na equação de base para Y_{ijkl} há sete tipos de efeitos:

- três efeitos principais de cada factor, α_i , β_j e γ_k .
- três efeitos de interação 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 interação nas células onde se cruzam níveis dos três factores: $(\alpha\beta\gamma)_{ijk}$

Para evitar um excesso de parâmetros, Consideram-se nulos os efeitos em que pelo menos um índice é igual a 1.

(No) 1. O modelo factorial a três factores

A equação de base do modelo é agora da forma:

$$Y_{ijkl} = \mu_{1111} + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkl} .$$

Com as restrições, o modelo tem *abc* parâmetros.

A Soma de Quadrados Total é agora decomposta em *oito parcelas*:

$$SQT = SQA + SQB + SQC + SQAB + SQAC + SQBC + SQABC + SQRE .$$

As sete *SQs* associadas a efeitos são *definidas pela diferença das Somas de Quadrados Residuais de modelos onde se vão sucessivamente omitindo os efeitos correspondentes*.

Há *sete testes*: um para cada tipo de efeitos. As estatísticas dos sete testes são todas do tipo $F = \frac{QM_x}{QMRE}$, onde *x* designa o tipo de efeitos a ser testado.

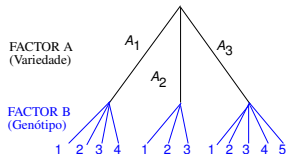
(No) 2. Delineamentos hierarquizados (nested designs)

São delineamentos com dois (ou mais) factores, em que os níveis de um dos factores variam consoante os níveis do outro factor.

Exemplo: dois factores, variedades e génotipos.

Um delineamento factorial é impossível.

Mas pode considerar-se uma estrutura hierárquica, representada no dendrograma à direita.



A equação base do modelo inclui efeitos de nível do Factor A e efeitos de nível do factor B, subordinado a A:

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk} .$$

Não faz sentido falar em efeitos do nível j do Factor B, sem especificar qual o nível do Factor A a que nos referimos. Nem faz sentido falar em efeitos de interacção: os níveis de cada factor não são, em geral, cruzados.

Haverá agora dois testes F : um para cada tipo de efeitos (α_i e $\beta_{j(i)}$). As estatísticas de teste obtêm-se de forma análoga, a partir da decomposição

$$SQT = SQA + SOB(A) + SOB$$

(No) 3. Outros tipos de delineamentos experimentais

Existem numerosos outros tipos de delineamentos mais complexos.

Alguns delineamentos visam reduzir o número de situações experimentais que é necessário estudar.

Exemplo: **quadrados latinos** ou **greco-romanos**.

Outros delineamentos visam ultrapassar dificuldades práticas na execução de uma experiência, como é o caso dos delineamentos em **parcelas divididas** (*split plots*).

(No) 4. Métodos não paramétricos de tipo ANOVA

Uma forma alternativa de estudar problemas análogos aos objectivos de ANOVAs resulta da utilização de **métodos não paramétricos**:

- Não exigem pressupostos tão restritivos como os métodos clássicos, (e.g., a Normalidade ou homogeneidade de variâncias).
- Em contrapartida têm 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.
- Frequentemente, substituem os valores observados da variável resposta pelas **ordens (ranks)** dessas observações. As estatísticas de teste são então funções dessas ordens.

(No) 4. Métodos não paramétricos de tipo ANOVA (cont.)

O teste de Kruskal-Wallis é uma alternativa não paramétrica à ANOVA a 1 Factor, em que:

- 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).
- 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, havendo uma distribuição exacta e uma distribuição assintótica para grandes amostras.

O teste de Kruskal-Wallis é equivalente a um teste ANOVA a um Factor sobre as ordens das observações.

Análise de Covariância: uma introdução

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 variáveis preditoras são **factores** (variáveis qualitativas, ou categóricas).
- Numa **Análise de Covariância**, entre as variáveis preditoras encontramos **quer variáveis numéricas, quer factores**.

Comparando rectas de regressão em diferentes níveis dum factor

A Análise de Covariância será discutida num contexto específico frequentar e de interesse prático, associado à Regressão Linear.

Pretende-se comparar as rectas de regressão linear entre uma variável numérica Y e um preditor numérico x , em vários contextos definidos pelos níveis dum dado factor.

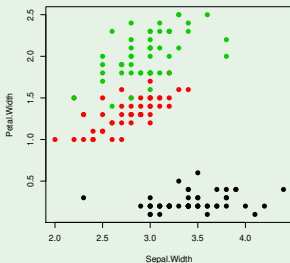
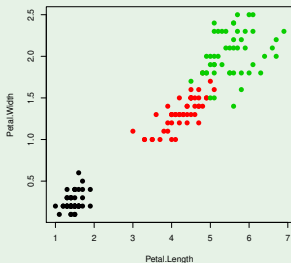
Assim, tem-se:

- uma variável resposta numérica Y ;
- um preditor numérico x ;
- um factor preditor, que define os diferentes contextos onde se deseja comparar a relação linear entre Y e x .

Um exemplo

Previendo largura de pétalas nos lírios - com espécies

Previendo a largura das pétalas **com base no seu comprimento** (à esquerda) era bom modelo para as três espécies em conjunto. E em separado?



Previendo a largura das pétalas **com base na largura das sépalas** (à direita) é um mau modelo para as três espécies em conjunto. E em separado?

A Análise de Covariância como ferramenta

O problema será formulado admitindo:

- um modelo ANCOVA que corresponde a uma relação linear entre Y e x específica para cada nível do factor;
- diferentes submodelos correspondem a admitir que alguns parâmetros dessas rectas são iguais em diferentes níveis do factor.

Tratando-se de Modelos Lineares, a teoria de que já dispomos permitirá optar entre o modelo completo e cada submodelo nesta Análise de Covariância.

Exemplifica-se o problema admitindo (como no exemplo) $k = 3$ níveis do factor. Mas a abordagem é automaticamente extensível a qualquer número $k \in \mathbb{N}$ de níveis.

A Análise de Covariância para o exemplo dado

Admita-se uma relação linear entre a variável resposta Y e o preditor x , que pode ser diferente em cada um dos 3 níveis do **factor espécies** dos lírios:

- Contexto 1: $Y = \beta_0 + \beta_1 x + \varepsilon$
- Contexto 2: $Y = \beta_0^* + \beta_1^* x + \varepsilon$
- Contexto 3: $Y = \beta_0^{**} + \beta_1^{**} x + \varepsilon$

Considere-se o primeiro contexto como **nível de referência** e escrevam-se os parâmetros dos contextos restantes à custa dos primeiros:

$$\begin{aligned} \beta_0^* &= \beta_0 + \alpha_{0:2} & ; & & \beta_1^* &= \beta_1 + \alpha_{1:2} \\ \beta_0^{**} &= \beta_0 + \alpha_{0:3} & ; & & \beta_1^{**} &= \beta_1 + \alpha_{1:3} \end{aligned}$$

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

Fazem-se n observações para ajustar o modelo, sendo n_i em cada nível ($i = 1, 2, 3$). Como na ANOVA a um factor, use-se a **dupla indexação** para identificar os níveis de origem: Y_{ij} e x_{ij} .

Tomem-se as **variáveis indicatrizes** $\vec{\mathcal{J}}_i$ de pertença aos níveis.

Definam-se ainda **vectores com os valores do preditor x num dado nível i ($i > 1$) e zero noutras posições**, representados por $\vec{x} \circ \vec{\mathcal{J}}_i$.

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

$$\vec{\mathcal{J}}_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \vec{x} \circ \vec{\mathcal{J}}_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ x_{21} \\ x_{22} \\ x_{23} \\ x_{23} \\ 0 \\ 0 \end{bmatrix}, \quad \vec{\mathcal{J}}_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}, \quad \vec{x} \circ \vec{\mathcal{J}}_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ x_{31} \\ x_{32} \end{bmatrix}$$

A equação do modelo ANCOVA

Podemos agora escrever a relação de base entre o vector \vec{Y} das n observações da variável resposta, e o preditor X , da seguinte forma:

$$\vec{Y} = \beta_0 \vec{1}_n + \beta_1 \vec{x} + \alpha_{0:2} \vec{\mathcal{J}}_2 + \alpha_{0:3} \vec{\mathcal{J}}_3 + \alpha_{1:2} (\vec{x} \circ \vec{\mathcal{J}}_2) + \alpha_{1:3} (\vec{x} \circ \vec{\mathcal{J}}_3) + \vec{\epsilon}.$$

No exemplo, e usando a notação vectorial/matricial $\vec{Y} = X\vec{\beta} + \vec{\epsilon}$:

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \\ Y_{31} \\ Y_{32} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & 0 & 0 & 0 & 0 \\ 1 & x_{12} & 0 & 0 & 0 & 0 \\ 1 & x_{13} & 0 & 0 & 0 & 0 \\ 1 & x_{21} & 1 & 0 & x_{21} & 0 \\ 1 & x_{22} & 1 & 0 & x_{22} & 0 \\ 1 & x_{23} & 1 & 0 & x_{23} & 0 \\ 1 & x_{24} & 1 & 0 & x_{24} & 0 \\ 1 & x_{31} & 0 & 1 & 0 & x_{31} \\ 1 & x_{32} & 0 & 1 & 0 & x_{32} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \alpha_{0:2} \\ \alpha_{0:3} \\ \alpha_{1:2} \\ \alpha_{1:3} \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{24} \\ \epsilon_{31} \\ \epsilon_{32} \end{bmatrix}$$

A equação do modelo ANCOVA

O modelo do slide 345 ajusta, às observações de cada um dos três contextos, uma recta de regressão distinta.

$$Y_{ij} = \begin{cases} \beta_0 + \beta_1 x_{1j} + \varepsilon_{1j}, & \text{se } i = 1 \\ (\beta_0 + \alpha_{0:2}) + (\beta_1 + \alpha_{1:2}) x_{2j} + \varepsilon_{2j}, & \text{se } i = 2 \\ (\beta_0 + \alpha_{0:3}) + (\beta_1 + \alpha_{1:3}) x_{3j} + \varepsilon_{3j}, & \text{se } i = 3. \end{cases} \quad (1)$$

Caso os parâmetros de acréscimo $\alpha_{i:j}$ sejam *todos* iguais a zero, a recta de regressão é a mesma, para os três contextos.

Com os pressupostos usuais sobre os erros aleatórios, este modelo ANCOVA é um **modelo linear** com $3 \times 2 = 6$ parâmetros (e variáveis predictoras \vec{x} , $\vec{\mathcal{I}}_2$, $\vec{\mathcal{I}}_3$, $\vec{x} \circ \vec{\mathcal{I}}_2$, $\vec{x} \circ \vec{\mathcal{I}}_3$).

Em geral, para k níveis do factor haverá $2k$ parâmetros.

Alguns submodelos interessantes

$$\vec{Y} = \beta_0 \vec{1}_n + \beta_1 \vec{x} + \alpha_{0:2} \vec{\mathcal{J}}_2 + \alpha_{0:3} \vec{\mathcal{J}}_3 + \alpha_{1:2} (\vec{x} \circ \vec{\mathcal{J}}_2) + \alpha_{1:3} (\vec{x} \circ \vec{\mathcal{J}}_3) + \vec{\varepsilon}$$

- A hipótese **duma única recta nos 3 contextos** é a hipótese $\alpha_{0:2} = \alpha_{0:3} = \alpha_{1:2} = \alpha_{1:3} = 0$.
- A hipótese de **três rectas paralelas** (i.e., declive igual), mas podendo ter diferentes ordenadas na origem, é a hipótese $\alpha_{1:2} = \alpha_{1:3} = 0$.
- 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, $\alpha_{1:2} - \alpha_{1:3} = 0$.
- A hipótese de **três rectas com igual ordenada na origem**, mas talvez declives diferentes, é a hipótese $\alpha_{0:2} = \alpha_{0:3} = 0$.

Estas hipóteses (ou outras análogas) podem ser testadas através de testes F e t – *Student* já vistos no estudo geral do modelo linear.

Cruzando factores com variáveis numéricas no

No R, um modelo ANCOVA de regressão de y sobre x , admitindo rectas diferentes para cada nível do **factor** f , é dado pela fórmula: $y \sim x * f$.

ANCOVA com os lírios

```
> modespecie.lm <- lm(Petal.Length ~ Sepal.Length * Species, data=iris)
> 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

A recta das *setosa* (nível de referência) tem declive significativamente diferente das outras espécies.

Um exemplo no \mathbb{R} . As 3 rectas.

ANCOVA com os lírios (cont.)

As três rectas ajustadas pelo modelo ANCOVA:

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.

As 3 rectas em regressões lineares separadas

ANCOVA com os lírios (cont.)

As três rectas ajustadas pelo modelo ANCOVA:

Espécie *Setosa*: $PL = 0.8031 + 0.1316 SL$

Espécie *Versicolor*: $PL = 0.1851 + 0.6865 SL$

Espécie *Virginica*: $PL = 0.6105 + 0.7501 SL$

As três rectas em regressões lineares separadas:

```
> coef(lm(Petal.Length ~ Sepal.Length , data=iris[1:50,]))
```

```
(Intercept) Sepal.Length  
0.8030518    0.1316317
```

```
> coef(lm(Petal.Length ~ Sepal.Length , data=iris[51:100,]))
```

```
(Intercept) Sepal.Length  
0.1851155    0.6864698
```

```
> coef(lm(Petal.Length ~ Sepal.Length , data=iris[101:150,]))
```

```
(Intercept) Sepal.Length  
0.6104680    0.7500808
```

A matriz \mathbf{H} por blocos na ANCOVA

Esta igualdade resulta da **estrutura especial** da matriz de projecção ortogonal \mathbf{H} , associada ao modelo ANCOVA do slide 345.

Seja \mathbf{H}_i a matriz de projecção ortogonal $n_i \times n_i$ sobre o subespaço $\mathcal{C}(X_{[i]}) \subset \mathbb{R}^{n_i}$ gerado apenas pelas observações do nível i do factor.

A matriz \mathbf{H} do modelo ANCOVA é então uma matriz **diagonal por blocos** (se as linhas de \mathbf{X} estiverem arrumadas por nível do factor):

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{H}_k \end{bmatrix}$$

Os valores ajustados do vector $\vec{\mathbf{Y}} = \mathbf{H}\vec{\mathbf{Y}}$ só dependem da matriz \mathbf{H}_i do respectivo nível i .

Um exemplo no . Recta única?

Uma recta única para as três espécies é admissível?

```
> modunico.lm <- lm(Petal.Length ~ Sepal.Length, data=iris)
```

```
> 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.

Outro exemplo no . Rectas paralelas?

Mo , uma regressão de y sobre x com rectas paralelas, mas admitindo diferentes ordenadas na origem para cada nível do factor f , é dado pela fórmula: $y \sim x + f$

Modelo para rectas paralelas nos lírios

```
> modparalelas.lm <- lm(Petal.Length ~ Sepal.Length + Species, data=iris)
> 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 \mathbb{R} : as 3 rectas paralelas

Rectas paralelas nos lírios

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 . Rectas paralelas? (cont.)

Mas é admissível considerar as três rectas sejam paralelas?

Vamos fazer um teste F parcial, comparando o submodelo de rectas paralelas e o modelo de rectas diferentes.

Teste F parcial para validar rectas paralelas

```
> 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.

Um alerta sobre os pressupostos

Os testes anteriormente referidos são válidos caso se verifiquem os pressupostos já admitidos nos Modelos Lineares:

- $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \forall i, j;$
- erros aleatórios independentes.

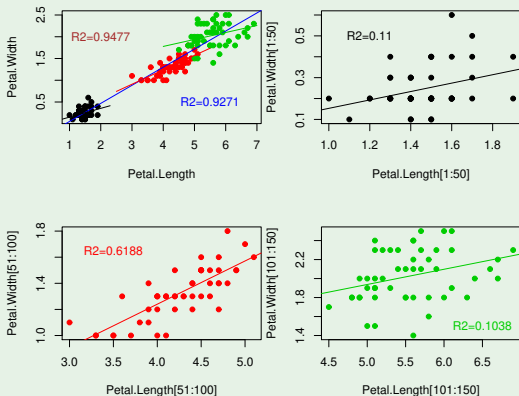
Trata-se **quase** dos mesmos pressupostos que seria necessário supor para ajustar cada recta, de forma separada, usando apenas as n_i observações de cada contexto.

Aqui há **pressupostos adicionais**: a independência e a homogeneidade das variâncias dos erros aleatórios têm de ser válidas **no conjunto** dos 3 contextos comparados.

Uma prevenção

Misturar subpopulações pode criar ilusões

Eis as nuvens de pontos e os R^2 de **largura sobre comprimento de pétala** (**Petal.Width vs. Petal.Length**): **única**, **ANCOVA** e separada, para as três espécies de lírios (**setosa**, **versicolor** e **virginica**).



O R^2 dum modelo ANCOVA

É possível relacionar os Coeficiente de Determinação do modelo ANCOVA, R^2 , e de k modelos dum único nível, $R_{[i]}^2$. Tem-se:

$$R^2 = \frac{\sum_{i=1}^k R_{[i]}^2 SQT_i + SQF}{\sum_{i=1}^k SQT_i + SQF}.$$

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 k 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 $R_{[i]}^2$ (com pesos SQT_i). Neste caso, $R^2 \approx 1$ só se a generalidade dos $R_{[i]}^2 \approx 1$.
- para SQF grande (i.e., efeitos significativos do Factor sobre Y), a separação das médias de Y em cada grupo predomina na expressão. $SQF \gg \sum_{i=1}^k SQT_i \Rightarrow R^2 \approx 1$, independentemente dos $R_{[i]}^2$.

Ainda o exemplo do slide 357

Os valores de cada Soma de Quadrados, e dos R^2 , para cada um dos modelos referidos no acetato 357, 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 |
| Ancova | 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 da ANCOVA.

AVISO: o modelo dum recta única não surge nesta comparação.

Generalizando

Como referido, é automática a extensão a k níveis do factor.

A ideia de fundo usada para comparar rectas de regressão linear em k contextos diferentes pode ser generalizada para estudar qualquer regressão linear múltipla em k 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.