

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 uma ou mais variáveis qualitativas (categóricas), ou seja, de um ou mais factores. Por exemplo, podemos querer relacionar o rendimento de uma cultura com os tipos de adubo disponíveis no mercado.

Em tais situações pode ser útil uma **Análise de Variância (ANOVA)**, metodologia estatística desenvolvida nos anos 30 na Estação Experimental Agrícola de Rothamstead (Reino Unido), por **R.A. Fisher**.

A ANOVA como caso particular do Modelo Linear

É possível formular a Análise de Variância como uma técnica distinta da Regressão Linear. Mas ambas são particularizações do chamado **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 : “valores” (distintas categorias) do factor, ou seja, diferentes situações experimentais onde se farão observações de Y .

A ANOVA a um Factor

Começamos por analisar o mais simples de todos os modelos ANOVA: a **ANOVA a um Factor** (totalmente casualizado).

Consideramos que **a variável resposta (numérica) Y depende de um único factor**. Admite-se que os valores de Y poderão variar por corresponderem a níveis diferentes do factor, ou ainda devido a flutuação aleatória.

As n observações

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

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

Embora fosse possível continuar a indexar as n observações de Y com um único índice, variando de 1 a n (como se fez na Regressão), é preferível utilizar dois índices para indexar as observações de Y :

- um para indicar o nível do factor a que a observação corresponde;
- outro para distinguir cada observação dentro de um dado nível.

As n observações (cont.)

Em geral, Y_{ij} indica a j -ésima observação no i -ésimo nível do factor, com $i = 1, \dots, k$ e $j = 1, \dots, n_i$.

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**.

Os delineamentos equilibrados são aconselháveis, por várias razões.

A modelação de Y

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 em cada uma das k situações (níveis do factor) em que é observado.

Uma primeira formulação do modelo pode assim ser dada pela equação de base:

$$E[Y_{ij}] = \mu_j .$$

A modelação de Y (cont.)

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

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

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

Cada α_i é designado o efeito do nível i .

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

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

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

O modelo ANOVA como um Modelo Linear

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

Seja

- \mathbf{Y} o vector n -dimensional com a totalidade das observações da variável resposta. Admite-se que as n_1 primeiras correspondem ao nível 1 do factor, as n_2 seguintes ao nível 2, e assim de seguida.
- $\mathbf{1}_n$ o vector de n uns, já considerado na regressão.
- \mathbf{I}_i a variável indicatriz de pertença ao 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.
- $\boldsymbol{\varepsilon}$ o vector dos n erros aleatórios.

As 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, as variáveis I_2 e I_3 serão:

$$I_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad I_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

A relação de base em notação vectorial

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

$$Y = \mu \cdot \mathbf{1}_n + \alpha_1 \cdot \mathbf{l}_1 + \alpha_2 \cdot \mathbf{l}_2 + \alpha_3 \cdot \mathbf{l}_3 + \boldsymbol{\varepsilon} .$$

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

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

$$\Leftrightarrow Y = X \cdot \beta + \boldsymbol{\varepsilon}$$

O problema do excesso de parâmetros

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

Existe um **excesso de parâmetros** no modelo. **Soluções possíveis:**

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

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:

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \\ Y_{31} \\ Y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} \mu_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix}$$

Agora μ_1 é o valor médio das observações do nível $i = 1$:

$$E[Y_{1j}] = \mu_1 \quad \forall j = 1, \dots, n_1$$

$$E[Y_{2j}] = \mu_1 + \alpha_2 \quad \forall j = 1, \dots, n_2$$

$$E[Y_{3j}] = \mu_1 + \alpha_3 \quad \forall j = 1, \dots, n_3$$

Cada α_i ($i > 1$) representa um **acrécimo** à média do primeiro nível.

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

Neste contexto, a matriz \mathbf{X} tem por colunas os vectors $\mathbf{1}_n, \mathbf{l}_2, \mathbf{l}_3, \dots, \mathbf{l}_k$.

A matriz \mathbf{X} em ANOVAs é chamada a **matriz do delineamento**, pois indica quais as observações correspondentes a cada nível do factor.

Como na Regressão, os valores ajustados de Y resultam de projectar ortogonalmente os valores observados da variável resposta Y sobre o subespaço de \mathbb{R}^n gerado pelas colunas da matriz \mathbf{X} .

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

O subespaço $\mathcal{L}(\mathbf{X})$ numa ANOVA a um factor

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

$$\mathbf{a}_1 \cdot \mathbf{1}_n + \mathbf{a}_2 \cdot \mathbf{l}_2 + \mathbf{a}_3 \cdot \mathbf{l}_3 + \dots + \mathbf{a}_k \cdot \mathbf{l}_k = \begin{bmatrix} \mathbf{a}_1 \\ \dots \\ \mathbf{a}_1 \\ \hline \mathbf{a}_1 + \mathbf{a}_2 \\ \dots \\ \mathbf{a}_1 + \mathbf{a}_2 \\ \hline \mathbf{a}_1 + \mathbf{a}_3 \\ \dots \\ \mathbf{a}_1 + \mathbf{a}_3 \\ \hline (\dots) \\ \hline \mathbf{a}_1 + \mathbf{a}_k \\ \dots \\ \mathbf{a}_1 + \mathbf{a}_k \end{bmatrix}$$

Os estimadores dos parâmetros

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

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

Devido à natureza das colunas da matriz \mathbf{X} , tem-se:

$$\mathbf{X}^t\mathbf{X} = \begin{bmatrix} n & n_2 & n_3 & n_4 & \cdots & n_k \\ n_2 & n_2 & 0 & 0 & \cdots & 0 \\ n_3 & 0 & n_3 & 0 & \cdots & 0 \\ n_4 & 0 & 0 & n_4 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ n_k & 0 & 0 & 0 & \cdots & n_k \end{bmatrix}$$

Os estimadores dos parâmetros (cont.)

Tem-se também:

$$(\mathbf{X}^t \mathbf{X})^{-1} = \frac{1}{n_1} \begin{bmatrix} 1 & -1 & -1 & -1 & \cdots & -1 \\ -1 & \frac{n_1+n_2}{n_2} & 1 & 1 & \cdots & 1 \\ -1 & 1 & \frac{n_1+n_3}{n_3} & 1 & \cdots & 1 \\ -1 & 1 & 1 & \frac{n_1+n_4}{n_4} & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & 1 & 1 & 1 & \cdots & \frac{n_1+n_k}{n_k} \end{bmatrix}$$

$$\mathbf{X}^t \mathbf{Y} = \begin{bmatrix} \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij} \\ \sum_{j=1}^{n_2} Y_{2j} \\ \sum_{j=1}^{n_3} Y_{3j} \\ \vdots \\ \sum_{j=1}^{n_k} Y_{kj} \end{bmatrix}$$

Os estimadores dos parâmetros (cont.)

Logo,

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

onde $\bar{Y}_i. = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$ é a média das n_i observações de Y no nível i .

Ou seja, os parâmetros são estimados pelas quantidades amostrais correspondentes.

Os estimadores das médias de nível

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

$$\begin{aligned}\hat{\mu}_1 &= \bar{Y}_1. \\ \hat{\mu}_2 &= \bar{Y}_2. \\ \hat{\mu}_3 &= \bar{Y}_3. \\ &\vdots \\ \hat{\mu}_k &= \bar{Y}_k.\end{aligned}$$

sendo \bar{Y}_i a média das n_i observações de Y no nível i do factor.

Qualquer observação no nível i tem por valor ajustado $\hat{Y}_{ij} = \hat{\mu}_i = \bar{Y}_i$.

O modelo para efeitos inferenciais

Para se poder fazer inferência neste modelo, admite-se não apenas que cada observação individual Y_{ij} é da forma

$$Y_{ij} = \mu_1 + \alpha_i + \varepsilon_{ij}, \quad \forall i = 1, \dots, k, \quad \forall j = 1, \dots, n_i,$$

com $E[\varepsilon_{ij}] = 0$ e $\alpha_1 = 0$.

Admite-se ainda que os erros aleatórios ε_{ij} têm as mesmas propriedades que no modelo de regressão linear: Normais, de variância constante e independentes.

O modelo ANOVA a um factor

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

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

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

O modelo ANOVA a um factor - notação vectorial

De forma equivalente, em notação vectorial,

Modelo ANOVA a um factor - notação vectorial

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

- 1 $\mathbf{Y} = \mu_1 \cdot \mathbf{1}_n + \alpha_2 \cdot \mathbf{l}_2 + \alpha_3 \cdot \mathbf{l}_3 + \dots + \alpha_k \cdot \mathbf{l}_k + \boldsymbol{\varepsilon}$, sendo $\mathbf{1}_n$ o vector de n uns e $\mathbf{l}_2, \mathbf{l}_3, \dots, \mathbf{l}_k$ as variáveis indicatrizes dos níveis indicados.
- 2 $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \cdot \mathbf{l}_{n \times n})$, sendo $\mathbf{l}_{n \times n}$ a matriz identidade $n \times n$.

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

O modelo ANOVA a um factor - notação vectorial/matricial

Uma terceira forma equivalente de escrever o Modelo:

Modelo ANOVA a um factor - notação vectorial/matricial

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

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

onde $\mathbf{X} = [\mathbf{1}_n \mid \mathbf{l}_2 \mid \mathbf{l}_3 \mid \cdots \mid \mathbf{l}_k]$ e $\boldsymbol{\beta} = (\mu_1, \alpha_2, \alpha_3, \dots, \alpha_k)^t$,
sendo $\mathbf{1}_n$ o vector de n uns e $\mathbf{l}_2, \mathbf{l}_3, \dots, \mathbf{l}_k$ as variáveis indicatrizes dos níveis indicados.

2 $\boldsymbol{\varepsilon} \cap \mathcal{N}_n(\mathbf{0}, \sigma^2 \cdot \mathbf{I}_{n \times n})$, sendo $\mathbf{I}_{n \times n}$ a matriz identidade $n \times n$.

O teste aos efeitos do factor

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

$$\alpha_2 = \alpha_3 = \dots = \alpha_k = 0.$$

É possível testar esta hipótese, através dum teste F de ajustamento global do modelo (ver acetato 154).

As Somas de Quadrados têm, neste contexto, fórmulas específicas.

Os resíduos e SQRE

Viu-se antes (acetato 217) que $\hat{Y}_{ij} = \hat{\mu}_i = \bar{Y}_{i.}$, pelo que o resíduo da observação Y_{ij} é dado por:

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

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

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

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 no i -ésimo nível do factor.

SQRE mede variabilidade no seio dos k níveis.

A Soma de Quadrados associada ao Factor

A Soma de Quadrados associada à Regressão toma, neste contexto, a designação **Soma de Quadrados associada ao Factor** e será representada por **SQF**. É dada por:

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

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

SQF mede **variabilidade** entre as médias amostrais de cada nível.

A relação entre Somas de Quadrados

A relação fundamental entre as três Somas de Quadrados ganha, neste contexto, um significado particular:

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

onde:

SQT – numerador da variância amostral S_Y^2 da totalidade das n observações de Y ;

SQF – medida da variabilidade das k médias de nível (variabilidade inter-níveis);

$SQRE$ – soma ponderada das variâncias amostrais de Y em cada um dos k níveis (variabilidade intra-níveis).

Os graus de liberdade

Neste contexto, o **número de parâmetros do modelo** é $p + 1 = k$. Logo, os graus de liberdade associados a cada Soma de Quadrados são:

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

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

O quadro-resumo da ANOVA a 1 Factor

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	$n - 1$	$SQT = (n - 1) s_y^2$	—	—

O Teste F aos efeitos do factor numa ANOVA

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

Teste F aos efeitos do factor

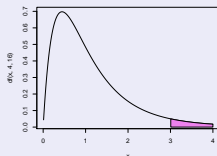
Hipóteses: $H_0 : \alpha_i = 0 \quad \forall i=2,\dots,k$ vs. $H_1 : \exists i=2,\dots,k$ t.q. $\alpha_i \neq 0$.
[FACTOR NÃO AFECTA] vs. [FACTOR AFECTA Y]

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


Nível de significância do teste: γ

Região Crítica (Região de Rejeição): Unilateral direita


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



ANOVAs a um Factor no

Para efectuar uma ANOVA a um Factor no , convém **organizar 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).

O  reconhece objectos do tipo **factor**, que são criados através do comando **factor**, aplicado a um vector de tipo `character`, que contenha os nomes dos vários níveis:

```
> factor(c("Adubo 1", "Adubo 1", ... , "Adubo k"))
```

NOTA: Explore o comando **rep** para instruções curtas que criam **repetições** de valores.

ANOVAs a um Factor no (cont.)


As fórmulas utilizadas no  para indicar as ANOVAs pretendidas são semelhantes às usadas na regressão linear, admitindo a indicação de nomes de factores.

Por exemplo, se pretendemos efectuar uma ANOVA de comprimentos das pétalas sobre espécies, nos dados relativos aos $n = 150$ lírios, a fórmula é:

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

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 uma ANOVA seja um caso particular do Modelo Linear, e seja possível usar o comando `lm` do  para efectuar uma ANOVA, existe outra função que organiza a informação da forma mais tradicional numa ANOVA: a função `aov`.

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

```
> aov(Petal.Length ~ Species)
```

É produzido o seguinte resultado (diferente do do comando `lm`):

```
Call:    aov(formula = Petal.Length ~ Species, data=iris)
```

```
Terms:
```

	Species	Residuals
Sum of Squares	437.1028	27.2226
Deg. of Freedom	2	147

```
Residual standard error: 0.4303345
```

ANOVAs a um factor no (cont.)

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

```
> iris.aov <- aov(Petal.Length ~ Species, data=iris)
> summary(iris.aov)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Species	2	437.10	218.55	1180.2	< 2.2e-16 ***
Residuals	147	27.22	0.19		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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.

No exemplo dos lírios, temos:

```
> coef(iris.aov)
      (Intercept) Speciesversicolor Speciesvirginica
              1.462                2.798                4.090
```

Estes são os **valores estimados** dos parâmetros

- $\hat{\mu}_1$: **média amostral** de comprimentos de pétalas *setosa*;
- $\hat{\alpha}_2$: **acréscimo** que, somado à média amostral da 1a. espécie, nos dá a média amostral dos comprimentos de pétalas *versicolor*;
- $\hat{\alpha}_3$: **acréscimo** que, somado à média amostral da 1a. espécie, nos dá a média amostral dos comprimentos de pétalas *virginica*.

Parâmetros estimados no (cont.)

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

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

```
Tables of means
```

```
Grand mean
```

```
3.758
```

```
Species
```

```
Species
```

```
setosa versicolor virginica
```

```
1.462      4.260      5.552
```

 ordena os níveis de um factor por ordem alfabética.

ANOVAs como modelo Linear no

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

```
> summary(lm(Petal.Length ~ Species , data=iris))
Call: lm(formula = Petal.Length ~ Species, data=iris)
Residuals:
    Min       1Q   Median       3Q      Max
-1.260 -0.258  0.038  0.240  1.348

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.46200    0.06086   24.02  <2e-16 ***
Speciesversicolor  2.79800    0.08607   32.51  <2e-16 ***
Speciesvirginica  4.09000    0.08607   47.52  <2e-16 ***
---
Residual standard error: 0.4303 on 147 degrees of freedom
Multiple R-squared:  0.9414, Adjusted R-squared:  0.9406
F-statistic: 1180 on 2 and 147 DF,  p-value: < 2.2e-16
```

A exploração ulterior de H_1

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

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

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

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

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

A exploração ulterior de H_1 (cont.)

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

Mesmo com $k = 3$, a rejeição de H_0 pode dever-se a:

$$\mu_1 = \mu_2 \neq \mu_3 \quad \text{i.e.,} \quad \alpha_2 = 0 ; \alpha_3 \neq 0$$

$$\mu_1 = \mu_3 \neq \mu_2 \quad \text{i.e.,} \quad \alpha_3 = 0 ; \alpha_2 \neq 0$$

$$\mu_1 \neq \mu_2 = \mu_3 \quad \text{i.e.,} \quad \alpha_2 = \alpha_3 \neq 0;$$

$$\mu_i \text{ todos diferentes} \quad \text{i.e.,} \quad \alpha_2 \neq \alpha_3 \text{ e } \alpha_2, \alpha_3 \neq 0.$$

Como optar entre estas diferentes alternativas?

A exploração ulterior de H_1 (cont.)

Uma hipótese consiste em efectuar testes aos α_i s, com base na teoria já estudada anteriormente.

Mas quanto maior for k , mais sub-hipóteses alternativas existem, mais testes haverá para fazer.

Não se trata apenas de uma questão de serem necessários muitos testes. A multiplicação do número de testes faz perder o controlo do nível de significância γ global para o conjunto de todos os testes.

As comparações múltiplas

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

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

O mais usado desses testes é o **teste de Tukey**.

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


Distribuição de Tukey para Amplitudes Studentizadas

Teorema (Distribuição de Tukey)

Sejam $\{\mathbf{W}_i\}_{i=1}^k$ variáveis aleatórias independentes, com distribuição Normal, de iguais parâmetros: $\mathbf{W}_i \sim \mathcal{N}(\mu_W, \sigma_W^2)$, $\forall i = 1, \dots, k$.

- Seja S_W^2 um estimador da variância comum σ_W^2 , tal que $\frac{v S_W^2}{\sigma_W^2} \sim \chi_v^2$.
- Seja $R_W = \max_i \mathbf{W}_i - \min_i \mathbf{W}_i$ a *amplitude amostral*.
- Sejam S_W e R_W independentes.

Então, a *amplitude Studentizada*, $\frac{R_W}{S_W}$, tem a *distribuição de Tukey*, que depende de *dois parâmetros*: k e v .

Os valores da função distribuição cumulativa e os quantis duma distribuição de Tukey são calculados no , através das funções `ptukey` e `qtukey`, respectivamente.

A utilidade da distribuição de Tukey

Numa ANOVA a um factor, admitimos que

$$Y_{ij} = \underbrace{\mu_1 + \alpha_j}_{=\mu_j} + \varepsilon_{ij}, \quad (\alpha_1 = 0),$$

pelo que (com os pressupostos relativos aos erros aleatórios do modelo ANOVA)

$$Y_{ij} \cap \mathcal{N}(\mu_j, \sigma^2).$$

Logo, a **média amostral de cada nível**, $\bar{Y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$, tem distribuição

$$\bar{Y}_{i.} \cap \mathcal{N}\left(\mu_i, \frac{\sigma^2}{n_i}\right) \quad \Leftrightarrow \quad \bar{Y}_{i.} - \mu_i \cap \mathcal{N}\left(0, \frac{\sigma^2}{n_i}\right)$$

A utilidade da distribuição de Tukey (cont.)

Caso o **delineamento seja equilibrado**, isto é,

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

as k diferenças $\bar{Y}_i - \mu_i$ terão a mesma distribuição $\mathcal{N}(0, \sigma^2/n_c)$, e serão as variáveis \mathbf{W}_i do Teorema no acetato (240).

Um **estimador da variância comum** σ^2/n_c é dado por $QMRE/n_c$, e:

$$(n-k) \cdot \frac{QMRE/n_c}{\sigma^2/n_c} = \frac{SQRE}{\sigma^2} \cap \chi_{n-k}^2,$$

(acetatos 134 e 135, pois no modelo ANOVA há k parâmetros). Os valores ajustados \bar{Y}_i e os resíduos que definem $SQRE$ são independentes, logo, **a amplitude amostral**

$$R = \max_i(\bar{Y}_i - \mu_i) - \min_j(\bar{Y}_j - \mu_j)$$

é independente do estimador da variância comum, $QMRE/n_c$.

Aplica-se o Teorema do acetato (240).

Assim,

$$\frac{R}{S} = \frac{\max_i(\bar{Y}_i - \mu_i) - \min_j(\bar{Y}_j - \mu_j)}{\sqrt{\frac{QMRE}{n_c}}}$$

tem a distribuição de Tukey, com parâmetros k e $n - k$.
O quociente $\frac{R}{S}$ não pode ser negativo, por definição.

Este resultado pode ser usado para construir testes de hipóteses ou intervalos de confiança para o conjunto de todas as diferenças de médias de nível de Y , $\mu_i - \mu_j$.

Intervalos de Confiança para $\mu_i - \mu_j$

Seja $q_{\gamma(k,n-k)}$ o valor que numa distribuição de Tukey com parâmetros k e $n - k$, deixa à direita uma região de probabilidade γ . Então, por definição:

$$P \left[\frac{R}{S} < q_{\gamma(k,n-k)} \right] = 1 - \gamma$$

Logo, um intervalo de confiança a $(1 - \gamma) \times 100\%$ para a amplitude R é dado por:

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

Intervalos de Confiança para $\mu_i - \mu_j$ (cont.)

Mas $R = \max_i(\bar{Y}_i - \mu_i) - \min_j(\bar{Y}_j - \mu_j)$ é a maior de todas as diferenças do tipo $|(\bar{Y}_i - \mu_i) - (\bar{Y}_j - \mu_j)|$, para qualquer $i, j = 1, \dots, k$.

Logo, para todos os pares de níveis i e j , tem-se, com grau de confiança global $(1 - \gamma) \times 100\%$,

$$\begin{aligned} |(\bar{y}_i - \bar{y}_j) - (\mu_i - \mu_j)| &\leq R < q_{\gamma(k, n-k)} \cdot \sqrt{\frac{QMRE}{nc}} \\ \Leftrightarrow (\bar{y}_i - \bar{y}_j) - q_{\gamma(k, n-k)} \cdot \sqrt{\frac{QMRE}{nc}} &< (\mu_i - \mu_j) < \\ &(\bar{y}_i - \bar{y}_j) + q_{\gamma(k, n-k)} \cdot \sqrt{\frac{QMRE}{nc}} \end{aligned}$$

Testes de Hipóteses para $\mu_i - \mu_j = 0$, $\forall i, j$

Alternativamente, a partir do resultado do acetato (243) é possível testar a Hipótese Nula de que **todas** as diferenças de pares de médias de nível, $\mu_i - \mu_j$, sejam nulas, em cujo caso


$$|\bar{Y}_{i.} - \bar{Y}_{j.}| < q_{\gamma(k, n-k)} \cdot \sqrt{\frac{QMRE}{n_c}},$$

com probabilidade $(1 - \gamma) \times 100\%$. Qualquer diferença de médias amostrais de nível, $\bar{Y}_{i.} - \bar{Y}_{j.}$, que exceda o limiar

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

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

Comparações Múltiplas de Médias no

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

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

```
> qtukey(1- $\gamma$ , k, n-k)
```

(com os valores numéricos de γ , k e $n - k$).

O valor de \sqrt{QMRE} é dado pelo comando `aoV`, sob a designação “Residual standard error”.

Comparações Múltiplas de Médias no (cont.)

Os intervalos de Confiança a $(1 - \gamma) \times 100\%$ para as diferenças de médias são obtidos através do comando **TukeyHSD**. Por exemplo, para os dados dos lírios,

```
> TukeyHSD(aov(Sepal.Width ~ Species, data=iris))
  Tukey multiple comparisons of means
    95% family-wise confidence level

$Species
              diff            lwr            upr      p adj
versicolor-setosa -0.658 -0.81885528 -0.4971447 0.0000000
virginica-setosa   -0.454 -0.61485528 -0.2931447 0.0000000
virginica-versicolor 0.204  0.04314472  0.3648553 0.0087802
```

Neste exemplo, nenhum dos intervalos inclui o valor zero, pelo que consideramos que $\mu_i \neq \mu_j$, para qualquer $i \neq j$, ou seja, todas as médias de espécie são diferentes.

Comparações Múltiplas de Médias no (cont.)


O valor de prova indicado (p adj) deve ser interpretado como o valor de γ para o qual cada diferença de médias, $\bar{y}_i - \bar{y}_j$, seria, pela primeira vez, considerado não significativo.

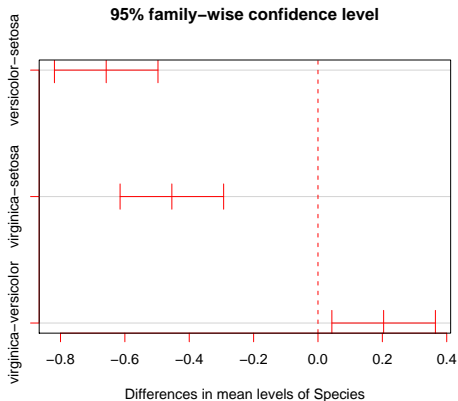
```
> TukeyHSD(aov(Sepal.Width ~ Species, data=iris))
  Tukey multiple comparisons of means
    95% family-wise confidence level

$Species
              diff            lwr            upr      p adj
versicolor-setosa -0.658 -0.81885528 -0.4971447 0.0000000
virginica-setosa   -0.454 -0.61485528 -0.2931447 0.0000000
virginica-versicolor 0.204  0.04314472  0.3648553 0.0087802
```

Assim, para $\gamma = 0.00878$, a diferença de médias amostrais para as espécies *virginica* e *versicolor* já seria considerada não significativa. Ou seja, o intervalo a $(1 - \gamma) \times 100\% = 0.99122\%$ de confiança para essa diferença de médias já conteria o valor zero.


Representação gráfica das comparações múltiplas

O  disponibiliza ainda um auxiliar gráfico para visualizar as comparações das médias de nível, através da função `plot`, aplicada ao resultado da função `TukeyHSD`.



Delineamentos não equilibrados

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

Mas, para delineamentos em que o desequilíbrio no número de observações não seja muito acentuado, é possível ajustar os valores da distribuição de Tukey. A função `TukeyHSD` do  incorpora essas correcções.

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. Mas há algumas particularidades.

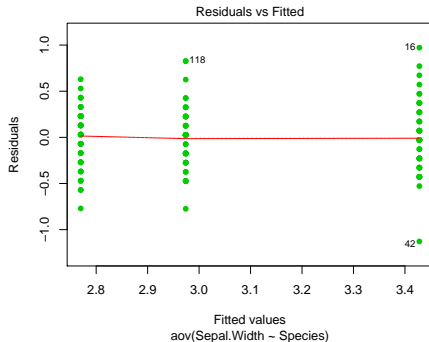
Numa ANOVA a um factor, os resíduos aparecem empilhados em k colunas nos gráficos de \hat{y}_{ij} vs. e_{ij} , porque qualquer valor ajustado \hat{y}_{ij} é igual para observações num mesmo nível do factor.

Este padrão **não** indicia qualquer violação aos pressupostos do modelo.

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

Padrão de resíduos numa ANOVA a 1 Factor

(o exemplo considerado é $\text{Sepal.Width} \sim \text{Species}$, nos lírios)



Inspeccionando a homogeneidade de variâncias

Outra particularidade da ANOVA, resultante do facto de haver n_i repetições em cada um dos k níveis do factor: **é possível testar formalmente se as variâncias dos erros aleatórios diferem entre os níveis do factor.**

O **Teste de Bartlett** testa as hipóteses

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$$

vs.

$$H_1 : \exists i, i' \text{ t.q. } \sigma_i^2 \neq \sigma_{i'}^2 ,$$

sendo σ_i^2 a variância comum dos erros aleatórios ε_{ij} do nível i .

Médias aritméticas e médias geométricas

Relação geral entre a média aritmética e a média geométrica (mesmo que ponderadas) de quaisquer k números positivos.

Sejam $\tau_1, \tau_2, \dots, \tau_k$ números positivos, e p_1, p_2, \dots, p_k números entre 0 e 1, de soma 1.

A **média aritmética** (ponderada com pesos p_i) dos τ_i s é

$$MA = \sum_{i=1}^k p_i \tau_i .$$

A **média geométrica** (ponderada com pesos p_i) dos τ_i s é

$$MG = \prod_{i=1}^k \tau_i^{p_i} .$$

Quando $p_i = \frac{1}{k}, \forall i$, temos as médias aritmética e geométrica simples.

A desigualdade entre média aritmética e geométrica

Quaisquer que sejam os valores (positivos) dos τ_i e das ponderações p_i , tem-se a seguinte desigualdade entre a média aritmética e geométrica dos k valores de τ :

$$MG \leq MA \quad (4)$$

A igualdade em (4) verifica-se se e só se os k valores de τ são iguais:

$$\tau_1 = \tau_2 = \cdots = \tau_k .$$

Quanto maior for a dispersão dos τ , maior será a diferença entre média geométrica e média aritmética.

O nosso contexto

Admita-se que os erros aleatórios, e portanto as observações Y_{ij} , do nível i do factor têm **variância comum** $V[\varepsilon_{ij}] = V[Y_{ij}] = \sigma_i^2$, podendo, no entanto os σ_i^2 diferir entre níveis.

Sejam MA e MG as médias, respectivamente aritmética e geométrica, das k variâncias de nível, $\{\sigma_i^2\}_{i=1}^k$, para um dado conjunto de pesos p_i . Tem-se sempre

$$\frac{MA}{MG} \geq 1,$$

com a igualdade se e só se for verdadeira a Hipótese Nula de que os σ_i^2 são todos iguais.

Estimando as variâncias de nível

Os σ_i^2 são desconhecidos. Mas podem ser estimados pelas variâncias amostrais das observações de Y , i.e., cada σ_i^2 pode ser estimado por:

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

Se usarmos como ponderações

$$p_i = \frac{n_i - 1}{\sum_i (n_i - 1)} = \frac{n_i - 1}{n - k} ,$$

a média aritmética ponderada dos estimadores S_i^2 é o Quadrado Médio Residual da ANOVA (ver o Acetato 223):

$$MA = \sum_{i=1}^k \frac{n_i - 1}{n - k} \cdot S_i^2 = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2}{n - k} = QMRE .$$

A ideia subjacente ao teste de Bartlett

A média geométrica dos k estimadores de variâncias de nível é:

$$MG = \prod_{i=1}^k (S_i^2)^{\frac{n_i-1}{n-k}} .$$

Sabemos que $MA/MG \geq 1$. Quanto maior for este quociente, maior será a variabilidade dos S_i^2 , e portanto mais duvidosa será a Hipótese Nula da igualdade dos σ_i^2 .

Logo, o quociente $\frac{MA}{MG}$ é um candidato a estatística do teste à igualdade de variâncias, com Região Crítica unilateral direita. Mas é necessário conhecer a distribuição de probabilidades duma estatística do Teste, sob H_0 .

O teste de Bartlett

Bartlett demonstrou que, sob H_0 , uma transformação monótona crescente do quociente MA/MG tem distribuição assintoticamente χ^2 , caso as variáveis subjacentes às variâncias tenham distribuição Normal. Concretamente, demonstrou que

$$K = \frac{n-k}{C} \cdot \log \left[\frac{MA}{MG} \right] = \frac{n-k}{C} \cdot (\log MA - \log MG),$$

tem, assintoticamente distribuição χ_{k-1}^2 , sendo

$$C = 1 + \frac{1}{3(k-1)} \left[\sum_{i=1}^k \frac{1}{n_i - 1} - \frac{1}{n-k} \right].$$

O Teste de Bartlett

Teste de Bartlett à homogeneidade de variâncias

Hipóteses: $H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$ vs. $H_1 : \exists i, i' \text{ t.q. } \sigma_i^2 \neq \sigma_{i'}^2$
[Variâncias homogéneas] [Var. heterogéneas]

Estatística do Teste:

$$K = \frac{(n-k) \log QMRE - \sum_{i=1}^k (n_i - 1) \log S_i^2}{C} \sim \chi_{k-1}^2$$


$$\text{onde } C = 1 + \frac{1}{3(k-1)} \left[\sum_{i=1}^k \frac{1}{n_i - 1} - \frac{1}{n-k} \right].$$

Nível de significância do teste: γ

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se $K_{\text{calc}} > \chi_{\gamma(k-1)}^2$

O Teste de Bartlett no

No , o teste de Bartlett é invocado pelo comando `bartlett.test`, tendo por argumento uma fórmula (análoga à usada no comando `aov` para indicar a variável resposta e o factor). E.g.,

```
> bartlett.test(Sepal.Width ~ Species, data=iris)
```

```
Bartlett test of homogeneity of variances
```

```
data: Sepal.Width by Species
```

```
Bartlett's K-squared = 2.0911, df = 2, p-value = 0.3515
```

Neste caso, o teste de Bartlett indica a não rejeição de H_0 , ou seja, é admissível a hipótese de igualdade nas variâncias em cada nível do factor.

Precauções

Duas precauções na utilização do teste de Bartlett:

- O teste de Bartlett é fortemente sensível à Normalidade das observações subjacentes.
- A distribuição χ^2 é apenas assintótica. Uma regra comum é considerar que o teste apenas deve ser usado caso $n_i \geq 5$, $\forall i = 1, \dots, k$.

Violações aos pressupostos da ANOVA

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

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

Uma advertência

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

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

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

Esta condição alternativa:

- muda a forma de interpretar os parâmetros (μ é agora uma espécie de média geral das observações e α_i o desvio médio das observações do nível i em relação a essa média geral);
- Muda os estimadores dos parâmetros.
- **Não** muda o resultado do teste F à existência de efeitos do factor, nem a qualidade global do ajustamento.
- **A nossa formulação**, além de generalizável a modelos com mais Factores, **permite aproveitar directamente os resultados da Regressão Linear Múltipla.**

Unidades experimentais

No delineamento das experiências para posterior análise através duma ANOVA (ou regressão linear), é frequente que as n observações da variável resposta correspondam a n diferentes indivíduos, ou parcelas de terreno, ou outra entidade que se pode designar uma **unidade experimental**.

As **unidades experimentais** nas quais se recolhem os dados devem ser **tão homogêneas quanto possível**,

tendo sido controladas de forma a **eliminar variabilidade** que possa afectar a variável resposta, **para além da variação nos preditores que se estejam a analisar**.

Unidades experimentais (cont.)

Qualquer **variabilidade não controlada** nas unidades experimentais (isto é, que não se pode atribuir aos preditores) é considerada no modelo como **variação aleatória**, pelo que irá contribuir 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,

heterogeneidade não controlada nas unidades experimentais contribui para esconder a presença de eventuais efeitos do factor.

Controlar a heterogeneidade

Na prática, é frequentemente impossível controlar totalmente todos os factores que afectam as unidades experimentais.

A natural variabilidade de plantas, animais, terrenos, localidades geográficas, células, etc. significa que em muitas situações existirá variabilidade indesejada entre unidades experimentais.

Alguma protecção contra efeitos não controlados resulta dos princípios de:

- repetição;
- casualização.

Deve-se associar níveis do factor às unidades experimentais de forma aleatória (casualizada).