Análise de Variância (ANOVA) de efeitos fixos

Adaptado de Cadima, J. (2021). O Modelos Linear. ISA-UL.

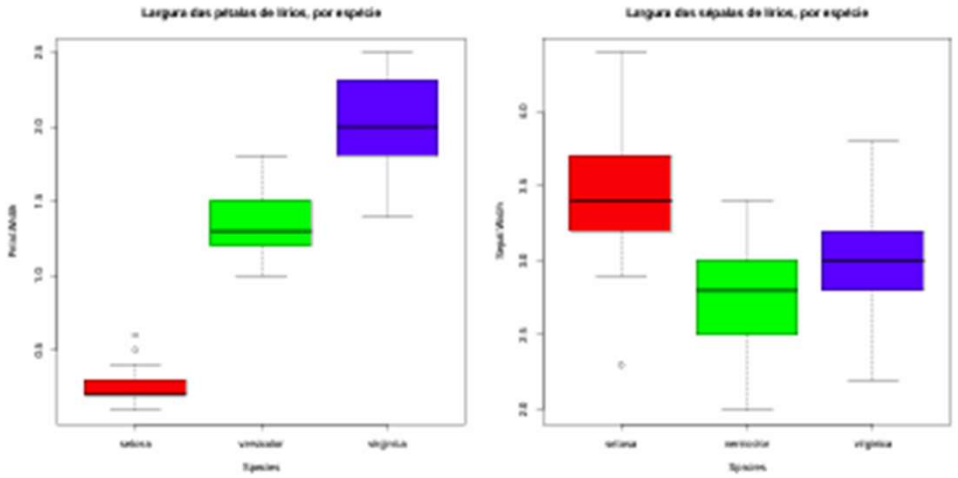
A Regressão Linear visa modelar uma variável resposta numérica (quantitativa), à custa de uma ou mais variáveis preditoras, igualmente numéricas.

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. Eis as médias amostrais:

$$\overline{y}_{seto} = 3.428 \quad ; \quad \overline{y}_{vers} = 2.770 \quad ; \quad \overline{y}_{virg} = 2.974$$

As diferenças serão apenas um acaso da amostra?

Objectivo: Testar a igualdade das médias populacionais de cada espécie.

A ANOVA como caso particular do Modelo Linear

A Análise de Variância (ANOVA) lida com variáveis preditoras (explicativas) qualitativas. Surgiu historicamente como um método autónomo. Mas, tal como a Regressão Linear, é uma particularização 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.

Nos exemplos, o factor Espécie tem $k = 3$ níveis.

A ANOVA a um Factor - notação

Na ANOVA a um Factor (totalmente casualizado), a modelação da variável resposta baseia-se numa única variável preditora categórica.

Admitimos que o factor tem k níveis (no exemplo dos lírios, $k=3$).

Admitimos que há n observações independentes de Y, sendo n_i $(i=1,\ldots,k)$ correspondentes ao nível i do factor. Logo, $\sum_{i=1}^{k} n_i = n$.

Delineamentos equilibrados

No caso de igual número de observações em cada nível,

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

diz-se que estamos perante um delineamento equilibrado.

Os delineamentos equilibrados são aconselháveis (mas não obrigatórios), por várias razões que adiante se discutem.

A dupla indexação de Y

Na regressão linear indexam-se as n observações de Y com um único índice, variando de 1 a n $(\{Y_i\}_{i=1}^{n})$.

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

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

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

O modelo ANOVA como um Modelo Linear

A equação geral $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, nas n_1 observações do nível $i = 1$ fica:

$$Y_{1j} = \mu + \alpha_1 + \varepsilon_{1j} ,$$

nas n_2 observações efectuadas no nível $i = 2$ fica:

$$Y_{2j} = \mu + \alpha_2 + \varepsilon_{2j} ,$$

etc.. Este conjunto de k equações pode ser escrita como uma única equação geral, que é a equação dum modelo linear:

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

onde $\mathscr{I}_m$ é a variável indicatriz do nível m do factor:

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

A relação de base em notação vectorial

Em notação matricial/vectorial, a equação de base será:

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

As colunas de $\mathbf{X}$ são: o vector $\vec{\mathbf{1}}_n$ e os vectores das indicatrizes $\vec{\mathscr{I}}_i$.
O vector dos parâmetros $\vec{\boldsymbol{\beta}}$ tem elementos: μ e os efeitos α_i.

Num exemplo com $n_1 = 3$, $n_2 = 4$ e $n_3 = 2$ observações:

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

O problema do excesso de parâmetros

Existe um problema "técnico": as colunas desta matriz **X** são linearmente dependentes (a soma das indicatrizes é o vector dos n uns) , pelo que a matriz $\mathbf{X}^t\mathbf{X}$ não é invertível. Há um excesso de parâmetros no modelo.

Soluções possíveis na equação $Y_{ij} = \mu + \alpha_1 \mathscr{I}_{1_{ij}} + \alpha_2 \mathscr{I}_{2_{ij}} + ... + \alpha_k \mathscr{I}_{k_{ij}} + \varepsilon_{ij}$:

1. retirar o parâmetro μ do modelo.
   - corresponde a retirar a coluna de uns da matriz **X**;
   - cada α_i equivalerá a μ_i, a média do nível;
   - não se pode generalizar a situações mais complexas;
   - mais difícil de encaixar na teoria já dada do Modelo Linear.
2. 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. tomar $\alpha_1 = 0$: será a solução utilizada.
   - corresponde a excluir a 1a. variável indicatriz do modelo (e de **X**);
   - permite aproveitar a teoria do Modelo Linear e é generalizável.

Cada solução tem implicações na forma de interpretar os parâmetros.

A matriz do modelo com a restrição $\alpha_1 = 0$

Com a restrição $\alpha_1 = 0$, a matriz do modelo **X** tem colunas $\vec{\mathbf{1}}_n, \vec{\mathscr{I}}_2, \ldots, \vec{\mathscr{I}}_k$.
No exemplo anterior, tem-se:

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

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

$$\begin{array}{lcccccccl} Y_{1j} & = & \mu + \varepsilon_{1j} & \Rightarrow & \mu_1 & = & E[Y_{1j}] & = & \mu & , \forall j = 1, \ldots, n_1 \\ Y_{2j} & = & \mu + \alpha_2 + \varepsilon_{2j} & \Rightarrow & \mu_2 & = & E[Y_{2j}] & = & \mu_1 + \alpha_2 & , \forall j = 1, \ldots, n_2 \\ Y_{3j} & = & \mu + \alpha_3 + \varepsilon_{3j} & \Rightarrow & \mu_3 & = & E[Y_{3j}] & = & \mu_1 + \alpha_3 & , \forall j = 1, \ldots, n_3 \end{array}$$

Os efeitos de nível α_i

Na equação duma ANOVA a um factor (acetato 228), e com a restrição $\alpha_1 = 0$, cada α_i ($i > 1$) representa o acréscimo que transforma a média do primeiro nível na média do nível i:

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

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

O modelo ANOVA a 1 factor para efeitos inferenciais

Para completar o modelo ANOVA a um factor, admite-se que os erros aleatórios ε_{ij} têm as mesmas propriedades que numa regressão linear:

Modelo ANOVA a um factor, com k níveis

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

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

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

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

O modelo ANOVA a um factor - notação vectorial

De forma equivalente, em notação vectorial,

Modelo ANOVA a um factor - notação vectorial

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

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

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

O teste aos efeitos do factor

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

$$\alpha_2 = \alpha_3 = \ldots = \alpha_k = 0$$
$$\Leftrightarrow \quad \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$$

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 $\vec{\mathscr{I}}_i$) são nulos.

O Teste F aos efeitos do factor numa ANOVA

Muda-se a designação de QMR para QMF (Quadrado Médio do Factor):

Teste F aos efeitos do factor

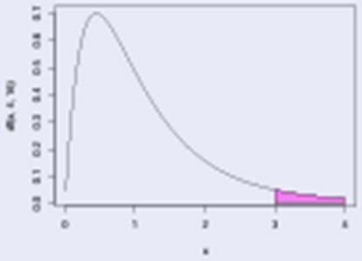
Hipóteses: $H_0 : \alpha_i = 0 \quad \forall i=2,\ldots,k$ vs. $H_1 : \exists i=2,..,k \text{ t.q. } \alpha_i \neq 0.$
[FACTOR NÃO AFECTA] vs. [FACTOR AFECTA Y]

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

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

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



Notação e graus de liberdade

Neste contexto, existem fórmulas simples para algumas quantidades.

Numa ANOVA a um factor, usamos SQF, em vez de SQR, para indicar a Soma de Quadrados associada aos efeitos do Factor, embora a sua definição seja idêntica (numerador da variância dos valores ajustados).

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

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

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

Estimadores de parâmetros na ANOVA a um factor

Na ANOVA a um factor, as k colunas de $\mathbf{X}$ são os vectores $\vec{1}_n$, $\vec{\mathscr{I}}_2$, $\vec{\mathscr{I}}_3$, ... , $\vec{\mathscr{I}}_k$. A matriz identifica as observações de cada nível do factor.

Dada a natureza especial da matriz $\mathbf{X}$, a fórmula dos parâmetros ajustados, $\hat{\vec{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{Y}}$ gera estimadores dos parâmetros populacionais que são as quantidades amostrais análogas. Sendo $\overline{Y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$ a média amostral das n_i observações de Y no nível i, tem-se:

$$
\begin{array}{rcll}
\mu_1 & & \longrightarrow & \hat{\mu}_1 = \overline{Y}_{1\cdot} \\
\alpha_2 = \mu_2 - \mu_1 & & \longrightarrow & \hat{\alpha}_2 = \overline{Y}_{2\cdot} - \overline{Y}_{1\cdot} \\
\alpha_3 = \mu_3 - \mu_1 & & \longrightarrow & \hat{\alpha}_3 = \overline{Y}_{3\cdot} - \overline{Y}_{1\cdot} \\
\vdots & & & \vdots \quad \vdots \quad \vdots \\
\alpha_k = \mu_k - \mu_1 & & \longrightarrow & \hat{\alpha}_k = \overline{Y}_{k\cdot} - \overline{Y}_{1\cdot}
\end{array}
$$

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

Valores ajustados $\hat{Y}_{ij}$

Do que foi visto, decorre que qualquer observação tem valor ajustado igual à média amostral das observações do seu nível:

$$\hat{Y}_{ij} \quad = \quad \underbrace{\hat{\mu}_1 + \hat{\alpha}_i}_{=\hat{\mu}_i} \quad = \quad \overline{Y}_{1\cdot} + (\overline{Y}_{i\cdot} - \overline{Y}_{1\cdot}) \quad = \quad \overline{Y}_{i\cdot} \, .$$

Os valores ajustados $\hat{Y}_{ij}$ são iguais para todas as observações num mesmo nível i do factor. Tal como na Regressão, estes valores resultam de projectar ortogonalmente o vector $\vec{\mathbf{Y}}$ dos valores observados da variável resposta, sobre o subespaço $\mathscr{C}(\mathbf{X}) \subset \mathbb{R}^n$ gerado pelas colunas da matriz $\mathbf{X}$: $\hat{\vec{\mathbf{Y}}} = \mathbf{H}\vec{\mathbf{Y}}$.

Numa ANOVA a um factor, o subespaço $\mathscr{C}(\mathbf{X})$ tem natureza especial: todos os vectores de $\mathscr{C}(\mathbf{X})$ têm de ter valor igual nas posições correspondentes a observações dum mesmo nível do factor.

Os resíduos e *SQRE*

Vimos que $\hat{Y}_{ij} = \hat{\mu}_i = \overline{Y}_{i\cdot}$.

O resíduo da observação Y_{ij} é dado pela sua diferença em relação à média amostral de nível:

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

A Soma de Quadrados dos Resíduos é dada por:

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

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

SQRE mede variabilidade no seio dos k níveis.

Fórmulas para delineamentos equilibrados

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

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

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

Assim, em delineamentos equilibrados, o Quadrado Médio Residual é a média (simples) das k variâncias de nível da variável resposta Y.

Em delineamentos não equilibrados, o QMRE é uma média ponderada dos S_i^2 (tendo cada parcela o peso $n_i - 1$).

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*. Sendo $\overline{Y}_{..} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}$ a média da totalidade das n observações, tem-se:

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

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

SQF mede variabilidade entre as médias amostrais de cada nível.

Fórmulas para delineamentos equilibrados

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

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

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

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

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

A relação entre Somas de Quadrados

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

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

onde:

- $SQT = (n-1)s_y^2$ mede a variabilidade total das n observações de Y;
- SQF mede a variabilidade entre diferentes níveis do factor (variabilidade inter-níveis);
- $SQRE$ mede a variabilidade no seio dos níveis - 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. Aqui, as causas podem ser o efeito do factor ou outras não explicadas pelo modelo (residuais).

O quadro de síntese da ANOVA a 1 Factor

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

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

A exploração ulterior de H_1

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

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

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

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

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:

$$\begin{aligned}
\mu_1 = \mu_2 \neq \mu_3 \quad & \text{i.e.,} \quad \alpha_2 = 0\,;\, \alpha_3 \neq 0 \\
\mu_1 = \mu_3 \neq \mu_2 \quad & \text{i.e.,} \quad \alpha_3 = 0\,;\, \alpha_2 \neq 0 \\
\mu_1 \neq \mu_2 = \mu_3 \quad & \text{i.e.,} \quad \alpha_2 = \alpha_3 \neq 0; \\
\mu_i \text{ todos diferentes} \quad & \text{i.e.,} \quad \alpha_2 \neq \alpha_3 \text{ e } \alpha_2, \alpha_3 \neq 0.
\end{aligned}$$

Como optar entre estas diferentes alternativas?

A exploração ulterior de H_1 (cont.)

Podem efectuar-se testes *t-Student* aos α_is, com base na teoria já estudada anteriormente (recorde-se que um modelo ANOVA é um modelo linear).

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

A multiplicação do número de testes faz perder o controlo do nivel de significância α global para o conjunto de todos os testes.

Testes de hipóteses alternativos, relativos a todas as diferenças $\mu_i - \mu_j$ de pares de médias populacionais de Y, permitem controlar o nível de significância global α do conjunto dos testes. Tais testes chamam-se testes de comparações múltiplas de médias.

As comparações múltiplas

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

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

A mais frequente abordagem de comparações múltiplas leva o nome de Tukey, embora em rigor só seja válido para delineamentos equilibrados.

Testes de Tukey na ANOVA a um factor

Dado um delineamento a um factor, equilibrado.

Teste de Tukey às diferenças de médias de nível

Hipóteses: $H_0 : \mu_i = \mu_j, \forall\, i,j$ vs. $H_1 : \exists i,j \text{ t.q. } \mu_i \neq \mu_j$.
[FACTOR NÃO AFECTA] vs. [FACTOR AFECTA Y]

Nível de significância (global) do teste: α

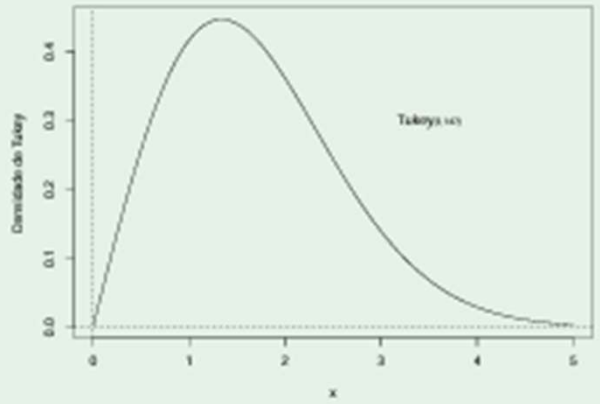
Regra: Rejeitar $\mu_i = \mu_j$ se $|\overline{Y}_{i\cdot} - \overline{Y}_{j\cdot}| > q_{\alpha(k,n-k)} \sqrt{\frac{QMRE}{n_c}}$,
sendo $q_{\alpha(k,n-k)}$ o valor que numa distribuição de Tukey com parâmetros k e $n-k$, deixa à direita uma região de probabilidade α.

O teste permite não apenas rejeitar H_0 globalmente, como identificar o(s) par(es) de níveis (i,j) responsáveis pela rejeição (a diferença das respectivas médias amostrais excede o termo de comparação), permitindo assim conclusões sobre diferenças significativas em cada par de médias.

Distribuição de Tukey

Distribuição Tukey na ANOVA a um factor: lírios

Eis a função densidade da distribuição de Tukey, correspondente ao exemplo dos lírios, com $k = 3$ e $n - k = 147$:



Na *webpage* da disciplina encontra-se uma tabela da distribuição de Tukey.

Intervalos de Confiança para $\mu_i - \mu_j$

Alternativamente, podem construir-se intervalos de confiança para todas as diferenças de pares de médias de nível, $\mu_i - \mu_j$, com um grau de confiança global $(1-\alpha)\times 100\%$.

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

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

Se para qualquer par (i,j) de níveis, o intervalo correspondente não contém o valor zero, então $\mu_i = \mu_j$ não é admissível.

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 $\hat{y}_{ij} = \overline{y}_{i.}$ é igual para observações num mesmo nível do factor.

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

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

Violações aos pressupostos da ANOVA

As n_i repetições em cada um dos k níveis do factor, permitem testar formalmente se as variâncias dos erros aleatórios diferem entre os níveis do factor (testes de Bartlett ou de Levene, que não são dados).

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

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

Uma advertência

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

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

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

Esta condição alternativa:

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

Considere a experiência seguidamente descrita.

Avaliou-se a largura das sépalas em 3 espécies de lírios num campo. Concretamente, de cada espécie foram aleatoriamente avaliadas 50 plantas.

Exercício 1: Indique qual o tipo de delineamento experimental em causa. Explicite o modelo correspondente e todas as hipóteses adicionais que sejam necessárias à consideração do problema em estudo.

Exercício 2: Para a largura das sépalas, eis os valores obtidos para as médias e variâncias para cada espécie e para a totalidade das observações:

Nível do factor	Média	Variância
setosa	3.428	0.143689796
versicolor	2.77	0.098469388
virginica	2.974	0.104004082

$s_y^2 = 0.189979418$

Construa a tabela-resumo da análise de variância correspondente a este caso.

Exemplo dos Lírios no SAS

```
proc glm data=iris PLOTS(UNPACK)=DIAGNOSTICS;
Class Species;
  model SepalWidth = Species/solution;
 lsmeans Species;
  means Species/ tukey alpha=0.05 cldiff lines;
 output out=residuals_data r=residual p=predicted;
run;
```

O quadro de síntese da ANOVA a 1 Factor

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

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

The SAS System

The GLM Procedure

Dependent Variable: SepalWidth

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	11.34493333	5.67246667	49.16	<.0001
Error	147	16.96200000	0.11538776		
Corrected Total	149	28.30693333			

Exercício 3: descreva, em pormenor, o teste aos efeitos do factor. O que conclui?

Neste caso, rejeita-se claramente a hipótese de que os acréscimos de nível, α_i, sejam todos nulos, pelo que se rejeita a hipótese de larguras médias de sépalas iguais em todas as espécies. Conclusão: o factor (espécie) afecta a variável resposta (largura da sépala).

Exercício 4: interprete os seguintes resultados:

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	2.974000000	B	0.04803910	61.91	<.0001
Species setosa	0.454000000	B	0.06793755	6.68	<.0001
Species versicol	-0.204000000	B	0.06793755	-3.00	0.0031
Species virginic	0.000000000	B	.	.	.

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

Exercício 5: a) Intreprete os seguintes resultados:

Species	SepalWidth LSMEAN
setosa	3.42800000
versicol	2.77000000
virginic	2.97400000

The GLM Procedure

Tukey's Studentized Range (HSD) Test for SepalWidth

Note: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	147
Error Mean Square	0.115388
Critical Value of Studentized Range	3.34842
Minimum Significant Difference	0.1609

Comparisons significant at the 0.05 level are indicated by ***.

Species Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
setosa - virginic	0.45400	0.29314	0.61486	***
setosa - versicol	0.65800	0.49714	0.81886	***
virginic - setosa	-0.45400	-0.61486	-0.29314	***
virginic - versicol	0.20400	0.04314	0.36486	***
versicol - setosa	-0.65800	-0.81886	-0.49714	***
versicol - virginic	-0.20400	-0.36486	-0.04314	***

Exercício 5: b) Intreprete os seguintes resultados:

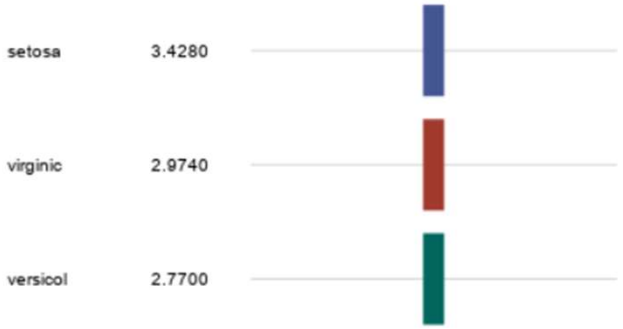
Tukey's Studentized Range (HSD) Test for SepalWidth

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

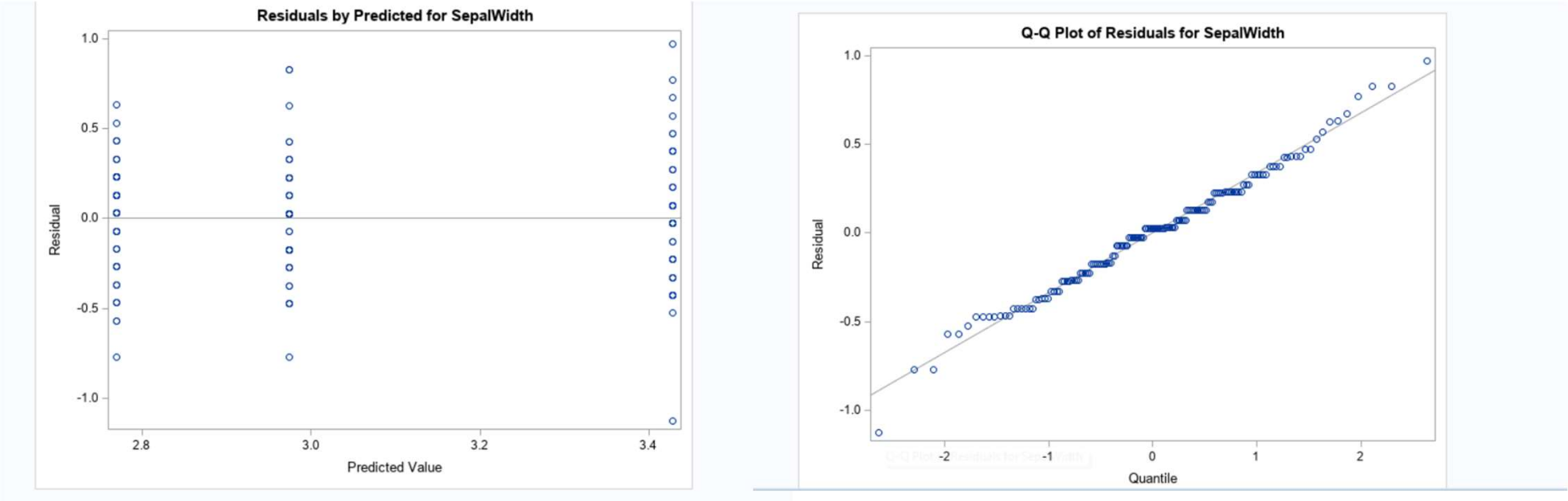
Alpha	0.05
Error Degrees of Freedom	147
Error Mean Square	0.115388
Critical Value of Studentized Range	3.34842
Minimum Significant Difference	0.1609

SepalWidth Tukey Grouping for Means of Species (Alpha = 0.05)

Means covered by the same bar are not significantly different.



Exercício 6: a) Comente os seguintes gráficos



Exercício 6: b) Identifique o nível do factor onde foi feita a observação cujo resíduo é, em módulo, mais elevado.