Análise de variância a 2 factores de efeitos fixos

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

Representação delineamento factorial (2 factores)

Um delineamento factorial é um delineamento em que há observações para todas as possíveis combinações de níveis de cada factor.

		Factor B				
	Níveis	B_1	B_2	B_3	…	B_b
FACTOR A	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 situação experimental. Nesta esquematização, há *ab* células, cada uma com 3 observações.

Modelos ANOVA a 2 Factores: notação

Admita-se a existência de:

- Uma variável resposta Y;
- Um Factor A, com a níveis;
- Um Factor B, com b níveis;
- n observações, com pelo menos uma em cada uma das ab situações experimentais (células).

O número de observações na célula correspondente ao nível i do factor A, e j do factor B é representado por n_{ij}.

O número total de observações é: $n = \sum_{i=1}^{a} \sum_{j=1}^{b} n_{ij}$.

Notação

Cada observação da variável resposta é identificada com três índices,

$$Y_{ijk}$$

onde:

- i indica o nível i do Factor A $\quad (i = 1, 2, ..., a)$.
- j indica o nível j do Factor B $\quad (j = 1, 2, ..., b)$.
- k indica a repetição k na célula (i, j) $\quad (k = 1, 2, ..., n_{ij})$.

Delineamento equilibrado

Se o número de observações for igual em todas as células, $n_{ij} = n_c$, $\forall i, j$, estamos perante um delineamento equilibrado.

Estudaremos dois diferentes modelos ANOVA para um delineamento factorial com 2 factores.

Modelo ANOVA a 2 factores (sem interacção)

Um primeiro modelo prevê a existência de dois diferentes tipos de efeitos associados aos níveis de cada factor. Admite-se que o valor esperado de cada observação Y_{ijk} é da forma:

$$E[Y_{ijk}] \quad = \quad \mu_{ij} \quad = \quad \mu + \alpha_i + \beta_j \quad , \quad \forall i,j,k.$$

O parâmetro μ é comum a todas as observações.

Cada parâmetro α_i é 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 é um acréscimo que pode diferir entre níveis do Factor B, e é designado o efeito do nível j do factor B.

Admite-se que todos estes parâmetros são constantes.

Admite-se que a variação de Y_{ijk} em torno do seu valor médio é aleatória e dada por um erro aleatório aditivo, ε_{ijk} (com $E[\varepsilon_{ijk}] = 0$):

$$Y_{ijk} \quad = \quad \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \ ,$$

As variáveis indicatrizes de nível de cada factor

A equação de base do modelo ANOVA a 2 factores (sem interacção) também pode ser escrita na forma vectorial, recorrendo a variáveis indicatrizes de pertença a cada nível de cada factor.

- $\vec{\mathbf{Y}}$ o vector aleatório *n*-dimensional com a totalidade das observações da variável resposta.
- $\vec{\mathbf{1}}_n$ o vector de *n* uns.
- $\vec{\mathscr{I}}_{A_i}$ a variável indicatriz de pertença ao nível *i* do Factor A.
- $\vec{\mathscr{I}}_{B_j}$ a variável indicatriz de pertença ao nível *j* do Factor B.
- $\vec{\boldsymbol{\varepsilon}}$ o vector aleatório dos *n* erros aleatórios.

A equação-base em notação vectorial (cont.)

Se se admitissem efeitos para todos os níveis de ambos os factores, temos a equação-base:

$$\vec{\mathbf{Y}} = \mu\vec{\mathbf{1}}_n + \alpha_1\vec{\mathscr{I}}_{A_1} + \alpha_2\vec{\mathscr{I}}_{A_2} + \ldots + \alpha_a\vec{\mathscr{I}}_{A_a} + \beta_1\vec{\mathscr{I}}_{B_1} + \beta_2\vec{\mathscr{I}}_{B_2} + \ldots + \beta_b\vec{\mathscr{I}}_{B_b} + \vec{\boldsymbol{\varepsilon}}$$

A matriz do modelo **X** definida com base nesta equação teria como colunas os vectores $\vec{\mathbf{1}}_n$, $\vec{\mathscr{I}}_{A_1}$, $\vec{\mathscr{I}}_{A_2}$, ..., $\vec{\mathscr{I}}_{A_a}$, $\vec{\mathscr{I}}_{B_1}$, $\vec{\mathscr{I}}_{B_2}$, ..., $\vec{\mathscr{I}}_{B_b}$.

Nessa matriz haveria dependências lineares por duas diferentes razões:

- a soma das indicatrizes do Factor A daria a coluna dos uns, $\vec{\mathbf{1}}_n$;
- a soma das indicatrizes do Factor B daria a coluna dos uns, $\vec{\mathbf{1}}_n$.

Agora, são necessárias duas restrições aos parâmetros, não podendo estimar-se parâmetros α_i e β_j para todos os níveis de cada Factor.

A matriz **X** sem restrições no modelo

$$
\mathbf{X} = \left[\begin{array}{c|ccc|ccc|ccc}
1 & 1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\
1 & 1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\
1 & 1 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \\
1 & 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \\
\hline
1 & 0 & 1 & \cdots & 0 & 1 & 0 & \cdots & 0 \\
1 & 0 & 1 & \cdots & 0 & 1 & 0 & \cdots & 0 \\
1 & 0 & 1 & \cdots & 0 & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 1 \\
1 & 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 1 \\
\hline
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\hline
1 & 0 & 0 & \cdots & 1 & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 1 \\
1 & 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 1 \\
\uparrow & \uparrow & \uparrow & & \uparrow & \uparrow & \uparrow & & \uparrow \\
\vec{1}_n & \vec{\mathscr{I}}_{A_1} & \vec{\mathscr{I}}_{A_2} & \cdots & \vec{\mathscr{I}}_{A_a} & \vec{\mathscr{I}}_{B_1} & \vec{\mathscr{I}}_{B_2} & \cdots & \vec{\mathscr{I}}_{B_b}
\end{array}\right]
$$

A exclusão da coluna $\vec{1}_n$ não resolve o problema.

Equação em notação vectorial, com restrições

Excluímos da equação do modelo as parcelas associadas ao primeiro nível de cada Factor, isto é, impõem-se as duas restrições:

$$\alpha_1 = 0 \qquad \text{e} \qquad \beta_1 = 0 \ ,$$

o que corresponde a excluir as colunas $\vec{\mathscr{I}}_{A_1}$ e $\vec{\mathscr{I}}_{B_1}$ da matriz **X**.

A equação-base do modelo ANOVA a 2 Factores, sem interacção, fica:

$$\vec{\mathbf{Y}} = \mu \vec{\mathbf{1}}_{\mathrm{n}} + \alpha_2 \vec{\mathscr{I}}_{A_2} + \ldots + \alpha_a \vec{\mathscr{I}}_{A_a} + \beta_2 \vec{\mathscr{I}}_{B_2} + \ldots + \beta_b \vec{\mathscr{I}}_{B_b} + \vec{\boldsymbol{\varepsilon}}$$

O parâmetro μ fica o valor esperado das observações na célula $(1,1)$:

$$Y_{11k} = \mu + \varepsilon_{11k} \qquad \Rightarrow \qquad E[Y_{11k}] = \mu = \mu_{11} \ .$$

A matriz do delineamento na ANOVA a 2 Factores (sem interacção), com as restrições $\alpha_1 = 0$ e $\beta_1 = 0$

$$
\mathbf{X} = \left[\begin{array}{c|ccc|ccc}
1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\
1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\
1 & 0 & \cdots & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & \cdots & 0 & 0 & \cdots & 1 \\
1 & 0 & \cdots & 0 & 0 & \cdots & 1 \\
\hline
1 & 1 & \cdots & 0 & 0 & \cdots & 0 \\
1 & 1 & \cdots & 0 & 0 & \cdots & 0 \\
1 & 1 & \cdots & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
1 & 1 & \cdots & 0 & 0 & \cdots & 1 \\
1 & 1 & \cdots & 0 & 0 & \cdots & 1 \\
\hline
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\hline
1 & 0 & \cdots & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & \cdots & 1 & 0 & \cdots & 1 \\
1 & 0 & \cdots & 1 & 0 & \cdots & 1
\end{array}\right]
\begin{array}{c}
\\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\
\uparrow \\ \vec{1}_n
\end{array}
\begin{array}{ccc}
\\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\
\uparrow & & \uparrow \\ \vec{\mathcal{I}}_{A_2} & \cdots & \vec{\mathcal{I}}_{A_a}
\end{array}
\begin{array}{ccc}
\\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\
\uparrow & & \uparrow \\ \vec{\mathcal{I}}_{B_2} & \cdots & \vec{\mathcal{I}}_{B_b}
\end{array}
$$

O modelo ANOVA a dois factores, sem interacção

Juntando os pressupostos necessários à inferência,

Modelo ANOVA a dois factores, sem interacção

Existem n observações, Y_{ijk}, n_{ij} das quais associadas à célula (i,j) $(i=1,...,a;\ j=1,...,b)$. Tem-se:

1. $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \varepsilon_{ijk}$, $\forall i=1,...,a;\ j=1,...,b;\ k=1,...,n_{ij}$ $\quad (\alpha_1 = 0; \beta_1 = 0)$.
2. $\varepsilon_{ijk} \frown \mathcal{N}(0, \sigma^2)$, $\quad \forall i,j,k$
3. $\{\varepsilon_{ijk}\}_{i,j,k}$ v.a.s independentes.

O modelo tem $a+b-1$ parâmetros desconhecidos:

- o parâmetro μ_{11};
- os $a-1$ acréscimos α_i $(i>1)$; e
- os $b-1$ acréscimos β_j $(j>1)$.

Testando a existência de efeitos

Um teste de ajustamento global do modelo tem como hipótese nula que todos os efeitos, quer do factor A, quer do Factor B são simultaneamente nulos, mas não distingue entre os efeitos de cada factor.

Mais útil será testar separadamente a existência dos efeitos de cada factor. Seria útil dispôr de dois testes, para as hipóteses:

- Teste I: $H_0 : \alpha_i = 0\,, \quad \forall i = 2, ..., a$;
- Teste II: $H_0 : \beta_j = 0\,, \quad \forall j = 2, ..., b$.

Teste aos efeitos do Factor B

O modelo ANOVA a 2 Factores, sem interacção

$$\vec{\mathbf{Y}} = \mu\vec{\mathbf{1}}_{\mathrm{n}} + \alpha_2\vec{\mathscr{I}}_{A_2} + \ldots + \alpha_a\vec{\mathscr{I}}_{A_a} + \beta_2\vec{\mathscr{I}}_{B_2} + \ldots + \beta_b\vec{\mathscr{I}}_{B_b} + \vec{\boldsymbol{\varepsilon}}$$

Sendo um Modelo Linear pode-se aplicar a teoria conhecida para este tipo de modelos e testar as hipóteses:

$$H_0 : \beta_j = 0, \quad \forall j = 2, \ldots, b \qquad \text{vs.} \qquad H_1 : \exists j \quad \text{tal que } \beta_j \neq 0 \ ,$$

através dum teste F parcial comparando o modelo completo

$$(\text{Modelo } M_{A+B}) \qquad Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \varepsilon_{ijk} \ ,$$

com o submodelo de equação de base

$$(\text{Modelo } M_A) \qquad Y_{ijk} = \mu_{11} + \alpha_i + \varepsilon_{ijk} \ ,$$

que é um modelo ANOVA a 1 Factor (factor A).

A construção do teste aos efeitos do Factor B

Assim,

- Ajusta-se o modelo completo M_{A+B} e o submodelo M_A.
- Obtêm-se as respectivas Somas de Quadrados Residuais, que designamos $SQRE_{A+B}$ e $SQRE_A$.
- Efectua-se o teste F parcial indicado. A estatística de teste é:

$$\text{(Efeitos Factor B)} \quad F = \frac{\overbrace{\frac{SQRE_A - SQRE_{A+B}}{b-1}}^{=SQB}}{\frac{SQRE_{A+B}}{n-(a+b-1)}} = \frac{QMB}{QMRE}$$

  definindo $QMB = \frac{SQB}{b-1} = \frac{SQRE_A - SQRE_{A+B}}{b-1}$.

- F tem distribuição $F_{[b-1,\,n-(a+b-1)]}$ sob $H_0 : \beta_j = 0,\ \forall j$.

A construção do teste aos efeitos do Factor A

Consideremos também um teste aos efeitos do Factor A, definido de forma um pouco diferente.

Defina-se:

- $SQA = SQF_A$, a Soma de Quadrados do Factor no Modelo M_A;
- $QMA = \frac{SQA}{a-1}$, o Quadrado Médio do Factor no Modelo M_A;
- $SQRE_{A+B}$ e $QMRE = \frac{SQRE_{A+B}}{n-(a+b-1)}$, como antes.

É possível provar que, caso $\alpha_i = 0$, $\forall_{i=2,\ldots,a}$, a estatística

$$F = \frac{QMA}{QMRE} = \frac{\frac{SQA}{a-1}}{\frac{SQRE_{A+B}}{n-(a+b-1)}}$$

tem distribuição $F_{(a-1,n-(a+b-1))}$.

O Teste F aos efeitos do factor A

Sendo válido o Modelo de ANOVA a dois factores, sem interacção:

Teste F aos efeitos do factor A

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

[A NÃO AFECTA Y] vs. [A AFECTA Y]

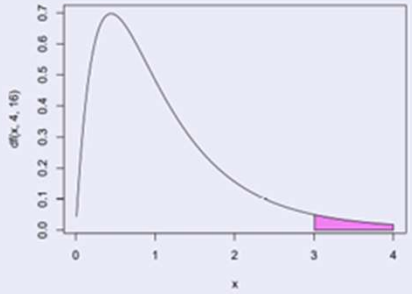
Estatística do Teste: $F = \frac{QMA}{QMRE} \frown F_{(a-1,n-(a+b-1))}$ se H_0.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se

$F_{calc} > f_{\alpha(a-1,n-(a+b-1))}$



O Teste F aos efeitos do factor B

Sendo válido o Modelo de ANOVA a dois factores, sem interacção:

Teste F aos efeitos do factor B

Hipóteses: $H_0 : \beta_j = 0 \quad \forall j=2,\ldots,b$ vs. $H_1 : \exists j=2,..,b$ t.q. $\beta_j \neq 0$.

[B NÃO AFECTA Y] vs. [B AFECTA Y]

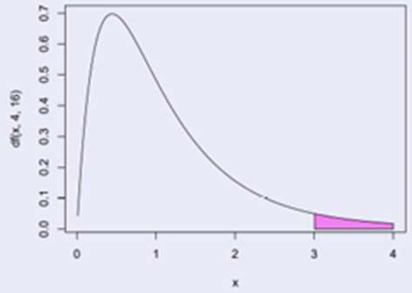
Estatística do Teste: $F = \frac{QMB}{QMRE} \frown F_{(b-1,\,n-(a+b-1))}$ se H_0.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se

$F_{calc} > f_{\alpha(b-1,n-(a+b-1))}$



A nova decomposição de SQT

Tendo em conta as Somas de Quadrados antes definidas, tem-se:

$$\begin{aligned} SQB &= SQRE_A - SQRE_{A+B} \\ SQA &= SQF_A = SQT - SQRE_A \end{aligned}$$

Somando estas SQs a $SQRE_{A+B}$, obtém-se:

A decomposição de SQT

$$SQA + SQB + SQRE_{A+B} = SQT$$

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.

Quadro-resumo ANOVA a 2 Factores (sem interacção)

Fonte	g.l.	SQ	QM	f_{calc}
Factor A	$a-1$	$SQA = SQF_A$	$QMA = \frac{SQA}{a-1}$	$\frac{QMA}{QMRE}$
Factor B	$b-1$	$SQB = SQRE_A - SQRE_{A+B}$	$QMB = \frac{SQB}{b-1}$	$\frac{QMB}{QMRE}$
Resíduos	$n-(a+b-1)$	$SQRE = SQRE_{A+B}$	$QMRE = \frac{SQRE}{n-(a+b-1)}$	
Total	$n-1$	$SQT = (n-1)\,s_y^2$	–	–

Trocando a ordem dos factores

Atenção: A forma como foram definidas as Somas de Quadrados de cada factor é diferente: $SQB = SQRE_A - SQRE_{A+B}$ e $SQA = SQF_A$.

A troca do papel dos factores A e B produz resultados diferentes em delineamentos não equilibrados. Designando por M_B o modelo ANOVA a um factor, mas apenas com o factor que temos chamado B, tem-se:

$$\begin{aligned} SQB &= SQF_B = SQT - SQRE_B \\ SQA &= SQRE_B - SQRE_{A+B} \, . \end{aligned}$$

Continua a ser verdade que SQT se pode decompor na forma

$$SQT = SQA + SQB + SQRE_{A+B} \, .$$

Justificam-se testes análogos aos dos acetatos 285 e 286.

Mas as duas formas alternativas de definir SQA e SQB apenas produzem resultados iguais no caso de delineamentos equilibrados, pelo que só nesse caso a ordem dos factores é arbitrária.

As várias médias amostrais

Sejam, num delineamento equilibrado:

- $\overline{Y}_{i..}$ a média amostral das $b\,n_c$ observações do nível i do Factor A, $\overline{Y}_{i..} = \frac{1}{b\,n_c} \sum_{j=1}^{b} \sum_{k=1}^{n_c} Y_{ijk}$
- $\overline{Y}_{.j.}$ a média amostral das $a\,n_c$ observações do nível j do Factor B, $\overline{Y}_{.j.} = \frac{1}{a\,n_c} \sum_{i=1}^{a} \sum_{k=1}^{n_c} Y_{ijk}$
- $\overline{Y}_{...}$ a média amostral da totalidade das $n = a\,b\,n_c$ observações, $\overline{Y}_{...} = \frac{1}{n} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_c} Y_{ijk}$.

SQA e *SQB* em delineamentos equilibrados

Num delineamento equilibrado, SQA é igual à Soma de Quadrados do Factor (SQF_A) do Modelo M_A, apenas com o Factor A

Nesse modelo, os valores ajustados são $\hat{Y}_{ijk} = \overline{Y}_{i..}$ (acetato 240). Assim, num delineamento equilibrado, tem-se:

$$SQF_A = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_c}(\underbrace{\hat{Y}_{ijk}}_{=\overline{Y}_{i..}} - \overline{Y}_{...})^2 = b\,n_c \cdot \sum_{i=1}^{a}(\overline{Y}_{i..} - \overline{Y}_{...})^2 = SQA\ .$$

Da mesma forma, num delineamento equilibrado, *SQB* é a Soma de Quadrados do Factor (SQF_B) do Modelo M_B, apenas com o Factor B. Nesse modelo, os valores ajustados são $\hat{Y}_{ijk} = \overline{Y}_{.j.}$, logo:

$$SQF_B = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_c}(\underbrace{\hat{Y}_{ijk}}_{=\overline{Y}_{.j.}} - \overline{Y}_{...})^2 = a\,n_c \cdot \sum_{j=1}^{b}(\overline{Y}_{.j.} - \overline{Y}_{...})^2 = SQB\ .$$

Fórmulas para delineamentos equilibrados (cont.)

Se o delineamento é equilibrado, ou seja, $n_{ij} = n_c \ , \ \forall i,j \ ,$ tem-se:

- $\hat{\mu}_{11} = \overline{Y}_{1..} + \overline{Y}_{.1.} - \overline{Y}_{...}$
- $\hat{\alpha}_i = \overline{Y}_{i..} - \overline{Y}_{1..}$
- $\hat{\beta}_j = \overline{Y}_{.j.} - \overline{Y}_{.1.}$

Tendo em conta a equação base do Modelo, os valores ajustados de cada observação dependem apenas das médias dos respectivos níveis em cada factor e da média geral de todas as observações:

$$\hat{Y}_{ijk} = \hat{\mu}_{11} + \hat{\alpha}_i + \hat{\beta}_j = \overline{Y}_{i..} + \overline{Y}_{.j.} - \overline{Y}_{...} \quad , \quad \forall i,j,k$$

Aviso: Ao contrário do que sucede na ANOVA a um factor, os valores ajustados $\hat{Y}_{ijk}$ não são a média das observações de Y na célula (i,j).

O quadro-resumo da ANOVA a 2 Factores (sem interacção; delineamento equilibrado)

Fonte	g.l.	SQ	QM	f_{calc}
Factor A	$a-1$	$SQA = b\,n_c \cdot \sum_{i=1}^{a} (\overline{y}_{i..} - \overline{y}_{...})^2$	$QMA = \frac{SQA}{a-1}$	$\frac{QMA}{QMRE}$
Factor B	$b-1$	$SQB = a\,n_c \cdot \sum_{j=1}^{b} (\overline{y}_{.j.} - \overline{y}_{...})^2$	$QMB = \frac{SQB}{b-1}$	$\frac{QMB}{QMRE}$
Resíduos	$n-(a+b-1)$	$SQRE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_c} [y_{ijk} - (\overline{y}_{i..} + \overline{y}_{.j.} - \overline{y}_{...})]^2$	$QMRE = \frac{SQRE}{n-(a+b-1)}$	
Total	$n-1$	$SQT = (n-1)\,s_y^2$	–	–

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 **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:

$$Y_{11k} = \mu_{11} + \underbrace{\alpha_1}_{=0} + \underbrace{\beta_1}_{=0} + \varepsilon_{11k} \qquad \Longrightarrow \qquad E[Y_{11k}] = \mu_{11}$$

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_i

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:

$$Y_{i1k} = \mu_{11} + \alpha_i + \underbrace{\beta_1}_{=0} + \varepsilon_{i1k} \qquad \Longrightarrow \qquad \mu_{i1} = E[Y_{i1k}] = \mu_{11} + \alpha_i$$

O parâmetro $\alpha_i = \mu_{i1} - \mu_{11}$ 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_i

Tabela com médias populacionais de célula (situação experimental):

		Factor B				
	Níveis	B_1	B_2	B_3	...	B_b
FACTOR A	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}
	⋮	⋮	⋮	⋮	⋱	⋮
	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:

$$Y_{1jk} = \mu_{11} + \underbrace{\alpha_1}_{=0} + \beta_j + \varepsilon_{1jk} \qquad \Longrightarrow \qquad \mu_{1j} = E[Y_{1jk}] = \mu_{11} + \beta_j$$

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				
	Níveis	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:

$$Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \varepsilon_{ijk} \qquad \Longrightarrow \qquad \mu_{ij} = E[Y_{ijk}] = \mu_{11} + \alpha_i + \beta_j.$$

Todas as parcelas destes valores esperados de Y já foram usados. Não há flexibilidade para descrever as médias 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., $n_{ij} = 1, \forall\, i,j$.

Modelos com interacção

Um modelo ANOVA a 2 Factores, sem interacção, foi considerado para um delineamento factorial, isto é, em que se cruzam todos os níveis de um e outro factor. Mas trata-se dum modelo pouco flexível.

Na presença de repetições nas células, a forma mais natural de modelar um delineamento com dois factores é a de prever a existência de um terceiro tipo de efeitos: os efeitos de interacção.

A ideia é incorporar na equação base do modelo para Y_{ijk} uma parcela $(\alpha\beta)_{ij}$ que permita que em cada célula haja um efeito específico associado à combinação dos níveis i do Factor A e j do Factor B:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \ .$$

Os valores esperados de Y_{ijk} (modelo com interacção)

Vamos admitir as seguintes restrições aos parâmetros:

$$\alpha_1 = 0 \quad ; \quad \beta_1 = 0 \quad ; \quad (\alpha\beta)_{1j} = 0 \ , \ \forall j \quad ; \quad (\alpha\beta)_{i1} = 0 \ , \ \forall i.$$

Tem-se, a partir da equação $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$:

- Para a primeira célula ($i = j = 1$): $\mu_{11} = E[Y_{11k}] = \mu$.
- Nas restantes células $(1, j)$ do primeiro nível do Factor A:
  $\mu_{1j} = E[Y_{1jk}] = \mu_{11} + \beta_j$.
- Nas restantes células $(i, 1)$ do primeiro nível do Factor B:
  $\mu_{i1} = E[Y_{i1k}] = \mu_{11} + \alpha_i$.
- Nas células genéricas (i, j), com $i > 1$ e $j > 1$,
  $\mu_{ij} = E[Y_{ijk}] = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij}$.

Os efeitos α_i e β_j designam-se efeitos principais de cada Factor.

Os valores esperados de Y_{ijk} (modelo com interacção)

Efeito das restrições $\alpha_1 = 0$; $\beta_1 = 0$; $(\alpha\beta)_{ij} = 0$ se $i = 1$ ou $j = 1$:

		Factor B				
	Níveis	B_1	B_2	B_3	...	B_b
FACTOR A	A_1	× × ×	× × ×	× × ×	...	× × ×
	A_2	× × ×	× × ×	× × ×	...	× × ×
	A_3	× × ×	× × ×	× × ×	...	× × ×
	⋮	⋮	⋮	⋮	⋱	⋮
	A_a	× × ×	× × ×	× × ×	...	× × ×

As observações que não estão associadas a A_1 (primeira linha) têm efeitos α_i.

As observações que não estão associadas a B_1 (primeira coluna) têm efeitos β_j.

As observações que não são da primeira coluna nem da primeira linha têm efeitos de interacção $(\alpha\beta)_{ij}$.

O modelo ANOVA a dois factores, com interacção

Juntando os pressupostos necessários à inferência,

Modelo ANOVA a dois factores, com interacção (Modelo M_{A*B})

Existem n observações, Y_{ijk}, n_{ij} das quais associadas à célula (i,j) $(i = 1,...,a; j = 1,...,b)$. Tem-se:

1. $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$, $\forall i{=}1,...,a \; ; \; j{=}1,...,b \; ; \; k{=}1,...,n_{ij}$
   $(\alpha_1{=}0 \; ; \; \beta_1{=}0 \; ; \; (\alpha\beta)_{ij}{=}0$, se $i{=}1$ e/ou $j{=}1)$.
2. $\varepsilon_{ijk} \frown \mathcal{N}(0, \sigma^2)$
3. $\{\varepsilon_{ijk}\}_{i,j,k}$ v.a.s independentes.

O modelo tem ab parâmetros desconhecidos:

- a 1 média da célula de referência, μ_{11};
- os $a-1$ acréscimos α_i $(i > 1)$;
- os $b-1$ acréscimos β_j $(j > 1)$; e
- os $(a-1)(b-1)$ efeitos de interacção $(\alpha\beta)_{ij}$, para $i > 1, j > 1$.

Variáveis indicatrizes de célula

A versão vectorial da equação do modelo com interacção associa os novos efeitos $(\alpha\beta)_{ij}$ a variáveis indicatrizes das respectivas células.

A equação-base do modelo ANOVA a 2 Factores, com interacção, é:

$$\begin{aligned}\vec{\mathbf{Y}} \;=\; & \mu\vec{\mathbf{1}}_{\mathrm{n}} + \alpha_2\vec{\mathscr{I}}_{A_2} + \ldots + \alpha_a\vec{\mathscr{I}}_{A_a} + \beta_2\vec{\mathscr{I}}_{B_2} + \ldots + \beta_b\vec{\mathscr{I}}_{B_b} + \\ & + (\alpha\beta)_{22}\vec{\mathscr{I}}_{A_2:B_2} + (\alpha\beta)_{23}\vec{\mathscr{I}}_{A_2:B_3} + \ldots + (\alpha\beta)_{ab}\vec{\mathscr{I}}_{A_a:B_b} + \vec{\boldsymbol{\varepsilon}}\end{aligned}$$

onde $\vec{\mathscr{I}}_{A_i:B_j}$ representa a variável indicatriz da célula correspondente ao nível *i* do Factor A e nível *j* do factor B.

Este modelo com *ab* parâmetros é designado modelo M_{A*B}

Modelo ANOVA a 2 factores, com interacção (cont.)

A matriz **X** do delineamento é agora constituída por ab colunas:

- uma coluna de uns, $\vec{\mathbf{1}}_n$, associada ao parâmetro μ_{11}.
- $a-1$ colunas de indicatrizes de nível do factor A, $\vec{\mathscr{I}}_{A_i}$, $(i>1)$, associadas aos parâmetros α_i.
- $b-1$ colunas de indicatrizes de nível do factor B, $\vec{\mathscr{I}}_{B_j}$, $(j>1)$, associadas aos parâmetros β_j.
- $(a-1)(b-1)$ colunas de indicatrizes de célula, $\vec{\mathscr{I}}_{A_i:B_j}$, $(i,j>1)$, associadas aos efeitos de interacção $(\alpha\beta)_{ij}$.

Como em modelos anteriores, $\hat{\vec{\mathbf{Y}}} = \mathbf{H}\vec{\mathbf{Y}}$, sendo **H** a matriz que projecta ortogonalmente sobre o espaço $\mathscr{C}(\mathbf{X})$ gerado pelas colunas desta matriz **X**.

E também, $SQRE_{A*B} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(Y_{ijk}-\hat{Y}_{ijk})^2$.

Os três testes ANOVA

Neste delineamento, desejamos fazer um teste à existência de cada um dos três tipos de efeitos:

- Teste I: $H_0 : (\alpha\beta)_{ij} = 0\,, \quad \forall i = 2,...,a\ ,\ \forall j = 2,...,b$;
- Teste II: $H_0 : \alpha_i = 0\,, \quad \forall i = 2,...,a$; e
- Teste III: $H_0 : \beta_j = 0\,, \quad \forall j = 2,...,b$.

As estatísticas de teste para cada um destes três testes obtêm-se a partir da decomposição da Soma de Quadrados Total (ou seja, da *análise da variancia*) em parcelas convenientes.

Testando efeitos de interacção

Para testar a existência de efeitos de interacção,

$$H_0 : (\alpha\beta)_{ij} = 0\,, \quad \forall i = 2,...,a\ ,\ \forall j = 2,...,b\ ,$$

pode efectuar-se um teste F parcial comparando o modelo

$$(\text{Modelo } M_{A*B}) \qquad Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}\ ,$$

com o submodelo sem efeitos de interacção

$$(\text{Modelo } M_{A+B}) \qquad Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \varepsilon_{ijk}\ ,$$

Designa-se Soma de Quadrados associada à interacção à diferença

$$SQAB = SQRE_{A+B} - SQRE_{A*B}$$

Testando os efeitos principais de cada Factor

Para testar os efeitos principais dos Factor B ($H_0 : \beta_j = 0\,,\ \forall j = 2, ..., b$) e do Factor A ($H_0 : \alpha_i = 0\,,\ \forall i = 2, ..., a$) pode partir-se dos modelos

$$\begin{aligned} (\text{Modelo } M_{A+B}) \qquad Y_{ijk} &= \mu_{11} + \alpha_i + \beta_j + \varepsilon_{ijk} \\ (\text{Modelo } M_A) \qquad Y_{ijk} &= \mu_{11} + \alpha_i + \varepsilon_{ijk}\ , \end{aligned}$$

Defina-se:

$$\begin{aligned} SQB &= SQRE_A - SQRE_{A+B} \\ SQA &= SQF_A = SQT - SQRE_A \end{aligned}$$

Nota: Estas duas Somas de Quadrados definem-se da mesma forma que no modelo sem efeitos de interacção.

A decomposição de SQT

Definimos :

$$\begin{aligned} SQAB &= SQRE_{A+B} - SQRE_{A*B} \\ SQB &= SQRE_{A} - SQRE_{A+B} \\ SQA &= SQF_A = SQT - SQRE_A \end{aligned}$$

Somando estas Somas de Quadrados a $SQRE_{A*B}$, obtém-se:

$$SQT = SQRE_{A*B} + SQAB + SQA + SQB$$

Esta decomposição de SQT gera as quantidades nas quais se baseiam as estatísticas dos três testes associados ao Modelo M_{A*B}.

O quadro-resumo

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 efeito são o número de parâmetros desse tipo que sobram após a imposição das restrições.

Como em qualquer modelo linear, 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]

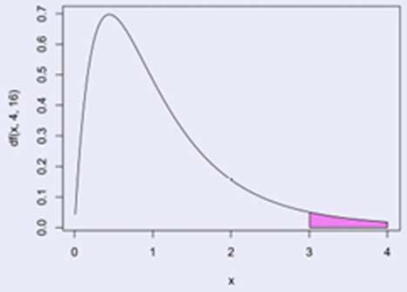
Estatística do Teste: $F = \frac{QMAB}{QMRE} \frown F_{((a-1)(b-1),n-ab)}$ se H_0.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se

$F_{calc} > f_{\alpha((a-1)(b-1),n-ab)}$



O Teste F aos efeitos principais do factor A

Sendo válido o Modelo ANOVA a 2 factores com interacção tem-se:

Teste F aos efeitos principais do factor A

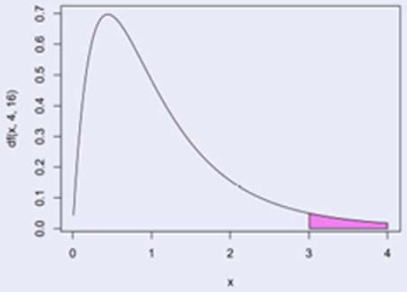
Hipóteses: $H_0 : \alpha_i = 0 \quad \forall i=2,...,a$ vs. $H_1 : \exists i=2,..,a \text{ t.q. } \alpha_i \neq 0.$
[$\nexists$ EFEITOS DE A] vs. [$\exists$ EFEITOS DE A]

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

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

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



O Teste F aos efeitos principais do factor B

Sendo válido o Modelo ANOVA a 2 factores com interacção tem-se:

Teste F aos efeitos principais do factor B

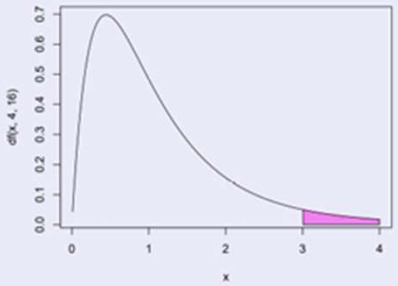
Hipóteses: $H_0 : \beta_j = 0 \quad \forall j=2,\ldots,b$ vs. $H_1 : \exists j=2,..,b \text{ t.q. } \beta_j \neq 0.$
[$\nexists$ EFEITOS DE B] vs. [$\exists$ EFEITOS DE B]

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

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

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



Estimação da interacção necessita de repetições

Para se poder estudar efeitos de interacção, é necessário que haja repetições nas células.

Os graus de liberdade do *SQRE* neste modelo são $n - ab$. Se houver uma única observação em cada célula, tem-se $n = ab$, ou seja, tantos parâmetros quantas as observações existentes. Nesse caso, nem sequer será possível definir o Quadrado Médio Residual, *QMRE*.

Num delineamento com uma única observação por célula é obrigatório optar por um modelo sem interacção.

Havendo repetições, é mais natural considerar um modelo com interacção e deixar que a conclusão sobre a existência, ou não, desse tipo de efeitos resulte do estudo do modelo.

Não constando do modelo, eventuais efeitos de interacção irão inflacionar a variabilidade residual, não explicada pelo modelo.

Valores ajustados de Y no modelo com interacção

Às médias já definidas no estudo do modelo a dois Factores, sem efeitos de interacção, (acetato 292):

- $\overline{Y}_{i..}$ - nível i do Factor A;
- $\overline{Y}_{.j.}$ - nível j do Factor B;
- $\overline{Y}_{...}$ - global;

acrescentam-se agora as médias de cada célula:

$$\overline{Y}_{ij.} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} Y_{ijk} \ .$$

Os valores ajustados $\hat{Y}_{ijk}$ são iguais para todas as observações numa mesma célula, e são dados pela média amostral da célula:

$$\hat{Y}_{ijk} = \overline{Y}_{ij.} \ .$$

Estimadores de parâmetros

Os estimadores dos parâmetros num modelo ANOVA a 2 Factores, com interacção, são dadas pelas quantidades amostrais correspondentes às definições populacionais de cada parâmetro (ver acetato 303):

- $\mu = \mu_{11} \quad \Rightarrow \quad \hat{\mu} = \hat{\mu}_{11} = \overline{Y}_{11\cdot}$
- $\alpha_i = \mu_{i1} - \mu_{11} \quad \Rightarrow \quad \hat{\alpha}_i = \overline{Y}_{i1\cdot} - \overline{Y}_{11\cdot} \qquad (i > 1)$
- $\beta_j = \mu_{1j} - \mu_{11} \quad \Rightarrow \quad \hat{\beta}_j = \overline{Y}_{1j\cdot} - \overline{Y}_{11\cdot} \qquad (j > 1)$
- $(\alpha\beta)_{ij} = \mu_{ij} - \cancel{\mu_{11}} - \underbrace{\alpha_i}_{=\mu_{i1}-\cancel{\mu_{11}}} - \underbrace{\beta_j}_{=\mu_{1j}-\mu_{11}} = \mu_{ij} + \mu_{11} - \mu_{i1} - \mu_{1j}$

  $\Rightarrow \quad \widehat{(\alpha\beta)}_{ij} = (\overline{Y}_{ij\cdot} + \overline{Y}_{11\cdot}) - (\overline{Y}_{i1\cdot} + \overline{Y}_{1j\cdot}) \qquad (i,j > 1)$

Intervalos de confiança ou testes de hipóteses para qualquer parâmetro individual, ou combinações lineares desses parâmetros, podem ser efectuados utilizando a teoria geral do Modelo Linear.

Soma de Quadrados Residual

Como os valores ajustados correspondem às medias amostrais da célula onde se efectuaram as observações, $\hat{Y}_{ijk} = \overline{Y}_{ij.}$, tem-se:

$$SQRE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} (Y_{ijk} - \hat{Y}_{ijk})^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} (Y_{ijk} - \overline{Y}_{ij.})^2$$

$$\Leftrightarrow \quad SQRE = \sum_{i=1}^{a}\sum_{j=1}^{b} (n_{ij} - 1) S_{ij}^2 \ ,$$

sendo S_{ij}^2 a variância amostral das observações de Y na 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:

$$QMRE = \frac{SQRE}{n - ab} = \frac{\cancel{n_c - 1}}{ab(\cancel{n_c - 1})} \sum_{i=1}^{a}\sum_{j=1}^{b} S_{ij}^2 = \frac{1}{ab} \sum_{i=1}^{a}\sum_{j=1}^{b} S_{ij}^2 \ .$$

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 correspondem (tal como no modelo sem efeitos de interacção) às Somas de Quadrados associadas a cada factor, caso se ajustasse (aos mesmos dados) um modelo ANOVA apenas com esse factor:

$$SQA = bn_c \sum_{i=1}^{a} (\overline{Y}_{i..} - \overline{Y}_{...})^2$$

$$SQB = an_c \sum_{j=1}^{b} (\overline{Y}_{.j.} - \overline{Y}_{...})^2$$

Comparações múltiplas de médias de células

Havendo *ab* células, a comparação das médias de cada par de células envolve $\binom{ab}{2}$ comparações.

O número potencialmente grande de comparações possíveis entre médias de célula aconselha a utilização de métodos de comparação múltipla, que permitam controlar globalmente o nível de significância do conjunto de testes de hipóteses (ou grau de confiança do conjunto de intervalos de confiança).

O mais utilizado dos métodos de comparação múltipla está associado ao nome de Tukey. Foi já introduzido no estudo de delineamentos a 1 Factor. Adapta-se facilmente à comparação múltipla de médias de células.

O Teste de Tukey

Teste de Tukey para médias de células

Admite-se que o delineamento é equilibrado, com $n_c > 1$ repetiçoes em todas as ab células.

Rejeita-se a igualdade das médias das células (i,j) e (i',j'), a favor da hipótese $\mu_{ij} \neq \mu_{i'j'}$, se

$$|\overline{Y}_{ij\cdot} - \overline{Y}_{i'j'\cdot}| \quad > \quad q_{\alpha\,(ab,n-ab)} \cdot \sqrt{\frac{QMRE}{n_c}} \ ,$$

sendo $q_{\alpha\,(ab,n-ab)}$ o valor que deixa à direita uma região de probabilidade α numa distribuição de Tukey com parâmetros $k = ab$ (o número total de médias de célula) e $\nu = n - ab$ (os graus de liberdade associados ao $QMRE$).

Intervalos de Confiança para $\mu_{ij} - \mu_{i'j'}$

Intervalos de Confiança de Tukey

Com grau de confiança global $(1-\alpha) \times 100\%$, todas as diferenças de médias de pares de células, $\mu_{ij} - \mu_{i'j'}$, estão em intervalos da forma:

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

Conclui-se que $\mu_{ij} \neq \mu_{i'j'}$ se o intervalo correspondente a este par de células não contém o valor zero.

Visualização gráfica de efeitos de interacção

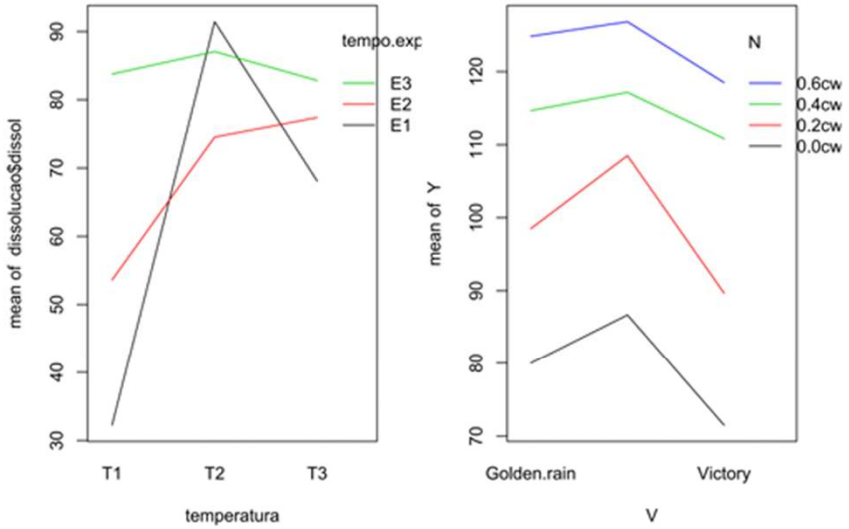
A existência de efeitos de interacção em delineamentos factoriais a dois factores transparece em gráficos onde:

- O eixo horizontal é associado aos níveis de um factor (e.g., *fA*);
- no eixo vertical são 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*).

A cada problema correspondem sempre dois possíveis gráficos de interacção, pois é arbitrária a escolha de qual o factor associado ao eixo horizontal, e qual o que define os pontos a serem unidos.

Como ler os gráficos de interacção

Havendo interacção, as linhas estarão longe de qualquer paralelismo (exemplo à esquerda). A inexistência de interacção significativa produz linhas aproximadamente "paralelas" (exemplo à direita).



A confirmação da significância dos efeitos de interacção exige que se efectue o respectivo teste *F*.

Análise dos Resíduos

A validade dos pressupostos do Modelo relativos aos erros aleatórios pode ser estudada de forma análoga ao que foi visto para um delineamento a 1 Factor.

Os resíduos relativos a uma mesma célula aparecem em *ab* colunas verticais num gráfico de E_{ijk} vs. $\hat{Y}_{ijk}$.

A hipótese de heterogeneidade de variâncias entre diferentes células pode ser testada recorrendo a testes de hipóteses (como o Teste de Bartlett), mas essa matéria não será leccionada.

Uma advertência

Na formulação clássica do modelo ANOVA a dois Factores, com interacção, e a partir da equação-base $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$, em vez de impor as condições $\alpha_1 = \beta_1 = (\alpha\beta)_{i1} = (\alpha\beta)_{1j} = 0$ $(\forall i,j)$, admitem-se as restrições:

- $\sum_i \alpha_i = 0$;
- $\sum_j \beta_j = 0$;
- $\sum_i (\alpha\beta)_{ij} = 0\ , \qquad \forall j$;
- $\sum_j (\alpha\beta)_{ij} = 0\ , \qquad \forall i$.

Estas condições alternativas:

- mudam a forma de interpretar os parâmetros;
- mudam os estimadores dos parâmetros;
- não mudam o resultado dos testes F à existência de efeitos.

Delineamentos factoriais com vários factores

Um delineamento factorial (isto é, com observações para todas as combinações de níveis de cada factor) pode ser definido com qualquer número de factores.

Num delineamento factorial a três factores – A, B e C – cada observação da variável resposta indexa-se com quatro índices: Y_{ijkl} indica a observação *l* no nível *i* do Factor A, nível *j* do Factor B e nível *k* do Factor C. A equação de base para Y_{ijkl} prevê a existência de sete tipos de efeitos:

- três efeitos principais de cada factor, α_i, β_j e γ_k.
- três efeitos de interacção dupla associados a cada combinação de níveis de dois Factores diferentes: $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$ e $(\beta\gamma)_{jk}$.
- um efeito de tripla interacção para as células onde se cruzam níveis dos três factores: $(\alpha\beta\gamma)_{ijk}$

O modelo factorial a três factores

A equação de base do modelo é agora:

$$Y_{ijkl} = \mu_{111} + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkl} \ .$$

A Soma de Quadrados Total é decomposta em oito parcelas: *SQA*, *SQB*, *SQC*, *SQAB*, *SQAC*, *SQBC*, *SQABC* e *SQRE*, de forma análoga ao visto antes.

Os graus de liberdade associados a cada tipo de efeito generalizam conceitos anteriores.

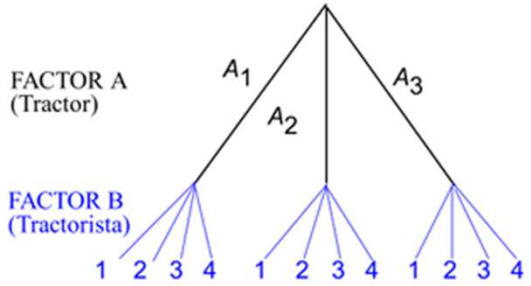
Há sete testes: um para cada tipo de efeitos. As estatísticas desses sete testes são todas do tipo $\frac{QMx}{QMRE}$, onde *x* designa o tipo de efeitos em questão.

As estatísticas desses testes terão, sob H_0, distribuição F com graus de liberdade dados pelos g.l. do numerador e do denominador, respectivamente.

Delineamentos hierarquizados (cont.)

Existe uma hierarquia dos factores: só identificamos os níveis de um factor (factor subordinado) após ter identificado o nível do outro factor (factor dominante) com que se trabalha.

	Tractor A_1	Tractor A_2	Tractor A_3
Tractorista A_1 1	×	-	-
Tractorista A_1 2	×	-	-
Tractorista A_1 3	×	-	-
Tractorista A_1 4	×	-	-
Tractorista A_2 1	-	×	-
Tractorista A_2 2	-	×	-
Tractorista A_2 3	-	×	-
Tractorista A_2 4	-	×	-
Tractorista A_3 1	-	-	×
Tractorista A_3 2	-	-	×
Tractorista A_3 3	-	-	×
Tractorista A_3 4	-	-	×



Um tal delineamento diz-se hierarquizado (*nested*, em inglês).

Um delineamento hierarquizado pode ser visto como um delineamento factorial (muito) incompleto. Deixa de fazer sentido falar em efeitos de interacção entre os níveis de cada Factor.

O modelo a 2 Factores, hierarquizados

Seja b_i o número de níveis do Factor B (folhas terminais do dendrograma), subordinados ao nível i do Factor A (ramo). b_i pode ser diferente para cada nível i do factor dominante.

Cada observação é representada por uma v.a. com três índices, Y_{ijk}:

- i nível do factor dominante ($i = 1, ..., a$);
- j nível do factor subordinado ($j = 1, ..., b_i$);
- k repetição para a célula (i,j), com $k = 1, ..., n_{ij}$.

A equação base do modelo inclui efeitos de nível do Factor A e efeitos de nível do factor B (subordinado):

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk} \ ,$$

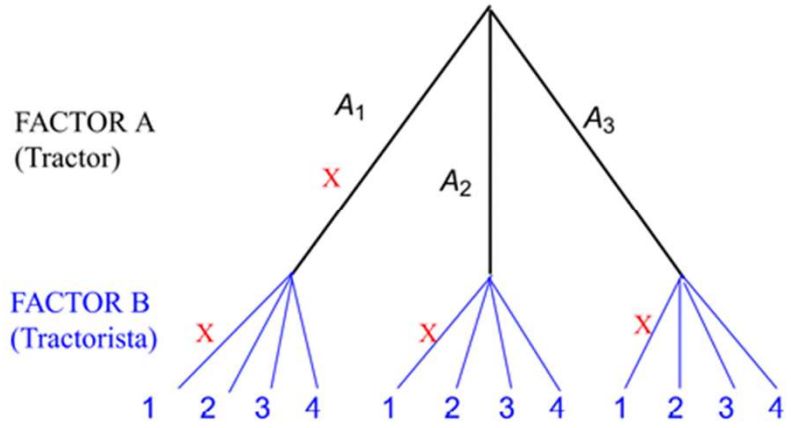
com $\alpha_1 = 0$ e $\beta_{1(i)} = 0, \forall i$. Com estas restrições, $\mu = \mu_{11}$.

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.

Restrições nos delineamentos hierarquizados

Cada ramo associado ao Factor dominante excepto o primeiro tem efeito α_i.

Cada folha terminal associada ao Factor subordinado excepto a primeira de cada ramo tem efeito $\beta_{j(i)}$.



Os valores esperados de Y_{ijk}

Tem-se:

- Para a primeira célula ($i = j = 1$): $E[Y_{11k}] = \mu = \mu_{11}$.
- Nas restantes células do primeiro nível do Factor A ($i = 1; j > 1$): $\mu_{1j} = E[Y_{1jk}] = \mu_{11} + \beta_{j(1)}$.
- Nos restantes primeiros níveis do factor B ($i > 1; j = 1$): $\mu_{i1} = E[Y_{i1k}] = \mu_{11} + \alpha_i$.
- Nas células genéricas (i,j), com $i > 1$ e $j > 1$, $\mu_{ij} = E[Y_{ijk}] = \mu_{11} + \alpha_i + \beta_{j(i)}$.

Os efeitos α_i e $\beta_{j(i)}$ designam-se efeitos dos níveis de cada Factor.

Variáveis indicatrizes e número de parâmetros

Como em modelos anteriores, a cada parâmetro associa-se uma variável indicatriz das observações correspondentes. Assim:

- um parâmetro μ_{11}, associado à coluna de uns, $\vec{\mathbf{1}}_n$.
- $(a-1)$ parâmetros α_i, associados às indicatrizes $\vec{\mathscr{I}}_{A_i}$ de cada nível $i > 1$ do Factor A.
- $\sum_{i=1}^{a} (b_i - 1)$ parâmetros $\beta_{j(i)}$, associados às indicatrizes $\vec{\mathscr{I}}_{B_{j(i)}}$ de cada nível $j > 1$ do Factor B, para $i = 1, ..., a$.

O no. de parâmetros é igual ao no. de situações experimentais:

$$1 + (a-1) + \sum_{i=1}^{a} (b_i - 1) = \cancel{1} + \cancel{a} - \cancel{1} + \sum_{i=1}^{a} b_i - \underbrace{\sum_{i=1}^{a} 1}_{=a} = \sum_{i=1}^{a} b_i$$

Se houver sempre $b = b_i$ níveis do Factor B, em cada nível i do Factor A, haverá ab parâmetros no modelo.

O modelo ANOVA a dois factores, hierarquizados

Juntando os pressupostos necessários à inferência,

Modelo ANOVA a dois factores, hierarquizados (Modelo $M_{A/B}$)

Seja A o Factor dominante e B o Factor subordinado.
Existem n observações, Y_{ijk}, n_{ij} das quais associadas à célula (i,j) $(i = 1,...,a\ ;\ j = 1,...,b_i)$. Tem-se:

1. $Y_{ijk} = \mu_{11} + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}\ , \quad \forall i{=}1,...,a\ ;\ j{=}1,...,b_i\ ;\ k{=}1,...,n_{ij}$
   $(\alpha_1 = 0\ ;\ \beta_{1(i)} = 0\ ,\ \forall i)$.
2. $\varepsilon_{ijk} \frown \mathcal{N}(0, \sigma^2)\ , \qquad \forall i,j,k$
3. $\{\varepsilon_{ijk}\}_{i,j,k}$ v.a.s independentes.

Os dois testes ANOVA

Neste delineamento, pretende-se testar a existência de cada um dos dois tipos de efeitos previstos no modelo:

- $H_0 : \alpha_i = 0, \quad \forall i = 2, ..., a$; e
- $H_0 : \beta_{j(i)} = 0, \quad \forall i = 1, ..., a$ e $j = 2, ..., b_i$.

As estatísticas de teste para cada um destes testes obtêm-se a partir da decomposição da Soma de Quadrados Total em três parcelas, correspondentes aos dois tipos de efeito e à variabilidade residual.

As Somas de Quadrados associadas a cada tipo de efeito definem-se de forma análoga à usada em delineamentos anteriores.

A decomposição de SQT

Para efectuar a decomposição da Soma de Quadrados Total, consideremos os modelos

$$\begin{aligned} (\text{Modelo } M_{A/B}) \qquad & Y_{ijk} = \mu_{11} + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk} \ , \\ (\text{Modelo } M_A) \qquad & Y_{ijk} = \mu_{11} + \alpha_i + \varepsilon_{ijk} \ , \end{aligned}$$

Designa-se Soma de Quadrados associada aos efeitos de B a

$$SQB(A) = SQRE_A - SQRE_{A/B}$$

e Soma de Quadrados associada aos efeitos de A a

$$SQA = SQF_A = SQT - SQRE_A$$

Juntamente com $SQRE_{A/B}$, tem-se:

$$SQT = SQA + SQB(A) + SQRE_{A/B}$$

Algumas fórmulas

Como $SQA = SQF_A$ (Modelo 1 Factor):

$$SQA = \sum_{i=1}^{a}\sum_{j=i}^{b_i}\sum_{k=1}^{n_{ij}}(\underbrace{\hat{Y}_{ijk}}_{=\overline{Y}_{i..}} - \overline{Y}_{...})^2 = \sum_{i=1}^{a}\sum_{j=i}^{b_i} n_{ij}\,(\overline{Y}_{i..} - \overline{Y}_{...})^2 \,.$$

Num delineamento equilibrado, tem-se: $SQA = n_c \sum_{i=1}^{a} b_i\,(\overline{Y}_{i..} - \overline{Y}_{...})^2$

No modelo a 2 factores hierarquizado também se tem:

$$\hat{Y}_{ijk} = \overline{Y}_{ij.}$$

Logo, a Soma de Quadrados Residual também é soma ponderada das variâncias de célula $S_{ij}^2 = \frac{1}{n_{ij}-1}\sum_{k=1}^{n_{ij}}(Y_{ijk} - \overline{Y}_{ij.})^2$:

$$SQRE = \sum_{i=1}^{a}\sum_{j=i}^{b_i}\sum_{k=1}^{n_{ij}}(Y_{ijk} - \underbrace{\hat{Y}_{ijk}}_{=\overline{Y}_{ij.}})^2 = \sum_{i=1}^{a}\sum_{j=i}^{b_i}(n_{ij}-1)\,S_{ij}^2 \,.$$

Graus de liberdade

Os graus de liberdade associados a cada tipo de efeito são dados por:

- $g.l.(SQA) = a - 1$, o número de parâmetros associados aos efeitos de nível de A.
- $g.l.[SQB(A)] = \sum_{i=1}^{a} (b_i - 1)$, o número de parâmetros associados aos efeitos de nível de B.
- $g.l.(SQRE) = n - \sum_{i=1}^{a} b_i$, o número de observações menos o número total de parâmetros do modelo.

Quadro-resumo da ANOVA a 2 Factores hierarquizados

Fonte	g.l.	SQ	QM	f_{calc}
Factor A	$a-1$	SQA	$QMA = \frac{SQA}{a-1}$	$\frac{QMA}{QMRE}$
Factor B(A)	$\sum_{i=1}^{a} (b_i - 1)$	$SQB(A)$	$QMB(A) = \frac{SQB(A)}{\sum_{i=1}^{a} (b_i - 1)}$	$\frac{QMB(A)}{QMRE}$
Resíduos	$n - \sum_{i=1}^{a} b_i$	$SQRE$	$QMRE = \frac{SQRE}{n - \sum_{i=1}^{a} b_i}$	
Total	$n-1$	$SQT = (n-1)\, S_y^2$	–	–

O Teste F aos efeitos do factor A (dominante)

Sendo válido o Modelo de ANOVA a 2 factores hierarquizados, tem-se:

Teste F aos efeitos do factor A (dominante)

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

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

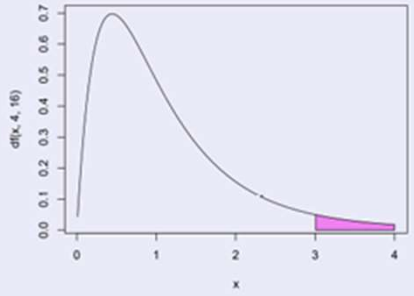
Estatística do Teste: $F = \frac{QMA}{QMRE} \frown F_{(a-1,\, n-\sum_i b_i)}$ se H_0.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se

$F_{calc} > f_{\alpha(a-1,\, n-\sum_i b_i)}$



O Teste F aos efeitos do factor B (subordinado)

Sendo válido o Modelo de ANOVA a dois factores hierarquizado,

Teste F aos efeitos do factor B (subordinado)

Hipóteses: $H_0 : \beta_{j(i)} = 0 \quad \forall j=2,\ldots,b_i \ , \ i=1,\ldots,a$ vs. $H_1 : \exists i,j \text{ t.q. } \beta_{j(i)} \neq 0$.

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

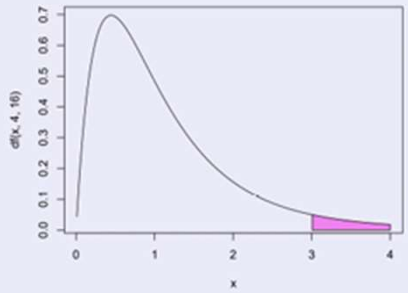
Estatística do Teste: $F = \frac{QMB(A)}{QMRE} \frown F_{(\sum_i (b_i-1),\, n-\sum_i b_i)}$ se H_0.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se

$F_{calc} > f_{\alpha(\sum_i (b_i-1),\, n-\sum_i b_i)}$



Comparações múltiplas de médias

Caso se conclua pela existência de efeitos do factor subordinado, é natural querer comparar médias da variável resposta nas $\sum_{i=1}^{a} b_i$ diferentes situações experimentais.

Comparações múltiplas de Tukey podem ser efectuadas, caso o delineamento seja equilibrado, isto é, se houver o mesmo número de observações em cada situação experimental.

Neste caso, os parâmetros da distribuição de Tukey serão

- o número de situações experimentais, $k = \sum_{i=1}^{a} b_i$; e
- os graus de liberdade associados ao $QMRE$, $\nu = n - \sum_{i=1}^{a} b_i$.

ANÁLISE DE COVARIÂNCIA DE EFEITOS FIXOS

Um exemplo de Análise de Covariância

A Regressão Linear e as Análises de Variância estudadas até aqui, são casos particulares do Modelo Linear, que inclui também as Análises de Covariância.

Em qualquer destas três situações se procura modelar uma variável resposta quantitativa (numérica) Y. O que distingue as três situações é a natureza das variáveis preditoras.

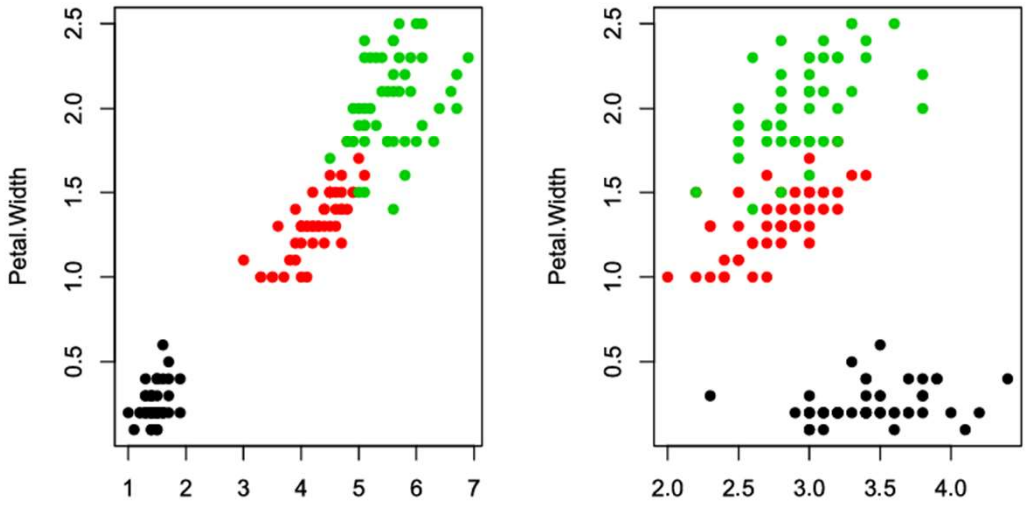
- Numa Regressão Linear, as variáveis preditoras são variáveis igualmente quantitativas (numéricas).
- Numa Análise de Variância, as variáves preditoras são factores (variáveis qualitativas, ou categóricas).
- Numa Análise de Covariância, entre as variáveis preditoras encontramos quer variáveis numéricas, quer factores.

Um exemplo de Análise de Covariância (cont.)

A Análise de Covariância será apenas vista no contexto dum problema específico de interesse prático, associado à Regressão Linear. Admita que se verificou ser válida uma regressão linear simples entre uma variável Y e um preditor x, num dado contexto. Surge de forma natural a questão de saber se a recta de regressão teórica é, ou não, idêntica, noutros contextos aparentados, ou seja, noutros níveis de um dado factor.

Um exemplo de Análise de Covariância (cont.)

No exemplo dos lírios (já considerado anteriormente), a relação entre Largura de Pétala e Comprimento de Pétala talvez (gráfico à esquerda) seja comum para as três espécies de lírios (*setosa*, *versicolor* e *virginica*). Já a relação entre Largura de Pétala e Largura de Sépala é claramente diferente para cada espécie (e até inexistente, enquanto relação linear, para o conjunto das três espécies - gráfico à direita):



Um exemplo de Análise de Covariância (cont.)

O problema em questão pode ser formulado como um problema de Análise de Covariância pois consiste no estudo duma relação linear entre *y* e *x*, mas influenciada também por uma variável qualitativa: o factor espécie, que tem três níveis, ou seja, três diferentes espécies.

O problema será formulado de tal forma que admitir a existência de uma única relação nas três espécies seja admitir a igualdade entre um modelo de regressão linear completo e um seu submodelo - permitindo assim usar a teoria de que já dispomos para esse efeito.

Um exemplo de Análise de Covariância (cont.)

Considere-se o exemplo de três contextos aparentados (e.g. espécies, localidades, anos, etc.), nas quais a relação entre uma variável resposta Y e uma preditora X seja dada, respectivamente, por:

- Contexto 1:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

- Contexto 2:

$$Y = \beta_0^* + \beta_1^* x + \varepsilon$$

- Contexto 3:

$$Y = \beta_0^{**} + \beta_1^{**} x + \varepsilon$$

Vamos considerar que o primeiro contexto é o nível de referência e escrever os parâmetros dos contextos restantes à custa dos primeiros:

$$\beta_0^* = \beta_0 + \alpha_{0:2} \quad ; \quad \beta_1^* = \beta_1 + \alpha_{1:2}$$
$$\beta_0^{**} = \beta_0 + \alpha_{0:3} \quad ; \quad \beta_1^{**} = \beta_1 + \alpha_{1:3}$$

$$\beta_0^* = \beta_{0:2}; \beta_1^* = \beta_{1:2}; \beta_0^{**} = \beta_{0:3}; \beta_1^{**} = \beta_{1:3}$$

As hipóteses de interesse

Com os parâmetros de cada recta escritos desta forma, a hipótese de que as três rectas de regressão sejam iguais é a hipótese

$$\alpha_{0:2} = \alpha_{0:3} = \alpha_{1:2} = \alpha_{1:3} = 0 \ .$$

Vamos arranjar um modelo de regressão múltipla que contenha os

parâmetros $\alpha_{i:j}$ ($i = 0, 1$ e $j = 2, 3$), de forma a poder tirar proveito deste facto.

As variáveis associadas aos acréscimos

Considere que se fazem n observações para ajustar o modelo, sendo

- n_1 correspondentes ao primeiro contexto;
- n_2 correspondentes ao segundo contexto;
- n_3 correspondentes ao terceiro contexto.

Definam-se as variáveis indicatrizes de pertença aos níveis (como na Análise de Variância). Definam-se também vectores com os valores da variável X num dado contexto i ($i > 1$) e zero noutras posições, que serão representados por $x \star I_i$:

$$I_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad x \star I_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ 0 \\ 0 \end{bmatrix}, \quad I_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}, \quad x \star I_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ x_8 \\ x_9 \end{bmatrix}$$

A equação de base no nosso exemplo

Podemos agora escrever a relação de base entre o vector $\vec{\mathbf{Y}}$ das n observações da variável resposta, e o preditor X, da seguinte forma:

$$\vec{\mathbf{Y}} = \beta_0 \cdot \vec{\mathbf{1}}_n + \beta_1 \cdot \vec{\mathbf{x}} + \alpha_{0:2} \cdot \mathbf{I}_2 + \alpha_{0:3} \cdot \mathbf{I}_3 + \alpha_{1:2} \cdot \vec{\mathbf{x}} \star \mathbf{I}_2 + \alpha_{1:3} \cdot \vec{\mathbf{x}} \star \mathbf{I}_3 \ .$$

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

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \\ Y_9 \end{bmatrix} = \begin{bmatrix} 1 & x_1 & 0 & 0 & 0 & 0 \\ 1 & x_2 & 0 & 0 & 0 & 0 \\ 1 & x_3 & 0 & 0 & 0 & 0 \\ 1 & x_4 & 1 & 0 & x_4 & 0 \\ 1 & x_5 & 1 & 0 & x_5 & 0 \\ 1 & x_6 & 1 & 0 & x_6 & 0 \\ 1 & x_7 & 1 & 0 & x_7 & 0 \\ 1 & x_8 & 0 & 1 & 0 & x_8 \\ 1 & x_9 & 0 & 1 & 0 & x_9 \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \alpha_{0:2} \\ \alpha_{0:3} \\ \alpha_{1:2} \\ \alpha_{1:3} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \end{bmatrix}$$

A equação de base no nosso exemplo (cont.)

Isto é,

$$Y_i = \begin{cases} \beta_0 + \beta_1 x_i + \varepsilon_i, & \text{se } i = 1, \ldots, 3 \\ (\beta_0 + \alpha_{0:2}) + (\beta_1 + \alpha_{1:2}) x_i + \varepsilon_i, & \text{se } i = 4, \ldots, 7 \\ (\beta_0 + \alpha_{0:3}) + (\beta_1 + \alpha_{1:3}) x_i + \varepsilon_i, & \text{se } i = 8, \ldots, 9 \, . \end{cases}$$

O modelo do slide 360 ajusta, às observações de cada um dos três contextos, uma recta de regressão distinta.

Caso os parâmetros de acréscimo $\alpha_{i:j}$ sejam *todos* iguais a zero, a recta de regressão é a mesma, para os três contextos.

A relação de base para comparar 3 rectas

Temos assim uma equação do tipo modelo linear com $3 \times 2 = 6$ parâmetros (e variáveis preditoras $\vec{\mathbf{x}}$, $\mathbf{I}_2$, $\mathbf{I}_3$, $\vec{\mathbf{x}} \star \mathbf{I}_2$, $\vec{\mathbf{x}} \star \mathbf{I}_3$), que ajusta rectas de regressão diferentes para as observações de cada um dos 3 contextos.

$$\vec{\mathbf{Y}} = \beta_0 \cdot \vec{\mathbf{1}}_n + \beta_1 \cdot \vec{\mathbf{x}} + \alpha_{0:2} \cdot \mathbf{I}_2 + \alpha_{0:3} \cdot \mathbf{I}_3 + \alpha_{1:2} \cdot \vec{\mathbf{x}} \star \mathbf{I}_2 + \alpha_{1:3} \cdot \vec{\mathbf{x}} \star \mathbf{I}_3 + \vec{\boldsymbol{\varepsilon}}$$

$$\vec{\mathbf{Y}} = \beta_0 \cdot \vec{\mathbf{1}}_n + \beta_1 \cdot \vec{\mathbf{x}} + \vec{\boldsymbol{\varepsilon}}$$

Um teste *F* parcial permite testar a admissibilidade duma recta única para os três contextos considerados.

A relação de base para comparar 3 rectas

Temos assim uma equação do tipo modelo linear com $3 \times 2 = 6$ parâmetros (e variáveis preditoras $\vec{\mathbf{x}}$, $\mathbf{I}_2$, $\mathbf{I}_3$, $\vec{\mathbf{x}} \star \mathbf{I}_2$, $\vec{\mathbf{x}} \star \mathbf{I}_3$), que ajusta rectas de regressão diferentes para as observações de cada um dos 3 contextos. Caso $\alpha_{0:2} = \alpha_{0:3} = \alpha_{1:2} = \alpha_{1:3} = 0$, obtém-se o submodelo correspondente a ajustar uma única recta aos 3 contextos:

$$\begin{aligned} \vec{\mathbf{Y}} &= \beta_0 \cdot \vec{\mathbf{1}}_n + \beta_1 \cdot \vec{\mathbf{x}} + \alpha_{0:2} \cdot \mathbf{I}_2 + \alpha_{0:3} \cdot \mathbf{I}_3 + \alpha_{1:2} \cdot \vec{\mathbf{x}} \star \mathbf{I}_2 + \alpha_{1:3} \cdot \vec{\mathbf{x}} \star \mathbf{I}_3 + \vec{\boldsymbol{\varepsilon}} \\ \vec{\mathbf{Y}} &= \beta_0 \cdot \vec{\mathbf{1}}_n + \beta_1 \cdot \vec{\mathbf{x}} + \vec{\boldsymbol{\varepsilon}} \end{aligned}$$

Um teste *F* parcial permite testar a admissibilidade duma recta única para os três contextos considerados.

O teste para 3 regressões simples diferenciadas

Teste F a 3 rectas diferentes

Teste F de comparação de um modelo com 3 rectas de regressão linear diferentes e o submodelo de recta única

Hipóteses:

$$H_0 : \alpha_{i:j} = 0, (\forall i{=}0,1; j{=}2,3) \quad \text{vs.} \quad H_1 : \exists (i,j) \text{ t.q. } \alpha_{i:j} \neq 0.$$

[RECTA ÚNICA] [RECTAS DIFERENTES]

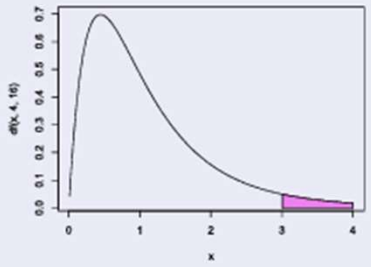
Estatística do Teste:

$$F = \frac{(SQRE_S - SQRE_C)/4}{SQRE_C/(n-6)} \quad \cap \quad F_{(4,n-6)}, \text{ sob } H_0.$$

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se $F_{calc} > f_{\alpha(4,n-6)}$



Outras comparações no exemplo

É possível fazer outras comparações, com base no modelo

$$\vec{\mathbf{Y}} = \beta_0 \cdot \vec{\mathbf{1}}_n + \beta_1 \cdot \vec{\mathbf{x}} + \alpha_{0:2} \cdot \mathbf{I}_2 + \alpha_{0:3} \cdot \mathbf{I}_3 + \alpha_{1:2} \cdot \vec{\mathbf{x}} \star \mathbf{I}_2 + \alpha_{1:3} \cdot \vec{\mathbf{x}} \star \mathbf{I}_3 + \vec{\boldsymbol{\varepsilon}}$$

- A hipótese de três rectas paralelas (i.e., com o mesmo declive), mas podendo ter diferentes ordenadas na origem, é a hipótese $\alpha_{1:2} = \alpha_{1:3} = 0$.
- A hipótese de três rectas com igual ordenada na origem, mas declives diferentes, é a hipótese $\alpha_{0:2} = \alpha_{0:3} = 0$.
- A hipótese de a primeira e segunda recta terem o mesmo declive, é a hipótese $\alpha_{1:2} = 0$.
- A hipótese de a segunda e terceira recta terem o mesmo declive, é a hipótese $\alpha_{1:2} = \alpha_{1:3}$, ou seja, $\alpha_{1:2} - \alpha_{1:3} = 0$.

Estas hipóteses (ou outras análogas) podem ser testadas através de testes já vistos no estudo geral do modelo linear.

A comparação de s rectas de regressão

Generalizando, a comparação de s modelos de regressão linear simples, cada um com n_i ($i = 1, \ldots, s$) observações ($n_1 + \cdots + n_s = n$):

$$Y_i = \begin{cases} \beta_0 + \beta_1 x_i + \varepsilon_i, & i=1,\ldots,n_1 \\ (\beta_0 + \alpha_{0:2}) + (\beta_1 + \alpha_{1:2}) x_i + \varepsilon_i, & i=n_1+1,\ldots,n_1+n_2 \\ \cdots & \\ (\beta_0 + \alpha_{0:s}) + (\beta_1 + \alpha_{1:s}) x_i + \varepsilon_i, & i=n_1+\ldots+n_{s-1}+1,\ldots,n_1+\ldots+n_{s-1}+n_s \end{cases},$$

usando a notação $\vec{\boldsymbol{\beta}}^t = (\beta_0, \beta_1, \alpha_{0:2}, \ldots, \alpha_{0:s}, \alpha_{1:2}, \ldots, \alpha_{1:s})$.

Admitir uma recta única nas s situações é admitir a hipótese

$$H_0 : \alpha_{0:2} = \cdots = \alpha_{0:s} = \alpha_{1:2} = \cdots = \alpha_{1:s} = 0.$$

Modelo com *s* rectas diferenciadas – notação vectorial

Um modelo que prevê a possibilidade de existirem *s* rectas de regressão linear simples diferentes em cada um de *s* contextos, tem a seguinte equação de base:

$$\vec{\mathbf{Y}} = \beta_0 \cdot \vec{\mathbf{1}}_n + \beta_1 \cdot \vec{\mathbf{x}} + \alpha_{0:2} \cdot \mathbf{I}_2 + \alpha_{0:3} \cdot \mathbf{I}_3 + \cdots + \alpha_{0:s} \cdot \mathbf{I}_s + \\ + \alpha_{1:2} \cdot \vec{\mathbf{x}} \star \mathbf{I}_2 + \alpha_{1:3} \cdot \vec{\mathbf{x}} \star \mathbf{I}_3 + \cdots + \alpha_{1:s} \cdot \vec{\mathbf{x}} \star \mathbf{I}_s + \vec{\boldsymbol{\varepsilon}} \ .$$

Este modelo tem $2s$ parâmetros.

Admitir uma recta única nas *s* situações é admitir que este modelo equivale ao seu submodelo:

$$\vec{\mathbf{Y}} = \beta_0 \cdot \vec{\mathbf{1}}_n + \beta_1 \cdot \vec{\mathbf{x}} + \vec{\boldsymbol{\varepsilon}} \ .$$

O submodelo tem 2 parâmetros.

Modelo com *s* rectas – notação matricial

O modelo diferenciado resulta de admitir, em notação matricial,

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times 2s}\vec{\boldsymbol{\beta}}_{2s\times 1} + \vec{\boldsymbol{\varepsilon}}_{n\times 1}$$

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ \vdots \\ Y_{n-1} \\ Y_n \end{bmatrix}, \vec{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \alpha_{0:2} \\ \vdots \\ \alpha_{0:s} \\ \alpha_{1:2} \\ \vdots \\ \alpha_{1:s} \end{bmatrix}, \vec{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \vdots \\ \vdots \\ \varepsilon_{n-1} \\ \varepsilon_n \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} \vec{\mathbf{1}}_{\mathrm{n}} & \vec{\mathbf{x}} & \mathbf{I}_2 & \cdots & \mathbf{I}_s & \mathbf{I}_2 * \vec{\mathbf{x}} & \cdots & \mathbf{I}_s * \vec{\mathbf{x}} \\ | & | & | & \cdots & | & | & \cdots & | \end{bmatrix}$$

Recta única ou *s* rectas?

A comparação dos modelos faz-se pelo teste *F* parcial a submodelos:

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times 2} \times \vec{\boldsymbol{\beta}}_{2\times 1} + \vec{\boldsymbol{\varepsilon}}_{n\times 1} \text{ (submodelo – recta única)}$$
$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times 2s} \times \vec{\boldsymbol{\beta}}_{2s\times 1} + \vec{\boldsymbol{\varepsilon}}_{n\times 1} \text{ (modelo – s rectas),}$$

O submodelo é a recta (única) de regressão com base na totalidade das *n* observações, sendo

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X}_{n\times 2} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \vec{\boldsymbol{\beta}}_{2\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \vec{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

O teste para s regressões simples diferenciadas

Teste F: s rectas diferentes ou uma recta única?

Teste F de comparação de um modelo com s rectas de regressão linear diferentes (índice D) e o submodelo de recta única (índice U)

Hipóteses:

$$H_0 : \alpha_{i:j} = 0 \text{, } (i{=}0,1; j{=}2,3,\ldots,s) \quad \text{vs.} \quad H_1 : \exists (i,j) \text{ t.q. } \alpha_{i:j} \neq 0.$$

[RECTA ÚNICA] [RECTAS DIFERENTES]

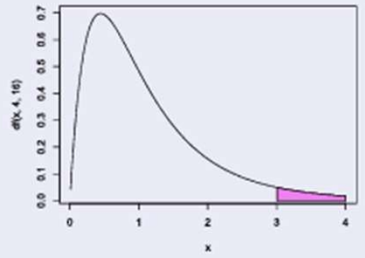
Estatística do Teste:

$$F = \frac{(SQRE_U - SQRE_D)/(2s-2)}{SQRE_D/(n-2s)} \quad \cap \quad F_{(2s-2,n-2s)}, \text{ sob } H_0.$$

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se $F_{calc} > f_{\alpha(2s-2,n-2s)}$



s rectas paralelas?

Tal como no caso inicial, com apenas 3 rectas, também no caso geral se pode testar a hipótese de as *s* rectas de regressão linear simples serem paralelas, isto é, terem o mesmo declive (podendo, no entanto, ter diferentes ordenadas na origem).
O modelo completo tem $2s$ parâmetros.

$$\vec{\mathbf{Y}} = \beta_0 \cdot \vec{\mathbf{1}}_n + \beta_1 \cdot \vec{\mathbf{x}} + \alpha_{0:2} \cdot \mathbf{I}_2 + \alpha_{0:3} \cdot \mathbf{I}_3 + \cdots + \alpha_{0:s} \cdot \mathbf{I}_s + \\ + \alpha_{1:2} \cdot \vec{\mathbf{x}} \star \mathbf{I}_2 + \alpha_{1:3} \cdot \vec{\mathbf{x}} \star \mathbf{I}_3 + \cdots + \alpha_{1:s} \cdot \vec{\mathbf{x}} \star \mathbf{I}_s + \vec{\boldsymbol{\varepsilon}} \ .$$

Admitir *s* rectas paralelas nas *s* situações é admitir que

$$\alpha_{1:2} = \alpha_{1:3} = \ldots = \alpha_{1:s} = 0$$

logo, que o modelo equivale ao submodelo (com $s+1$ parâmetros):

$$\vec{\mathbf{Y}} = \beta_0 \cdot \vec{\mathbf{1}}_n + \beta_1 \cdot \vec{\mathbf{x}} + \alpha_{0:2} \cdot \mathbf{I}_2 + \alpha_{0:3} \cdot \mathbf{I}_3 + \cdots + \alpha_{0:s} \cdot \mathbf{I}_s + \vec{\boldsymbol{\varepsilon}} \ .$$

O teste para s rectas de regressão paralelas

Teste F: s rectas paralelas ou s rectas diferentes?

Teste F de comparação do modelo com s rectas de regressão linear diferentes (índice D) e o submodelo de s rectas paralelas (índice P)

Hipóteses:

$$H_0 : \alpha_{i:j} = 0, (\forall i{=}1; j{=}2,3,\ldots,s) \quad \text{vs.} \quad H_1 : \exists j \text{ t.q. } \alpha_{1:j} \neq 0.$$

[RECTAS PARALELAS] [NÃO PARALELAS]

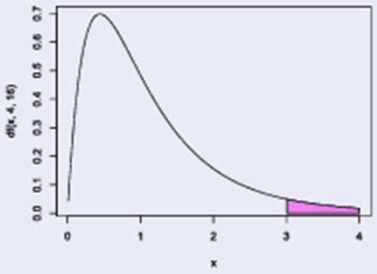
Estatística do Teste:

$$F = \frac{(SQRE_P - SQRE_D)/(s-1)}{SQRE_D/(n-2s)} \quad \cap \quad F_{(s-1,n-2s)}, \text{ sob } H_0.$$

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se $F_{calc} > f_{\alpha(s-1,n-2s)}$



Outras comparações no exemplo

É possível fazer outras comparações, com base no modelo

$$\begin{aligned}\vec{\mathbf{Y}} &= \beta_0 \cdot \vec{\mathbf{1}}_n + \beta_1 \cdot \vec{\mathbf{x}} + \alpha_{0:2} \cdot \mathbf{I}_2 + \alpha_{0:3} \cdot \mathbf{I}_3 + \cdots + \alpha_{0:s} \cdot \mathbf{I}_s + \\ &\quad + \alpha_{1:2} \cdot \vec{\mathbf{x}} \star \mathbf{I}_2 + \alpha_{1:3} \cdot \vec{\mathbf{x}} \star \mathbf{I}_3 + \cdots + \alpha_{1:s} \cdot \vec{\mathbf{x}} \star \mathbf{I}_s + \vec{\boldsymbol{\varepsilon}}\end{aligned}$$

- A hipótese de as s rectas terem igual ordenada na origem, mas declives diferentes, é a hipótese $\alpha_{0:2} = \alpha_{0:3} = \cdots = \alpha_{0:s} = 0$.
- 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 $\alpha_{1:2} = \alpha_{1:3}$.

Estas hipóteses (ou outras análogas) podem ser testadas através de testes já vistos no estudo geral do modelo linear.

Os pressupostos

Os testes anteriormente referidos são válidos caso se verifiquem os pressupostos já admitidos nos Modelos Lineares, i.e., que os erros aleatórios da equação do modelo verificam:

- $\varepsilon_i \cap \mathcal{N}(0, \sigma^2), \ \forall i$;
- erros aleatórios independentes.

Trata-se (quase) dos mesmos pressupostos que seria necessário supor para ajustar cada recta, de forma separada, usando apenas as n_i observações correspondentes ao seu contexto.

Mas há um pressuposto adicional em relação ao ajustamento em separado: a homogeneidade das variâncias dos erros aleatórios tem de ser comum aos *s* contextos.

Modelo com *s* rectas ou *s* regressões simples?

Qual a relação entre as rectas ajustadas

- pelo modelo que admite *s* rectas diferenciadas para os vários níveis de um factor (descrito no slide 360); e
- pelos *s* modelos de regressão linear simples em separado (usando apenas as observações de um dado nível do factor)?

As estimativas dos parâmetros das rectas são iguais nas duas abordagens.

Ou seja, as *s* rectas ajustadas através da Análise de Covariância são as mesmas *s* rectas que se obteriam caso fossem feitas *s* regressões separadas, usando apenas as observações de um dado contexto.

O modelo conjunto e *s* regressões individuais (cont.)

Portanto,

- os valores ajustados de *y* em cada recta são iguais nas duas abordagens;
- os resíduos são iguais nas duas abordagens;
- a soma de quadrados dos resíduos na abordagem conjunta é a soma dos *s* *SQRE*s de cada modelo separado.

Ou seja,

$$SQRE_{conjunto} = SQRE_1 + SQRE_2 + \cdots + SQRE_s \, .$$

Modelo com *s* rectas ou *s* regressões simples? (cont.)

- o Quadrado Médio Residual no modelo conjunto é uma média ponderada dos *QMRE*s de cada modelo separado, sendo os pesos na média ponderada dados pelos graus de liberdade de cada *QMRE* separado.

Ou seja,

$$
\begin{aligned}
& SQRE_{conjunto} = SQRE_1 + SQRE_2 + \cdots + SQRE_s \\
\Rightarrow\ & SQRE_{conjunto} = QMRE_1 \cdot (n_1 - 2) + QMRE_2 \cdot (n_2 - 2) + \cdots + QMRE_s \cdot (n_s - 2) \\
\Leftrightarrow\ & QMRE_{conjunto} = \frac{QMRE_1 \cdot (n_1 - 2) + QMRE_2 \cdot (n_2 - 2) + \cdots + QMRE_s \cdot (n_s - 2)}{n - 2s}
\end{aligned}
$$

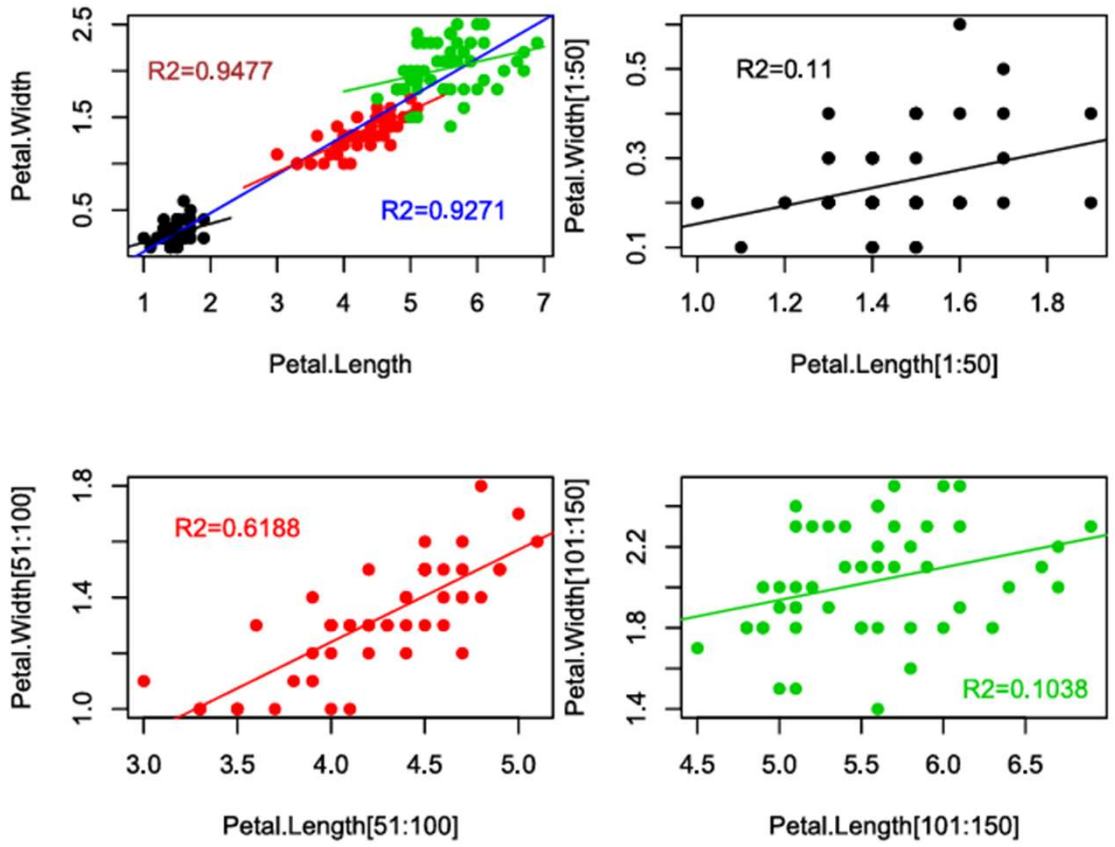
que é uma média ponderada dos *QMRE*, pois a soma das ponderações é $\sum_{i=1}^{s} (n_i - 2) = n - 2s$.

Modelo com s rectas ou s regressões simples? (cont.)

- Os Coeficientes de Determinação, R^2, dos modelos separados e do modelo conjunto são mais difíceis de relacionar. O R^2 do modelo conjunto mede a relação linear da nuvem de pontos obtida com a totalidade dos n pontos. Pode ser maior ou menor do que qualquer dos valores individuais de R^2 só das observações de um dado nível do factor.

Não esquecer que o valor do Coeficiente de Determinação é sempre dado por $R^2 = \frac{SQR}{SQT}$, em que os valores de SQR e SQT (e $SQRE$) se referem sempre ao conjunto de pontos usados no ajustamento.

Um exemplo com as relações entre Largura de Pétala e Comprimento de Pétala única, diferenciada e separada, para as três espécies de lírios (*setosa*, *versicolor* e *virginica*). Atenção aos R^2!



Comparando os *SQT*

A relação entre o *SQT* do modelo conjunto das *s* rectas e os SQT_i de cada um dos *s* modelos individuais, obtidos ajustando apenas os n_i pontos de cada situação, envolve a decomposição de *SQT* que resulta de efectuar a ANOVA a 1 Factor, sendo o factor dado pela distinção das *s* situações analisadas.
Seja *SQF* a Soma dos Quadrados do Factor nessa ANOVA relacionando *Y* e o factor. Tem-se:

$$SQT = \sum_{i=1}^{s} SQT_i + SQF \ .$$

Comparando os *SQR*

Tendo em conta a relação fundamental de qualquer regressão, $SQT = SQR + SQRE$, e tendo ainda em conta a relação entre o *SQRE* do modelo conjunto e os s $SQRE_i$ de cada modelo, visto no slide 380, tem-se a seguinte relação entre o *SQR* do modelo conjunto e as s Somas de Quadrado da Regressão, associadas às s regressões individuais:

$$SQR = \sum_{i=1}^{s} SQR_i + SQF \ .$$

Comparando os Coeficientes de Determinação

As relações dos slides anteriores permitem agora relacionar o valor do Coeficiente de Determinação R^2 do modelo conjunto, com os s Coeficientes de Determinação R_i^2 de cada modelo individual. Tem-se:

$$R^2 = \frac{\sum_{i=1}^{s} SQR_i + SQF}{\sum_{i=1}^{s} SQT_i + SQF} = \frac{\sum_{i=1}^{s} R_i^2 SQT_i + SQF}{\sum_{i=1}^{s} SQT_i + SQF}.$$

Note-se que:

- se $SQF \approx 0$ (i.e., se o Factor não tem efeitos significativos sobre Y), R^2 será aproximadamente uma média ponderada dos R_i^2 (sendo as ponderações dadas pelos SQT_i). Neste caso, R^2 só pode ser próximo de 1 se a generalidade dos R_i^2 for próxima de 1.
- para SQF grande (i.e., efeitos significativos do Factor sobre Y), R^2 será próximo de 1: a separação das médias de Y em cada grupo vai predominar na expressão.

Generalizando para qualquer número de preditores

A ideia de fundo usada para comparar rectas de regressão linear em *s* contextos diferentes pode ser generalizada para estudar qualquer regressão linear múltipla em *s* contextos diferentes.

Para cada preditor, admite-se a possibilidade de haver acréscimos no respectivo coeficiente (em relação ao coeficiente do primeiro contexto), diferentes em cada um dos restantes contextos.

ANÁLISE DE VARIÂNCIA DE EFEITOS ALEATÓRIOS

Efeitos aleatórios em modelos tipo ANOVA

Nos modelos ANOVA estudados até aqui, admitiu-se sempre que as parcelas de efeitos nas equações dos modelos eram constantes. Este tipo de modelos dizem-se de efeitos fixos.

Uma outra grande classe de modelos alternativos designam-se modelos de efeitos aleatórios.

Não sendo, em rigor, modelos lineares do tipo considerado até aqui, têm pontos de contacto importantes, em particular no caso dum modelo a um único factor.

Modelos ANOVA com efeitos aleatórios (cont.)

Se um factor tem um número muito grande, ou mesmo uma infinidade, de possíveis níveis, não sendo possível estudar todos, pode optar-se por estudar apenas uma amostra aleatória de níveis do factor, na tentativa de extrair conclusões para o factor na sua totalidade.

Esta situação surge com frequência quando os níveis de um factor são terrenos, genótipos ou outras entidades para as quais se admite variabilidade, mas em que não é possível estudar a totalidade dos possíveis casos (níveis do factor).

Efeitos de blocos, ou de factores hierarquizados subordinados são, com muita frequência, mais correctamente descritos por efeitos aleatórios.

Um campo importante de aplicação do modelo a um factor com efeitos aleatórios é o de estudos de melhoramento vegetal e animal.

Modelos ANOVA com efeitos aleatórios (cont.)

Nesses casos, os efeitos dos níveis seleccionados aleatoriamente para o estudo são melhor descritos por variáveis aleatórias, e não por constantes.

Por exemplo, a equação base de um modelo a um factor com efeitos aleatórios, com k níveis do factor, será

$$Y_{ij} = \mu + \mathbf{u}_i + \boldsymbol{\varepsilon}_{ij} \ ,$$

sendo $\mathbf{u}_i$ uma variável aleatória que indica o efeito do nível que vier a ser aleatoriamente seleccionado como nível i do factor.

Podem ser considerados modelos com vários factores em que todos, ou apenas alguns, são de efeitos aleatórios. Um modelo com factores de efeitos fixos e outros de efeitos aleatórios diz-se um modelo misto.

Modelos ANOVA com efeitos aleatórios (cont.)

A existência de novas variáveis aleatórias (além dos erros aleatórios) na equação de base de um modelo com efeitos aleatórios exige novos pressupostos para possibilitar o estudo do modelo.

Os pressupostos usuais em modelos com efeitos aleatórios são que os efeitos aleatórios do tipo $\mathbf{u}_i$:

- têm distribuição Normal;
- têm média zero;
- têm variância $\sigma^2_{\mathbf{u}}$;
- são independentes entre si e independentes dos erros aleatórios.

Estas hipóteses correspondem a admitir que a distribuição dos efeitos de nível do factor é $\mathbf{u}_i \cap \mathcal{N}(0, \sigma^2_{\mathbf{u}})$ e que os níveis amostrados são seleccionados de forma independente.

Modelo ANOVA a 1 Factor com efeitos aleatórios

Modelo ANOVA a um factor, de efeitos aleatórios

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

1. $Y_{ij} = \mu + \mathbf{u}_i + \boldsymbol{\varepsilon}_{ij}$, $\forall i = 1,...,k$, $\forall j = 1,...,n_i$.
2. $\mathbf{u}_i \cap \mathcal{N}(0, \sigma^2_{\mathbf{u}}), \forall i$
3. $\boldsymbol{\varepsilon}_{ij} \cap \mathcal{N}(0, \sigma^2_{\varepsilon}), \forall i,j$
4. $\{\{\mathbf{u}_i\}_i, \{\boldsymbol{\varepsilon}_{ij}\}_{i,j}\}$ são $k+n$ v.a.s independentes.

O índice ε na variância dos erros aleatórios apenas serve para distinguir da nova variância, $\sigma^2_{\mathbf{u}}$, dos efeitos do factor.

O modelo tem 2 parâmetros desconhecidos: as variâncias $\sigma^2_{\mathbf{u}}$ e σ^2_{ε}.

A inexistência de efeitos do factor corresponde à hipótese $\sigma^2_{\mathbf{u}} = 0$.

Admitiremos que se tem um delineamento equilibrado: $n_i = n_c$, $\forall i$.

As componentes da variância

No modelo a 1 Factor de efeitos aleatórios, agora referido, tem-se que a variância de qualquer observação Y_{ij} é dada por:

$$V[Y_{ij}] = V[\mu + \mathbf{u}_i + \boldsymbol{\varepsilon}_{ij}] = \sigma^2_{\mathbf{u}} + \sigma^2_{\varepsilon} \ .$$

As duas parcelas desta expressão designam-se as componentes da variância, expressão que é por vezes usada para indicar ANOVAs com efeitos aleatórios.

Note-se que neste modelo,

$$E[Y_{ij}] = \mu \ ,$$

para qualquer observação Y_{ij}.

Teste a efeitos aleatórios do factor

Um teste à existência de efeitos do factor tem as hipóteses:

$$H_0 : \sigma^2_{\mathbf{u}} = 0 \qquad \text{vs.} \qquad H_1 : \sigma^2_{\mathbf{u}} > 0$$

Embora este modelo a um factor não seja um Modelo Linear do mesmo tipo que o modelo de efeitos fixos antes estudado, o teste envolve uma estatística equivalente.

Em geral, com delineamentos mais complexos, testes à existência de efeitos aleatórios envolvem quocientes de Quadrados Médios, com distribuição F sob H_0, mas nem sempre as estatísticas dos testes são iguais aos correspondentes casos de efeitos fixos.

Teste a efeitos com um único factor

No caso concreto do modelo a um factor, o ponto de partida é igualmente a decomposição de $SQT = SQF + SQRE$, com somas de quadrados definidas de forma igual que no caso de efeitos fixos.

Definindo ainda:

$$QMF = \frac{SQF}{k-1} \qquad \text{e} \qquad QMRE = \frac{SQRE}{n-k} \ ,$$

tem-se:

$$\frac{\sigma_\varepsilon^2}{n_c\,\sigma_{\mathbf{u}}^2 + \sigma_\varepsilon^2} \cdot \frac{QMF}{QMRE} \sim F_{(k-1,n-k)} \ .$$

Assim, sob H_0, tem-se à mesma

$$\frac{QMF}{QMRE} \sim F_{(k-1,n-k)} \ .$$

O Teste F aos efeitos aleatórios dum factor

Teste F - efeitos aleatórios dum factor - delin. equilibrado

Hipóteses: $H_0 : \sigma^2_{\mathbf{u}} = 0$ vs. $H_1 : \sigma^2_{\mathbf{u}} > 0$.

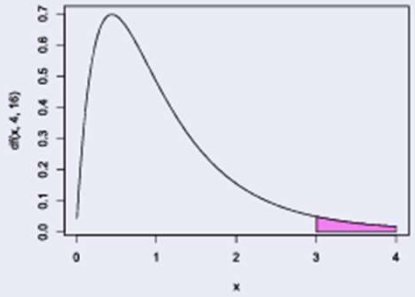
[FACTOR NÃO AFECTA] vs. [FACTOR AFECTA Y]

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



Estimação dos parâmetros com efeitos aleatórios

Caso se tenha rejeitado H_0, interessa estudar a variância $\sigma^2_{\mathbf{u}}$ dos efeitos do factor.

No modelo a um factor de efeitos aleatórios, tem-se:

$$E[QMRE] = \sigma^2_\varepsilon \qquad \text{e} \qquad E[QMF] = \sigma^2_\varepsilon + n_c\,\sigma^2_{\mathbf{u}} \,.$$

Podem definir-se como estimadores dos parâmetros do modelo:

$$\hat{\sigma}^2_\varepsilon = QMRE \quad \text{(como de costume)}$$
$$\hat{\sigma}^2_{\mathbf{u}} = \frac{QMF - QMRE}{n_c}$$

Aviso: O estimador $\hat{\sigma}^2_{\mathbf{u}}$ pode tomar valores negativos, mas é impossível que $\sigma^2_{\mathbf{u}}$ seja negativo. Estimativas negativas surgem em situações em que não se rejeita H_0. Nesses casos toma-se $\sigma^2_{\mathbf{u}} = 0$.

Covariâncias entre observações de Y

Uma diferença deste modelo de efeitos aleatórios em relação ao correspondente modelo de efeitos fixos é que diferentes observações de Y num mesmo nível do factor não são independentes:

$$\begin{aligned} cov(Y_{ij}, Y_{ij'}) &= cov(\mu + \mathbf{u}_i + \boldsymbol{\varepsilon}_{ij}, \mu + \mathbf{u}_i + \boldsymbol{\varepsilon}_{ij'}) \\ &= \underbrace{cov(\mathbf{u}_i, \mathbf{u}_i)}_{=V[\mathbf{u}_i]=\sigma^2_{\mathbf{u}}} + \underbrace{cov(\mathbf{u}_i, \boldsymbol{\varepsilon}_{ij'})}_{=0} + \underbrace{cov(\boldsymbol{\varepsilon}_{ij}, \mathbf{u}_i)}_{=0} + \underbrace{cov(\boldsymbol{\varepsilon}_{ij}, \boldsymbol{\varepsilon}_{ij'})}_{=0} \\ &= \sigma^2_{\mathbf{u}} \end{aligned}$$

Apenas observações de níveis diferentes ($i \neq i'$) são independentes:

$$cov(Y_{ij}, Y_{i'j'}) = 0$$