Análise de Covariância (ANCOVA)

Elsa Gonçalves

(Adaptado,Cadima,J.(2021).O Modelo Linear. ISA, ULisboa)

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.

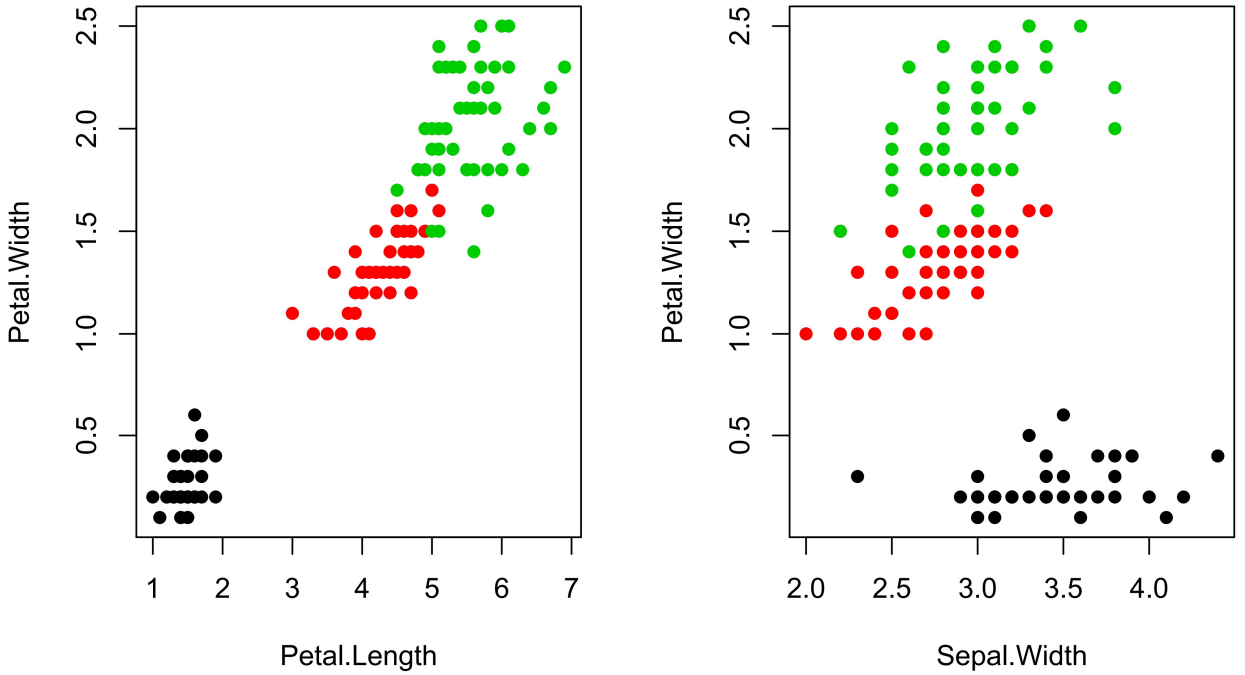
- Numa Regressão Linear, as variáveis preditoras são variáveis igualmente quantitativas (numéricas).
- Numa Análise de Variância, as variáves preditoras são factores (variáveis qualitativas, ou categóricas).
- Numa Análise de Covariância, entre as variáveis preditoras encontramos quer variáveis numéricas,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

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

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,

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

O modelo do slide 366 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.

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

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

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:

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

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]

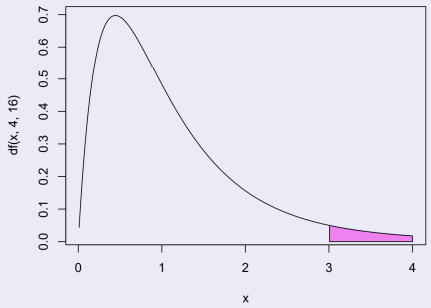
Estatística do Teste:

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

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:

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

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:

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

[RECTA ÚNICA] [RECTAS DIFERENTES]

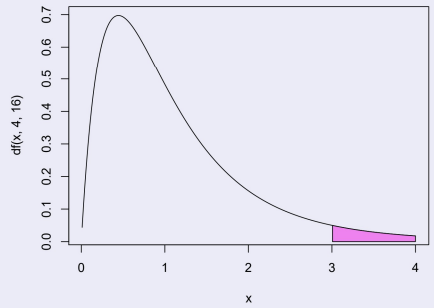
Estatística do Teste:

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

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.

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

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]

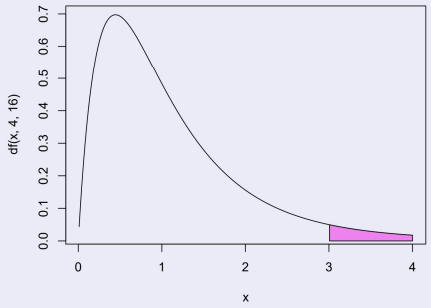
Estatística do Teste:

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

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.

Cruzando factores com variáveis numéricas no R

No R, um modelo de regressão de y sobre x, admitindo rectas diferentes para cada nível do factor f, é indicado pela fórmula

$$y \sim x * f$$

No exemplo dos $n = 150$ lírios,

```
> modespecie.lm <- lm(Petal.Length ~ Sepal.Length * Species)
> summary(modespecie.lm)
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                    0.8031     0.5310   1.512    0.133
Sepal.Length                   0.1316     0.1058   1.244    0.216
Speciesversicolor             -0.6179     0.6837  -0.904    0.368
Speciesvirginica              -0.1926     0.6578  -0.293    0.770
Sepal.Length:Speciesversicolor  0.5548     0.1281   4.330 2.78e-05 ***
Sepal.Length:Speciesvirginica   0.6184     0.1210   5.111 1.00e-06 ***
---
Residual standard error: 0.2611 on 144 degrees of freedom
Multiple R-squared: 0.9789,     Adjusted R-squared: 0.9781
F-statistic:  1333 on 5 and 144 DF,  p-value: < 2.2e-16
```

Um exemplo no R. Recta única?

De novo o exemplo dos 150 lírios. Pretende-se modelar Comprimento das Pétalas, à custa de Comprimento das Sépalas.
Recta única ou rectas diferenciadas por espécie?

```
> modunico.lm <- lm(Petal.Length ~ Sepal.Length)
> modespecie.lm <- lm(Petal.Length ~ Sepal.Length*Species)
> anova(modunico.lm, modespecie.lm)
Analysis of Variance Table
Model 1: Petal.Length ~ Sepal.Length
Model 2: Petal.Length ~ Sepal.Length * Species
  Res.Df     RSS  Df Sum of Sq     F    Pr(>F)
1    148 111.459
2    144   9.818   4   101.641 372.7 < 2.2e-16 ***
```

Rejeita-se a hipótese de uma recta única, em favor de rectas diferentes.

Um exemplo no R. Rectas paralelas?

No R, um modelo de regressão de y sobre x, que admite rectas paralelas, mas com diferentes ordenadas na origem para cada nível de um factor f, pode ser indicado na forma

$$y \sim x + f$$

```
> modparalelas.lm <- lm(Petal.Length ~ Sepal.Length + Species)
> summary(modparalelas.lm)
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.70234    0.23013  -7.397 1.01e-11 ***
Sepal.Length       0.63211    0.04527  13.962  < 2e-16 ***
Speciesversicolor  2.21014    0.07047  31.362  < 2e-16 ***
Speciesvirginica   3.09000    0.09123  33.870  < 2e-16 ***
---
Residual standard error: 0.2826 on 146 degrees of freedom
Multiple R-squared: 0.9749,     Adjusted R-squared: 0.9744
F-statistic:  1890 on 3 and 146 DF,  p-value: < 2.2e-16
```

Um exemplo no R. Rectas paralelas? (cont.)

Mas é admissível que as três rectas sejam paralelas?
Vamos fazer um teste aos modelos encaixados que admitem rectas paralelas e rectas diferentes.

```
> modparalelas.lm <- lm(Petal.Length ~ Sepal.Length + Species)
> modespecie.lm <- lm(Petal.Length ~ Sepal.Length * Species)
> anova(modparalelas.lm,modespecie.lm)

Analysis of Variance Table

Model 1: Petal.Length ~ Sepal.Length + Species
Model 2: Petal.Length ~ Sepal.Length * Species
  Res.Df     RSS  Df Sum of Sq      F    Pr(>F)
1    146 11.6571
2    144  9.8179   2    1.8393 13.489 4.272e-06 ***
```

Rejeita-se a hipótese de rectas paralelas.

Os pressupostos

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

- $\varepsilon_i \cap \mathcal{N}(0, \sigma^2), \ \forall i$;
- 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 366); 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 $\sum_{i=1}^{s} (n_i - 2) = n - 2s$.

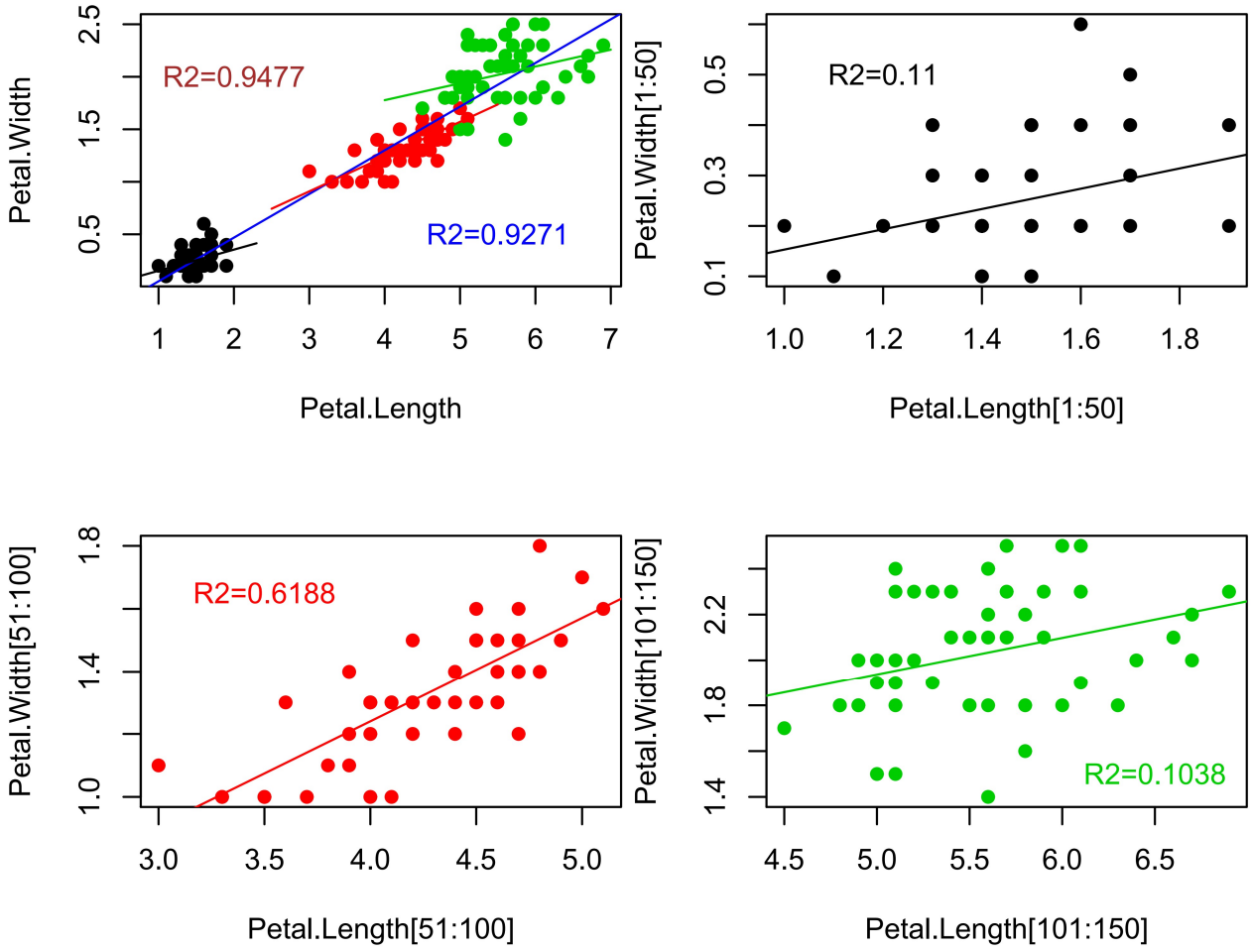
Modelo com s rectas ou s regressões simples? (cont.)

- 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

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,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.

Ainda o exemplo do slide 388

Os valores de cada Soma de Quadrados, bem como do Coeficiente de Determinação, para cada um dos modelos referidos no exemplo do Acetato 388, são:

	SQT	SQR	SQRE	QMRE	R2
setosa	0.54420	0.05985029	0.4843497	0.01009062	0.1099785
versicolor	1.91620	1.18583401	0.7303660	0.01521596	0.6188467
virginica	3.69620	0.38349444	3.3127056	0.06901470	0.1037537
conjunto	86.56993	82.04251207	4.5274213	0.03144043	0.9477022

Resultados ANOVA a 1 Factor: Petal.Width ~ Species

SQF=80.41333 SQRE=6.15660

É o valor elevado de *SQF* que gera um valor elevado do R^2 conjunto.

NOTA: o modelo único não surge nesta comparação.

Generalizando para qualquer número de preditores

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

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

Análise de Variância a 1 factor de efeitos aleatórios

Elsa Gonçalves

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.

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. $H_1 : \sigma^2_{\mathbf{u}} > 0$.

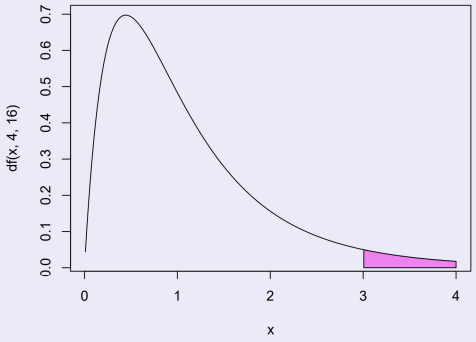
[FACTOR NÃO AFECTA] vs. [FACTOR AFECTA Y]

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 σ_α^2 dos efeitos do factor.

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

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

Podem definir-se como estimadores dos parâmetros do modelo:

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

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

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:

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

Diversidade genética e melhoramento

Um campo importante de aplicação do modelo a um factor com efeitos aleatórios é o de estudos de melhoramento vegetal e animal. O factor em questão são aí diferentes genótipos, que representam uma amostra aleatória duma infinidade de possíveis genótipos.

A variância dos efeitos do factor, $V[\mathbf{u}_i] = \sigma^2_{\mathbf{u}}$ é conhecida, neste contexto, por variabilidade genotípica.

A variância das médias de nível amostrais é, neste contexto e admitindo um delineamento equilibrado, dada por $V[\overline{Y}_{i.}] = \sigma^2_{\mathbf{u}} + \frac{\sigma^2_\varepsilon}{n_c}$. Esta variância é conhecida por variabilidade fenotípica ao nível da média dos genótipos.

A heritabilidade

A razão das variabilidades genotípica e fenotípica é conhecida, em alguns estudos de melhoramento vegetal (com reprodução assexuada), por heritabilidade (em sentido lato):

$$H^2 = \frac{\sigma^2_{\mathbf{u}}}{\sigma^2_{\mathbf{u}} + \frac{\sigma^2_\varepsilon}{n_c}}$$

A heritabilidade mede a proporção da variabilidade observada (fenotípica) de origem genética. Quanto maior este quociente, mais interesse haverá em seleccionar genótipos para fazer melhoramento.

A heritabilidade pode ser estimada usando os estimadores de $\sigma^2_{\mathbf{u}}$ e σ^2_ε referidos no slide 404. Nesse caso, verifica-se que a heritabilidade estimada é função da estatística do teste F aos efeitos aleatórios:

$$\hat{H}^2 = 1 - \frac{1}{F_{calc}}$$

Um exemplo

Num estudo inicial efectuado pelo ISA para melhoramento da casta Tinta Miúda (uma casta frequente na região Oeste), foram usados $k = 100$ genótipos. Foi registado o rendimento (em kg/planta) médio nos anos de 1993 a 2001.

A tabela-resumo da ANOVA a um factor correspondente é:

```
> summary(aov(rend ~ genotipo , data=TintaMiuda))
             Df Sum Sq Mean Sq F value Pr(>F)
genotipo     99 222.44  2.2468    20.4 <2e-16 ***
Residuals   300  33.03  0.1101
```

Assim, a hipótese de inexistência de variabilidade genética ($H_0 : \sigma^2_{\mathbf{u}} = 0$) é claramente rejeitada, existindo um índice de heritabilidade de $H^2 = 1 - 1/20.4 = 0.95098$. Ou seja, mais de 95% da variabilidade observada é de origem genética. O melhoramento é promissor. Os genótipos com melhor desempenho neste estudo foram submetidos a um estudo com mais repetições e em vários locais, para a escolha final de genótipos merecedores de utilização intensiva.