CE075 - Análise de Dados Longitudinais

Silva, J.L.P.

26 de agosto, 2019

Perspectiva histórica

Múltiplas Amostras: ANOVA *Split-Plot*

O caso de múltiplas amostras, também chamado de desenho *split-plot*, é comum em ensaios clínicos aleatorizados, nos quais os indivíduos são aleatorizados a diferentes grupos de tratamento e seguidos ao longo do tempo.

Com $h = 1, \ldots, s$ grupos, $i = 1, \ldots, N_h$ indivíduos no grupo h (com $N = \sum_{h=1}^{s} N_h$), e $j = 1, \ldots, n$ ocasiões, o modelo é:

$$Y_{hij} = \mu + \gamma_h + \tau_j + (\gamma\tau)_{hj} + \alpha_{i(h)} + \varepsilon_{hij},$$

em que:

Múltiplas Amostras: ANOVA *Split-Plot*

- μ é a média geral;
- γ_h é o efeito do grupo h ($\sum_h \gamma_h = 0$)
- τ_j é o efeito do tempo j ($\sum_j \tau_j = 0$);
- $(\gamma\tau)_{hj}$ é a interação do tempo j e grupo h ($\sum_h \sum_j (\gamma\tau)_{hj} = 0$);
- $\alpha_{i(h)}$ é o componente do indivíduo i aninhado no grupo h;
- ε_{hij} é o termo de erro para o indivíduo i no grupo h no tempo j.

Múltiplas Amostras: ANOVA *Split-Plot*

As suposições distribucionais são

$$\alpha_{i(h)} \sim N(0, \sigma_\alpha^2) \quad \text{e} \quad \varepsilon_{hij} \sim N(0, \sigma_\varepsilon^2),$$

que implica na mesma estrutura de simetria composta anterior.

O modelo é misto porque os indivíduos são considerados efeitos aleatórios e grupo e tempo são considerados efeitos fixos.

Os dados são assumidos balanceados em termos de n (tempos), mas não necessariamente em termos de N_h (tamanhos de grupo).

Múltiplas Amostras: ANOVA *Split-Plot*

Tabela 1: Representação dos dados

		Tempo			
Grupo	Indivíduo	1	2	...	n
1	1	y_{111}	y_{112}	$\cdots$	y_{11n}
1	2	y_{121}	y_{122}	$\cdots$	y_{12n}
1	.	.	.	...	.
1	N_1	y_{1N_11}	y_{1N_12}	$\cdots$	y_{1N_1n}
.	.	.	.	...	.
.	.	.	.	...	.
s	1	y_{s11}	y_{s12}	$\cdots$	y_{s1n}
s	s	y_{s21}	y_{s22}	$\cdots$	y_{s2n}
s	.	.	.	...	.
s	N_s	y_{sN_s1}	y_{sN_s2}	$\cdots$	y_{sN_sn}

Múltiplas Amostras: ANOVA *Split-Plot*

Se a interação grupo versus tempo for rejeitada, concluímos:

- as diferenças entre os grupos não são as mesmas ao longo do tempo;
- as curvas entre os grupos não são paralelas; e
- os efeitos de tempo e grupo são confundidos com a interação e não podem ser testados separadamente.

Assim, não há efeito geral de tempo, pois este varia com o grupo.

Se a hipótese de interação não for rejeitada, testes de efeitos principais de tempo e grupo podem ser testados separadamente e independentemente:

$$
\begin{aligned}
H_T :& \tau_1 = \tau_2 = \ldots = \tau_n = 0. \\
H_G :& \gamma_1 = \gamma_2 = \ldots = \gamma_s = 0.
\end{aligned}
$$

Múltiplas Amostras: ANOVA *Split-Plot*

Assim como no caso de uma amostra, podemos testar a significância dos efeitos aleatórios de indivíduos.

É comum assumir $\sigma_\alpha^2 > 0$, que nos leva à estimação do coeficiente de correlação intra classe como:

$$ICC = \frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_\alpha^2 + \hat{\sigma}_\varepsilon^2}.$$

A suposição de simetria composta (variâncias e covariâncias iguais ao longo do tempo) é bastante restritiva e frequentemente não realística (especialmente quando n cresce).

Múltiplas Amostras: ANOVA *Split-Plot*

Simetria composta é um caso particular de uma condição chamada *esfericidade*, sob a qual os testes F da ANOVA para os termos relacionados com o tempo são válidos.

A forma mais geral de definir esfericidade é dizer que todas as variâncias das diferença duas a duas entre as variáveis são iguais:

$$\begin{aligned} Var(Y_{ij} - Y_{ij'}) =& Var(Y_{ij}) + Var(Y_{ij'}) - 2Cov(Y_{ij}, Y_{ij'}) \\ =& \text{ constante } \ \forall j \text{ e } j'. \end{aligned}$$

Simetria composta satisfaz esta condição pois todas as variâncias são iguais, assim como as covariâncias.

Múltiplas Amostras: ANOVA *Split-Plot*

Se a esfericidade não se mantém, os testes F são, em geral, muito liberais.

Se a suposição de esfericidade for rejeitada, pode-se usar ANOVA multivariada de medidas repetidas (MANOVA), a qual permite uma forma geral para $Var(\boldsymbol{Y}_i)$.

Outras alternativas clássicas incluem correções nos graus de liberdade da estatística F, como as propostas por Greenhouse-Geisser e Huynh-Feldt.

Ilustração: Bock (1975)

Como ilustração, considere dados de aquisição de vocabulário medidos em uma coorte de 64 estudantes avaliados em um laboratório da Universidade de Chicago.

Os dados longitudinais são oriundos de um teste de leitura aplicados a alunos do oitavo ao décimo primeiro ano (série).

Como a faixa de idade avaliada marca o período no qual o crescimento físico começa a desacelerar, o pesquisador tem como hipótese que também ocorra uma desaceleração da aquisição de novo vocabulário.

Ilustração: Bock (1975)

```r
library(tidyverse)
dados <- read.table("BockData.txt",sep="",h=TRUE)
head(dados)
```

```
  SUBJECT VOCAB1 VOCAB2 VOCAB3 VOCAB4
1       1   1.75   2.60   3.76   3.68
2       2   0.90   2.47   2.44   3.43
3       3   0.80   0.93   0.40   2.27
4       4   2.42   4.15   4.56   4.21
5       5  -1.31  -1.31  -0.66  -2.22
6       6  -1.56   1.67   0.18   2.33
```

Ilustração: Bock (1975)

```r
dadosl <-  dados %>% gather("VOCAB1", "VOCAB2", "VOCAB3",
                            "VOCAB4",key="grade",value="score")
head(dadosl)
```

```
  SUBJECT  grade score
1       1 VOCAB1  1.75
2       2 VOCAB1  0.90
3       3 VOCAB1  0.80
4       4 VOCAB1  2.42
5       5 VOCAB1 -1.31
6       6 VOCAB1 -1.56
```

Ilustração: Bock (1975)

```r
dadosl %>% group_by(grade) %>% summarise(n=n(),
                                  Mean=mean(score), SD=sd(score))
```

```
# A tibble: 4 x 4
  grade      n  Mean    SD
  <chr>  <int> <dbl> <dbl>
1 VOCAB1    64  1.14  1.89
2 VOCAB2    64  2.54  2.08
3 VOCAB3    64  2.99  2.17
4 VOCAB4    64  3.47  1.93
```

```r
#
round(cor(dados[,-1]),3)
```

```
       VOCAB1 VOCAB2 VOCAB3 VOCAB4
VOCAB1  1.000  0.810  0.867  0.785
VOCAB2  0.810  1.000  0.785  0.757
VOCAB3  0.867  0.785  1.000  0.811
VOCAB4  0.785  0.757  0.811  1.000
```

Ilustração: Bock (1975)

```
ggplot(dadosl,aes(x=grade,y=score)) + geom_point() +
  geom_line(aes(group=SUBJECT)) + theme_bw()
```

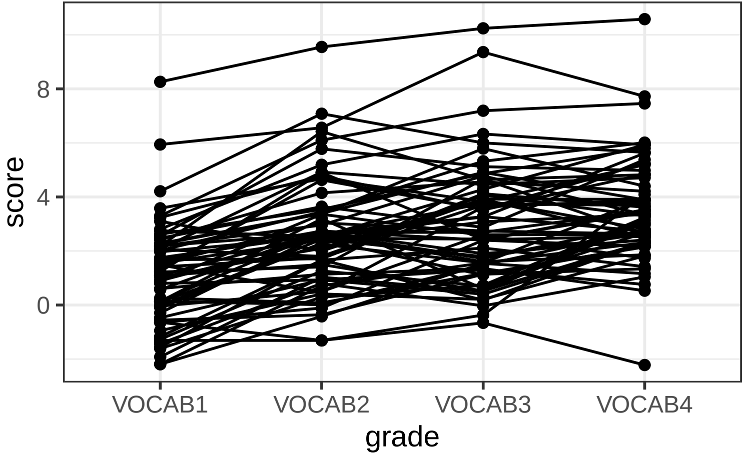

Ilustração: Bock (1975)

```
dadosl %>% group_by(grade) %>% summarise(n=n(), Mean=mean(score),
 SD=sd(score), SE=SD/sqrt(n)) %>% ggplot(aes(x=grade, y=Mean)) +
  geom_errorbar(aes(ymin=Mean-SE, ymax=Mean+SE), width=.1) +
  geom_line(aes(group=1)) + geom_point() + ylim(c(0,4)) + theme_bw()
```

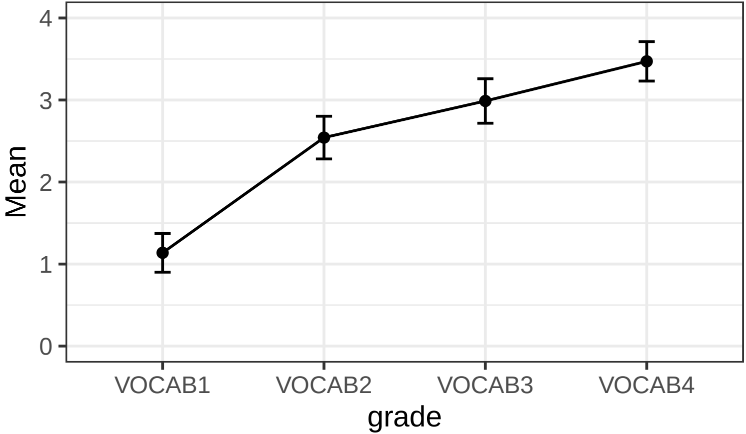

Ilustração: Bock (1975)

```r
options(contrasts=c("contr.sum","contr.poly"))
modelAOV <- aov(score~grade+Error(factor(SUBJECT)),
                data = dadosl)
summary(modelAOV)
```

```

Error: factor(SUBJECT)
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 63  873.6   13.87

Error: Within
           Df Sum Sq Mean Sq F value Pr(>F)
grade       3  194.3   64.78   79.02 <2e-16 ***
Residuals 189  154.9    0.82
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ilustração: Bock (1975)

- A Tabela da ANOVA revela que devemos rejeitar a hipótese nula para o efeito de ano.

- Isto está na direção do que vimos pelos gráficos, que mostraram que o vocabulário aumenta com a idade.

- Antes de procedermos uma análise mais aprofundada da tendência temporal, vamos utilizar a função `lmer` para calcularmos o coeficiente de correlação intraclasse.

Ilustração: Bock (1975)

```r
library(lme4)
fit <- lmer(score~grade + (1|SUBJECT), data=dadosl)
anova(fit)
```

```
Analysis of Variance Table
      Df Sum Sq Mean Sq F value
grade  3 194.34  64.779  79.019
```

```r
VarCorr(fit)
```

```
 Groups   Name        Std.Dev.
 SUBJECT  (Intercept) 1.80603
 Residual             0.90543
```

Ilustração: Bock (1975)

```
re_dat = as.data.frame(VarCorr(fit)); re_dat
```

```
       grp        var1 var2      vcov     sdcor
1  SUBJECT (Intercept) <NA> 3.2617301 1.8060260
2 Residual        <NA> <NA> 0.8197975 0.9054267
```

```
(sub_vcov = re_dat[1,'vcov'])
```

```
[1] 3.26173
```

```
(resid_vcov = re_dat[2,'vcov'])
```

```
[1] 0.8197975
```

```
(ICC=sub_vcov/(resid_vcov+sub_vcov))
```

```
[1] 0.7991444
```

Há um grande efeito de indivíduos na variabilidade: 80% da variação no vocabulário não explicada pela série do aluno (tempo) é atribuível aos indivíduos.

Ilustração: Bock (1975)

- Como vimos, rejeitamos a hipótese nula de que não existe efeito de tempo.

- Uma análise mais aprofundada envolve a construção de contrastes para testar efeito linear, quadrático ou cúbico.

- Vamos proceder com a utilização de contrastes polinomiais ortogonais.

Ilustração: Bock (1975)

```
dadosl$grade <- as.factor(dadosl$grade)
contrasts(dadosl$grade) <- contr.poly(4)
modelAOV <- aov(score~grade+Error(factor(SUBJECT)), data = dadosl)
summary(modelAOV,split=list(grade=list("Linear"=1,"Quadratic"=2,
                                       "Cubic"=3)))
```

```
Error: factor(SUBJECT)
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 63  873.6   13.87

Error: Within
                    Df Sum Sq Mean Sq F value   Pr(>F)
grade                3 194.34   64.78  79.019  < 2e-16 ***
  grade: Linear      1 177.59  177.59 216.630  < 2e-16 ***
  grade: Quadratic   1  13.58   13.58  16.564 6.91e-05 ***
  grade: Cubic       1   3.17    3.17   3.862   0.0509 .
Residuals          189 154.94    0.82
---
```

Ilustração: Bock (1975)

- O termo linear é altamente significativo, assim como o termo quadrático.
- O termo cúbico é apenas marginalmente significativo sugerindo que a desaceleração reverte em certa medida com o aumento da idade.
- Podemos concluir que existe uma tendência de desaceleração positiva com a idade, sustentando a noção de que a aquisição de vocabulário diminui à medida que os estudantes alcançam a maturidade.

Limitações - ANOVA

1. Não se aplica em situações desbalanceadas;
2. Usualmente a correlação tende a diminuir à medida que aumentamos a distância temporal;
3. Difícil (impossível?) ser utilizada na presença de covariáveis contínuas.
4. Resposta com distribuição Normal.

Razões Históricas - Planejamento de Experimentos

1. A matriz de simetria composta tem uma justificativa em termos da aleatorização em Planejamento de Experimentos.

2. Usualmente, não tem a dimensão temporal e, simplesmente, medidas repetidas.

3. Facilidade computacional em termos históricos. Basta uma calculadora para construir a ANOVA.

MANOVA para Medidas Repetidas

Na MANOVA as n medidas repetidas são tratadas como um vetor de respostas $\boldsymbol{Y}_i$ de dimensão $n \times 1$.

Devido à natureza multivariada da análise, os indivíduos com qualquer y_{ij} ausente são omitidos da análise.

Para o caso de uma amostra, o modelo é dado por

$$\boldsymbol{Y}_i = \boldsymbol{\mu} + \boldsymbol{\varepsilon},$$

em que $\boldsymbol{\mu}$ é o vetor de médias para os tempos, e $\boldsymbol{\varepsilon}$ é o vetor de erros, com distribuição $N(\mathbf{0}, \boldsymbol{\Sigma})$ na população.

Esta especificação permite que a matriz de variância covariância seja completamente geral.

MANOVA para Medidas Repetidas

Para o caso de múltiplos grupos, seja $h = 1, \ldots, s$ grupos, $i = 1, \ldots, N_h$ indivíduos no grupo h, $j = 1, \ldots, n$ tempos, e $N = \sum N_h$ o número total de indivíduos.

O número de indivíduos pode variar por grupo, mas cada indivíduo é medido em n ocasiões.

O modelo é escrito como:

$$\boldsymbol{Y}_{hi} = \boldsymbol{\mu} + \gamma_h + \varepsilon_{hi},$$

em que:

MANOVA para Medidas Repetidas

- $\boldsymbol{\mu}$ é o vetor $n \times 1$ de médias para os tempos;
- $\boldsymbol{\gamma}_h$ é o vetor $n \times 1$ de efeitos para a população da qual a grupo h foi amostrado;
- $\boldsymbol{\varepsilon}_{hi}$ é o vetor $n \times 1$ de erros distribuído como $N(\mathbf{0}, \boldsymbol{\Sigma})$ em cada uma das populações.

O modelo assume homogeneidade de variâncias-covariâncias entre os s grupos.

MANOVA para Medidas Repetidas

Testar o efeito geral de tempo e efeitos de interação tempo versus grupo envolve testes multivariados.

Várias estatísticas de teste estão disponíveis para este fim, como lambda de Wilk, traço de Lawley-Hotelling, traço de Pillai e maior autovalor.

MANOVA tem, essencialmente, as mesmas limitações da ANOVA em relação à dados longitudinais e medidas repetidas.

Na sequência, estudaremos modelos mais gerais que não possuem as limitações dos procedimentos tradicionais ANOVA e MANOVA.