Universidade Estadual de Maringá

Aplicação de modelos mistos em dados do FEV

Arthur Cesar de Moura Rocha , Isolde Previdelli[1]
Departamento de Estatística - Universidade Estadual de Maringá

Keywords: Modelos mistos, R, Estrutura de covariância

I. INTRODUÇÃO

Esse trabalho foi feito com intuito de avaliar e aprimorar o conhecimento sobre a aplicação da metodologia de modelos mistos, sendo parte do programa da disciplina de Modelos Mistos, ministrada pela professora Dra. Isolde Previdelli aos discentes do curso de graduação em Estatística pela Universidade Estadual de Maringá.

A. Os dados

Para realização desse projeto, conta-se com um banco de dados que possui a variável resposta FEV, referente a uma qualidade respiratória medida ao longo de 8 horas além da medição imediata pós intervenção das drogas A,C e placebo.

Por se tratar de um problema de dados longitudinais (a ordem do tempo importa), foram feitas algumas análises, além da descritiva, com abordagens diferentes, sendo as primeiras equivocadas (não considerando a dependência) e as últimas partindo da ideia de modelos mistos.

II. METODOLOGIA

A. Análise de variância

A primeira parte desse trabalho consiste na utilização de técnicas previamente conhecidas, a análise de variância baseada em modelos lineares, nos quais são modelados apenas efeitos fixos.

1. Análise de variância one-way

Esse modelo inicial considera apenas uma covariável para a explicação da resposta e ficaria na forma:

$$Y = \mu + \beta_1(tratamento) + \epsilon \tag{1}$$

Em que μ representa uma média geral e β_1 indica o efeito do tratamento, além de que o erro $\epsilon \sim N(0, \sigma^2)$.

Para o caso de mais covariáveis, basta acrescentar os coeficientes $\beta's$ relativos aos efeitos delas e considerar a mesma estrutura de modelo.

2. Análise de variância com blocos

Considerando que existe algum fator que possui algum efeito conhecido na variância da resposta, de forma que impede a casualização do modelo, esse fator é considerado como um bloco.

Para o presente caso, o modelo ficaria na forma:

$$\begin{aligned} Y = \mu + \beta_1(tratamento) + \beta_2(hora) + \\ \beta_3(tratamentohora) + \beta_4(Individuo) + \epsilon \end{aligned} \tag{2}$$

Em que μ representa uma média geral, os $\beta's$ são relativos aos efeitos dos fatores do modelo e o erro $\epsilon \sim N(0, \sigma^2)$.

3. Análise de variância para medidas repetidas

Para esse cenário, considera-se uma estrutura de medida repetida, isto é, há dependência nos dados, sendo eles medidos no mesmo indivíduo em alguns momentos distintos, porém com a ordem de medição não relevante.

O modelo proposto para essa situação é dada pela equação:

$$\begin{aligned} Y = \mu + \beta_1(tratamento) + \beta_2(hora) + \\ \beta_3(tratamentohora) + \epsilon \end{aligned} \tag{3}$$

Sendo que os parâmetros tem interpretação análoga ao modelo anterior, porém o erro ϵ tem uma distribuição normal multivariada com vetor de médias constantes iguais a zero e uma estrutura de covariância Σ, no caso, Compound Symmetry. Esse tipo de estrutura corresponde ao caso de uma variância constante dentro de cada momento e uma correlação comum (constante) entre os momentos.

B. Modelos mistos

Os modelos lineares mistos são uma extensão dos modelos lineares, em que é possível assumir uma parte de efeitos aleatórios no modelo além da parte de efeitos fixos já presentes em modelos lineares comuns, permitindo assim a composição de uma estrutura de covariância (dependência) na análise e dessa forma modelar dados de natureza um pouco mais complexa.

1. Modelo

Nesse caso, só iremos considerar um efeito aleatório no intercepto relativo a cada indivíduo do estudo, sendo que o modelo misto, para a situação presente pode ser expresso como:

$$Y = \mu + b_0(indivíduo) + \beta_1(tratamento) + \beta_2(hora) + \beta_3(tratamento \times hora) + \beta_4(baseline) + \epsilon \quad (4)$$

Tais quais os $\beta's$ são relativos aos efeitos dos fatores fixos do modelo e o b_0 relativo ao efeito aleatório de cada indivíduo. Além disso ϵ representam os erros teórico com distribuição normal de médias iguais a 0 e variâncias constantes σ^2.

2. Estimação

Dado que estamos tratando de uma modelagem um pouco mais complexa do que a linear comum, faz-se necessário o uso de métodos de estimação diferenciados. Em todos os casos desse trabalho utilizou-se a estimação pelo método de máxima verossimilhança restrita, que, ao contrário do método de máxima verossimilhança usual, não produz estimativas viesadas (subestimadas) das variâncias.

3. Comparação entre modelos

Para comparar os diferentes modelos mistos com estruturas de covariância diferentes, optou-se pelo teste da razão entre verossimilhanças, que permite essa comparação desde que os modelos sejam aninhados, isto é, as estruturas de covariância devem ser casos particulares uma das outras.

C. Comparações múltiplas

Em todos os casos utilizou-se o teste para comparações múltiplas (*post hoc*) de Tukey, sendo utilizada a adaptação adequada para cada modelo.

Todas as análises foram feitas utilizando o ambiente estatístico R[1], fixando o nível de significância em 5% para todos os testes.

É importante comentar que todos os resultados são apresentados com seus devidos códigos em R e suas respectivas saídas.

III. RESULTADOS

A. Análise descritiva

O primeiro passo de uma análise é verificar o comportamento dos dados de forma descritiva, portanto essa sessão é dedicada a esse fim.

Verifica-se que parece ser razoavel aceitar que a correlação é constante para a medida de base (BASEFEV1), isto é, como não há tempo para a droga agir espera-se que essa quantidade seja constante. Já para os outros momentos, percebe-se que a correlação decai lentamente conforme o tempo, o que faz sentido do ponto de vista de dados longitudinais.

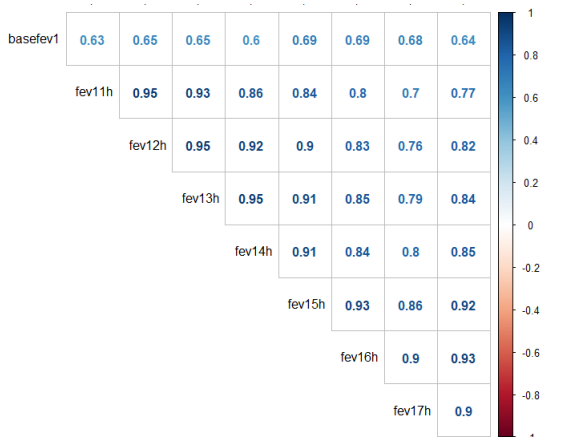

Figura 1. Matriz de correlação amostral.

```
$a
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.170   2.688   3.090   3.072   3.590   4.490

$c
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.210   2.688   3.225   3.251   3.922   4.990

$p
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.950   2.455   2.845   2.792   3.175   3.980
```

Nota-se um desempenho aparentemente superior da droga a e c, não tendo muita diferença em suas distribuições, ao contrário da droga p, que apresentou as medidas de posição um pouco abaixo das demais.

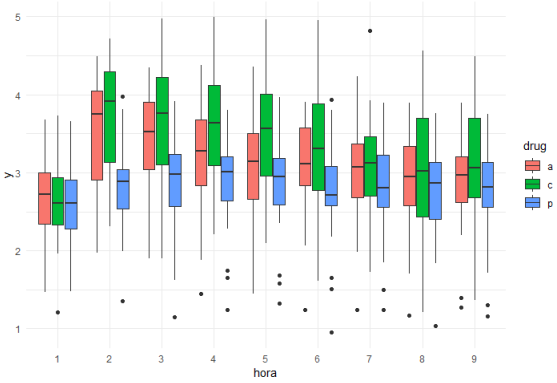

Figura 2. Distribuição do FEV conforme droga e hora.

Ao se analisar o comportamento da resposta conforme a droga e a hora pela Figura 2, nota-se que há um pico na resposta na segunda hora de aplicação e depois há uma queda, sendo o momento 7 o de maior queda e a partir dalí há uma estabilização, isso da indicações de que o remédio tem um efeito que dura aproximadamente 6 horas. Também é possível ver um aumento na respostas em todos os casos, inclusive no placebo. Além disso, nota-se que no tempo inicial existe uma similaridade nas distribuições do FEV entre as drogas, o que permite a comparação entre elas.

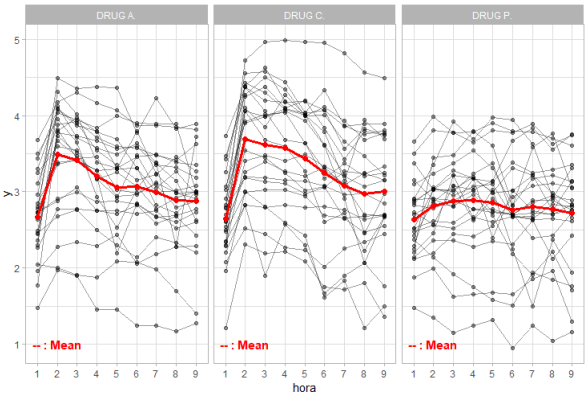

Figura 3. Gráfico de perfis dos indivíduos.

A partir dos gráficos de perfis, ilustrados pela Figura 3, é possível perceber que o intercepto é diferente para cada indivíduo, isto é, cada paciente começa com uma resposta distinta. Além disso, constata-se o que foi visto anteriormente de que as maiores médias da resposta são de indivíduos que utilizaram a droga C e A, respectivamente, sendo que há menos indivíduos que são resistentes ao remédio C do que ao A.

1. Abordagem de análise de variância

Primeiramente será feita uma análise considerando modelos mais simples e posteriormente modelos mistos serão aplicados nos dados

```
## - ANOVA one way - errado -

modelo1=aov(y~drug,data = dados1)
modelo1 %>%
  summary()

             Df Sum Sq Mean Sq F value  Pr(>F)
drug          2  25.78  12.891   24.51 6.1e-11
Residuals   573 301.35   0.526
---

Tukey multiple comparisons of means
    95% family-wise confidence level

$drug
          diff         lwr        upr     p adj
c-a  0,2036979  0,02977412  0,3776217 0,0167939
p-a -0,3108333 -0,48475713 -0,1369095 0,0000918
p-c -0,5145313 -0,68845505 -0,3406074 0,0000000

## - ANOVA two way - errado -

modelo2=aov(y~drug+hora,data = dados1)
modelo2 %>%
summary()

             Df Sum Sq Mean Sq F value   Pr(>F)
drug          2  25.78  12.891  25.675 2.12e-11
hora          7  17.17   2.453   4.885 2.28e-05
Residuals   566 284.18   0.502
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level


$drug
          diff         lwr        upr     p adj
c-a  0,2036979  0,03375498  0,3736409 0,0138935
p-a -0,3108333 -0,48077627 -0,1408904 0,0000602
p-c -0,5145313 -0,68447419 -0,3445883 0,0000000
```

Esses dois modos equivocados de análise, por conta de não considerarem a dependência presente nos dados, tiveram resultados similares, em que foi possível averiguar significância no efeito da droga e da hora.

No teste de comparações múltiplas, ambos captaram diferenças significativas entre todos as drogas, sendo a droga c superior às demais.

```
## - ANOVA com indivíduo como bloco -

modelo3=aov(y~drug*hora+patient,data = dados1)
modelo3 %>% summary()

             Df Sum Sq Mean Sq F value   Pr(>F)
drug          2  25.78  12.891 204.212  < 2e-16
hora          7  17.17   2.453  38.857  < 2e-16
patient      69 247.41   3.586  56.801  < 2e-16
drug:hora    14   6.28   0.449   7.106 1.92e-13
Residuals   483  30.49   0.063
--
  Tukey multiple comparisons of means
    95% family-wise confidence level
$drug
          diff         lwr         upr p adj
c-a  0,2036979  0,1434119  0,2639839     0
p-a -0,3108333 -0,3711193 -0,2505473     0
p-c -0,5145313 -0,5748173 -0,4542452     0
```

Ao se considerar o paciente como bloco, é possível retirar um possível efeito de dependência e assim ser um pouco mais "correto"na análise desses dados. Os resultados do *post hoc* foram similares aos anteriores.

```
## - ANOVA para medidas repetidas -

modelo4=aov(y ~ basefev1+ drug*hora +
            Error(patient/hora),data=dados1)

modelo4 %>% summary()

Error: patient
          Df Sum Sq Mean Sq F value   Pr(>F)
basefev1   1 131.89  131.89  76.984 8.64e-13
drug       2  24.81   12.41   7.241  0.00141
Residuals 68 116.50    1.71
---

Error: patient:hora
           Df Sum Sq Mean Sq F value   Pr(>F)
hora        7  17.17  2.4529  38.857  < 2e-16
drug:hora  14   6.28  0.4486   7.106 1.92e-13
```

```
Residuals 483  30.49  0.0631
---
```

Considerando que a estrutura de covariância é esférica, torna-se viável considerar esse modelo. É possível perceber significância nos efeitos fixos e também no erro composto.

2. Abordagem de modelos mistos

```
### Modelo com estrutura de covariância
Compound Symmetry

mixed1=lme(y ~ basefev1+hora*drug,
random= ~1 | patient,
method="REML",
correlation = corCompSymm(),
data=dados1)

anova(mixed1)
```

	numDF	denDF	F-value	p-value
(Intercept)	1	483	3204.098	<.0001
basefev1	1	68	76.984	<.0001
hora	7	483	38.857	<.0001
drug	2	68	7.241	0.0014
hora:drug	14	483	7.106	<.0001

```
 AIC: 403.2902
--
Multiple Comparisons of Means: Tukey Contrasts
```

	Estimate	Std. Error	z value	Pr(>|z|)
c - a == 0	0,2184	0,1499	1,458	0,145
p - a == 0	-0,6444	0,1499	-4,300	3,42e-05
p - c == 0	-0,8629	0,1498	-5,759	2,54e-08

```
---
```

Considerando que a estrutura de covariância é Compound Symmetry, torna-se viável considerar esse modelo. É possível perceber significância nos efeitos fixos e aleatórios, sendo que no teste de comparações múltiplas foi possível averiguar diferença significativa entre o placebo e as drogas, mas não entre os medicamentos A e P.

```
### Modelo com estrutura de covariância AR(1)
mixed2=lme(y ~ basefev1+hora*drug,
random= ~1 | patient,
method="REML",
correlation = corAR1(),
data=dados1)

anova(mixed2)
```

	numDF	denDF	F-value	p-value
(Intercept)	1	483	3245.828	<.0001
basefev1	1	68	76.574	<.0001
hora	7	483	17.101	<.0001
drug	2	68	7.783	9e-04
hora:drug	14	483	3.942	<.0001

```
 AIC: 303.0306
 --
 Multiple Comparisons of Means: Tukey Contrasts
```

	Estimate	Std. Error	z value	Pr(>|z|)
c - a == 0	0,2182	0,1494	1,460	0,144
p - a == 0	-0,6447	0,1495	-4,313	3,22e-05
p - c == 0	-0,8629	0,1494	-5,775	2,31e-08

Utilizando a estrutura de covariância AR(1) obteve-se os mesmos resultados que para o caso anterior, sendo a interpretação análoga.

```
# Modelo com estrutura de covariância
não estruturada #

mixed3=lme(y ~ basefev1+hora*drug,
random= ~1 | patient,
method="REML",
correlation = corCymm(),
data=dados1)

anova(mixed3)
```

	numDF	denDF	F-value	p-value
(Intercept)	1	483	3526,622	<.0001
basefev1	1	68	83,029	<.0001
hora	7	483	13,251	<.0001
drug	2	68	9,283	3e-04
hora:drug	14	483	3,963	<.0001

```
 AIC: 263,4179
 --
 Multiple Comparisons of Means: Tukey Contrasts
```

	Estimate	Std. Error	z value	Pr(>|z|)
c - a == 0	0,2181	0,1485	1,469	0,142
p - a == 0	-0,6448	0,1485	-4,341	2,84e-05
p - c == 0	-0,8629	0,1485	-5,811	1,87e-08

O resultado foi análogo aos outros dois anteriores.

OBS: O modelo com estrutura de covariância Toeplitz não tinha implementação em R de fácil acesso.

3. Comparando estruturas de covariância

Para verificar qual a estrutura de covariância pode ser escolhida é usual utilizar o teste da razão entre verossimilhanças, desde que se tratem de estruturas "aninhadas", isto é, uma pode ser escrita como caso particular da outra. No caso, só podeemos comparar a CS com a não estruturada e a AR(1) com a não estruturada.

```
# CS versus UN #

anova.lme(mixed1,mixed3)
```

```
       Model df      AIC      BIC     logLik
       Test  L.Ratio p-value
mixed1     1 28 403,2902 524,0187 -173,64508
mixed3     2 55 263,4179 500,5633  -76,70896
1 vs 2 193,8722  <.0001
--
anova.lme(mixed2,mixed3)

# AR(1) versus UN #

       Model df      AIC      BIC     logLik
       Test  L.Ratio p-value
mixed2     1 28 303,0306 423,7592 -123,51532
mixed3     2 55 263,4179 500,5633  -76,70896
1 vs 2 93,61271  <.0001
```

O valor p do teste da razão entre verossimilhanças foi menor do que o nível α=0,05 de significância ($< 0,001$) para ambos os testes. Indicando que há uma diferença significativa entre os modelos. Desta forma, opta-se pelo modelo com mais parâmetros (não estruturado).

IV. CONCLUSÃO

Adotando a classe de modelos mistos com estruturas de covariância CS e AR(1), verifica-se que o segundo tem menor AIC e pela forma da estrutura da matriz de correlação amostral parece ser mais adequado, porém, pelo teste da razão entre as verossimilhanças, há indícios de uma diferença significativa entre esses modelos e o não estruturado (com mais parâmetros).

REFERÊNCIAS

[1]R. C. Team *et al.*, (2013).
[2]D. C. Montgomery, *Design and analysis of experiments* (Wiley New York, 1984).
[3]Y. Pawitan, *In all likelihood: statistical modelling and inference using likelihood* (Oxford University Press, 2001).
[4]R. C. Littell, J. Pendergast, and R. Natarajan, Statistics in medicine **19**, 1793 (2000).