MAT02035 - Modelos para dados correlacionados

Modelos lineares de efeitos mistos: exemplo

Rodrigo Citton P. dos Reis
citton.padilha@ufrgs.br

Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Departamento de Estatística

Porto Alegre, 2022

Six Cities Study of Air Pollution and Health

Six Cities Study of Air Pollution and Health

- Dados longitudinais sobre o crescimento da função pulmonar em crianças e adolescentes do *Six Cities Study of Air Pollution and Health*.
- Os dados são de uma coorte de 300 meninas em idade escolar que moram em Topeka, Kansas, que, na maioria dos casos, estavam matriculadas na primeira ou segunda série (com idades entre seis e sete anos).
- As meninas foram medidas anualmente até a formatura do ensino médio (aproximadamente aos dezoito anos) ou perda de acompanhamento, e cada menina forneceu no mínimo uma e no máximo 12 observações.

Six Cities Study of Air Pollution and Health

- A cada exame, as medidas da função pulmonar foram obtidas a partir da espirometria simples.
- Uma medida amplamente utilizada, calculada a partir da espirometria simples, é o volume de ar expirado no primeiro segundo da manobra, FEV_1 **(variável resposta de interesse)**.
- As variáveis **idade (anos)** e **altura (metros)** também foram registradas em cada uma das visitas.

Estudo Six Cities

```
library(haven)

fev <- read_dta(file = here::here("data", "fev1.dta"))

# cria variável log(fev1/ht)
fev$fev1 <- exp(fev$logfev1)
fev$logfh <- log(fev$fev1/fev$ht)

# uma observação atípica (?)
fev <- fev[- which(fev$logfh < -0.5), ]

names(fev) <- c("id", "altura", "idade", "altura_basal",
                "idade_basal", "logFEV_1", "FEV_1", "logfh")
```

Estudo Six Cities

id	altura	idade	altura_basal	idade_basal	logFEV_1	FEV_1	logfh
13	1.25	7.5483	1.25	7.5483	0.32930	1.389995	0.1061564
13	1.60	17.6099	1.25	7.5483	1.10194	3.010000	0.6319364
13	1.60	18.6092	1.25	7.5483	1.06815	2.909991	0.5981464
14	1.27	7.4251	1.27	7.4251	0.37156	1.449995	0.1325431
15	1.25	7.8412	1.25	7.8412	0.23111	1.259998	0.0079665
15	1.59	15.0746	1.25	7.8412	0.95551	2.599996	0.4917760
15	1.60	16.0739	1.25	7.8412	1.17557	3.239989	0.7055664
15	1.60	16.9281	1.25	7.8412	1.19392	3.299992	0.7239164
15	1.60	17.9877	1.25	7.8412	1.18173	3.260009	0.7117264
16	1.20	6.5079	1.20	6.5079	0.23902	1.270004	0.0566984
16	1.24	7.4606	1.20	6.5079	0.37156	1.449995	0.1564486
16	1.30	8.4846	1.20	6.5079	0.44469	1.560007	0.1823258
16	1.36	9.4812	1.20	6.5079	0.66269	1.940004	0.3552053
16	1.42	10.5681	1.20	6.5079	0.77932	2.179989	0.4286632
16	1.49	11.5811	1.20	6.5079	0.94391	2.570010	0.5451339
16	1.55	13.5250	1.20	6.5079	1.09861	2.999993	0.6603551
16	1.56	14.5325	1.20	6.5079	1.16938	3.219996	0.7246942
16	1.56	15.6167	1.20	6.5079	1.16315	3.199997	0.7184642
16	1.57	16.4709	1.20	6.5079	1.13783	3.119991	0.6867544
16	1.57	17.4648	1.20	6.5079	1.17557	3.239989	0.7244944

Estudo Six Cities

- Observe que, embora as meninas com apenas uma única observação não forneçam diretamente informações sobre mudanças longitudinais ou intraindividuais ao longo do tempo, suas observações em uma única ocasião contribuem para a análise.
  - Por exemplo, essas observações contribuem com informações para a estimativa de variâncias e efeitos entre-indivíduos.
- A tabela anterior revela que esses dados são inerentemente desbalanceados ao longo do tempo, e o grau de desequilíbrio é ainda mais acentuado quando a idade da criança é usada como a variável de tempo.
  - Ou seja, neste conjunto de dados, as crianças entram no estudo em diferentes idades e também têm diferentes ocasiões de medição.

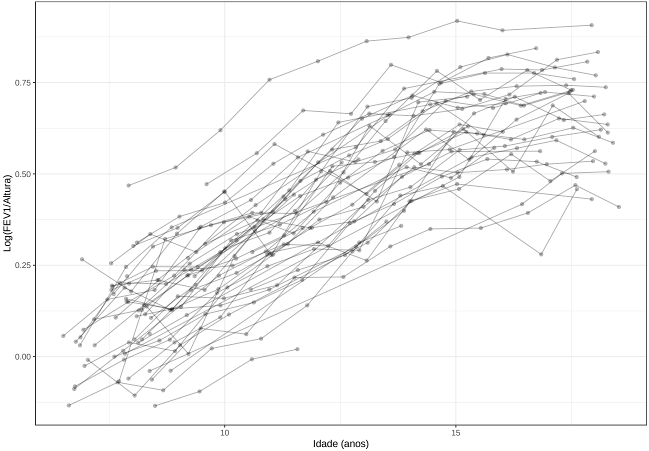
Estudo Six Cities

```
library(ggplot2)

set.seed(5)
id.s <- sample(x = unique(fev$id), size = 50,
               replace = FALSE)

p <- ggplot(data = fev[which(fev$id %in% id.s),],
            mapping = aes(x = idade, y = logfh,
                          group = id)) +
  geom_point(alpha = 0.3) +
  geom_line(alpha = 0.3) +
  labs(x = "Idade (anos)",
       y = "Log(FEV1/Altura)") +
  theme_bw()
p
```

Estudo Six Cities



Estudo Six Cities: modelagem

Quando a idade é usada como a variável de tempo, existem **duas fontes de informação** sobre a relação entre FEV_1 e idade.

1. Há **informações "transversais"** ou entre indivíduos que surgem porque as crianças entram no estudo em idades diferentes.
    - Por exemplo, há informações sobre como o FEV_1 muda com a idade nas observações da linha de base (ou $\text{tempo} = 0$).
2. Há **informações "longitudinais"** ou intra-indivíduo que surgem porque as crianças são medidas repetidamente ao longo do tempo, produzindo medições de FEV_1 em diferentes idades.

Estudo Six Cities: modelagem

- É importante modelar esses dados de uma forma que permita a estimativa separada dos efeitos "transversais" e "longitudinais" da idade do FEV_1.
- É então possível testar se existem diferenças entre os efeitos transversais e longitudinais da idade na FEV_1 e relatar efeitos separados quando necessário ou estimar um efeito combinado, com base em ambas as fontes de informação, se apropriado.
- Observe que os mesmos problemas surgem ao examinar a relação entre FEV_1 e altura.

Estudo Six Cities: modelagem

- O **objetivo da análise** foi determinar como as mudanças na função pulmonar (FEV_1) ao longo do tempo estão relacionadas à idade e altura da criança.
- Pesquisas anteriores indicaram que $\log(\text{FEV}_1)$ tem uma relação aproximadamente linear com idade e $\log(\text{altura})$ em crianças e adolescentes.
- Para distinguir entre os efeitos transversais e longitudinais de idade e $\log(\text{altura})$ em $\log(\text{FEV}_1)$, os valores basais e atuais dessas covariáveis foram incluídos no modelo para a média.
- Como esses dados são inerentemente desbalanceados, explicar a covariância entre as observações repetidas na mesma criança por meio de uma estrutura de efeitos aleatórios é muito adequado.

Estudo Six Cities: modelagem

- Neste examplo, vamos considerar que o intercepto e a inclinação para a idade variem aleatoriamente de uma criança para outra.
- Especificamente, consideramos o seguinte modelo para $\log(\text{FEV}_1)$:

$$\begin{aligned} \text{E}\left(Y_{ij}|b_i\right) &= \beta_1 + \beta_2\text{idade}_{ij} + \beta_3\log(\text{altura})_{ij} + \beta_4\text{idade}_{i1} + \beta_5\log(\text{altura})_{i1} \\ & \quad + b_{i1} + b_{i2}\text{idade}_{ij}, \end{aligned}$$

em que Y_{ij} é o $\log(\text{FEV}_1)$ para o i-ésima chiança na j-ésima ocasião, e idade_{i1} e $\log(\text{altura})_{i1}$ são a idade e $\log(\text{altura})$ basal para a i-ésima criança.

Estudo Six Cities: modelagem

- Neste modelo, β_2 e β_3 são os efeitos longitudinais da idade e log(altura), respectivamente, enquanto $(\beta_2 + \beta_4)$ e $(\beta_3 + \beta_5)$ são os efeitos transversais correspondentes.
- Ou seja, β_4 e β_5 representam as diferenças entre os efeitos longitudinal e transversal da idade e log(altura), respectivamente.

Estudo Six Cities: ajuste

```
library(nlme)

mod1 <- lme(fixed = logFEV_1 ~ idade + log(altura) +
              idade_basal + log(altura_basal),
            random = ~ idade | id,
            # method = "REML",
            data = fev,
            na.action = na.omit)
```

Estudo Six Cities: ajuste

```
# library(lme4)
#
# mod <- lmer(formula = logFEV_1 ~ idade + log(altura) +
#                idade_basal + log(altura_basal) +
#                (1 + idade | id),
#             REML = TRUE,
#             data = fev,
#             na.action = na.omit)
```

Estudo Six Cities: estimativas

As estimativas REML dos efeitos fixos são exibidas na tabela a seguir.

	Estimativa	EP	Z
(Intercept)	-0.2883	0.0387	-7.45
idade	0.0235	0.0014	16.86
log(altura)	2.2372	0.0435	51.39
idade_basal	-0.0165	0.0075	-2.21
log(altura_basal)	0.2182	0.1455	1.50

Estudo Six Cities: estimativas

- Com base na magnitude das estimativas de β_4 e β_4, em relação aos seus erros padrão, há evidências de uma diferença significativa entre os efeitos longitudinais e transversais da idade, mas não de log(altura).
- As magnitudes dos efeitos longitudinais e transversais de log(altura) são bastante semelhantes (2.24 versus 2.46), enquanto as magnitudes dos efeitos longitudinais e transversais da idade são notavelmente diferentes (0.024 versus 0.007).
- Ou seja, os efeitos longitudinais e transversais da idade sobre as mudanças no FEV_1 ($e^{0.024} \approx 1.025$ versus $e^{0.007} \approx 1.007$) são discerníveis diferentes.
- Isso pode ser devido, em parte, à quantidade relativamente pequena de variabilidade nas idades na linha de base (em relação à variabilidade nas idades ao longo da duração do estudo).
- A partir dos efeitos longitudinais da idade e log(altura), há evidências claras de que as mudanças no $\log(\text{FEV}_1)$ estão relacionadas tanto com a idade quanto com a altura.

Estudo Six Cities: interpretação das estimativas dos efeitos fixos

O modelo para a média, calculado sobre a distribuição dos efeitos aleatórios indivíduo-específico, é dado por

$$E\left(Y_{ij}\right) = \beta_1 + \beta_2 \text{idade}_{ij} + \beta_3 \log(\text{altura})_{ij} + \beta_4 \text{idade}_{i1} + \beta_5 \log(\text{altura})_{i1}.$$

Além disso, este modelo pode ser **reexpresso em** termos de **dois modelos**, um **modelo transversal** e um **modelo longitudinal**.

Estudo Six Cities: interpretação das estimativas dos efeitos fixos

O primeiro é dada por

$$\begin{aligned} \mathrm{E}\left(Y_{i1}\right) &= \beta_1 + \beta_2 \text{idade}_{i1} + \beta_3 \log(\text{altura})_{i1} + \beta_4 \text{idade}_{i1} + \beta_5 \log(\text{altura})_{i1} \\ &= \beta_1 + (\beta_2 + \beta_4)\text{idade}_{i1} + (\beta_3 + \beta_5) \log(\text{altura})_{i1}, \end{aligned}$$

enquanto o segundo é dado por

$$\begin{aligned} \mathrm{E}\left(Y_{ij} - Y_{i1}\right) &= \beta_1 + \beta_2 \text{idade}_{ij} + \beta_3 \log(\text{altura})_{ij} + \beta_4 \text{idade}_{i1} + \beta_5 \log(\text{altura})_{i1} \\ &\quad - \{\beta_1 + \beta_2 \text{idade}_{i1} + \beta_3 \log(\text{altura})_{i1} + \beta_4 \text{idade}_{i1} + \beta_5 \log(\text{altura})_{i1}\} \\ &= \beta_2(\text{idade}_{ij} - \text{idade}_{i1}) + \beta_3[\log(\text{altura})_{ij} - \log(\text{altura})_{i1}]. \end{aligned}$$

Estudo Six Cities: interpretação das estimativas dos efeitos fixos

- O efeito longitudinal de log(altura), β_3, tem interpretação em termos de mudanças na média do $\log(\text{FEV}_1)$ para um aumento de uma unidade em log(altura), para qualquer mudança na idade (por exemplo, durante um intervalo de dois anos).
- Da mesma forma, o efeito longitudinal da idade, β_2, tem interpretação em termos das mudanças a média do $\log(\text{FEV}_1)$ para um aumento de um ano na idade, para qualquer mudança no log(altura).

Estudo Six Cities: interpretação das estimativas dos efeitos fixos

- O coeficiente para $\log(\text{altura})$, 2.24, não é diretamente interpretável porque uma mudança de uma única unidade em $\log(\text{altura})$ corresponde a um aumento de quase três vezes (ou $e^{1.0} \approx 2.7$) na altura.
- Em vez disso, provavelmente é mais significativo considerar o efeito de um aumento de 10% na altura.
- Na escala logarítmica isso corresponde a um aumento de 0.1 na $\log(\text{altura})$, já que $e^{0.1} \approx 1.1$.
- Portanto, um aumento de 10% na altura está associado a um aumento de 0.224 no $\log(\text{FEV}_1)$.
  - Observe que um aumento de 0.224 no $\log(\text{FEV}_1)$ corresponde a um aumento de 25% no FEV_1 (já que $e^{0.224} = 1.25$).

Estudo Six Cities: interpretação das estimativas dos efeitos fixos

- Por outro lado, o coeficiente para a idade, 0.024, é mais diretamente interpretável.
- A estimativa do efeito longitudinal da idade indica que um aumento de um ano na idade está associado a um aumento de 0.024 no $\log(\text{FEV}_1)$ ou um aumento de aproximadamente 2,5% ($e^{0.024} \approx 1.025$) no FEV_1, para qualquer mudança fixa na altura.

Estudo Six Cities: estimativas dos componentes de variância

```
getVarCov(mod1, type = "random.effects")
```

```
## Random effects variance covariance matrix
##              (Intercept)       idade
## (Intercept)   0.01220700 -4.3253e-04
## idade        -0.00043253  5.0103e-05
##   Standard Deviations: 0.11049 0.0070784
```

```
summary(mod1)$sigma^2
```

```
## [1] 0.003628602
```

```
VarCorr(mod1)
```

```
## id = pdLogChol(idade)
##             Variance     StdDev      Corr
## (Intercept) 1.220705e-02 0.110485541 (Intr)
## idade       5.010347e-05 0.007078381 -0.553
## Residual    3.628602e-03 0.060237881
```

Estudo Six Cities: estimativas dos componentes de variância

- A **covariância marginal** ($\text{Cov}(Y_i) = \Sigma_i$) entre as observações repetidas é uma função desses parâmetros de variância e covariância (e σ^2) e das idades da criança quando as observações foram obtidas.

Estudo Six Cities: estimativas dos componentes de variância

- As correlações estimadas para medições anuais de 7 a 18 anos são exibidas na tabela a seguir

7	8	9	10	11	12	13	14	15	16	17	18
1.00	0.70	0.69	0.68	0.67	0.66	0.64	0.62	0.60	0.58	0.56	0.54
0.70	1.00	0.70	0.69	0.68	0.67	0.66	0.65	0.63	0.61	0.60	0.58
0.69	0.70	1.00	0.70	0.70	0.69	0.68	0.67	0.66	0.64	0.63	0.61
0.68	0.69	0.70	1.00	0.70	0.70	0.70	0.69	0.68	0.67	0.66	0.64
0.67	0.68	0.70	0.70	1.00	0.71	0.71	0.70	0.70	0.69	0.68	0.67
0.66	0.67	0.69	0.70	0.71	1.00	0.72	0.72	0.71	0.71	0.71	0.70
0.64	0.66	0.68	0.70	0.71	0.72	1.00	0.73	0.73	0.73	0.72	0.72
0.62	0.65	0.67	0.69	0.70	0.72	0.73	1.00	0.74	0.74	0.74	0.74
0.60	0.63	0.66	0.68	0.70	0.71	0.73	0.74	1.00	0.75	0.75	0.75
0.58	0.61	0.64	0.67	0.69	0.71	0.73	0.74	0.75	1.00	0.76	0.76
0.56	0.60	0.63	0.66	0.68	0.71	0.72	0.74	0.75	0.76	1.00	0.77
0.54	0.58	0.61	0.64	0.67	0.70	0.72	0.74	0.75	0.76	0.77	1.00

Estudo Six Cities: estimativas dos componentes de variância

- Esse padrão de correlação reforça a observação que fizemos nas aulas anteriores: a correlação entre medições repetidas de muitos desfechos de saúde raramente decai a zero, mesmo quando estão separados por muitos anos.

Estudo Six Cities: modelagem alternativa

$$
\begin{aligned}
\mathrm{E}\left(Y_{ij}|b_i\right) &= \beta_1 + \beta_2 \mathrm{idade}_{ij} + \beta_3 \log(\mathrm{altura})_{ij} + \beta_4 \mathrm{idade}_{i1} + \beta_5 \log(\mathrm{altura})_{i1} \\
&\quad + b_{i1} + b_{i2} \log(\mathrm{altura})_{ij},
\end{aligned}
$$

```
mod2 <- lme(fixed = logFEV_1 ~ idade + log(altura) +
              idade_basal + log(altura_basal),
            random = ~ log(altura) | id,
            # method = "REML",
            data = fev,
            na.action = na.omit)
```

Estudo Six Cities: modelagem alternativa

	Estimativa	EP	Z
(Intercept)	-0.2846	0.0390	-7.29
idade	0.0233	0.0012	18.65
log(altura)	2.2523	0.0461	48.82
idade_basal	-0.0163	0.0074	-2.19
log(altura_basal)	0.1808	0.1455	1.24

Estudo Six Cities: modelagem alternativa

Qual o modelo mais apropriado?

- Como ambos os modelos têm o mesmo número de parâmetros de covariância, podemos fazer esse julgamento com base em uma comparação direta de suas log-verossimilhanças REML maximizadas.

```
logLik(mod1)
```

```
## 'log Lik.' 2283.941 (df=9)
```

```
logLik(mod2)
```

```
## 'log Lik.' 2294.737 (df=9)
```

- A comparação das log-verossimilhanças maximizadas indica que o modelo com inclinações aleatórias para log(altura) deve ser preferido.

Estudo Six Cities: modelagem alternativa

Um modelo com inclinações aleatórias para idade e log(altura).

```
mod3 <- lme(fixed = logFEV_1 ~ idade + log(altura) +
              idade_basal + log(altura_basal),
            random = ~ idade + log(altura) | id,
            # method = "REML",
            data = fev,
            na.action = na.omit)

logLik(mod3)
```

```
## 'log Lik.' 2294.95 (df=12)
```

Estudo Six Cities: modelagem alternativa

```
library(varTestnlme)

vt <- varCompTest(m1 = mod3, m0 = mod2)

library(EnvStats)

print(vt)
```

```
##
## Results of Hypothesis Test
## --------------------------
##
## Null Hypothesis:                  variance of the random effect associated to
##
## Alternative Hypothesis:           variance of the random effect associated to
##
## Test Name:                        Likelihood ratio test for variance componen
##
## Data:
##
## Test Statistic:                   LRT = 0.4261284
##
```

Estudo Six Cities: modelagem alternativa

```
## Test Statistic Parameters:          df        = 2, 3
##                                     weights   = 0.5, 0.5
##                                     sdweights = 0, 0
##                                     FIM       = NA
##
## P-values:                           pvalue.weights    = 0.871449
##                                     pvalue.sample     =       NA
##                                     pvalue.lowerbound = 0.871449
##                                     pvalue.upperbound = 0.871449
```

```
summary(vt)
```

```
## Variance components testing in mixed effects models
## Testing that:
##  variance of the random effect associated to idade is equal to 0
## against the alternative that:
##  variance of the random effect associated to idade > 0
##
##  Likelihood ratio test statistic:
##  LRT =  0.4261284
##
##  Limiting distribution:
##  mixture of 2 chi-bar-square distributions with degrees of freedom 2 3
```

Estudo Six Cities: modelagem alternativa

```
##  associated (exact) weights: 0.5 0.5
##
##  p-value of the test:
##  from exact weights: 0.871449
```

Bons estudos!

Instituto de MATEMÁTICA E ESTATÍSTICA

60 anos UFRGS