Estudo de Caso - Modelo Misto

José Luiz Padilha da Silva

01 de outubro de 2018

Introdução

O estudo *Six Cities Study of Air Pollution and Health* foi um estudo longitudinal destinado a caracterizar o desenvolvimento pulmonar, medido por mudanças na função pulmonar em crianças e adolescentes, e os fatores que influenciam tal desenvolvimento. Uma coorte de 13.379 crianças nascidas a partir de 1967 abrangeu seis comunidades dos Estados Unidos (estados de Massachusetts, Tennessee, Missouri, Ohio, Wisconsin e Kansas).

A maioria das crianças estava matriculada na primeira ou segunda série (entre seis e sete anos) e as medidas dos participantes do estudo foram obtidos anualmente até a conclusão do ensino médio ou perda de acompanhamento. Em cada exame anual foi realizada uma espirometria, medição da função pulmonar; e um questionário de saúde respiratória foi preenchido por um pai ou responsável pela criança.

O exercício básico na espirometria é a inspiração máxima (ou a respiração) seguida de exalação forçada em um recipiente fechado. Muitas medidas podem ser derivadas da curva espirométrica do volume exalado versus tempo. Uma medida amplamente utilizada é o volume total de ar exalado no primeiro segundo (FEV_1).

Os dados considerados aqui consistem de todas as medidas de FEV_1, altura e idade obtidos de um subconjunto aleatoriamente selecionado ($N = 300$) dentre as meninas.

Análise Exploratória

```r
dados=read.table("fev1.csv",h=TRUE,dec=',',sep=";"); head(dados)
```

```
##   Id Height     Age Initial.Height Initial.Age LogFEV1
## 1  1   1.20  9.3415            1.2      9.3415 0.21511
## 2  1   1.28 10.3929            1.2      9.3415 0.37156
## 3  1   1.33 11.4524            1.2      9.3415 0.48858
## 4  1   1.42 12.4600            1.2      9.3415 0.75142
## 5  1   1.48 13.4182            1.2      9.3415 0.83291
## 6  1   1.50 15.4743            1.2      9.3415 0.89200
```

```r
summary(dados)
```

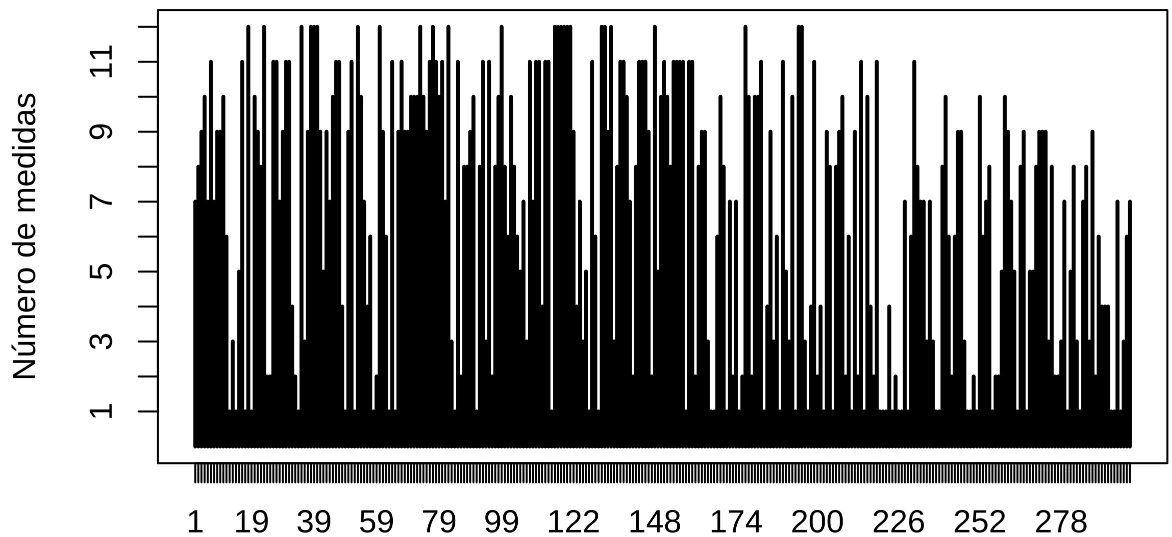
```
##        Id            Height           Age         Initial.Height
##  Min.   :  1.0   Min.   :1.110   Min.   : 6.434   Min.   :1.110
##  1st Qu.: 69.0   1st Qu.:1.370   1st Qu.: 9.717   1st Qu.:1.220
##  Median :129.5   Median :1.540   Median :12.595   Median :1.260
##  Mean   :135.8   Mean   :1.497   Mean   :12.566   Mean   :1.276
##  3rd Qu.:199.0   3rd Qu.:1.620   3rd Qu.:15.366   3rd Qu.:1.320
##  Max.   :300.0   Max.   :1.790   Max.   :18.691   Max.   :1.720
##   Initial.Age        LogFEV1
##  Min.   : 6.434   Min.   :-0.6932
##  1st Qu.: 7.136   1st Qu.: 0.5481
##  Median : 7.781   Median : 0.8671
##  Mean   : 8.030   Mean   : 0.8152
##  3rd Qu.: 8.449   3rd Qu.: 1.0978
##  Max.   :14.067   Max.   : 1.5953
```

```r
dim(dados)
```

```
## [1] 1994    6
```

O gráfico a seguir mostra o número de medidas repetidas por `id`.

```r
plot(table(dados$Id),yaxt="n",ylab="Número de medidas"); axis(2,at=1:12)
```



Note que, embora meninas com apenas uma observação não forneçam diretamente informação sobre a mudança longitudinal, elas também contribuem para a análise (na estimação das variâncias e efeitos entre-indivíduos).

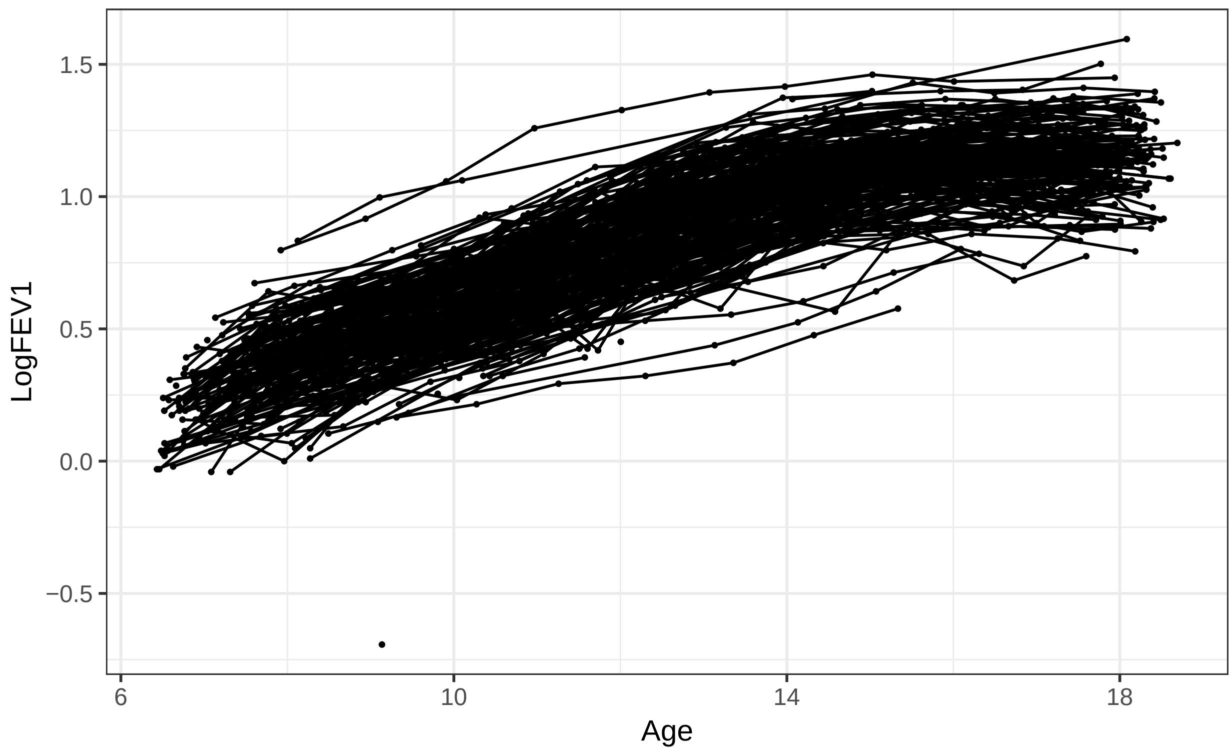
```r
head(dados,15)#duas primeiras crianças
```

```
##    Id Height     Age Initial.Height Initial.Age LogFEV1
## 1   1   1.20  9.3415           1.20      9.3415 0.21511
## 2   1   1.28 10.3929           1.20      9.3415 0.37156
## 3   1   1.33 11.4524           1.20      9.3415 0.48858
## 4   1   1.42 12.4600           1.20      9.3415 0.75142
## 5   1   1.48 13.4182           1.20      9.3415 0.83291
## 6   1   1.50 15.4743           1.20      9.3415 0.89200
## 7   1   1.52 16.3723           1.20      9.3415 0.87129
## 8   2   1.13  6.5873           1.13      6.5873 0.30748
## 9   2   1.19  7.6496           1.13      6.5873 0.35066
## 10  2   1.49 12.7392           1.13      6.5873 0.75612
## 11  2   1.53 13.7741           1.13      6.5873 0.86710
## 12  2   1.55 14.6940           1.13      6.5873 1.04732
## 13  2   1.56 15.8220           1.13      6.5873 1.15373
## 14  2   1.57 16.6680           1.13      6.5873 0.92426
## 15  2   1.57 17.6318           1.13      6.5873 1.13462
```

Os dados são inerentemente desbalanceados (diferentes tempos e diferentes ocasiões de medida), e o grau de desbalanceamento é ainda mais notável quando a idade é usada como substituta para o tempo.

A seguir, um gráfico de perfis.

```
library(ggplot2)
p1=ggplot(dados, aes(x=Age, y=LogFEV1)) + geom_line(aes(group=Id))  + theme_bw() +
   theme(legend.position="top") + scale_x_continuous(breaks=c(6,10,14,18))
p1 + geom_point(size=0.5)
```



Quando a idade é usada como tempo, existem duas fontes de informação sobre o relacionamento entre FEV_1 e idade.

- Primeiro, a informação "transversal" (ou entre indivíduos) que surge porque as crianças entram no estudo em diferentes idades. Por exemplo, há informações sobre como FEV_1 muda com a idade no *baseline* (ou $tempo = 0$).
- Em segundo lugar, a "longitudinal" (ou dentro do indivíduo) que surge porque as crianças são mensuradas repetidamente ao longo do tempo.

Como há duas fontes de informação potencialmente conflitantes sobre o relacionamento entre FEV_1 e idade, é importante modelar os dados de forma a obter estimativas separadas sobre estes efeitos da idade e FEV_1.

É possível testar se existem diferenças entre os efeitos "transversais" e "longitudinais" de idade na FEV_1, e relatar efeitos separados, quando necessário, ou estimar uma combinação de efeitos, com base em ambas as fontes de informação, se apropriado. A mesma questão surge ao examinar a relação entre FEV_1 e altura.

Ajuste dos Modelos

O objetivo do estudo é determinar como mudanças na função pulmonar (determinada pela FEV_1) ao longo do tempo estão relacionadas com a idade e altura da criança. Pesquisas anteriores indicaram que $log(FEV_1)$

tem uma relação aproximadamente linear com idade e $log(altura)$ em crianças e adolescentes.

Para distinguir entre efeitos "transversais" e "longitudinais" de idade e $log(altura)$ em $log(FEV_1)$, valores *baseline* e atuais destas covariáveis foram incluídos no modelo para a média. Como os dados são inerentes desbalanceados, modelar a covariância entre as observações repetidas via uma estrutura de modelos mistos é bastante atraente.

Considere um modelo com intercepto e inclinação para idade variando aleatoriamente de criança para criança:

$$E(Y_{ij}|b_i) = \beta_0 + \beta_1 Idade_{ij} + \beta_2 log(Altura)_{ij} + \beta_3 Idade_{i1} + \beta_4 log(Altura)_{i1} + b_{0i} + b_{i1} Idade_{ij},$$

em que

- Y_{ij} é a $log(FEV_1)$ da i-ésima criança na j-ésima ocasião,
- $Idade_{i1}$ e $log(Altura)_{i1}$ são a idade inicial e a $log(Altura)$ inicial para a i-ésima criança.

Neste modelo β_1 e β_2 são os efeitos longitudinais de $Idade$ e $log(Altura)$, respectivamente, enquanto $(\beta_1 + \beta_3)$ e $(\beta_2 + \beta_4)$ são os correspondentes efeitos transversais.

Um análise preliminar revelou que uma medida de FEV_1 era claramente um *outlier*. Esta medida, de uma menina com apenas a avaliação *baseline*, foi removida. Todas as análises subsequentes são baseadas em dados de 299 meninas (com um total de 1993 medidas).

```
dados=subset(dados,Id!=197)
```

As estimativas de REMV são dadas a seguir:

```
library(nlme)
m1 = lme(LogFEV1~Age+log(Height)+Initial.Age+log(Initial.Height),random=~Age|Id,data=dados)
round(coef(summary(m1)),4)
```

```
##                       Value Std.Error   DF t-value p-value
## (Intercept)         -0.2883    0.0387 1692 -7.4470  0.0000
## Age                  0.0235    0.0014 1692 16.8623  0.0000
## log(Height)          2.2372    0.0435 1692 51.3859  0.0000
## Initial.Age         -0.0165    0.0075  296 -2.2136  0.0276
## log(Initial.Height)  0.2182    0.1455  296  1.4995  0.1348
```

- Há evidências de diferença entre os efeitos longitudinais e transversais de $Idade$ mas não de $log(Altura)$.
- As magnitudes dos efeitos longitudinais e transversais de $log(Altura)$ são similares $(2,24$ versus $2.46)$, mas as as magnitudes destes efeitos para $Idade$ são bastante diferentes $(0,024$ versus $0,007)$. Isto é, os efeitos longitudinais e transversais de $Idade$ em FEV_1 $(e^{0,024} \approx 1,025$ versus $e^{0,007} \approx 1,007)$ são diferentes.
- Com relação aos efeitos longitudinais de $Idade$ e $log(Altura)$, há evidências de que mudanças em $log(FEV_1)$ estejam relacionadas tanto com $Idade$ como com $Altura$.

Vamos agora considerar a interpretação das estimativas dos efeitos fixos. O modelo para a média, ponderando sobre a distribuição dos efeitos aleatórios, é dados por:

$$E(Y_{ij}) = \beta_0 + \beta_1 Idade_{ij} + \beta_2 log(Altura)_{ij} + \beta_3 Idade_{i1} + \beta_4 log(Altura)_{i1}.$$

Além disso, este modelo pode ser reexpresso em termos de dois modelos, um modelo transversal e um modelo longitudinal. O primeiro é dado por:

$$\begin{aligned} E(Y_{i1}) =& \beta_0 + \beta_1 Idade_{i1} + \beta_2 log(Altura)_{i1} + \beta_3 Idade_{i1} + \beta_4 log(Altura)_{i1} \\ =& \beta_0 + (\beta_1 + \beta_3) Idade_{i1} + (\beta_2 + \beta_4) log(Altura)_{i1}, \end{aligned}$$

enquanto o segundo é dado por

$$\begin{aligned} E(Y_{ij} - Y_{i1}) =& \beta_0 + \beta_1 Idade_{ij} + \beta_2 log(Altura)_{ij} + \beta_3 Idade_{i1} + \beta_4 log(Altura)_{i1} \\ & - \{\beta_0 + \beta_1 Idade_{i1} + \beta_2 log(Altura)_{i1} + \beta_3 Idade_{i1} + \beta_4 log(Altura)_{i1}\} \\ =& \beta_1 (Idade_{ij} - Idade_{i1}) + \beta_2 \{log(Altura)_{ij} - log(Altura)_{i1}\}. \end{aligned}$$

O efeito longitudinal de $log(Altura)$, β_2, tem interpretação em termos de mudança na média de $log(FEV_1)$ para o aumento de uma unidade de $log(Altura)$, dada qualquer mudança em $Idade$ (ex: durante um intervalo de dois anos). Similarmente, o efeito longitudinal de $Idade$, β_1, tem interpretação em termos de mudança na média de $log(FEV_1)$ para um aumento unitário na $Idade$, dada qualquer mudança em $log(Altura)$.

O coeficiente para $log(Altura)$, $\beta_2 = 2,24$, não é diretamente interpretável porque uma mudança de uma unidade em $log(Altura)$ corresponde a um aumento quase triplicado (ou $e^{1,0} \approx 2.7$) na $Altura$. Ao invés disso, é provavelmente mais razoável considerar o efeito de um aumento de 10% na $Altura$. Nesta escala logarítmica, isso corresponde a um aumento de $0,1$ em $log(Altura)$, já que $e^{0,1} \approx 1,1$. Portanto, um aumento de 10% na $Altura$ (correspondendo a um aumento de aproximadamente $0,1$ em $log(Altura)$) está associado com um aumento de $0,224$ em $log(FEV_1)$. Note que um aumento de $0,224$ em $log(FEV_1)$ corresponde a um aumento de 25% em FEV_1 (pois $e^{0,224} = 1,25$).

Por outro lado, o coeficiente para $Idade$, $\beta_1 = 0,024$, é mais diretamente interpretável. A estimativa do efeito longitudinal da $Idade$ indica que o aumento de um ano está associado com um aumento de $0,024$ em $log(FEV_1)$ ou aproximadamente 2,5% ($e^{0,024} \approx 1,025$) em FEV_1, para qualquer mudança fixa em $Altura$.

Considere agora as estimativas da variância residual e dos componentes de variância dos efeitos aleatórios.

```
VarCorr(m1)
```

```
## Id = pdLogChol(Age)
##             Variance     StdDev      Corr
## (Intercept) 1.220705e-02 0.110485538 (Intr)
## Age         5.010347e-05 0.007078381 -0.553
## Residual    3.628602e-03 0.060237879
```

A covariância marginal entre as observações repetidas é função destes parâmetros e das idades das crianças nas quais as observações foram medidas. Para crianças de 7 a 18 anos temos as correlações:

```
s2_erro=3.628602e-03; s2_b0=1.220705e-02; s2_b1=5.010347e-05
s_b01=0.110485538*0.007078381*(-0.553)
#
Z=cbind(1,7:18)
cov_b=matrix(c(s2_b0,s_b01,s_b01,s2_b1),2,2)
cov_y=Z%*%cov_b%*%t(Z)+s2_erro*diag(nrow(Z))
corr_y=cov2cor(cov_y);rownames(corr_y)=colnames(corr_y)=7:18
round(corr_y,2)
```

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

Estes resultados indicam uma forte correlação positiva entre as medidas de $log(FEV_1)$ que diminui pouco após um período de 11 anos de acompanhamento. Como discutido anteriormente, a correlação entre medidas repetidas raramente decai para zero, mesmo que as observações estejam separadas por muitos anos.

Finalmente, note que a correlação entre as medidas repetidas foi modelada pela introdução de efeitos aleatórios no intercepto e inclinação de $Idade$. Alternativamente, poderíamos considerar um modelo com inclinações aleatórias para $log(Altura)$. Assumir que as inclinações para $log(Altura)$ variam aleatoriamente também induzia covariância entre as medidas repetidas mas com correlações que são funções não da $Idade$ mas da $Altura$ das crianças.

Considere, então, o seguinte modelo:

$$E(Y_{ij}|b_i) = \beta_0 + \beta_1 Idade_{ij} + \beta_2 log(Altura)_{ij} + \beta_3 Idade_{i1} + \beta_4 log(Altura)_{i1} + b_{0i} + b_{i1} log(Altura)_{ij},$$

As estimativas REMV são dados a seguir:

```
m2 = lme(LogFEV1~Age+log(Height)+Initial.Age+log(Initial.Height),random=~log(Height)|Id,
         data=dados)
round(coef(summary(m2)),4)
```

```
##                       Value Std.Error   DF t-value p-value
## (Intercept)         -0.2846    0.0390 1692 -7.2950  0.0000
## Age                  0.0233    0.0012 1692 18.6549  0.0000
## log(Height)          2.2523    0.0461 1692 48.8239  0.0000
## Initial.Age         -0.0163    0.0074  296 -2.1908  0.0292
## log(Initial.Height)  0.1808    0.1455  296  1.2427  0.2150
```

Os valores são qualitativamente muito simulares àqueles encontrados anteriormente. Qual dos modelos é então mais apropriado aos dados?

Já que ambos possuem o mesmo número de parâmetros de covariância, vamos compará-los com base nas log-verossimilhanças, AIC e BIC.

```
anova(m1,m2)
```

```
##    Model df       AIC       BIC   logLik
## m1     1  9 -4549.882 -4499.528 2283.941
## m2     2  9 -4571.473 -4521.119 2294.737
```

O modelo com inclinações aleatórias para $log(Altura)$ deve ser preferido. Para fins ilustrativos, vamos ajustar um modelo com inclinações aleatórias tanto para $Idade$ e $log(Altura)$. Neste caso, as covariâncias entre as medidas repetidas são funções tanto da $Idade$ como da $Altura$ das crianças.

```
m3 = lme(LogFEV1~Age+log(Height)+Initial.Age+log(Initial.Height),
         random=~(Age+log(Height))|Id,data=dados)
anova(m1,m2,m3)
```

```
##    Model df       AIC       BIC   logLik   Test   L.Ratio p-value
## m1     1  9 -4549.882 -4499.528 2283.941
## m2     2  9 -4571.473 -4521.119 2294.737
## m3     3 12 -4565.899 -4498.761 2294.950 2 vs 3 0.4261294  0.9348
```

Não notamos uma melhora discernível com relação a um modelo com inclinações aleatórias apenas para $log(Altura)$. A seguir calculamos o valor p do teste considerando uma mistura de distribuições χ^2.

```
chisq=2*(m3$logLik-m2$logLik);chisq
```

```
## [1] 0.4261294
```

```
pchisq(chisq, 3, lower.tail = FALSE)
```

```
## [1] 0.9347934
```

```
0.5*pchisq(chisq, 2, lower.tail = FALSE) + 0.5*pchisq(chisq, 3, lower.tail = FALSE)
```

```
## [1] 0.8714486
```