# Exemplo: GEE e Modelo Misto

*José Luiz Padilha da Silva*

*05 de setembro de 2018*

## Exemplo: Dados de Crescimento

Vamos revisitar os dados de crescimento de Potthoff & Roy (1964). No exemplo, são apresentadas medidas de crescimento de 11 meninas e 16 meninos. As medidas referem-se à distância entre dois marcos faciais (do centro da pituitária à fissura do maxilar) em quatro idades (8, 10, 12 e 14 anos). O objetivo era descrever e comparar o crescimento de meninos e meninas.

### Análise Exploratória

Os dados estão disponíveis no `R` no pacote `mice` e podem ser acessados como:

```
library(mice); library(plyr); library(ggplot2); library(gridExtra); library(nlme)
data(potthoffroy)
```

Resumo dos dados por sexo:

```
with(potthoffroy,by(potthoffroy[,-c(1,2)],sex,summary,digits=3))
```

```
## sex: F
##        d8             d10             d12             d14
##  Min.   :16.5   Min.   :19.0   Min.   :19.0   Min.   :19.5
##  1st Qu.:20.2   1st Qu.:21.0   1st Qu.:21.8   1st Qu.:22.8
##  Median :21.0   Median :22.5   Median :23.0   Median :24.0
##  Mean   :21.2   Mean   :22.2   Mean   :23.1   Mean   :24.1
##  3rd Qu.:22.2   3rd Qu.:23.5   3rd Qu.:24.2   3rd Qu.:25.8
##  Max.   :24.5   Max.   :25.0   Max.   :28.0   Max.   :28.0
## --------------------------------------------------------
## sex: M
##        d8             d10             d12             d14
##  Min.   :17.0   Min.   :20.5   Min.   :22.5   Min.   :25.0
##  1st Qu.:21.9   1st Qu.:22.4   1st Qu.:23.9   1st Qu.:26.0
##  Median :23.0   Median :23.5   Median :25.0   Median :26.8
##  Mean   :22.9   Mean   :23.8   Mean   :25.7   Mean   :27.5
##  3rd Qu.:24.1   3rd Qu.:25.1   3rd Qu.:26.6   3rd Qu.:28.8
##  Max.   :27.5   Max.   :28.0   Max.   :31.0   Max.   :31.5
```

Correlações marginais:

```
cor(potthoffroy[,-c(1:2)])
```

```
##            d8       d10       d12       d14
## d8  1.0000000 0.6255833 0.7108079 0.5998338
## d10 0.6255833 1.0000000 0.6348775 0.7593268
## d12 0.7108079 0.6348775 1.0000000 0.7949980
## d14 0.5998338 0.7593268 0.7949980 1.0000000
```

Correlações por sexo:

```r
with(potthoffroy,by(potthoffroy[,-c(1,2)],sex,cor))
```

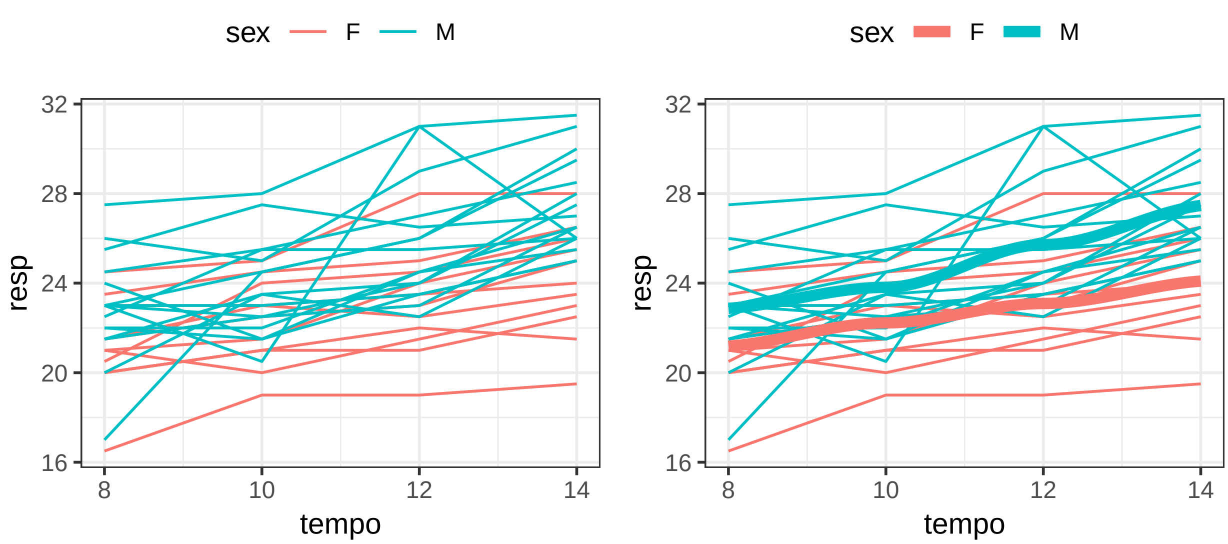
```
## sex: F
##            d8       d10       d12       d14
## d8  1.0000000 0.8300900 0.8623146 0.8413558
## d10 0.8300900 1.0000000 0.8954156 0.8794236
## d12 0.8623146 0.8954156 1.0000000 0.9484070
## d14 0.8413558 0.8794236 0.9484070 1.0000000
## --------------------------------------------------------
## sex: M
##            d8       d10       d12       d14
## d8  1.0000000 0.4373932 0.5579310 0.3152311
## d10 0.4373932 1.0000000 0.3872909 0.6309234
## d12 0.5579310 0.3872909 1.0000000 0.5859866
## d14 0.3152311 0.6309234 0.5859866 1.0000000
```

Transformação dos dados para o formato longo:

```r
dados=reshape(data=potthoffroy,direction="long", idvar="id", v.names="resp",
              varying = list(names(potthoffroy)[3:6]), time= c(8,10,12,14), timevar="tempo")
dados=arrange(dados, id) #Ordenamos os dados por ID, função do pacote plyr
```

Gráficos de perfis:

```r
p1=ggplot(dados, aes(x=tempo, y=resp, color=sex)) + theme_bw() +
   geom_line(aes(group=id)) + theme(legend.position="top") +
   scale_x_continuous(breaks=unique(dados$tempo))
p2 = p1 + geom_smooth(method="loess",se=FALSE,size=2)
grid.arrange(p1,p2,ncol=2)
```




O modelo a ajustado é dado pela expressão a seguir:

$$E(Y_{ij}) = \beta_0 + \beta_1 \times sexo_i + \beta_2 \times tempo_j + \beta_3 \times tempo_j \times sexo_i.$$

## Estimador GEE

Considere os quatro ajustes GEE realizados.

```r
library(geepack)
gee2.ind<-geeglm(resp ~ sex*tempo, id=id, corstr="independence", data=dados) #Independente
gee2.exch<-geeglm(resp ~ sex*tempo, id=id, corstr="exchangeable", data=dados) #Simetria composta
gee2.ar1<-geeglm(resp ~ sex*tempo, id=id, corstr="ar1", data=dados) #AR(1)
gee2.unst<-geeglm(resp ~ sex*tempo, id=id, corstr="unstructured", data=dados) #Não estruturada
```

Resultados:

```r
# Independente
round(coef(summary(gee2.ind)),3)
```

```
##             Estimate Std.err    Wald Pr(>|W|)
## (Intercept)   17.373   0.725 573.869    0.000
## sexM          -1.032   1.378   0.561    0.454
## tempo          0.480   0.063  57.697    0.000
## sexM:tempo     0.305   0.117   6.803    0.009
```

```r
# Simetria composta
round(coef(summary(gee2.exch)),3)
```

```
##             Estimate Std.err    Wald Pr(>|W|)
## (Intercept)   17.373   0.725 573.869    0.000
## sexM          -1.032   1.378   0.561    0.454
## tempo          0.480   0.063  57.697    0.000
## sexM:tempo     0.305   0.117   6.803    0.009
```

```r
# AR(1)
round(coef(summary(gee2.ar1)),3)
```

```
##             Estimate Std.err    Wald Pr(>|W|)
## (Intercept)   17.312   0.792 477.573    0.000
## sexM          -0.659   1.526   0.186    0.666
## tempo          0.484   0.063  58.979    0.000
## sexM:tempo     0.283   0.124   5.216    0.022
```

```r
# Não estruturada
round(coef(summary(gee2.unst)),3)
```

```
##             Estimate Std.err    Wald Pr(>|W|)
## (Intercept)   17.397   0.724 576.702    0.000
## sexM          -1.074   1.376   0.609    0.435
## tempo          0.478   0.064  56.023    0.000
## sexM:tempo     0.310   0.117   6.997    0.008
```

Conclusões:

- As estimativas de erro padrão dos coeficientes são similares entre as diferentes estruturas, o que mostra a robustez do método GEE à má especificação da estrutura de dependência entre as medidas repetidas.
- O efeito de interação é significativo em todas as análises.
- Podemos concluir que meninos e meninas crescem em ritmos distintos.

## Modelo Misto

Vamos agora repetir a análise, mas considerando o modelo misto. Ajustaremos dois modelos, um com intercepto aleatório apenas, e outro com intercepto e inclinação aleatórios.

Modelo com intercepto aleatório:

$$E(Y_{ij}|b_i) = \beta_0 + b_{0i} + \beta_1 \times sexo_i + \beta_2 \times tempo_j + \beta_3 \times tempo_j \times sexo_i.$$

Modelo com intercepto e inclinação aleatórios:

$$E(Y_{ij}|b_i) = \beta_0 + b_{0i} + \beta_{1i} \times sexo_i + \beta_2 \times tempo_j + b_{1i} \times tempo_j + \beta_3 \times tempo_j \times sexo_i.$$

```
lme1 <-lme(resp~sex*tempo, random= ~1|id,data=dados)
lme2 <-lme(resp~sex*tempo, random= ~tempo|id,data=dados)
```

Para o primeiro modelo, obtemos:

```
summary(lme1)
```

```
## Linear mixed-effects model fit by REML
##  Data: dados
##        AIC      BIC    logLik
##   445.7572 461.6236 -216.8786
##
## Random effects:
##  Formula: ~1 | id
##         (Intercept) Residual
## StdDev:    1.816214 1.386382
##
## Fixed effects: resp ~ sex * tempo
##                 Value Std.Error DF   t-value p-value
## (Intercept) 17.372727 1.1835071 79 14.679023  0.0000
## sexM        -1.032102 1.5374208 25 -0.671321  0.5082
## tempo        0.479545 0.0934698 79  5.130483  0.0000
## sexM:tempo   0.304830 0.1214209 79  2.510520  0.0141
##  Correlation:
##            (Intr) sexM   tempo
## sexM       -0.770
## tempo      -0.869  0.669
## sexM:tempo  0.669 -0.869 -0.770
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -3.59804400 -0.45461690  0.01578365  0.50244658  3.68620792
##
## Number of Observations: 108
## Number of Groups: 27
```

Para o segundo modelo, obtemos:

```
summary(lme2)
```

```
## Linear mixed-effects model fit by REML
##  Data: dados
##        AIC      BIC    logLik
##   448.5817 469.7368 -216.2908
```

```
##
## Random effects:
##  Formula: ~tempo | id
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev    Corr
## (Intercept) 2.4055009 (Intr)
## tempo       0.1803455 -0.668
## Residual    1.3100396
##
## Fixed effects: resp ~ sex * tempo
##                 Value Std.Error DF   t-value p-value
## (Intercept) 17.372727 1.2283958 79 14.142614  0.0000
## sexM        -1.032102 1.5957329 25 -0.646789  0.5237
## tempo        0.479545 0.1037193 79  4.623492  0.0000
## sexM:tempo   0.304830 0.1347353 79  2.262432  0.0264
##  Correlation:
##            (Intr) sexM   tempo
## sexM       -0.770
## tempo      -0.880  0.678
## sexM:tempo  0.678 -0.880 -0.770
##
## Standardized Within-Group Residuals:
##          Min           Q1          Med           Q3          Max
## -3.168077732 -0.385939009  0.007104087  0.445154545  3.849463576
##
## Number of Observations: 108
## Number of Groups: 27
```

Os dois modelos permitem conclusões semelhantes com relação aos efeitos fixos. Os efeitos aleatórios podem ser acessados via:

```r
ranef(lme1)
```

```
##    (Intercept)
## 1  -1.11090166
## 2   0.30748171
## 3   0.96212019
## 4   1.94407791
## 5  -0.01983753
## 6  -1.32911449
## 7   0.30748171
## 8   0.63480095
## 9  -1.32911449
## 10 -3.62034916
## 11  3.25335486
## 12  2.42761769
## 13 -1.39110677
## 14 -0.62736188
## 15  1.44565997
## 16 -1.71842601
## 17  1.22744715
## 18 -1.06378753
## 19 -0.95468112
## 20  0.13638302
## 21  3.95510748
```

```
## 22 -1.17289394
## 23 -0.62736188
## 24 -0.62736188
## 25 -0.08182981
## 26  0.79102150
## 27 -1.71842601
```

```r
ranef(lme2)
```

```
##    (Intercept)        tempo
## 1  -0.64132024 -0.044754845
## 2  -0.66020223  0.090293750
## 3  -0.24892689  0.113565208
## 4   1.66111350  0.028212826
## 5   0.57096833 -0.054963721
## 6  -0.82630563 -0.048057940
## 7   0.05820188  0.023482879
## 8   1.41328613 -0.071778786
## 9  -0.53894398 -0.074782288
## 10 -2.98417340 -0.062697171
## 11  2.19630252  0.101480090
## 12  1.58201968  0.081009128
## 13 -1.15234167 -0.023562635
## 14 -0.43305242 -0.018682891
## 15  2.97663819 -0.140968498
## 16 -1.64534099 -0.008474015
## 17  1.35484459 -0.010649850
## 18 -0.94670401 -0.011926906
## 19  0.36707567 -0.123853840
## 20 -0.43216727  0.053007723
## 21  3.45164067  0.050682093
## 22  0.32577111 -0.140519109
## 23 -1.15145653  0.048127980
## 24 -3.88139215  0.302009290
## 25  0.67597475 -0.070554939
## 26 -0.30825358  0.103003530
## 27 -0.78325605 -0.088647061
```

A comparação entre os dois modelos é realizada a seguir. O teste usual da razão de verossimilhança retorna:

```r
anova(lme1,lme2)
```

```
##      Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## lme1     1  6 445.7572 461.6236 -216.8786
## lme2     2  8 448.5817 469.7368 -216.2908 1 vs 2 1.175588  0.5556
```

Como a hipótese testada está no limite do espaço paramétrico, um teste com melhores propriedades é dado por:

```r
chisq=2*(lme2$logLik-lme1$logLik);chisq
```

```
## [1] 1.175588
```

```r
0.5*pchisq(chisq, 1, lower.tail = FALSE) + 0.5*pchisq(chisq, 2, lower.tail = FALSE)
```

```
## [1] 0.4169038
```

A hipótese nula não é rejeitada: devemos ficar com o modelo que possui apenas intercepto aleatório.

Conclusões gerais:

- Modelos mistos, aqui considerando que cada indivíduo tenha um intercepto aleatório, permitem que os dados sejam modelados de forma flexível.
- Pelos valores-p obtidos podemos responder às questões iniciais do estudo dizendo que existe diferença estatisticamente significativas entre as medidas ao longo do tempo.
- Podemos dizer que os sexos diferem com relação à resposta, sendo que os meninos apresentam as maiores medidas.
- O efeito da interação indica que essa diferença entre os sexos não pode ser considerada marginalmente, mas que depende do tempo em que é avaliada.