Modelos Marginais: Estimadores GLS e GEE

José Luiz Padilha da Silva

16 de setembro de 2019

Exemplo 2: Dados de Crescimento

Potthoff & Roy (1964) apresentaram um conjunto de dados 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 é 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(nlme); library(geepack)
data(potthoffroy); head(potthoffroy)
```

```
##   id sex   d8  d10  d12  d14
## 1  1   F 21.0 20.0 21.5 23.0
## 2  2   F 21.0 21.5 24.0 25.5
## 3  3   F 20.5 24.0 24.5 26.0
## 4  4   F 23.5 24.5 25.0 26.5
## 5  5   F 21.5 23.0 22.5 23.5
## 6  6   F 20.0 21.0 21.0 22.5
```

A seguir um 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
```

Notamos que as meninas possuem menores valores médios que os meninos. As correlações marginais são dadas a seguir no geral e por sexo.

```
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
```

Os dados mostram forte correlação positiva.

```
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
```

Contudo, as meninas apresentam correlação entre as medidas repetidas consideravelmente maiores que os meninos. Além disso, as correlações para o grupo dos meninos é comparativamente mais variável enquanto para as meninas é mais homogênea.

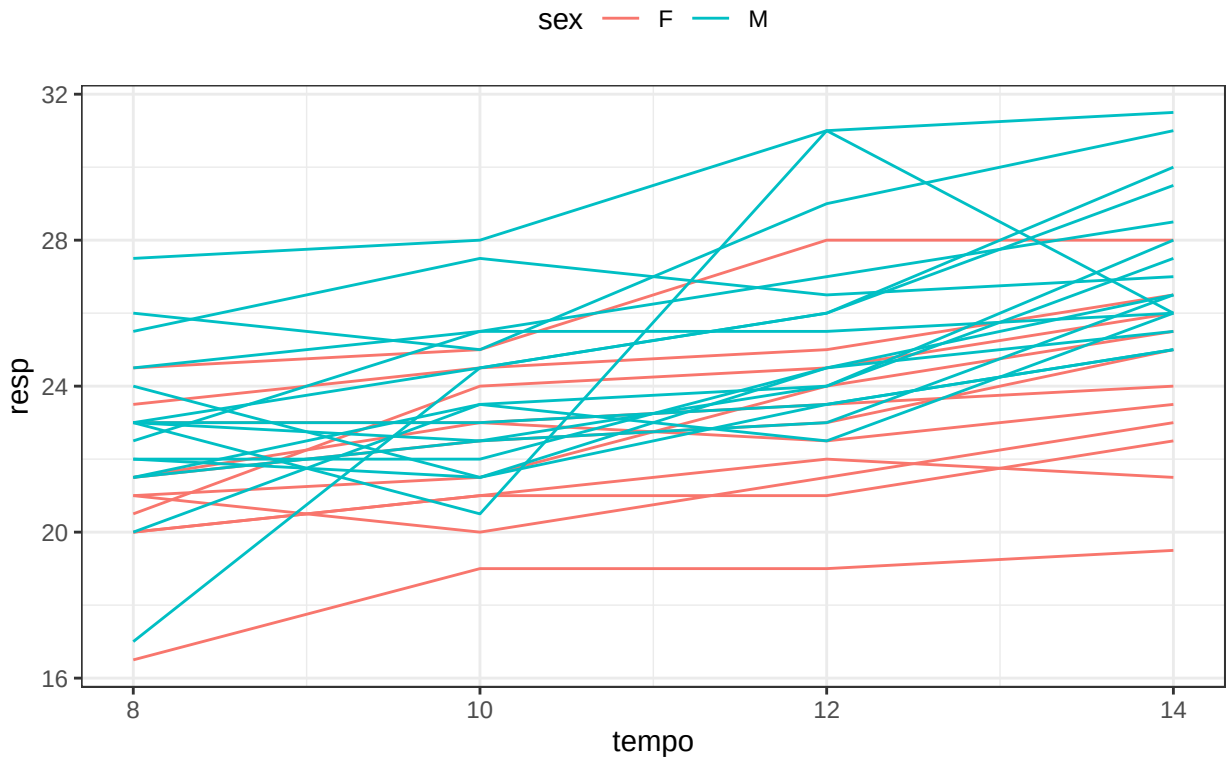
A seguir transformamos os dados para o formato longo.

```
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
head(dados, 8)
```

```
##   id sex tempo resp
## 1  1   F     8 21.0
## 2  1   F    10 20.0
## 3  1   F    12 21.5
## 4  1   F    14 23.0
## 5  2   F     8 21.0
## 6  2   F    10 21.5
## 7  2   F    12 24.0
## 8  2   F    14 25.5
```

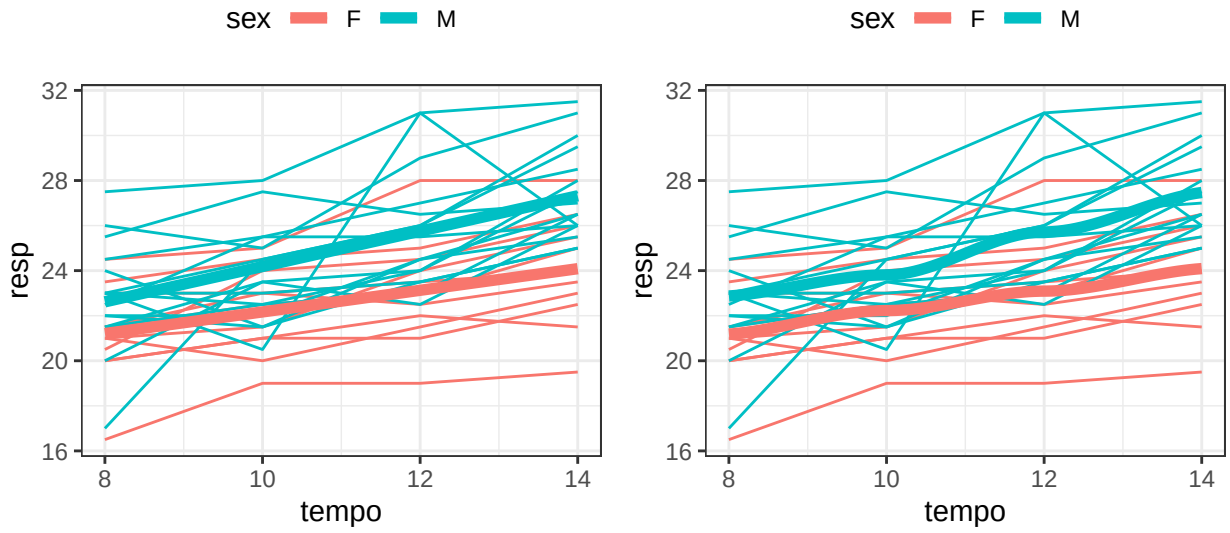
Na sequência o gráfico de perfis:

```
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))
p1
```



Uma linha de regressão linear ou suavizada pode ser adicionada ao gráfico fazendo

```
library(gridExtra)
p11 <- p1 + geom_smooth(method="lm",se=FALSE,size=2)
p12 <- p1 + geom_smooth(method="loess",se=FALSE,size=2)
grid.arrange(p11,p12,ncol=2)
```



Como vemos, o comportamento longitudinal é aproximadamente linear e um modelo com interação sexo e tempo parece ser adequado. O modelo a ser ajustado é dado por

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

Estimador GLS

Consideraremos novamente as estruturas de correlação do tipo *independente*, *simetria composta*, *AR(1)* e *não estruturada*. Para fins de análise as idades foram centradas em um valor comum, no caso a média de 11 anos.

```
dados$tempo=dados$tempo-11
gls2.ind<-gls(resp ~ sex*tempo, data=dados) #Independente
gls2.exch<-gls(resp ~ sex*tempo, correlation=corCompSymm(form=~1|id), data=dados) #Simetria composta
gls2.ar1<-gls(resp ~ sex*tempo, correlation=corAR1(form=~1|id), data=dados) #AR(1)
gls2.unst<-gls(resp ~ sex*tempo, correlation=corSymm(form=~1|id), data=dados) #Não estruturada
```

Os resultados dos ajustes são mostrados a seguir:

```
# Independente
round(coef(summary(gls2.ind)),3)
```

```
##              Value Std.Error t-value p-value
## (Intercept) 22.648     0.340  66.562   0.000
## sexM         2.321     0.442   5.251   0.000
## tempo        0.480     0.152   3.152   0.002
## sexM:tempo   0.305     0.198   1.542   0.126
```

```
# Simetria composta
round(coef(summary(gls2.exch)),3)
```

```
##              Value Std.Error t-value p-value
## (Intercept) 22.648     0.586  38.639   0.000
## sexM         2.321     0.761   3.048   0.003
## tempo        0.480     0.093   5.130   0.000
## sexM:tempo   0.305     0.121   2.511   0.014
```

```
# AR(1)
round(coef(summary(gls2.ar1)),3)
```

```
##              Value Std.Error t-value p-value
## (Intercept) 22.643     0.529  42.797   0.000
## sexM         2.418     0.687   3.519   0.001
## tempo        0.484     0.141   3.430   0.001
## sexM:tempo   0.285     0.183   1.558   0.122
```

```
# Não estruturada
round(coef(summary(gls2.unst)),3)
```

```
##              Value Std.Error t-value p-value
## (Intercept) 22.645     0.585  38.697   0.000
## sexM         2.355     0.760   3.098   0.003
## tempo        0.476     0.099   4.791   0.000
## sexM:tempo   0.348     0.129   2.696   0.008
```

Note como as estimativas das estruturas *independente* e *simetria composta* são similares. Interessante notar como o valor p é bastante pequeno para *simetria composta* e *não estruturada* e alto para as demais estruturas. Assim, diferentes escolhas para a correlação levam a diferentes inferências quanto ao efeito de interação.

```
gls2.ind$modelStruct$corStruct
```

```
## NULL
```

```
gls2.exch$modelStruct$corStruct
```

```
## Correlation structure of class corCompSymm representing
```

```
##       Rho
## 0.6318381
```

```
gls2.ar1$modelStruct$corStruct
```

```
## Correlation structure of class corAR1 representing
##       Phi
## 0.6244888
```

```
gls2.unst$modelStruct$corStruct
```

```
## Correlation structure of class corSymm representing
##  Correlation:
##   1     2     3
## 2 0.575
## 3 0.638 0.574
## 4 0.515 0.749 0.721
```

Das correlações marginais vimos que as estruturas independente e autorregressiva não são adequadas a esses dados. Vamos comparar as diferentes estruturas via medidas de informação e testes formais:

```
anova(gls2.unst, gls2.exch)
```

```
##           Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## gls2.unst     1 11 448.1706 477.2589 -213.0853
## gls2.exch     2  6 445.7572 461.6236 -216.8786 1 vs 2 7.586616  0.1805
```

```
anova(gls2.unst, gls2.ar1)
```

```
##           Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## gls2.unst     1 11 448.1706 477.2589 -213.0853
## gls2.ar1      2  6 456.5874 472.4538 -222.2937 1 vs 2 18.41681  0.0025
```

```
anova(gls2.exch, gls2.ar1)
```

```
##           Model df      AIC      BIC    logLik
## gls2.exch     1  6 445.7572 461.6236 -216.8786
## gls2.ar1      2  6 456.5874 472.4538 -222.2937
```

A estrutura escolhida por ambos é a simetria composta.

Estimador GEE

Ajustamos agora as mesmas estruturas de correlação e estimamos os modelos pelo método GEE.

```
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
```

As estimativas são dados por:

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

```
##             Estimate Std.err     Wald Pr(>|W|)
## (Intercept)   22.648   0.605 1400.761    0.000
## sexM           2.321   0.750    9.583    0.002
## 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)   22.648   0.605 1400.761    0.000
## sexM           2.321   0.750    9.583    0.002
## 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)   22.641   0.618 1341.792    0.000
## sexM           2.452   0.758   10.458    0.001
## 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)   22.656   0.599 1431.397    0.000
## sexM           2.337   0.736   10.077    0.002
## tempo          0.478   0.064   56.023    0.000
## sexM:tempo     0.310   0.117    6.997    0.008
```

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.

```r
round(summary(gee2.ind)$corr,3)
```

```
## [1] Estimate Std.err
## <0 rows> (or 0-length row.names)
```

```r
round(summary(gee2.exch)$corr,3)
```

```
##       Estimate Std.err
## alpha    0.618   0.131
```

```r
round(summary(gee2.ar1)$corr,3)
```

```
##       Estimate Std.err
## alpha    0.759   0.096
```

```r
round(summary(gee2.unst)$corr,3)
```

```
##           Estimate Std.err
## alpha.1:2    0.501   0.133
## alpha.1:3    0.736   0.138
## alpha.1:4    0.515   0.192
## alpha.2:3    0.555   0.226
## alpha.2:4    0.621   0.090
## alpha.3:4    0.779   0.163
```

Agora o efeito de interação é significativo em todas as análises. Podemos concluir que meninos e meninas crescem em ritmos distintos.

Comentários sobre a coincidência entre as estimativas de independência e simetria composta

Como vimos, as análises independente e simetria composta retornam as mesmas estimativas e erro padrão robusto porque os dados são balanceados no tempo. Vamos criar alguns "dados ausentes" e ver o que acontece. Deletamos as últimas duas observações dos primeiros cinco indivíduos para criar desbalanceamento.

```
dados2 <- dados[-c(3,4,7,8,11,12,15,16,19,20),]
head(dados2)
```

```
##    id sex tempo resp
## 1   1   F    -3 21.0
## 2   1   F    -1 20.0
## 5   2   F    -3 21.0
## 6   2   F    -1 21.5
## 9   3   F    -3 20.5
## 10  3   F    -1 24.0
```

```
gee3.ind <- geeglm(resp ~ sex*tempo, id=id, corstr="independence", data=dados2) #Independente
gee3.exch <- geeglm(resp ~ sex*tempo, id=id, corstr="exchangeable", data=dados2) #Simetria composta
round(coef(summary(gee3.ind)),3)
```

```
##             Estimate Std.err    Wald Pr(>|W|)
## (Intercept)   22.408   0.779 827.602    0.000
## sexM           2.561   0.896   8.169    0.004
## tempo          0.369   0.127   8.469    0.004
## sexM:tempo     0.416   0.160   6.723    0.010
```

```
round(coef(summary(gee3.exch)),3)
```

```
##             Estimate Std.err     Wald Pr(>|W|)
## (Intercept)   22.518   0.656 1179.456    0.000
## sexM           2.451   0.791    9.597    0.002
## tempo          0.415   0.073   32.167    0.000
## sexM:tempo     0.370   0.123    9.102    0.003
```

Por conta do desbalanceamento os resultados são diferentes para as duas estruturas.