CE075 - Análise de Dados Longitudinais

Silva, J.L.P.

02 de setembro, 2019

Modelos para a Estrutura de Covariância

Modelos para a Estrutura de Covariância

Temos as possibilidades:

1. Não Estruturado: somente é adequada para desenhos balanceados com poucos tempos. No caso heterocedástico, para n medidas repetidas por unidade, há $n(n+1)/2$ parâmetros.

2. Estruturando a Covariância: simetria composta, AR(1), etc. Usualmente adequada para desenhos balanceados com poucos tempos.

3. Modelos de Efeitos Aleatórios: a estrutura de covariância é função dos efeitos aleatórios.

Modelos para a Estrutura de Covariância

Nos casos extremos:

- **Nenhuma estrutura imposta**: pode haver muitos parâmetros para uma quantidade limitada de dados. Isso afeta a precisão com que β é estimado.
- **Estrutura imposta**: é possível melhorar a precisão com que β é estimado! Contudo, se muita restrição é imposta, há um risco potencial de má especificação que pode resultar em inferência enganosas sobre β.

Temos o clássico problema de *trade-off* entre viés e precisão. Deve-se buscar equilíbrio entre essas duas forças.

1. Simetria Composta

Possui apenas um parâmetro de correlação, independente do número de medidas: $Corr(Y_{ij}, Y_{ik}) = \rho, \ \forall j, k$.

$$V_0 = [(1-\rho)I_n + \rho 1_n 1_n'],$$

em que I_n é a matriz identidade e 1_n é um vetor de 1's, ambos de dimensão n.

Assim:

$$V_0 = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}_{n \times n}$$

1. Simetria Composta

- Tem justificativa teórica quando a média depende de uma combinação de parâmetros populacionais $\boldsymbol{\beta}$, e um único efeito aleatório referente ao indivíduo.
- É parcimonioso: são dois parâmetros (uma variância e uma correlação).
- Restrição: $\rho \geq 0$. A suposição de que ρ é o mesmo pode não ser realístico pois se espera um decaimento com o tempo.

Justificativa

Considere o modelo de efeitos aleatórios, em que intercepto é o único termo com variação aleatória

$$Y_{ij} = X_{ij}'\beta + U_i + \varepsilon_{ij}.$$

A diferença entre os indivíduos está explicada pelo intercepto aleatório:

$$U_i \sim N(0, \sigma_u^2)$$
$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2),$$

em que U_i e ε_{ij} são independentes.

Justificativa

Temos:

$$\begin{aligned} Cov(Y_{ij}, Y_{il}) =& Cov(\boldsymbol{X}'_{ij}\beta + U_i + \varepsilon_{ij}, \boldsymbol{X}'_{il}\beta + U_i + \varepsilon_{il}) \\ =& Cov(U_i + \varepsilon_{ij}, U_i + \varepsilon_{il}) \\ =& Cov(U_i, U_i) + Cov(\varepsilon_{ij}, \varepsilon_{il}) \\ =& \sigma_u^2 + \sigma_\varepsilon^2 I(j = l). \end{aligned}$$

Logo,

$$Var(Y_{ij}) = \sigma_u^2 + \sigma_\varepsilon^2,$$

o que implica:

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2}.$$

2. Correlação AR(1)

Temos $Corr(Y_{ij}, Y_{i,j+k}) = \rho^k, \ \forall j, k.$

$$V_0 = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}_{n \times n}$$

- É muito parcimonioso (são dois parâmetros).
- Apropriado para intervalos de tempo igualmente espaçados.
- Correlações decaem para zero, mas em muitos estudos o decaimento ocorre em ritmo menor que o previsto por tal estrutura.

Justificativa

Quando os erros surgem de um processo autorregressivo:

$$\varepsilon_{ij} = \rho\varepsilon_{ij-1} + \omega_{ij}, \quad \omega_{ij} \sim N(0, \sigma^2(1-\rho^2)),$$

em que ε_{ij} e ω_{ij} são independentes.

Então

$$\begin{aligned} Var(\varepsilon_{ij}) =& \rho^2\sigma^2 + \sigma^2(1-\rho^2) = \sigma^2 \\ Cov(\varepsilon_{ij}, \varepsilon_{ij-1}) =& Cov(\rho\varepsilon_{ij-1} + \omega_{ij}, \varepsilon_{ij-1}) = \rho\sigma^2 \end{aligned}$$

e para defasagens maiores que 1,

$$Cov(\varepsilon_{ij}, \varepsilon_{ij-k}) = \rho^k\sigma^2.$$

3. Correlação Exponencial

Temos $Corr(Y_{ij}, Y_{ik}) = \rho^{|t_{ij}-t_{ik}|}, \ \forall j, k.$

$$V_0 = \begin{bmatrix} 1 & \rho^{|t_{i1}-t_{i2}|} & \rho^{|t_{i1}-t_{i3}|} & \dots & \rho^{|t_{i1}-t_{in}|} \\ \rho^{|t_{i2}-t_{i1}|} & 1 & \rho^{|t_{i2}-t_{i3}|} & \dots & \rho^{|t_{i2}-t_{in}|} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{|t_{in}-t_{i1}|} & \rho^{|t_{in}-t_{i2}|} & \rho^{|t_{in}-t_{i3}|} & \dots & 1 \end{bmatrix}_{n\times n}$$

- Os tempos não precisam ser igualmente espaçados.
- Assume que a correlação é um se as medidas são tomadas repetidamente na mesma ocasião: corresponde à situação que as respostas são medidas sem erro.
- As correlações decaem rapidamente para zero.

Decaimento Exponencial

O decaimento para zero ocorre de maneira bem rápida:

	Distância				
ρ	1	2	3	4	5
0,9	0,90	0,81	0,73	0,66	0,59
0,7	0,70	0,49	0,34	0,24	0,17
0,5	0,50	0,25	0,13	0,06	0,03
0,3	0,30	0,09	0,03	0,01	0,00

Tal comportamento é raramente observado em estudos longitudinais.

4. Toeplitz

Assume que qualquer par de respostas igualmente espaçadas no tempo tem a mesmo correlação. $Corr(Y_{ij}, Y_{i,j+k}) = \rho_k, \ \forall j, k.$

$$V_0 = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1 \end{bmatrix}_{n \times n}$$

- É uma extensão da estrutura AR(1), com $n-1$ parâmetros de correlação.
- Como assume que a correlação entre ocasiões adjacentes no tempo é constante, ρ_1, é apropriada para intervalos de tempo igualmente espaçados.

5. Banded

Características:

- A correlação é zero além de um período especificado de tempo.
- Pode ser aplicado a qualquer estrutura vista anteriormente.

Por exemplo, um padrão de correlação *banded* de tamanho 2 assume $Corr(Y_{ij}, Y_{i,j+k}) = 0, \ \forall k \geq 2$. Neste caso, para uma estrutura Toeplitz, temos:

$$V_0 = \begin{bmatrix} 1 & \rho_1 & 0 & \cdots & 0 \\ \rho_1 & 1 & \rho_1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}_{n \times n}$$

6. Modelos Híbridos

Considere a combinação:

$$Cov(Y_i) = \sigma_1^2 V_1 + \sigma_2^2 V_2.$$

Seja, por exemplo, V_1 simetria composta e V_2 autorregressivo (exponencial).

Para este modelo híbrido, tempos:

$$\begin{aligned} Var(Y_{ij}) =& \sigma_1^2 + \sigma_2^2 \\ Cov(Y_{ij}, Y_{ik}) =& \rho_1 \sigma_1^2 + \rho_2^{|t_{ij} - t_{ik}|} \sigma_2^2 \\ Corr(Y_{ij}, Y_{ik}) =& \frac{\rho_1 \sigma_1^2 + \rho_2^{|t_{ij} - t_{ik}|} \sigma_2^2}{\sigma_1^2 + \sigma_2^2} \end{aligned}$$

6. Modelos Híbridos

- A correlação entre réplicas no mesmo indivíduo na mesma ocasião é

$$\frac{\rho_1\sigma_1^2+\sigma_2^2}{\sigma_1^2+\sigma_2^2}<1,\ \text{quando}\ \rho_1<1$$

- À medida que a separação no tempo aumenta, a correlação não decai para zero, mas tem um mínimo em

$$\frac{\rho_1\sigma_1^2}{\sigma_1^2+\sigma_2^2}>0,\ \text{quando}\ \rho_1>0.$$

6. Modelos Híbridos

- Simetria composta é também um modelo de efeitos aleatórios, assim

$$\sigma_1^2 = \sigma_u^2 + \sigma_\epsilon^2 \ \text{ e } \ \rho_1 = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2}.$$

Ou seja, podemos pensar na variância total como a soma da variância autorregressiva, σ_2^2, a variabilidade entre indivíduos σ_u^2, e a variabilidade do erro de medida, σ_ε^2.

Estimador de Mínimos Quadrados Generalizados

Estimador de Mínimos Quadrados Generalizados

Considere o modelo linear:

$$Y_i = \boldsymbol{X}_i \beta + \varepsilon_i$$

tal que $E(\varepsilon_i) = \mathbf{0}$, $Var(\varepsilon_i) = \boldsymbol{V} = \sigma^2 \boldsymbol{V}_0$, com $\boldsymbol{V}_0$ conhecida.

O *Estimador de Mínimos Quadrados Generalizados* (GLS) minimiza

$$(Y - \boldsymbol{X}\beta)' \boldsymbol{V}^{-1} (Y - \boldsymbol{X}\beta),$$

para o qual obtemos

$$\hat{\beta}_{GLS} = (\boldsymbol{X}' \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}' \boldsymbol{V}^{-1} \boldsymbol{Y}.$$

Estimador de Mínimos Quadrados Generalizados

O estimador GLS é não-viciado, consistente, eficiente e assintoticamente normal com

$$E(\hat{\beta}) = \beta \quad \text{e} \quad Var(\hat{\beta}) = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}.$$

As propriedades acima valem em grandes amostras mesmo quando $\boldsymbol{Y}_i$ não tem uma distribuição normal multivariada.

O estimador GLS é equivalente a aplicar o estimador de mínimos quadrados ordinários em uma versão transformada dos dados.

Estimador de Mínimos Quadrados Generalizados

Toda matriz positiva definida pode ser escrita como

$$\boldsymbol{V} = \boldsymbol{K}\boldsymbol{K}', \quad \boldsymbol{K} \text{ é não singular}$$

Redefina o modelo como $\boldsymbol{Z} = \boldsymbol{B}\beta + \boldsymbol{\eta}$, em que:

$$\begin{aligned} \boldsymbol{Z} &= \boldsymbol{K}^{-1}\boldsymbol{Y} \\ \boldsymbol{B} &= \boldsymbol{K}^{-1}\boldsymbol{X} \\ \eta &= \boldsymbol{K}^{-1}\varepsilon \end{aligned}$$

Estimador de Mínimos Quadrados Generalizados

Assim,

$$\begin{aligned} Var(\boldsymbol{\eta}) &= Var(\boldsymbol{K}^{-1}\boldsymbol{\varepsilon}) \\ &= \boldsymbol{K}^{-1} Var(\boldsymbol{\varepsilon}) \boldsymbol{K}^{-1'} \\ &= \sigma^2 \boldsymbol{K}^{-1} \boldsymbol{K} \boldsymbol{K}' (\boldsymbol{K}')^{-1} \\ &= \sigma^2 I_{Nn} \end{aligned}$$

Desta forma, retornamos à condição de Mínimos Quadrados Ordinários.

Estimador de Mínimos Quadrados Generalizados

Considere as equações normais:

$$
\begin{aligned}
\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} &= (\boldsymbol{Z} - \boldsymbol{B}\beta)'(\boldsymbol{Z} - \boldsymbol{B}\beta) \\
&= (\boldsymbol{Y} - \boldsymbol{X}\beta)'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\beta) \\
&= \sum_{i=1}^{N} (\boldsymbol{Y}_i - \boldsymbol{X}_i\beta)'\boldsymbol{V}_0^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta).
\end{aligned}
$$

Resolver o sistema de equações:

$$
\begin{aligned}
\frac{\partial \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{\partial \beta} &= 2\boldsymbol{X}'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\beta) \\
&= \sum_{i=1}^{N} \boldsymbol{X}_i'\boldsymbol{V}_0^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\beta}) = \boldsymbol{0}.
\end{aligned}
$$

Equações Normais

Então

$$\begin{aligned}\hat{\beta} &= (\boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{Z} \\ &= (\boldsymbol{X}'\boldsymbol{K}^{-1'}\boldsymbol{K}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{K}^{-1'}\boldsymbol{K}^{-1}\boldsymbol{Y} \\ &= (\boldsymbol{X}'\boldsymbol{K}^{-1'}\boldsymbol{K}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{K}\boldsymbol{K}')^{-1}\boldsymbol{Y} \\ &= (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y}\end{aligned}$$

e

$$Var(\hat{\beta}) = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}.$$

Exemplo

Exemplo: Dados de Crescimento

Retornemos aos dados de crescimento de Potthoff e Roy (1964):

- Foram avaliadas mudanças nas medidas ortodônticas ao longo do tempo de 11 meninas e 16 meninos.
- Resposta: distância do centro da pituitária à fissura do maxilar.
- As medições ocorreram aos 8, 10, 12 e 14 anos.
- Como objetivo citamos: (i) como essa distância cresce com a idade, (ii) testar se há diferença entre os valores para os meninos e meninas e (iii) se existe interação entre essas variáveis.

Exemplo: Dados de Crescimento

```r
library(mice); library(reshape); library(plyr)
library(nlme); library(ggplot2)
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
```

Exemplo: Dados de Crescimento

```
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
```

Exemplo: Dados de Crescimento

```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
```

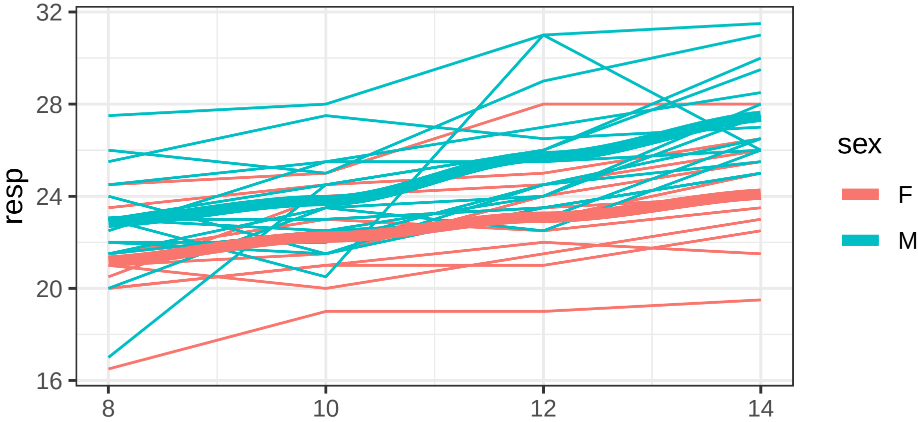
Exemplo: Dados de Crescimento

```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)
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
```

Exemplo: Dados de Crescimento

```
ggplot(dados, aes(x=tempo,y=resp,color=sex)) + geom_line(aes(group=id)) +
geom_smooth(method="loess",se=FALSE,size=2) +  labs(x="") +
scale_x_continuous(breaks=unique(dados$tempo)) + theme_bw()
```



Exemplo: Dados de Crescimento

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

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

Tal modelo permite estimar taxas de crescimento diferente para meninos e meninas.

Exemplo: Dados de Crescimento

Como ilustração, considere 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)#Sim. comp.
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 estrut.
```

Exemplo: Dados de Crescimento

```r
# 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
```

```r
# 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
```

Exemplo: Dados de Crescimento

```r
# 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
```

```r
# 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
```

Modelo Marginal

A validade do estimador GLS fica comprometida quando $\boldsymbol{V}$ não for conhecida.

Em situações reais o que devemos fazer?

1. Resposta Normal: Utilizar o Método de Máxima Verossimilhança (usual ou restrito) para estimar β e também os componentes de variância. Ou seja, os parâmetros da média e também da estrutura escolhida de covariância.
2. Sem especificar distribuição para a resposta: Investigar qual é o impacto ao utilizarmos $\boldsymbol{W}$ ao invés de $\boldsymbol{V}$. Ideia do GEE. ($\boldsymbol{W}$ a princípio pode ser aquela mais adequada para modelar a estrutura de covariância dos dados.)