MAT02035 - Modelos para dados correlacionados

Modelando a covariância

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, 2019

Modelando a covariância

- Os dados longitudinais apresentam **dois aspectos** dos dados que **requerem modelagem**: a **resposta média** ao longo do tempo e a **covariância**.
- Embora esses dois aspectos dos dados possam ser modelados separadamente, eles estão **inter-relacionados**.
- As escolhas de modelos para resposta média e covariância são **interdependentes**.
- Um modelo para a covariância deve ser escolhido com base em algum modelo assumido para a resposta média.
- **Lembre-se:** a covariância entre qualquer par de resíduos, digamos $[Y_{ij} - \mu_{ij}(\beta)]$ e $[Y_{ik} - \mu_{ik}(\beta)]$, depende do modelo para a média, ou seja, depende de β.

Modelando a covariância

Três abordagens gerais podem ser distinguidas:

(1) padrão de covariância “não estruturado” ou arbitrário
(2) modelos de padrões de covariância
(3) estrutura de covariância de efeitos aleatórios

Covariância não estruturada

- Apropriado quando o delineamento é "balanceado" e o número de ocasiões de medições é relativamente pequena.
- Nenhuma estrutura explícita é assumida além da homogeneidade de covariância entre diferentes indivíduos, $\mathrm{Cov}\,(Y_i) = \Sigma_i = \Sigma$.
- **Principal vantagem:** nenhuma suposição sobre os padrões de variância e covariâncias.

Covariância não estruturada

- Com n ocasiões de medição, a matriz de covariância "não estruturada" possui $n(n+1)/2$ parâmetros:
  - as n variâncias e $n \times (n-1)/2$ covariâncias (ou correlações) duas-a-duas,

$$\mathrm{Cov}\,(Y_i) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix}.$$

Covariância não estruturada

Potenciais desvantagens:

- O número de parâmetros de covariância cresce rapidamente com o número de ocasiões de medição:
  - $n = 3$, o número de parâmetros de covariância é 6
  - $n = 5$, o número de parâmetros de covariância é 15
  - $n = 10$, o número de parâmetros de covariância é 55
- Quando o número de parâmetros de covariância é grande, em relação ao tamanho da amostra, é provável que a estimativa seja muito instável.
- O uso de uma covariância não estruturada é atraente apenas quando N é grande em relação a $n \times (n + 1)/2$.
- A covariância não estruturada é problemática quando há medições irregulares no tempo.

Modelos de padrão de covariância

- Ao tentar impor alguma estrutura à covariância, um **equilíbrio** sutil precisa ser alcançado.
- Com **pouca estrutura**, pode haver **muitos parâmetros** a serem estimados com quantidade limitada de dados.
- Com **muita estrutura**, risco potencial de **erros de especificação do modelo** e inferências enganosas a respeito de β.
- Clássico **tradeoff** entre **viés** e **variância**.
- Os modelos de padrões de covariância têm como base modelos de correlação serial originalmente desenvolvidos para dados de **séries temporais**.

(Relação entre a covariância e a correlação)

- Lembre que $\rho_{jk} = \text{Corr}(Y_{ij}, Y_{ik})$ pode ser expresso como

$$\begin{aligned} \rho_{jk} &= \frac{\text{Cov}(Y_{ij}, Y_{ik})}{\sqrt{\text{Var}(Y_{ij})}\sqrt{\text{Var}(Y_{ik})}} \\ &= \frac{\sigma_{jk}}{\sigma_j \sigma_k} \\ &\Rightarrow \sigma_{jk} = \rho_{jk}\sigma_j\sigma_k. \end{aligned}$$

- Se assumirmos que $\text{Var}(Y_{i1}) = \text{Var}(Y_{i2}) = \ldots \text{Var}(Y_{in}) = \sigma^2$, então $\sigma_j = \sigma_k$ e $\sigma_j\sigma_k = \sigma^2$, e assim $\sigma_{jk} = \rho_{jk}\sigma^2$.

Simetria composta

- Assume que a variância é constante entre as ocasiões, digamos σ^2, e $\text{Corr}(Y_{ij}, Y_{ik}) = \rho$ para todo j e k.

$$\text{Cov}(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots & \\ \rho & \rho & \rho & \cdots & 1 \end{pmatrix}.$$

- **Parcimônia:** dois parâmetros, independentemente do número de ocasiões de medição.
- Suposições fortes sobre variância e correlação geralmente não são válidas com dados longitudinais.

Toeplitz

- Assume que a variância é constante entre as ocasiões, digamos σ^2, e $\text{Corr}\left(Y_{ij}, Y_{i,j+k}\right) = \rho_k$ para todo j e k.

$$\text{Cov}\left(Y_i\right) = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{n-3} \\ \vdots & \vdots & \ddots & \vdots & \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1 \end{pmatrix}.$$

- Assume que a correlação entre as respostas em ocasiões adjacentes de medição é constante, ρ_1.

Toeplitz

- Toeplitz é apropriado apenas quando as medições são feitas em intervalos de tempo iguais (ou aproximadamente iguais).
- A covariância Toeplitz possui n parâmetros (1 parâmetro de variância e $n-1$ parâmetros de correlação).
- Um caso especial da covariância Toeplitz é a covariância **autoregressiva** (de primeira ordem).

Autoregressiva

- Assume que a variância é constante entre as ocasiões, digamos σ^2, e $\text{Corr}\,(Y_{ij}, Y_{i,j+k}) = \rho^k$ para todo j e k, e $\rho \geq 0$.

$$\text{Cov}\,(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \ddots & \vdots & \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{pmatrix}.$$

- **Parcimônia:** apenas 2 parâmetros, independentemente do número de ocasiões de medição.
- Somente apropriado quando as medições são feitas em intervalos de tempo iguais (ou aproximadamente iguais).

Comentário

- Simetria composta, Toeplitz e covariância autoregressiva assumem que as variâncias são constantes ao longo do tempo.
- Essa suposição pode ser relaxada considerando-se versões desses modelos com variâncias heterogêneas, $\text{Var}\left(Y_{ij}\right) = \sigma_j^2$.
- Um padrão heterogêneo de covariância autoregressiva:

$$\text{Cov}\left(Y_i\right) = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho^2\sigma_1\sigma_3 & \cdots & \rho^{n-1}\sigma_1\sigma_n \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \rho\sigma_2\sigma_3 & \cdots & \rho^{n-2}\sigma_2\sigma_n \\ \rho^2\sigma_1\sigma_3 & \rho\sigma_2\sigma_3 & \sigma_3^2 & \cdots & \rho^{n-3}\sigma_3\sigma_n \\ \vdots & \vdots & \ddots & \vdots & \\ \rho^{n-1}\sigma_1\sigma_n & \rho^{n-2}\sigma_2\sigma_n & \rho^{n-3}\sigma_3\sigma_n & \cdots & \sigma_n^2 \end{pmatrix}.$$

e possui $n+1$ parâmetros (n parâmetros de variância e 1 parâmetro de correlação).

Banded (em faixas)

- Assume que a correlação é zero além de algum intervalo especificado.
- Por exemplo, um padrão de covariância em faixas com tamanho de banda 3 pressupõe que $\mathrm{Corr}\left(Y_{ij}, Y_{i,j+k}\right) = 0$ para $k \geq 3$.
- É possível aplicar um padrão em faixas a qualquer um dos modelos de padrões de covariância considerados até o momento.

Banded (em faixas)

- Um padrão de covariância de Toeplitz em faixas com um tamanho de banda 2 é dado por,

$$\mathrm{Cov}\,(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho_1 & 0 & \cdots & 0 \\ \rho_1 & 1 & \rho_1 & \cdots & 0 \\ 0 & \rho_1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

em que $\rho_2 = \rho_3 = \ldots = \rho_{n-1} = 0$.

- As faixas fazem suposições muito fortes sobre a rapidez com que a correlação cai para zero com o aumento da separação de tempo.

Exponencial

- Quando as ocasiões de medição não são igualmente espaçadas ao longo do tempo, o modelo autorregressivo pode ser generalizado da seguinte maneira.
- Deixe $\{t_{i1}, \ldots, t_{in}\}$ denotam os tempos de observação para o *i*-ésimo indivíduo e assuma que a variância é constante em todas as ocasiões de medição, digamos σ^2, e

$$\text{Corr}\,(Y_{ij}, Y_{ik}) = \rho^{|t_{ij} - t_{ik}|}, \text{ para } \rho \geq 0.$$

- A correlação entre qualquer par de medidas repetidas diminui exponencialmente com as separações de tempo entre elas.

Exponencial

- Referida como covariância "exponencial" porque pode ser reexpressa como

$$\begin{aligned} \mathrm{Cov}\,(Y_{ij}, Y_{ik}) &= \sigma^2 \rho^{|t_{ij} - t_{ik}|} \\ &= \sigma^2 \exp(-\theta |t_{ij} - t_{ik}|), \end{aligned}$$

em que $\theta = -\log(\rho)$ ou $\rho = \exp(-\theta)$ para $\theta \geq 0$.

- O modelo de covariância exponencial é invariante sob transformação linear da escala de tempo.
  - Se substituirmos t_{ij} por $(a + bt_{ij})$ (por exemplo, se substituirmos o tempo medido em "semanas" pelo tempo medido em "dias"), a mesma forma para a matriz de covariância se mantém.

Escolha entre modelos de padrões de covariância

- A escolha de modelos para covariância e média são interdependentes.
- A escolha do modelo de covariância deve ser baseada em um modelo **"maximal"** para a média que minimiza qualquer possível erro de especificação.
- Com delineamentos balanceados e um número muito pequeno de covariáveis discretas, escolha "modelo saturado" para a resposta média.
- Modelo saturado: inclui os efeitos principais do tempo (considerado um fator dentro do indivíduo) e todos os outros efeitos principais, além de suas interações duas-a-duas e de ordem superiores.

Escolha entre modelos de padrões de covariância

- O modelo maximal deve ser, em certo sentido, o modelo mais elaborado para a resposta média que consideraríamos do ponto de vista do indivíduo.
- Uma vez escolhido o modelo maximal, a variação residual e a covariância podem ser usadas para selecionar o modelo apropriado para covariância.
- Para modelos de padrões de covariância aninhados, pode ser construída uma estatística de teste de razão de verossimilhanças que compara os modelos "completo" e "reduzido".

Escolha entre modelos de padrões de covariância

- Lembre-se: dois modelos são aninhados quando o modelo "reduzido" é um caso especial do modelo "completo".
- Por exemplo, o modelo de simetria composta é aninhado dentro do modelo Toeplitz, desde que se o primeiro o último vale então o anterior necessariamente deve valer, com $\rho_1 = \rho_2 = \ldots = \rho_{n-1}$.
- O teste da razão de verossimilhanças é obtido tomando-se o dobro da diferença das respectivas log-verossimilhanças REML maximizadas,

$$G^2 = 2(\hat{l}_{comp} - \hat{l}_{redu}),$$

e comparando a estatística com uma distribuição qui-quadrado com graus de liberdade igual à diferença entre o número de parâmetros de covariância nos modelos completo e reduzido.

Escolha entre modelos de padrões de covariância

- Para comparar modelos não aninhados, uma abordagem alternativa é o Critério de Informação de Akaike (Akaike Information Criterion - AIC).
- De acordo com a AIC, dado um conjunto de modelos concorrentes para a covariância, deve-se selecionar o modelo que minimiza

$$
\begin{aligned}
AIC &= -2(\text{log-vero. maximizada}) + 2(\text{número de parâmetros}) \\
&= -2(\hat{l} - c),
\end{aligned}
$$

em que $\hat{l}$ é a log-verossimilhança REML maximizada e c é o número de parâmetros de covariância.

Exemplo: Teste de Terapia por Exercício

- Os indivíduos foram designados para um dos dois programas de levantamento de peso para aumentar a força muscular.
- Tratamento 1: o número de repetições dos exercícios foi aumentado à medida que os indivíduos se tornaram mais fortes.
- No tratamento 2, o número de repetições foi mantido constante, mas a quantidade de peso foi aumentada à medida que os indivíduos se tornaram mais fortes.
- As medidas de força corporal foram realizadas na linha de base e nos dias 2, 4, 6, 8, 10 e 12.
- Vamos nos concentrar apenas nas medidas de força obtidas na linha de base (ou no dia 0) e nos dias 4, 6, 8 e 12.

Exemplo: Teste de Terapia por Exercício

- Antes de considerar modelos para a covariância, é necessário escolher um modelo maximal para a resposta média.
- Escolhemos o modelo maximal como o modelo saturado para a média.
- Primeiro, consideramos uma matriz de covariância não estruturada.
- Observe que a variância parece ser maior no final do estudo, quando comparada à variância na linha de base.
- As correlações diminuem à medida que a separação do tempo entre as medidas repetidas aumenta.

Carregando os dados

```r
# ----------------------------------------------------
# Carregando pacotes do R
library(here)
library(haven)
library(tidyr)
library(ggplot2)
library(dplyr)
# ----------------------------------------------------
# Carregando o arquivo de dados
af <- read_dta(
  file = here::here("data", "exercise.dta"))
af
```

Carregando os dados

```
## # A tibble: 37 x 9
##       id group    y0    y2    y4    y6    y8   y10   y12
##    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1     1     1    79    NA    79    80    80    78    80
##  2     2     1    83    83    85    85    86    87    87
##  3     3     1    81    83    82    82    83    83    82
##  4     4     1    81    81    81    82    82    83    81
##  5     5     1    80    81    82    82    82    NA    86
##  6     6     1    76    76    76    76    76    76    75
##  7     7     1    81    84    83    83    85    85    85
##  8     8     1    77    78    79    79    81    82    81
##  9     9     1    84    85    87    89    NA    NA    86
## 10    10     1    74    75    78    78    79    78    78
## # ... with 27 more rows
```

Transformando os dados

```
names(af)[which(names(af) == "group")] <- "trt"
af.longo <- gather(data = af,
                   key = "tempo",
                   value = "fc", -id, -trt)
af.longo
```

```
## # A tibble: 259 x 4
##        id   trt tempo    fc
##     <dbl> <dbl> <chr> <dbl>
##  1      1     1 y0       79
##  2      2     1 y0       83
##  3      3     1 y0       81
##  4      4     1 y0       81
##  5      5     1 y0       80
##  6      6     1 y0       76
##  7      7     1 y0       81
```

Transformando os dados

```
## 8      8      1 y0        77
## 9      9      1 y0        84
## 10    10      1 y0        74
## # ... with 249 more rows
```

```
af.longo <- subset(af.longo, tempo != "y2" & tempo != "y10")

af.longo$dia <- factor(af.longo$tempo,
                       labels = c(0, 12, 4, 6, 8))
af.longo$dia <- factor(af.longo$dia,
                       levels = c("0", "4", "6", "8", "12"))
af.longo$tempo <- as.numeric(
  factor(af.longo$dia))
af.longo$trt <- factor(af.longo$trt)
af.longo
```

Transformando os dados

```
## # A tibble: 185 x 5
##         id trt   tempo    fc dia
##      <dbl> <fct> <dbl> <dbl> <fct>
##  1      1 1         1    79 0
##  2      2 1         1    83 0
##  3      3 1         1    81 0
##  4      4 1         1    81 0
##  5      5 1         1    80 0
##  6      6 1         1    76 0
##  7      7 1         1    81 0
##  8      8 1         1    77 0
##  9      9 1         1    84 0
## 10     10 1         1    74 0
## # ... with 175 more rows
```

Um modelo spline linear (velha guarda)

```
af.longo <- as.data.frame(af.longo)
library(nlme)
# matriz de covariância não estruturada
mod1 <- gls(fc ~ trt*dia,
            na.action = na.omit,
            corr = corSymm(form = ~ tempo | id),
            weights = varIdent(form = ~ 1 | tempo),
            method = "REML",
            data = af.longo)
# matriz de covariância autoregressiva
mod2 <- gls(fc ~ trt*dia,
            na.action = na.omit,
            corr = corAR1(form = ~ tempo | id),
            method = "REML",
            data = af.longo)
```

Matriz de covariância não estruturada estimada

```
library(lavaSearch2)

knitr::kable(
  getVarCov2(mod1)$Omega,
  digits = 3)
```

1	2	3	4	5
9.668	10.175	8.974	9.812	9.407
10.175	12.550	11.091	12.580	11.928
8.974	11.091	10.642	11.686	11.101
9.812	12.580	11.686	13.991	13.121
9.407	11.928	11.101	13.121	13.945

Comparando os dois modelos {.allowframebreaks} ->

```r
anova(mod1, mod2)
```

```
##      Model df      AIC      BIC    logLik   Test  L.Ratio p
## mod1     1 25 647.3399 724.6837 -298.6699
## mod2     2 12 645.0718 682.1968 -310.5359 1 vs 2 23.73189
```

Exercício

- Ajuste o modelo de covariância **exponencial** e compare com os dois modelos já ajustados.

Pontos fortes e fracos dos modelos de covariância

- Os modelos de padrões de covariância tentam caracterizar a covariância com um número relativamente pequeno de parâmetros.
- No entanto, muitos modelos (por exemplo, autoregressivo, Toeplitz e em faixas) são apropriados apenas quando medições repetidas são obtidas em intervalos iguais e não podem lidar com medições irregulares no tempo.
- Embora exista uma grande variedade de modelos para correlações, a escolha de modelos para variâncias é limitada.
- Eles não são adequados para modelar dados de delineamentos longitudinais inerentemente desbalanceados.

Exercícios

- Realize os exercícios do Capítulo 7 do livro "**Applied Longitudinal Analysis**" (páginas 186 e 187).

Avisos

- **Próxima semana:** Jornada 60 anos do IME; inscrevam-se e participem.
- **Na semana seguinte a próxima:** Semana acadêmica; participem do Salão UFRGS.
- **Próxima aula (29/10):** Modelos lineares de efeitos mistos.
- **Para casa:** ler o Capítulo 7 do livro "**Applied Longitudinal Analysis**".
  - Caso ainda não tenha lido, leia também os Caps. 1, 2, 3, 4, 5 e 6.

Bons estudos!

Instituto de
MATEMÁTICA
E ESTATÍSTICA
60 anos
UFRGS