Análise de Dados Longitudinais
Aula 13.08.2018

José Luiz Padilha da Silva - UFPR
www.docs.ufpr.br/∼jlpadilha

Sumário

Modelos Marginais para Dados Longitudinais

1. Modelar a resposta média $E(Y_i)$.

2. Modelar a estrutura de Variância-Covariância $Var(Y_i),\ i = 1, \ldots N$.

3. Assumir uma distribuição (normal) para a resposta (dispensável).

Modelos Lineares para Dados Longitudinais

Dois caminhos:

1. Assumir resposta normal: usar MQG ou MV (usual ou restrita).
2. Não assumir distribuição para a resposta: usar GEE: “Generalized Estimation Equations”.

Modelos Lineares para Dados Longitudinais

Modelo Linear Geral p-covariáveis

$$Y_{ij} = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} + \varepsilon_{ij}; \quad i = 1, \ldots, N; \; j = 1, \ldots, n,$$

em que $X_{ij1} = 1$. Escrevendo em forma matricial:

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{pmatrix} = X_i \beta + \varepsilon_i$$

em que X_i tem dimensão $n \times p$ e β é um vetor p-variado.

Modelos para a Estrutura de Covariância

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.

Modelos para a Estrutura de Covariância

- **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.

Estrutura de Variância-Covariância

$$Var(Y_i) = \sigma^2 V_0 \quad \text{(supondo homocedasticiadade)}$$

e desta forma V_0 é a matriz de correlação de Y_i.

Como as unidades formam uma amostra aleatória da população, temos que:

$$V = \begin{pmatrix} V_0 & 0 & \cdots & 0 \\ 0 & V_0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & V_0 \end{pmatrix}$$

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 β, 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:

$$Y_{ij} = X'_{ij}\beta + U_i + \varepsilon_{ij}$$

O intercepto é o único termo com variação aleatória.

A diferença entre os indivíduos está explicada pelo intercepto aleatório:

$$\begin{aligned} U_i &\sim N(0, \sigma_u^2) \\ \varepsilon_{ij} &\sim N(0, \sigma^2), \end{aligned}$$

em que U_i e ε_{ij} são independentes.

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.
- 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:

$$\varepsilon_{ij} = \rho\varepsilon_{ij-1} + \omega_{ij}$$

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

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

Então

$$Var(\varepsilon_{ij}) = \rho^2\sigma^2 + \sigma^2(1-\rho^2) = \sigma^2$$

$$Cov(\varepsilon_{ij}, \varepsilon_{ij-1}) = Cov(\rho\varepsilon_{ij-1} + \omega_{ij}, \varepsilon_{ij-1}) = \rho\sigma^2 \quad \text{e}$$

para lags maiores que 1,

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

3. Correlação Exponencial

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

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

- 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

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

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

Por exemplo, um padrão *banded* de tamanho 2 assume $Corr(Y_{ij}, Y_{i,j+k}) = 0, \ \forall k > 2$. 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

Temos:

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

Considere V_1 como simetria composta e V_2 como 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

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

  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.