MAT02035 - Modelos para dados correlacionados

Visão geral de modelos lineares para dados longitudinais

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 média

Introdução

- Distinguem-se duas abordagens principais:
  1. a análise de perfis de resposta;
  2. curvas paramétricas ou semi-paramétricas.
- A análise dos dados longitudinais enfoca as mudanças na resposta média ao longo do tempo e a relação dessas mudanças com as covariáveis.
- O fato de as medidas obtidas no mesmo indivíduo não serem independentes, mas estarem correlacionadas positivamente é uma consideração importante em suas análises, mas para a maioria dos estudos longitudinais a correlação geralmente **não é de interesse científico** em si.
- Antes de discutir abordagens para modelar a resposta média ao longo do tempo, é importante esclarecer a distinção entre **parâmetros substantivos** e **incômodos (de perturbação)** no contexto de um estudo longitudinal.

Parâmetros substantivos e de incômodo para dados longitudinais

- Nos modelos de regressão para dados longitudinais, os parâmetros de regressão β relacionam as mudanças na resposta média ao longo do tempo às covariáveis e são geralmente considerados de **interesse primário** ou intrínseco.
  - Podem ser definidos para resumir aspectos importantes das **questões de pesquisa**.
  - Nos referimos a esses parâmetros como **parâmetros substantivos**.
- Por outro lado, em muitas aplicações, parâmetros que resumem aspectos da covariância ou correlação entre as medidas repetidas são considerados de **interesse secundário**.
  - Os parâmetros associados a esses aspectos secundários dos dados costumam ser chamados de parâmetros **incômodo**.
  - Para a análise de dados longitudinais, os parâmetros de correlação ou covariância são frequentemente considerados parâmetros de incômodo, uma vez que não há interesse intrínseco neles.

Parâmetros substantivos e de incômodo para dados longitudinais

SECUNDÁRIO $\neq$ NEGLIGENCIÁVEL

- Em algumas configurações em que dados correlacionados surgem, pode haver uma reversão completa de funções.
- Em estudos de família, o objetivo é determinar se a presença de doença em um membro da família aumenta o risco de doença para os parentes.
  - As **correlações** entre irmãos e entre pais e filhos são de **interesse principal** (e os parâmetros de regressão β são de incômodo) porque suas magnitudes relativas podem ser usadas para fornecer evidências indiretas de risco genético para a doença devido ao compartilhamento do mesmo conjunto de genes.
- Um exemplo adicional surge quando os pesquisadores estão interessados na heterogeneidade de um efeito de tratamento em uma população.

Modelando a resposta média ao longo do tempo

- **Análise de perfis:** permite padrões arbitrários na resposta média ao longo do tempo.
  - Nenhuma tendência de tempo específica é assumida. Em vez disso, os tempos de medição são considerados como níveis de um fator discreto.
  - Só é aplicável quando todos os indivíduos são medidos no mesmo conjunto de ocasiões e o número de ocasiões geralmente é pequeno.
- **Curva paramétrica:** assume uma tendência linear ou quadrática, por exemplo, para a resposta média ao longo do tempo.
  - Pode reduzir drasticamente o número de parâmetros do modelo.
  - Descrevem a resposta média como uma função explícita do tempo.
  - Não há necessidade de exigir que todos os indivíduos no estudo tenham o mesmo conjunto de tempos de medição, nem mesmo o mesmo número de medições repetidas.

Modelando a covariância

Modelando a covariância

- A contabilização da correlação entre medidas repetidas completa a especificação de qualquer modelo de regressão para dados longitudinais e geralmente aumenta a eficiência ou a precisão com a qual os parâmetros de regressão podem ser estimados.
- Quando um modelo apropriado para a covariância é adotado, erros padrão corretos são obtidos e inferências válidas sobre os parâmetros de regressão podem ser feitas.
- Além disso, quando há dados ausentes, a modelagem correta da covariância é frequentemente um requisito para obter estimativas válidas dos parâmetros de regressão.
- Distinguem-se três abordagens principais:

  1. covariância não estruturada;
  2. modelos de padrões de covariância;
  3. estruturas de covariância de efeitos aleatórios.

Covariância não estruturada

- Permite qualquer padrão arbitrário de covariância entre as medidas repetidas.
  - Isso resulta no que normalmente é chamado de covariância "não estruturada".
  - Assim, quando existem n medidas repetidas, as n variâncias em cada ocasião e $n \times (n-1)/2$ covariâncias (ou correlações) aos pares são estimadas.
- Historicamente, a matriz de covariância não estruturada tem sido o modelo de escolha para a covariância na análise de perfis de resposta (mas também pode ser usada na análise de curvas paramétricas).

Covariância não estruturada

Existem duas **desvantagens** em potencial com essa abordagem.

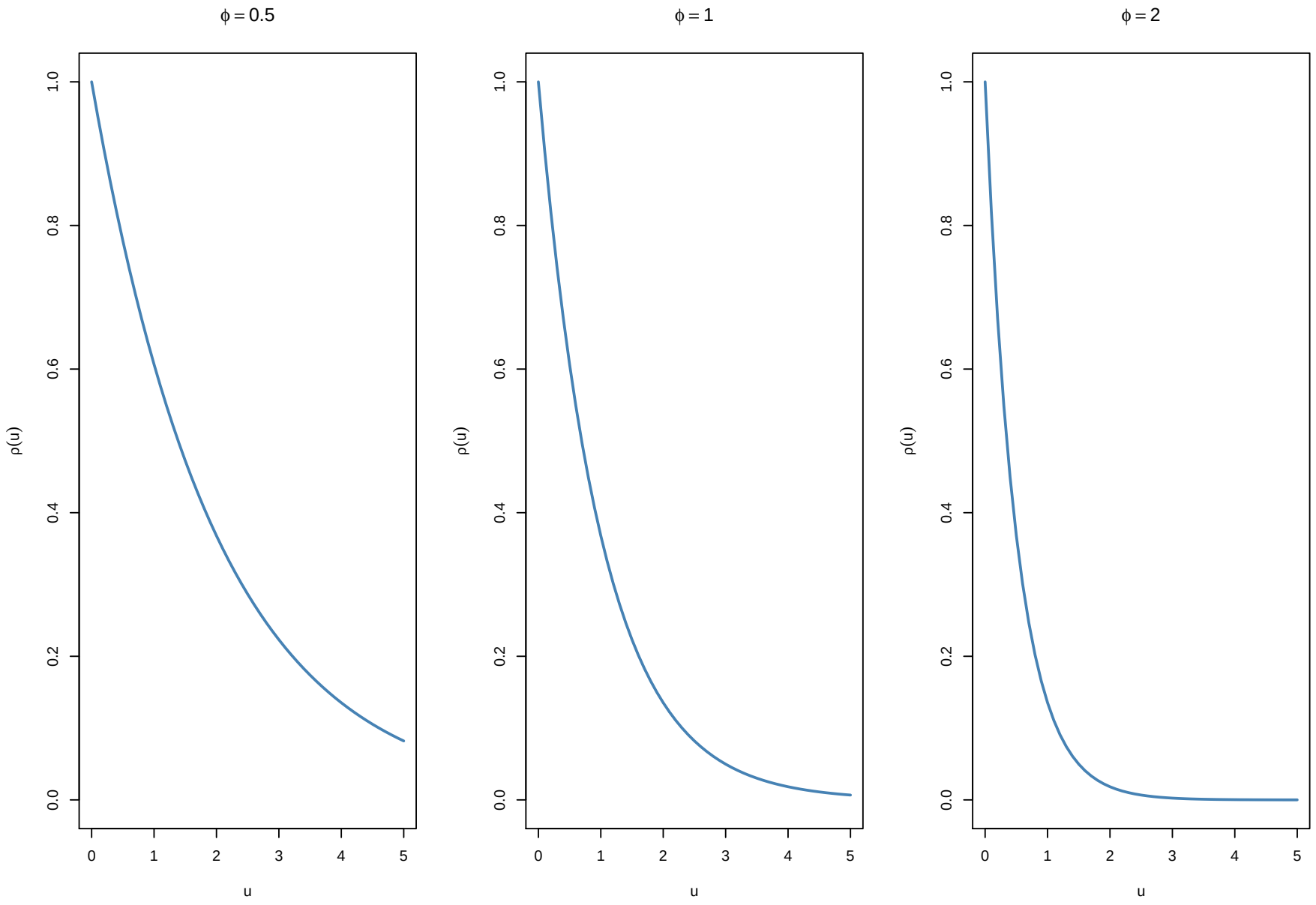
1. O número de parâmetros de covariância pode ser bastante grande.
   - Se houver n ocasiões de medição, a matriz de covariância $n \times n$ terá $n \times (n+1)/2$ parâmetros únicos.
   - Assim, em um estudo longitudinal com 10 ocasiões de medição, uma covariância não estruturada possui 55 parâmetros (10 variâncias e 45 covariâncias).
   - Quando o número de parâmetros de covariância a ser estimado é grande em relação ao tamanho da amostra, é provável que as estimativas sejam **instáveis**.
2. É aplicável apenas quando todos os indivíduos são medidos no mesmo conjunto de ocasiões.

Modelos de padrões de covariância

- Esta abordagem toma emprestado ideias da literatura da análise de séries temporais.
- Espera-se que medidas repetidas tomadas mais próximas no tempo sejam mais altamente correlacionadas do que medidas repetidas mais distantes no tempo.
  - Um caso especial é o **modelo de decaimento exponencial**

$$\rho(u) = \exp\{-\phi u\}, \phi > 0.$$

Modelos de padrões de covariância



Modelos de padrões de covariância

- Isso implica que as correlações decaem à medida que a separação do tempo aumenta.
- Muitas vezes, a correlação entre medidas repetidas é expressa como uma função explícita da separação do tempo.
- Esses modelos podem ser usados com observações desigualmente espaçadas.
- **(Parcimônia)** Modelos paramétricos podem descrever adequadamente a estrutura de covariância entre as medidas repetidas com apenas alguns parâmetros.

Estruturas de covariância de efeitos aleatórios

- Uma estratégia alternativa e um tanto indireta para impor estrutura à covariância é através da introdução de **efeitos aleatórios**.
  - Uma das primeiras abordagens para analisar dados de medidas repetidas.
- No chamado modelo **ANOVA univariada de medidas repetidas**, a correlação entre medidas repetidas é explicada pela inclusão de um único efeito aleatório específico individual.
  - Esse efeito pode ser pensado como um intercepto variando aleatoriamente, representando uma agregação de todos os fatores não observados ou não medidos que tornam alguns indivíduos "altos respondedores" e outros "baixos respondedores".
- A consequência de adicionar um único efeito aleatório específico do indivíduo a todas as medidas em qualquer indivíduo é que as medidas repetidas resultantes serão **correlacionadas positivamente**.
  - Assim, a inclusão de efeitos aleatórios impõe estrutura à covariância.

Abordagens históricas

Abordagens históricas

- De uma perspectiva histórica, três métodos para a análise de dados de medidas repetidas podem ser destacados:
  (1) análise de variância univariada de medidas repetidas (ANOVA)
  (2) análise multivariada de variância de medidas repetidas (MANOVA)
  (3) métodos baseados em medidas sumárias.
- Todas essas três abordagens tiveram graus variados de popularidade e algumas ainda são amplamente utilizadas em diferentes áreas de aplicação.

Abordagens históricas

- Muitas dessas abordagens são desnecessariamente restritivas em suas suposições e objetivos analíticos.
  - Por exemplo, ANOVA e MANOVA se concentram na comparação de grupos em termos de sua tendência de resposta média ao longo do tempo, mas fornecem poucas informações sobre como os indivíduos mudam ao longo do tempo.
  - Além disso, como veremos mais adiante, a ANOVA e a MANOVA têm inúmeras características que limitam sua utilidade para a análise de dados longitudinais.

Análise de medidas repetidas por ANOVA

- Uma das primeiras propostas para analisar respostas correlacionadas foi a **análise de variância** de medidas repetidas **(ANOVA)**, às vezes referida como análise de variância "univariada" ou "modelo misto".
- O paradigma da análise de variância foi desenvolvido no início do século XX por **Ronald A. Fisher**.
- Embora muitas das primeiras aplicações da ANOVA fossem para experimentos delineados na agricultura, desde então, ela foi amplamente difundida em muitas outras disciplinas.

Análise de medidas repetidas por ANOVA

- No modelo ANOVA de medidas repetidas, presume-se que a correlação entre medições repetidas decorra da contribuição aditiva de um **efeito aleatório** específico **do indivíduo** para cada medição em qualquer indivíduo.
- Assim, o modelo assume que a correlação entre medições repetidas ocorre porque cada sujeito possui um **nível de resposta subjacente** (ou latente) que **persiste** ao longo do tempo e **influencia** todas as medições repetidas nesse assunto.
- Esse efeito específico do indivíduo é considerado uma **variável aleatória**.

Análise de medidas repetidas por ANOVA

- Uma característica notável dos modelos ANOVA é que a resposta está relacionada a um conjunto de covariáveis discretas ou fatores.
- No paradigma da ANOVA, as ocasiões de medição são tratadas como um fator adicional intra-individual.
- Assim, se permitirmos que X_{ij} denote o vetor de variáveis indicadoras para os fatores de estudo (por exemplo, grupo de tratamento, tempo e interação), o modelo ANOVA de medidas repetidas pode ser expresso como

$$Y_{ij} = X'_{ij}\beta + b_i + \epsilon_{ij},$$

em que b_i é um efeito específico do indivíduo aleatório e ϵ_{ij} é um erro de medição dentro do indivíduo (é implicitamente assumido que $X_{ij1} = 1$ para todo i e j).

Análise de medidas repetidas por ANOVA

- Embora ambos os b_i e ϵ_{ij} são aleatórios, eles são assumidos como **independentes** um do outro.
- Especificamente, o b_i presume-se que tenha uma **distribuição normal**, com média zero e variância, $\text{Var}(b_i) = \sigma_b^2$.
- Presume-se que os erros ϵ_{ij} também tenham uma **distribuição normal** com média zero, mas com variância, $\text{Var}(\epsilon_{ij}) = \sigma_\epsilon^2$.
- Como b_i e ϵ_{ij} têm média zero, o modelo para a resposta média, calculado sobre as duas fontes de variabilidade, é dado por

$$\text{E}(Y_{ij}|X_{ij}) = X'_{ij}\beta.$$

Análise de medidas repetidas por ANOVA

- Assim, no modelo ANOVA de medidas repetidas, presume-se que a resposta para o i-ésimo indivíduo seja diferente da média da população, μ_{ij}, por um efeito aleatório específico do indivíduo, b_i, que persiste em todas as ocasiões de medição e por um erro de medição intra-indivíduo, ϵ_{ij}.
- Ou seja, o modelo ANOVA de medidas repetidas distingue **duas fontes principais de variação** nos dados:
  - **variação entre indivíduos**, σ_b^2;
  - e variação intra-indivíduo, σ_ϵ^2.
- A variação entre indivíduos reconhece o simples fato de os indivíduos responderem de maneira diferente; alguns são respondedores "altos", outros são respondedores "baixos" e outros são respondedores "médios".
- A variação intra-indivíduo reconhece que existem flutuações aleatórias que surgem do processo de medição, por exemplo, devido a erro de medição e/ou variabilidade da amostra.

Análise de medidas repetidas por ANOVA

- Dadas essas suposições sobre as duas principais fontes de variação, a **matriz de covariância** das medidas repetidas tem a seguinte estrutura:

$$\text{Cov}\,(Y_i) = \begin{pmatrix} \sigma_b^2 + \sigma_\epsilon^2 & \sigma_b^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma_\epsilon^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 + \sigma_\epsilon^2 & \cdots & \sigma_b^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \cdots & \sigma_b^2 + \sigma_\epsilon^2 \end{pmatrix}.$$

Análise de medidas repetidas por ANOVA

- É importante notar que as variância em todas as ocasiões são iguais, $(\sigma_b^2 + \sigma_\epsilon^2)$, assim como as covariâncias, σ_b^2.
- Consequentemente, a correlação entre qualquer par de medidas repetidas,

$$\text{Corr}\,(Y_{ij}, Y_{ik}) = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\epsilon^2}$$

é positiva (em virtude do fato de que as variâncias,σ_b^2 e σ_ϵ^2 devem ser positivas) e constante, independentemente do tempo decorrido entre as ocasiões de medição.

Análise de medidas repetidas por ANOVA

- Essa estrutura de covariância específica também é conhecida como **simetria composta** e tem uma justificativa de aleatorização em certos delineamentos de medidas repetidas (por exemplo, delineamentos de parcelas subdivididas).
  - Em um experimento em que o fator dentro do indivíduo é alocado aleatoriamente para os indivíudos, argumentos de aleatorização podem ser feitos para mostrar que a variância constante e as condições de correlação constantes são válidas.
- Historicamente, isso forneceu uma justificativa atraente para o uso da análise de medidas repetidas por ANOVA em experimentos aleatorizados.
- O argumento de aleatorização simplesmente não é justificável na configuração de dados longitudinal; **ocasiões de medição** não podem ser alocadas aleatoriamente aos indivíduos.

Análise de medidas repetidas por ANOVA

- Como resultado, a suposição de simetria composta para a covariância é frequentemente inadequada para dados longitudinais.
- Além disso, a suposição de variância constante ao longo do tempo muitas vezes não é realista.
- Finalmente, como originalmente concebido, o modelo ANOVA de medidas repetidas foi desenvolvido para a análise de dados de experimentos delineados, onde as medidas repetidas são obtidas em um conjunto de ocasiões comuns a todos os indivíduos, as covariáveis são fatores discretos (por exemplo, grupo de tratamento e tempo) e os dados são completos.
- Assim, a ANOVA de medidas repetidas não pôde ser prontamente aplicada a dados longitudinais com espaçamento irregular, incompletos ou quando era interessante incluir covariáveis quantitativas na análise.

Análise de medidas repetidas por ANOVA

- Talvez uma das principais razões para seu uso generalizado tenha sido porque a formulação ANOVA levou a fórmulas computacionais relativamente simples que podiam ser executadas com uma calculadora de mesa ou de bolso (ou, de fato, com caneta, papel e muita perseverança).
- Historicamente, a ANOVA de medidas repetidas foi provavelmente um dos poucos modelos que poderiam ser ajustados realisticamente a dados longitudinais em um momento em que a computação estava em sua infância.
- No entanto, com a moderna computação e a ampla disponibilidade de *software* estatístico para ajustar uma classe mais ampla de modelos para dados correlacionados, há poucas razões para analisar dados longitudinais sob as limitações e restrições inerentes impostas pelo modelo ANOVA de medidas repetidas.

Exercícios

Exercícios

- Com o auxílio do computador, faça os exercícios do Capítulo 2 do livro "**Applied Longitudinal Analysis**" (páginas 44 e 45).
- **Enviar soluções** pelo Moodle através do fórum.

Avisos

- **Para casa:** ler o Capítulo 3 do livro "**Applied Longitudinal Analysis**".
  - Resumir a Seção 3.6 (Abordagens históricas para análise de dados longitudinais).
  - Caso ainda não tenha lido, leia também os Caps. 1 e 2.
- **Próxima aula:** Estimação e inferência estatística.

Bons estudos!