MAT02035 - Modelos para dados correlacionados

Dados longitudinais: conceitos básicos

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

Introdução

Introdução

- Apresentamos uma visão geral dos principais objetivos da análise longitudinal e algumas das características definidoras dos dados longitudinais.
- O objetivo é enfatizar que o foco principal da análise de dados longitudinais está na **avaliação de mudanças individuais** na variável resposta **ao longo do tempo**.
- Também revisamos as **características** mais importantes dos delineamentos de estudos longitudinais, introduzimos alguma **notação** para dados longitudinais e destacamos os **principais aspectos** dos dados longitudinais que complicam sua análise.

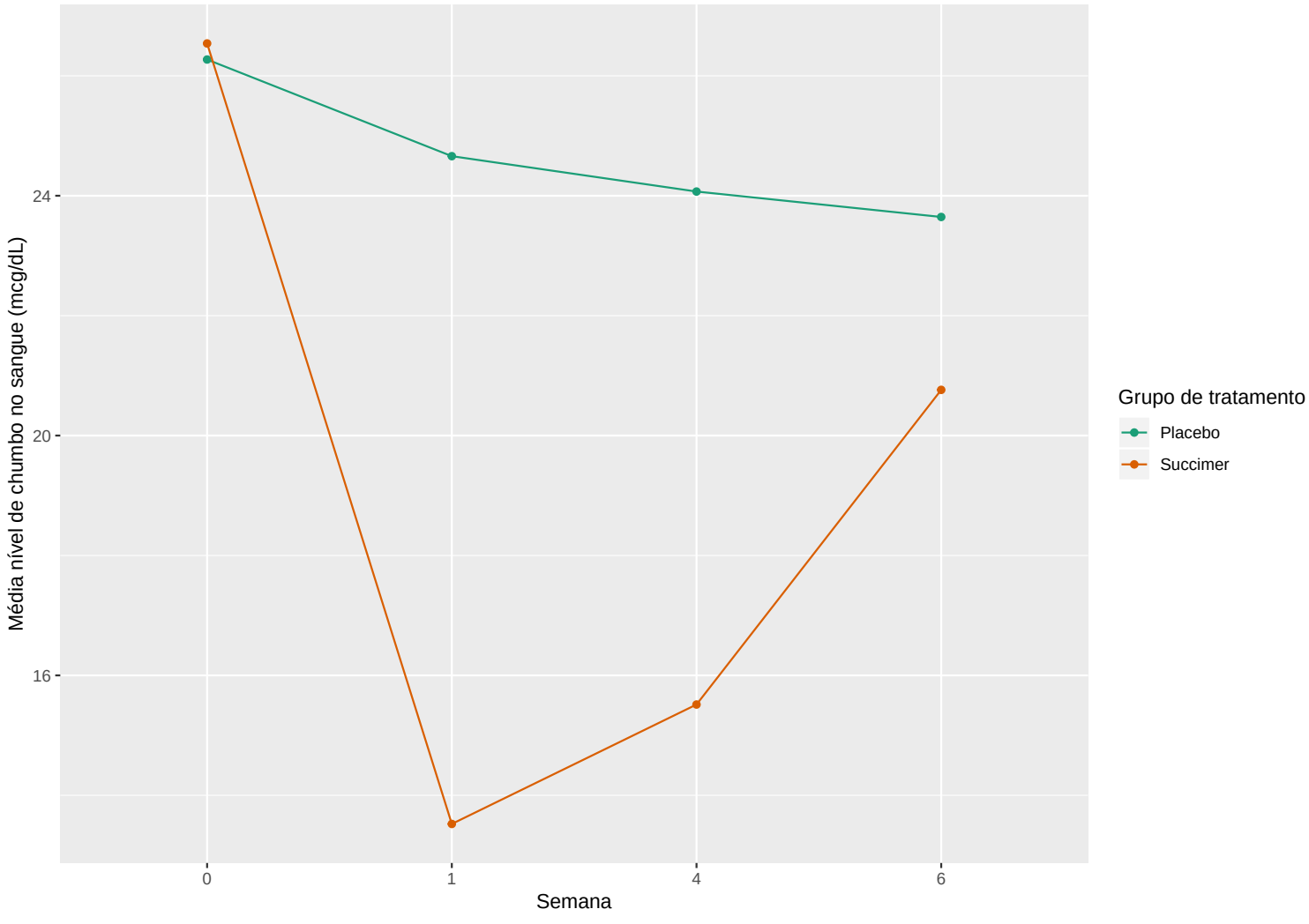
Objetivos da análise longitudinal

Objetivos da análise longitudinal

- A característica distintiva dos estudos longitudinais é que os participantes do estudo são medidos repetidamente ao longo da duração do estudo, permitindo assim a avaliação direta de **mudanças na variável resposta ao longo do tempo**.
- O objetivo principal de um estudo longitudinal é caracterizar a mudança na resposta ao longo do tempo.
- Embora a mensuração de alterações dentro de cada indivíduo seja um objetivo fundamental de um estudo longitudinal, também é interessante determinar se essas mudanças dentro de cada indivíduo na resposta estão relacionadas a covariáveis selecionadas.

Objetivos da análise longitudinal

Exemplo: Tratamento em Crianças Expostas a Chumbo



Objetivos da análise longitudinal

- Uma análise longitudinal das mudanças dentro do indivíduo procede em dois estágios conceitualmente distintos.
  1. A mudança individual na resposta é caracterizada em termos de algum resumo apropriado das mudanças nas medidas repetidas em cada indivíduo durante o período de observação (por exemplo, usando "escores de diferença" ou alguma forma de "trajetória de resposta").
  2. Essas estimativas de mudanças individuais são relacionadas a diferenças interindividuais em covariáveis selecionadas.
- Embora essas duas etapas da análise sejam conceitualmente distintas, elas podem ser **combinadas em um modelo estatístico** para dados longitudinais.
- Um único modelo estatístico para dados longitudinais pode ser usado tanto para capturar como os indivíduos mudam ao longo do tempo quanto para relacionar mudanças individuais na resposta a covariáveis selecionadas.

Objetivos da análise longitudinal

- É um fato inescapável que a avaliação das mudanças dentro do sujeito na resposta ao longo do tempo pode ser alcançada apenas dentro de um delineamento de estudo longitudinal.
- Um estudo transversal simplesmente não pode estimar como os indivíduos mudam com o tempo, já que a resposta é medida em uma única ocasião.
- Um estudo longitudinal pode estimar como os indivíduos mudam e também o fazem **com grande precisão**, porque cada indivíduo age como seu próprio controle.
- Ao comparar as respostas de cada indivíduo em duas ou mais ocasiões, uma análise longitudinal pode remover fontes alheias, mas inevitáveis, de variabilidade entre os indivíduos.

Objetivos da análise longitudinal

- O ponto chave aqui é que há heterogeneidade natural entre os indivíduos em muitas variáveis alheias.
- Embora essas variáveis externas não sejam de interesse substantivo, elas podem ter um impacto na variável resposta.
- A beleza de um delineamento de estudo longitudinal é que quaisquer fatores externos (independentemente de terem sido medidos) influenciam a resposta, e cuja influência persiste, mas permanece relativamente estável durante toda a duração do estudo (por exemplo, gênero, status socioeconômico e muitos fatores genéticos, ambientais, sociais e comportamentais) são eliminados ou bloqueados quando as respostas de um indivíduo são comparadas em duas ou mais ocasiões.
- Ao eliminar essas principais fontes de variabilidade ou "ruído" da estimativa de mudança individual, uma estimativa muito precisa da mudança pode ser obtida.

Características definidoras dos dados longitudinais

Terminologia: indivíduos e tempos

- Em um estudo longitudinal, os participantes, ou, mais geralmente, as unidades em estudo, são referidas como **indivíduos** ou **sujeitos**.
  - Em muitos, mas não em todos, estudos longitudinais, os indivíduos são sujeitos humanos.
  - Em outros estudos longitudinais, os indivíduos podem ser animais.
- Como já mencionado, em um estudo longitudinal, os indivíduos são medidos repetidamente em diferentes **ocasiões** ou **tempos**.
- Assim, adotando a terminologia introduzida até agora, a característica definidora de um delineamento de estudo longitudinal é que as medidas da variável resposta são tomadas nos mesmos indivíduos em várias ocasiões.

Terminologia: dados balanceados

- O número de observações repetidas e seu tempo podem variar muito de um estudo longitudinal para outro.
  - Por exemplo, um ensaio clínico projetado para examinar a eficácia de um novo agente analgésico pode tomar medidas repetidas de uma escala de dor autorreferida no início e no final de seis intervalos de 15 minutos.
  - Isso resultaria em **sete medidas** repetidas que são **igualmente separadas** no tempo.

Terminologia: dados balanceados

- Por outro lado, um estudo observacional do crescimento humano pode fazer medições de altura e peso em intervalos de 3 meses, do nascimento até a idade de 2 anos, seguido por observações anuais desde a infância até a idade adulta jovem.
  - Por delineamento, este último estudo resultaria em uma **sequência de medidas** repetidas de altura e peso que são **desigualmente separadas** no tempo.
- Em ambos os exemplos, o **número** e o **momento** das medições repetidas **são os mesmos** para todos os indivíduos, independentemente de as ocasiões de medição serem igualmente ou desigualmente distribuídas ao longo da duração do estudo.
- Empregando a terminologia estatística emprestada do campo do delineamento experimental, nos referimos aos últimos estudos como sendo **"equilibrados"** (balanceados) ao longo do tempo.

Terminologia: dados desbalanceados

- É uma característica quase inescapável dos estudos longitudinais nas ciências da saúde, especialmente aqueles onde as medidas repetidas se estendem por um período relativamente longo, que alguns indivíduos perderão sua visita programada ou data de observação.
- Em alguns estudos, isso pode exigir que as observações sejam feitas algum tempo antes ou depois do momento programado.
  - Consequentemente, a sequência dos tempos de observação não é mais comum a todos os indivíduos no estudo.
- Nesse caso, nos referimos aos dados como **"desequilibrados"** (desbalanceados) ao longo do tempo.

Terminologia: dados ausentes

- **Dados ausentes** é um problema comum e desafiador em estudos longitudinais.
  - A ocorrência de dados ausentes em estudos longitudinais é a regra, e não a exceção.
- Por exemplo, os participantes do estudo nem sempre aparecem para uma observação agendada, ou podem simplesmente deixar o estudo antes de sua conclusão.
- Quando faltam algumas observações, os dados são necessariamente desequilibrados ao longo do tempo, uma vez que nem todos os indivíduos têm o mesmo número de medições repetidas obtidas em um conjunto comum de ocasiões.

Terminologia: dados ausentes

- Para distinguir dados ausentes em um estudo longitudinal de outros tipos de dados desequilibrados, esses conjuntos de dados são geralmente chamados de **"incompletos"**.
  - Essa distinção é importante e enfatiza o fato de que uma medida pretendida em um indivíduo não pode ser obtida.
- Uma das consequências da falta de equilíbrio e/ou ausência de dados é que requer alguns cuidados para recuperar a mudança dentro de cada indivíduo.
- Por exemplo, considere uma configuração em que cada indivíduo é medido em cada uma das n ocasiões.
  - Em seguida, considere traçar a resposta média em cada ocasião.
  - Diferenças na resposta média ao longo do tempo medem a mudança dentro de cada indivíduo.
  - Isso ocorre porque a diferença nas médias também é a média das diferenças quando cada assunto é medido em todas as ocasiões.

Terminologia: dados ausentes

- Quando há dados ausentes, um gráfico da resposta média sobre o tempo pode ser enganador.
  - Mudanças ao longo do tempo podem refletir o padrão de ausência de dados, e não de mudança individual.
- Como discutiremos ao longo do curso, será necessário examinar cuidadosamente as suposições e a adequação da análise para determinar a validade das inferências com delineamentos desequilibrados e/ou dados ausentes.

Comentários

- Dados longitudinais podem ser balanceados e completos quando todos os indivíduos são medidos em um conjunto comum de ocasiões e não há dados faltantes.
- Dados longitudinais nas ciências da saúde raramente são equilibrados e completos, a menos que os indivíduos não tenham a vontade humana (por exemplo, ratos de laboratório) ou a duração do estudo seja relativamente curta (por exemplo, um estudo longitudinal da eficácia de um analgésico medições repetidas podem ser obtidas em uma única visita de estudo).
- É muito mais comum ter dados longitudinais desequilibrados e/ou incompletos.
- Como resultado, para ser de uso prático real, métodos para a análise de dados longitudinais devem ser capazes de lidar com dados que são desequilibrados ao longo do tempo e possivelmente incompletos.

Terminologia: correlação

- Um aspecto dos dados longitudinais que aparece com destaque em sua análise estatística é que as medidas repetidas no mesmo indivíduo são geralmente **positivamente correlacionadas**.
- As observações correlacionadas são uma característica positiva dos dados longitudinais porque fornecem estimativas mais precisas da taxa de mudança ou o efeito das covariáveis nessa taxa de mudança do que seria obtido a partir de um número igual de observações independentes de indivíduos diferentes.
- No entanto, a correlação entre medidas repetidas viola a suposição fundamental de independência que é fundamental em tantas técnicas de regressão padrão.

Notação: variável resposta

- Seja Y_{ij} a variável resposta para o i-ésimo indivíduo ($i = 1, \ldots, N$) na j-ésima ocasião do tempo ($j = 1, \ldots, n$).
  - Variável aleatória: Y_{ij}
  - Realização da variável aleatória: y_{ij}

Indivíduo	Ocasião				
	1	2	3	$\cdots$	n
1	y_{11}	y_{12}	y_{13}	$\cdots$	y_{1n}
2	y_{21}	y_{22}	y_{23}	$\cdots$	y_{2n}
.	.	.	.	$\cdots$	.
.	.	.	.	$\cdots$	.
.	.	.	.	$\cdots$	.
N	y_{N1}	y_{N2}	y_{N3}	$\cdots$	y_{Nn}

Notação: variável resposta

- As n medidas repetidas da variável resposta em cada indivíduo pode ser agrupada em um vetor resposta $n \times 1$, denotado por

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{pmatrix}.$$

- Uma forma equivalente é dada por

$$Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in})'.$$

Notação: resposta média

- Na análise de dados de um estudo longitudinal, o principal interesse está na resposta média.
  - Mudanças ao longo do tempo na resposta média.
  - Como estas mudanças dependem de covariáveis (grupo de tratamento, exposições).
- Denotamos a média ou o valor esperado de cada resposta Y_{ij} por

$$\mu_j = \mathsf{E}(Y_{ij}).$$

- Podemos pensar em μ_j como uma média sobre uma grande **população** de indivíduos na j-ésima ocasião do tempo.

Notação: resposta média

- Em muitos estudos longitudinais o principal objetivo é relacionar mudanças na resposta média sobre o tempo à covariáveis.
- Para permitir adicionalmente que a resposta média e as mudanças na resposta média variem de indivíduo para indivíduo em função de covariáveis de nível individual, utilizaremos a seguinte notação

$$\mu_{ij} = \mathrm{E}(Y_{ij}).$$

- Aqui, o valor esperado denota uma média sobre uma grande **subpopulação** de indivíduos que compartilham valores semelhantes das covariáveis (por exemplo, indivíduos designados para o grupo de tratamento ativo, sujeitos não expostos) na j-ésima ocasião do tempo.
  - Diremos que μ_{ij} é a resposta média condicional na j-ésima ocasião, em que o termo condicional é usado para denotar a dependência da média nas covariáveis.

Dependência e Correlação

- As noções de dependência e independência têm significados precisos nas estatísticas.
- Duas variáveis são consideradas independentes se a distribuição condicional de uma delas não depender da outra.
- Por exemplo, o nível de colesterol LDL seria considerado independente do sexo se a distribuição do nível de colesterol LDL fosse a mesma para homens e mulheres.

Dependência e Correlação

- Muitas técnicas estatísticas padrão (por exemplo, regressão linear e análise de variância para uma resposta única e univariada) fazem a suposição de que as **observações** do estudo são realizações de variáveis aleatórias que são **independentes** umas das outras.
- Esta suposição será bastante razoável:
  1. Quando o delineamento do estudo exigir que uma observação seja obtida de cada indivíduo **e** os indivíduos sejam selecionados aleatoriamente de uma população maior.
  2. Quando o estudo exige que uma observação seja obtida de cada indivíduo e os indivíduos sejam aleatoriamente designados para diferentes condições de tratamento.
- Além disso, a suposição de observações independentes pode ser justificada em bases puramente físicas ou científicas.
- Ou seja, a resposta de um indivíduo não influencia nem é influenciada pela resposta do outro.

Dependência e Correlação

- No caso em que mais de uma única observação é obtida no mesmo indivíduo, a suposição de observações independentes é simplesmente **insustentável**.
- Ou seja, a resposta de um indivíduo em uma ocasião é muito provável que seja preditiva da resposta do mesmo indivíduo em uma ocasião futura.
  - Por exemplo, um indivíduo com um alto nível de colesterol LDL em uma ocasião é muito provável que também tenha um alto nível de colesterol LDL na próxima ocasião.
- Além disso, com uma **variável resposta quantitativa**, essa dependência entre as medidas repetidas no mesmo indivíduo pode ser caracterizada pela sua **correlação**.

Dependência e Correlação

- Considere um delineamento longitudinal simples que é balanceado e completo, com n medidas repetidas da variável resposta em um conjunto comum de ocasiões em N indivíduos.
- Se denotamos a **média condicional** de Y_{ij} por $\mu_{ij} = \mathrm{E}(Y_{ij})$, então a **variância condicional** de Y_{ij} é definida como

$$\sigma_j^2 = \mathrm{E}\{Y_{ij} - \mathrm{E}(Y_{ij})\}^2 = \mathrm{E}(Y_{ij} - \mu_{ij})^2.$$

- O **desvio-padrão** condicional $\sigma_j = \sqrt{\sigma_j^2}$.
- **Observação:** note que em nossa discussão da variância assumimos que esta pode variar de ocasião para ocasião (σ_j^2).

Dependência e Correlação

A **covariância condicional** entre as respostas em duas diferentes ocasiões, Y_{ij} e Y_{ik}, é denotada por

$$\sigma_{jk} = \mathrm{E}\{(Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})\},$$

e fornece uma medida de **dependência linear** entre Y_{ij} e Y_{ik}, dado as covariáveis.

- A covariância entre Y_{ij} e Y_{ik} pode assumir valores positivos ou negativos.
- Quando a covariância é zero, então não há dependência linear entre as respostas nas duas ocasiões (dado as covariáveis).
- A magnitude da covariância é de difícil interpretação.

Dependência e Correlação

A correlação condicional entre Y_{ij} e Y_{ik} é denotada por

$$\rho_{jk} = \frac{\mathrm{E}\{(Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})\}}{\sigma_j \sigma_k},$$

em que σ_j e σ_k são os desvios-padrão de Y_{ij} e Y_{ik}, respectivamente.

- A correlação pode variar entre -1 e 1.
- Correlação positiva implica que uma variável aumenta conforme a outra aumenta.

Dependência e Correlação

- Embora duas variáveis que são estatisticamente independentes uma da outra sejam necessariamente não correlacionadas, as variáveis podem ser não correlacionadas sem serem independentes (uma vez que a correlação apenas mede a dependência linear).
- A independência estatística é uma condição mais forte que a correlação zero.
  - Implica "nenhuma dependência", isto é, nenhuma dependência linear ou não linear entre as variáveis.
- Por outro lado, a correlação quantifica o grau em que duas variáveis são relacionadas ou dependentes, desde que a dependência seja linear.

Dependência e Correlação

- Quando n medidas repetidas são coletadas em um vetor $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in})$, podemos definir a **matriz de variância-covariância**:

$$\text{Cov}\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{pmatrix} = \begin{pmatrix} \text{Var}(Y_{i1}) & \text{Cov}(Y_{i1}, Y_{i2}) & \ldots & \text{Cov}(Y_{i1}, Y_{in}) \\ \text{Cov}(Y_{i2}, Y_{i1}) & \text{Var}(Y_{i2}) & \ldots & \text{Cov}(Y_{i2}, Y_{in}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(Y_{in}, Y_{i1}) & \text{Cov}(Y_{in}, Y_{i2}) & \ldots & \text{Var}(Y_{in}) \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_{nn} \end{pmatrix}.$$

Dependência e Correlação

- Note que $\text{Cov}(Y_{ij}, Y_{ik}) = \sigma_{jk} = \sigma_{kj} = \text{Cov}(Y_{ij}, Y_{ik})\text{Cov}(Y_{ik}, Y_{ij})$ (**simetria**).
- Ainda, $\sigma_{kk} = \text{Cov}(Y_{ik}, Y_{ik}) = \text{Var}(Y_{ik}) = \sigma_k^2$.
- Dessa forma podemos nos referir a matriz de variância-covariância de Y_i como a (matriz) covariância de Y_i, ou $\text{Cov}(Y_i)$

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

Dependência e Correlação

- Também podemos definir a matriz de **correlação**, $\text{Corr}\,(Y_i)$

$$\text{Corr}\,(Y_i) = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1n} \\ \rho_{21} & 1 & \dots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \dots & 1 \end{pmatrix}.$$

Comentários

- Em dados longitudinais, os pressupostos usuais da análise de regressão padrão não são válidos.
- A heterogeneidade da variância ao longo do tempo pode ser explicada ao permitir que os elementos na diagonal principal da matriz de covariância sejam diferentes.
- A falta de independência entre as medições repetidas é explicada por permitir que os elementos fora da diagonal das matrizes de covariância e correlação sejam diferentes de zero.
- Além disso, espera-se que as correlações sejam positivas e a natureza sequencial dos dados longitudinais implica que pode haver um padrão para as correlações.
  - Por exemplo, espera-se que um par de medidas repetidas que foram obtidas juntas no tempo sejam mais altamente correlacionadas do que um par de medidas repetidas separadas no tempo.
  - Em geral, espera-se que a correlação entre as medidas repetidas diminua com o aumento da separação de tempo.

Avisos

- **Para casa:** ler o Capítulo 1 do livro "**Applied Longitudinal Analysis**". Caso já tenha lido o Cap. 1, leia o Capítulo 2.
- **Próxima aula:** Dados longitudinais - exemplo, fontes de variação e consequências de ignorar a correlação entre dados longitudinais.

Bons estudos!