

MAT02035 - Modelos para dados correlacionados

Dados longitudinais: conceitos básicos

Rodrigo Citton Padilha dos Reis
citton.padilha@ufrgs.br

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DEPARTAMENTO DE ESTATÍSTICA

Porto Alegre, 2023

Objetivos da análise longitudinal

Objetivos da análise longitudinal

- ▶ Nas ciências da saúde (e em outras áreas do conhecimento, como economia e ciências sociais), os **estudos longitudinais** desempenham um papel importante em aumentar nossa compreensão do **desenvolvimento** e **persistência** da doença.
- ▶ Há muita **heterogeneidade** natural **entre os indivíduos** em termos de como as doenças se desenvolvem e progridem.



Objetivos da análise longitudinal

Delineamento de estudo longitudinal

Permite a descoberta de características individuais que podem explicar essas **diferenças interindividuais** nas **mudanças nos desfechos** (**respostas de saúde ao longo do tempo**).

- ▶ 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**¹.

¹Em estudos transversais não é possível avaliar as mudanças individuais com base em uma **única “fotografia” da resposta** do indivíduo em um determinado momento.

Objetivos da análise longitudinal

A característica definidora

de um estudo longitudinal é **que duas ou mais observações da variável resposta, realizadas em momentos diferentes, são obtidas em pelo menos alguns dos participantes do estudo.**

- ▶ Normalmente, o delineamento longitudinal exige um **número fixo de medidas repetidas** a serem feitas **em todos os participantes** do estudo **em um conjunto de pontos de tempo comuns.**
 - ▶ As ocasiões de medição **não são** necessariamente **distribuídas uniformemente** ao longo da duração do estudo.

Objetivos da análise longitudinal

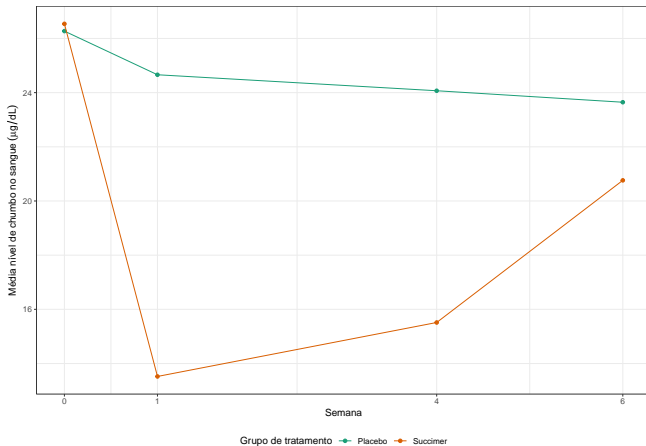
O objetivo principal

de um estudo longitudinal é **caracterizar a mudança na resposta ao longo do tempo**.

- ▶ Também é interessante **determinar se essas mudanças** dentro de cada indivíduo na resposta **estão relacionadas a covariáveis selecionadas**.
 - ▶ **Exemplos:** grupo de tratamento, grupo etário, grupo de exposição a determinado fator (social, econômico, clínico, etc.).

Objetivos da análise longitudinal

Relembrando: exemplo TLC



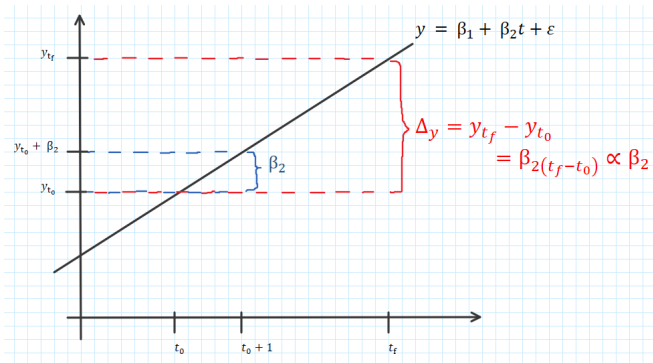
Abordagens de análise

- ▶ Na forma mais elementar, uma medida da **mudança** observada **na resposta dentro de indivíduo** pode ser conceituada em termos de “**escores de mudança**” ou “**escores de diferença**”.
 - ▶ Por exemplo, as diferenças (Δ_y) entre as medidas de pós-tratamento e pré-tratamento na resposta (y):

$$\Delta_y = y_{t_f} - y_{t_0}.$$

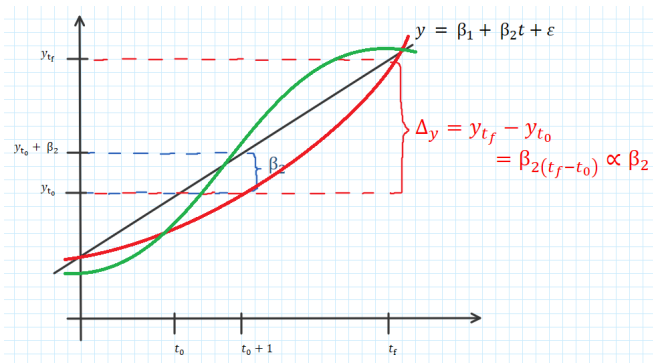
Abordagens de análise

- ▶ Essa noção simples de mudança intraindividual se estende naturalmente de “escores de diferença” para “**trajetórias de resposta**” mais gerais ao longo do tempo.



Abordagens de análise

- Note que outros tipos de trajetórias de resposta, **lineares por partes** ou **curvilíneas**, podem ser usados para suavizar parcimoniosamente e resumir as mudanças intraindividuais na resposta ao longo da duração do estudo.



Abordagens de análise

Uma análise longitudinal das mudanças dentro de 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.**

Abordagens de análise

- ▶ 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 **(1) capturar como os indivíduos mudam ao longo do tempo** quanto para **(2) relacionar mudanças individuais na resposta a covariáveis selecionadas**.

Abordagens de análise

- ▶ Um estudo longitudinal pode estimar como os indivíduos mudam e também o fazem **com grande precisão**, pois 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 indivíduos.

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 estudos longitudinais, os indivíduos são seres humanos.
 - ▶ Em outros casos, os indivíduos podem ser animais ou objetos.
- ▶ Em um estudo longitudinal os indivíduos são medidos repetidamente em diferentes **ocasiões** ou **tempos**.

característica definidora

Assim, adotando a terminologia introduzida até agora, a característica definidora de um estudo longitudinal é que as **medidas da** variável **resposta** são **obtidas nos mesmos indivíduos em diversas ocasiões**.

Terminologia: dados balanceados

- ▶ O **número de observações repetidas** e **seus tempos (momentos de ocorrência)** podem variar muito de um estudo longitudinal para outro.

- ▶ Por exemplo, um ensaio clínico delineado para examinar a eficácia de um novo agente analgésico pode realizar medidas repetidas de uma escala de dor autorreferida na linha de base e ao final de seis intervalos de 15 minutos.
 - ▶ Isso resultaria em **sete medidas** repetidas que são **igualmente separadas** no tempo.
- ▶ 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 **separadas de forma desigual** no tempo.

Terminologia: dados balanceados

- ▶ Em ambos os exemplos, o **número** e o **momento** das medições repetidas **são os mesmos** para todos os indivíduos, independentemente das 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 **“balanceados” (equilibrados)** ao longo do tempo.

Terminologia: dados balanceados

Exemplo TLC

Tabela 1: Níveis de chumbo no sangue de dez crianças do estudo TLC

ID	Grupo	Linha de base	Semana 1	Semana 4	Semana 6
44	Succimer	26.8	20.4	19.3	23.8
46	Placebo	35.4	30.4	26.5	28.1

Exemplo ensaio clínico medicamento anti-epilético

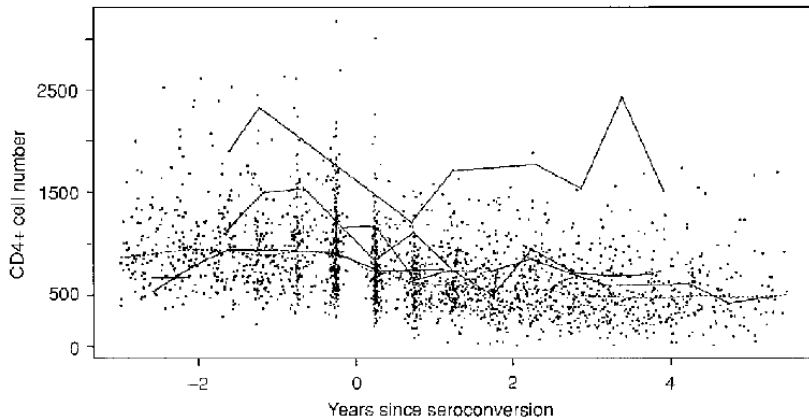
Tabela 2: Número de convulsões por semana

ID	Grupo	Linha de base	Sem 1	Sem 2	Sem 3	Sem 4
43	Progabide	46	11	14	25	15
9	Placebo	23	5	6	6	5

Terminologia: dados desbalanceados

- ▶ É uma característica dos estudos longitudinais, especialmente aqueles em que 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 “**desbalanceados**” ao longo do tempo.

Terminologia: dados desbalanceados



Terminologia: dados ausentes

- ▶ **Dados ausentes** são 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 (**padrão dropout**).
- ▶ Quando faltam algumas observações, os dados são necessariamente desbalanceados ao longo do tempo, uma vez que nem todos os indivíduos têm o mesmo número de medidas repetidas obtidas em um conjunto comum de ocasiões.

Terminologia: dados ausentes

- ▶ Para distinguir dados ausentes em um estudo longitudinal de outros tipos de dados desbalanceados, 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 pôde ser obtida.
- ▶ Uma das consequências da falta de balanceamento e/ou ausência de dados é que requer alguns cuidados para recuperar a mudança dentro de cada indivíduo.

Terminologia: dados ausentes

- ▶ **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 indivíduo 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 desbalanceados e/ou ocorrência de dados ausentes.

Terminologia: correlação

- ▶ Um aspecto dos dados longitudinais é que as medidas repetidas no mesmo indivíduo são geralmente **positivamente correlacionadas**.
- ▶ As observações correlacionadas são uma característica positiva dos dados longitudinais, pois fornecem estimativas mais precisas da taxa de mudança² 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 de independência** que é fundamental em tantas técnicas de regressão padrão.

²Ou o efeito das covariáveis nessa taxa de mudança.

Notação: variável resposta

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

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

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

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

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

- ▶ 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

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

- ▶ 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

- ▶ Na estatística, as noções de dependência e independência têm significados precisos.
- ▶ **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 indivíduos do sexo feminino e masculino.

Dependência e Correlação

- ▶ Muitas técnicas estatísticas (por exemplo, regressão linear e análise de variância para uma resposta única e univariada) assumem 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.

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 $\mu_{ij} = E(Y_{ij})$, então a **variância condicional** de Y_{ij} é definida como

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

- ▶ O **desvio-padrão** condicional $\sigma_j = \sqrt{\sigma_j^2}$.
- ▶ **Observação:** note que assumimos que a variância 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

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

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).

Dependência e Correlação

A magnitude da covariância é de difícil interpretação. A **correlação condicional** entre Y_{ij} e Y_{ik} é denotada por

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

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.
 - ▶ Correlação negativa implica que uma variável aumenta conforme a outra diminui.

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 $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{in})$, podemos definir a **matriz de variância-covariância**:

$$\begin{aligned} \text{Cov} \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{pmatrix} &= \begin{pmatrix} \text{Var}(Y_{i1}) & \text{Cov}(Y_{i1}, Y_{i2}) & \dots & \text{Cov}(Y_{i1}, Y_{in}) \\ \text{Cov}(Y_{i2}, Y_{i1}) & \text{Var}(Y_{i2}) & \dots & \text{Cov}(Y_{i2}, Y_{in}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(Y_{in}, Y_{i1}) & \text{Cov}(Y_{in}, Y_{i2}) & \dots & \text{Var}(Y_{in}) \end{pmatrix} \\ &= \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_{nn} \end{pmatrix}. \end{aligned}$$

Dependência e Correlação

- ▶ Note que $\text{Cov}(Y_{ij}, Y_{ik}) = \sigma_{jk} = \sigma_{kj} = \text{Cov}(Y_{ik}, Y_{ij})$ (**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.

Comentários

- ▶ 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 próximas 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!

