MAT02035 - Modelos para dados correlacionados

Introdução aos dados longitudinais e agrupados

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, 2021

Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

Introdução

Introdução

A pesquisa estatística em métodos para o delineamento e análise de investigações humanas expandiu explosivamente na segunda metade do Século XX.

- Nos EUA (no começo dos anos 1950): mudança de suporte para pesquisa da área militar para a área biomédica.
- Nos EUA (1944): aprovação da Lei do Serviço de Saúde Pública $\Rightarrow$ crescimento do *National Institutes of Health* (NIH[1]).
  - Orçamento do NIH: 1947 - \$8 milhões $\Rightarrow$ 1966 - \$1 bilhão.
- O NIH patrocinou vários dos estudos epidemiológicos importantes e ensaios clínicos daquele período, incluindo o ***Framingham Heart Study***[2] (FHS).

[1] **Um pouco da história do NIH:** www.nih.gov/about-nih/who-we-are/history
[2] **Um pouco da história do FHS:** www.ncbi.nlm.nih.gov/pmc/articles/PMC4159698/

Introdução

- O foco destes primeiros estudos foi a **morbidade** e, especialmente, a **mortalidade**.
- Pesquisadores procuravam **identificar as causas** da morte prematura e avaliar a efetividade dos tratamentos para atrasar a morte e a morbidade.
  - **Regressão logística** (1960s).
  - A análise de dados de **tempo até o evento** foi revolucionada pelo artigo de 1972 de **D. R. Cox**[3], descrevendo **modelo de riscos proporcionais**[4].

[3] Reid, N. A Conversation with Sir David Cox. *Statistical Science*, 9:439-455, 1994.

[4] Cox, D.R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B*, 34:187-220, 1972.

Introdução

- Embora o delineamento do FHS exigisse a **medição periódica** das características do paciente, consideradas determinantes de doenças crônicas, o interesse pelos níveis e **padrões de mudança** dessas características **ao longo do tempo** foi inicialmente limitado.
- Com avanços da pesquisa, investigadores começaram a fazer perguntas sobre o **comportamento** desses **fatores de risco**.
  - No FHS os pesquisadores começaram a perguntar se os **níveis de pressão arterial** na infância eram preditivos de hipertensão na vida adulta.
  - No *Coronary Artery Risk Development in Young Adults* (CARDIA) *Study*[5], os investigadores procuraram identificar os determinantes da transição do estado normotenso[6] ou normocolesterolêmico[7] no início da vida adulta para hipertensão e hipercolesterolemia na meia-idade.

---

[5] Friedman, G.D. *et al.* CARDIA: study design, recruitment, and some characteristics of the examined subjects. *Journal of Clinical Epidemiology*, 41:1105-1116, 1988.

[6] "pressão arterial normal"

[7] "colesterol normal"

Introdução

- No tratamento de artrite, asma e outras doenças que normalmente não ameaçam a vida, os pesquisadores começaram a estudar os efeitos dos tratamentos na mudança ao longo do tempo em medidas de gravidade da doença.
  - Questões semelhantes estavam sendo colocadas em todos os contextos de doenças.
- Os pesquisadores começaram a **acompanhar populações** de todas as idades ao longo do tempo, tanto em estudos observacionais quanto em ensaios clínicos, para **entender o desenvolvimento e a persistência da doença e para identificar fatores que alteram o curso do desenvolvimento da doença**.

Introdução

- Esse interesse nos padrões temporais de mudança nas características humanas ocorreu em um período em que os avanços no poder da computação tornaram novas e mais intensivas abordagens computacionais para a análise estatística disponíveis no *desktop*.

Introdução

- Na década de 1980 **Laird e Ware** propuseram o uso do algoritmo EM para ajustar uma classe de **modelos lineares de efeitos mistos** apropriados para a análise de medidas repetidas[8].
  - **Jennrich e Schluchter** propuseram uma variedade de algoritmos alternativos, incluindo algoritmos de Fisher e Newton-Raphson[9]..
- Mais tarde, na mesma década, **Liang e Zeger** introduziram as **equações de estimativas generalizadas** na literatura bioestatística e propuseram uma família de modelos lineares generalizados para ajustar observações repetidas de dados binários e contagens[10,11].

---

[8] Laird, N. M., Ware, J. H. Random-effects models for longitudinal data. *Biometrics*, 38:963-974, 1982.

[9] Jennrich, R.I., Schluchter, M.D. Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42:805-820, 1986.

[10] Liang, K.Y., Zeger, S.L. Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, 73:13-22, 1986.

[11] Zeger, S.L., Liang, K.Y. Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics*, 42:121-130, 1986.

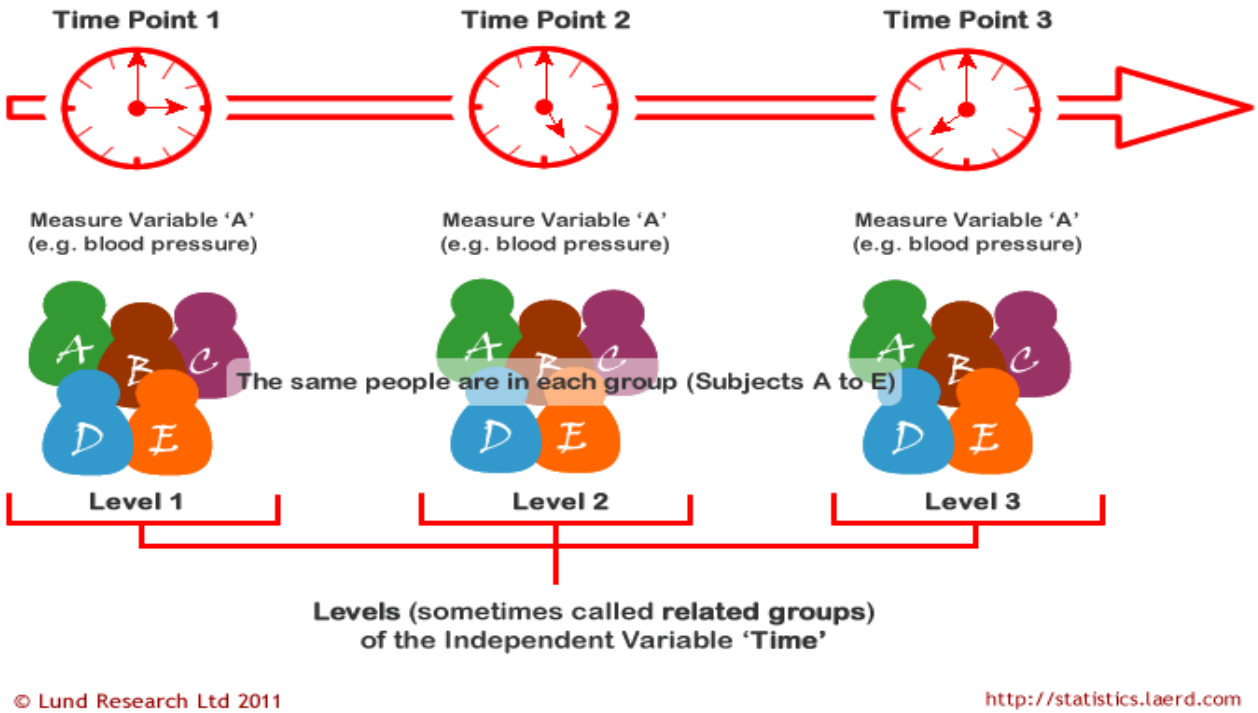
Introdução

- Muitos outros pesquisadores que escrevem na literatura biomédica, educacional e psicométrica contribuíram para o rápido desenvolvimento de metodologia para a análise desses dados "longitudinais".
- Os últimos 40 anos assistiram a progressos consideráveis no desenvolvimento de métodos estatísticos para a análise de dados longitudinais.
- Este curso apresentará parte desta literatura de maneira rigorosa apontando para as possibilidades de aplicação destas técnicas.

Dados longitudinais e agrupados

Estudos longitudinais

- A característica definidora de estudos longitudinais é que medidas dos mesmos indivíduos são tomadas **repetidamente** através do tempo, permitindo, assim, o estudo direto da **mudança ao longo do tempo**.



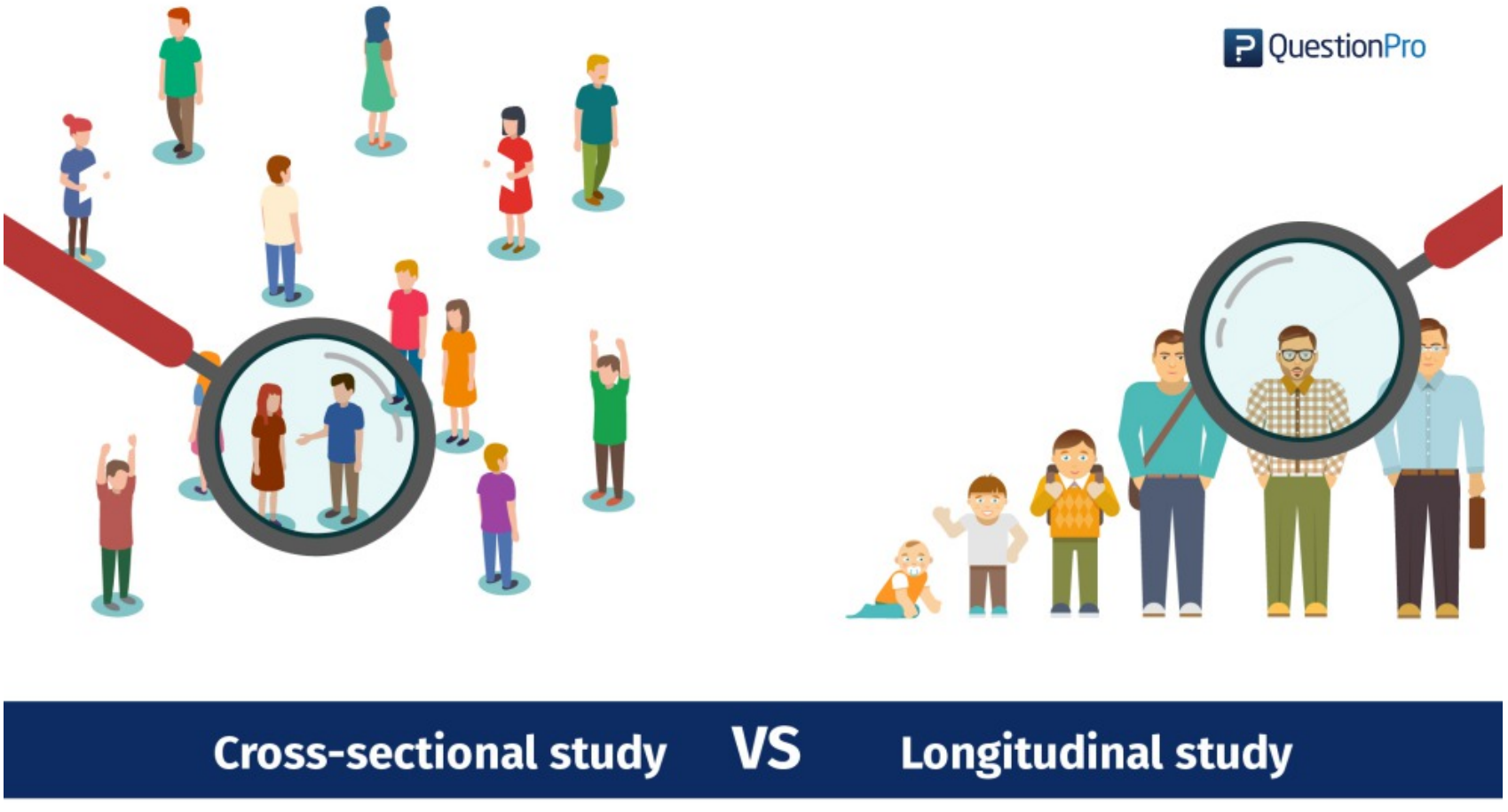
- O **objetivo principal** de um estudo longitudinal é caracterizar a mudança na resposta ao longo do tempo e fatores que influenciam a mudança.

Estudos longitudinais vs. estudos transversais

- Com medidas repetidas realizadas nos indivíduos, é possível capturar a **mudança dentro de indivíduo**[12].
- Em um **estudo transversal**, em que a resposta é medida em apenas uma ocasião, é possível apenas estimar **diferenças** nas respostas **entre indivíduos**.
  - Permite comparações entre subpopulações que diferem em idade, mas não fornece qualquer informação a respeito de como os indivíduos mudam durante o correspondente período.

[12] Intraindividual.

Estudos longitudinais vs. estudos transversais



Um exemplo

Para destacar essa importante distinção entre delineamentos de estudos transversais e longitudinais, considere o seguinte exemplo.

- Acredita-se que a gordura corporal nas meninas aumenta pouco antes ou ao redor da menarca, estabilizando-se aproximadamente 4 anos após a menarca.
- Suponha que os investigadores estejam interessados em determinar o aumento da gordura corporal nas meninas após a menarca.

Um exemplo: estudo transversal

Em um delineamento de estudo transversal, os pesquisadores podem obter medidas de porcentagem de gordura corporal em dois grupos separados de meninas:

- um grupo de meninas de 10 anos (uma coorte[13] pré-menarca);
- e um grupo de meninas de 15 anos de idade (uma coorte pós-menarca).

Neste delineamento de estudo transversal, a comparação direta da porcentagem de gordura corporal média nos dois grupos de meninas pode ser feita usando um **teste t de duas amostras (não pareado)**.

[13] Coorte é um conjunto de pessoas que tem em comum um evento que se deu no mesmo período.

Um exemplo: estudo transversal

- Essa comparação não fornece uma estimativa da mudança na gordura corporal quando as meninas têm entre 10 e 15 anos.
- O **efeito do crescimento** ou **envelhecimento**, um inerente efeito individual, simplesmente não pode ser estimado a partir de um estudo transversal que não obtenha medidas de como os indivíduos mudam com o tempo.
- Em um estudo transversal, o efeito do envelhecimento é potencialmente **confundido** com possíveis efeitos de coorte.
  - Há muitas características que diferenciam as meninas nestes dois grupos etários distintos (hábitos alimentares de "duas gerações", por exemplo) que poderiam distorcer a relação entre a idade e a gordura corporal.

Um exemplo: estudo longitudinal

Por outro lado, um estudo longitudinal que mede uma única coorte de meninas nas idades (momentos) de 10 e 15 anos pode fornecer uma **estimativa** válida **da mudança** na gordura corporal à medida que as meninas envelhecem.

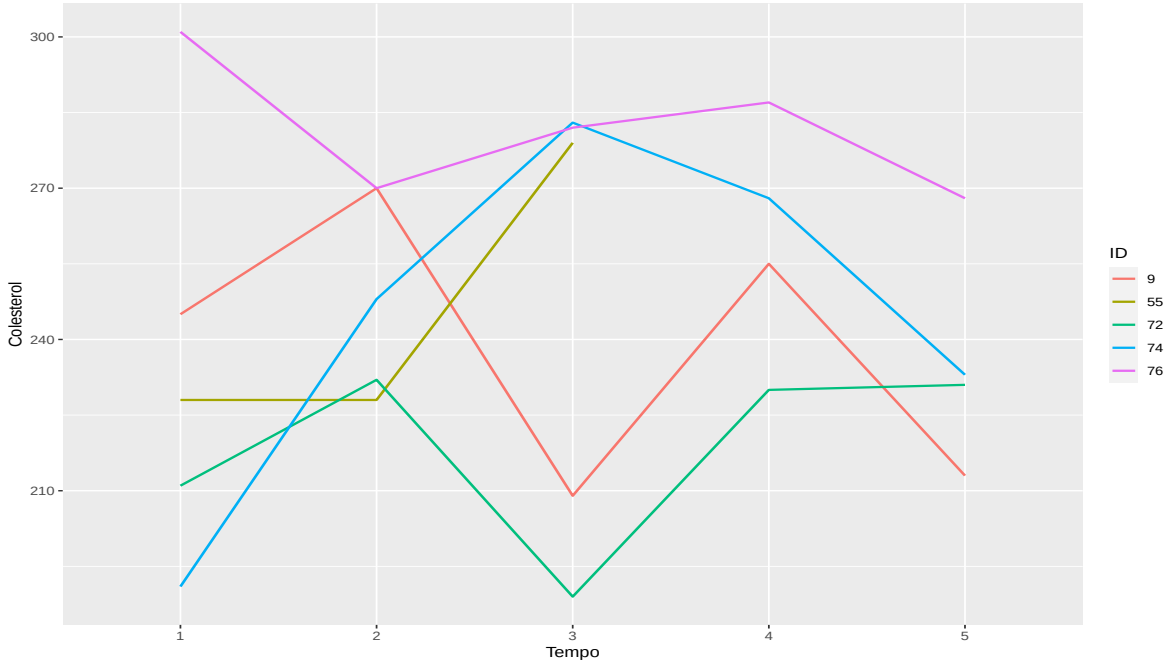
Neste caso, a análise seria baseada em um **teste** t **pareado**, usando a diferença ou mudança na porcentagem de gordura corporal dentro de cada menina como a variável de desfecho[14].

- Esta comparação dentro do indivíduo fornece uma estimativa válida da mudança na gordura corporal quando as meninas envelhecem de 10 a 15 anos.

[14] Variável resposta.

Dados longitudinais são dados agrupados

- Uma **característica distintiva** de dados longitudinais é que eles são **agrupados**.



- **Cada indivíduo forma um grupo de observações repetidas ao longo do tempo.**
- Dados longitudinais também possuem um **ordenamento temporal**, e este tem implicações importantes para a análise.

Dados agrupados

- Observações dentro de um grupo tipicamente irão exibir **correlação positiva**, e esta correlação **deve ser levada em conta** na análise.
- Dados agrupados podem surgir de experimentos aleatorizados por grupo (intervenções de saúde pública) ou de amostragens aleatórias de grupos que se caracterizam naturalmente na população:
  - Famílias;
  - Domicílios;
  - Enfermarias hospitalares;
  - Práticas médicas;
  - Vizinhanças;
  - Escolas.
- Ainda, dados agrupados podem surgir quando a resposta de interesse é simultaneamente obtida de **múltiplos avaliadores** (juízes, examinadores) ou de **diferentes instrumentos de medição**.

Correlação

- Nestes casos, esperamos que medidas em unidades dentro de um grupo são mais similares que as medidas em unidades de grupos diferentes.
- O **grau de agrupamento** pode ser expresso em termos da **correlação** entre as medidas em unidades do mesmo grupo.
- Esta correlação **invalida a suposição crucial de independência** que é o pilar de tantas técnicas estatísticas.
  - Modelos estatísticos para dados correlacionados devem explicitamente descrever e levar em conta esta correlação.

Comentários

- Dados longitudinais são um caso especial de dados agrupados (com uma ordenação natural das medições dentro de um grupo).
  - Faremos a descrição dos métodos de análise para dados agrupados, mais amplamente definidos.
- Um dos objetivos deste curso é demonstrar que os métodos para a análise de dados longitudinais são, mais ou menos, casos especiais de métodos de regressão mais gerais para dados agrupados.
  - Como resultado, uma compreensão abrangente de métodos para a análise de dados longitudinais fornece a base para uma compreensão mais ampla de métodos para analisar a ampla gama de dados agrupados.

Comentários

- Os exemplos descritos anteriormente consideram apenas um único nível de agrupamento.
- Mais recentemente, pesquisadores desenvolveram metodologia para a análise de **dados multiníveis**, em que as observações podem ser agrupadas em mais de um nível.
  - Os dados podem consistir em medições repetidas em pacientes agrupados por clínica.
  - Os dados podem consistir em observações sobre crianças aninhadas dentro de salas de aula, aninhadas dentro das escolas.
- Os dados multiníveis serão discutidos no final do curso.

Um exemplo

Tratamento em Crianças Expostas a Chumbo

- A intoxicação por chumbo em crianças é tratável no sentido de que existirem intervenções médicas, conhecidas como terapias de quelação, que podem ajudar a criança a excretar o chumbo ingerido.
- Até recentemente, o tratamento de quelação de crianças com altos níveis de chumbo no sangue era administrado por injeção e requeria hospitalização.
- Um novo agente quelante, o *succimer*, aumenta a excreção urinária de chumbo e tem a vantagem de poder ser administrado oralmente, em vez de ser administrado por injeção.
- Na década de 1990, o **Grupo de Estudo de Tratamento de Crianças Expostas a Chumbo** conduziu um **estudo aleatorizado** de *succimer*, e controlado por placebo, em crianças com níveis de chumbo no sangue confirmados de 20 a 44 μg/dL (níveis altos)[15,16].

[15] Treatment of Lead-Exposed Children (TLC) Trial Group. Safety and efficacy of succimer in toddlers with blood leads of 20–44 μg/dL. *Pediatric Research*, 48:593-599, 2000.

[16] Rogan, W.J. *et al.* The effect of chelation therapy with succimer on neuropsychological development in children exposed to lead. *New England Journal of Medicine*, 344:1421-1426, 2001.

Tratamento em Crianças Expostas a Chumbo

- As crianças tinham idade entre 12 e 33 meses no momento da entrada no estudo e viviam em habitações em estado de deterioração interna.
- A idade média das crianças no momento da aleatorização[17] foi de 2 anos e o nível médio de chumbo no sangue foi de 26 μg/dL.
- As crianças receberam até três cursos de 26 dias de *succimer* ou placebo e foram acompanhados por 3 anos.

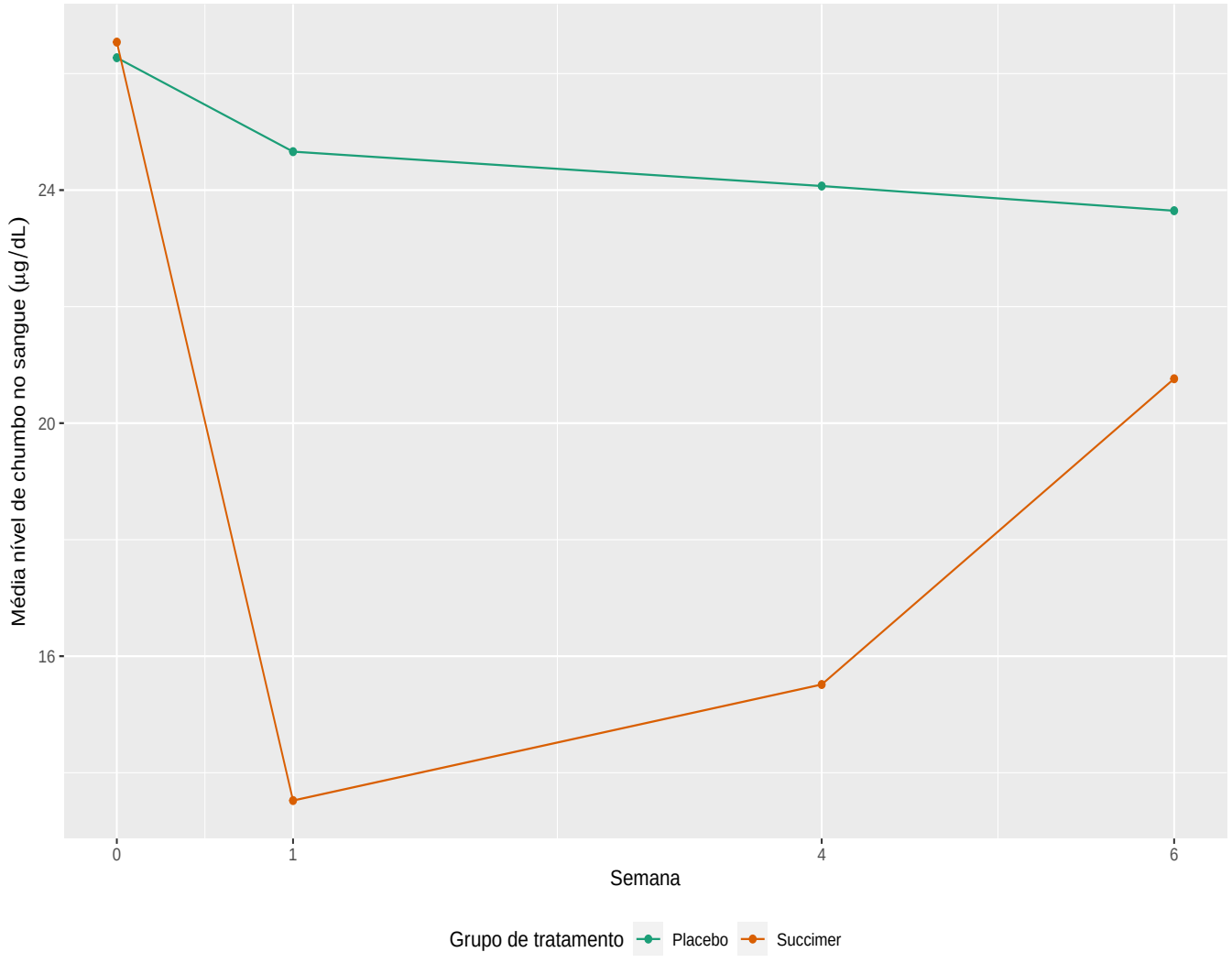
[17] Ou seja, no início do estudo

Tratamento em Crianças Expostas a Chumbo

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

ID	Grupo	Linha de base	Semana 1	Semana 4	Semana 6
9	Placebo	19.7	14.9	15.3	14.7
74	Placebo	22.4	22.0	19.1	18.7
76	Placebo	24.9	23.6	21.2	21.1
55	Placebo	28.5	32.6	27.5	22.8
72	Succimer	21.8	10.6	14.4	18.7
54	Succimer	34.7	39.0	28.8	34.7
39	Succimer	24.4	16.4	11.6	16.6
83	Placebo	20.0	22.7	21.2	20.5
88	Placebo	22.1	21.1	21.5	20.6
15	Placebo	21.1	20.3	18.4	20.8

Tratamento em Crianças Expostas a Chumbo



Modelos de regressão para respostas correlacionadas

Modelos de regressão para respostas correlacionadas

- Nos últimos anos, vimos avanços notáveis nos métodos de análise de dados longitudinais e agrupados.
- Em particular, agora temos uma classe **ampla** e **flexível de modelos** para dados correlacionados com base em um **paradigma de regressão**.
- Todos os métodos descritos nas próximas aulas podem ser considerados modelos de regressão para respostas correlacionadas.

Modelos de regressão para respostas correlacionadas

- Modelos de regressão são amplamente utilizados e fornecem uma abordagem muito geral e versátil para a análise de dados.
- Nosso uso do termo "modelo de regressão" aqui **não é estritamente limitado** ao **modelo de regressão linear** padrão para uma variável **resposta contínua**.
- Usamos esse termo mais amplamente para nos referir a qualquer modelo que descreva a **dependência da média de uma variável resposta em um conjunto de covariáveis** em termos de alguma forma de equação de regressão.

Modelos de regressão para dados correlacionadas

- O caso mais simples é o familiar **modelo de regressão linear** para uma variável resposta contínua.
- No entanto, existem muitas **generalizações** possíveis.
- Modelos de regressão foram desenvolvidos para outras variáveis resposta, como respostas binárias ou contagens.
  - Para a variável **resposta binária**, a **regressão logística** tem sido amplamente utilizada para muitas aplicações.
  - Para **contagens**, a **regressão de Poisson** ou **log-linear** é frequentemente apropriada.
- Outra generalização importante é para observações que não podem ser consideradas estatisticamente independentes umas das outras, ou seja, **modelos de regressão para respostas correlacionadas**.

O modelo linear

- Note que o termo "linear" apareceu em todos os três exemplos de modelos de regressão considerados até agora.
- A linearidade neste cenário tem um significado muito preciso e refere-se ao fato de que todos esses modelos para a média (ou alguma transformação da média) são **lineares nos parâmetros (coeficientes)** da regressão.

O modelo linear

Por exemplo, se Y denota a variável resposta e X uma covariável (variável explicativa ou preditora), os três modelos a seguir para a resposta média

$$\mathrm{E}(Y|X) = \beta_1 + \beta_2 X,$$
$$\mathrm{E}(Y|X) = \beta_1 + \beta_2 \log(X),$$
$$\mathrm{E}(Y|X) = \beta_1 + \beta_2 X + \beta_3 X^2,$$

são todos casos em que a média é linear nos parâmetros de regressão.

- Lembrando que $\mathrm{E}(Y|X)$ denota a média condicional de Y dado X.

O modelo linear

Os dois modelos a seguir

$$\mathrm{E}(Y|X) = \beta_1 + e^{\beta_2 X},$$

$$\mathrm{E}(Y|X) = \frac{\beta_1}{1 + \beta_2 e^{-\beta_3 X}},$$

são casos em que a média é não-linear nos parâmetros de regressão.

- Estes casos não serão considerados em nosso curso.

O modelo linear

- **Observação:** isto não impede relações entre a resposta média e covariáveis que são curvilíneas ou não lineares.
- Este tipo de não-linearidade pode ser acomodado tomando transformações apropriadas da resposta média (transformação logarítmica na regressão de Poisson) e as covariáveis (log(dose)) e/ou incluindo polinômios.
  - Por exemplo, uma tendência quadrática na resposta média ao longo do tempo pode ser incorporada incluindo a covariável *tempo* e o $tempo^2$ no modelo de regressão.
- A inclusão de covariáveis transformadas não viola a "linearidade" do modelo de regressão.
  - O modelo ainda é linear nos parâmetros de regressão.

O modelo linear

- Em resumo, vemos o paradigma de regressão como uma abordagem muito flexível e versátil para analisar dados longitudinais e correlacionados que surgem de muitos diferentes tipos de estudos.
- Modelos de regressão podem fornecer uma descrição parcimoniosa ou explicação de como a resposta média em um estudo longitudinal muda com o tempo, e como essas mudanças estão relacionadas a covariáveis de interesse (sendo estas contínuas ou discretas).
- O nosso principal objetivo é fornecer uma descrição simples dos padrões discerníveis de mudança na resposta ao longo do tempo e sua relação com as covariáveis, através de coeficientes de regressão que se baseiam diretamente nas questões científicas de interesse principal.

Avisos

- **Para casa:** ler o Capítulo 1 do livro "**Applied Longitudinal Analysis**".
- **Próxima aula:** Revisão de vetores e propriedades do valor esperado e variância.

Bons estudos!