MAT02035 - Modelos para dados correlacionados

Dados longitudinais: conceitos básicos (continuação)

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

Exemplo: Tratamento em Crianças Expostas a Chumbo

Introdução

- Consideramos os dados do **estudo sobre tratamento de crianças expostas ao chumbo** (TLC).
- Este ensaio TLC foi um estudo aleatorizado, e controlado por placebo, de succimer em crianças com níveis de chumbo no sangue de 20 a 44 μg/dL (níveis altos de exposição).
- Os dados consistem em quatro medições repetidas dos níveis de chumbo no sangue obtidos na linha de base (ou semana 0), semana 1, semana 4 e semana 6 em 100 crianças que foram aleatoriamente designadas para tratamento de quelação com succimer ou placebo.
- Esses dados são balanceados.

Objetivos da análise

- Em geral, o principal objetivo de uma análise longitudinal é descrever as **mudanças na resposta média ao longo do tempo** e como essas mudanças estão relacionadas às covariáveis de interesse.
- No estudo TLC, os investigadores estavam interessados em determinar se o tratamento de quelação com succimer reduz os níveis de chumbo no sangue ao longo do tempo em relação a quaisquer alterações observadas no grupo placebo.

Objetivos da análise

- Existem muitas maneiras possíveis de expressar essa pergunta em termos de alterações intra-individuais nos níveis de chumbo no sangue.
- Por exemplo, a hipótese nula de **nenhum efeito do tratamento** nas mudanças nos níveis de chumbo no sangue ao longo do tempo pode ser expressa como

$$H_0 : \mu_j(S) = \mu_j(P), \text{ para todo } j = 1, \ldots, 4$$

em que $\mu_j(S)$ e $\mu_j(P)$ denota a resposta média na j-ésima ocasião nos grupos succimer e placebo.

Objetivos da análise

- Esta hipótese nula afirma que as respostas médias em **todos os momentos** coincidem ou são iguais nos dois grupos de tratamento.
- A abordagem de regressão para modelar dados longitudinais pode ser formulada de tal maneira que certos parâmetros de regressão correspondam à **questão científica de interesse**.
  - Aqui, um modelo de regressão para os dados do nível de chumbo no sangue pode incluir efeitos principais para o grupo de tratamento e tempo, além de sua interação.
- A hipótese nula dada acima pode então ser expressa em termos dos parâmetros de regressão para o efeito principal do grupo de tratamento e o tempo pela interação do grupo de tratamento.

Objetivos da análise

- Alternativamente, a hipótese nula de nenhum efeito do tratamento nas alterações dos níveis de chumbo no sangue ao longo do tempo pode ser expressa como

$$H_0 : \mu_j(S) - \mu_1(S) = \mu_j(P) - \mu_1(P), \text{ para todo } j = 2, \ldots, 4.$$

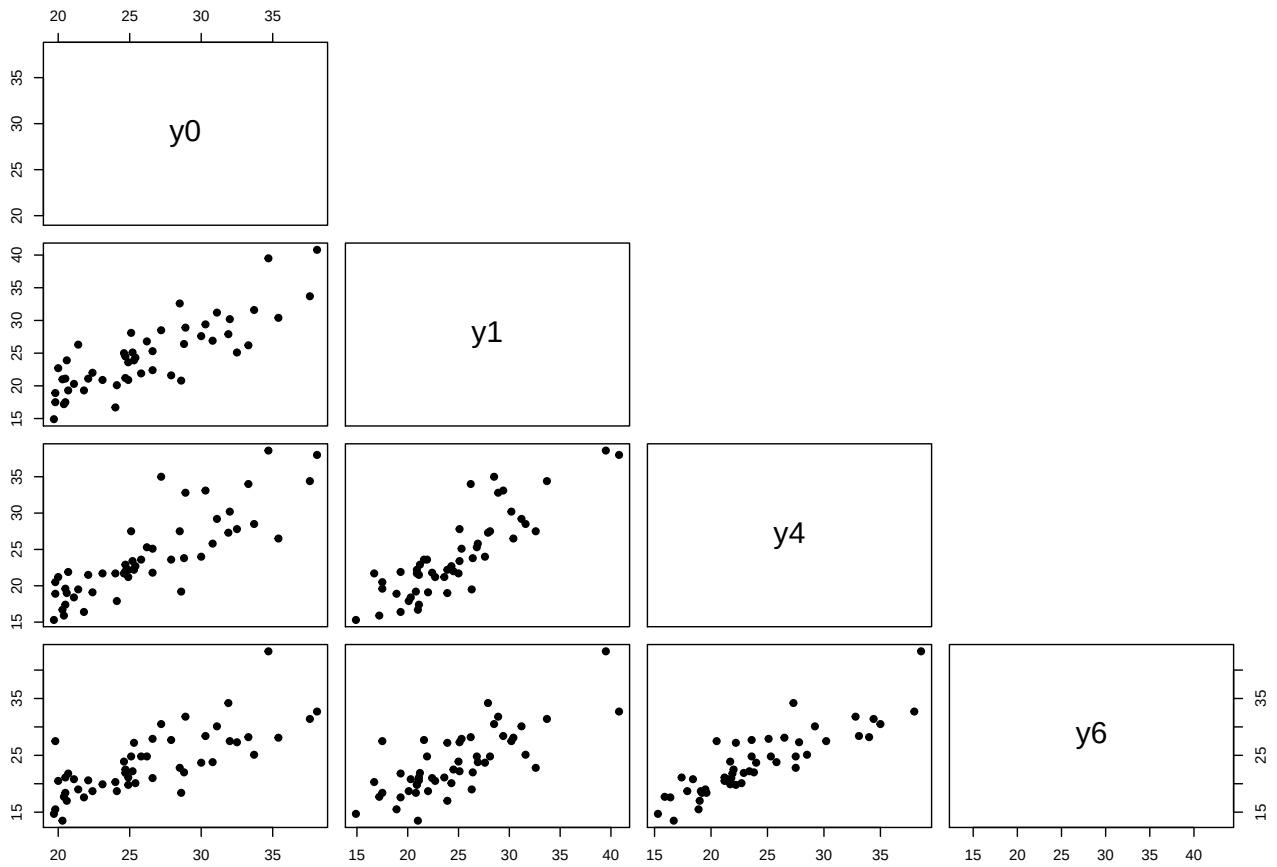
- Esta hipótese nula afirma que todas as alterações na resposta média da linha de base são iguais nos dois grupos de tratamento.
- A segunda versão é um pouco menos restritiva, pois os grupos de tratamento podem ter diferenças de médias na linha de base, mas alterações idênticas da linha de base ao longo do tempo.
- Mais uma vez, um modelo de regressão pode ser formulado correspondendo a esta segunda versão da hipótese nula.

Correlação e covariância

- Para facilitar a exposição, restringimos a atenção aos dados longitudinais do grupo tratado com **placebo** neste estudo.
- Portanto, para o subconjunto de 50 crianças que foram aleatoriamente designadas para o grupo placebo, deixe Y_{ij} denotar o nível de chumbo no sangue para o i-ésimo indivíduo ($i = 1, \ldots, 50$) na j-ésima ocasião ($j = 1, \ldots, 4$).

Correlação e covariância

- Os diagramas de dispersão para todos os seis possíveis pares das quatro medidas repetidas indicam que há uma relação positiva relativamente forte entre as medidas repetidas de níveis de chumbo no sangue ao longo do tempo.



Correlação e covariância

Matriz de covariância

	y0	y1	y4	y6
y0	25.2	22.7	24.3	21.4
y1	22.7	29.8	27.0	23.4
y4	24.3	27.0	33.1	28.2
y6	21.4	23.4	28.2	31.8

- A diagonal principal da matriz de covariância revela que a variância aumenta ao longo do tempo.
  - A variância não-constante dos dados longitudinais é outro tipo de incômodo que é não-padrão na maioria de modelos de regressão.

Correlação e covariância

Matriz de correlação

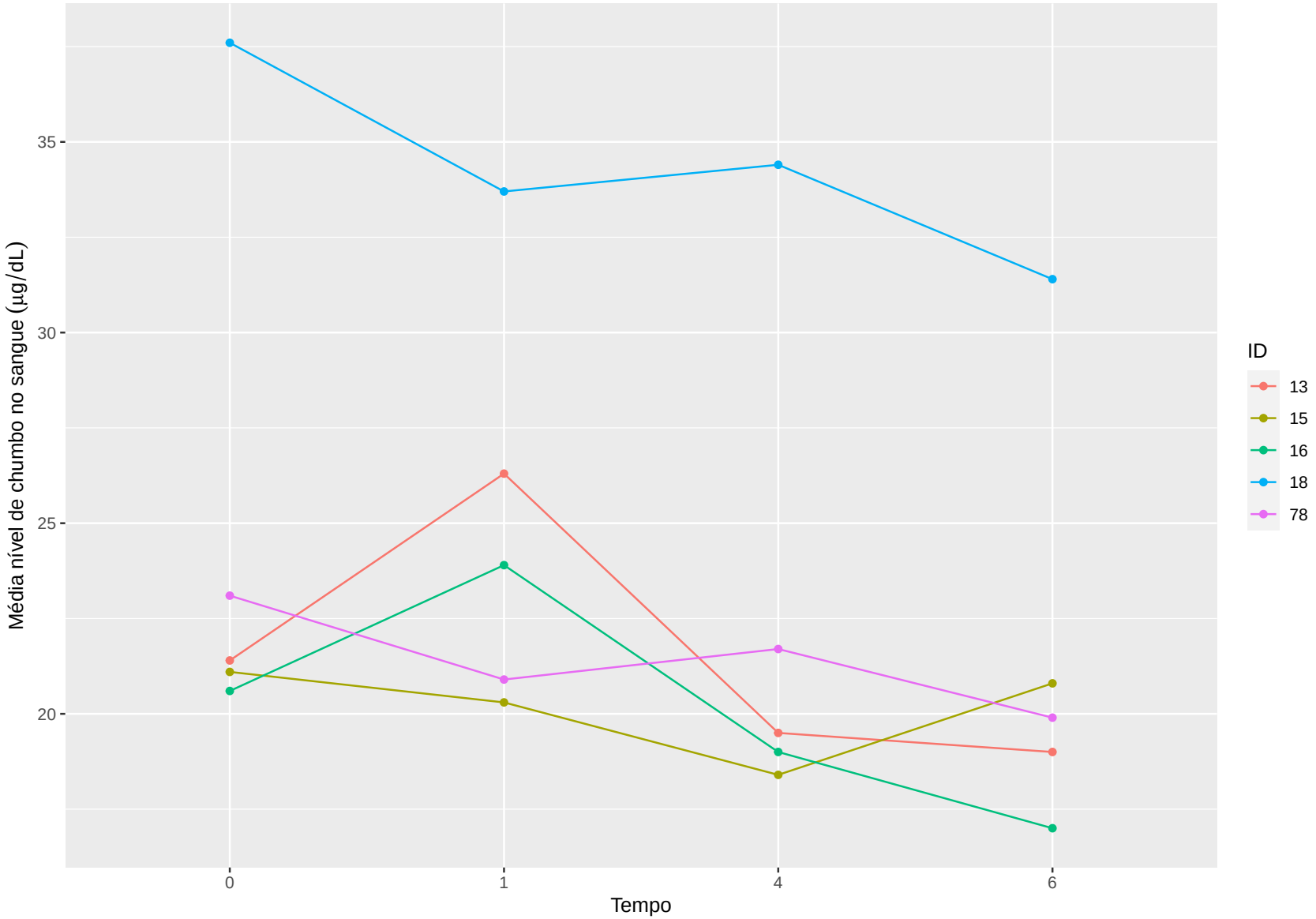
	y0	y1	y4	y6
y0	1.00	0.83	0.84	0.76
y1	0.83	1.00	0.86	0.76
y4	0.84	0.86	1.00	0.87
y6	0.76	0.76	0.87	1.00

- A matriz de correlação confirma que as correlações são todas positivas.
- Ainda, a correlação mostra uma tendência a diminuir com o aumento da separação dos tempos de mensuração.

Correlação e covariância

- Há outra maneira, embora menos óbvia, de avaliar graficamente a dependência entre as medidas repetidas.
- Isso pode ser obtido usando um único gráfico de dispersão que apresenta as respostas no eixo vertical e os tempos das medidas no eixo horizontal, com medidas repetidas sucessivas no mesmo indivíduo unidas por linhas retas.
  - Nos referimos a este gráfico como um *time plot*.
- A dependência entre as medidas repetidas é avaliada comparando a quantidade relativa de variabilidade entre indivíduos e dentro de indivíduos.

Correlação e covariância



Correlação e covariância

- Com base em cinco indivíduos selecionados aleatoriamente do grupo placebo no ensaio de TLC, vemos que há uma variabilidade substancial dentro de indivíduo nos níveis de chumbo no sangue.
- Além disso, há também uma variabilidade substancial entre os indivíduos.
- À primeira vista, esta parece ser uma forma muito indireta de avaliar o grau de dependência entre medidas repetidas e geralmente não é a exibição gráfica mais satisfatória ou informativa dessa dependência.
- No entanto, fornece uma explicação direta para uma das principais fontes da correlação entre medidas repetidas, a saber, entre a heterogeneidade individual.

Fontes de variabilidade em estudos longitudinais

Fontes de variabilidade em estudos longitudinais

- Há geralmente três potenciais fontes de variabilidade que têm impacto na correlação entre as medidas repetidas no mesmo indivíduo:
  1. Variação entre-unidades;
  2. Variação intra-unidade;
  3. Erro de medição.

Variação entre-unidades

- Em qualquer estudo longitudinal alguns indivíduos consistentemente têm uma resposta acima da média, enquanto outros consistentemente têm resposta abaixo da média.
- Uma causa da correlação positiva entre as medidas repetidas é a heterogeneidade ou variabilidade na resposta entre os diferentes indivíduos.
- Um par de medidas repetidas de um mesmo indivíduo tende a ser mais similar que observações únicas obtidas de dois indivíduos aleatoriamente selecionados.

Variação entre-unidades

- Há também heterogeneidade entre os indivíduos quanto as suas trajetórias no tempo.
- Mudanças na resposta ao longo do tempo – devido aos efeitos de tratamento, intervenções ou exposição – não afetam de forma completamente uniforme todos os indivíduos.
- Isso influencia não apenas ter correlação positiva mas também um padrão decrescente de correlação à medida que o tempo aumenta.

Variação entre-unidades

- Nos modelos estatísticos, podemos levar em conta variabilidade entre os indivíduos pela introdução de "efeitos aleatórios" (por exemplo, interceptos e inclinações aleatórios).
- Isto é, alguns efeitos ou coeficientes de regressão são tratados como aleatórios.
- Modelos com efeitos aleatórios serão tratados com detalhe ao longo deste curso.

Variação intra-unidades

- A inerente variabilidade biológica de muitas respostas é uma importante fonte de variabilidade que impacta a correlação entre medidas repetidas.
- Por exemplo, variáveis respostas, tais como pressão sanguínea e dor auto-reportada, flutuam consideravelmente mesmo em intervalos pequenos de tempo.
- Muitas variáveis (ex: níveis séricos de colesterol, pressão sanguínea, ritmo cardíaco, etc) podem ser pensadas como realizações de algum processo biológico ou uma combinação de processos biológicos operando no indivíduo e que variam no tempo.
  - Sucessivos desvios aleatórios não podem ser considerados independentes (variação sistemática).

Variação intra-unidades

- Como consequência, medidas tomadas muito próximas no tempo tipicamente serão mais altamente correlacionadas que medidas mais separadas no tempo.
- Como exemplo, considere que a pressão sanguínea é medida repetidamente em intervalos de 30 minutos.
  - Medições adjacentes serão mais altamente correlacionadas que medidas repetidas tomadas com semanas ou meses de distância.

Erro de medição

- Para algumas respostas de saúde, por exemplo, altura e peso, a variação devido ao erro de medida pode ser negligenciável.
- Para muitas outras, contudo, esta variabilidade pode ser substancial.
- Considere que tomamos duas medidas simultaneamente do mesmo indivíduo, excluindo a possibilidade de qualquer variabilidade biológica, os valores não são esperados serem coincidentes devido à imprecisão do instrumento de medida.

Erro de medição

- Por exemplo, suponha que a variável de interesse seja ingestão de nutrientes, determinada por um biomarcador particular no sangue.
- Além disso, suponha que uma amostra de sangue seja coletada de cada indivíduo e o frasco de sangue seja dividido em duas subamostras, cada uma submetida à medição laboratorial do biomarcador de interesse.
- Em geral, não se espera que essas duas medidas replicadas do biomarcador concordem devido ao erro de medição aleatório.

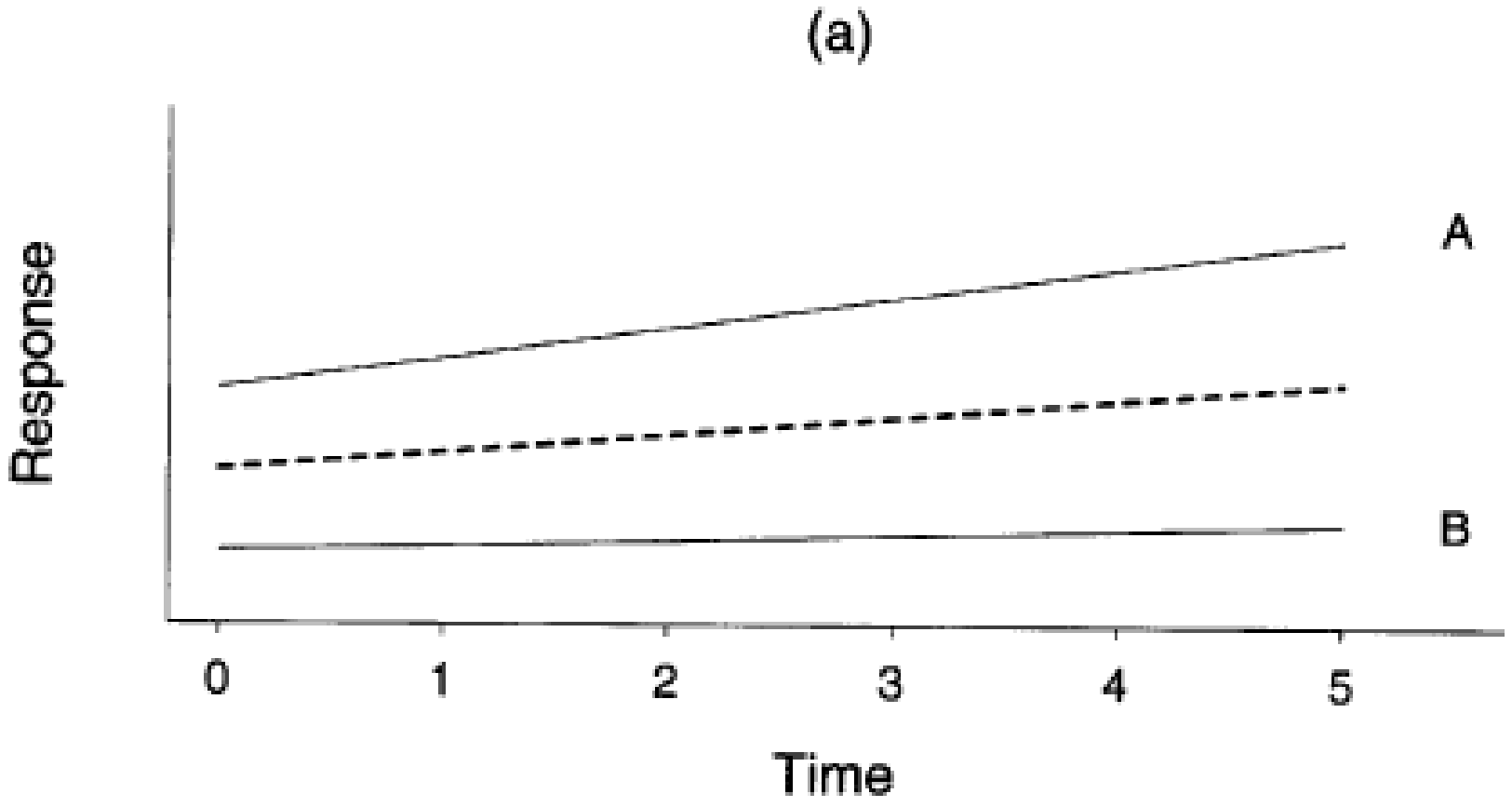
Erro de medição

- Dada a presença de erro de medida, qual o impacto potencial desta variabilidade nas correlações?
  - Em geral, o impacto será de "atenuar" ou "encolher" as correlações em direção ao zero.
- Muitos estudos longitudinais não terão dados suficientes para estimar estas fontes distintas de variabilidade.
  - Elas serão combinadas em um único componente de variabilidade intra-indivíduo.

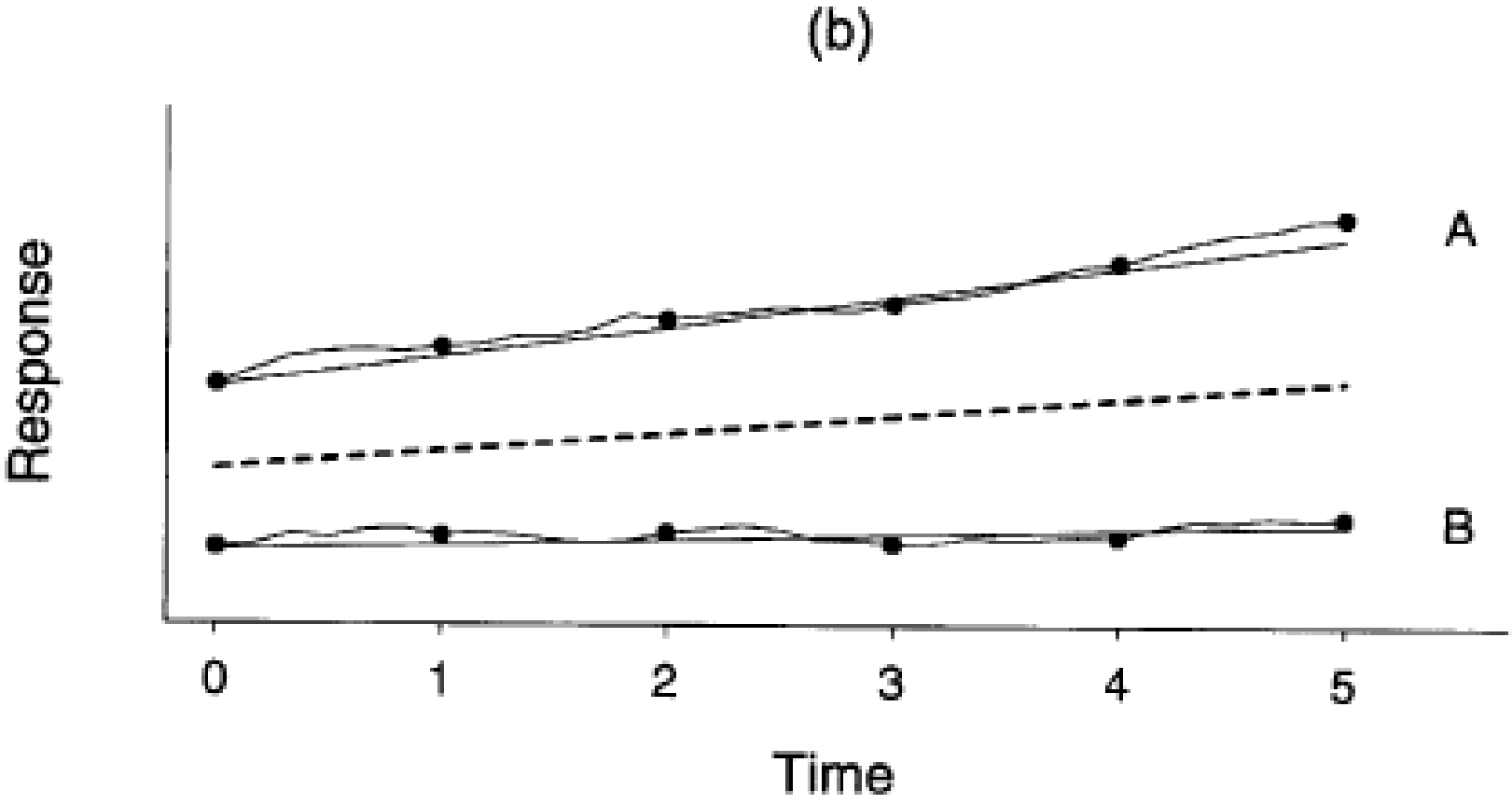
Fontes de variabilidade em estudos longitudinais

- Estas três fontes de variação podem ser visualizadas de forma gráfica.
  - pontos pretos são respostas livre de erro de medição;
  - pontos brancos são as respostas observadas;
  - A e B são diferentes indivíduos.

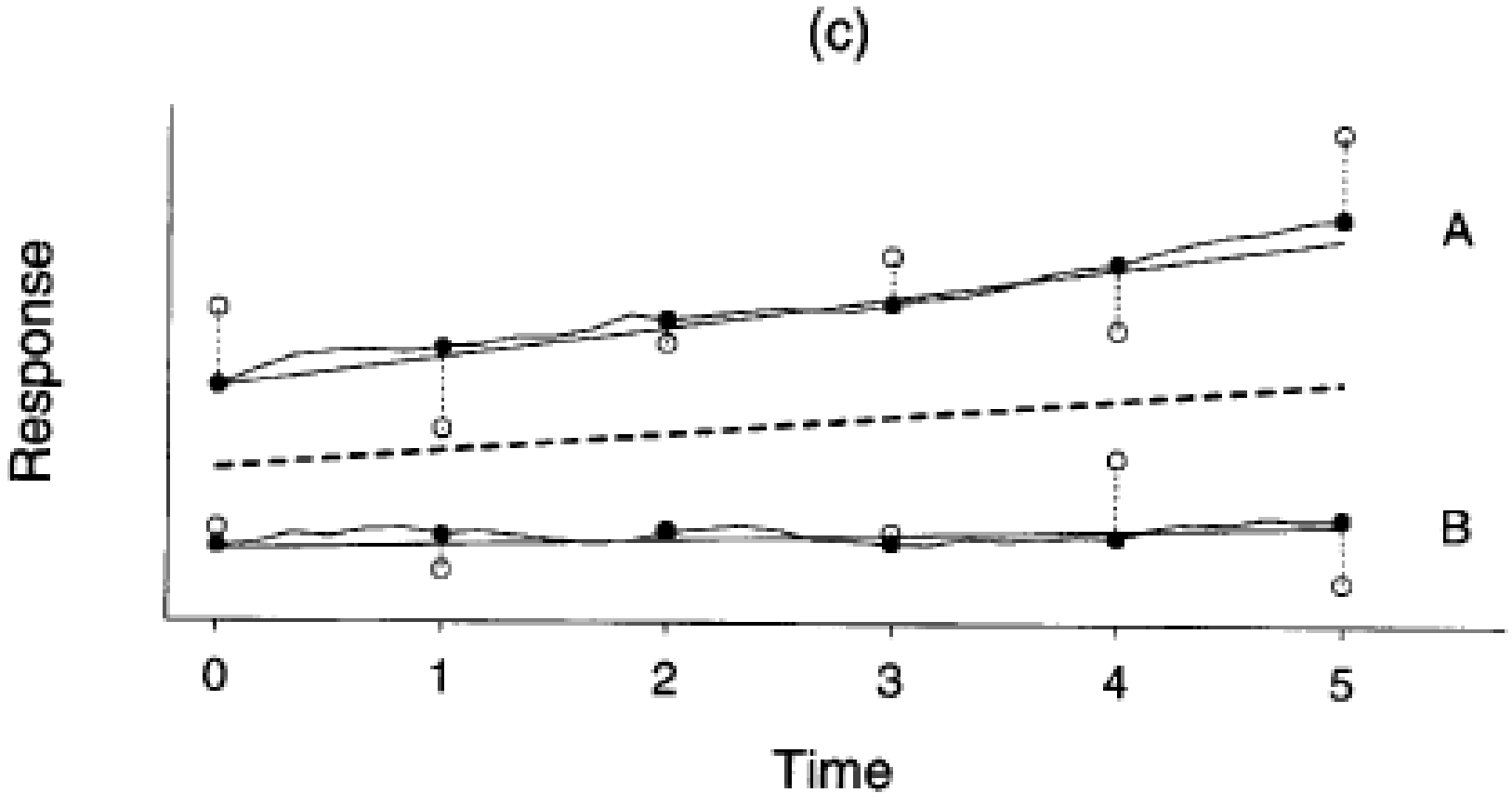
Fontes de variabilidade: entre-unidades



Fontes de variabilidade: intra-unidades



Fontes de variabilidade: erro de medição



Consequências de ignorar a correlação entre dados longitudinais

- Nós vimos que dados longitudinais são, usualmente, positivamente correlacionados, e que a força da correlação é, em geral, uma função decrescente da distância entre os tempos de mensuração.
- Agora, vamos considerar as potenciais implicações de ignorar a correlação entre as medidas repetidas.
- Ao longo do curso, vamos discutir este tópico em maiores detalhes.
- Por hora, veremos o potencial impacto de ignorar a correlação com um exemplo simples usando os dados do **estudo TLC**.

Consequências de ignorar a correlação entre dados longitudinais

- Considere somente as duas primeiras medidas do estudo: linha de base (semana 0) e semana 1.
- Suponha que é de interesse determinar se existe uma mudança na resposta média ao longo do tempo.
- Uma estimativa da mudança na reposta média é dada por:

$$\hat{\delta} = \hat{\mu}_2 - \hat{\mu}_1,$$

$$\hat{\mu}_j = \frac{1}{N} \sum_{i=1}^{N} Y_{ij}, \ j = 1, 2.$$

Consequências de ignorar a correlação entre dados longitudinais

- Para os dados do estudo, no grupo **succimer**, temos $\hat{\delta} = 13.5 - 26.5 = -13$.
- Precisamos de uma medida de incerteza para esta estimativa (erro padrão - EP).
- A expressão da variância de $\hat{\delta}$ é dada por

$$\text{Var}\,(\hat{\delta}) = \text{Var}\left\{\frac{1}{N}\sum_{i=1}^{N}(Y_{i2} - Y_{i1})\right\} = \frac{1}{N}(\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}).$$

- Note que a expressão acima inclui o termo $-2\sigma_{12}$.
  - Este termo é o responsável por levar em consideração a correlação entre as duas primeiras medidas repetidas.

Consequências de ignorar a correlação entre dados longitudinais

- Para os dados do estudo, no grupo **succimer**, temos $\hat{\sigma}_1^2 = 25.2$, $\hat{\sigma}_2^2 = 58.9$ e $\hat{\sigma}_{12} = 15.5$, e portanto:

$$\hat{\text{Var}}(\hat{\delta}) = \frac{1}{50}(25.2 + 58.9 - 2(15.5)) = 1.06.$$

- Se ignorássemos o fato que os dados são correlacionados e e procedêssemos com uma análise assumindo que todas as observações são independentes (e portanto, não correlacionados, com covariância zero), teríamos a seguinte estimativa **(incorreta)** da variância da mudança na resposta média

$$\frac{1}{50}(25.2 + 58.9) = 1.68.$$

Consequências de ignorar a correlação entre dados longitudinais

- Ao ignorar a correlação entre os dados, obtemos uma estimativa da variância da mudança na resposta média 1.6 vezes maior que a estimativa correta.
- Isso acarretará em:
  - Erros padrões muito grandes (superestimados);
  - Intervalos de confiança muito largos;
  - Valores p para o teste $H_0 : \delta = 0$ muito grandes.
- Em resumo, não levar em conta a correlação entre as medidas repetidas irá, em geral, resultar em estimativas incorretas da variabilidade amostral, que levam a inferências bastante enganosas.

Exercícios

Exercícios

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

Avisos

- **Para casa:** ler o Capítulo 2 do livro "**Applied Longitudinal Analysis**".
- **Próxima aula:** Modelos lineares para dados longitudinais (resposta contínua) - visão geral, suposições distribucionais e análise descritiva.

Por hoje é só! Bons estudos!