MAT02035 - Modelos para dados correlacionados

Modelos multiníveis

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, 2023

Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

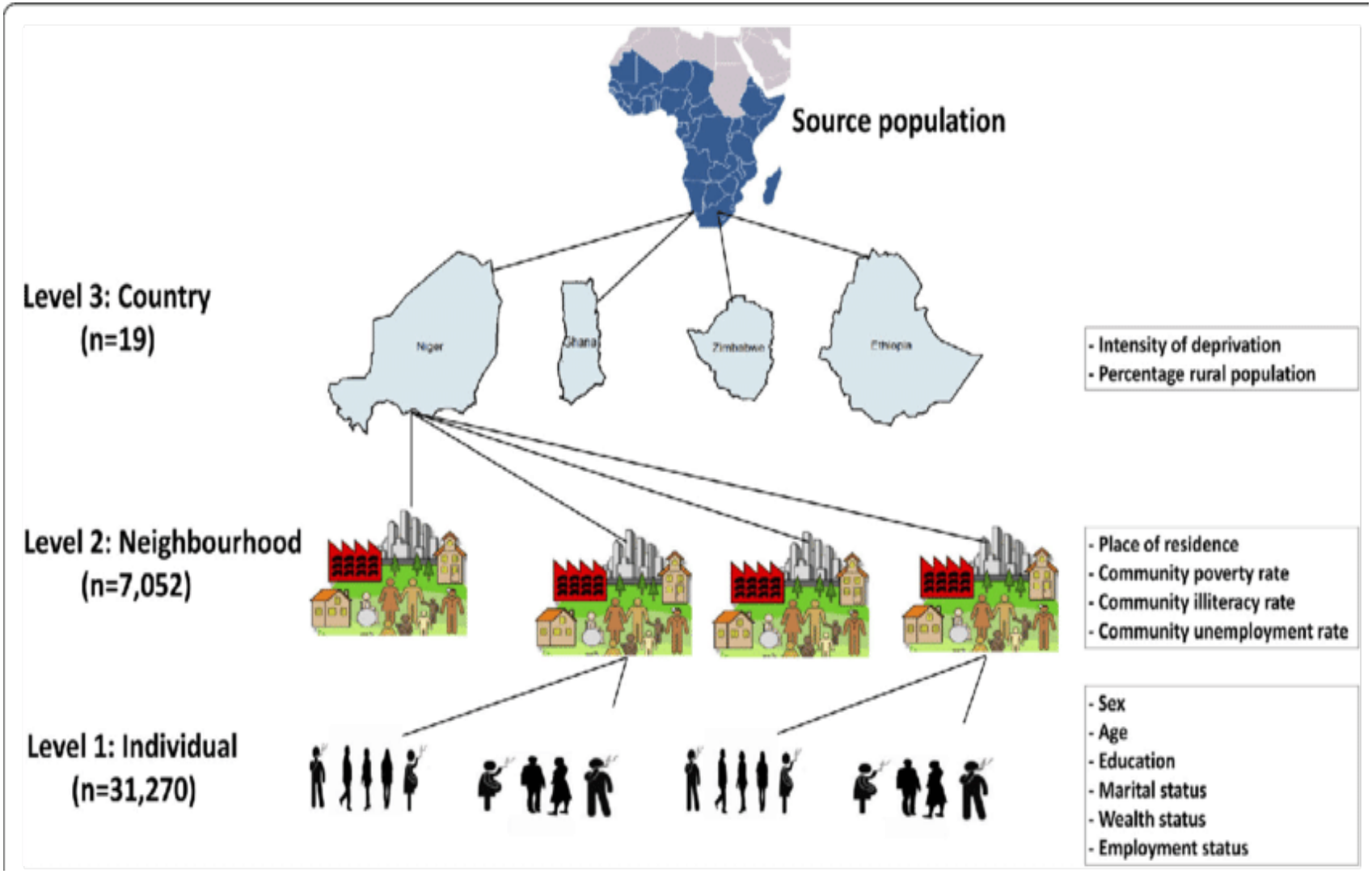
Introdução

- Até hoje, este curso se concentrou na análise de dados longitudinais.
- Modelos mistos também podem ser usados para analisar **dados multiníveis**.
- Os **dados hierárquicos** ou **multiníveis** surgem quando há uma estrutura agrupada nos dados.
- Dados desse tipo frequentemente surgem nas ciências sociais, comportamentais e de saúde, uma vez que os indivíduos podem ser agrupados de muitas maneiras diferentes.
- Por exemplo, em estudos de serviços de saúde e desfechos, frequentemente são obtidas avaliações da qualidade do atendimento de pacientes **aninhados** em diferentes clínicas.

Introdução

- Esses dados podem ser considerados hierárquicos/multiníveis, com pacientes referidos como unidades de nível 1 e clínicas como unidades de nível 2.
  - Neste exemplo, existem dois níveis na hierarquia de dados e, por convenção, o nível mais baixo da hierarquia é referido como nível 1.
- O termo "nível", usado neste contexto, significa a posição de uma unidade de observação dentro de uma hierarquia.
- O agrupamento de dados em vários níveis pode ser devido a uma hierarquia que ocorre naturalmente na população-alvo ou a uma consequência do desenho do estudo (ou às vezes a ambos).

Introdução


Source population
Level 3: Country
(n=19)
Niger
Ghana
Zimbabwe
Ethiopia
- Intensity of deprivation
- Percentage rural population
Level 2: Neighbourhood
(n=7,052)
- Place of residence
- Community poverty rate
- Community illiteracy rate
- Community unemployment rate
Level 1: Individual
(n=31,270)
- Sex
- Age
- Education
- Marital status
- Wealth status
- Employment status

Hierarquias de dados que ocorrem "naturalmente"

- **Estudos de núcleos familiares:** observações sobre mãe, pai e filhos (unidades de nível 1) aninhadas nas famílias (unidades de nível 2).
- **Estudos de serviços/desfechos de saúde:** observações em pacientes (unidades de nível 1) aninhadas nas clínicas (unidades de nível 2).
- **Estudos de educação:** observações em crianças (unidades de nível 1) aninhadas nas salas de aula (unidades de nível 2).
- **Observação:** as estruturas hierárquicas de dados que ocorrem "naturalmente" podem ter mais de dois níveis, por exemplo, crianças (unidades de nível 1) aninhadas nas salas de aula (unidades de nível 2), aninhadas nas escolas (unidades de nível 3).

Agrupamento como consequência do desenho do estudo

- **Estudos longitudinais:** os agrupamentos são compostos de medidas repetidas obtidas de um único indivíduo em diferentes ocasiões.
  - Em estudos longitudinais, as unidades de nível 1 são as repetidas ocasiões de medição e as unidades de nível 2 são os indivíduos.
- **Ensaios clínicos aleatorizados por *cluster*:** grupos (unidades de nível 2) de indivíduos (unidades de nível 1), em vez de os próprios indivíduos, são aleatoriamente designados para diferentes tratamentos ou intervenções.

Agrupamento como consequência do desenho do estudo

- **Pesquisas de amostras complexas:** muitas pesquisas nacionais usam amostragem em vários estágios, por exemplo, a **Pesquisa Nacional em Saúde (PNS)**.
  - Por exemplo, na 1ª etapa, as "unidades de amostragem primária" (UPAs) são definidas com base nos setores censitários. Uma amostra aleatória de UPAs de primeiro estágio é selecionada. No 2º estágio, dentro de cada UPA selecionada, uma amostra aleatória de domicílis é selecionada. No terceiro estágio, dentro dos domicílos selecionados, um indivíduo é aleatoriamente selecionado.
- Os dados resultantes podem ser considerados hierárquicos, sendo os indivíduos as unidades de nível 1, os domícilios as unidades de nível 2 e os setores censitários as unidades de nível 3.

Agrupamento como consequência do desenho do estudo

- Finalmente, o agrupamento pode ser devido ao desenho do estudo e às hierarquias que ocorrem naturalmente na população-alvo.
- **Exemplo:** Os ensaios clínicos são frequentemente realizados em muitos centros diferentes para garantir um número suficiente de pacientes e/ou avaliar a eficácia do tratamento em diferentes contextos.
- As observações de um **ensaio clínico longitudinal multicêntrico** podem ser consideradas dados hierárquicos com 3 níveis, com ocasiões de medição repetidas (unidades de nível 1) aninhadas nos indivíduos (unidades de nível 2) aninhadas nas clínicas (unidades de nível 3).

Característica distintiva de múltiplos níveis de dados

- A característica distintiva dos dados multiníveis é que eles estão **agrupados**.
- Uma consequência desse agrupamento é que a medição em unidades dentro de um agrupamento é **mais semelhante** do que as medidas em unidades em diferentes agrupamentos.
  - Por exemplo, espera-se que duas crianças selecionadas aleatoriamente da mesma família respondam de maneira mais semelhante do que duas crianças selecionadas aleatoriamente de famílias diferentes.
- O agrupamento pode ser expresso em termos de **correlação** entre as medidas em unidades dentro do mesmo agrupamento.
- Modelos estatísticos para dados hierárquicos devem levar em consideração a correlação intra-agrupamento em cada nível; não fazer isso pode resultar em inferências enganosas.

Modelos lineares multiníveis

- A abordagem dominante à análise de dados multiníveis emprega um tipo de **modelo linear de efeitos mistos** conhecido como **modelo linear hierárquico**[1].
- A **correlação induzida** pelo agrupamento é descrita por **efeitos aleatórios** em cada nível da hierarquia.
- **Observação:** Em um modelo multinível, a resposta é obtida no primeiro nível, mas as covariáveis podem ser medidas em qualquer nível.
  - Por exemplo, se estamos estudando o IMC[2], podemos medir dietas individuais, atitudes familiares sobre hábitos alimentares e de compras e atributos da comunidade, como a densidade de restaurantes de *fast-food*.

[1] Também conhecido como **modelo multinível**.

[2] *Índice de massa corporal*; $IMC = \frac{\text{peso em kg}}{(\text{altura em m})^2}$.

Modelos lineares multiníveis

- A combinação de covariáveis medidas em diferentes níveis da hierarquia em um único modelo de regressão é central para a modelagem hierárquica.
- Começamos introduzindo as ideias com o modelo de dois níveis.
- Posteriormente, passamos ao modelo de três níveis para ilustrar a abordagem geral.

Modelos lineares de dois níveis

Notação

- Deixe i indexar o nível 1 e j indexar unidades de nível 2 (por convenção, os subscritos são ordenados do nível mais baixo ao mais alto).
- Assumimos n_2 unidades de nível 2 na amostra.
- Cada um desses agrupamentos ($j = 1, 2, \ldots, n_2$) é composto por n_{1j} unidades de nível 1.
  - Por exemplo, em um estudo de dois níveis de clínicas médicas, estudaríamos as n_2 clínicas, com n_{1j} pacientes na j-ésima clínica.

Modelos lineares de dois níveis

- Deixe Y_{ij} denotar a resposta para o paciente i na j-ésima clínica.
- Associado a cada Y_{ij} está um vetor de $1 \times p$ (linha) de covariáveis, X_{ij}.
- Considere o seguinte modelo para a média:

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

- Por exemplo, em um ensaio clínico multicêntrico comparando dois tratamentos, podemos supor que:

$$\mathrm{E}(Y_{ij}) = \beta_1 + \beta_2 Trt_{ij},$$

em que Trt_{ij} é uma variável indicadora para o grupo de tratamento (ou Trt_j, se o tratamento for constante dentro da clínica).

Modelos lineares de dois níveis

- O modelo linear hierárquico de dois níveis pressupõe que a correlação nas clínicas possa ser descrita por um efeito aleatório.
- Assim, assumimos que

$$Y_{ij} = X_{ij}\beta + b_j + \epsilon_{ij}.$$

Ou, de maneira mais geral,

$$Y_{ij} = X_{ij}\beta + Z_{ij}b_j + \epsilon_{ij},$$

com mais de um efeito aleatório.

Características do modelo linear de dois níveis

1. O modelo define **duas fontes de variação**.
   - As magnitudes da variação dentro e entre grupos determinam o grau de agrupamento/correlação.
2. Para uma determinada unidade de nível 2, efeitos aleatórios são assumidos constantes nas unidades de nível 1.
3. A esperança condicional de Y_{ij}, dada a identidade do grupo de nível 2, é

$$X_{ij} + Z_{ij} b_j.$$

4. Presume-se que as observações de nível 1 sejam **condicionalmente independentes**, dado os efeitos aleatórios.

- O modelo de dois níveis é idêntico ao modelo linear misto com estrutura de correlação intraclasse para medidas repetidas (embora com a inversão do subscrito!).

Modelos lineares de três níveis

Considere um ensaio clínico longitudinal de três níveis em que

1. as clínicas médicas são aleatorizadas para tratamento,
2. os pacientes estão aninhados nas clínicas e
3. os pacientes são medidos no início e em três ocasiões após o tratamento.

- O nível 1 é a ocasião, o nível 2 é o paciente e o nível 3 é a clínica.

Modelos lineares de três níveis

- Deixe Y_{ijk} denotar a resposta na i-ésima observação do j-ésimo paciente na k-ésima clínica.
- As covariáveis podem ser medidas em qualquer um dos três níveis. No entanto, agora introduza efeitos aleatórios para representar agrupamentos nos níveis 2 e 3.
- O modelo linear geral de três níveis é escrito da seguinte maneira:

$$Y_{ijk} = X_{ijk}\beta + Z^{(3)}_{ijk} b^{(3)}_{k} + Z^{(2)}_{ijk} b^{(2)}_{jk} + \epsilon_{ijk}$$

Exemplo: modelo de três níveis para o ensaio clínico longitudinal multinível

- Vamos denotar t_{ijk} o tempo a partir da linha de base na qual Y_{ijk} é obtido.
- Trt_{ij} deve denotar o tratamento dado ao j-ésimo paciente na i-ésima ocasião.
- O tratamento pode ser constante nas ocasiões para um determinado paciente (Trt_j).
- Um modelo hierárquico de três níveis para a resposta é dado por

$$Y_{ijk} = \beta_1 + \beta_2 t_{ijk} + \beta_3 (Trt_j \times t_{ijk}) + b_k^{(3)} + b_{jk}^{(2)} + \epsilon_{ijk}.$$

- Este modelo assume um intercepto comum e separa tendências lineares ao longo do tempo nos dois grupos de tratamento.

Exemplo: modelo de três níveis para o ensaio clínico longitudinal multinível

E se $\text{Var}\,(b_k^{(3)}) = G^{(3)}$, $\text{Var}\,(b_{jk}^{(2)}) = G^{(2)}$ e $\text{Var}\,(\epsilon_{ijk}) = \sigma^2$, e todos os efeitos aleatórios são assumidos como independentes, então

$$\text{Var}\,(Y_{ijk}) = G^{(2)} + G^{(3)} + \sigma^2$$

e a covariância entre duas observações do mesmo paciente é $G^{(2)} + G^{(3)}$.

Exemplo: modelo de três níveis para o ensaio clínico longitudinal multinível

- Assim, as observações para um determinado paciente têm uma estrutura de correlação intraclasse, com

$$\text{Corr}\,(Y_{ijk}, Y_{ijl}) = \frac{G^{(2)} + G^{(3)}}{G^{(2)} + G^{(3)} + \sigma^2}.$$

- Por ser um modelo misto linear,

$$\text{E}\,(Y_{ijk}) = \beta_1 + \beta_2 t_{ijk} + \beta_3 (Trt_{ij} \times t_{ijk}).$$

Estimação

- Para o modelo linear de três níveis, as suposições distribucionais padrão são: $b_k^{(3)} \sim N(0, G^{(3)})$, $b_{jk}^{(2)} \sim N(0, G^{(2)})$, e $\epsilon_{ijk} \sim N(0, \sigma^2)$.
- Dadas essas suposições, a estimação dos parâmetros do modelo é relativamente direta. O estimador MQG é dada por

$$\hat{\beta} = \left\{ \sum_{k=1}^{n_3} (X_k' V_k^{-1} X_k) \right\}^{-1} \sum_{k=1}^{n_3} (X_k' V_k^{-1} X_k),$$

em que Y_k é um vetor coluna de comprimento $\sum_{j=1}^{n_{2k}} n_{1jk}$, o número de observações no k-ésimo agrupamento. X_k é a matriz correspondente de covariáveis e V_k é a matriz de covariância de Y_k.

Estimação

- Como anteriormente, usamos o **método da máxima verossimilhança restrita (REML)**, ou MV, para obter estimativas de $G^{(3)}$, $G^{(2)}$ e σ^2.
- Uma vez obtidas essas estimativas, podemos estimar as matrizes de covariância, V_k, e substituir essas estimativas na expressão do estimador MQG de β.
- Este procedimento de estimativa está disponível nos pacotes `nlme` e `lme4`.

Exemplo: TVSFP

- Embora a prevalência do tabagismo tenha diminuído entre os adultos nas últimas décadas, um número substancial de jovens começa a fumar e se torna viciado em tabaco.
- O *Television, School and Family Smoking Prevention and Cessation Project* (TVSFP) foi um estudo concebido para determinar a eficácia de um currículo escolar de prevenção do tabagismo em conjunto com um programa de prevenção baseado na televisão, em termos de prevenção do início do tabagismo e aumento da cessação do tabagismo.
- O estudo utilizou um desenho fatorial 2×2, com quatro condições de intervenção determinadas pela classificação cruzada de um currículo escolar de resistência social (`CC: codificado 1 = sim, 0 = não`) com um programa de prevenção baseado em televisão (`TV : codificado 1 = sim, 0 = não`).
- A randomização para uma das quatro condições de intervenção ocorreu em nível escolar, enquanto grande parte da intervenção foi realizada em nível de sala de aula.

Exemplo: TVSFP

- O estudo original envolveu 6.695 alunos em 47 escolas no sul da Califórnia.
- Nossa análise se concentra em um subconjunto de 1.600 alunos da sétima série de 135 turmas em 28 escolas de Los Angeles.
- A variável resposta, uma escala de conhecimento sobre tabaco e saúde (THKS), foi administrada antes e após a randomização das escolas para uma das quatro condições de intervenção.
  - A escala avaliou o conhecimento de um estudante sobre tabaco e saúde.

Exemplo: TVSFP

- Consideramos um modelo linear para o escore THKS pós-intervenção, com o escore THKS basal ou pré-intervenção como covariável.
- Esse modelo para a mudança ajustada nos escores do THKS incluiu os principais efeitos de CC e TV e a interação $CC \times TV$.
- Os **efeitos da escola e da sala de aula foram modelados pela incorporação de efeitos aleatórios nos níveis 3 e 2**, respectivamente.

Exemplo: TVSFP

- Deixando Y_{ijk} denotar a pontuação THKS pós-intervenção do i-ésimo aluno dentro da j-ésima sala de aula dentro da k-ésima escola, nosso modelo é dado por

$$Y_{ijk} = \beta_1 + \beta_2 Pre-THKS + \beta_3 CC + \beta_4 TV + \beta_5 CC \times TV + b_k^{(3)} + b_{jk}^{(2)} + \epsilon_{ijk},$$

em que $\epsilon_{ijk} \sim N(0, \sigma_1^2)$, $b_{jk}^{(2)} \sim N(0, \sigma_2^2)$ e $b_k^{(3)} \sim N(0, \sigma_3^2)$.

Exemplo: TVSFP

```
# ------------------------------------------------
# Carregando pacotes do R

library(haven)
library(tidyr)
library(dplyr)
library(nlme)

# ------------------------------------------------
# Carregando o arquivo de dados

thks <- read_dta(file = here::here("data", "tvsfp.dta"))

# ------------------------------------------------
# Formata dados

thks$curriculum <- factor(thks$curriculum)
thks$tvprevent <- factor(thks$tvprevent)
```

Exemplo: TVSFP

```
thks
```

```
## # A tibble: 1,600 x 6
##       sid    cid curriculum tvprevent prescore postscore
##     <dbl>  <dbl> <fct>      <fct>         <dbl>     <dbl>
##  1    403 403101 1          0                 2         3
##  2    403 403101 1          0                 4         4
##  3    403 403101 1          0                 4         3
##  4    403 403101 1          0                 3         4
##  5    403 403101 1          0                 3         4
##  6    403 403101 1          0                 4         3
##  7    403 403101 1          0                 2         2
##  8    403 403101 1          0                 4         4
##  9    403 403101 1          0                 5         5
## 10    403 403101 1          0                 3         4
## # ... with 1,590 more rows
```

Exemplo: TVSFP

```
mod1 <- lme(
  fixed = postscore ~ prescore + curriculum*tvprevent,
            random = ~1 | sid/cid, # (turmas aninhadas em escola
            data = thks,
            na.action = na.omit)
```

Exemplo: TVSFP

- As estimativas REML dos efeitos fixos são exibidas na tabela a seguir.

	E	stimativa	EP	Z
(Intercept)		1.7020	0.1254	13.57
prescore		0.3054	0.0259	11.79
curriculum1		0.6413	0.1609	3.98
tvprevent1		0.1821	0.1572	1.16
curriculum1:tvprevent1		-0.3309	0.2246	-1.47

Exemplo: TVSFP

- As estimativas REML das três fontes de variabilidade indicam que há variabilidade tanto nos níveis de sala de aula quanto na escola, com quase duas vezes mais variabilidade entre as salas de aula dentro de uma escola do que entre as escolas.

```
VarCorr(mod1)
```

```
##                Variance     StdDev
## sid =          pdLogChol(1)
## (Intercept) 0.03864004   0.1965707
## cid =          pdLogChol(1)
## (Intercept) 0.06466148   0.2542862
## Residual    1.60229395   1.2658175
```

Exemplo: TVSFP

- A correlação entre as pontuações do THKS para colegas (ou crianças da mesma sala de aula na mesma escola) é de aproximadamente 0.061 (ou $\frac{0.039+0.065}{0.039+0.065+1.602}$), enquanto a correlação entre as pontuações do THKS para crianças de diferentes salas de aula dentro da mesma escola é de aproximadamente 0.023 (ou $\frac{0.039}{0.039+0.065+1.602}$).

Exemplo: TVSFP

- As estimativas dos efeitos fixos para as condições de intervenção, quando comparadas com seus erros padrão, indicam que:
  - **nem a intervenção na mídia de massa (TV) nem sua interação com o currículo de sala de aula de resistência social (CC) têm impacto nas mudanças ajustadas pelas pontuações THKS da linha de base.**
- Há um efeito significativo do currículo de sala de aula de resistência social, com crianças designadas ao currículo de resistência social mostrando maior conhecimento sobre tabaco e saúde.

Breve conclusão

Políticas educacionais abrangentes são mais eficazes para promover o conhecimento sobre o assunto do que campanhas de mídia de massa, que dependem fortemente da motivação individual. Ver, por exemplo, Rose (2001)[3].

[3] Rose G. Sick individuals and sick populations. Int J Epidemiol. 2001 Jun;30(3):427-32; discussion 433-4. doi: 10.1093/ije/30.3.427. PMID: 11416056.

Exemplo: TVSFP

- A estimativa do efeito principal do CC, no modelo que exclui a interação $CC \times TV$, é de 0.47 ($EP = 0.113$, $p < 0.001$).

```
mod2 <- lme(
  fixed = postscore ~ prescore + curriculum + tvprevent,
            random = ~1 | sid/cid,
            data = thks,
            na.action = na.omit)
```

	Estimativa	EP	Z
(Intercept)	1.7849	0.1129	15.80
prescore	0.3052	0.0259	11.79
curriculum1	0.4715	0.1133	4.16
tvprevent1	0.0196	0.1133	0.17

Exemplo: TVSFP

- As correlações intra-cluster nos níveis de escola e sala de aula são relativamente pequenas.
  - Podemos ficar tentados a considerar isso como uma indicação de que o agrupamento desses dados é irrelevante.
  - **No entanto, tal conclusão seria errônea.**
- Embora as correlações intra-cluster sejam relativamente pequenas, elas têm um impacto substancial nas inferências sobre os efeitos das condições de intervenção.

Exemplo: TVSFP

- Para ilustrar isso, considere o seguinte modelo para as mudanças ajustadas nas pontuações do THKS:

$$Y_{ijk} = \beta_1 + \beta_2 Pre - THKS + \beta_3 CC + \beta_4 TV + \beta_5 CC \times TV + \epsilon_{ijk},$$

em que $\epsilon_{ijk} \sim N(0, \sigma^2)$.

- Este modelo ignora o agrupamento dos dados nos níveis de sala de aula e escola; é um modelo de **regressão linear padrão** e **assume observações independentes** e **variância homogênea**.

Exemplo: TVSFP

```
mod3 <- lm(
  formula = postscore ~ prescore + curriculum*tvprevent,
            data = thks)
```

	Estimativa	EP	Z
(Intercept)	1.6613	0.0844	19.69
prescore	0.3252	0.0259	12.58
curriculum1	0.6406	0.0921	6.95
tvprevent1	0.1987	0.0900	2.21
curriculum1:tvprevent1	-0.3216	0.1303	-2.47

Exemplo: TVSFP

- As estimativas dos efeitos fixos são semelhantes às relatadas no Modelo 1.
- No entanto, os erros padrão baseados no modelo (supondo que não haja agrupamento) são **enganosamente pequenos** para os efeitos da intervenção randomizada e **levam a conclusões substancialmente diferentes** sobre os efeitos das condições de intervenção.

Exemplo: TVSFP

- **Isso destaca uma lição importante:** o impacto do agrupamento depende tanto da magnitude da correlação intra-cluster quanto do tamanho do cluster.
  - Para os dados do TVSFP, os tamanhos dos clusters variam de 1 a 13 salas de aula dentro de uma escola e de 2 a 28 alunos dentro de uma sala de aula.
  - Com tamanhos de cluster relativamente grandes, mesmo uma correlação intra-cluster muito modesta pode ter um impacto discernível nas inferências.

Generalizações

- O modelo multinível pode ser generalizado para um número arbitrário de níveis.
- **Modelos lineares generalizados de efeitos mistos (GLMMs)** também foram desenvolvidos para a análise de desfechos binários e contagens no cenário multinível.

Observações

A modelagem multinível pode ser difícil:

- Uma covariável pode operar em diferentes níveis.
- Nem sempre é claro em como combinar covariáveis em um único modelo.
- Embora modelos lineares hierárquicos com efeitos aleatórios sejam atraentes, a extensão para modelos lineares generalizados levanta problemas difíceis de interpretação.
- Como discutido anteriormente, modelos marginais e modelos de efeitos mistos podem dar resultados bastante diferentes no **cenário não-linear**.

Resumindo

- Apesar de certas complexidades, os modelos multiníveis agora são amplamente utilizados.
- Nos experimentos delineados e nos estudos dos efeitos de fatores familiares/comunitários na saúde, os modelos multiníveis fornecem uma abordagem geralmente eficaz para a análise de dados que explica as correlações induzidas pelo agrupamento.
- Modelos multiníveis não são, em certo sentido, diferentes de modelos longitudinais.
- Diferentemente da regressão logística e da análise de sobrevivência, onde o conceito de análise de regressão pode ser aplicado com bastante robustez e com poucas opções, a análise longitudinal e multinível exige uma reflexão mais cuidadosa sobre a escolha e o significado dos modelos.
- Esse é o desafio e a recompensa deles!

Avisos

- **Próxima aula:** atividade de avaliação 03.
- **Para casa:** ler o Capítulo 22 do livro "**Applied Longitudinal Analysis**".
  - Caso ainda não tenha lido, leia também os Caps. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 e 13.

Bons estudos!