UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DEPARTAMENTO DE ESTATÍSTICA
MAT02035 - Modelos para dados correlacionados

Avaliação 01

Exercício 1. (1 ponto) Apresente as principais características de um estudo longitudinal.

Em um estudo longitudinal as medidas dos mesmos indivíduos são tomadas repetidamente através do tempo, permitindo, assim, o estudo direto da mudança da variável resposta ao longo do tempo. Os dados longitudinais podem ser considerados um caso especial de dados agrupados. Estes fornecem observações dentro de um grupo: um indivíduo representa um agrupamento em um estudo longitudinal; uma família/domicílio representa um agrupamento. Tipicamente, dados agrupados irão exibir correlação positiva, e esta correlação deve ser levada em conta na análise.

Outra característica presente em estudos longitudinais é a perda de dados. Em estudos com muitos participantes e/ou com um longo período de acompanhamento é comum que algumas medidas não sejam realizadas. A perda de uma visita (ao centro de investigação), ou o acompanhamento (indivíduo não deseja mais participar do estudo) são alguns dos motivos para a perda de informação.

Exercício 2. (1 ponto) Explique por que dados longitudinais podem ser vistos como dados agrupados. Dê um exemplo de outro tipo de estudo que pode fornecer dados agrupados.

As medidas repetidas ao longo do tempo em um mesmo indivíduo podem ser vistas como um agrupamento das observações ao nível de indivíduo.

Um **estudo aleatorizado por *clusters*** é outro exemplo de uma estrutura de dados agrupados. As observações são aleatorizadas em grupos. Por exemplo, turmas de alunos que recebem um tratamento (ensino com apoio de dispositivo eletrônico) vs. turmas de alunos que recebem o tratamento controle (ensino tradicional). Neste exemplo, as turmas são aleatorizadas nos grupos de tratamentos, e cada turma representa um agrupamento de observções (no caso, as respostas dos alunos).

Exercício 3. (1 ponto) Descreva as diferenças (incluindo possíveis vantagens e desvantagens) entre estudos longitudinais e estudos transversais.

Em um estudo longitudinal, em que medidas repetidas realizadas nos indivíduos, é possível capturar a mudança dentro de indivíduo. Já em um estudo transversal, em que a resposta é medida em apenas uma ocasião, é possível apenas estimar diferenças nas respostas entre indivíduos. Ou seja, os estudos transversais permitem comparações entre subpopulações que

diferem em idade, mas não fornecem qualquer informação a respeito de como os indivíduos mudam durante o correspondente período.

Exercício 4. (1 ponto) Para um estudo longitudinal, descreva as diferenças entre os delineamentos balanceado e desbalanceado. Dê um exemplo para cada um dos delineamentos.

Quando o número e o momento das medidas repetidas são os mesmos para todos os indivíduos do estudo, dizemos que o estudo segue um delineamento balanceado. Em estudos com delineamento balanceado, as ocasiões de medição podem ser igualmente ou desigualmente distribuídas ao longo da duração do estudo. Quando a sequência dos tempos de observação não é mais comum a todos os indivíduos no estudo, dizemos que o estudo possui um delinemanto desabalanceado. Isto geralmente ocorre em alguns estudos longitudinais, especialmente aqueles em que as medidas repetidas se estendem por um período relativamente longo, e alguns indivíduos perdem a sua visita programada ou data de observação.

Exercício 5. (1 ponto) Quais as consequências de se ignorar (na análise) a correlação presente entre os dados de medidas repetidas?

Ao ignorar a correlação entre as medidas repetidas irá, em geral, resultar em estimativas incorretas da variabilidade amostral, que levam a inferências bastante enganosas. Ou seja, a análise (incorreta) implicará em erros padrões muito grandes, intervalos de confiança muito largos, e valores p para testes de hipóteses para avaliar a mudança na resposta média muito grandes.

Exercício 6. (2 pontos) O estudo *Treatment of Lead-Exposed Children* (TLC) foi um estudo aleatorizado e controlado por placebo de *succimer* (um agente quelante) em crianças com níveis de chumbo no sangue de 20 a 44 microgramas/dL. Lembre-se de que 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. Para este conjunto de problemas, focamos apenas nas 50 crianças designadas para tratamento de quelação com *succimer*.

Os dados brutos são armazenados em um arquivo externo: `lead.dta`. Cada linha do conjunto de dados contém as 5 variáveis a seguir: `ID`, `Y1`, `Y2`, `Y3`, `Y4`.

a. Leia os dados do arquivo externo e calcule as médias da amostra, os desvios padrão e as variâncias dos níveis de chumbo no sangue em cada ocasião.

Lendo, formatando e filtrando os dados.

```
# Carrega pacotes
library(haven)
library(tidyr)
library(dplyr)

# Carrega dados
chumbo <- read_dta(file = here::here("data","tlc.dta"))
```

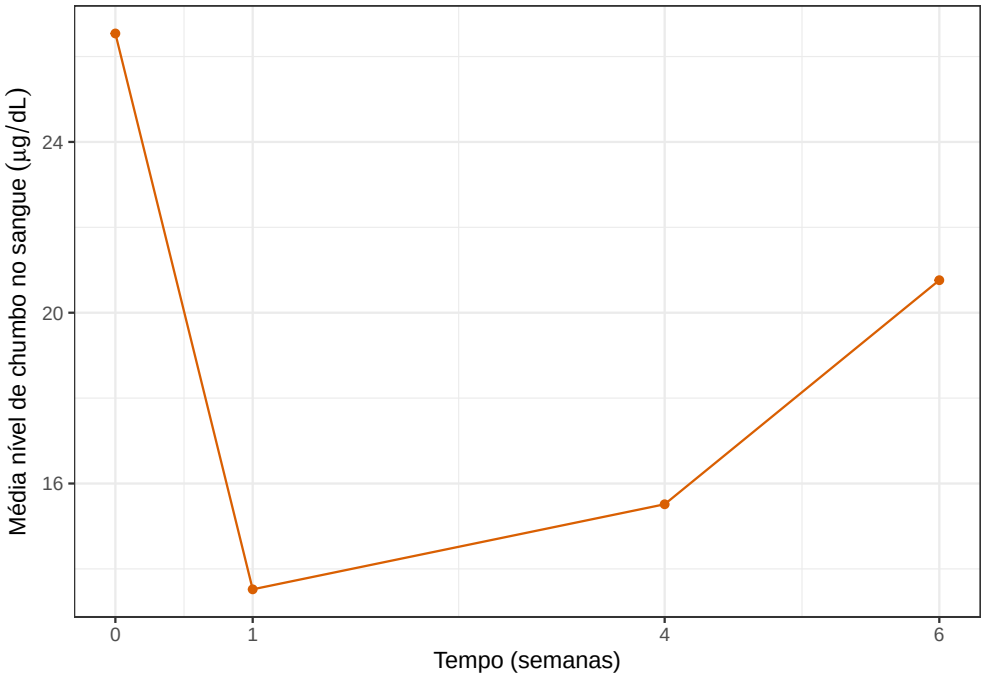
```r
# Largo para longo
chumbo.longo <- gather(data = chumbo,
                       key = "tempo",
                       value = "chumbo", -id, -trt)

# Transforma dados
chumbo.longo$tempo <- as.numeric(
  as.character(factor(chumbo.longo$tempo,
                      labels = c(0, 1, 4, 6))))

chumbo.longo$trt <- factor(chumbo.longo$trt,
                           labels = c("Placebo", "Succimer"))

# Filtra dados do grupo succimer
chumbo.longo <- chumbo.longo %>%
  filter(trt == "Succimer")
```

A seguir são apresentados os desvios-padrões e variâncias dos níveis de chumbo no sangue em cada ocasião para os indivíduos do grupo *Succimer*.

Tempo	Média	Desvio Padrão	Variância
0	26.5	5.0	25.2
1	13.5	7.7	58.9
4	15.5	7.9	61.7
6	20.8	9.2	85.5

b. Construa um gráfico de tempo dos níveis médios de chumbo no sangue versus tempo (em semanas). Descreva as características gerais da tendência temporal.

O gráfico de perfis médios é apresentado a seguir.



É possível perceber um decaimento acentuado no nível médio de chumbo no sangue logo após o começo do estudo. No entanto, após a semana 1, os níveis médios de chumbo no sangue volta a subir, embora não atinjam o mesmo nível apresentado na linha de base.

c. Calcule as matrizes de covariância e correlação 4×4 para as quatro medidas repetidas dos níveis de chumbo no sangue.

A matriz de covariância das medidas repetidas dos níveis de chumbo no sangue é apresentada a seguir.

	y0	y1	y4	y6
y0	25.21	15.47	15.14	22.99
y1	15.47	58.87	44.03	35.97
y4	15.14	44.03	61.66	33.02
y6	22.99	35.97	33.02	85.49

E a seguir, é apresentada a matriz de correlação das medidas repetidas dos níveis de chumbo no sangue é apresentada a seguir.

	y0	y1	y4	y6
y0	1.00	0.40	0.38	0.50
y1	0.40	1.00	0.73	0.51
y4	0.38	0.73	1.00	0.45
y6	0.50	0.51	0.45	1.00

d. Verifique se os elementos diagonais da matriz de covariâncias são as variâncias comparando com a estatística descritiva obtida no item (a).

Como sabemos, a $\text{Cov}(Y_j, Y_j) = \text{Var}(Y_j)$. Assim, como era de se esperar, verifica-se que os elementos diagonais da matriz de covariâncias (25.2, 58.9, 61.7, 85.5) são as variâncias das medidas em cada ocasião de tempo apresentadas no item (a).

Exercício 7. (3 pontos) Seja $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})'$ um vetor de respostas de medidas repetidas, X_i uma matriz $n_i \times p$ de covariáveis associadas ao vetor de respostas Y_i, para $i = 1, \ldots, N$, em que N é o número de indivíduos em um certo estudo. Considere o seguinte modelo:

$$Y_i = X_i\beta + e_i, i = 1, \ldots, N,$$

em que $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ é um vetor de coeficientes de regressão desconhecidos, e $e_i = (e_{i1}, e_{i2}, \ldots, e_{in_i})'$ é um vetor de erros aleatórios. Ainda, suponha que $e_i \sim N_{n_i}(0, \Sigma_i)$, em que Σ_i é uma matriz de covariâncias.

a. Qual a distribuição condicional de $Y_i|X_i$?

Pela propriedade da distribuição normal, temos:

$$Y_i|X_i \sim N_{n_i}(X_i\beta, \Sigma_i).$$

b. Escreva a função de verossimilhança de β (considere Σ_i conhecido).

$$L(\beta) = \prod_{i=1}^{N} f(y_i),$$

em que

$$f(y_i) = (2\pi)^{-n_i/2}|\Sigma_i|^{-1/2} \exp\left\{-\frac{1}{2}(y_i - \mu_i)'\Sigma_i^{-1}(y_i - \mu_i)\right\}.$$

Logo,

$$L(\beta) = \left\{\prod_{i=1}^{N} (2\pi)^{-n_i/2}|\Sigma_i|^{-1/2}\right\} \times \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\left[(y_i - X_i\beta)'\Sigma_i^{-1}(y_i - X_i\beta)\right]\right\}.$$

c. Utilizando o método da máxima verossimilhança, encontre a expressão do estimador de β (considere Σ_i conhecido). (Apresente o desenvolvimento)

Da função de log-verossimilhança, temos que:

$$\ell = -\frac{K}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{N}\log|\Sigma_i| - \frac{1}{2}\left\{\sum_{i=1}^{N}(y_i - X_i\beta)'\Sigma_i^{-1}(y_i - X_i\beta)\right\},$$

em que $K = \sum_{i=1}^{N} n_i$ é o **número total de observações**. Note que maximizar a log-verossimilhança em relação a β é equivalente a minimizar

$$\sum_{i=1}^{N} (y_i - X_i\beta)'\Sigma_i^{-1}(y_i - X_i\beta).$$

$$\begin{aligned}
&\frac{\partial}{\partial\beta}\sum_{i=1}^{N} (y_i - X_i\beta)'\Sigma_i^{-1}(y_i - X_i\beta) = 0 \\
&\Rightarrow \sum_{i=1}^{N} (X_i'\Sigma_i^{-1}X_i)\hat{\beta} = \sum_{i=1}^{N} (X_i'\Sigma_i^{-1}y_i) \\
&\Rightarrow \hat{\beta} = \left\{\sum_{i=1}^{N} (X_i'\Sigma_i^{-1}X_i)\right\}^{-1} \sum_{i=1}^{N} (X_i'\Sigma_i^{-1}y_i).
\end{aligned}$$

d. Demonstre que o estimador de máxima verossimilhança $\widehat{\beta}$ é não-enviesado para β. Em palavras, explique as implicações práticas deste resultado.

$$\begin{aligned}
\mathrm{E}[\hat{\beta}] &= \mathrm{E}\left[\left\{\sum_{i=1}^{N} (X_i'\Sigma_i^{-1}X_i)\right\}^{-1} \sum_{i=1}^{N} (X_i'\Sigma_i^{-1}Y_i)\right] \\
&= \left\{\sum_{i=1}^{N} (X_i'\Sigma_i^{-1}X_i)\right\}^{-1} \sum_{i=1}^{N} (X_i'\Sigma_i^{-1}\mathrm{E}[Y_i]) \\
&= \left\{\sum_{i=1}^{N} (X_i'\Sigma_i^{-1}X_i)\right\}^{-1} \sum_{i=1}^{N} (X_i'\Sigma_i^{-1}X_i)\beta \\
&= I_p\beta = \beta.
\end{aligned}$$

Diferentes amostras, de mesmo tamanho, considerando que o modelo especificado é modelo correto, produzem diferentes estimativas (as estimativas variam). No entanto, a média das estimativas é igual ao parâmetro ("em média acertam o alvo").

e. Suponha que este modelo foi ajustado para um certo conjunto de dados, considerando três covariáveis (portanto, $\beta = (\beta_1, \beta_2, \beta_3)'$) e que $\widehat{\beta}_2 = 1.4$. A matriz de covariância estimada de $\widehat{\beta}$ é apresentada a seguir:

$$\widehat{\mathrm{Cov}}(\widehat{\beta}) = \begin{bmatrix} 0.50 & -0.50 & -0.12 \\ -0.50 & 1.01 & 0.12 \\ -0.12 & 0.12 & 0.63 \end{bmatrix}.$$

Apresente um intervalo de confiança (IC) de 95% para β_2. Justifique este método de construção do IC.

Pelas propriedades dos estimadores de máxima verossimilhança do modelo linear geral para dos longitudinais, temos que um IC de 95% para β_2 pode ser obtido da seguinte forma

$$1.4 \pm 1,96\sqrt{1.01} = (-0.57,\ 3.37),$$

em que $\widehat{\text{Var}}(\hat{\beta}_2) = 1.01$.