INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
(IME/UFRGS)

Lista 1 Modelos para Dados Correlacionados (2019/2)

Autores: José Angelo Rosa Bastos (278049) e Vítor Coutinho Borges (278056)

Porto Alegre
08 de Outubro de 2019

Lista 1 Modelos para Dados Correlacionados (2019/2)

Autores: José Angelo Rosa Bastos (278049) e Vítor Coutinho Borges (278056)

Relatório apresentado junto ao Curso de Estatística da
Universidade Federal do Rio Grande do Sul
como requisito parcial à obtenção do título de Bacharel em Estatística.

Orientador: Prof. Rodrigo Citton dos Reis

Porto Alegre
08 de Outubro de 2019

Sumário:

Questões

1. Apresente exemplos de dados correlacionados.
2. Apresente as principais características de um estudo longitudinal.
3. Apresente os principais objetivos em uma análise de dados longitudinais.
4. Para um estudo longitudinal, descreva as diferenças entre os delineamentos balanceado e desbalanceado.
5. Considere a seção "Propriedades dos Valores Esperados e Variâncias", dos *slides* da **Aula 02**. Dadas as condições no *slide* 44, demonstre as propriedades de **1 a 5** para o valor esperado (*slide* 46) e as propriedades **1 a 5** da variância (*slide* 47).
6. Considere $Y_1, ..., Y_N$ variáveis aleatórias **i.i.d.** com média μ e variância σ^2. Considere o seguinte estimador para a média: $\hat{\mu} = \bar{Y}_N = \frac{1}{N}\sum_{i=1}^{N} Y_i$. Encontre o erro padrão de $\hat{\mu}$. O que esta medida representa?
7. Com o auxílio do computador, faça os exercícios do **Capítulo 2** do livro "*Applied Longitudinal Analysis*" (páginas **44 e 45**):

   I. Ler a base e calcular as médias, desvios-padrão e variâncias do nível de chumbo no sangue em cada ocasião.
   II. Construir um gráfico temporal da média do nível do sangue no tempo *versus* o tempo (em semanas). Descrever as características gerais da tendência temporal.
   III. Calcule as matrizes de covariância e correlação 4×4 para as quatro medidas repetidas de nível de chumbo no sangue.
   IV. Verifique que os elementos da diagonal da matriz de covariância são as variâncias, comaprando com as estatísticas descritivas obtidas no item **I**.
   V. Verifique que a **correlação** entre o nível de chumbo no sangue na linha de base (**semana 0**) e na **semana 1** é igual à **covariância** entre o nível de chumbo no sangue na linha de base e na semana 1 dividida pelo produto dos desvios-padrão do nível de chumbo no sangue na linha de base e na semana 1.
8. Considere o modelo de regressão linear para o vetor de respostas médias $\mathbb{E}(Y_i|X_i) = X_i\beta_i,\ i = 1, ..., N$, em que Y_i é um vetor $n_i \times 1$ de respostas, X_i é uma matriz $n_i \times 1$ de covariáveis, β é um vetor $\rho \times 1$ de coeficientes de regressão desconhecidos e N é o número de indivíduos observados. Ainda, assuma que Y_i tem uma distribuição condicional normal multivariada com média dada por $\mathbb{E}(Y_i|X_i)$ e matriz de covariância $Cov(Y_i|X_i) = \sum_i (n_i \times n_i)$. Suponha $\sum_i$ desconhecida.
   a) Escreva a função de verossimilhança do modelo.
   b) Considere o estimador de mínimos quadrados generalizados $\hat{\beta} = \{\sum_{i=1}^{N}(X_i' \sum^{-1} X_i)\}^{-1} \sum_{i=1}^{N}(X_i' \sum^{-1} y_i)$.
      I. Justifique o fato de que, neste caso, este é o estimador de máxima verossimilhança.
      II. Demonstre que $\mathbb{E}(\hat{\beta}) = \beta$.
      III. Demonstre que $Cov(\hat{\beta}) = \{\sum_{i=1}^{N}(X_i' \sum^{-1} X_i)\}^{-1}$.
      IV. Considere o caso em que $\sum_i = \sigma^2 I$, onde I é a matriz identidade.
      Demonstre que $\hat{\beta}$ é reduzido ao estimador de mínimos quadrados ordinários.
9. Considere os dados do estudo dos níveis de chumbo no sangue (TLC). Proponha um modelo de regressão linear para a média com base nas questões de pesquisa. Com o auxílio do computador, **utilize gráficos** para justificar a sua proposta de modelo.

Respostas

Questão 1:

Observações de um mesmo indivíduo ao longo do tempo (característica definidora de estudos longitudinais) são um bom exemplo. Alguns outros exemplos seriam o número de vendas mensais de algum produto ou o movimento de uma partícula com relação a algum referencial.

Questão 2:

Em estudos longitudinais as medidas dos indivíduos na amostra são tomadas repetidas vezes ao longo do tempo, o que faz com que os dados sejam **agrupados** (cada indivíduo é um "grupo" ao longo do tempo, cujas medições são ordenadas naturalmente), sendo que observações dentro de um mesmo grupo terão **correlação positiva**.

Questão 3:

Em contraste com os estudos **transversais**, onde deseja-se analisar as diferenças **entre** os indivíduos amostrados com relação a alguma característica de interesse, nos estudos **longitudinais** tem-se como principal objetivo analisar as diferenças **dentro** dos indivíduos com relação à característica de interesse ao longo do tempo. Então, pode-se analisar o efeito do tempo na característica de interesse na amostra observada.

Questão 4:

Um estudo longitudinal tem delineamento dito **balanceado** quando todos os grupos (indivíduos) têm o mesmo número de medições e estas são tomadas nos mesmos períodos de tempo. Se algum indivíduo na amostra possuir mais ou menos medições que os outros ou dois ou mais indivíduos possuírem medidas tomadas em momentos diferentes, o delineamento do estudo é dito **desbalanceado**.

Questão 5:

Propriedades para o **valor esperado** de Y:

1. $\mathbb{E}(a) = \int_{-\infty}^{+\infty} y\, f_Y(y)\, dy = a \int_{-\infty}^{+\infty} \mathbb{P}(Y = a)\, dy = a\mathbb{P}(Y = a) = 1a = a$
2. $\mathbb{E}(bY) = \int_{-\infty}^{+\infty} by\, f_Y(y)\, dy = b \int_{-\infty}^{+\infty} y\, f_Y(y)\, dy = b\mathbb{E}(Y)$
3. $\mathbb{E}(a + bX) = \int_{-\infty}^{+\infty} a + by\, f_Y(y)\, dy = \int_{-\infty}^{+\infty} ady + \int_{-\infty}^{+\infty} by\, f_Y(y)\, dy = a + \int_{-\infty}^{+\infty} by\, f_Y(y)\, dy = a + b\mathbb{E}(Y)$
4. $\mathbb{E}(aX+bY) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (ax+by) f(x,y) dy dx = \int_{-\infty}^{+\infty} ax(\int_{-\infty}^{+\infty} f(x,y)dy)dx + \int_{-\infty}^{+\infty} by(\int_{-\infty}^{+\infty} f(x,y)dy)dx = \int_{-\infty}^{+\infty} axf_X(x,y)dx + \int_{-\infty}^{+\infty} byf_Y(x,y)dy = a\mathbb{E}(X) + b\mathbb{E}(Y)$
5. Para duas variáveis aleatórias X e Y independentes, temos que $f(x,y) = f_X(x)f_Y(y)$, portanto podemos obter $\mathbb{E}(XY) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xyf(x,y)dxdy = \int_{-\infty}^{+\infty} xf_X(x)(\int_{-\infty}^{+\infty} yf_Y(y)dy)dx = (\int_{-\infty}^{+\infty} yf_Y(y)dy)(\int_{-\infty}^{+\infty} xf_X(x)) = \mathbb{E}(X)\mathbb{E}(Y)$. Se X e Y não forem independentes, porém, $f(x,y) \neq f_X(x)f_Y(y)$ e, portanto, $\mathbb{E}(XY) \neq \mathbb{E}(X)\mathbb{E}(Y)$.

Propriedades para a **variância** de Y:

1. $Var(a) = \mathbb{E}\{a - \mathbb{E}(a)\}^2 = \mathbb{E}(a - a)^2 = 0^2 = 0$
2. $Var(bY) = \mathbb{E}\{bY - b\mathbb{E}(Y)\}^2 = \mathbb{E}\{b[Y - \mathbb{E}(Y)]\}^2 = b^2\mathbb{E}\{Y - \mathbb{E}(Y)\} = b^2Var(Y)$
3. $Var(a + bY) = \mathbb{E}\{a + bY - a - b\mathbb{E}(Y)\}^2 = \mathbb{E}\{bY - b\mathbb{E}(Y)\}^2 = \mathbb{E}\{b[Y - \mathbb{E}(Y)]\}^2 = b^2\mathbb{E}\{Y - \mathbb{E}(Y)\} = b^2Var(Y)$
4. $Var(aX + bY) = \mathbb{E}\{aX + bY - \mathbb{E}(aX + bY)\}^2 = \mathbb{E}\{aX + bY - a\mathbb{E}(X) - b\mathbb{E}(Y)\}^2 = \mathbb{E}\{a[X - \mathbb{E}(X)] + b[Y - \mathbb{E}(Y)]\}^2 = \mathbb{E}\{a^2[X - \mathbb{E}(X)]^2 + b^2[Y - \mathbb{E}(Y)]^2 + 2ab[X - \mathbb{E}(X)][Y - \mathbb{E}(Y)]\} = a^2\mathbb{E}[X - \mathbb{E}(X)]^2 + b^2\mathbb{E}[Y - \mathbb{E}(Y)]^2 + 2ab\mathbb{E}\{[X - \mathbb{E}(X)][Y - \mathbb{E}(Y)]\} = a^2Var(X) + b^2Var(Y) + 2abCov(X,Y)$

5. $Var(aX - bY) = \mathbb{E}\{aX - bY - \mathbb{E}(aX - bY)\}^2 = \mathbb{E}\{aX - bY - a\mathbb{E}(X) + b\mathbb{E}(Y)\}^2 = \mathbb{E}\{a[X - \mathbb{E}(X)] - b[Y - \mathbb{E}(Y)]\}^2 = \mathbb{E}\{a^2[X - \mathbb{E}(X)]^2 + b^2[Y - \mathbb{E}(Y)]^2 - 2ab[X - \mathbb{E}(X)][Y - \mathbb{E}(Y)]\} = a^2\mathbb{E}[X - \mathbb{E}(X)]^2 + b^2\mathbb{E}[Y - \mathbb{E}(Y)]^2 - 2ab\mathbb{E}\{[X - \mathbb{E}(X)][Y - \mathbb{E}(Y)]\} = a^2Var(X) + b^2Var(Y) - 2abCov(X, Y)$

Questão 6:

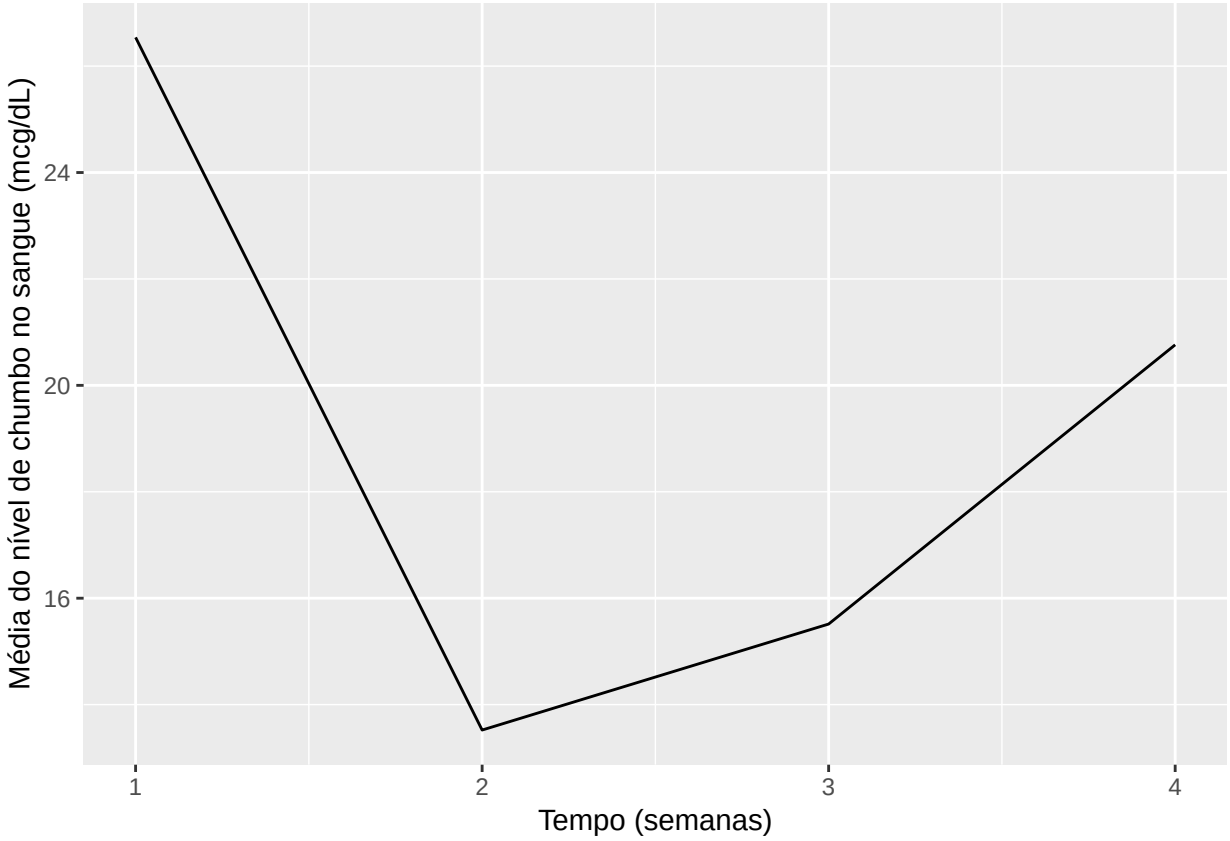
O Erro Padrão representa a variação de uma média amostral com relação à media populacional.
Para o estimador $\hat{\mu}$ descrito no enunciado da questão, a fórmula para o Erro Padrão fica da seguinte forma: $EP(Y) = \frac{\sqrt{\frac{\sum_{i=1}^{N}\{Y_i - \hat{\mu}\}^2}{N-1}}}{\sqrt{N}} = \frac{\sqrt{\frac{\sum_{i=1}^{N}\{Y_i - \frac{1}{N}\sum_{j=1}^{N} Y_j\}^2}{N-1}}}{\sqrt{N}}$, onde N é o tamanho amostral.

Questão 7:

I. Descritivas da base:

```
##     Estatistica Linha_de_Base  Semana_1  Semana_2  Semana_3
## 1         Média     26.540000 13.522000 15.514000 20.762000
## 2 Desvio-Padrão      5.020936  7.672487  7.852207  9.246332
## 3     Variância      2.240744  2.769925  2.802179  3.040778
```

II. Gráfico dos níveis de chumbo pelo tempo:



III. Matriz de covariância:

```
##           y1       y2       y3       y4
## y1 25.20979 15.46543 15.13800 22.98543
## y2 15.46543 58.86706 44.02907 35.96595
## y3 15.13800 44.02907 61.65715 33.02197
## y4 22.98543 35.96595 33.02197 85.49465
```

Matriz de correlação:

```
##           y1        y2        y3        y4
## y1 1.0000000 0.4014589 0.3839654 0.4951063
## y2 0.4014589 1.0000000 0.7308221 0.5069743
## y3 0.3839654 0.7308221 1.0000000 0.4548224
## y4 0.4951063 0.5069743 0.4548224 1.0000000
```

Questão 8:

Questão 9: