CE075 - Análise de Dados Longitudinais

Silva, J.L.P.

21 de agosto, 2019

Efeitos transversais e longitudinais

Notação para Dados Longitudinais

Notação (Estrutura Balanceada)

$$\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in})', \quad i = 1, \ldots, N,$$

é o vetor de respostas do i-ésimo indivíduo.

- N: número de indivíduos;
- Número total de observações: Nn;
- $E(Y_i) = (E(Y_{i1}), \ldots, E(Y_{in}))'$;
- $\mu_{ij} = E(Y_{ij})$;
- σ_j^2: variância de Y_{ij};
- σ_{jk}: covariância entre Y_{ij} e Y_{ik}.

Estudos Transversais vs Longitudinais

Vetor de observações longitudinais para o i-ésimo indivíduo:

$$\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in})'$$

- No tempo inicial (linha de base, $j = 1$) foram selecionados indivíduos com diferentes idades.
- Os indivíduos foram acompanhados longitudinalmente.
- Desta forma temos duas fontes da variação da resposta com a idade (transversal e longitudinal).

Qual é a diferença dos efeitos?

- Efeito transversal: variação entre indivíduos. Variação da resposta média em função das idades dos indivíduos medida no tempo inicial.
- Efeito longitudinal: variação intra-indivíduo. Variação da resposta média em função da idade no mesmo indivíduo.
- O efeito de idade em um estudo transversal pode estar potencialmente confundido com efeito de coorte.

Estudos Transversais vs Longitudinais

Estudo Transversal (sem intercepto): $j = 1$

$$Y_{i1} = \beta_T x_{i1} + \varepsilon_{i1} \quad i = 1, \ldots, N$$

ou

$$E(Y_{i1}) = \beta_T x_{i1} \quad i = 1, \ldots, N$$

β_T representa a diferença da resposta média entre duas sub-populações que diferem por uma unidade em x.

Se x é a idade, representa o aumento (diminuição) na média de Y para cada incremento de um ano na idade.

Estudos Transversais vs Longitudinais

Estudo Longitudinal: a resposta média aumenta linearmente com mudanças na idade no mesmo indivíduo:

$$E(Y_{ij} - Y_{i1}) = \beta_L(x_{ij} - x_{i1}),$$

β_L representa a mudança esperada em Y para a mudança em uma unidade em x.

Modelo Linear com componentes transversais e longitudinais:[1]

$$E(Y_{ij}) = \beta_T x_{i1} + \beta_L(x_{ij} - x_{i1}).$$

[1] É necessário assumir $\beta_L = \beta_T$ para estimar mudança da resposta no tempo em estudos transversais

Exemplo: Fitzmaurice e colegas (2011, pag. 253)

- Três coortes de crianças com idades iniciais: 5, 6 e 7 anos.
- A resposta foi medida na linha de base e seguida por três anos.
- Suponha que o efeito transversal é linear:

$$E(Y_{i1}) = 0,75 \; x \; \text{idade}_{i1}$$

  e que esta relação também vale para $j = 2, 3, 4$.
- Suponha que a resposta média também cresce linearmente com as mudanças na idade em cada coorte. Ou seja,

$$E(Y_{ij} - Y_{i1}) = 0,25 \; x \; (\text{idade}_{ij} - \text{idade}_{i1})$$

Exemplo: Estudos Transversais vs Longitudinais

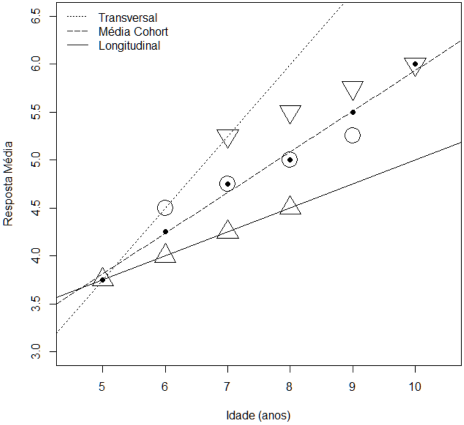

Figura 1: Resposta Média: transversal vs longitudinal. Transversal: 5,6 e 7 anos. Longitudinal: seguimento por 3 anos. $\beta_T = 0,75$ e $\beta_L = 0,25$.

Exemplo: Estudos Transversais vs Longitudinais

- Diferença grande entre os efeitos transversal (linha pontilhada) e longitudinal (linha sólida).

- Efeito de coorte introduz vício na estimativa transversal quando o efeito longitudinal é ignorado.

- Neste caso, o efeito medido é uma combinação ponderada entre β_L e β_T. Ou seja,

$$\hat{\beta} = (1 - w)\hat{\beta}_L + w\hat{\beta}_T,$$

em que w depende da proporção de variabilidade (intra e entre indivíduos) e correlação entre as observações intra indivíduo.

Exemplo: Estudos Transversais vs Longitudinais

```r
dados <- data.frame(
coorte=c(5,5,5,5,6,6,6,6,7,7,7,7),
idade=c(5,6,7,8,6,7,8,9,7,8,9,10),
tempo=c(1:4,1:4,1:4),
resp=c(3.75,4.00,4.25,4.50,4.50,4.75,5.00,5.25,5.25,5.50,
       5.75,6.00))
head(dados)
```

```
  coorte idade tempo resp
1      5     5     1 3.75
2      5     6     2 4.00
3      5     7     3 4.25
4      5     8     4 4.50
5      6     6     1 4.50
6      6     7     2 4.75
```

Exemplo: Estudos Transversais vs Longitudinais

```
library(ggplot2); library(ggpubr)
```

```
Loading required package: magrittr
```

```
dados$coorte <- as.factor(dados$coorte)
p1 <- ggplot(dados,aes(x=idade,y=resp,shape=coorte,color=coorte)) +
      geom_point() + theme_bw() + geom_point(size=3) +
      theme(legend.position="top")
p2 <- p1 + geom_line(linetype="dashed")
p3 <- p1 + geom_line(data=subset(dados,tempo==1),aes(group=tempo),
      colour="black") + geom_line(data=subset(dados,tempo==2),
      aes(group=tempo),colour="black") + geom_line(data=
      subset(dados,tempo==3),aes(group=tempo),colour="black") +
      geom_line(data=subset(dados,tempo==4), aes(group=tempo),
      colour="black")
```

Exemplo: Estudos Transversais vs Longitudinais

```
ggarrange(p1,p2,p3,ncol=3, common.legend=TRUE, labels="auto")
```

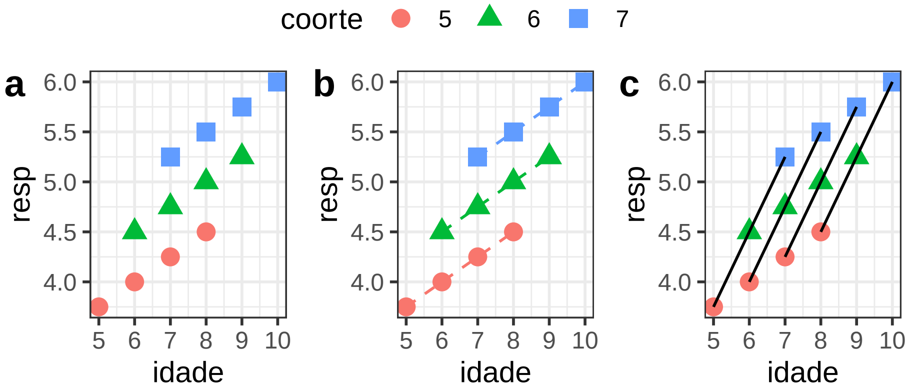

Figura 2: (a) dados combinados, (b) efeito longitudinal, (c) efeito transversal

Exemplo: Estudos Transversais vs Longitudinais

```r
### Efeito transversal
round(lm(resp~idade,data=subset(dados,tempo==1))$coef[2],2)
```

```
idade
 0.75
```

```r
round(lm(resp~idade,data=subset(dados,tempo==2))$coef[2],2)
```

```
idade
 0.75
```

```r
round(lm(resp~idade,data=subset(dados,tempo==3))$coef[2],2)
```

```
idade
 0.75
```

```r
round(lm(resp~idade,data=subset(dados,tempo==4))$coef[2],2)
```

```
idade
 0.75
```

Exemplo: Estudos Transversais vs Longitudinais

```r
### Efeito longitudinal
lm(resp~idade,data=subset(dados,coorte==5))$coef[2]
```

```
idade
 0.25
```

```r
lm(resp~idade,data=subset(dados,coorte==6))$coef[2]
```

```
idade
 0.25
```

```r
lm(resp~idade,data=subset(dados,coorte==7))$coef[2]
```

```
idade
 0.25
```

Perspectiva histórica

Perspectiva Histórica

Há dois enfoques clássicos para análise de dados longitudinais: ANOVA de medidas repetidas e ANOVA multivariada (MANOVA).

Ambos assumem respostas contínuas e erros normalmente distribuídos que são homogêneos entre os grupos. Em alguns casos, a normalidade pode ser obtida através de transformações.

Para ambos, o foco principal é a comparação de médias de grupos, e nenhum dos modelos é informativo quanto às curvas de crescimento individual (i.e., tendências específicas do indivíduo).

Perspectiva Histórica

Além disso, os pontos no tempo são considerados fixos e são tratados como uma variável de classificação no modelo ANOVA ou MANOVA.

Isso exclui a análise de desenhos desbalanceados em que diferentes indivíduos são medidos em diferentes ocasiões.

Ambos os modelos são baseados em estimação via mínimos quadrados e são afetados por *outliers* e dados ausentes.

Enquanto a ANOVA possa lidar com dados ausentes (i.e., há métodos para desenhos desbalanceados), a MANOVA não admite quaisquer dados ausentes.

Perspectiva Histórica

Em termos da estrutura de variância-covariância para as respostas $\boldsymbol{Y}_i$, enquanto o modelo ANOVA assume simetria composta (i.e., variâncias e covariâncias iguais no tempo), a MANOVA não faz esta suposição.

Esta vantagem da MANOVA sobre a ANOVA é amenizada pela limitação maior de requerer dados completos para todos os indivíduos.

A aplicação da MANOVA deve seguir à exclusão de todos os indivíduos sem dados completos e está propensa a viés substancial no sentido de que os indivíduos que completam o estudo podem ser muito diferentes daqueles indivíduos no tempo da aleatorização.

Uma Amostra: ANOVA com Blocos Aleatorizados

Não temos intervenção ou efeitos de grupos, apenas usamos o modelo para caracterizar mudanças no tempo.

Com $i = 1, \ldots, N$ indivíduos e $j = 1, \ldots, n$ ocasiões, a ANOVA é dada pelo modelo linear:

$$Y_{ij} = \mu + \alpha_i + \tau_j + \varepsilon_{ij},$$

em que

- μ é a média geral;
- α_i é o componente de efeito para o indivíduo i;
- τ_j é o efeito do tempo, assumido igual para todos os indivíduos;
- ε_{ij} é o termo de erro para o indivíduo i na ocasião j.

Uma Amostra: ANOVA com Blocos Aleatorizados

Assumimos:

- $\alpha_i \sim N(0, \sigma^2_\alpha)$, em que σ^2_α é a variância entre indivíduos;
- $\varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon)$, em que σ^2_ε é a variância intra-indivíduo.

Note que este é um modelo misto porque inclui tanto parâmetros aleatórios (α_i) quanto fixos (τ_j).

Observações:

- Este desenho é similar a um delineamento em blocos aleatorizados em que os indivíduos são os blocos.
- No caso simples em que $n = 2$, o desenho e a análise são idênticos a um teste t pareado, em termos de testar o efeito do tempo.

Uma Amostra: ANOVA com Blocos Aleatorizados

Tabela 1: Representação dos dados

	Tempo			
Indivíduo	1	2	$\cdots$	n
1	y_{11}	y_{12}	$\cdots$	y_{1n}
2	y_{21}	y_{21}	$\cdots$	y_{2n}
.	.	.	...	.
.	.	.	...	.
N	y_{N1}	y_{N2}	$\cdots$	y_{Nn}

Uma Amostra: ANOVA com Blocos Aleatorizados

Em termos de suposições do modelo, assumimos:

$$
\begin{aligned}
\sum_{j=1}^{n} \alpha_j =& 0 \\
E(Y_{ij}) =& \mu + \tau_j \\
V(Y_{ij}) =& V(\mu + \tau_j + \alpha_i + \varepsilon) = \sigma_\alpha^2 + \sigma_\varepsilon^2 \\
Cov(Y_{ij}, Y_{i',j}) =& 0, \ \forall i \neq i' \\
Cov(Y_{ij}, Y_{i,j'}) =& \sigma_\alpha^2, \ \forall j \neq j'
\end{aligned}
$$

- A primeira covariância indica que os indivíduos são independentes.
- A segunda indica que a covariância é σ_α^2 para quaisquer duas medidas dentro do mesmo indivíduo.

Uma Amostra: ANOVA com Blocos Aleatorizados

A correlação, que reflete a magnitude da associação intra-indivíduo, vale

$$Corr(Y_{ij}, Y_{i,j'}) = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2}.$$

Esta é a chamada *correlação intraclasse* e varia de zero a um:

- Vale zero se os indivíduos não explicam nada da variância (i.e., $\sigma_\alpha^2 = 0$), e
- Vale um se os indivíduos explicam toda a variância (i.e., $\sigma_\varepsilon^2 = 0$).

Uma Amostra: ANOVA com Blocos Aleatorizados

A matriz de variância-covariância tem a forma de *simetria composta*, a qual não é muito realística para dados longitudinais:

- Primeiro porque as variâncias geralmente mudam ao longo do tempo, sendo os indivíduos geralmente mais similares no começo do estudo que no final.
- Segundo porque as covariâncias em tempos mais próximos são geralmente maiores que as covariâncias em tempos mais separados.

Uma Amostra: ANOVA com Blocos Aleatorizados

Testes de hipoteses são construídos como:

$$
\begin{aligned}
H_S &: \qquad \sigma_\tau^2 = 0 \\
H_T &: \tau_1 = \tau_2 = \ldots = \tau_n = 0
\end{aligned}
$$

O foco geralmente é testar a significância do efeito de tempo, pois geralmente assumimos que $\sigma_\alpha^2 > 0$.

Para quantificar o efeito do indivíduo, a correlação intraclasse (ICC) descreve a magnitude (relativa) de σ_α^2:

$$
ICC = \frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_\alpha^2 + \hat{\sigma}_\varepsilon^2}.
$$