

MAT02035 - Modelos para dados correlacionados

Modelando a média: análise de perfis de respostas

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, 2022

Introdução

Introdução

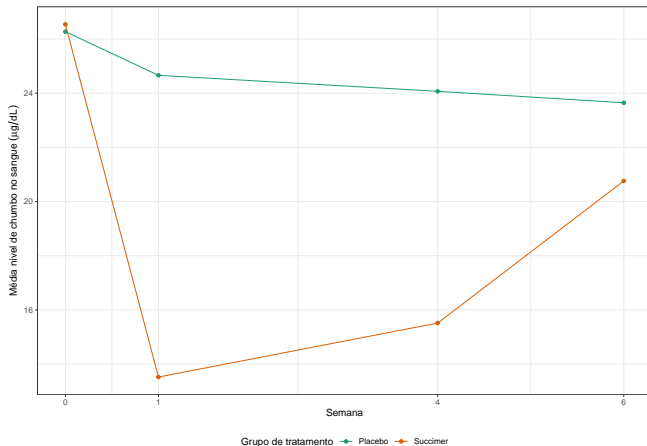
- ▶ Nesta aula apresentamos um método para analisar dados longitudinais que impõe uma estrutura mínima ou restrições na resposta média ao longo do tempo e na covariância entre as medidas repetidas.
- ▶ O método foca na **análise de perfis de respostas** e pode ser aplicado para dados longitudinais quando o **delineamento é balanceado**, com um conjunto de ocasiões de medidas comum para todos os indivíduos no estudo.
 - ▶ A análise de perfis de respostas também pode contemplar dados incompletos devido à perda (ou seja, estudos longitudinais incompletos com delineamentos balanceados).

Introdução

- ▶ Métodos para analisar perfis de respostas são atraentes quando existe uma única covariável categórica (grupo de tratamento ou exposição) e quando nenhum padrão específico a priori para diferenças em perfis de respostas entre grupos pode ser especificado.
- ▶ Os dados podem ser resumidos pela resposta média em cada ocasião de tempo, estratificado por níveis do fator de grupo.
- ▶ Em qualquer nível do fator de grupo, a sequência de médias no tempo é referida como o **perfil de resposta** médio.

Introdução

Relembrando: exemplo TLC



Introdução

- ▶ O principal objetivo da análise de perfis de respostas é caracterizar os padrões de **mudança** na resposta média ao longo do tempo nos grupos e para determinar quanto as formas dos perfis de respostas médias diferem entre os grupos.
- ▶ O método de perfis de respostas pode ser generalizado para estudos com mais de um único fator de grupo de tratamento (exposição) e quando existem covariáveis medidas na linha de base que precisam ser ajustadas.
 - ▶ Em estudos observacionais, os grupos são definidos por características dos indivíduos do estudo, tais como idade, sexo, ou nível de exposição.
 - ▶ Em estudos aleatorizados, os grupos são definidos por um mecanismo aleatório. Logo, espera-se que a distribuição de tais características (idade, sexo, etc.) seja equilibrada (balanceada) entre os grupos de tratamento.

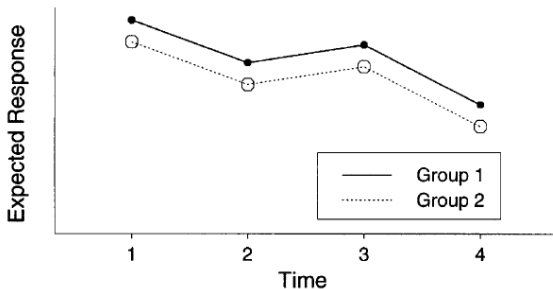
Hipótesis sobre perfiles de respuesta

Hipóteses sobre perfis de resposta

- ▶ Em nossa discussão sobre a análise dos perfis de resposta, focamos inicialmente no delineamento de dois grupos, mas as generalizações para mais de dois grupos são diretas.
- ▶ Dada uma sequência de n medidas repetidas em um número de grupos distintos de indivíduos, três questões principais relacionadas aos perfis de resposta podem ser colocadas:

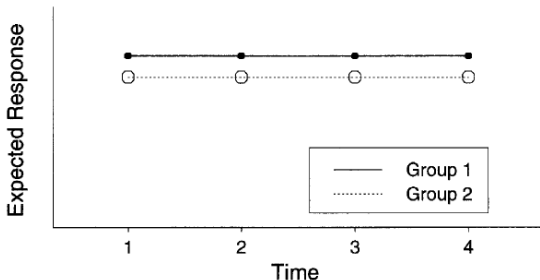
Hipóteses sobre perfis de resposta

1. Os perfis de resposta média são semelhantes nos grupos, no sentido de que os perfis de resposta média são paralelos?
 - ▶ Essa é uma pergunta que diz respeito ao **efeito de interação** *grupo × tempo*.



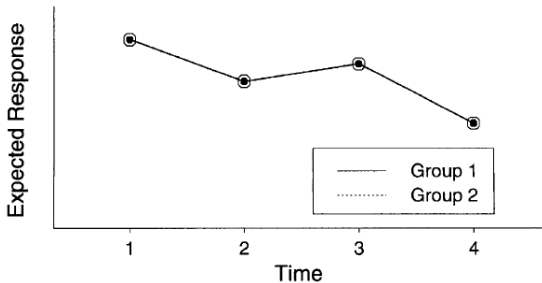
Hipóteses sobre perfis de resposta

2. Supondo que os perfis médios de resposta da população sejam paralelos, as médias são constantes ao longo do tempo?
- Esta é uma pergunta que diz respeito ao **efeito do tempo**.



Hipóteses sobre perfis de resposta

3. Supondo que os perfis médios de resposta da população sejam paralelos, eles também estão no mesmo nível, no sentido de que os perfis médios de resposta para os grupos coincidem?
- Esta é uma pergunta que diz respeito ao **efeito do grupo**.



Hipóteses sobre perfis de resposta

- ▶ Exceto em circunstâncias muito raras, não faz sentido fazer a segunda e a terceira perguntas se os perfis médios de resposta não são paralelos.
- ▶ Isso é consistente com o princípio geral de que os efeitos principais (por exemplo, efeitos de grupo ou tempo) normalmente não são de interesse quando há uma interação entre eles.
- ▶ Ou seja, quando há uma interação *grupo* \times *tempo*, os perfis médios de resposta nos grupos são diferentes (perfis não paralelos); consequentemente, sua forma pode ser descrita apenas com referência a um grupo específico, e seu nível pode ser descrito apenas com referência a um tempo específico.

Formulação do modelo linear geral

Perfis de respostas e o modelo linear geral

- ▶ Antes de ilustrarmos as principais ideias com um exemplo numérico, consideramos como a análise de perfis de respostas pode ser implementada no modelo linear geral

$$E(Y_i|X_i) = \mu_i = X_i\beta,$$

para uma escolha apropriada de X_i .

- ▶ Também descreveremos como as principais hipóteses de **ausência de efeito de interação grupo \times tempo** em termos de β .
- ▶ Considere n o número de medidas repetidas e N o número de indivíduos.
- ▶ Para expressar o modelo para o delineamento longitudinal com G grupos e n ocasiões de medição, **precisaremos de $G \times n$ parâmetros** para G perfis de respostas médias.

Perfis de respostas e o modelo linear geral

- ▶ Por exemplo, com **dois grupos** medidos em **três ocasiões**, há $2 \times 3 = 6$ parâmetros de média.
- ▶ Para o **primeiro grupo**, a matriz de delineamento 3×6 , X_i , fica

$$X_i = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

- ▶ Para o **segundo grupo**, a matriz de delineamento fica

$$X_i = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Perfis de respostas e o modelo linear geral: interpretação

- ▶ Em termos do modelo

$$E(Y_i|X_i) = \mu_i = X_i\beta,$$

em que $\beta = (\beta_1, \dots, \beta_6)'$ é um vetor 6×1 de coeficientes de regressão,

$$\mu(1) = \begin{pmatrix} \mu_1(1) \\ \mu_2(1) \\ \mu_3(1) \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix};$$

similarmente

$$\mu(2) = \begin{pmatrix} \mu_1(2) \\ \mu_2(2) \\ \mu_3(2) \end{pmatrix} = \begin{pmatrix} \beta_4 \\ \beta_5 \\ \beta_6 \end{pmatrix}.$$

Perfis de respostas e o modelo linear geral: hipóteses

- ▶ Como resultado, hipóteses sobre os perfis médios de resposta nos dois grupos podem ser facilmente expressas em termos de hipóteses sobre os componentes de β .
- ▶ Especificamente, a hipótese de ausência de efeito de interação $\text{grupo} \times \text{tempo}$ pode ser expressa como

$$H_{01} : (\beta_1 - \beta_4) = (\beta_2 - \beta_5) = (\beta_3 - \beta_6).$$

- ▶ Nesta parametrização, hipóteses sobre a interação $\text{grupo} \times \text{tempo}$ não podem ser expressas em termos de certos componentes de β como sendo zero.
- ▶ Em vez disso, essas hipóteses podem ser expressas em termos de $L\beta = 0$, para escolhas particulares de vetores ou matrizes L .

Perfis de respostas e o modelo linear geral: hipóteses

- ▶ Por exemplo, a hipótese nula de ausência de efeito de interação *grupo* \times *tempo*,

$$H_{01} : (\beta_1 - \beta_4) = (\beta_2 - \beta_5) = (\beta_3 - \beta_6),$$

pode ser expressa como

$$H_{01} : L\beta = 0,$$

em que

$$L = \begin{pmatrix} 1 & -1 & 0 & -1 & 1 & 0 \\ 1 & 0 & -1 & -1 & 0 & 1 \end{pmatrix}.$$

Perfis de respostas e o modelo linear geral: dados ausentes

- Uma característica atraente da formulação do modelo linear geral

$$E(Y_i|X_i) = \mu_i = X_i\beta,$$

é que ele pode lidar com configurações em que os dados de alguns indivíduos estão ausentes.

- Por exemplo, suponha que o i -ésimo indivíduo pertença ao primeiro grupo e esteja faltando a resposta na terceira ocasião.
- A matriz de delineamento apropriada para esse indivíduo é a seguinte matriz 2×6 , obtida pela remoção da última linha da matriz de delineamento de dados completa para os indivíduos do primeiro grupo:

$$X_i = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Perfis de respostas e o modelo linear geral: dados ausentes

- ▶ Para padrões mais gerais de perda de dados, a matriz de delineamento apropriada para o i -ésimo indivíduo é simplesmente obtida removendo linhas da matriz de delineamento de dados completa correspondentes às respostas ausentes.
- ▶ Isso permite que a análise dos perfis de resposta seja baseada em todas as observações disponíveis dos indivíduos.

Perfis de respostas e o modelo linear geral: categoria de referência

- Note que o modelo linear geral para dois grupos medidos em duas ocasiões,

$$E(Y_i|X_i) = \mu_i = X_i\beta,$$

poderia também ser expresso em termo das seguintes duas matrizes de delineamento:

Perfis de respostas e o modelo linear geral: categoria de referência

$$X_i = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \end{pmatrix},$$

para o primeiro grupo e

$$X_i = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \end{pmatrix},$$

para o segundo grupo.

Perfis de respostas e o modelo linear geral: categoria de referência

► Neste caso

$$\mu(2) = \begin{pmatrix} \mu_1(2) \\ \mu_2(2) \\ \mu_3(2) \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_1 + \beta_2 \\ \beta_1 + \beta_3 \end{pmatrix},$$

e

$$\mu(1) = \begin{pmatrix} \mu_1(1) \\ \mu_2(1) \\ \mu_3(1) \end{pmatrix} = \begin{pmatrix} \beta_1 + \beta_4 \\ (\beta_1 + \beta_4) + (\beta_2 + \beta_5) \\ (\beta_1 + \beta_4) + (\beta_3 + \beta_6) \end{pmatrix}.$$

Perfis de respostas e o modelo linear geral: categoria de referência

- ▶ Esta **parametrização é a mais utilizada** pelos softwares estatísticos.
- ▶ A escolha desta parametrização, e a categoria de referência, é uma escolha do usuário do software (no R veja a formulação `~ -1` e a função `relevel()`).
- ▶ Com esta parametrização as hipóteses de interesse de pesquisa podem ser reescritas:

Ausência de efeito de interação

$$H_{01} : \beta_5 = \beta_6 = 0.$$

Tal hipótese pode ser expressa como $H_{01} : L\beta = 0$, em que

$$L = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Perfis de respostas e o modelo linear geral: categoria de referência

- ▶ Quando a hipótese de perfis paralelos não pode ser rejeitada, hipóteses com respeito aos efeitos principais do tempo e/ou do grupo devem ser de interesse.

Ausência de efeito de mudança ao longo do tempo

$$H_{02} : \beta_2 = \beta_3 = 0.$$

Ou de forma equivalente $H_{02} : L\beta = 0$, em que

$$L = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

Perfis de respostas e o modelo linear geral: categoria de referência

Ausência de efeito de grupo

$$H_{03} : \beta_4 = 0.$$

Ou de forma equivalente $H_{03} : L\beta = 0$, em que

$$L = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

Perfis de respostas e o modelo linear geral: estrutura de covariância

- ▶ Finalmente, dado que a análise de perfis de respostas pode ser expressa em termos do modelo de regressão linear,

$$E(Y_i|X_i) = \mu_i = X_i\beta,$$

em que $\beta = (\beta_1, \dots, \beta_p)'$ é um vetor $p \times 1$ de coeficientes de regressão (com $p = G \times n$), a **estimação de máxima verossimilhança** de β , e a construção de testes de hipóteses para a interação *grupo* \times *tempo* (e efeitos principais de *tempo* e *grupo*), são possíveis uma vez que a **covariância de Y_i foi especificada**.

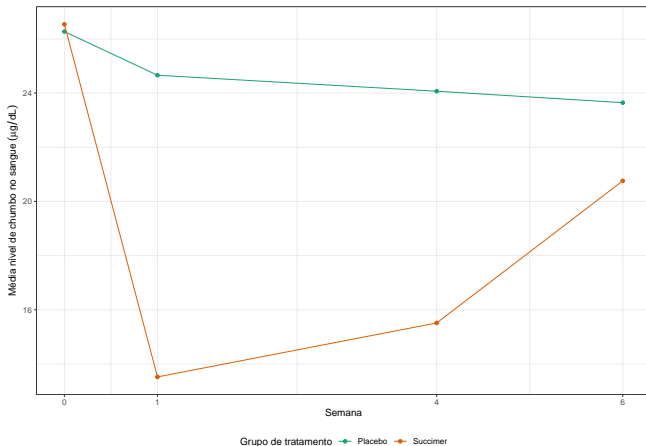
- ▶ Na análise de perfis, a covariância de Y_i é usualmente assumida ser **não estruturada** com nenhuma restrição para os $n(n+1)/2$ parâmetros de covariância.

Exemplo

Estudo de tratamento de crianças expostas ao chumbo

- ▶ Lembre-se de que o estudo TLC foi um estudo aleatorizado, controlado por placebo, de um agente quelante administrado por via oral, *succimer*, em crianças com níveis confirmados de chumbo no sangue de 20 a 44 $\mu\text{g}/\text{dL}$.
- ▶ As crianças do estudo tinham idades entre 12 e 33 meses e viviam em moradias deterioradas no centro da cidade.
- ▶ A análise a seguir é baseada em dados sobre os níveis de chumbo no sangue na linha de base (ou semana 0), semana 1, semana 4 e semana 6 durante o primeiro período de tratamento.

Estudo de tratamento de crianças expostas ao chumbo



Um modelo: perfis de respostas

- Lembre-se que na abordagem de perfis de respostas os tempos de medição são considerados como níveis de um fator discreto.

```
class(chumbo.df.longo$tempo)
```

```
## [1] "factor"
```

```
class(chumbo.df.longo$trt)
```

```
## [1] "factor"
```

Um modelo: perfis de respostas

```
library(nlme)

# modelo de perfis de respostas
# com matriz de covariância não estruturada
mod.pr <- gls(chumbo ~ trt * tempo,
              corr = corSymm(form = ~ 1 | id),
              weights = varIdent(form = ~ 1 | tempo),
              method = "REML",
              data = chumbo.df.longo)
```


Um modelo: perfis de respostas

```
summary(mod.pr)
```

```
## Generalized least squares fit by REML
##   Model: chumbo ~ trt * tempo
##   Data: chumbo.df.longo
##           AIC      BIC    logLik
##   2452.076 2523.559 -1208.038
##
## Correlation Structure: General
## Formula: ~1 | id
## Parameter estimate(s):
## Correlation:
##   1      2      3
## 2 0.571
## 3 0.570 0.775
## 4 0.577 0.582 0.581
## Variance function:
## Structure: Different standard deviations per stratum
## Formula: ~1 | tempo
## Parameter estimates:
##           0      1      4      6
## 1.000000 1.325888 1.370453 1.524826
```

Um modelo: perfis de respostas

```
##
## Coefficients:
##              Value Std.Error   t-value p-value
## (Intercept)    26.272  0.7102888   36.98777  0.0000
## trtSuccimer      0.268  1.0045000    0.26680  0.7898
## tempo1         -1.612  0.7919194   -2.03556  0.0425
## tempo4         -2.202  0.8149167   -2.70212  0.0072
## tempo6         -2.626  0.8885252   -2.95546  0.0033
## trtSuccimer:tempo1 -11.406  1.1199432  -10.18445  0.0000
## trtSuccimer:tempo4  -8.824  1.1524662   -7.65662  0.0000
## trtSuccimer:tempo6  -3.152  1.2565645   -2.50843  0.0125
##
## Correlation:
##              (Intr) trtScc tempo1 tempo4 tempo6 trtS:1 trtS:4
## trtSuccimer    -0.707
## tempo1         -0.218  0.154
## tempo4         -0.191  0.135  0.680
## tempo6         -0.096  0.068  0.386  0.385
## trtSuccimer:tempo1  0.154 -0.218 -0.707 -0.481 -0.273
## trtSuccimer:tempo4  0.135 -0.191 -0.481 -0.707 -0.272  0.680
## trtSuccimer:tempo6  0.068 -0.096 -0.273 -0.272 -0.707  0.386  0.385
##
```

Um modelo: perfis de respostas

```
## Standardized residuals:
```

```
##           Min           Q1           Med           Q3           Max
## -2.1756390 -0.6849960 -0.1515545  0.5294173  5.6327402
```

```
##
```

```
## Residual standard error: 5.0225
```

```
## Degrees of freedom: 400 total; 392 residual
```

Matriz covariância estimada

- Para obter a matriz de covariância estimada utilizamos a função `getVarCov`.

```
getVarCov(mod.pr)
```

```
## Marginal variance covariance matrix
##      [,1]  [,2]  [,3]  [,4]
## [1,] 25.226 19.107 19.699 22.202
## [2,] 19.107 44.346 35.535 29.675
## [3,] 19.699 35.535 47.377 30.620
## [4,] 22.202 29.675 30.620 58.652
## Standard Deviations: 5.0225 6.6593 6.8831 7.6584
```

Matriz covariância estimada

- ▶ Como a função `getVarCov` retorna uma matriz, podemos utilizar a função `kable` do pacote `knitr` para a geração de tabelas em markdown.

```
knitr::kable(x = matrix(getVarCov(mod.pr),  
                        ncol = 4),  
             digits = 1)
```

25.2	19.1	19.7	22.2
19.1	44.3	35.5	29.7
19.7	35.5	47.4	30.6
22.2	29.7	30.6	58.7

Matriz covariância estimada

- ▶ Observe o aumento perceptível na variância dos níveis de chumbo no sangue de pré a pós-aleatorização.
- ▶ Este aumento na variância da linha de base é provavelmente devido a dois fatores.
 - ▶ Primeiro, dentro de cada grupo de tratamento, pode haver heterogeneidade natural nas trajetórias de resposta individual ao longo do tempo.
 - ▶ Em segundo lugar, o estudo teve um critério de inclusão de que os níveis de chumbo no sangue no início do estudo estavam na faixa de 20 a 44 $\mu\text{g}/\text{dL}$; isso pode ser parcialmente responsável pela variação menor na linha de base.

Testando hipóteses (teste de Wald)

- ▶ A função Anova do pacote car apresenta o resultado de testes Wald multivariado de múltiplas hipóteses (H_{01} , H_{02} e H_{03} testadas separadamente).

```
library(car)
```

```
Anova(mod.pr)
```

```
## Analysis of Deviance Table (Type II tests)
```

```
##
```

```
## Response: chumbo
```

```
##           Df      Chisq Pr(>Chisq)
```

```
## trt         1    4.2266    0.0398 *
```

```
## tempo       3 184.4806    <2e-16 ***
```

```
## trt:tempo   3 107.7870    <2e-16 ***
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testando hipóteses (teste de Wald)

- Mais uma vez podemos gerar uma tabela em markdown.

```
knitr::kable(  
  Anova(mod.pr),  
  digits = c(0, 2, 4))
```

	Df	Chisq	Pr(>Chisq)
trt	1	4.23	0.0398
tempo	3	184.48	0.0000
trt:tempo	3	107.79	0.0000

Testando hipóteses (teste de razão de verossimilhanças)

- ▶ Para testar a hipótese de ausência de efeito de interação entre *grupo* e *tempo* também podemos utilizar o teste da razão de verossimilhanças.
 - ▶ Para tal, precisamos ajustar dois modelos: um completo e outro reduzido.
- ▶ **Importante:** a construção de testes de razão de verossimilhanças comparando modelos encaixados para a média deve sempre ser baseada na log-verossimilhança MV, e não no REML.

Testando hipóteses (teste de da razão de verossimilhanças)

```
mod.comp <- gls(chumbo ~ trt * tempo,  
               corr = corSymm(form = ~ 1 | id),  
               weights = varIdent(form = ~ 1 | tempo),  
               method = "ML",  
               data = chumbo.df.longo)  
  
mod.red <- gls(chumbo ~ trt + tempo,  
              corr = corSymm(form = ~ 1 | id),  
              weights = varIdent(form = ~ 1 | tempo),  
              method = "ML",  
              data = chumbo.df.longo)
```

Testando hipóteses (teste de razão de verossimilhanças)

```
anova(mod.comp, mod.red)
```

##	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
##	mod.comp	1 18	2461.368	2533.214	-1212.684			
##	mod.red	2 15	2529.555	2589.427	-1249.778	1 vs 2	74.18778	<.0001

Coefficientes estimados

- Os coeficientes de regressão estimados e os respectivos erros padrões também podem ser combinado para gerar uma saída em markdown.

```
knitr::kable(
  summary(mod.pr)$tTable[,-4],
  digits = c(3, 3, 2),
  col.names = c("Estimativa", "EP", "Z"))
```

	Estimativa	EP	Z
(Intercept)	26.272	0.710	36.99
trtSuccimer	0.268	1.005	0.27
tempo1	-1.612	0.792	-2.04
tempo4	-2.202	0.815	-2.70
tempo6	-2.626	0.889	-2.96
trtSuccimer:tempo1	-11.406	1.120	-10.18
trtSuccimer:tempo4	-8.824	1.152	-7.66
trtSuccimer:tempo6	-3.152	1.257	-2.51

Exercícios

1. Resolva os exercícios do Capítulo 5 do livro “**Applied Longitudinal Analysis**” (páginas 140 e 141).
 - ▶ O arquivo de dados (`cholesterol.dta`) estão no Moodle.
2. Construa intervalos de confiança de 95% para as estimativas do modelo do exemplo da aula.
3. Com base na leitura da Seção 5.8 (Applied Longitudinal Analysis), faça uma discussão dos pontos fortes e fracos da análise de perfis de resposta e poste no **fórum geral do Moodle**.

Avisos

- ▶ **Próxima aula:** Modelando a média através de curvas paramétricas.
- ▶ **Para casa:** ler o Capítulo 5 do livro “**Applied Longitudinal Analysis**”.
 - ▶ Caso ainda não tenha lido, leia também os Caps. 1, 2, 3 e 4.

Bons estudos!

