MAT02035 - Modelos para dados correlacionados

Modelando a média: curvas paramétricas

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, 2021

Introdução

Introdução

- Na aula anterior descrevemos uma abordagem para modelar dados longitudinais que efetivamente não impôs nenhuma estrutura na **tendência no tempo** da resposta média.
  - Esta abordagem tem algum apelo quando todos os indivíduos foram medidos nas mesmas ocasiões e o número de ocasiões é relativamente pequeno.
- Mas, conforme o número de ocasiões aumenta e as medidas repetidas são irregulares no tempo, analisar perfis de resposta se torna menos atraente.
  - Ainda, o teste da hipótese nula de **ausência de interação** *grupo* $\times$ *tempo* é um teste global e (a rejeição desta hipótese nula) não fornece nenhum indicativo de um padrão de mudança da média ao longo do tempo.
- A resposta média ao longo tempo pode, em geral, ser descrita por curvas paramétricas (**lineares** ou **quadráticas**) ou multiparamétricas (**splines lineares**) relativamente simples.

Introdução

- De um ponto de vista puramente substantivo, é improvável que o padrão de mudança na resposta média ao longo de um estudo longitudinal seja tão complicado que sua descrição exija tantos parâmetros quanto as ocasiões de medição.
- A análise dos perfis de resposta usa um modelo **saturado** para a resposta média ao longo do tempo e, portanto, produz um ajuste perfeito ao perfil de resposta média observado.
  - Ao fazer isso, o método falha em descrever os aspectos mais importantes das mudanças na resposta média ao longo do tempo em termos de algum padrão que pode ser interpretado de maneira substantiva ou teórica.
- Ou seja, na análise dos perfis de resposta, **não há redução na complexidade**.

Introdução

- Por outro lado, o ajuste de curvas paramétricas ou semiparamétricas a dados longitudinais pode ser justificado em bases substantivas e estatísticas.
- Substancialmente, em muitos estudos longitudinais, o verdadeiro **processo de resposta média subjacente** provavelmente **mudará ao longo do tempo** em um **padrão** relativamente **suave**, monotonicamente crescente ou decrescente, pelo menos durante a duração do estudo.
  - Como resultado, curvas paramétricas ou semiparamétricas simples podem ser usadas para descrever como a resposta média muda com o tempo.
- De uma perspectiva estatística, o ajuste de modelos parcimoniosos para a resposta média resultará em testes estatísticos de efeitos de covariáveis (por exemplo, interações tratamento x tempo) que têm **maior poder** do que em uma análise de perfis de resposta.

Introdução

- Finalmente, curvas paramétricas simples fornecem uma **descrição parcimoniosa** das mudanças na resposta média ao longo do tempo em termos de um número relativamente pequeno de parâmetros.
- Os resultados podem ser comunicados facilmente a pesquisadores.

Tendências polinomiais no tempo

Tendências lineares no tempo

- A curva mais simples possível para descrever mudanças na resposta média ao longo do tempo é uma **linha reta**.
    - Neste modelo, a inclinação do tempo tem interpretação direta em termos de uma **mudança constante** na resposta média para uma mudança de unitária única no tempo[1].
- Considere o estudo hipotético de **dois grupos** comparando um **novo tratamento** e um **controle**.

[1] Se $\mathrm{E}(Y_{ij}) = \beta_1 + \beta_2 t_{ij}$, então $\frac{d}{dt_{ij}}\mathrm{E}(Y_{ij}) = \beta_2$.

Tendências lineares no tempo

- Se a resposta média mudar de uma maneira aproximadamente linear ao longo da duração do estudo, podemos adotar o seguinte modelo de tendência linear:

$$\mathrm{E}\,(Y_{ij}) = \beta_1 + \beta_2\,Tempo_{ij} + \beta_3\,Grupo_i + \beta_4(\,Tempo_{ij} \times Grupo_i\,),$$

em que $Grupo_i = 1$ se o i-ésimo indivíduo foi designado ao novo tratamento, e $Grupo_i = 0$ caso contrário; $Tempo_{ij}$ denota o tempo de medição para a j-ésima ocasião no i-ésimo indivíduo.

- No modelo linear acima, a média dos indivíduos designados ao grupo controle é

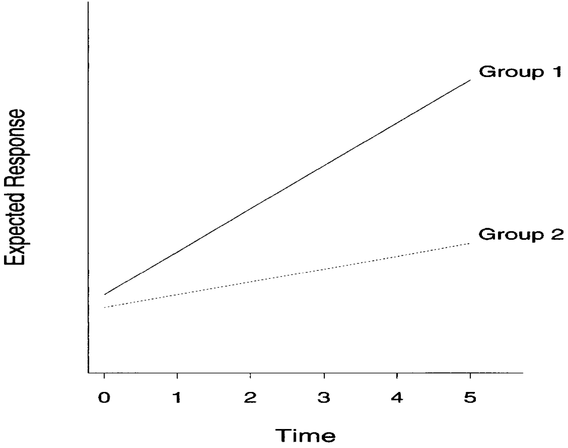
$$\mathrm{E}\,(Y_{ij}) = \beta_1 + \beta_2\,Tempo_{ij},$$

enquanto que para os indivíduos do grupo tratamento é

$$\mathrm{E}\,(Y_{ij}) = (\beta_1 + \beta_3) + (\beta_2 + \beta_4)\,Tempo_{ij}.$$

Tendências lineares no tempo

- Presume-se que a resposta média de cada grupo muda linearmente ao longo do tempo.



Tendências lineares no tempo

- Aqui β_1 é o intercepto no grupo de controle (o grupo de "referência"), enquanto $(\beta_1 + \beta_2)$ é o intercepto no grupo de tratamento.
- Os interceptos para cada um dos dois grupos têm interpretação em termos da resposta média quando $Tempo_{ij} = 0$.
- A menos que algum cuidado seja tomado com como as covariáveis são padronizadas (por exemplo, centralizando todas as covariáveis quantitativas antes da inclusão no modelo), β_1 nem sempre é prontamente interpretável e pode representar uma extrapolação para além dos dados em mãos.

Tendências lineares no tempo

- Finalmente, a inclinação, ou taxa constante de mudança na resposta média por unidade de mudança no tempo, é β_2 no grupo controle, enquanto a inclinação correspondente no grupo de tratamento é $(\beta_2 + \beta_4)$.
- Normalmente, em um estudo longitudinal a questão de interesse primário diz respeito a uma comparação das mudanças na resposta média ao longo do tempo; isso pode ser traduzido em uma comparação das inclinações.
- Assim, se $\beta_4 = 0$, então os dois grupos não diferem em termos de mudanças na resposta média ao longo do tempo.

Tendências quadráticas no tempo

- Quando mudanças na resposta média ao longo do tempo não são lineares, **tendências polinomiais de ordem mais alta** podem ser consideradas.
- Por exemplo, se as médias estão aumentando monotonicamente ou diminuindo ao longo do estudo, mas de forma curvilínea, um modelo com **tendências quadráticas** pode ser considerado.
- Em um modelo de tendência quadrática, as **mudanças na resposta média não são mais constantes** (como no modelo de tendência linear) ao longo da duração do estudo.
- Em vez disso, a taxa de mudança na resposta média depende do tempo[2].

[2] Se $\mathrm{E}(Y_{ij}) = \beta_1 + \beta_2 t_{ij} + \beta_3 t_{ij}^2$, então $\frac{d}{dt_{ij}}\mathrm{E}(Y_{ij}) = \beta_2 + 2\beta_3 t_{ij}$.

Tendências quadráticas no tempo

- Considere mais uma vez o estudo hipotético de **dois grupos** comparando um **novo tratamento** e um **controle**.
- Assumindo que as mudanças nas respostas médias podem ser aproximadas por tendências quadráticas, o seguinte modelo pode ser adotado:

$$
\begin{aligned}
\mathrm{E}\,(Y_{ij}) &= \beta_1 + \beta_2 \, Tempo_{ij} + \beta_3 \, Tempo_{ij}^2 + \beta_4 \, Grupo_i \\
&\quad + \beta_5 (Tempo_{ij} \times Grupo_i) + \beta_6 (Tempo_{ij}^2 \times Grupo_i).
\end{aligned}
$$

Tendências quadráticas no tempo

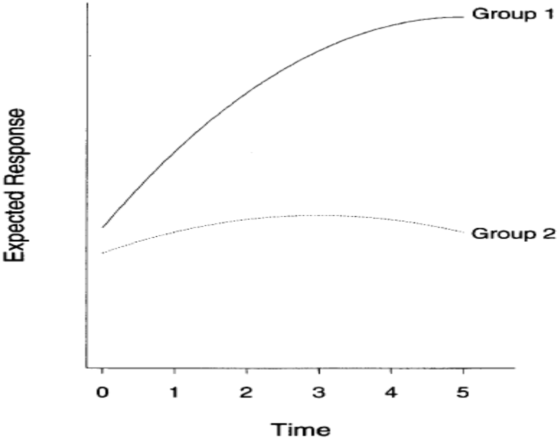
- Neste modelo, a média dos indivíduos designados ao grupo controle é

$$\mathrm{E}\left(Y_{ij}\right) = \beta_1 + \beta_2 Tempo_{ij} + \beta_3 Tempo_{ij}^2,$$

enquanto a resposta média correspondente para os indivíduos do grupo tratamento é

$$\mathrm{E}\left(Y_{ij}\right) = (\beta_1 + \beta_4) + (\beta_2 + \beta_5) Tempo_{ij} + (\beta_3 + \beta_6) Tempo_{ij}^2.$$

Tendências quadráticas no tempo



Tendências quadráticas no tempo

- Observe que no modelo de tendências quadráticas, a resposta média muda a uma taxa diferente, dependendo do $Time_{ij}$.
- Por exemplo, a taxa de mudança no grupo controle é dada por $\beta_2 + 2\beta_3 Time_{ij}$.
  - Assim, no início do estudo, quando $Tempo_{ij} = 1$, a taxa de mudança na resposta média é $\beta_2 + 2\beta_3$, enquanto mais tarde no estudo, digamos $Tempo_{ij} = 4$, a taxa de mudança na resposta média é $\beta_2 + 8\beta_3$.

Tendências quadráticas no tempo

- A taxa de mudança é diferente nas duas ocasiões e a magnitude e o sinal dos coeficientes de regressão β_2 e β_3 determinam se a resposta média está aumentando ou diminuindo ao longo do tempo e como a taxa de mudança depende no tempo.
- Os coeficientes de regressão, $(\beta_2 + \beta_5)$ e $(\beta_3 + \beta_6)$, têm interpretações semelhantes para o grupo de tratamento.

Tendências quadráticas no tempo

- Nos modelos de tendência polinomial, existe uma **hierarquia natural** de efeitos que tem implicações para testar hipóteses sobre tendências polinomiais lineares, quadráticas e de ordem superior.
- Termos de ordem superior devem ser testados (e, se apropriado, removidos do modelo) antes que os termos de ordem inferior sejam avaliados.

Tendências quadráticas no tempo

- Assim, no modelo quadrático, não é apropriado testar o coeficiente para a tendência linear, β_2, em um modelo que também inclui um coeficiente para a tendência quadrática, β_3.
- Em vez disso, um teste para a tendência quadrática (versus tendência linear) pode ser executada testando a hipótese nula de que $\beta_3 = 0$.
  - **Se essa hipótese nula não puder ser rejeitada**, é apropriado remover o termo quadrático do modelo e considerar o modelo apenas com tendência linear.
  - O teste para tendência linear é realizada testando a hipótese nula de que $\beta_2 = 0$ no modelo que inclui **apenas** o termo linear.

Splines lineares

Splines lineares

- Em algumas aplicações, as tendências longitudinais na resposta média não podem ser caracterizadas por polinômios de primeiro e segundo graus no tempo.
- Além disso, existem outras aplicações em que tendências não lineares na resposta média não podem ser bem aproximadas por polinômios de qualquer ordem.
- Isso ocorre com mais frequência quando a resposta média aumenta (ou diminui) rapidamente por algum tempo e depois mais lentamente depois (ou vice-versa).
- Quando esse tipo de padrão de alteração ocorre, muitas vezes pode ser tratado usando **modelos de splines lineares**.

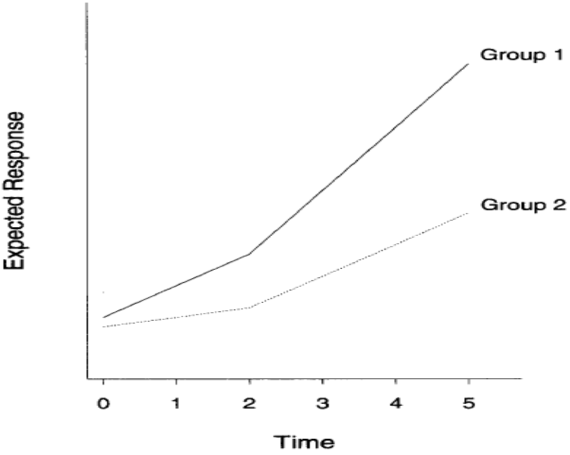
Splines lineares

- Se o modelo mais simples possível para a resposta média for uma linha reta, uma maneira de estender este modelo é ter uma **sequência de segmentos de retas** conectados que produzam um **padrão linear por partes**.
- Os modelos de spline linear fornecem uma maneira muito útil e flexível de acomodar muitas das tendências não lineares que não podem ser aproximadas por polinômios simples no tempo.
- A ideia básica por trás dos modelos de splines lineares é notavelmente simples:
  - Divida o eixo do tempo em uma série de segmentos e considere um modelo para a tendência ao longo do tempo, composta por tendências lineares por partes, com **diferentes inclinações** em cada segmento, mas unidas em tempos fixos.

Splines lineares

- Os locais em que as retas são interligadas são conhecidos como **“nós”**.
- Este modelo permite que a resposta média aumente ou diminua à medida que o tempo avança, dependendo do sinal e da magnitude das inclinações de regressão para os segmentos de retas.
- A curva linear por partes resultante é chamada de **spline**.

Splines lineares



Splines lineares

- O modelo de spline mais simples possível **possui apenas um nó** e pode ser parametrizado de **várias maneiras diferentes**.
- Voltando ao estudo hipotético de dois grupos, comparando um **novo tratamento** e um **controle**, se a resposta média mudar ao longo do tempo de maneira linear por partes, podemos ajustar o seguinte modelo de spline linear com o nó em t^*:

$$\begin{aligned} \mathrm{E}(Y_{ij}) &= \beta_1 + \beta_2 Tempo_{ij} + \beta_3 (Tempo_{ij} - t^*)_+ + \beta_4 Grupo_i \\ &+ \beta_5 (Tempo_{ij} \times Grupo_i) + \beta_6 \{(Tempo_{ij} - t^*)_+ \times Grupo_i\}, \end{aligned}$$

em que $(x)_+$, é conhecida como a **função de reta truncada**, e é definida como uma função que é **igual a x quando x é positivo** e é **igual a zero caso contrário**. Assim,

$$(Tempo_{ij} - t^*)_+ = \begin{cases} (Tempo_{ij} - t^*), & \text{se} \quad Tempo_{ij} > t^*; \\ 0, & \text{se} \quad Tempo_{ij} \leq t^*. \end{cases}$$

Splines lineares

- No modelo acima, as médias para os indivíduos no grupo controle são

$$E\,(Y_{ij}) = \beta_1 + \beta_2\,Tempo_{ij} + \beta_3(\,Tempo_{ij} - t^*)_+.$$

- Quando expressa em termos da reposta média antes e após t^*,

$$\begin{array}{lcll} E\,(Y_{ij}) & = & \beta_1 + \beta_2\,Tempo_{ij}, & Tempo_{ij} \leq t^*; \\ E\,(Y_{ij}) & = & (\beta_1 - \beta_3 t^*) + (\beta_2 + \beta_3)\,Tempo_{ij}, & Tempo_{ij} > t^*. \end{array}$$

- No grupo controle, a inclinação antes de t^* é β_2 e após t^* é $(\beta_2 + \beta_3)$.

Splines lineares

- De maneira similar para os indivíduos do grupo tratamento são dadas por

$$\mathrm{E}\left(Y_{ij}\right) = (\beta_1 + \beta_4) + (\beta_2 + \beta_5)\,Tempo_{ij} + (\beta_3 + \beta_6)(\,Tempo_{ij} - t^*)_{+}.$$

- Quando expressa em termos da reposta média antes e após t^*,

$$\begin{array}{rcll}
\mathrm{E}\left(Y_{ij}\right) & = & (\beta_1 + \beta_4) + (\beta_2 + \beta_5)\,Tempo_{ij}, & Tempo_{ij} \leq t^*; \\
\mathrm{E}\left(Y_{ij}\right) & = & \{(\beta_1 + \beta_4) - (\beta_3 + \beta_6)t^*)\} + (\beta_2 + \beta_3 + \beta_5 + \beta_6)\,Tempo_{ij}, & Tempo_{ij} > t^*.
\end{array}$$

Splines lineares

- Então, em termos de comparações entre grupos, a hipótese nula de não haver diferenças entre os grupos nos padrões de mudança ao longo do tempo pode ser expressa como $H_0 : \beta_5 = \beta_6 = 0$.
- Também são possíveis comparações dos grupos antes e depois de t^*.
  - Por exemplo, a hipótese nula de não haver diferenças de grupo nos padrões de mudança anteriores a t^* pode ser expressa como $H_0 : \beta_5 = 0$.

Splines lineares

- Como fazer a escolha da quantidade e localização de nós? Existe uma escolha ótima?
- É possível o ajuste de modelos de splines não-lineares (**polinômios por partes**; **splines quadráticos**; **splines cúbicos**)?
- Cenas de um capítulo futuro!

Splines lineares

- Como fazer a escolha da quantidade e localização de nós? Existe uma escolha ótima?
- É possível o ajuste de modelos de splines não-lineares (**polinômios por partes**; **splines quadráticos**; **splines cúbicos**)?
- Cenas de um capítulo futuro!
- No momento, vamos nos restringir aos splines lineares, especificando o número e a localização dos nós de acordo com o comportamento das respostas médias observadas.

Formulação do modelo linear geral

Curvas paramétricas e o modelo linear geral

- A seguir, demonstramos como os modelos de tendência polinomial e spline podem ser expressos em termos do modelo linear geral

$$\mathrm{E}\left(Y_i|X_i\right) = \mu_i = X_i\beta,$$

para uma escolha apropriada de X_i.

- Considere n_i o número de medidas repetidas no i-ésimo indivíduo $(i = 1, \ldots, N)$.
- Para ilustrar como o modelo de tendência polinomial pode ser expresso em termos do modelo linear geral, considere o estudo hipotético de dois grupos comparando um **novo tratamento** e um **controle** discutido anteriormente.

Curvas paramétricas e o modelo linear geral

- Vamos supor que a resposta média mude ao longo do tempo em uma **tendência quadrática**.
- Assim, a matriz de delineamento X_i tem a seguinte forma para o **grupo de controle**:

$$X_i = \begin{pmatrix} 1 & t_{i1} & t_{i1}^2 & 0 & 0 & 0 \\ 1 & t_{i2} & t_{i2}^2 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_{in_i} & t_{in_i}^2 & 0 & 0 & 0 \end{pmatrix},$$

Curvas paramétricas e o modelo linear geral

enquanto que para o **grupo tratamento** a matriz de delineamento é dada por

$$X_i = \left(\begin{array}{cccccc} 1 & t_{i1} & t_{i1}^2 & 1 & t_{i1} & t_{i1}^2 \\ 1 & t_{i2} & t_{i2}^2 & 1 & t_{i2} & t_{i2}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_{in_i} & t_{in_i}^2 & 1 & t_{in_i} & t_{in_i}^2 \end{array} \right).$$

- **Exercício:** como fica o vetor μ_i em função de X_i e $\beta = (\beta_1, \ldots, \beta_6)'$ para cada um dos grupos?

Curvas paramétricas e o modelo linear geral

- Para o **modelo de spline**, suponhamos que a resposta média mude ao longo do tempo de maneira linear por partes, com nó em $t^* = 4$.
- Assim, a matriz de delineamento X_i tem a seguinte forma para o **grupo de controle**:

$$X_i = \begin{pmatrix} 1 & t_{i1} & (t_{i1} - 4)_+ & 0 & 0 & 0 \\ 1 & t_{i2} & (t_{i2} - 4)_+ & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_{in_i} & (t_{in_i} - 4)_+ & 0 & 0 & 0 \end{pmatrix},$$

Curvas paramétricas e o modelo linear geral

enquanto que para o **grupo tratamento** a matriz de delineamento é dada por

$$X_i = \left(\begin{array}{cccccc} 1 & t_{i1} & (t_{i1}-4)_+ & 1 & t_{i1} & (t_{i1}-4)_+ \\ 1 & t_{i2} & (t_{i2}-4)_+ & 1 & t_{i2} & (t_{i2}-4)_+ \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_{in_i} & (t_{in_i}-4)_+ & 1 & t_{in_i} & (t_{in_i}-4)_+ \end{array}\right).$$

Curvas paramétricas e o modelo linear geral

- Como os modelos de tendência polinomial e spline podem ser expressos em termos do modelo de regressão linear geral,

$$\mathrm{E}\left(Y_i|X_i\right) = \mu_i = X_i\beta,$$

a estimativa de máxima verossimilhança restrita de β, e a construção de intervalos de confiança e testes de hipóteses, são possíveis quando a **covariância** de Y_i foi **especificada**.

- Diferentemente da análise dos perfis de resposta, modelos mais parcimoniosos para a covariância podem ser adotados.

Curvas paramétricas e o modelo linear geral

- De fato, o uso de curvas paramétricas para a resposta média é mais atraente em configurações em que os dados longitudinais são inerentemente desbalanceados ao longo do tempo.
    - Como resultado, uma matriz de covariância não estruturada pode não estar bem definida, muito menos estimada quando, em princípio, cada indivíduo pode ter uma sequência única de tempos de medição.

Curvas paramétricas e o modelo linear geral

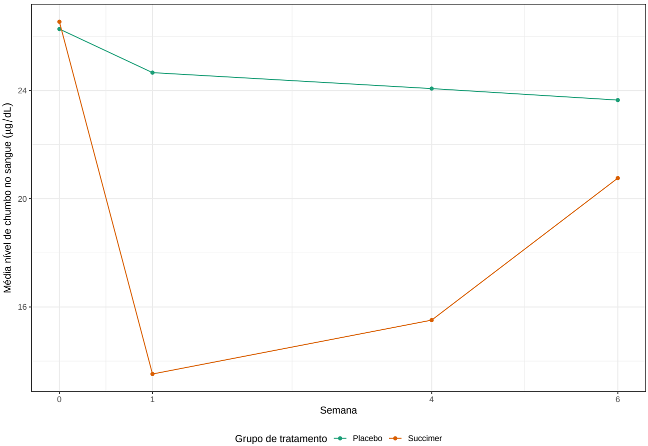
- No entanto, a discussão de modelos para a covariância é adiada para a próxima aula!
- Aqui assumimos simplesmente que algum modelo apropriado para a covariância foi adotado.
- Dados os modelos para a média e a covariância, as estimativas REML e seus erros padrão (com base na covariância estimada de $\hat{\beta}$), podem ser obtidos usando o método de estimação apresentado na Aula 09.

Estudo de caso

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 μg/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 spline linear (velha guarda)

- Observe que, a partir do gráfico das médias, nos parece que apenas os níveis médios de chumbo no sangue no grupo placebo podem ser descritos por uma tendência linear; a média no grupo succimer diminui da linha de base até a semana 1, mas depois aumenta.
- Dado que existem não linearidades nas tendências ao longo do tempo, modelos polinomiais de ordem superior (por exemplo, um modelo de tendência quadrática) podem ser ajustados aos dados.
- No entanto, para ilustrar a aplicação de modelos spline, acomodamos a não linearidade com um modelo linear por partes com nó comum na semana 1 ($t^* = 1$),

$$
\begin{aligned}
\mathrm{E}\,(Y_{ij}) &= \beta_1 + \beta_2 Semana_{ij} + \beta_3 (Semana_{ij} - 1)_+ + \beta_4 Grupo_i \\
&\quad + \beta_5 (Semana_{ij} \times Grupo_i) + \beta_6 \{(Semana_{ij} - 1)_+ \times Grupo_i\},
\end{aligned}
$$

em que $Grupo_i = 1$ se o i-ésimo indivíduo foi designado ao novo tratamento, e $Grupo_i = 0$ caso contrário.

Um modelo spline linear

```
chumbo.df.longo
```

```
## # A tibble: 400 x 5
##       id trt      tempo chumbo semana
##    <dbl> <fct>    <dbl>  <dbl>  <dbl>
##  1     1 Placebo      1   30.8      0
##  2     2 Succimer     1   26.5      0
##  3     3 Succimer     1   25.8      0
##  4     4 Placebo      1   24.7      0
##  5     5 Succimer     1   20.4      0
##  6     6 Succimer     1   20.4      0
##  7     7 Placebo      1   28.6      0
##  8     8 Placebo      1   33.7      0
##  9     9 Placebo      1   19.7      0
## 10    10 Placebo      1   31.1      0
## # ... with 390 more rows
```

Um modelo spline linear

```
library(nlme)

# modelo de curvas paramétricas
#     splines lineares
# com matriz de covariância não estruturada
mod.spline <- gls(chumbo ~ semana + I( (semana - 1) * (semana >= 1) ) +
                    semana:trt + I( (semana - 1) * (semana >= 1) ):trt,
                  corr = corSymm(form = ~ tempo | id),
                  weights = varIdent(form = ~ 1 | tempo),
                  method = "REML",
                  data = chumbo.df.longo)
```

Um modelo spline linear

```r
summary(mod.spline)
```

```
## Generalized least squares fit by REML
##   Model: chumbo ~ semana + I((semana - 1) * (semana >= 1)) + semana:trt +      I((semana
##   Data: chumbo.df.longo
##        AIC      BIC    logLik
##   2467.452 2527.136 -1218.726
##
## Correlation Structure: General
##  Formula: ~tempo | id
##  Parameter estimate(s):
##  Correlation:
##   1     2     3
## 2 0.569
## 3 0.560 0.768
## 4 0.574 0.576 0.553
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | tempo
##  Parameter estimates:
##        1        2        3        4
## 1.000000 1.329903 1.385011 1.544057
##
## Coefficients:
##                                                      Value Std.Error   t-value p-value
## (Intercept)                                      26.342207 0.4991198  52.77733  0.0000
```

Um modelo spline linear

```
## semana                                     -1.629603 0.7817592  -2.08453  0.0378
## I((semana - 1) * (semana >= 1))             1.430494 0.8777585   1.62971  0.1040
## semana:trtSuccimer                        -11.249985 1.0924522 -10.29792  0.0000
## I((semana - 1) * (semana >= 1)):trtSuccimer  12.582249 1.2278544  10.24735  0.0000
##
##  Correlation:
##                                             (Intr) semana I((s-1)*(>=1)) smn:tS
## semana                                      -0.154
## I((semana - 1) * (semana >= 1))              0.147 -0.988
## semana:trtSuccimer                           0.000 -0.699  0.691
## I((semana - 1) * (semana >= 1)):trtSuccimer  0.000  0.690 -0.699         -0.987
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.0020271 -0.6888161 -0.1136309  0.5520751  5.8000806
##
## Residual standard error: 4.999256
## Degrees of freedom: 400 total; 395 residual
```

Um modelo spline linear

- Uma forma mais moderna de especificar o mesmo modelo é utilizando a função `lspline` do pacote de mesmo nome.

```
library(lspline)

# modelo de curvas paramétricas
#     splines lineares
# com matriz de covariância não estruturada
mod.spline2 <- gls(chumbo ~ lspline(x = semana,
                                    knots = 1,
                                    marginal = TRUE) +
                     lspline(x = semana,
                             knots = 1,
                             marginal = TRUE):trt,
                   corr = corSymm(form = ~ tempo | id),
                   weights = varIdent(form = ~ 1 | tempo),
                   method = "REML",
                   data = chumbo.df.longo)
```

Um modelo spline linear

```r
summary(mod.spline2)
```

```
## Generalized least squares fit by REML
##   Model: chumbo ~ lspline(x = semana, knots = 1, marginal = TRUE) + lspline(x = semana,      knots = 1, marginal = TRUE):trt
##   Data: chumbo.df.longo
##        AIC      BIC    logLik
##   2467.452 2527.136 -1218.726
##
## Correlation Structure: General
##  Formula: ~tempo | id
##  Parameter estimate(s):
##  Correlation:
##   1     2     3
## 2 0.569
## 3 0.560 0.768
## 4 0.574 0.576 0.553
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | tempo
##  Parameter estimates:
##        1        2        3        4
## 1.000000 1.329903 1.385011 1.544057
##
## Coefficients:
##                                                                   Value Std.Error   t-value
## (Intercept)                                                   26.342207 0.4991198  52.77733
## lspline(x = semana, knots = 1, marginal = TRUE)1              -1.629603 0.7817592  -2.08453
## lspline(x = semana, knots = 1, marginal = TRUE)2               1.430494 0.8777585   1.62971
## lspline(x = semana, knots = 1, marginal = TRUE)1:trtSuccimer -11.249985 1.0924522 -10.29792
## lspline(x = semana, knots = 1, marginal = TRUE)2:trtSuccimer  12.582249 1.2278544  10.24735
##                                                              p-value
## (Intercept)                                                   0.0000
## lspline(x = semana, knots = 1, marginal = TRUE)1              0.0378
## lspline(x = semana, knots = 1, marginal = TRUE)2              0.1040
## lspline(x = semana, knots = 1, marginal = TRUE)1:trtSuccimer  0.0000
```

Um modelo spline linear

```
## lspline(x = semana, knots = 1, marginal = TRUE)2:trtSuccimer  0.0000
##
##  Correlation:
##                                                               (Intr) ls(=s,k=1,m=TRUE)1
## lspline(x = semana, knots = 1, marginal = TRUE)1             -0.154
## lspline(x = semana, knots = 1, marginal = TRUE)2              0.147 -0.988
## lspline(x = semana, knots = 1, marginal = TRUE)1:trtSuccimer  0.000 -0.699
## lspline(x = semana, knots = 1, marginal = TRUE)2:trtSuccimer  0.000  0.690
##                                                              ls(=s,k=1,m=TRUE)2
## lspline(x = semana, knots = 1, marginal = TRUE)1
## lspline(x = semana, knots = 1, marginal = TRUE)2
## lspline(x = semana, knots = 1, marginal = TRUE)1:trtSuccimer  0.691
## lspline(x = semana, knots = 1, marginal = TRUE)2:trtSuccimer -0.699
##                                                              l(=s,k=1,m=TRUE)1:
## lspline(x = semana, knots = 1, marginal = TRUE)1
## lspline(x = semana, knots = 1, marginal = TRUE)2
## lspline(x = semana, knots = 1, marginal = TRUE)1:trtSuccimer
## lspline(x = semana, knots = 1, marginal = TRUE)2:trtSuccimer -0.987
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.0020271 -0.6888161 -0.1136309  0.5520751  5.8000806
##
## Residual standard error: 4.999256
## Degrees of freedom: 400 total; 395 residual
```

Um modelo spline linear

- Neste modelo linear por partes, as médias dos indivíduos do **grupo placebo** são dadas por

$$\mathrm{E}\,(Y_{ij}) = \beta_1 + \beta_2 Semana_{ij} + \beta_3 (Semana_{ij} - 1)_+,$$

e para o **grupo succimer** as médias são dadas por

$$\mathrm{E}\,(Y_{ij}) = \beta_1 + (\beta_2 + \beta_4) Semana_{ij} + (\beta_3 + \beta_5)(Semana_{ij} - 1)_+.$$

Um modelo spline linear: coeficientes estimados

- As estimativas REML dos coeficientes de regressão para o modelo linear por partes são apresentadas na tabela a seguir.

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

	Estimativa	EP	Z
(Intercept)	26.3422	0.4991	52.78
semana	-1.6296	0.7818	-2.08
I((semana - 1) * (semana >= 1))	1.4305	0.8778	1.63
semana:trtSuccimer	-11.2500	1.0925	-10.30
I((semana - 1) * (semana >= 1)):trtSuccimer	12.5822	1.2279	10.25

Um modelo spline linear: médias estimados

- Quando expressas em termos da resposta média antes e depois da semana 1, as médias estimadas no **grupo placebo** são

$$\begin{array}{rcll} \hat{\mu}_{ij} & = & \hat{\beta}_1 + \hat{\beta}_2 Semana_{ij}, & Semana_{ij} \leq 1; \\ \hat{\mu}_{ij} & = & (\hat{\beta}_1 - \hat{\beta}_3) + (\hat{\beta}_2 + \hat{\beta}_3) Semana_{ij}, & Semana_{ij} > 1. \end{array}$$

- Então, no grupo placebo a inclinação antes da semana 1 é $\hat{\beta}_2 = -1.63$ e, após a semana 1, é $(\hat{\beta}_2 + \hat{\beta}_3) = -1.63 + 1.43 = -0.20$.

Um modelo spline linear: médias estimados

- Similarmente, quando expressas em termos da resposta média antes e depois da semana 1, as médias estimadas no **grupo succimer** são

$$
\begin{aligned}
\hat{\mu}_{ij} &= \hat{\beta}_1 + (\hat{\beta}_2 + \hat{\beta}_4) Semana_{ij}, & Semana_{ij} \leq 1; \\
\hat{\mu}_{ij} &= \hat{\beta}_1 - (\hat{\beta}_3 + \hat{\beta}_5) + (\hat{\beta}_2 + \hat{\beta}_3 + \hat{\beta}_4 + \hat{\beta}_5) Semana_{ij}, & Semana_{ij} > 1.
\end{aligned}
$$

Um modelo spline linear: médias estimados

- Uma maneira direta de se obter as médias estimadas a partir do objeto do modelo é utilizando a função ggpredict do pacote ggeffects.

```
library(ggeffects)

media_chap_df <- ggpredict(mod.spline, terms = c("semana", "trt"))
media_chap_df
```

```
##
## # Predicted values of id
## # x = semana
##
## # trt = Placebo
##
## x | Predicted |   SE |         95% CI
## -------------------------------------
## 0 |     26.34 | 0.50 | [25.36, 27.32]
## 1 |     24.71 | 0.78 | [23.19, 26.23]
## 4 |     24.12 |      |
## 6 |     23.72 |      |
##
## # trt = Succimer
```

Um modelo spline linear: médias estimados

```
##
## x | Predicted |   SE |         95% CI
## ---------------------------------------
## 0 |     26.34 | 0.86 | [24.66, 28.03]
## 1 |     13.46 | 0.88 | [11.74, 15.18]
## 4 |     16.86 |      |
## 6 |     19.13 |      |
##
## Adjusted for:
## * tempo =  2.50
## *    id = 50.50
```

Um modelo spline linear: médias estimados

- As estimativas das médias dos níveis de chumbo no sangue para os grupos placebo e succimer podem ser apresentadas em um gráfico.

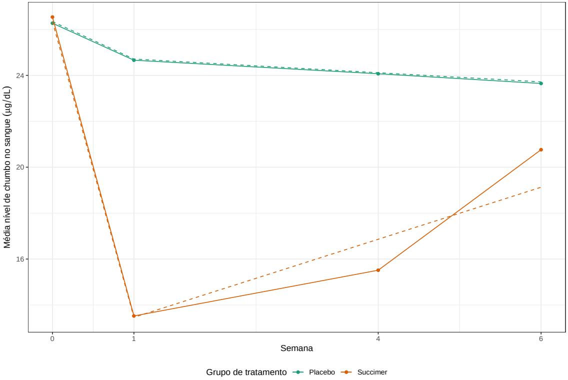
```
df_aux <- as.data.frame(media_chap_df)[c(6, 2, 1)]
names(df_aux) <- c("trt", "chumbo.m", "semana")

p2 <- p + geom_line(data = df_aux,
              linetype = "dashed")
```

Um modelo spline linear: médias estimados

- As médias estimadas (linhas tracejadas) do modelo linear por partes parece adequadamente ajustar os perfis de respostas médias observadas (linhas sólidas) para os dois grupos de tratamento.

```
p2
```



Comparando modelos: linear vs. linear por partes

- Note que o modelo de **tendências lineares** é um caso particular (encaixado) do modelo linear por partes, quando fazemos $\beta_3 = \beta_5 = 0$.
- Podemos utilizar o **teste da razão de verossimilhanças** para comparar os dois modelos.
- **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.

```
# modelo linear por partes
mod.comp <- gls(chumbo ~ semana + I( (semana - 1) * (semana >= 1) ) +
                      semana:trt + I( (semana - 1) * (semana >= 1) ):trt,
                  corr = corSymm(form = ~ tempo | id),
                  weights = varIdent(form = ~ 1 | tempo),
                  method = "ML",
                  data = chumbo.df.longo)

# modelo de tendências lineares
mod.red <- gls(chumbo ~ semana +
                      semana:trt,
                  corr = corSymm(form = ~ tempo | id),
                  weights = varIdent(form = ~ 1 | tempo),
                  method = "ML",
                  data = chumbo.df.longo)
```

Comparando modelos: linear vs. linear por partes

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

```
##          Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## mod.comp     1 15 2466.226 2526.098 -1218.113
## mod.red      2 13 2583.967 2635.856 -1278.983 1 vs 2 121.7409  <.0001
```

- O teste indica que o modelo linear por partes significativamente melhora o ajuste global da resposta média ao longo do tempo quando comparado com o modelo linear.

Comparando modelos: linear por partes vs. quadrático

- Embora os modelos **linear por partes** e **de tendência quadrática** (com intercepto comum para os dois grupos de tratamento) **não sejam encaixados**, ambos têm o mesmo número de parâmetros e, portanto, suas respectivas log-verossimilhanças podem ser comparadas diretamente.

```
# modelo linear por partes
mod.linpartes <- mod.comp

# modelo de tendências lineares
mod.lin <- mod.red

# modelo de tendência biquadrática
mod.quad <- gls(chumbo ~ semana + I(semana^2) +
                    semana:trt + I(semana^2):trt,
                  corr = corSymm(form = ~ tempo | id),
                  weights = varIdent(form = ~ 1 | tempo),
                  method = "ML",
                  data = chumbo.df.longo)
```

Comparando modelos: linear por partes vs. quadrático

```
logLik(mod.linpartes)
```

```
## 'log Lik.' -1218.113 (df=15)
```

```
logLik(mod.quad)
```

```
## 'log Lik.' -1255.842 (df=15)
```

```
logLik(mod.lin)
```

```
## 'log Lik.' -1278.983 (df=13)
```

Exercícios

Exercícios

1. Resolva os exercícios do Capítulo 6 do livro "**Applied Longitudinal Analysis**" (páginas 163 e 164).
   - O arquivo de dados (`rat.dta`) está no Moodle.
2. Com a ajuda do computador realize a análise do exemplo do Estudo Vlagtwedd-Vlaardingen (arquivo de dados FEV1) (páginas 154 à 157).

Avisos

Avisos

- **Próxima aula:** Modelando a estrutura de covariância.
- **Para casa:** ler o Capítulo 6 do livro "**Applied Longitudinal Analysis**".
  - Caso ainda não tenha lido, leia também os Caps. 1, 2, 3, 4 e 5.

Bons estudos!