Modelos aditivos generalizados

Elias T Krainski

Curso de Especialização em
Data Science & Big Data
Universidade Federal do Paraná

September 21, 2018

Introdução

Regressão polinomial

Regressão segmentada

Funções base

Funções base: *splines*

Bases e coeficientes

Modelos aditivos generalizados (*Generalized Additive Models* - GAM)

Introdução

Introdução

Introdução

- Em modelos de regressão temos a média de uma variável resposta a modelada em função de p variáveis preditoras

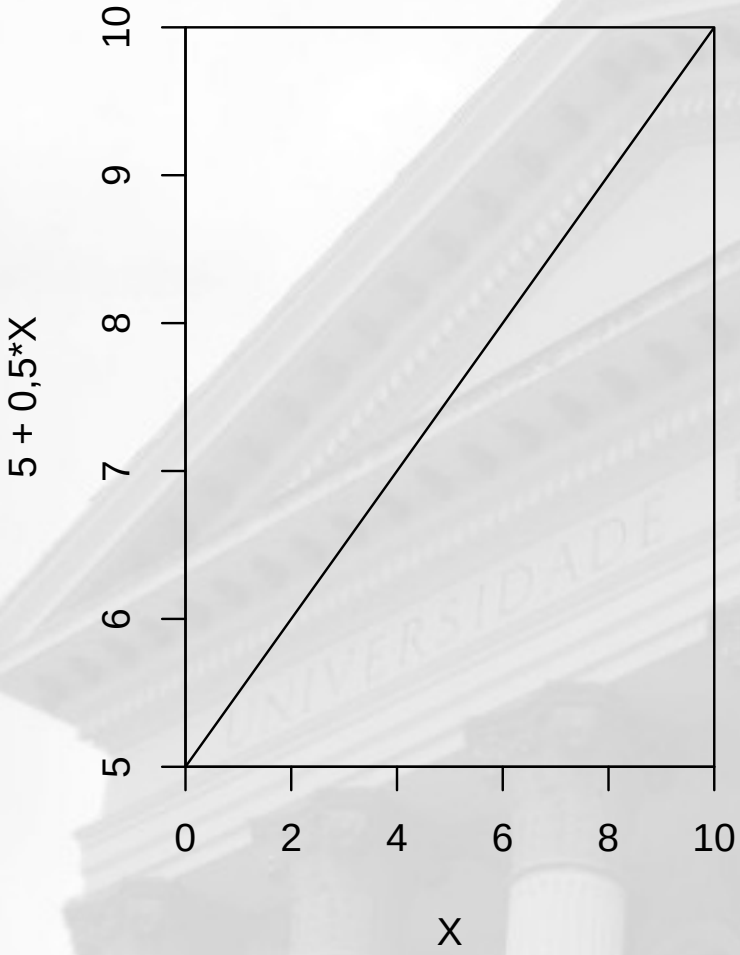
$$\begin{aligned} \mathrm{E}(y) &= g^{-1}(\eta) \\ \eta &= f(x_1, x_2, ..., x_p) \end{aligned} \tag{1}$$

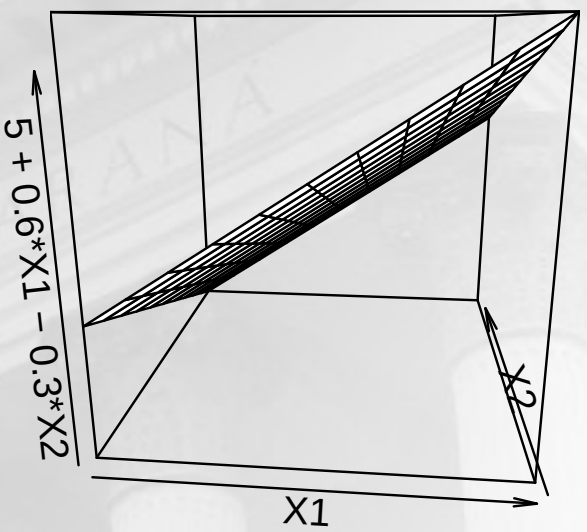
- Os modelos de regressão linear (múltipla) são aqueles onde

$$\eta = \alpha + \beta_1 X_1 + ... + \beta_p X_p$$

- Efeito de X_j é constante (β_j) ao longo dos valores de X_j
- É um hiper-plano p dimensional

Efeito constante: planos





Regressão polinomial

Regressão polinomial

- É possível ter polinômios para acomodar não linearidade. Exemplo:

$$\eta = \alpha + \beta_1 X_1 + \beta_2 X_1^2 + ... + \beta_d X_1^d$$

Regressão polinomial

- É possível ter polinômios para acomodar não linearidade. Exemplo:

$$\eta = \alpha + \beta_1 X_1 + \beta_2 X_1^2 + ... + \beta_d X_1^d$$

- É possível aproximar **qualquer** função por um polinômio**!!!**
- Problema: correlação entre as potências da variável
  - pode-se trocar por polinômios ortogonais
  - melhor alternativa é o uso de funções base

Dias chuvosos em Tokyo

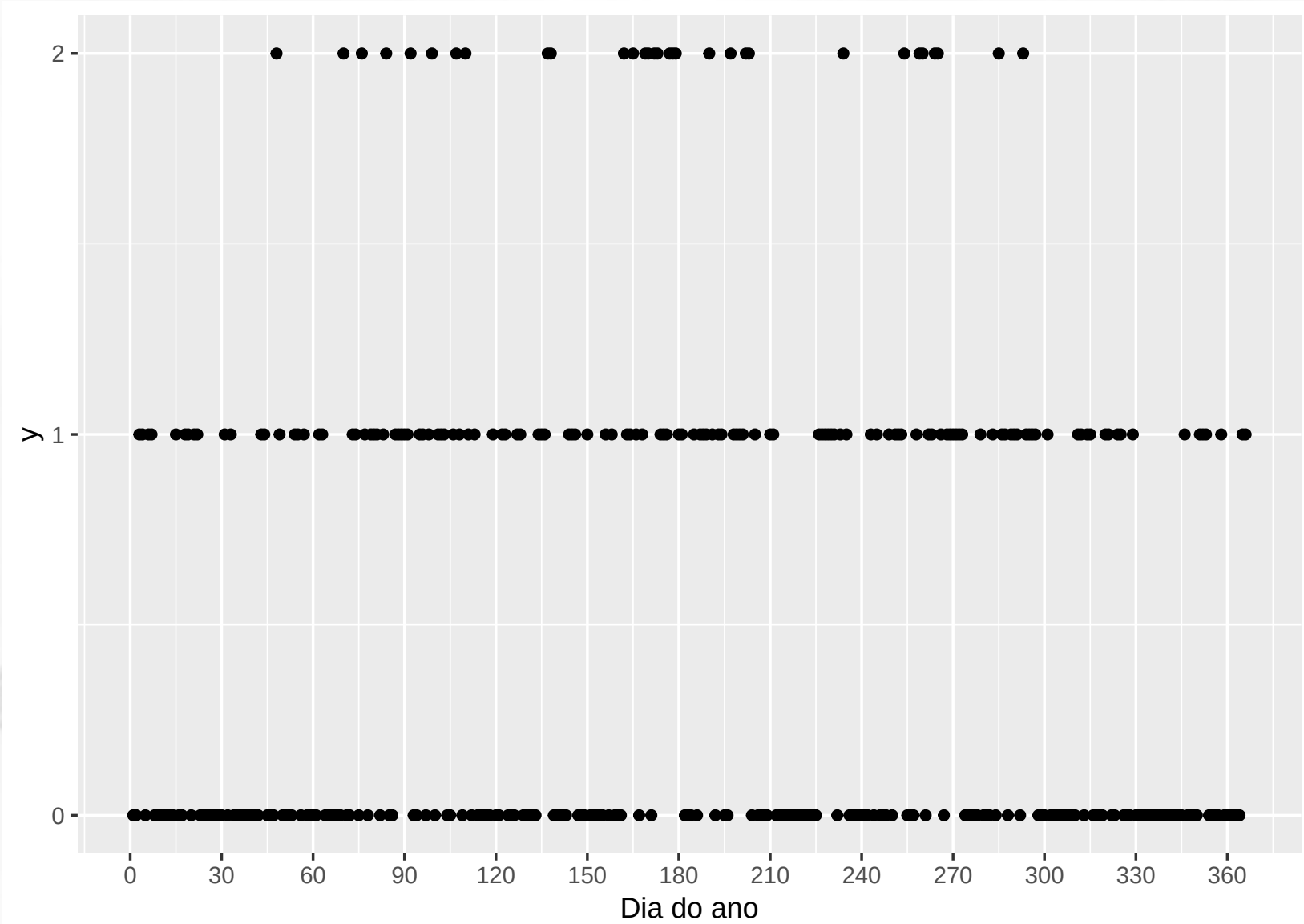


Figure 1: Choveu ou não em cada dia do ano durante dois anos.

Polinômios em t

- polinômio de ordem m

$$\eta_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + ... + \beta_m t_i^m$$

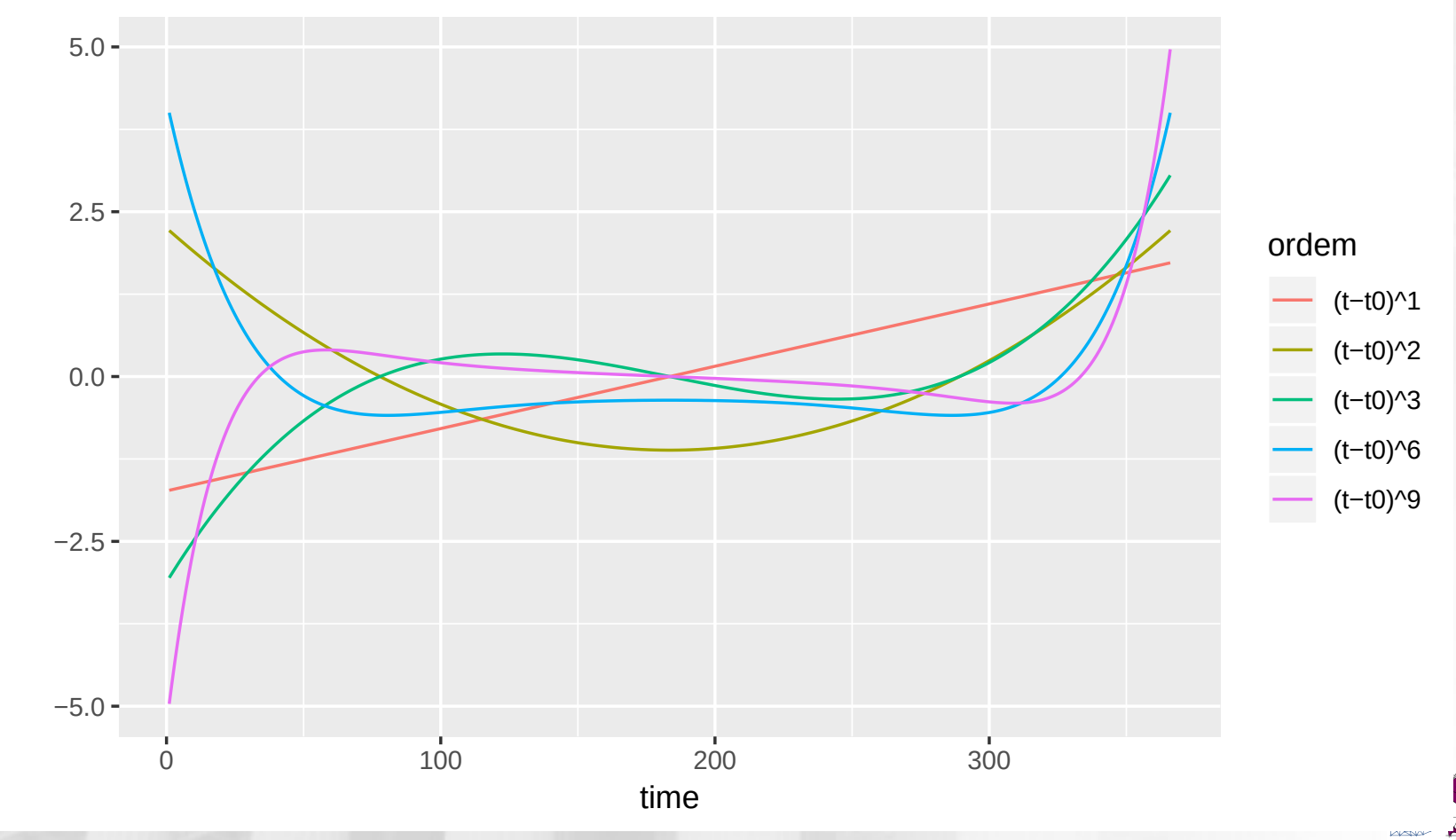
estimar os parâmetros $\beta_j, j = 1, ..., m$

Polinômios em t

- polinômio de ordem m

$$\eta_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + ... + \beta_m t_i^m$$

estimar os parâmetros $\beta_j, j = 1, ..., m$



Exemplo Toquio: considerando polinômios

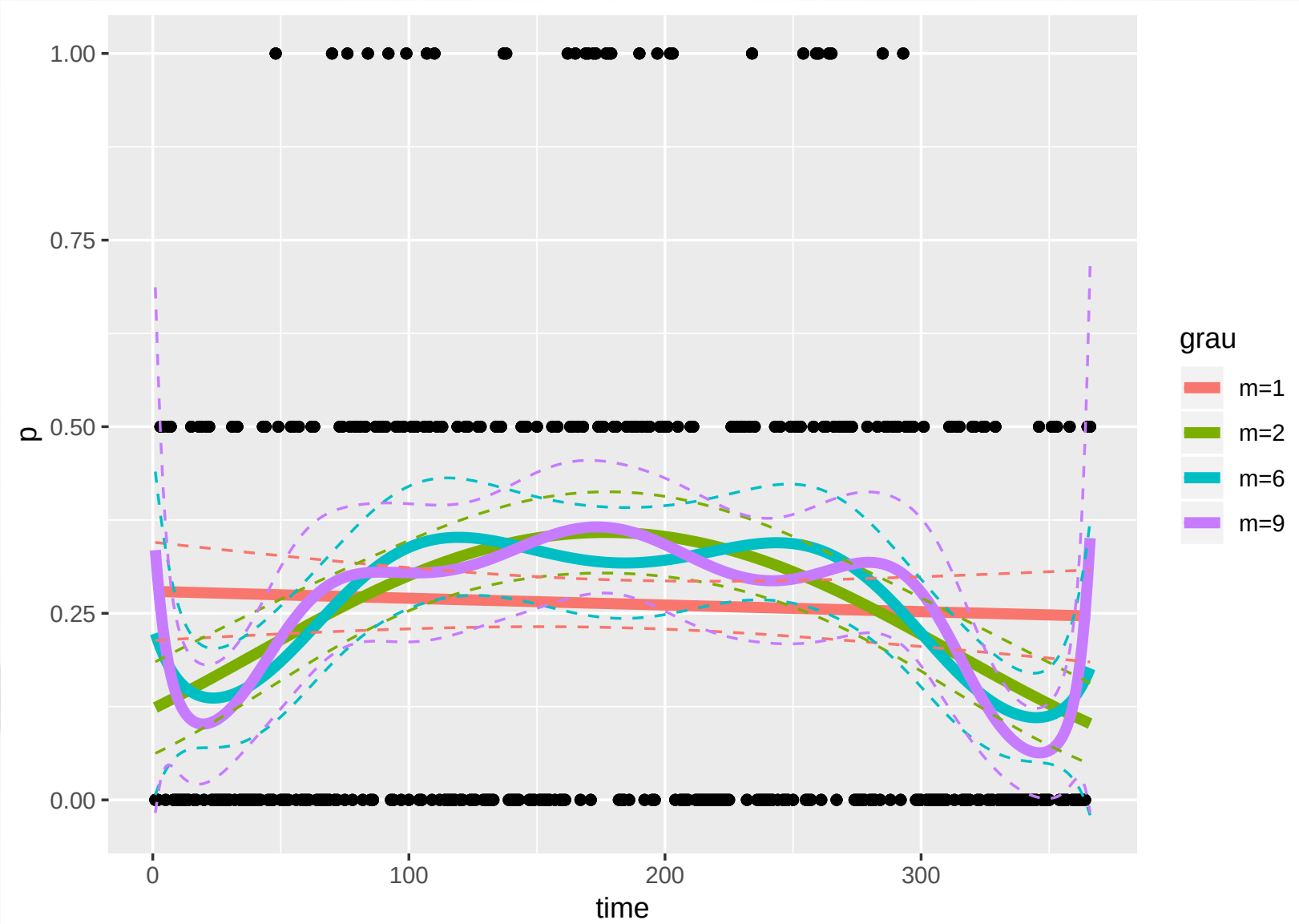


Figure 2: Curvas de predição (+ incerteza), para diferentes graus.

Regressão segmentada

Regressão segmentada

- Suponha o caso em que temos
  - $\eta = a_1 + b_1 x$, se $x \leq c$
  - $\eta = a_2 + b_2 x$, se $x > c$
- Modelo linear por partes, o ponto de corte c pode ser fixo ou estimado

Regressão segmentada

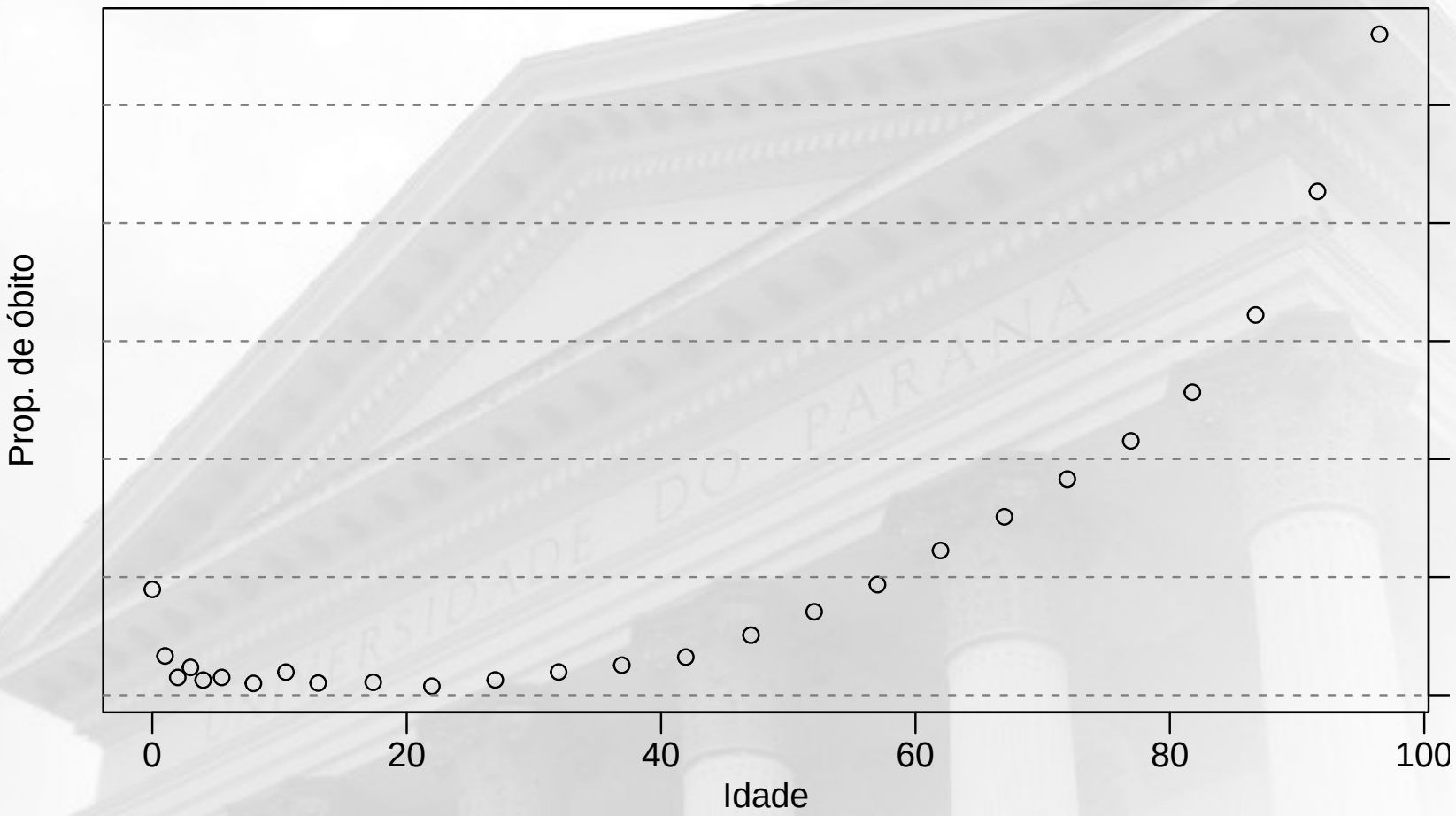
- Suponha o caso em que temos
  - $\eta = a_1 + b_1 x$, se $x \leq c$
  - $\eta = a_2 + b_2 x$, se $x > c$
- Modelo linear por partes, o ponto de corte c pode ser fixo ou estimado
- Esse modelo é equivalente a

$$\eta = \alpha_1 + \beta_1 x + \alpha_2 D + \beta_2 x D$$

onde

- $D = 0$ se $x \leq c$
- $D = 1$ se $x > c$

Exemplo: Morte versus Idade em internações no Paraná em Julho 2018



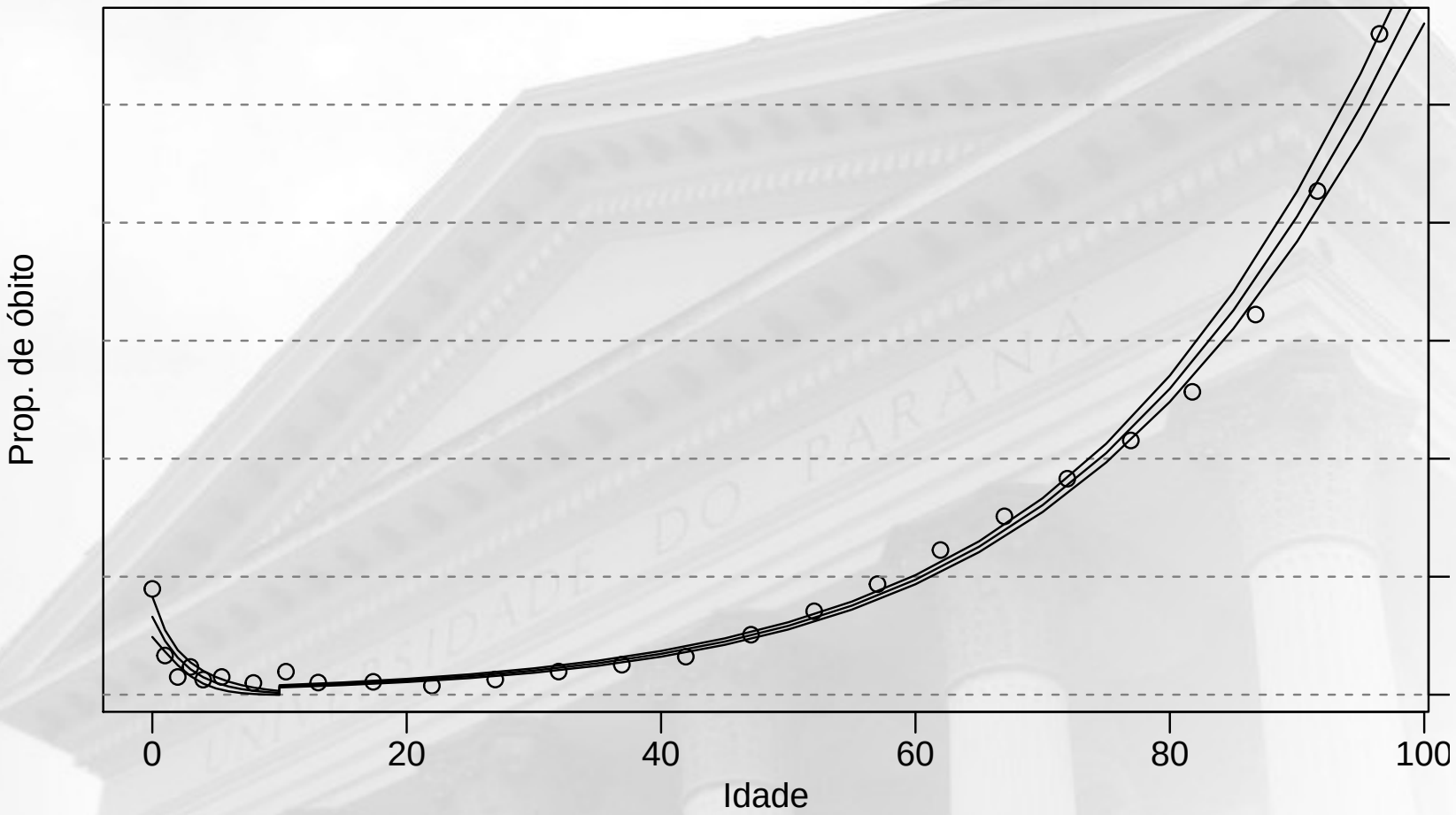
- A relação entre proporção de óbitos e idade é não monótona

Exemplo: Morte versus Idade (cont.)

```
dados <- data.frame(y = rd$MORTE)
dados$a1 <- (rd$IDADE<=7)+0
dados$a2 <- 1 - dados$a1
dados$idade1 <- rd$IDADE*dados$a1
dados$idade2 <- rd$IDADE*dados$a2

m1 <- glm(y ~ 0 + a1 + a2 + idade1 + idade2,
          family=binomial, data=dados)
coef(summary(m1))
##         Estimate Std. Error z value   Pr(>|z|)
## a1     -3.37752   0.134004 -25.205 3.568e-140
## a2     -6.17442   0.077563 -79.606  0.000e+00
## idade1 -0.38037   0.061711  -6.164  7.106e-10
## idade2  0.05338   0.001129  47.300  0.000e+00
```

Exemplo: Morte versus Idade (cont.)



- **Problema**: descontinuidade

Regressão segmentada com restrição

- Restrição: o valor das duas equações igual para $x = c$

$$a_1 + b_1 c = a_2 + b_2 c$$

- Isso é equivalente a ter

$$\eta = \alpha + \beta_1 X + \beta_2 (X - c)_+$$

onde $(X - c)_+$ é a parte positiva de $(X - c)$, ou seja, igual a zero se $X < 0$ e $(X - c)$ caso contrário

Regressão segmentada com restrição

- Restrição: o valor das duas equações igual para $x = c$

$$a_1 + b_1 c = a_2 + b_2 c$$

- Isso é equivalente a ter

$$\eta = \alpha + \beta_1 X + \beta_2 (X - c)_+$$

onde $(X - c)_+$ é a parte positiva de $(X - c)$, ou seja, igual a zero se $X < 0$ e $(X - c)$ caso contrário

- É possível considerar mais de um ponto de corte

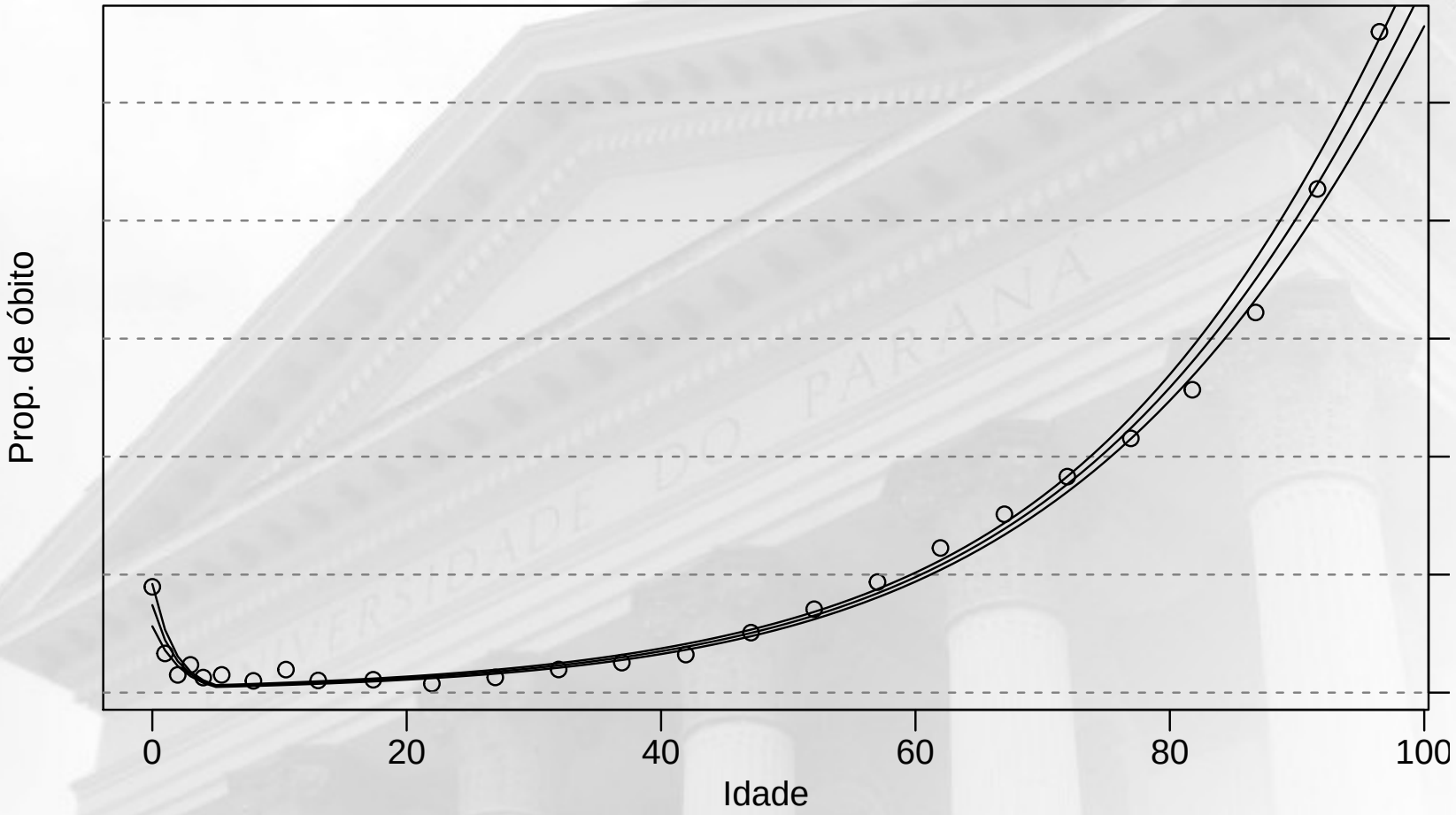
Exemplo: Morte versus Idade (cont.)

```
dados <- rd[c('MORTE', 'IDADE')]
m <- glm(MORTE ~ IDADE, binomial, data = dados)
c1 <- 5
dados$IdadeC <- (rd$IDADE>c1) * (rd$IDADE-c1)
ms <- update(m, .~.+ IdadeC, data=dados)

c(AIC(m), AIC(ms))
## [1] 22452 22216

coef(summary(ms))
##              Estimate Std. Error z value   Pr(>|z|)
## (Intercept)  -3.2571    0.12562  -25.93 3.222e-148
## IDADE        -0.5248    0.02982  -17.60  2.616e-69
## IdadeC        0.5778    0.03041   19.00  1.724e-80
```

Exemplo: Morte versus Idade (cont.)



Funções base

Funções base constantes

- Funções base constante por partes dividem o domínio de x em $k+1$ partes
  - considera $c_1, ..., c_k$ pontos de corte
  - cria uma função indicadora para cada parte:

$$\begin{aligned} b_1(X) &= I(X < c_1) \\ b_j(X) &= I(c_j \leq X < c_{j+1}), \quad j = 2, ..., k-1 \\ b_m(X) &= I(c_k < X) \end{aligned} \tag{2}$$

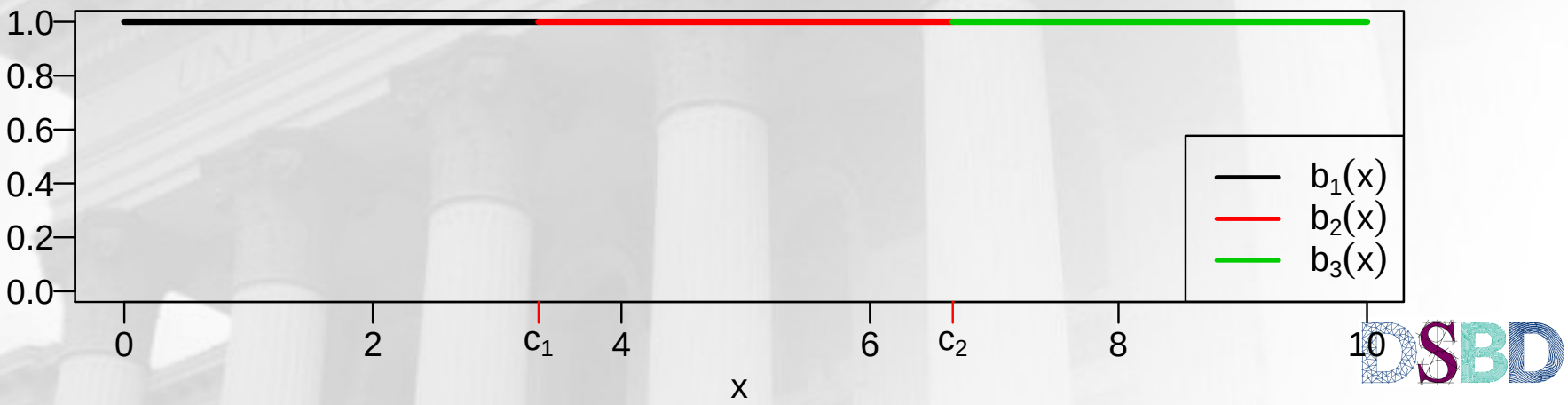
Funções base constantes

- Funções base constante por partes dividem o domínio de x em $k+1$ partes
  - considera $c_1, ..., c_k$ pontos de corte
  - cria uma função indicadora para cada parte:

$$\begin{aligned} b_1(X) &= I(X < c_1) \\ b_j(X) &= I(c_j \leq X < c_{j+1}), \quad j = 2, ..., k-1 \\ b_m(X) &= I(c_k < X) \end{aligned} \tag{2}$$

- Exemplo: três funções bases constantes por partes

$$b_1(X) = I(X < c_1), \quad b_2 = I(c_1 \leq X < c_2), \quad b_3 = I(X \geq c_3)$$



Funções base linear por partes

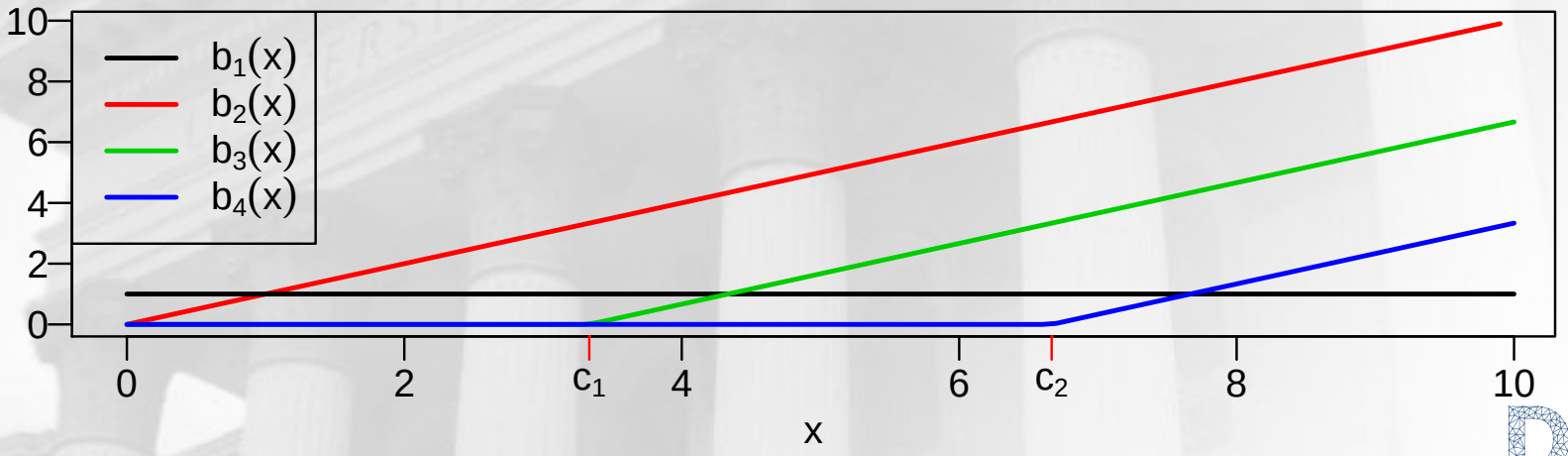
- Divide o domínio de X em $k+1$ partes $(c_1, ..., c_k)$
- Define: $b_j(x) = x^{j-1}, j = 1, ..., k-2$
- Define: $b_{m+j}(x) = b_j(x)(x - c_j)_+$

Funções base linear por partes

- Divide o domínio de X em $k+1$ partes $(c_1, ..., c_k)$
- Define: $b_j(x) = x^{j-1}, j = 1, ..., k-2$
- Define: $b_{m+j}(x) = b_j(x)(x - c_j)_+$

- Exemplo: $k = 2$

$$b_1(x) = 1, \quad b_2(x) = x$$
$$b_3(x) = (x - c_1)_+, \quad b_4(x) = (x - c_2)_+$$



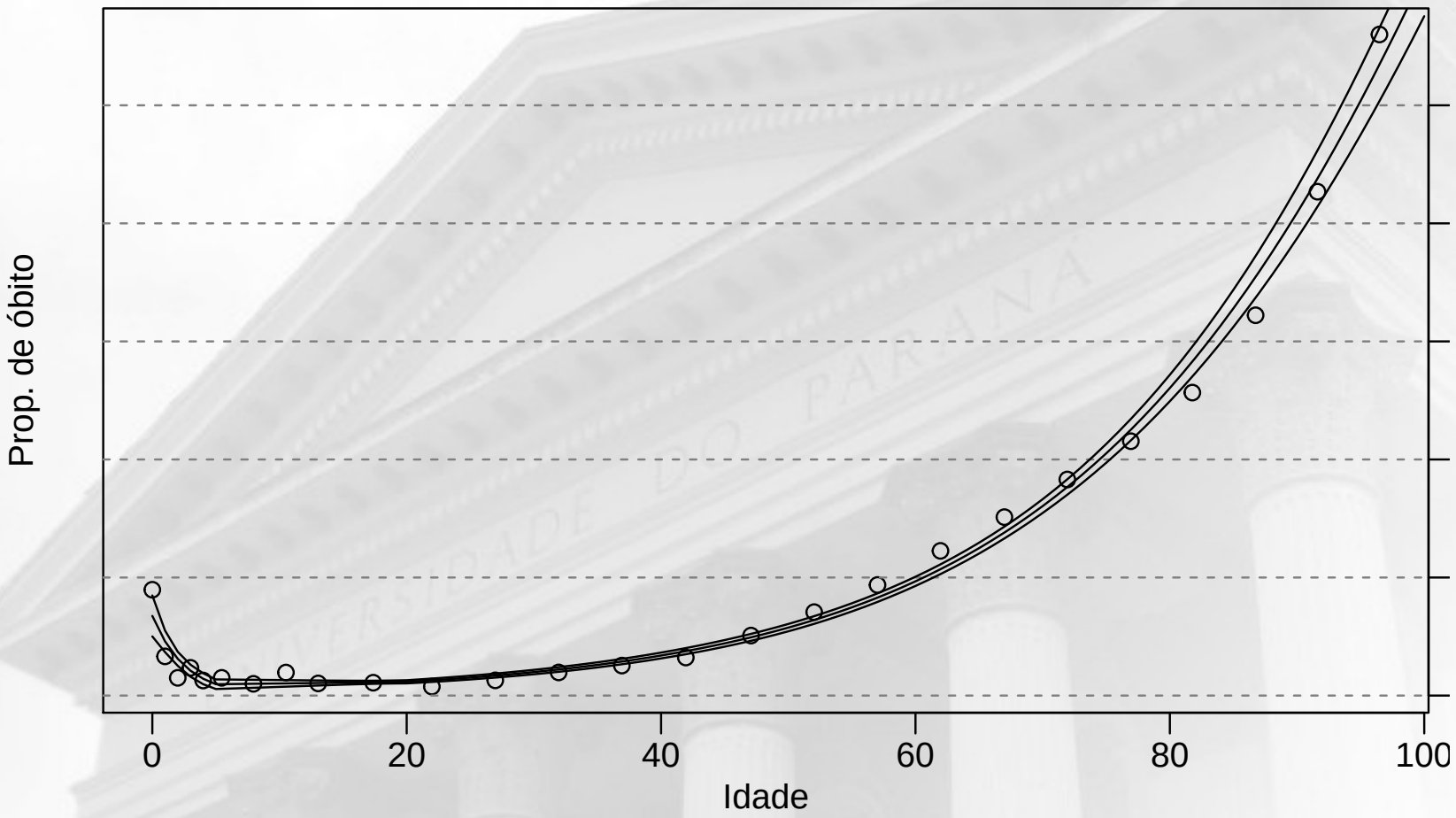
Exemplo: Morte versus Idade

```
c2 <- 20
dados$IdadeC2 <- (rd$IDADE>c2) * (rd$IDADE-c2)
ms2 <- update(ms, .~.+ IdadeC2, data=dados)

c(AIC(m), AIC(ms), AIC(ms2))
## [1] 22452 22216 22212

coef(summary(ms2))
##              Estimate Std. Error z value   Pr(>|z|)
## (Intercept) -3.35492    0.13310 -25.207 3.385e-140
## IDADE       -0.39584    0.05679  -6.970  3.165e-12
## IdadeC       0.40879    0.07018   5.825  5.728e-09
## IdadeC2      0.04107    0.01543   2.663  7.755e-03
```

Exemplo: Morte versus Idade (cont.)



Funções base: *splines*

Funções base: *splines*

- representar/subdividir o espaço da variável

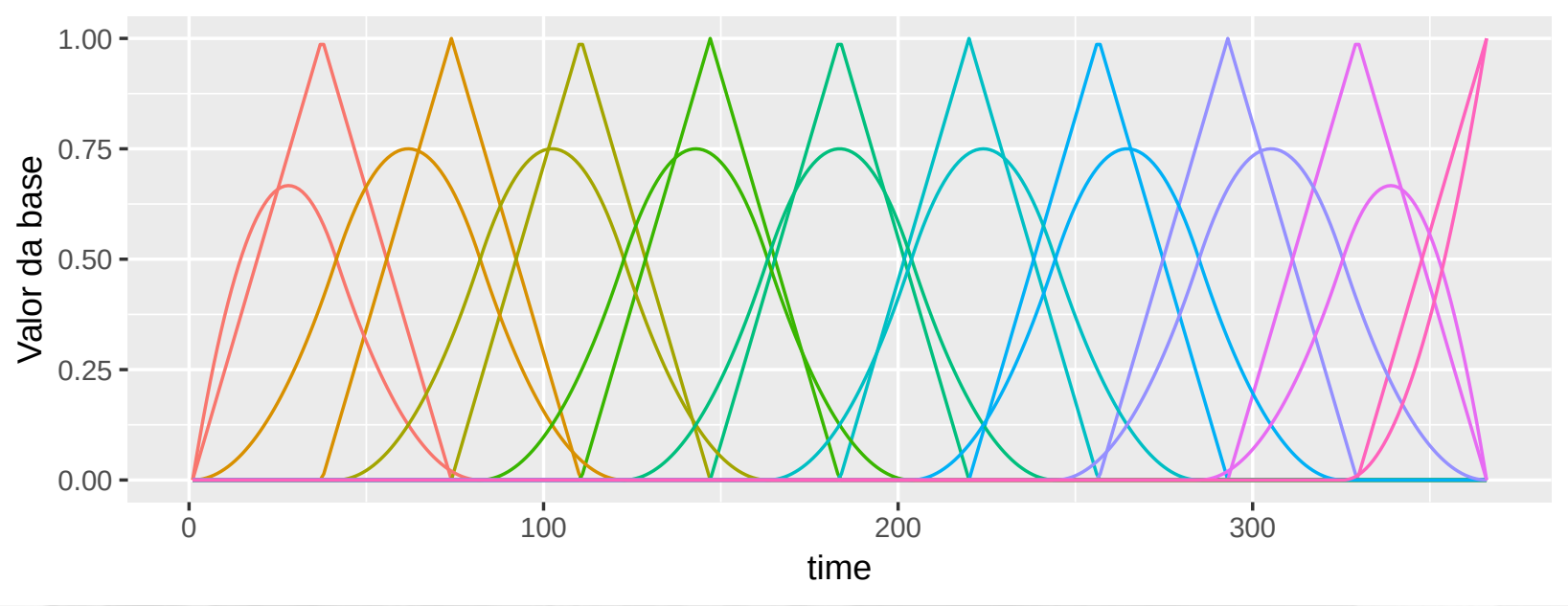


Figure 3: B-splines de 1^0 e 2^0 grau, com 8 graus de liberdade.

Funções base: *splines*

- representar/subdividir o espaço da variável

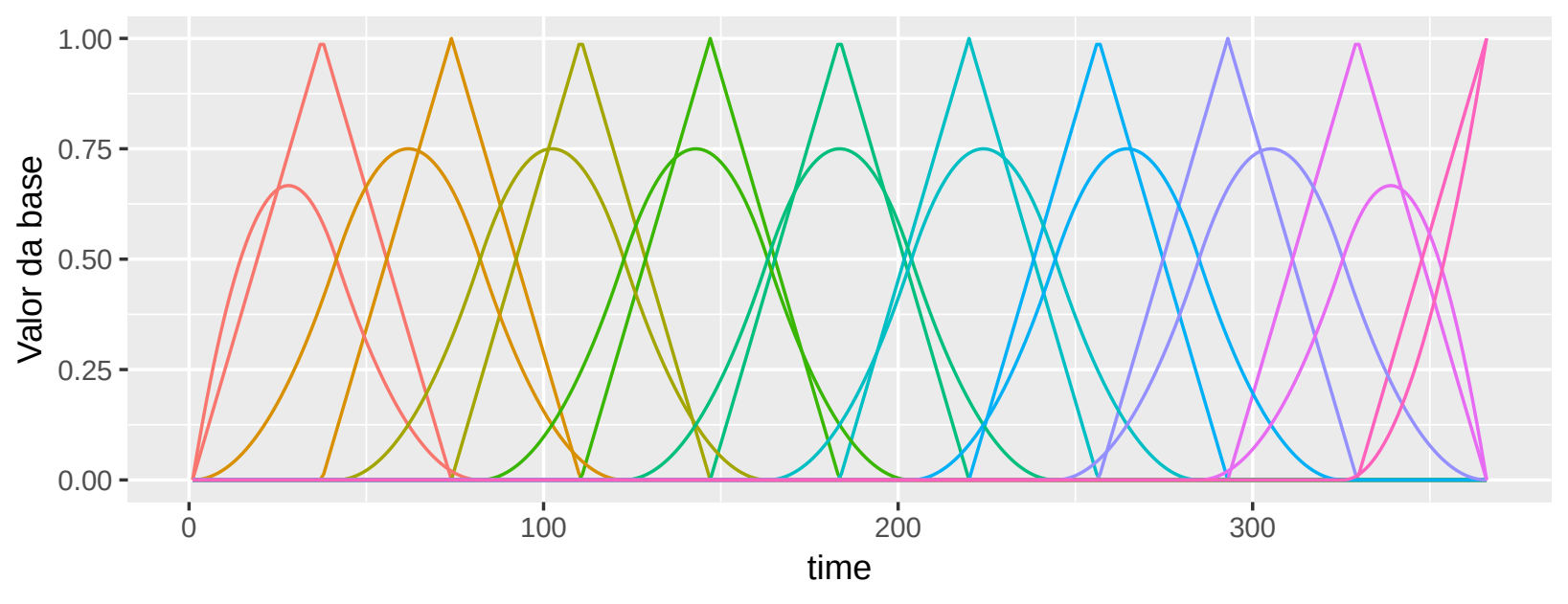


Figure 3: B-splines de 1^0 e 2^0 grau, com 8 graus de liberdade.

- suporte compacto
  - cada função base representa uma parte
  - valores não nulos em parte da variável
  - coeficientes de regressão: ativação naquela parte

Usando B-splines de grau 2

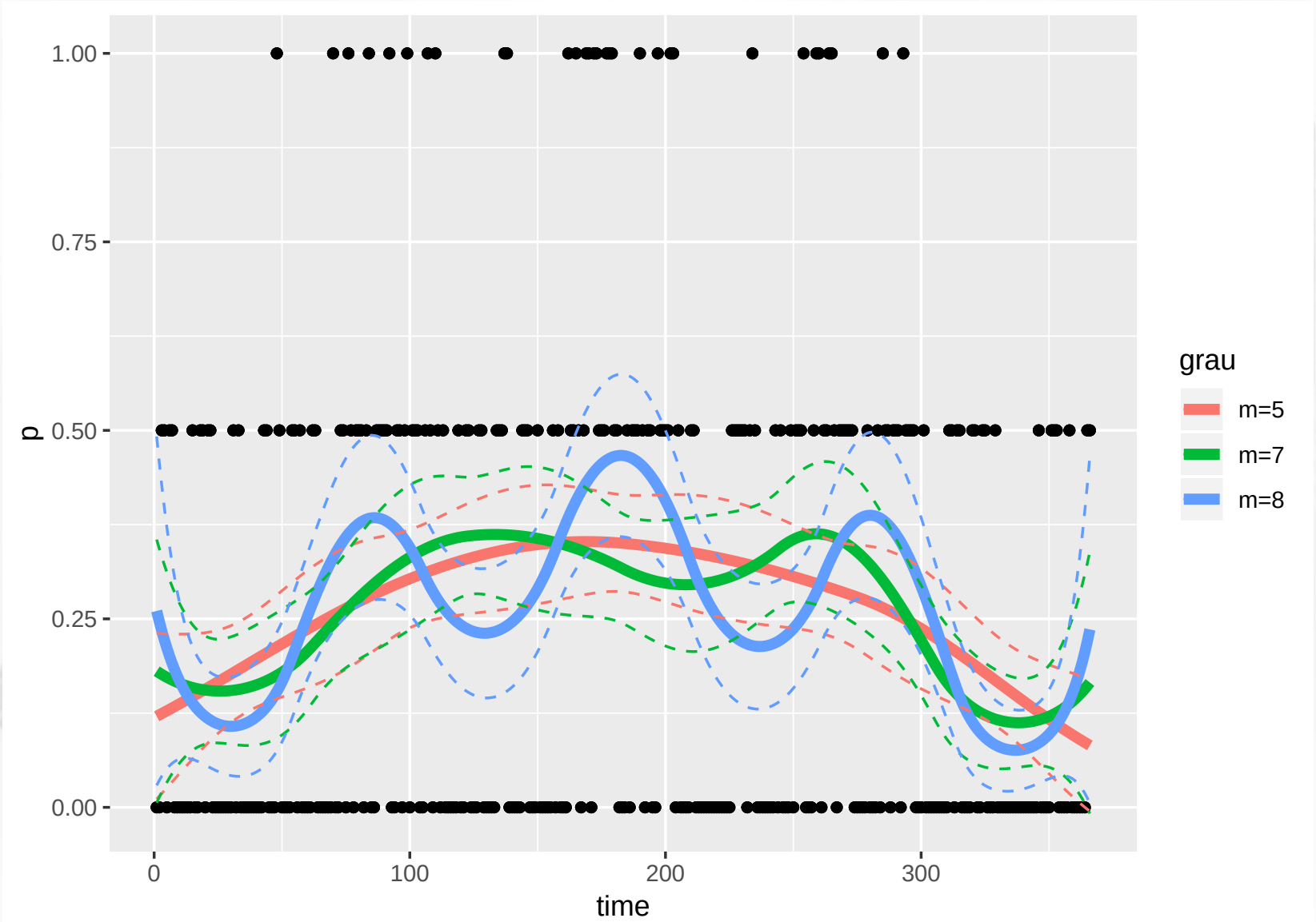


Figure 4: Curvas de predição (e bandas de incerteza), para diferentes graus de liberdade.

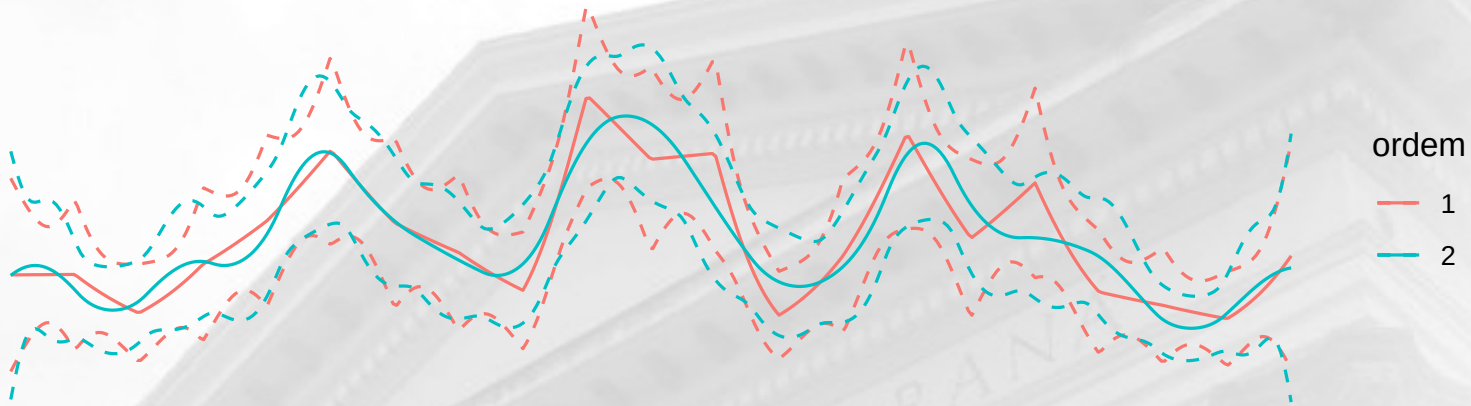
splines

- facilidade maior de interpretação que polinômios
- muitos graus de liberdade: *ajuste excessivo* aos dados
- no limite volta ao caso de análise de cada dia separadamente

Bases e coeficientes

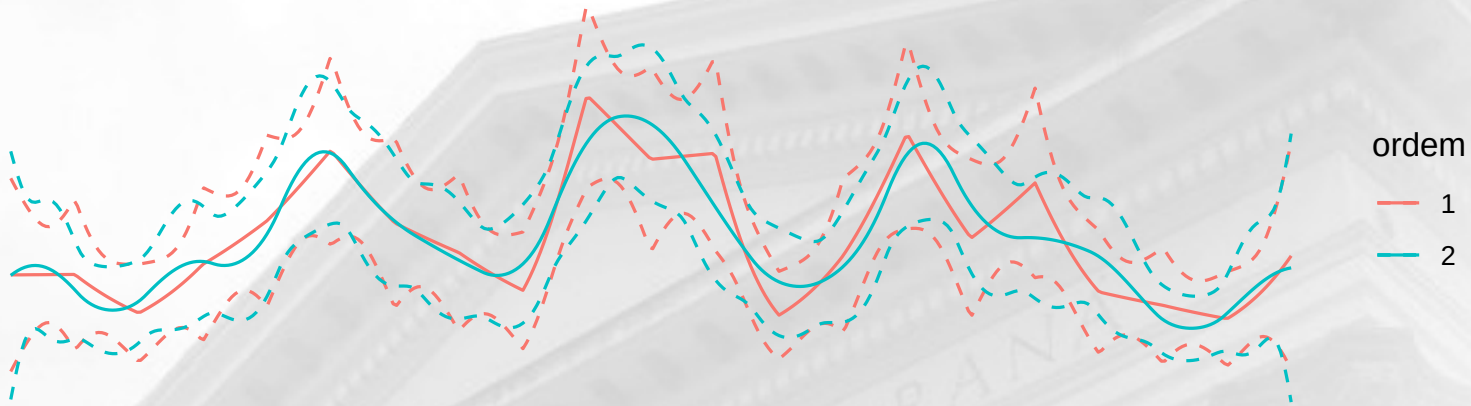
Exemplo: 20 *B-splines* de ordens 1 e 2

- Resultado:

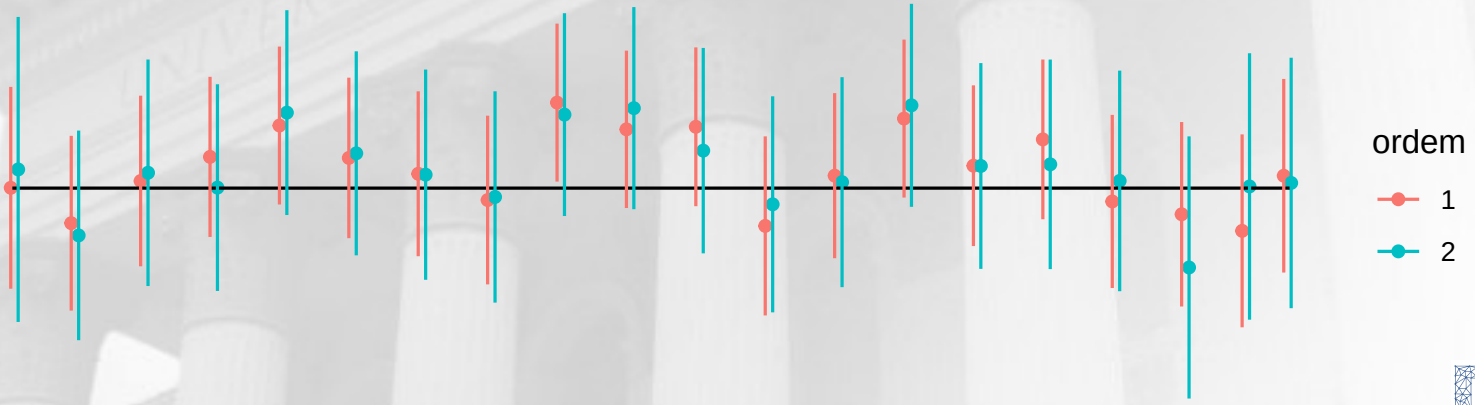


Exemplo: 20 *B-splines* de ordens 1 e 2

- Resultado:

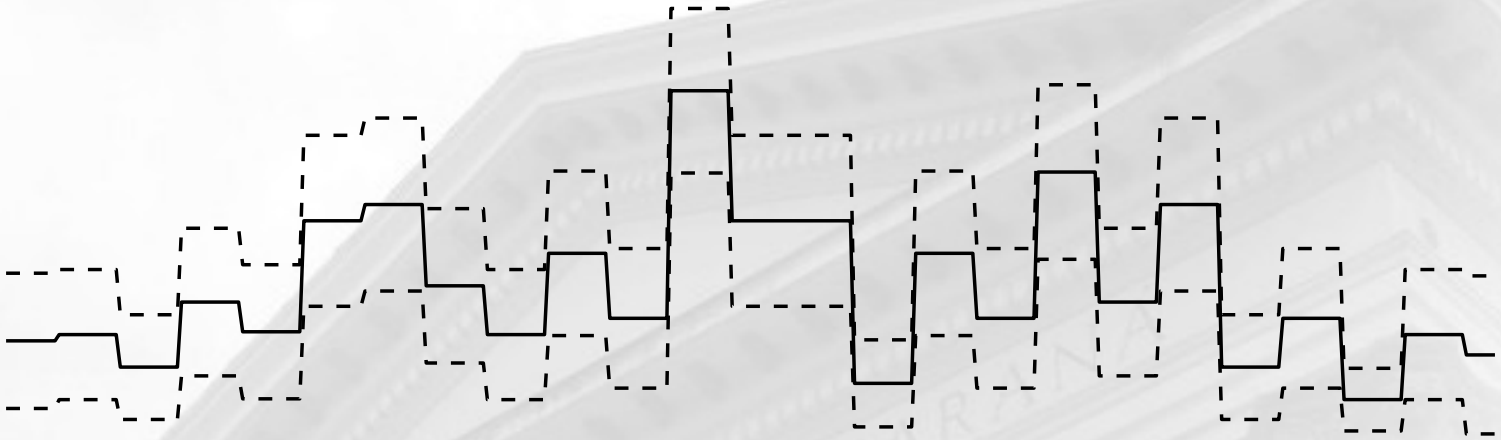


- O que ocorre com os coeficientes:



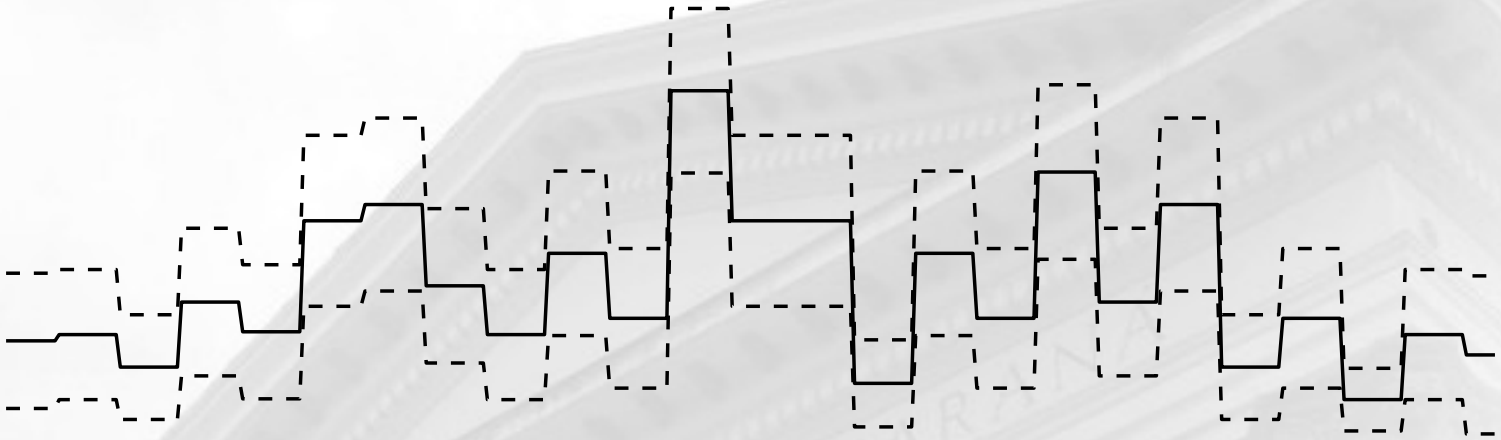
Tempo discretizado (15 dias) como fator

- Resultado:



Tempo discretizado (15 dias) como fator

- Resultado:



- O que ocorre com os coeficientes:

Modelos aditivos generalizados (*Generalized Additive Models* - GAM)

Modelos aditivos generalizados

- Suponha que $f(x_1, ..., x_p)$ pode ser escrita como

$$f(x_1, ..., x_p) = \sum_{j=1}^{p} f_j(x_j)$$

- $f_j(x_j)$ é uma função da covariável x_j
- Cada termo $f_j(x_j)$ é aditivo e é uma função (suave) qualquer

GAM, possibilidades

- Podemos ter

$$\eta = \alpha + \mathbf{Z}\beta + \sum_{j=1}^{k} f_j(x_j)$$

em que $\mathbf{Z}$ é uma matriz de desenho das variáveis com coeficientes de regressão β associados

GAM, possibilidades

- Podemos ter

$$\eta = \alpha + \mathbf{Z}\beta + \sum_{j=1}^{k} f_j(x_j)$$

em que $\mathbf{Z}$ é uma matriz de desenho das variáveis com coeficientes de regressão β associados

- Cada função pode depender de mais de uma variável
  - Exemplo: $f_1(x_1, x_2) + f_2(x_1, x_3) + f_3(x_2, x_3)$

GAM, possibilidades

- Podemos ter

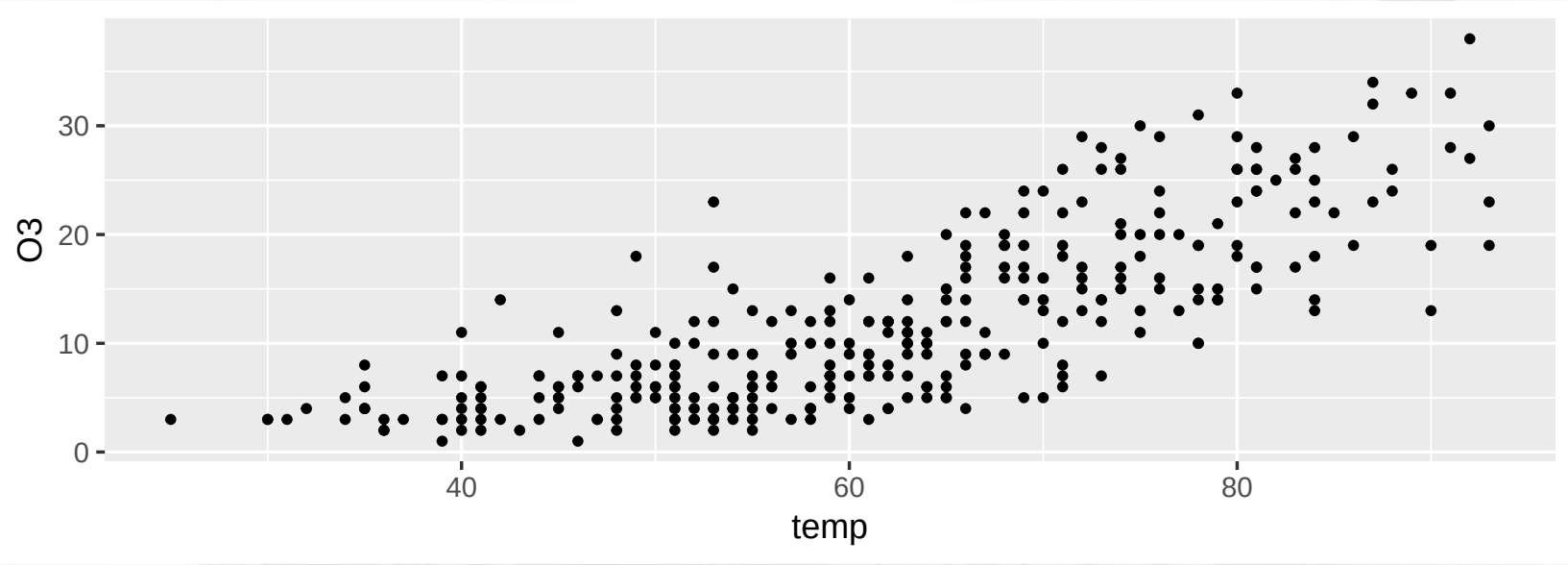
$$\eta = \alpha + \mathbf{Z}\beta + \sum_{j=1}^{k} f_j(x_j)$$

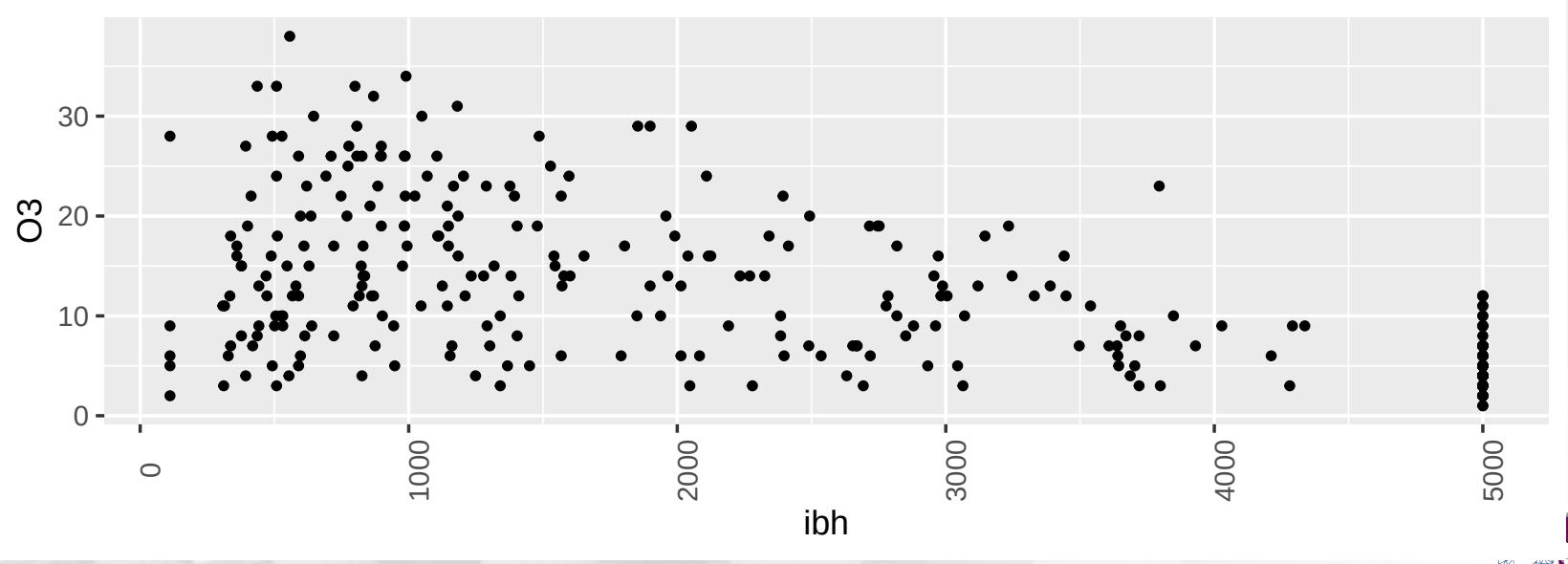
em que $\mathbf{Z}$ é uma matriz de desenho das variáveis com coeficientes de regressão β associados

- Cada função pode depender de mais de uma variável
  - Exemplo: $f_1(x_1, x_2) + f_2(x_1, x_3) + f_3(x_2, x_3)$

- É possível incluir efeitos aleatórios: Modelos aditivos generalizados mistos - GAMM
  - Exemplo: $f_1(idade) + \delta_{indivíduo} + \gamma_{tempo}$

Exemplo: Ozone





Exemplo: Ozone (Cont.)

```
library(mgcv) # pacote que implementa modelos GAM

# Um suave e outro não
am1 <- gam(O3 ~ temp+s(ibh),data=ozone)

# Estima considerando dois termos suaves
am2 <- gam(O3 ~ s(temp)+s(ibh),data=ozone)

# compara
anova(am1, am2, test="F")
## Analysis of Deviance Table
##
## Model 1: O3 ~ temp + s(ibh)
## Model 2: O3 ~ s(temp) + s(ibh)
##   Resid. Df Resid. Dev   Df Deviance    F  Pr(>F)
## 1       323       6950
## 2       319       6054 3.62      896 13.1 3.1e-09 ***
## ---
```

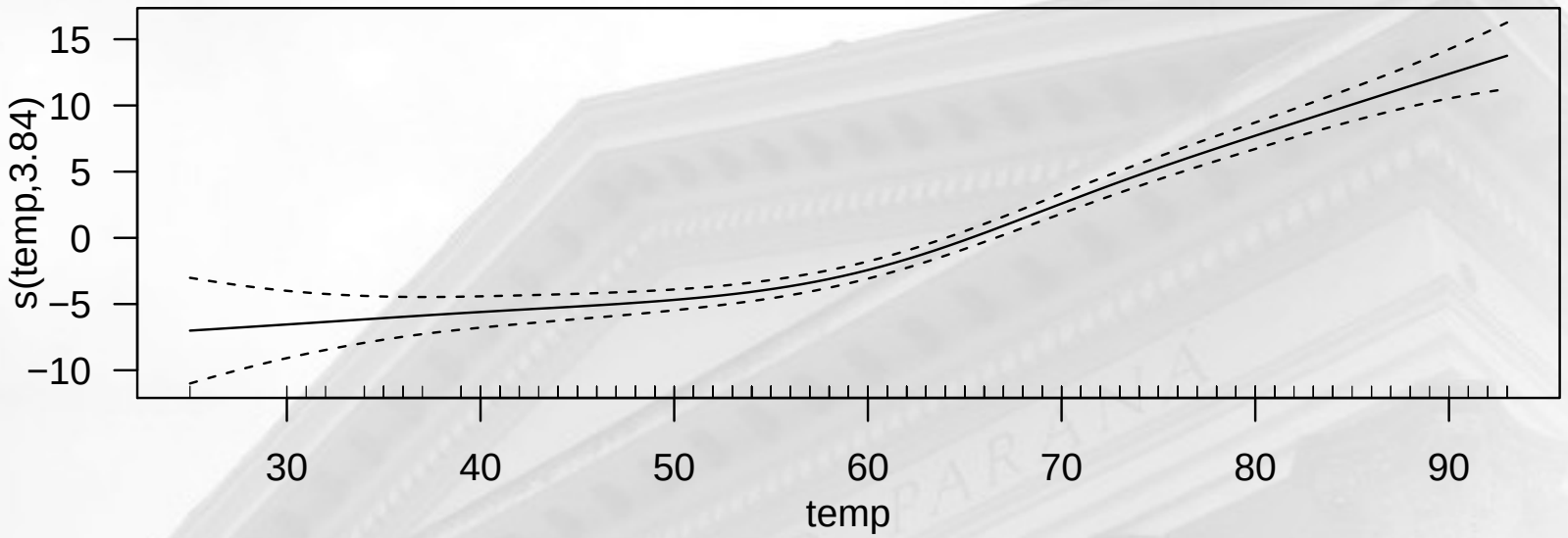
Exemplo: Ozone (Cont.)

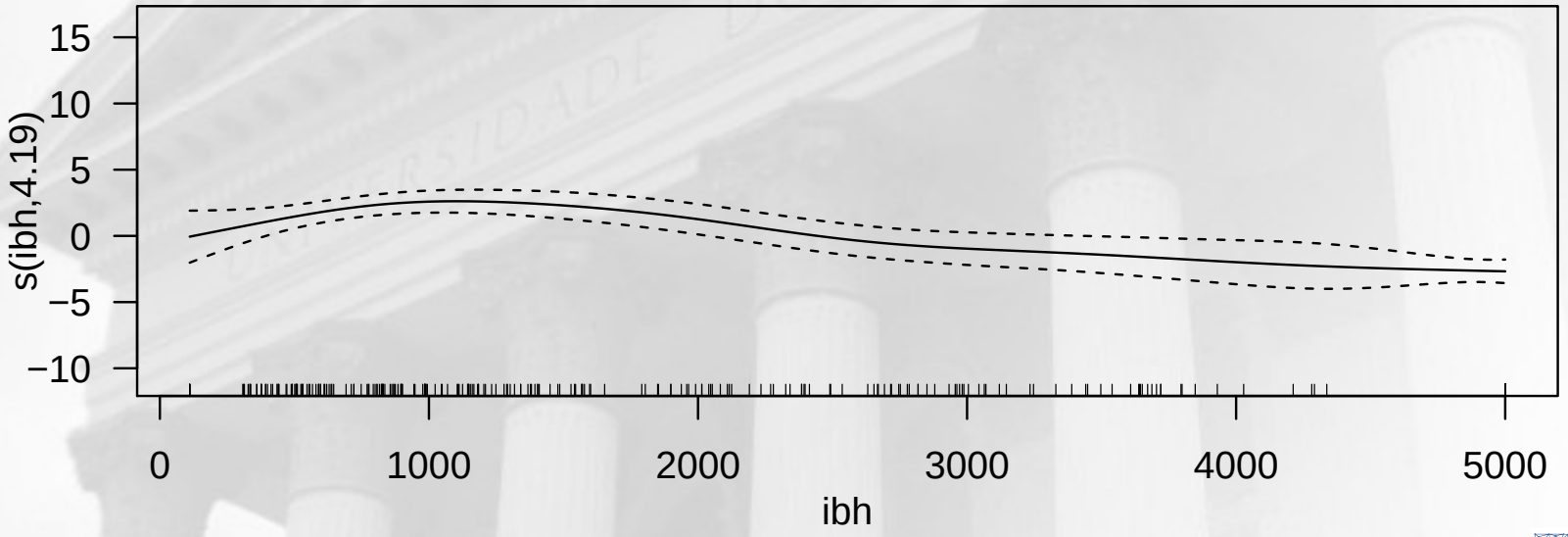
```
summary(am1)
##
## Family: gaussian
## Link function: identity
##
## Formula:
## O3 ~ temp + s(ibh)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.0156     1.3410   -7.47  7.6e-13 ***
## temp          0.3529     0.0213   16.55  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##         edf Ref.df    F p-value
## s(ibh) 4.33   5.26 10.9 5.4e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.665   Deviance explained = 67.1%
## GCV = 21.892  Scale est. = 21.472    n = 330
```

Exemplo: Ozone (Cont.)

```
summary(am2)
##
## Family: gaussian
## Link function: identity
##
## Formula:
## O3 ~ s(temp) + s(ibh)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   11.776      0.239    49.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##          edf Ref.df    F p-value
## s(temp) 3.84   4.79 74.7 < 2e-16 ***
## s(ibh)  4.19   5.09 11.1 4.8e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.706   Deviance explained = 71.3%
## GCV = 19.393  Scale est. = 18.862    n = 330
```

Exemplo: Ozone (Cont.)





Exemplo Bi-dimensional: Ozone

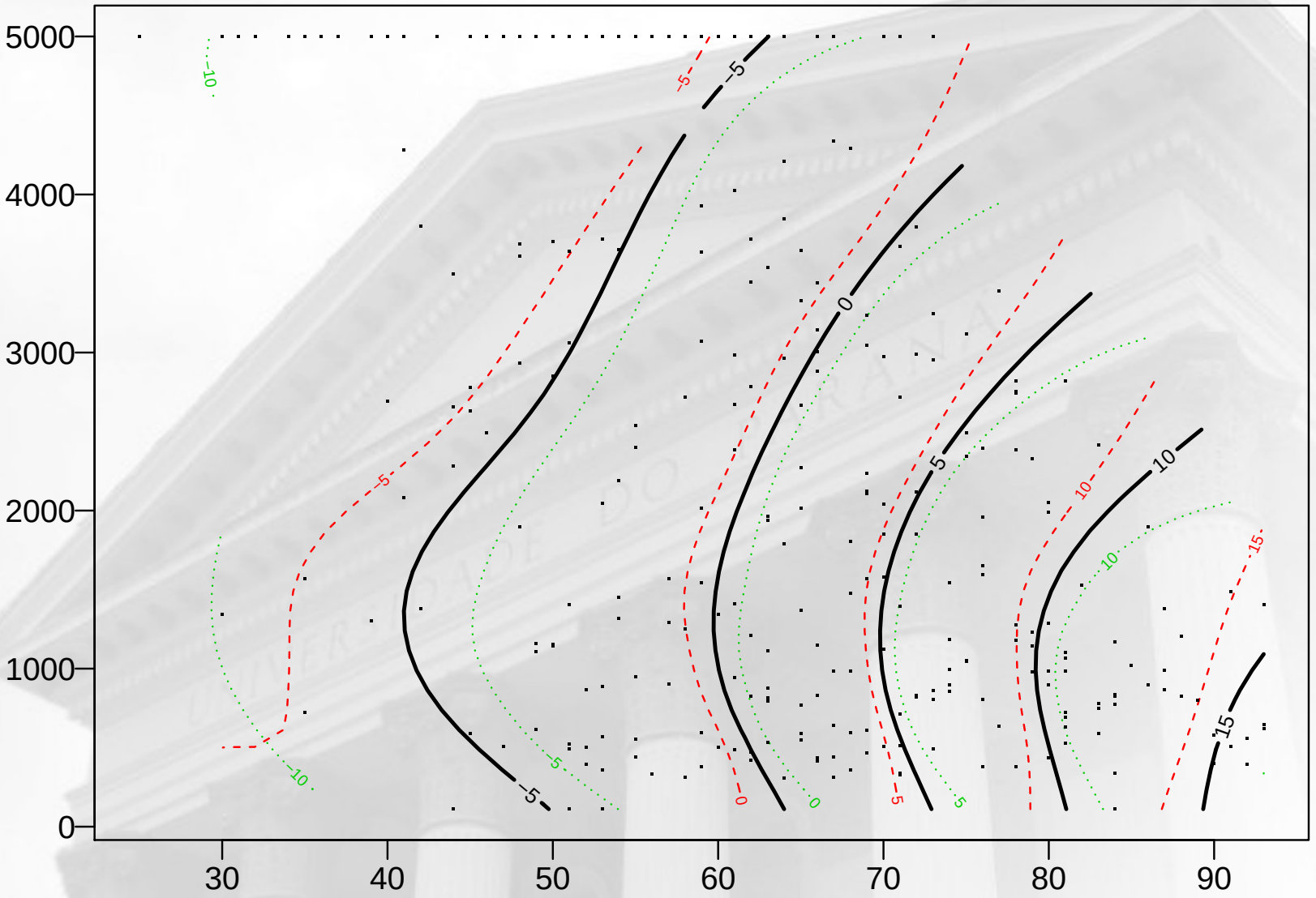
```
amint <- gam(O3 ~ te(temp, ibh), data=ozone)
summary(amint)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## O3 ~ te(temp, ibh)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   11.776      0.238    49.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##               edf Ref.df    F p-value
## te(temp,ibh) 10.5     13 61.4  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.708   Deviance explained = 71.7%
## GCV = 19.439  Scale est. = 18.764    n = 330
```

```
anova(am2, amint, test="F")
```

```
## Analysis of Deviance Table
##
## Model 1: O3 ~ s(temp) + s(ibh)
## Model 2: O3 ~ te(temp, ibh)
##   Resid. Df Resid. Dev   Df Deviance   F Pr(>F)
## 1       319       6054
## 2       316       5977 3.13     76.6 1.3   0.27
```

Exemplo Bi-dimensional: Ozone (Cont.)



Exemplo Bi-dimensional: Ozone (Cont.)