Data Mining

LABDATA

Coordenação:

Prof. Dr. Adolpho Walter Pimazzi Canton

Profa. Dra. Alessandra de Ávila Montini

Disciplina: Machine Learning

Prof. Carlos Eduardo Martins Relvas

Currículo

- Bacharel em Estatística, Universidade de São Paulo.
- Mestre em Estatística, Universidade de São Paulo.
- Doutorando em Ciência da Computação, Universidade de São Paulo.

- Itaú, 2010-2015. Principais atividades:

  - Consultoria estatística para várias áreas do banco com foco principal em melhorias no processo de modelagem de risco de crédito.

  - De 2013 a 2015, participação do projeto Big Data do banco usando tecnologia Hadoop e diversas técnicas de machine learning. Desenvolvemos diversos algoritmos em MapReduce usando R e Hadoop streaming, criando uma plataforma de modelagem estatística no Hadoop.

- Nubank, dede 2015. Equipe de Data Science, responsável por toda a parte de modelagem da empresa, desde modelos de crédito e até mesmo identificar motivos de atendimento.

Agenda

- **Modelos não lineares.**
- **GAM – Generalized Additive models.**

Como tornar o modelo não linear?

- Até o momento só vimos modelos lineares. Os modelos lineares, em geral, são mais simples e trazem a grande vantagem de uma interpretação mais clara. No entanto, em algumas situações, a performance preditiva não é boa o suficiente comparado com outras soluções.

- É possível criar soluções não lineares e continuar usando a regressão linear ou logística. Hoje, veremos alguns destes métodos:
  - Regressão polinomial
  - Step Functions (categorização)
  - Splines
  - GAM (Generalized additive models)

Regressão polinomial

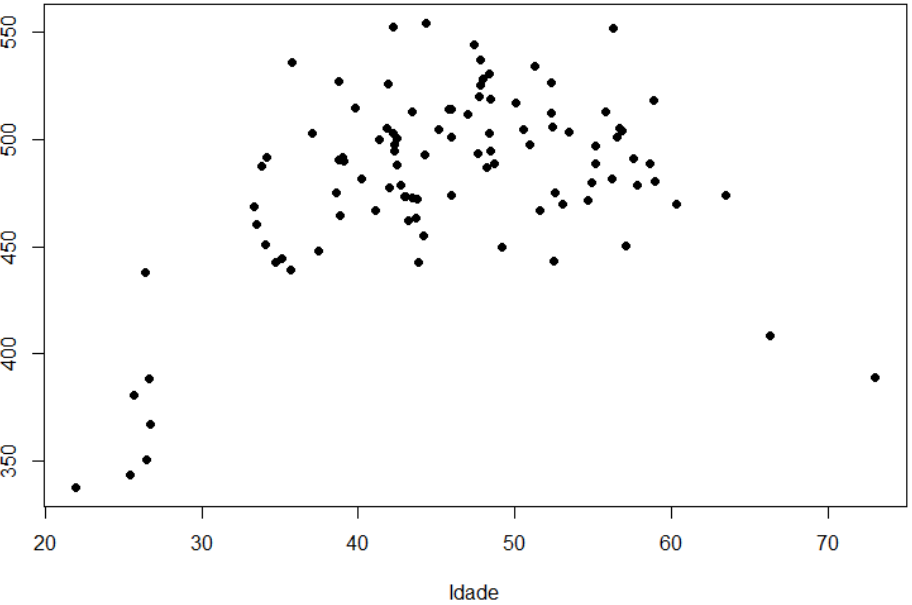
- Podemos extender o modelo de regressão linear simples

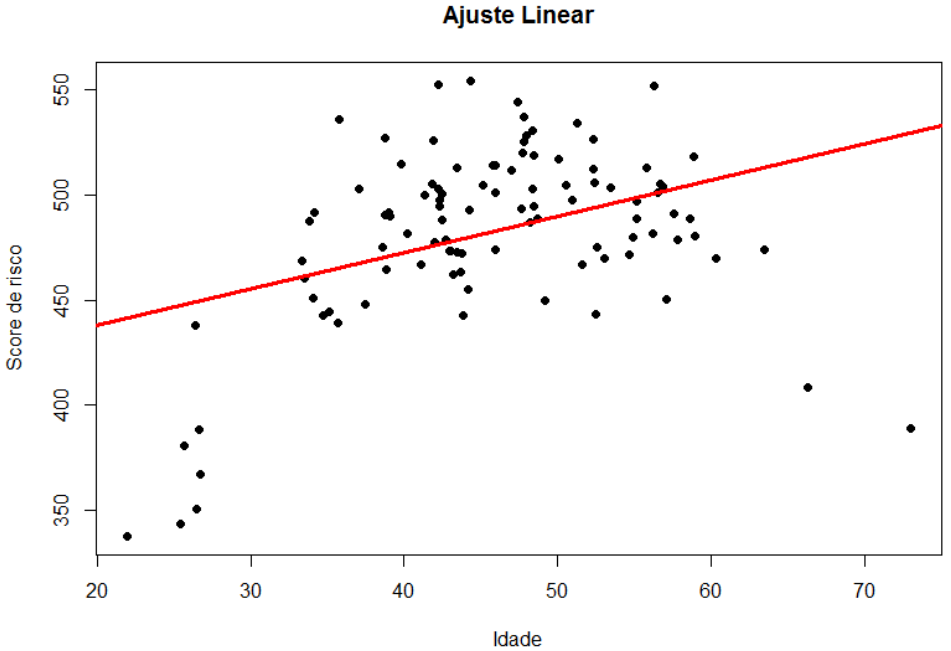
$$\mathrm{Y} = \beta_0 + \beta_1 X$$

Para um polinômio de grau d

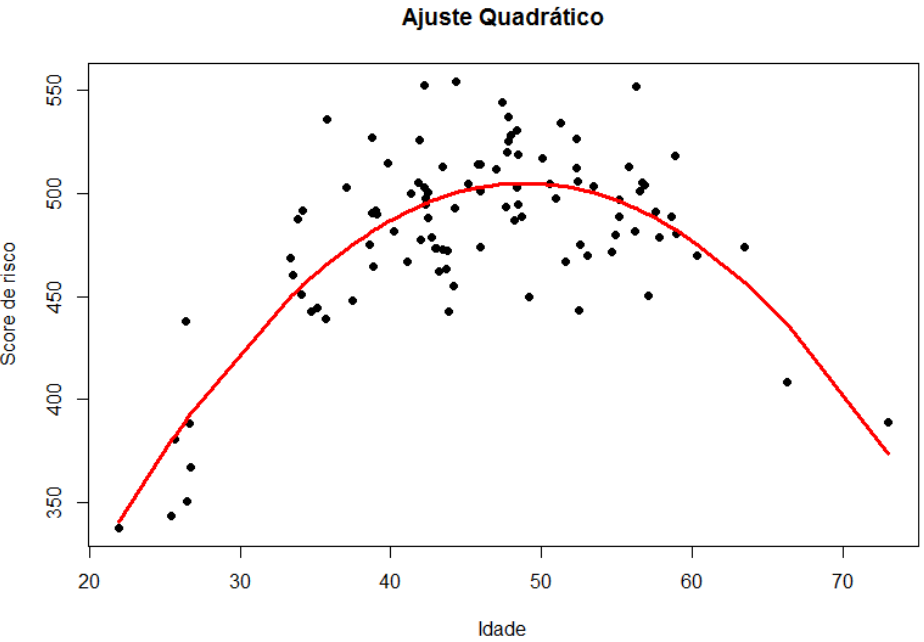
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \cdots + \beta_d X^d$$

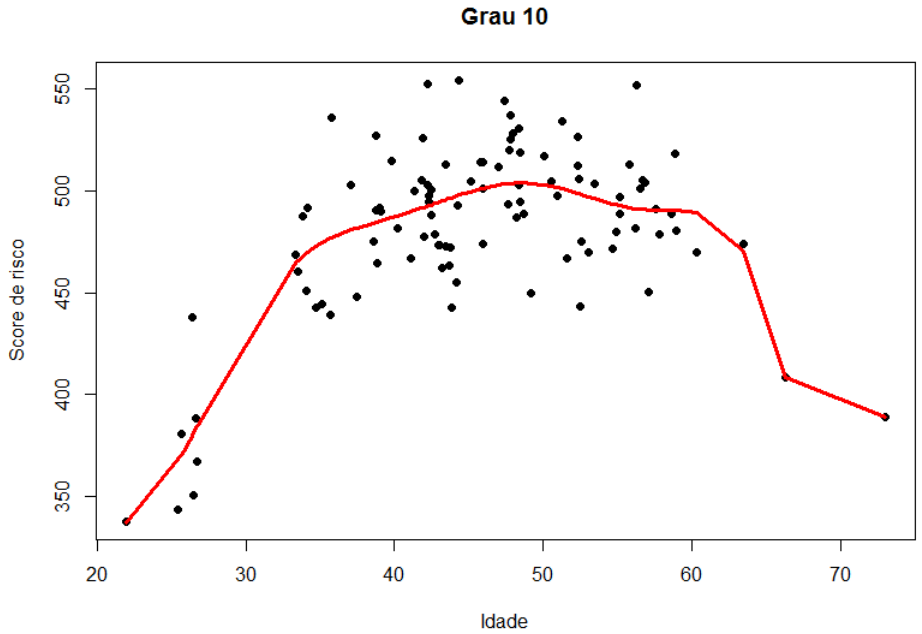
Regressão polinomial





Regressão Polinomial





Regressão polinomial

- Como escolhemos o grau do polinômio corretamente?

➢ Podemos escolher visualmente (e se tivermos 2000 variáveis?)

➢ Podemos usar algum processo de seleção de variáveis vistos anteriormente como stepwise ou lasso.

- Geralmente usa-se o grau de polinômio d no máximo igual a 3 ou 4 para se evitar o overfitting.

Step Functions

- Com a regressão polinomial, temos efeitos não lineares, mas devemos fixar a estrutura global da não linearidade (quadrático, cúbico, etc). Com o uso de step functions, podemos trazer não linearidades sem especificar a estrutura específica.

- Consiste em particionar o espaço em intervalos e ajustar um valor constante para cada intervalo. Seria a categorização de uma variável contínua.

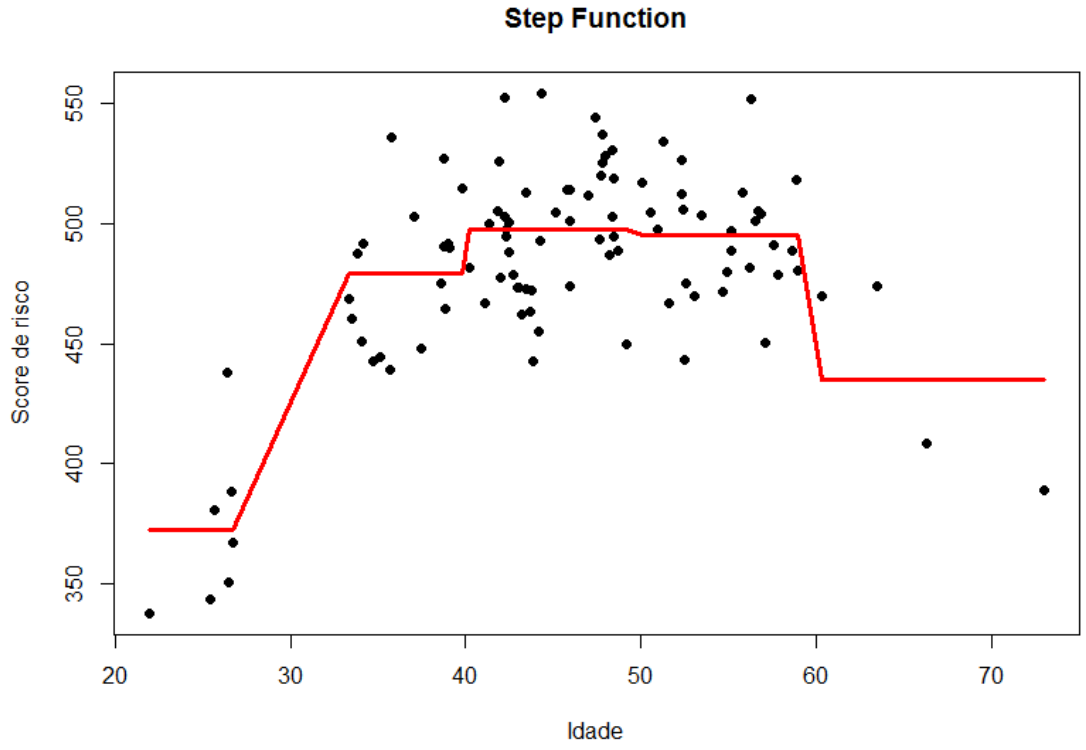
Step Functions

- Escolhemos pontos de corte $(c_1, c_2, c_3, \dots, c_k)$ e com isso construir novas variáveis dummies da seguinte forma:

$$
\begin{aligned}
C_0(X) &= I(X < c_1), \\
C_1(X) &= I(c_1 \leq X < c_2), \\
C_2(X) &= I(c_2 \leq X < c_3), \\
&\vdots \\
C_{K-1}(X) &= I(c_{K-1} \leq X < c_K), \\
C_K(X) &= I(c_K \leq X),
\end{aligned}
$$

- Em que $I(X < c_1)$ representa uma função indicadora que recebe o valor 1 se $X < c_1$ e 0 caso contrário (variável dummy).

Step Functions



Step Functions

- Como podemos escolher os pontos de corte? Se escolhermos muitos pontos, podemos ter um overfitting. Se escolhermos poucos pontos, podemos ter um underfitting. E onde colocamos estes pontos?

- Algumas técnicas encontradas na literatura são:

    - Análise visual
    - Percentis
    - Por meio de uma árvore de decisão.

Laboratório

Laboratório

- Base de dados dos distritos de Boston.
- O objetivo é prever valor médio das casas.
- Variáveis:

  Crime per capita, proporção de terrenos com mais de 25.000 pés ao quadrado, proporção de area de comércio, dummy se faz divisa com rio, concetração de nitrogênio, número médio de quartos, proporção de casas ocupadas antes de 1940, distância média para os centros de emprego de Boston, acessibilidade a estradas, taxa de imposto por propriedade, razão aluno professor, proporção de negros, valor médio das casas.

Exercício

- Base de dados sobre a qualidade do ar em Nova York

- Prever a concentração de Ozônio no ar de Nova York

- Variáveis:
- Ozônio
- Temperatura
- Vento
- Intensidade do sol

Exercício

- Carregue a base de dados

```
library(datasets)
data(airquality)
df <- airquality
```

- Remova os valores ausentes por meio do comando (df <- na.omit(df))
- Crie as amostras de treino e teste (30%) usando seed de 42.
- Faça uma análise descritiva para entender a relação das variáveis.
- Compare um modelo linear, polinomial e utilizando step functions.
- Interprete estes modelos.

Splines

- Ao invés de ajustar um polinômio com grau elevado, podemos ajustar polinômios de graus menores em partições da variável explicativa.

- Ou seja, podemos ajustar o seguinte modelo:

$$y_i = \begin{cases} \beta_{01} + \beta_{11} x_i + \beta_{21} x_i^2 + \beta_{31} x_i^3 + \epsilon_i & \text{if } x_i < c; \\ \beta_{02} + \beta_{12} x_i + \beta_{22} x_i^2 + \beta_{32} x_i^3 + \epsilon_i & \text{if } x_i \geq c. \end{cases}$$

- Em que o ponto c é chamado de nó (knot).

- Podemos ter vários nós além de escolher qualquer grau de polinômio.

Splines

- Só precisamos forçar que cada polinômio gere uma função contínua. Assim escrevemos o modelo, por exemplo, o cúbico, como:

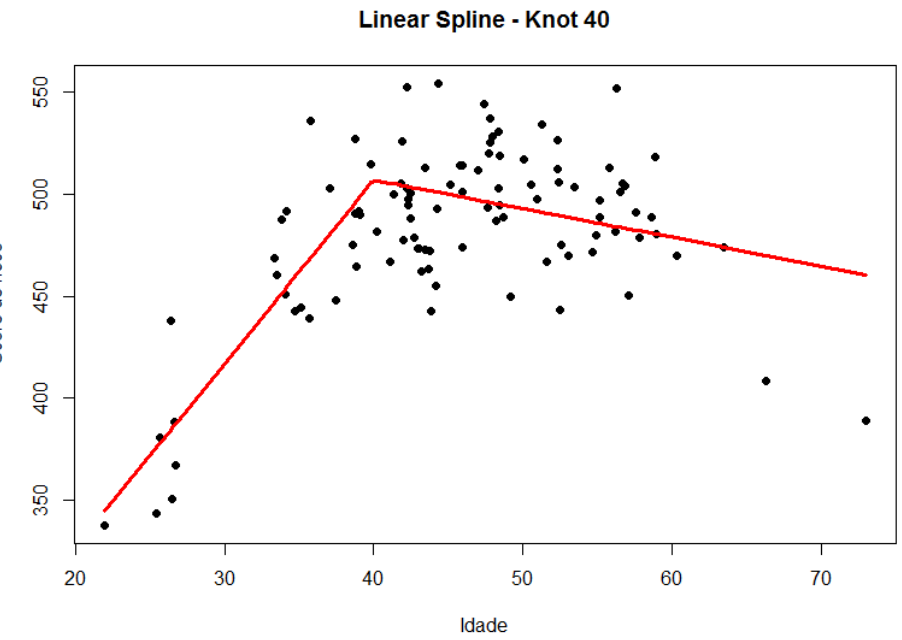
- $Y = \beta_o + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4\, h(X,\ c_1) + \cdots + \beta_m\, h(X,\ c_k)$

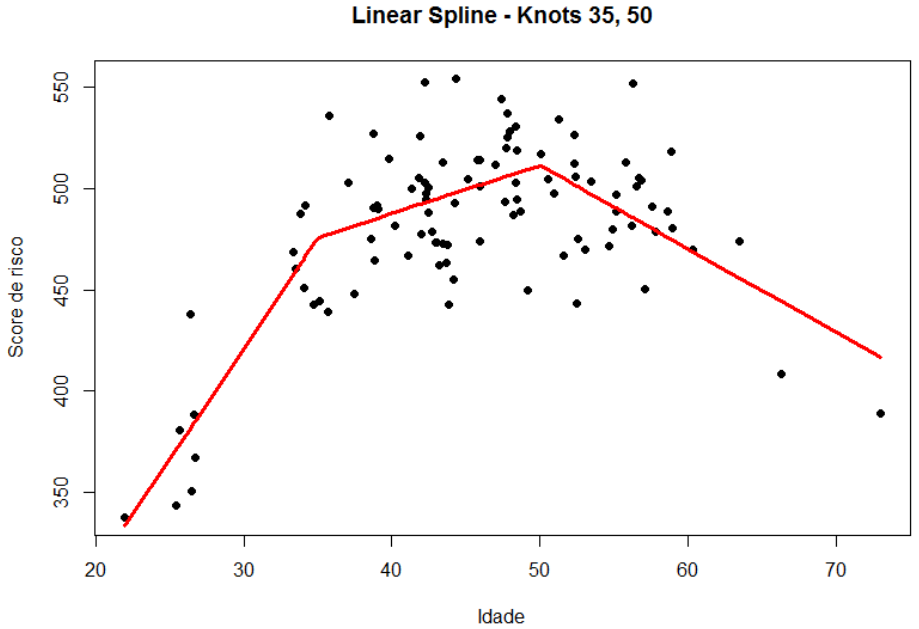
- Em que c_1, ..., c_k são os k nós e

$$h(X, c_k) = (X - c_k)_+^3 = \begin{cases} (X - c_k)^3, se\ X > c_k \\ 0, caso\ contrário \end{cases}$$

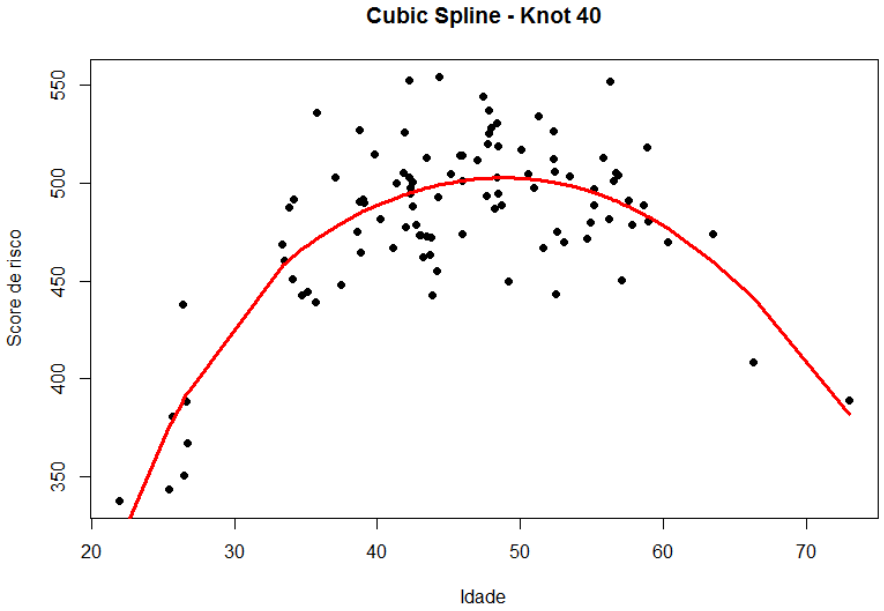
- O spline chamado de natural spline requer que o primeiro e o último polinômios sejam lineares (devido a variância).

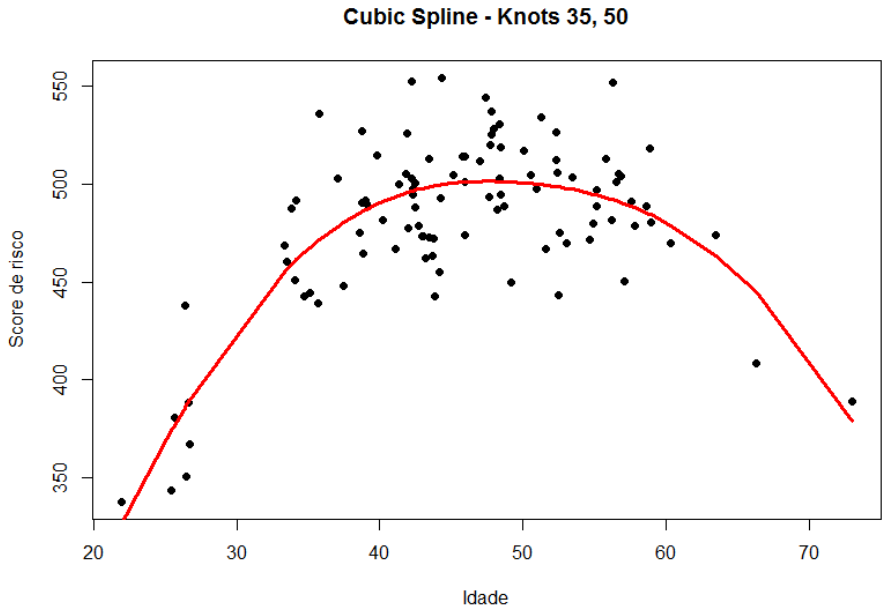
Splines


Linear Spline - Knot 40
350
400
450
500
550
20
30
40
50
60
70
Idade


Linear Spline - Knots 35, 50
Score de risco
350
400
450
500
550
20
30
40
50
60
70
Idade

Splines


Cubic Spline - Knot 40
Score de risco
350
400
450
500
550
20
30
40
50
60
70
Idade


Cubic Spline - Knots 35, 50
Score de risco
350
400
450
500
550
20
30
40
50
60
70
Idade

Splines

- Quanto mais nós e mais graus do polinômio, maior a flexibilidade do modelo em se ajustar aos dados (maior o risco de overfitting).

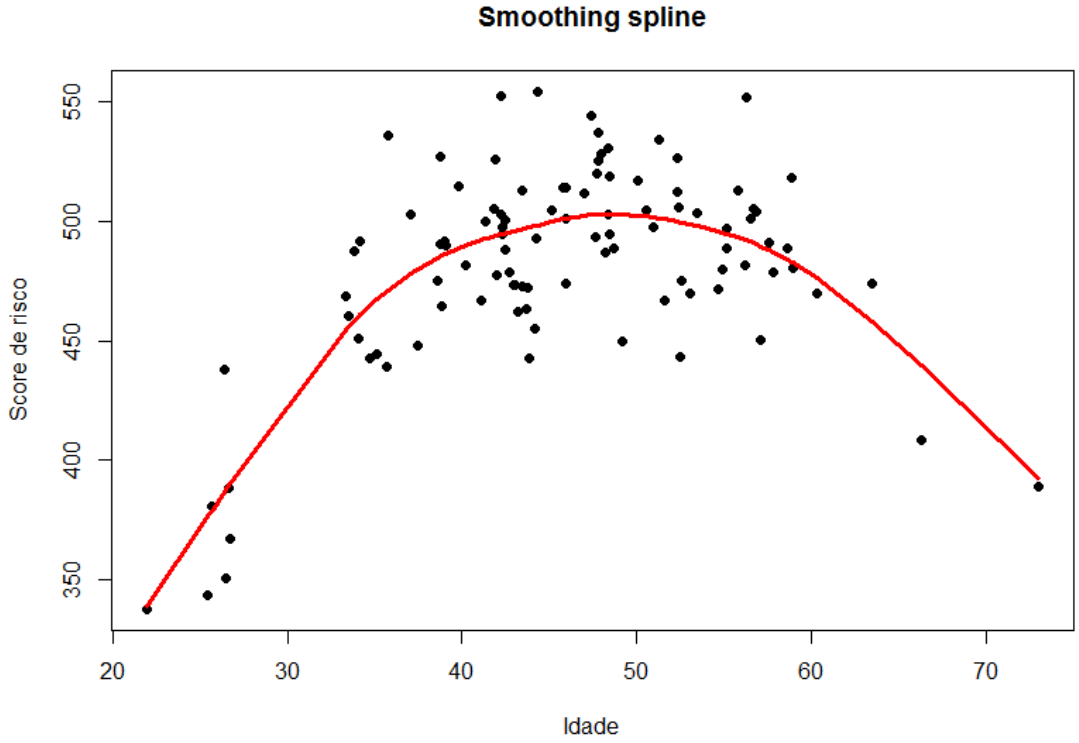
- Mas como escolher o grau do polinômio?
  - Geralmente se usa grau 2 ou 3.

- E os números de nós? Como escolhemos a quntidade e onde colocar?
  - Visualmente. É ideal colocar mais nós onde os pontos estão mais concentrados.
  - Testando valores diferentes e analisando o que produz melhores resultados.

Smoothing Spline

- E se considerarmos que cada ponto distinto de X é um nó?
- Teríamos praticamente uma interpolação dos dados.
- E se usarmos regularização?
- Essa é a ideia do smoothing spline. Ter todo ponto como um nó e usar regularização quadrática forçando vários dos coeficientes a ficarem bem próximos de 0.
- Temos que apenas otimizar a escolha do valor de lambda.

Smoothing Spline



Generalized Additive Model - GAM

- Consiste em usar splines de cada variável em um modelo linear. Assim, substituímos o modelo:

$$Y = \beta_0 + \beta_1 X_1 + \ \ldots \ + \beta_p X_p$$

Por

$$Y = \beta_0 + f_1(X_1) + \ \ldots + f_p(X_p)$$

Em que $f_1(X_1)$ representa uma smooth function da variável X_1, como por exemplo, um spline.

O nome aditivo vem do fato do modelo ser uma soma de smooth functions.

Generalized Additive Model - GAM

Pros:

- Ajusta efeitos não lineares sem a necessidade de especificar a estrutura.
- Em geral, previsões mais corretas
- Como o modelo é aditivos, ainda podemos interpretar cada variável mantendo todas as outras constantes.

Contras:

- Interpretação não é tão direta quanto nos modelos lineares
- Não capta efeito de interação entre as variáveis.

Laboratório

Laboratório

- Base de dados dos distritos de Boston.
- O objetivo é prever valor médio das casas.
- Variáveis:

  Crime per capita, proporção de terrenos com mais de 25.000 pés ao quadrado, proporção de area de comércio, dummy se faz divisa com rio, concetração de nitrogênio, número médio de quartos, proporção de casas ocupadas antes de 1940, distância média para os centros de emprego de Boston, acessibilidade a estradas, taxa de imposto por propriedade, razão aluno professor, proporção de negros, valor médio das casas.

Exercício

- Base de dados sobre a qualidade do ar em Nova York

- Prever a concentração de Ozônio no ar de Nova York

- Variáveis:
- Ozônio
- Temperatura
- Vento
- Intensidade do sol

Exercício

- Carregue a base de dados

```
library(datasets)
data(airquality)
df <- airquality
```

- Remova os valores ausentes por meio do comando (df <- na.omit(df))
- Crie as amostras de treino e teste (30%) usando seed de 42.
- Faça uma análise descritiva para entender a relação das variáveis.
- Ajuste um modelo GAM e compare com os resultados anteriores.
- Interprete este modelo.

Referências Bibliográficas

- Hastie, T., Tibshirani, R. & Friedman, J.H. (2001) “The Elements of Statistical Learning”
- Bishop, C.M. (2007) “Pattern Recognition and Machine Learning”
- Mitchell, T.M. (1997) “Machine Learning”
- Abu-Mostafa, Y., Magdon-Ismail, M., Lin, H.T (2012) “Learning from data”
- Theodoridis, S., Koutroumbas, K., (2008) “Pattern Recognition”
- Kuhn, M., Johnson, K., (2013) “Applied Predictive Modeling”
- Burns, P. (2011) “The R inferno”