

Universidade de São Paulo

Câmpus de Piracicaba

ESALQ - Escola Superior de Agricultura "Luiz de Queiroz"

Regressão Spline

Alex Rocha Soares

Responsável: Profa. Dra. Taciana Villela Savian

LCE - Laboratório de Ciências Exatas

ESALQ - USP

Novembro-2016

Sumário

1	Regresão Spline	2
1.1	Spline Linear	2
1.2	Spline Cúbico	4
1.3	B-Spline	5
1.4	Aplicação:	6
	Referências Bibliográficas	10

Capítulo 1

Regresão Spline

Um *spline* consiste em uma fita ou régua flexível usada para desenhar curvas que passam por pontos pré determinados, com o intuito de auxiliar na etapa de delineamento de objetos como cascos de navio, peças de avião, etc.

De acordo com o glossário Inglês-Português de Estatística e pela Sociedade Portuguesa de Estatística(SPE) determina *spline* como "função definida segmentadamente por polinômios(no sentido nato)".

Em outras palavras, *spline* é um conjunto de funções polinomiais, conectadas em determinados pontos de corte(onde definiremos mais adiante como nós) utilizados para ajustar uma curva a um conjunto de dados.

Técnicas diferentes são utilizados para representar esse ajuste por parte, como: *spline* linear, *b-spline*, *spline* cubico, entre outros. Neste trabalho, nosso intuito, é inserir conceitos básico da técnica de regressão *spline*.

1.1 Spline Linear

Também conhecido como função linear segmentada, consiste em uma técnica que baseia-se em um conjunto de funções lineares conectadas continuamente nos nós. Para cada intervalo intervalos entre os nós, existe um segmento de reta. Na conexão entre cada segmento de reta, os nós são suavizados e apresentam mudanças abruptas na direção dos segmentos de reta.

Considerando que o eixo x esta dividido em n parcelas, definidas pelos nós $\xi_1, \xi_2, \dots, \xi_{n-1}$, o modelo mais simples de *spline* linear, possui a seguinte forma:

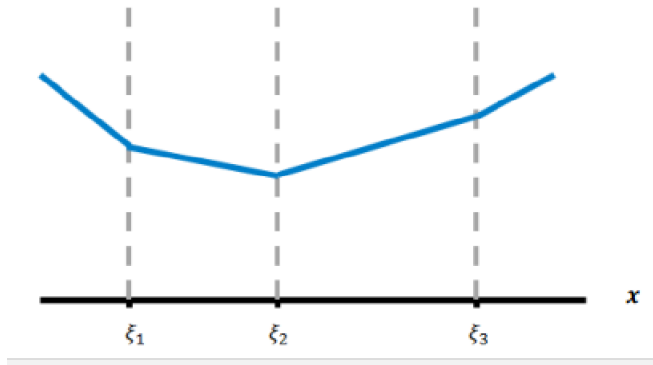
$$f(x) = \beta_0 + \beta_1(x) + \beta_2(x - \xi_1)_+ + \dots + \beta_{n+1}(x - \xi_n)_+$$

em que

$$(x - \xi_j)_+ = \begin{cases} (x - \xi_j), & (x - \xi_j) \geq 0 \\ 0, & (x - \xi_j) < 0 \end{cases}$$

Podemos notar que, conforme o valor no eixo x aumenta, novos termos são adicionados à função. Note que, se $x < \xi_1$, a equação que irá configurar esse espaço (entre o menos valor e ξ_1 será $f(x) = \beta_0 + \beta_1(x)$, agora, quando x estiver entre os dois nós seguintes, a equação será $f(x) = \beta_0 + \beta_1(x) + \beta_2(x - \xi_1)$ e assim sucessivamente.

A figura abaixo, representa o modelo de *spline* linear?



Para fazer um teste de hipótese com respeito a linearidade da variável resposta e o preditor, pode ser avaliado da seguinte maneira

$$H_0 = \beta_2 = \beta_3 = \dots = \beta_n$$

$$H_a = \beta_i \neq \beta_j, \text{ onde } i \neq j$$

Assim, se a hipótese for verdadeira, a contribuição dos termos *spline* correspondentes aos coeficientes de regressão $\beta_2 = \beta_3 = \dots = \beta_n$ é nula(ou muito pequena), restando apenas o intercepto β_0 e o coeficiente de regressão β_1 associado ao termo linear, ou seja, a linearidade em x está satisfeita.

O teste linearidade de preditores quantitativos, pode ser realizado por meio da estatística do teste da razão:

$$f(x) = \beta_0 + \beta_1(x)$$

e

$$f(x) = \beta_0 + \beta_1(x) + \beta_2(x - \xi_1) + \dots + \beta_{n+1}(x - \xi_n)$$

respectivamente denotados por

$$L_0(\beta_0, \beta_1; x) \text{ e } L_1 = (\beta_0, \beta_1, \dots, \beta_n; x).$$

Portanto, a estatística é da forma:

$$-2 \log \frac{L_0}{L_1} \sim \chi_{n-1}^2$$

onde o número de grau de liberdade é definido pela diferença entre o número de parâmetros β dos dois modelos.

1.2 Spline Cúbico

De forma análoga a como o *spline* linear é construído, nós construímos agora os *spline* cúbico, entretanto, a divergência se encontra no fato dos dados de cada intervalo serem ajustados por um polinômio de grau três.

O *spline* cúbico pode ser denominado como restrito ou não restrito. Ele será dito restrito se as caudas (partes do polinômio antes do primeiro nó e após o último nó) forem modeladas através de funções lineares, e caso contrário, será dita não restrita.

A função de *spline* cúbico restrito com k nós, $\xi_1, \xi_2, \dots, \xi_k$ pode ser escrita como:

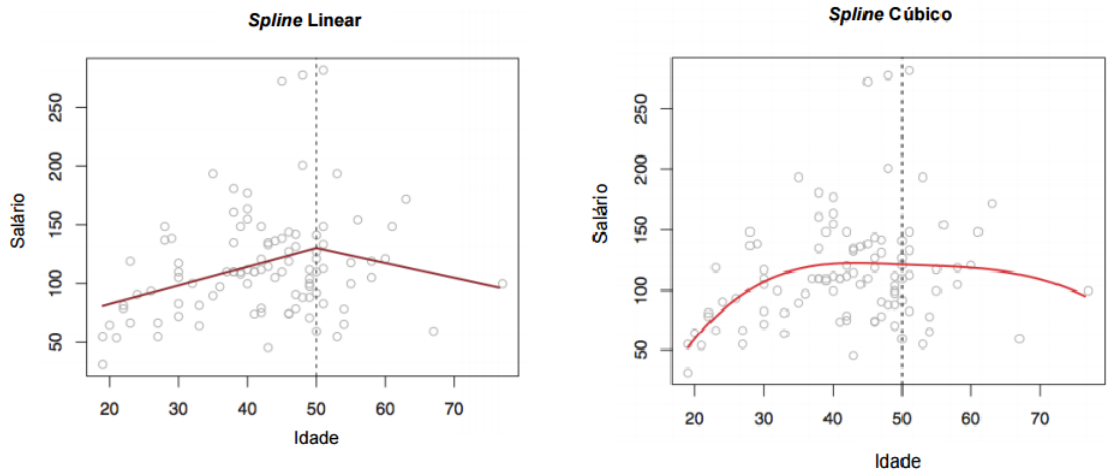
$$f(x) = \beta_0 + \beta_1 x + \beta_2(x - \xi_1)_+^3 + \beta_3(x - \xi_2)_+^3 + \dots + \beta_{k+1}(x - \xi_k)_+^3 \text{ onde}$$

$$(x - \xi_k)_+^3 = \begin{cases} (x - \xi_k)^3, & x - \xi_k \geq 0 \\ 0, & x - \xi_k < 0 \end{cases}$$

Todavia, alguns estudos (STONE et al, 1985, Additive regression and other nonparametric models) mostram que *splines* cúbicos não restritos não apresentam bom comportamento nas caudas, sendo estas, muito suscetíveis a mudanças. Portanto, por convenção, as funções que vem antes e depois do primeiro nó são forçadas a terem formato linear.

Note ainda que, por existir uma conexão em cada nó, a função *spline* acaba se tornando uma única curva contínua, entretanto, o que difere do *spline* linear, é a suavização em cada nó, o que converge em vantagens, como captar pequenas ondulações no comportamento dos dados.

A figura abaixo, representa um *spline* linear e um *spline* cúbico, todos ajustados aos mesmos dados, onde podemos perceber uma suavização no ajuste utilizando *spline* cúbico.



1.3 B-Spline

B-spline é uma transformação feita na variável x antes de se iniciar a etapa de ajuste do modelo. O $b-spline$ é um polinômio definido em partes de grau m definido para uma variável x . A principal propriedade de um $b-spline$ é a continuidade nos nós de forma não abrupta. o $spline$ cúbico é um caso particular de $b-spline$.

Outras propriedades são vistas em um $b-spline$ dado k número de nós e m sendo o grau do polinômio, que são:

- A função consiste em $k + 1$ pedaços do polinômio de grau m ;
- Os pedaços se unem nesses k nós, cujas derivadas nos pontos existem;
- o $b-spline$ é positivo no domínio de $k + 2$ nós e nulo nos restante;
- Exceto nas caudas, cada pedaço de polinômio se sobrepõe com $2xm$ pedaços de polinômios vizinhos;
- Os nós podem ser equidistantes ou não.

Podemos representar a curva ajustada para os dados como

$$f(x) = \sum_{j=1}^{k+m+1} a_j B_j(x)$$

em que $B_j(x)$ denota o valor do j -ésimo *b-spline* de grau m no ponto x para uma malha de k nós equidistantes

Vale observar que: para ao caso de nós não equidistantes, a equação sofre pequenas modificações.

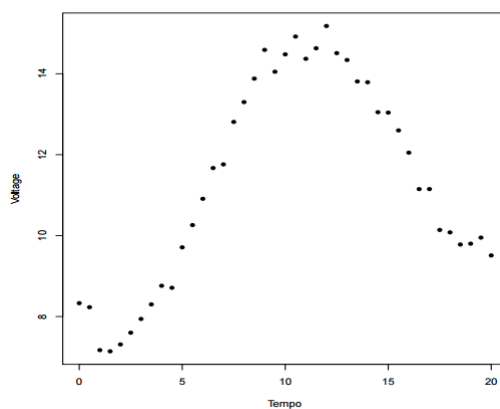
1.4 Aplicação:

Dado um conjunto de dados de um experimento em que representa a queda de tensão da bateria (em voltagem) de um motor de míssil guiado é observada ao longo do tempo(em segundos). A queda da tensão da bateria foi observada em 41 instantes da trajetória do míssil. O objetivo do estudo, é utilizar o modelo ajustado de regressão por partes para simular a trajetória do míssil.

Descrição dos dos dados:

Medida	Queda de Tensão
n	41
Média	11,34
D.Padrão	2,56
CV	22,6%
Mínimo	7,14
1º Quartil	9,51
Mediana	11,15
3º Quartil	13,81
Máximo	15,18

A figura abaixo representa a comportamento de queda do míssil:



Nota-se indícios de mudanças de comportamento da queda da tensão entre 6 e 10 segundos e 10 e 15 segundos. Uma proposta é ajustar uma regressão por partes dividindo o intervalo de tempo em três partes(até 6,5 segundos, entre 6,5 e 13 segundos e entre 13 e 20 segundos), ficando o seguinte modelo de regressão linear:

$$f(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 (x_i - 6,5)_+^3 + \beta_5 (x_i - 13)_+^3 + \varepsilon_i$$

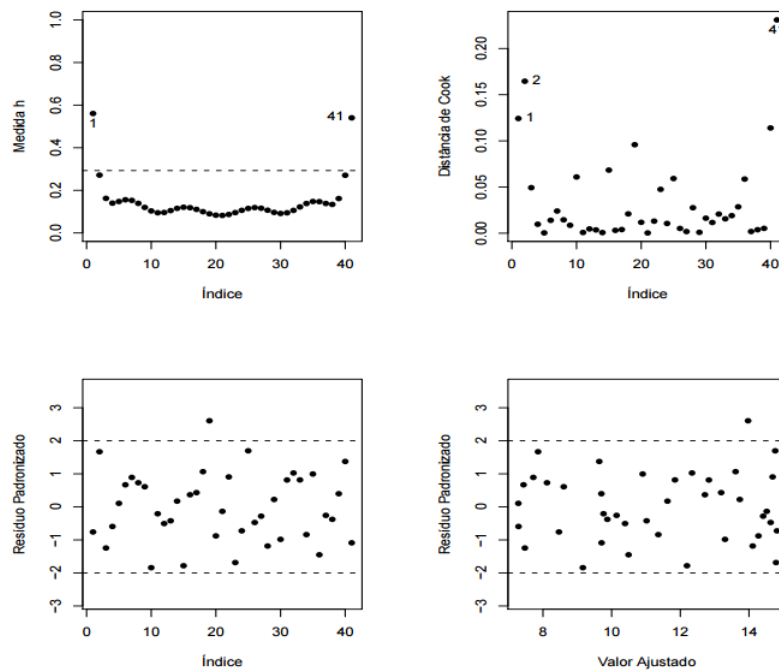
para $i = 1, \dots, 41$ em que $f(x_i)$ e x_i denotam, respectivamente, a queda de tensão e o tempo decorrido até o i -ésimo instante, com $\varepsilon \sim^{iid} N(0, \sigma^2)$

As estimativas do modelo estão representadas abaixo:

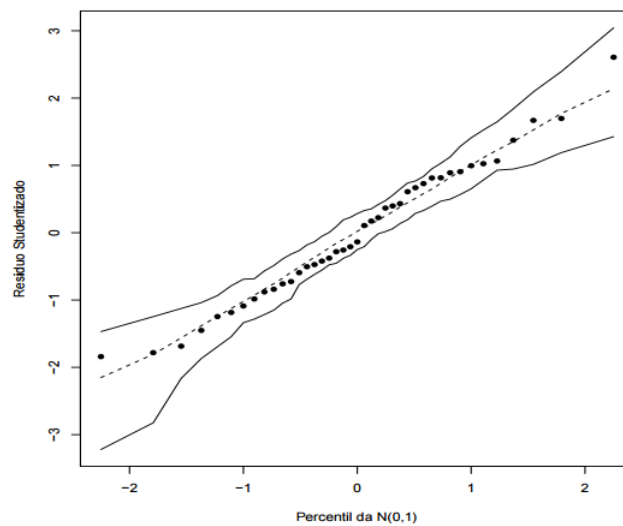
Efeito	Estimativa	E. Padrão	valor-t	valor-P
β_0	8,465678	0,200520	42,119	0,00
β_1	-1,453124	0,181586	-8,002	0,00
β_2	0,489889	0,043018	11,388	0,00
β_3	-0,029467	0,002848	-10,347	0,00
β_4	0,024706	0,004039	6,116	0,00
β_5	0,027112	0,003578	7,577	0,00
R^2	0,990			
R^2 -ajustado	0,989			
s	0,268			
F	725,5	(5 e 35 g.l.)		0,00

Os coeficientes são marginalmente significativos.

Agora, faremos uma análise diagnóstica, na qual esta representada abaixo:



Então, façamos a análise de resíduos



Nota-se pelos gráficos de diagnósticos 3 pontos de alavanca e 2 pontos potencialmente influentes. A eliminação das 3 observações altera numericamente algumas estimativas, todavia não muda a inferência. Todos os coeficientes continuam altamente significativos.

Então, obtemos o modelo ajustado:

- Parte 1

Para $0 \leq x \leq 6,5$ temos o seguinte modelo ajustado

$$\hat{\mu} = 8,466 - 1,453x + 0,490x^2 - 0,029x^3$$

- Parte 2

Para $6,5 < x \leq 13$ temos o seguinte modelo ajustado

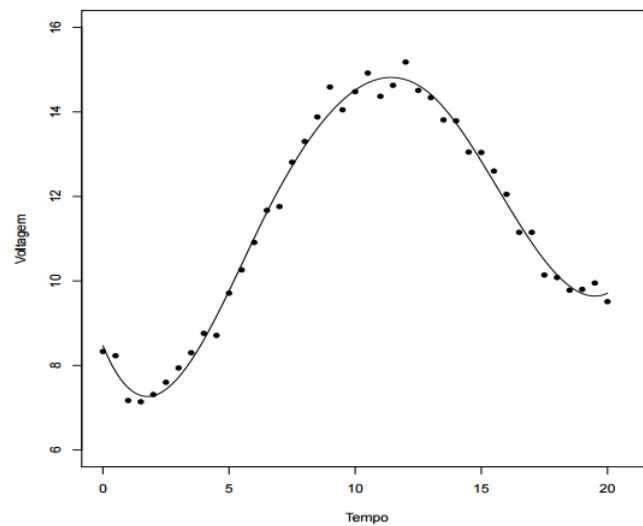
$$\hat{\mu} = 8,466 - 1,453x + 0,490x^2 - 0,029x^3 + 0,025(x - 6,5)^3$$

- Parte 3

Para $13 < x \leq 20$ temos o seguinte modelo ajustado

$$\hat{\mu} = 8,466 - 1,453x + 0,490x^2 - 0,029x^3 + 0,025(x - 6,5)^3 + 0,027(x - 13)^3.$$

Portanto, a curva ajustada é da seguinte forma:



CONCLUSÃO:

Este exemplo mostra a utilidade de regressão por partes através de um *spline* cúbico com dois pontos de mudança.

Referências Bibliográficas

- [1] < https://w.ime.usp.br/~giapaula/slides_exemplo_aterias.pdf > - acessado no dia 27/11/2016.
- [2] < <https://www.lume.ufrgs.br/bitstream/handle/10183/129437/000976776.pdf?sequence=1> > - acessado no dia 27/11/2016.
- [3] Montgomery, D. C.; Peck, E. A. e Vining, G. G. (2001). Introduction to Linear Regression Analysis, Third Edition. Hoboken: Wiley.