METÓDOS DE REGRESSÃO KERNEL

George Cavalcanti de Albuquerque Júnior [1]; Getúlio José Amorim do Amaral [2]

[1]Estudante do Curso de Estatística - CCEN – UFPE; E-mail: gcdaj1@de.ufpe.br,
[2] Docente/pesquisador do Depto de Estatística – CCEN – UFPE. E-mail: gjaa@de.ufpe.br.

Sumário: Na literatura estatística, os modelos de regressão baseados na distribuição normal são os mais frequentemente utilizados. Porém, essa classe de modelos têm suposições que muitas vezes não são satisfeitas por um conjunto de dados reais. Dessa forma, este projeto estudou o modelo de regressão Kernel que é não-paramétrico e representa melhor os padrões observados em dados reais. Foi feito comparações gráficas e análises de resíduos para poder classificar o "melhor" modelo relacionado em cada caso estudado. O objetivo desse projeto é mostrar e avaliar as vantagens de utilizar um modelo de regressão não paramétrico em dados reais, no nosso caso o modelo de regressão Kernel.

Palavras–chave: método Kernel; modelo normal; regressão

INTRODUÇÃO

Um problema muito encontrado na estatística é a construção de modelos que representem a relação entre duas variáveis, onde a variável de interesse y é chamada de variável resposta e a variável relacionada a ela x é denominada de variável explicativa. Uma das alternativas é utilizar o método Kernel que tenta modelar a relação x e y, mesmo que a distribuição de y seja desconhecida. Nem sempre em todos os casos teremos relações de variáveis explicativas com variáveis respostas de forma linear, então nem sempre será uma boa alternativa utilizar o método da regressão Linear.

MATERIAIS E MÉTODOS

Implementamos o algoritmo da Regressão Kernel na linguagem R para avaliar os resultados da modelagem da regressão Normal e da regressão Linear e assim classificar o "melhor" modelo em cada caso.

O algoritmo utilizado para estimação da variável resposta y do método regressão Kernel é apresentado a seguir.

Inicialmente a variável resposta é calculada por:

$$\hat{y}i = \sum_{j=1}^{n} W_{ij}\, y_j \tag{1}$$

Onde $\hat{y}$ será todos os valores estimados em relação a X e Y, W_{ij} será a matriz de suavizamento dos dados e y_i será os valores reais da variável resposta.

A matriz de suavizamento é calculada da seguinte forma:

$$W_{ij} = \frac{K\left(\frac{Xi-Xj}{h}\right)}{\sum_{k=1}^{n} K\left(\frac{Xi-Xk}{h}\right)} \tag{2}$$

Onde x são os dados independentes e K a função Kernel, definida pela função Gaussiana é dada por:

$$f(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)} \tag{3}$$

A variável b, em (2), será o que chamamos de janela de parametrização, que serve para aumentar ou diminuir os detalhes de suavizamento da curva de regressão.

Utilizando essas equações teremos um modelo que irá representar uma relação de duas variáveis x e y.

RESULTADOS

Tabela 1 – Velocidade e Saída do Moinho de vento.

Observações	**Y – Saída**	**X – Velocidade do Moinho**
1	1.582	5.00
2	1.822	6.00
3	1.057	3.40
4	0.500	2.70
5	2.236	10.00
6	2.386	9.70
7	2.294	9.55
8	0.558	3.05
9	2.166	8.15
10	1.866	6.20
11	0.653	2.90
12	1.930	6.35
13	1.562	4.60
14	1.737	5.80
15	2.088	7.40
16	1.137	3.60
17	2.179	7.85
18	2.112	8.80
19	1.800	7.00
20	1.501	5.45
21	2.303	9.10
22	2.310	10.20
23	1.194	4.10
24	1.144	3.95
25	0.123	2.45

A Tabela 1 apresenta os dados de duas amostras X e Y que representam a velocidade de um moinho de vento e sua saída respectivamente. (Veja Montgomery & Peck (2012)).
A Figura 1 apresenta o método da regressão normal em cima dos dados da Tabela 2.

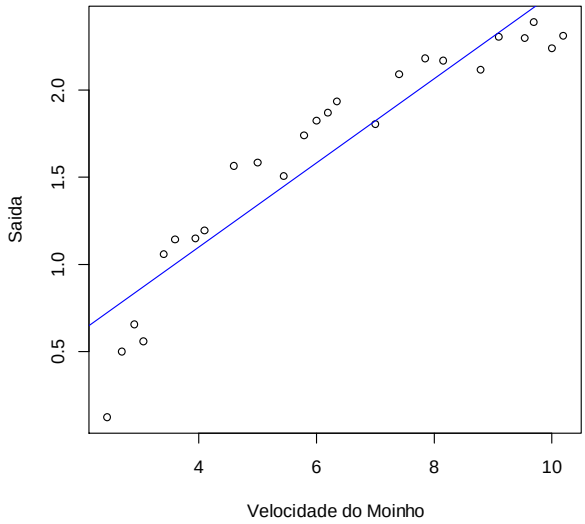

Figura 1 – Regressão Normal para relação do Moinho de Vento.

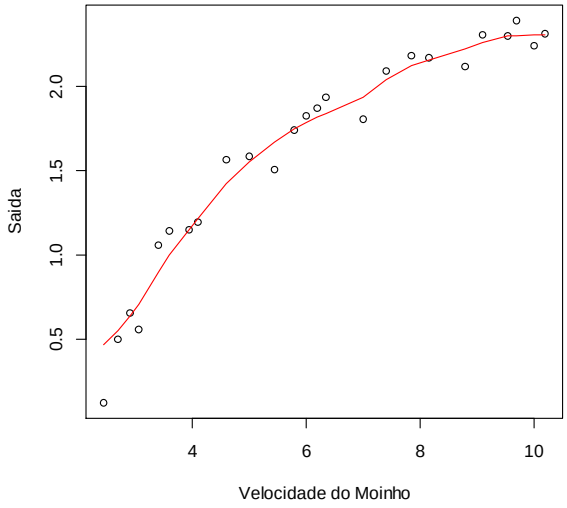

Figura 2 – Regressão Normal para relação do Moinho de Vento.

A figura 1 apresenta a relação x e y da tabela 1, e a reta em azul o método Linear para representar essa relação. Em relação à figura 2 que apresenta o modelo de regressão Kernel, a curva da figura 2 acompanha melhor a distribuição da relação de x e y, mostrando visualmente um modelo mais eficaz que o modelo Linear.
Observando as Figuras 3 e 4 que representam os resíduos da regressão Normal e regressão Kernel, respectivamente, vamos poder observar qual melhor método nesse caso.

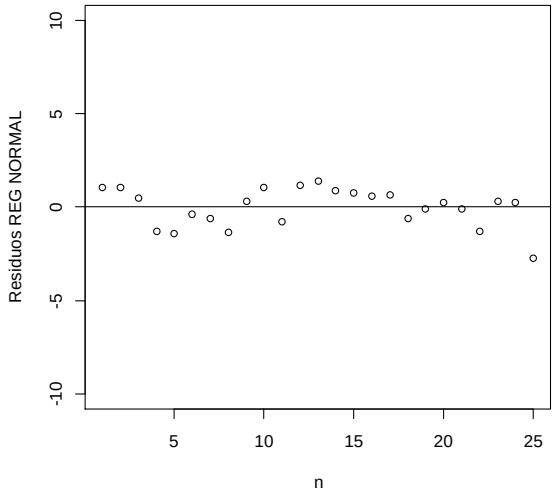

Figura 3 - Resíduos regressão Normal para Tabela 1.
A figura 3 apresenta os resíduos, que seria o erro da variável resposta y para a regressão Normal. Os resíduos não apresentam nenhuma das suposições que impedem a utilização deste método.

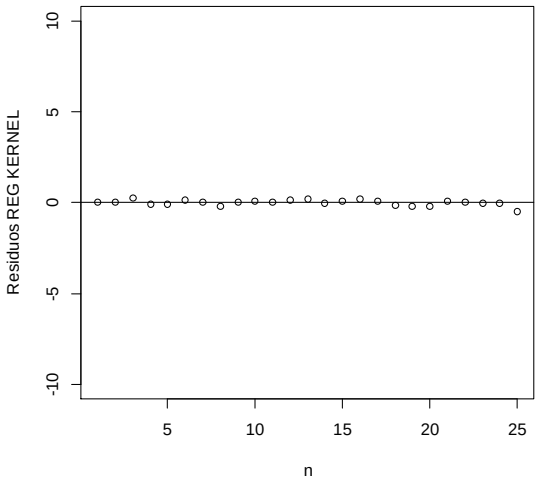

Figura 4 - Resíduos regressão Kernel para Tabela 1.
A figura 4 apresenta os resíduos, que seria o erro da variável resposta y para a regressão Kernel sem nenhuma suposição.

DISCUSSÃO

Podemos concluir que o método Kernel, visualmente pelos gráficos dos resíduos dos dois modelos de regressão, é mais eficiente, pois os valores dos resíduos estão mais próximos do valor zero do que os valores dos resíduos do método da regressão Normal.
A aplicação do método Kernel foi feito e aplicado à estatística de análise Procrustes, mas foi inviável o uso pela ineficiência em gerar uma boa regressão a partir de duas imagens de mesma origem.

CONCLUSÕES

O trabalho com a regressão Kernel tem vantagem em relação a regressão Normal por conseguir suavizar uma curva que representa melhor uma distribuição de dados do que uma reta. Estamos acostumados a utilizar sempre métodos paramétricos para modelagem de dados, descartando muitas vezes a vantagem de utilizar a estatística não paramétrica. Uma possibilidade de trabalho futuro é o estudo sobre as constantes de suavizamento que são utilizados no método Kernel.

AGRADECIMENTOS

Agradeço ao meu orientador pela oportunidade de iniciar na carreira científica, agradecer a paciência e experiência transmitida através desses meses. Agradeço a bolsa recebida que incentiva e ajuda a dar continuidade ao trabalho. Agradeço a meus pais e amigos que colaboraram a todo o momento.

REFERÊNCIAS

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Douglas, C. Montgomery & Elizabeth, A. Peck & G. Geoffrey Vining. 2012. Introduction to Linear Regression Analysis, 5th Edition. Editora Wiley.