Regressão Local *(LOESS)*

Wagner Hugo Bonat*

14 de novembro de 2007

1 Regressão Local

Regressão Local *(Loess)* é um método não paramétrico que estima curvas e superfícies através de suavização *(smoothing)*. Este método ganhou popularidade a partir da década de 70 com o desenvolvimento de computadores e a publicação dos estudos independentes de [8], [9] e [7]. Sendo que [9] desenvolveu o software *Lowess*, que foi implementado em diversos pacotes estatísticos. As idéias básicas do método podem ser observadas ao considerar-se o mais simples dos modelos de regressão, onde a variável dependente, y, e a independente, x, são relacionadas por:

$$y_i = g(x_i) + \epsilon_i$$

onde ϵ_i denota o termo de erro independente e identicamente distribuído com distribuição normal, média zero e variância constante.

Ao contrário dos métodos paramétricos que estimam a função globalmente, regressão local estima a função "g" na vizinhança de cada ponto de interesse $x = x_0$. Uma forma simples de estimar uma função localmente é considerar a média ponderada das observações que estão na vizinhança do ponto de interesse, x_0. Duas escolhas devem ser feitas para realizar esta estimativa. Primeiro, deve ser escolhido o tamanho da vizinhança, h, do ponto $x = x_0$ e , segundo, deve ser escolhida uma função K que pondera o conjunto de pontos vizinhos a x_0. A função K é denominada de núcleo (*Kernel*), enquanto que h é denominada de banda ou parâmetro de suavização. Com este procedimento, a equação para a média local ponderada por K é dada por:

$$\widehat{g}(x_0) = \frac{\sum_{i=1}^{n} K_h(x_i - x_0) y_i}{\sum_{i=1}^{n} K_h(x_i - x_0)}$$

*Graduando em estatística - UFPR

Este estimador de núcleo foi proposto inicialmente por [5] e [6]. Existem sérias limitações com a estimativa de uma constante localmente, como por exemplo, viés nas regiões de fronteira e no interior se a variável independente não for uniforme e se a função de regressão tiver curvatura. Uma maneira de resolver este problema é através de regressão local linear ponderada, proposta inicialmente por [8] e [9]. Ao estimar uma linha reta localmente ao invés de uma constante, o problema de viés de primeira ordem é eliminado, desta forma, regressão local linear resolve um problema de mínimos quadrados ponderados a cada ponto de interesse, x_0, conforme:

$$\min_{\alpha\beta} \sum_{i=1}^{n} K_h(x_i - x_0)[y_i - \alpha - \beta(x_i - x_0)]^2 \tag{1}$$

Regressão local linear será igual ao estimador de Nadaraya-Watson expresso pela equação (1) se o termo $\beta(x_i - x_0)$ for removido. Neste caso, uma constante será estimada localmente. Apesar de regressão local linear ser utilizado como técnica padrão por muitos autores [1], não há razões para não utilizar polinômios de ordem mais alta, mesmo porque a regressão local pode apresentar viés quando a função a ser estimada possui uma forte curvatura. Nestes casos uma polinomial de grau d pode ser estimada através da seguinte função:

$$\min_{\alpha\beta_j, j=1,\ldots,d} \sum_{i=1}^{n} K_h(x_i - x_0)[y_i - \alpha - \sum_{i=1}^{d} \beta(x_j - x_0)^j]^2$$

Portanto para modelar-se determinado processo por regressão local, deve-se de forma geral fazer três escolhas: a função núcleo, o parâmetro de suavização e o grau da polinomial. Existe ainda uma outra escolha que deve ser feita, que diz respeito a distribuição assumida para os termos de erro, no presente trabalho assume-se que os erros seguem uma distribuição gaussiana. Para uma discussão de regressão local com a consideração de outras distribuições do erro ver [4].

1.1 Parâmetro de suavização (banda).

O parâmetro de suavização (span ou bandwidth), h, controla o tamanho da vizinhança no entorno de x_0 no qual a função núcleo será aplicada. O parâmetro de suavização possui papel determinante na variabilidade e no viés da estimativa. Se o h escolhido for pequeno, a estimativa terá um viés reduzido, mas uma variabilidade elevada. Por outro lado, se o h escolhido for grande, a estimativa terá um viés elevado mas pequena variabilidade.O objetivo é produzir uma estimativa que seja a mais suave possível sem distorcer a

relação de dependência entre as variáveis em análise [2]. [3] discute e compara diferentes procedimentos para a escolha da banda, os quais são classificados em dois grupos. O primeiro, constituído pelos métodos clássicos, são baseados em extensões dos procedimentos já utilizados em regressão paramétrica, tais como validação cruzada (cross validation), critério de informação de Akaike e Cp de Mallows, que consistem basicamente em empregar alguma medida de aderência, como por exemplo, minimizar a média da integral do erro ao quadrado ou uma simplificação desta. Os métodos do segundo grupo são baseados em anexo (plug-ins). Estes consistem em escrever a função inicialmente estimada, $\widehat{g}$, como uma função g desconhecida e aproximada por uma expansão de Taylor ou outra expansão assintótica. Uma estimativa de g é então "anexada"(plugged-in) para derivar uma estimativa da tendenciosidade e uma estimativa da aderência, tal como, o erro quadrado médio integrado *(mean integrated squared error)*. Segundo [4] os métodos clássicos apresentam melhores resultados em termos práticos, bem como se ajustam a uma grande variedade de casos.

1.2 Grau do polinômio local.

Esta escolha também afeta a relação entre variância e viés, quanto maior o grau da polinomial menor será o viés e maior a variância para um mesmo parâmetro de suavização. De modo geral , o aumento da variância que decorre da utilização de polinomiais de ordem mais elevada pode ser compensado empregando-se um parâmetro de suavização maior. A utilização de polinomiais de baixa ordem é suficiente para produzir estimativas de ótima qualidade, normalmente são utilizados polinomiais com graus variando de zero a três. A escolha do grau da polinomial é, em sua maior parte, determinada pelos objetivos do trabalho e pelos dados, na prática a escolha do grau da polinomial pode ser realizada pela inspeção visual do gráfico com os dados originais e a estimativa de regressão local. De forma geral, a presença de "picos"ou "vales"nos dados são um indicativo de que d deve ser igual a dois ou três, enquanto que a presença de um padrão único indicam que d deve ser igual a um.

1.3 A função de *kernel*.

Esta função é responsável por ponderar as observações na vizinhança de cada ponto de interesse, x_0. Segundo [2] e [4] esta função deve ser contínua, simétrica, com maior peso em torno de x_0 e decrescente a medida em que x se afasta de x_0. Dentre as escolhas possíveis, destaca-se aqui a função tri-cúbica que será utilizada no trabalho.

Para analisar esta função vamos considerar uma variável transformada u_i definida por:

$$u_i = \frac{(x_i - x_0)}{h_i}$$

Então a função de peso K é obtida em função da variável u, isto é,

$$K(u) = K[\frac{(x_i - x_0)}{h_i}]$$

Um escolha comum é a função tri-cúbica que é obtida por:

$$\begin{aligned} k(u) &= (1 - |u|^3)^3 = \left(1 - \left|\frac{x - x_i}{h_i}\right|^3\right)^3 \quad \text{se} \quad |u| < 1 \\ &= 0 \quad \text{se} \quad |u| \geq 1 \end{aligned} \tag{2}$$

Muitas outras funções de *Kernel* ou pesos, podem ser utilizadas, ela vai depender dos objetivos do trabalho.

Referências

[1] BOWMAN ADRIAN ; AZZALINI ADELCHI. Applied smoothing techniques for data analysis: The kernel approach with s-plus illustrations. *Oxford: Oxford University Press*, 1997.

[2] CLEVELAND W. S. ; LOADER C. Smoothing by local regression: Principles and methods. *Physica-Verlag, p. 10-49*, 1996a.

[3] LOADER CLIVE. Old faithful erupts: Bandwidth selection reviewed. *Working paper, ATT Bell Laboratory*, 1995.

[4] LOADER CLIVE. Local regression and likelihood. *Springer-Verlag*, 1999.

[5] NADARAYA E.A. On estimating regression. *Theory of Probability and ilts Applications, v.9, p. 141-142*, 1964.

[6] WATSON G.S. Smooth regression analysis. *Sankhya, Series. A, v.26, p. 359-372*, 1964.

[7] STONE C. J. Optimal rates of convergence for nonparametric estimators. *The annals of Statistics, v.8 p.1348-1360*, 1980.

[8] C. J. STONE. Consistent nonparametric regression, with discussion. *The annals of Statistics, 5 p.549-645*, 1977.

[9] CLEVELAND WILLIAM. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, v.74, p. 829-836*, 1979.