ISSN 0102-9924

FACULDADE DE CIÊNCIAS ECONÔMICAS DA UFRGS

Análise Econômica

ANO 22

Nº 42

SETEMBRO, 2004

PROF. FERNANDO FERRARI FILHO
Revista *Análise Econômica* - Av. João Pessoa, 52
CEP 90040-000 PORTO ALEGRE - RS, BRASIL
Telefones: (051) 316-3513 – Fax: (051) 316-3990
E-mail: rae@ufrgs.br

Análise Econômica
Ano 22, nº 42, março, 2004 - Porto Alegre
Faculdade de Ciências Econômicas, UFRGS, 2004
Periodicidade semestral, março e setembro.
Tiragem: 500 exemplares
1. Teoria Econômica - Desenvolvimento Regional - Economia Agrícola - Pesquisa Teórica e Aplicada - Periódicos. I. Brasil.
Faculdade de Ciências Econômicas,
Universidade Federal do Rio Grande do Sul.

CDD 330.05
CDU 33 (81) (05)

Princípios e aplicações de regressão local

Adalmir Marquetti [1] *e Lorí Viali* [2]

Resumo: Este texto discute os princípios e realiza três aplicações de regressão local, um método não paramétrico que estima cuevas e superfícies através de suavização (*smoothing*). Na análise não paramétrica a função é estimada sem referência a uma forma funcional previamente estabelecida, permitindo que os "dados falem por si próprios". Regressão local pode ser utilizada para estudar modelos com uma ou mais variáveis independentes, bem como para estimar derivadas. Além disso, suas propriedades estatísticas têm sido estudadas, permitindo que se realizem inferências sobre os resultados. Regressão local tem sido implementada em uma série de softwares, possibilitando o fácil acesso dos pesquisadores a esta técnica.

Abstract: This paper presents the basic principles of local regression, a nonparametric procedure to estimate curves and surfaces through smoothing. In the nonparametric analysis, the relationship between the dependent and independents variables is estimated without reference to a priori functional form, allowing that the data speak for themselves. Local regression might be employed to study models with one or more independent variables as well as to estimate derivatives. Moreover, its statistical properties have been studied, permitting to realize inference over the results. Local regression has been implemented in a series of softwares, allowing that the researchers have access to this technique.

Palavras-chave: Análise não-paramétrica, regressão local, modelagem local, ajustamento de curvas e superfícies, *Loess e Lowess*.

Keywords: Nonparametric analysis, local regression, local modeling, smoothing, Loess, Lowess.

Jell classification: C14.

1 Introdução

O objetivo deste texto é apresentar os princípios básicos de regressão local. Regressão local é um método não-paramétrico que estima curvas e superfícies através de suavização (*smoothing*). Este

[1] Departamento de Economia, Pontifícia Universidade Católica do Rio Grande do Sul, PUCRS, Av. Ipiranga 6681, Porto Alegre, RS, Brasil, 90619-900, e-mail: aam@pucrs.br.
[2] Departamento de Matemática, Pontifícia Universidade Católica do Rio Grande do Sul, PUCRS, Av. Ipiranga 6681, Porto Alegre, RS, Brasil, 90619-900, e-mail: viali@pucrs.br.

método começou a ganhar popularidade a partir do final dos anos 1970 com o desenvolvimento dos computadores e a publicação dos estudos independentes de Stone (1977), Cleveland (1979) e Stone (1980). Em boa medida, esta popularidade deveu-se ao trabalho de Cleveland (1979) que desenvolveu o software *Lowess*, que foi implementado em diversos pacotes estatísticos e tornou-se um dos principais recursos para a análise de regressão não-paramétrica.

A análise não-paramétrica, em contraste com o método paramétrico, estima uma função média sem referência a uma forma funcional previamente estabelecida, permitindo que os "dados falem por si próprios". Tal característica é de grande interesse, pois, muitas vezes, a análise teórica não estabelece a forma estrutural entre as variáveis ou estabelece formas estruturais competitivas. Contudo, no momento de realizar o estudo empírico, a análise paramétrica assume a forma linear ou uma transformação que lineariza o modelo e o torna aditivo. Uma das grandes vantagens da análise não-paramétrica está na sua flexibilidade, permitindo facilmente a percepção de relações não lineares.

Regressão local tornou-se um dos principais métodos empregados em suavização por duas razões. Primeiro, sua implementação em diversos softwares comerciais e livres (S-PLUS, SAS, GAUSS, XploRe, R, entre outros). Segundo, suas vantagens do ponto de vista dos resultados estatísticos em relação aos métodos competitivos (Fan, 1992). Dias (2002) apresenta métodos alternativos de regressão não-paramétrica.

Deve-se ressalatar que há diversas técnicas não paramétricas competivas à regressão local. para se analisar ou predizer o comportamento de uma dependente em função de uma ou mais variáveis independentes. O campo da modelagem não paramétrica é vasto e, conforme Fan (2000) chama a atenção, têm tido um desenvolvimento notável nas últimas três décadas. Os métodos não paramétricas podem ser divididos em técnicas de suavização ou alisamento e em técnicas de redução de dimensionalidade.

As principais técnicas de suavização ou alisamento incluem médias móveis, *splines*, séries ortogonais e *wavelets*, estimadores de núcleo (*kernel*) e regressão local. Fan e Gijbels (1996, cap. 2) realizam uma comparação entre estas técnicas. Quando a regressão envolve duas ou mais dimensões, a suavização torna-se menos viável devido ao problema da dimensionalidade. Além da terceira dimensão não é possível visualizar uma superfície e, neste

caso, métodos que tentam capturar estruturas dimensionais mais baixas podem ser apropriados.

Muitas técnicas têm sido propostas para o ajuste de dados multivariados. Entre estas incluem-se a procura projetiva (*projection pursuit*), modelos aditivos, modelos de coeficientes variáveis, modelos de interação de baixa dimensão e modelos de índices múltiplos. Muitas destas técnicas permitem o desenvolvimento de modelos semi-paramétricos, os quais procuram aliar as vantagens da regressão linear clássica com o método não-paramétrico.

O presente artigo está organizado como segue. A seção 2 apresenta a concepção básica de regressão local. A seção 3 discute as escolhas que o pesquisador realiza ao utilizar regressão local. A seção 4 aborda a extensão de regressão local para modelos multivariados. A seção 6 analisa a estimativa de derivadas. A seção 6 trata das propriedades estatístisca de regressão local. A seção 7 discute três exemplos de aplicações de regressão local. Por fim, a seção 8 conclui o trabalho.

2 Concepção básica

Regressão local é um método não-paramétrico que utiliza suavização (*smoothing*) para ajustar curvas e superfícies. 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:

$$y_i = g(x_i) + \varepsilon_i \qquad (1)$$

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 relativamente 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:

$$\hat{g}(x_0) = \frac{\sum_{i=1}^{n} K_h(x_i - x_0) y_i}{\sum_{i=1}^{n} K_h(x_i - x_0)} \qquad (2)$$

Este estimador de núcleo foi proposto inicialmente por Nadaraya (1964) e Watson (1964). 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. Para uma comparação entre os diferentes estimadores de núcleo e uma análise das limitações do estimador de Nadaraya-Watson, consultar Hastie e Loader (1993) e Fan e Gijbels (1996, cap. 2).

Uma maneira de resolver este problema é através de regressão local linear ponderada, proposta inicialmente por Stone (1977) e Cleveland (1979). Ao estimar uma linha reta localmente ao invés de uma constante, o problema de viés de primeira ordem é eliminado. Regressão local linear resolve um problema de mínimos quadrados ponderados a cada ponto de interesse, x_0, conforme:

$$\min_{\alpha,\beta} \sum_{1}^{n} K_h(x_i - x_0)[y_i - \alpha - \beta(x_i - x_0)]^2 \qquad (3)$$

Regressão local linear será igual ao estimador de Nadaraya-Watson se o termo $\beta(x_i - x_0)$ for removido. Neste caso, uma constante será estimada localmente. Apesar de alguns autores utilizarem regressão local linear como técnica padrão (BOWMAN e AZZALINI, 1997), não há razões para não utilizar polinômios de ordem mais alta. Regressão local linear pode exibir viés quando a função a ser estimada possui forte curvatura. É possível estimar uma polinomial de grau d através da seguinte expressão:

$$\min_{\alpha,\beta j, j=1,\ldots,d} \sum_{1}^{n} K_h(x_i - x_0)[y_i - \alpha - \sum_{1}^{d} \beta(x_j - x_0)^j]^2 \qquad (4)$$

O grau da polinomial a ser estimada é escolhido pelo pesquisador. Portanto, ao modelar seus dados, o pesquisador deve fazer três escolhas: a função núcleo, o parâmetro de suavização e o grau da polinomial. Há, ainda, uma outra escolha importante a ser realizada, a da distribuição que os erros seguem. No presente texto assume-se que os erros seguem uma distribuição gaussiana.

Contudo, é possível assumir diversas famílias de distribuições. Para uma discussão de regressão local com a consideração de outras distribuições do erro, ver Loader (1999, cap. 4), Cleveland e Loader (1996) e Fan e Gijbels (1996, cap. 5).

3 Fazendo escolhas

Como dito anteriormente, o pesquisador deve fazer algumas escolhas quando emprega regressão local. Estas envolvem a amplitude da vizinhança, isto é, a banda, o grau do polinômio local, a função peso e as hipóteses sobre a distribuição resposta.

Antes de discutir cada uma destas escolhas, é importante chamar a atenção para a diferença sobre a interpretação dos resultados entre regressão linear paramétrica e regressão local. Na regressão linear paramétrica são estimados os coeficientes de uma forma funcional determinada previamente, e o pesquisador verifica quão bem os resultados se aproximam dos coeficientes reais (populacionais) por meio de um teste de hipóteses (análise de variância). Não há maior preocupação com a curva estimada. Na regressão local ocorre uma mudança de perspectiva, como a forma funcional não é previamente estabelecida, a curva estimada passa a ocupar o papel central na análise. De maneira semelhante, um teste de hipóteses pode ser empregado para verificar se a curva estimada reproduz a verdadeira função média. Portanto, um aspecto central da regressão local é a visualização. Cleveland (1993), em um trabalho já clássico, discute como o processo de visualização e o de regressão local podem ser associados na análise dos dados.

3.1 O parâmetro de suavização (banda)

A escolha do 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 um efeito fundamental na regressão local, pois 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 variabili-

dade. Quando h se aproxima de zero, a estimativa tende a interpolar as observações. Quando h aumenta, a curva estimada aproxima-se de uma regressão linear de grau d, o grau da função polinomial utilizada.

O ideal seria escolher uma banda diferente para cada ponto, isto é, tomar h_i para cada x_i, levando em consideração a densidade dos pontos no local. Entretanto, fazer isto na prática é muito complexo. Usualmente, a função banda é especificada com poucos ou então sem parâmetros.

A primeira decisão é a escolha entre um valor de h global, isto é, que sirva para todos os pontos de interesse, x_i. A escolha mais simples é considerar o parâmetro de suavização constante. Este procedimento é satisfatório nos casos em que a variável independente possui uma distribuição uniforme. O principal problema desta abordagem é o caso das vizinhanças vazias, que ocorrerem principalmente nas caudas das distribuições ou quando a estimativa envolve mais do que uma dimensão.

Uma segunda abordagem é a opção por uma banda local, $h(x)$, que pode ser escolhida de forma a sempre conter um número especificado de pontos e que resolve o problema das vizinhanças vazias. Esta abordagem é conhecida como a do vizinho mais próximo (*nearest neighbor bandwidth*).

A Figura 1 mostra estimativas de regressão local para diferentes parâmetros de suavização utilizando a abordagem do vizinho mais próximo para a relação entre as variáveis "taxa de fertilidade por mulher" (MARQUETTI, 1997) e "escolaridade média da população feminina acima de 25 anos" (BARRO e LEE, 2000). Foram utilizadas observações para 63 países em 1985. Como pode ser observado, à medida que o parâmetro de suavização aumenta a variabilidade da estimativa se reduz, mas o viés aumenta. Como esperado, a utilização de $h = 0,2$ resulta em uma estimativa com muito "ruído", enquanto as estimativas com $h = 0,8$ e $h = 1$ não capturam a mudança de relação entre escolaridade e fertilidade que ocorre em torno dos oito anos de educação. A melhor opção é a utilização de $h = 0,5$, que captura a mudança na relação entre escolaridade média da população feminina e taxa de natalidade que ocorre por volta de oito anos de escolarização. O objetivo é produzir uma estimativa que seja a mais suave possível sem distorcer a relação de dependência entre as variáveis em análise (CLEVELAND e LOADER, 1996a, p. 12).

Existe uma vasta literatura que discute outros métodos para a escolha de *h*. Loader (1995, 1999) discute esta literatura e compara os 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 Mallow. Estes consistem basicamente em empregar alguma medida de aderência, como, por exemplo, minimizar a média da integral do erro ao quadrado ou uma simplificação desta.

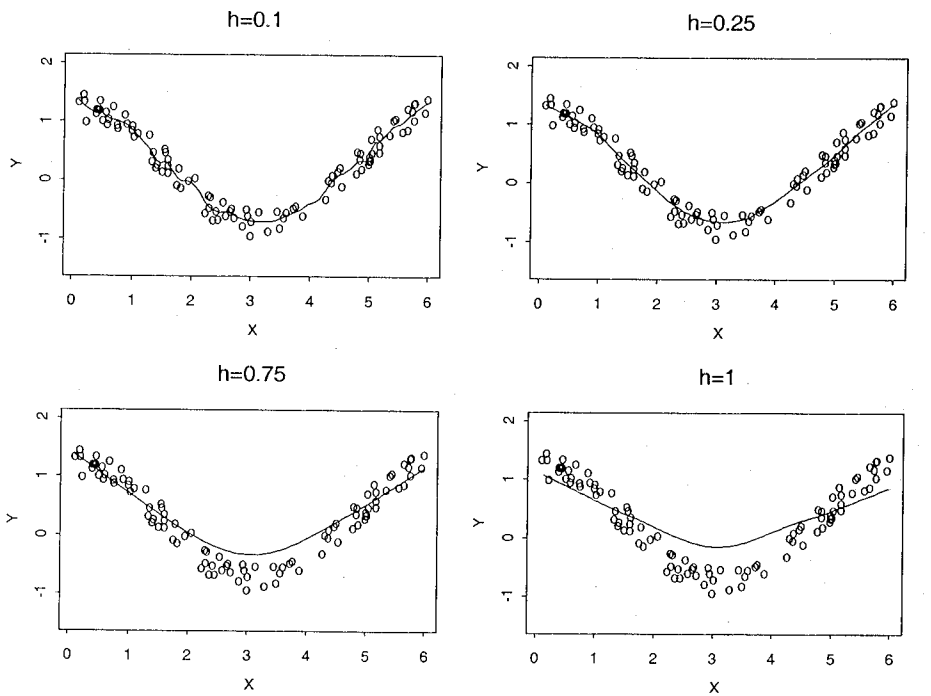

Figura 1: Estimativas de regressão local para diferentes parâmetros de suavização. Nos painéis há 100 pares (X_i, Y_i) gerados aleatoriamente por $Y = \cos(6X)+\delta$, $X \sim U[0, 1)$ e $\delta \sim N(0, 1/2)$.

Os métodos do segundo grupo são baseados em anexos (*plug-ins*). Estes consistem em escrever a função inicialmente estimada, $\hat{g}$, como uma função da g desconhecida e aproximada

por uma expansão em séries 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 Loader (1999, p. 178), os métodos clássicos apresentam melhores resultados em termos práticos, bem como se ajustam a uma grande variedade de casos.

3.2 O grau do polinômio local

Neste caso, a escolha a ser feita é o grau da polinomial local, d, a ser estimada. Esta escolha também afeta a relação entre variância e viés, quanto maior for 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. Bowman e Azzalini (1997), por exemplo, utilizam somente regressão local linear, ou seja, polinomiais de primeira ordem. Alguns estudos indicam que o emprego de polinomiais de ordem impar produz melhores resultados em termos de redução da variância e viés do que os de ordem par. Isto é claramente verdadeiro para estimativas com regressões locais constantes (ordem zero) e lineares (ordem um), não sendo necessariamente correto para estimativas de ordem quadrática e cúbica (CLEVELAND e LOADER, 1996a, p. 36).

A Figura 2 mostra a estimativa de regressão local para as polinomiais de ordem zero, um dois e três, mantido constante o parâmetro de suavização. É possível observar que a regressão local constante apresenta viés na escolaridade média da população feminina de zero a dois anos, contudo o viés não está presente quando a escolaridade média da população feminina é superior aos 8 anos. Este resultado é esperado, pois a taxa de natalidade mantém-se constante, ligeiramente abaixo de dois fi-

lhos, na população feminina com 8 ou mais anos de escolaridade média. Verifica-se também que a utilização de polinomiais de ordem mais elevada ocasiona um pequeno aumento da variância da estimativa. Neste caso, há a possibilidade de aumentar o parâmetro de suavização para reduzir a variância.

A escolha do grau da polinomial é, em sua maior parte, guiada pelos objetivos do pesquisador e pelos dados que estão sendo utilizados. 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 modo 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.

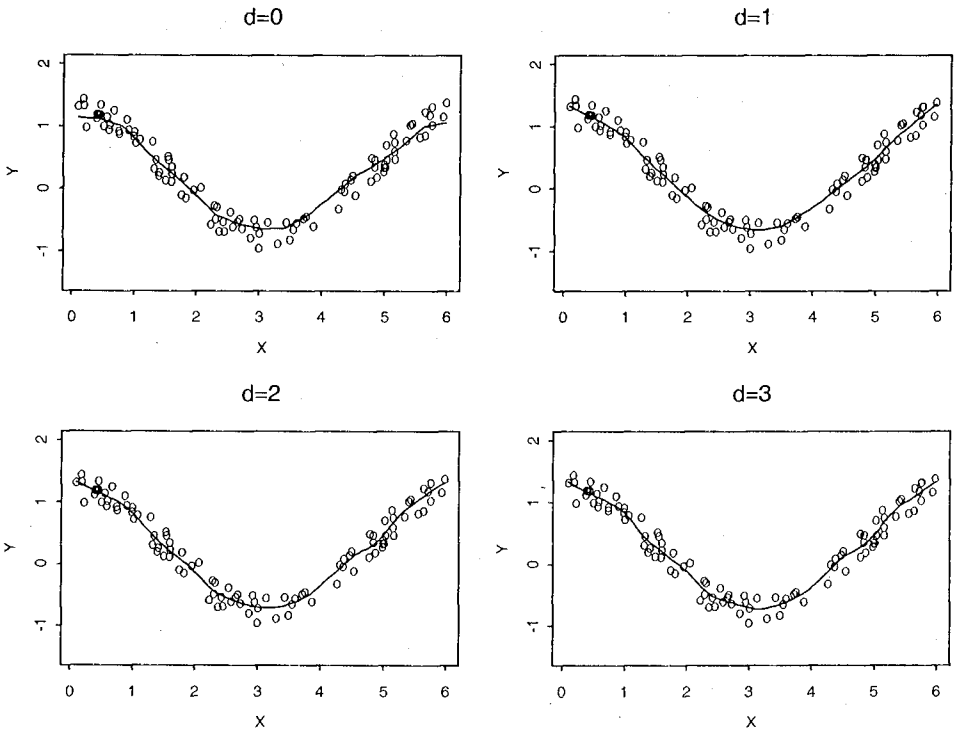

Figura 2: Estimativas de regressão local para diferentes ordens de polinomiais e o mesmo parâmetro de suavização

3.3 A função peso ou nuclear

Outra escolha a ser realizada é a função peso, a qual é responsável por ponderar as observações na vizinhança de cada ponto de interesse, x_0. Segundo Cleveland e Loader (1996a, p. 11) e Loader (1999, p. 23), 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, destacam-se as funções retangular, tri-cúbica, de Epanechnikov e a normal ou Gaussiana.

Para analisar cada uma destas funções, vamos considerar uma variável transformada u_i definida por:

$$u_i = (x_i - x_0)/h_i \qquad (5)$$

Então as funções pesos K são obtidas em função da variável u, isto é, $K(u) = K[(x_i - x_0)/h_i]$. A primeira destas funções é a retangular, a qual pondera as observações na vizinhança de interesse de x_0 com peso um e peso zero para as demais. Esta função é dada por $K(u) = I_{[-1,1]}(u)$, mas é pouco utilizada pois resulta em estimativas com descontinuidades. Uma definição alternativa para esta função é:

$$K = 1 \text{ se } |u| < 1 \qquad (6)$$
$$= 0 \text{ se } |u| \geq 1$$

Uma segunda escolha, mais comum, é a função peso tri-cúbica que é obtida por:

$$K(u) = \left(1 - |u|^3\right)^3 = \left(1 - \left|\frac{x - x_i}{h_i}\right|^3\right)^3 \quad \begin{matrix} \text{se } |u| < 1 \\ = 0 \text{ se } |u| \geq 1, \end{matrix} \qquad (7)$$

onde h_i é determinado por um dos métodos mencionados em 3.1.

Uma terceira escolha é a função peso atribuída a Epanechnikov, a qual é definida por:

$$K = 0{,}75(1 - u^2) \text{ se } |u| < 1 \qquad (8)$$
$$= 0 \text{ se } |u| \geq 1$$

A última escolha é representada pela função peso normal. Como a normal está centrada em x_0, a banda h é o desvio padrão da amostra. Assim, valores que estiverem situados a mais de dois

desvios (2h) receberão um peso negligenciável, pois a área da curva normal além dos dois desvios é muito pequena.

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2} se - \infty < u < +\infty \qquad (9)$$

A função peso a ser utilizada depende dos objetivos do pesquisador. A Figura 3 ilustra as várias funções peso.

3.4 A distribuição resposta

A última escolha a ser feita é a hipótese sobre a distribuição que a regressão local segue. A escolha mais comum e mais simples é que a distribuição seja gaussiana. Contudo, este pode não ser o caso, especialmente na presença de *outliers,* o que resulta em resíduos com distribuições que se afastam da normal nas caudas da distribuição. Neste caso, é possível considerar processos iterativos que reduzem o peso dos *outliers* na estimativa de regressão local. Este método é, por exemplo, empregado por Cleveland (1979).

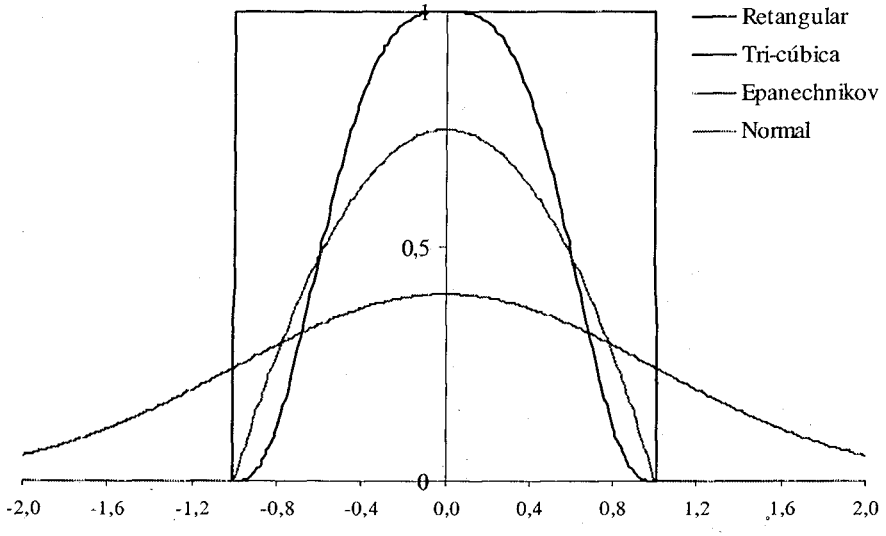

Figura 3: As várias funções pesos utilizadas em regressão local

Regressão local também pode ser expandida para modelos lineares generalizados, por exemplo, para dados binários. Loader (1999, cap. 4) apresenta a estimativa de máxima verossimilhança para regressão local considerando diversas distribuições.

Por fim, deve-se ressaltar que de maneira análoga à regressão paramétrica, uma série de testes gráficos deve ser utilizadas para testar as escolhas e hipóteses subjacentes à análise de regressão local. Cleveland e Devlin (1988, p. 600) e Loader (1999, p. 25) sugerem análise gráfica para testar as hipóteses de normalidade dos resíduos, a hipótese de variância constante, a presença de autocorrelação nos resíduos e de vieses na estimativa.

4 Modelos multivariados

Regressão local pode ser facilmente estendida para múltiplas variáveis explicativas. O modelo a ser estimado assume a forma:

$$y_i = g(x_{1,i} + x_{2,i} + ... + x_{k,i}) + \varepsilon_i \qquad (10)$$

onde há k variáveis independentes, i varia de 1 a n observações e e representa o termo de erro normalmente e independentemente distribuído com média zero e variância constante. No presente contexto, a vizinhança de interesse é definida no espaço das variáveis independentes, e cada observação é ponderada com uma função núcleo (*kernel*) de dimensão k. Uma polinomial de ordem d é estimada utilizando mínimos quadrados ponderados. As mesmas escolhas anteriores devem ser realizadas. Como aponta Jacoby (1998), o objetivo é mesmo que em regressão paramétrica: analisar a dependência da variável dependente para cada uma das k variáveis independentes, levando em consideração o efeito das demais variáveis.

O resultado é uma superfície suavizada de dimensão $k + 1$. Dois problemas surgem com o aumento da dimensão.

O primeiro é denominado de problema da dimensionalidade. À medida que o número de variáveis independentes aumenta, o estimador não paramétrico deve ponderar sobre regiões muito grandes do espaço, aumentando rapidamente o número de observações necessário para produzir uma estimativa de qualidade (HASTIE, TIBSHIRANI, FREDMAN, 2001, cap. 2). Uma forma de tratar o problema da dimensionalidade é a consideração de alguma

hipótese sobre a estrutura do modelo. Modelos aditivos generalizados (HASTIE e TIBSHIRANI 1990) são uma forma de solucionar o problema da dimensionalidade.

Um segundo problema é a visualização em dimensões maiores do que três. É possível utilizar gráficos de três dimensões, os quais podem apresentar problemas de interpretação. Uma alternativa para análises com mais de duas variáveis independentes é empregar gráficos condicionados, denominados por Cleveland (1993, p. 182) de *coplots*. Estes mostram a relação entre a variável dependente e uma variável independente, condicionada aos valores das demais independentes. Uma aplicação com modelos multivariados é realizada na sétima seção deste trabalho.

5 Estimando derivadas

Muitas vezes o interesse do pesquisador está na estimativa da derivada, tanto de primeira quanto de segunda ordem. A derivada primeira mostra o efeito da mudança da variável independente sobre a variável dependente. A derivada segunda mostra a curvatura da função estimada, em particular, se esta é côncava ou convexa. Na análise paramétrica a estimativa de derivadas é realizada a partir de formas funcionais previamente estabelecidas. Muitas vezes estes modelos possuem problemas de especificação, resultando em estimativas com elevado viés. A vantagem de utilizar regressão local na estimativa de derivadas é o menor viés que esta apresenta em relação à análise de regressão tradicional.

Loader (1999, p. 101) discute a estimativa de derivada em regressão local. Ele chama a atenção que, ao estimar derivadas em regressão local, obtém-se, na verdade, a inclinação local da curva estimada e não a derivada exata da curva estimada. Contudo, argumenta que, se a regressão local possui uma boa estimativa dos dados, então a inclinação local é uma aproximação consistente da derivada. Pagan e Ullah (1999, cap. 4) discutem diversos métodos não-paramétricos para estimar derivativas e apresentam alguns exemplos.

A Figura 4 apresenta a estimativa da derivada primeira para a relação entre as variáveis X e Y. Observa-se que a estimativa capta a mudança na curvatura de Y, pois torna-se positiva para valores de X maior do três.

6 Inferência estatística

A discussão até o presente tem enfatizado o ajustamento de uma curva entre as variáveis em análise. Contudo, uma das vantagens de regressão local é que suas propriedades estatísticas têm sido estudadas permitindo que se realizem inferências sobre os resultados obtidos. Estas propriedades permitem que se determinem intervalos de confiança e se realizem testes de hipóteses para as estimativas da regressão local.

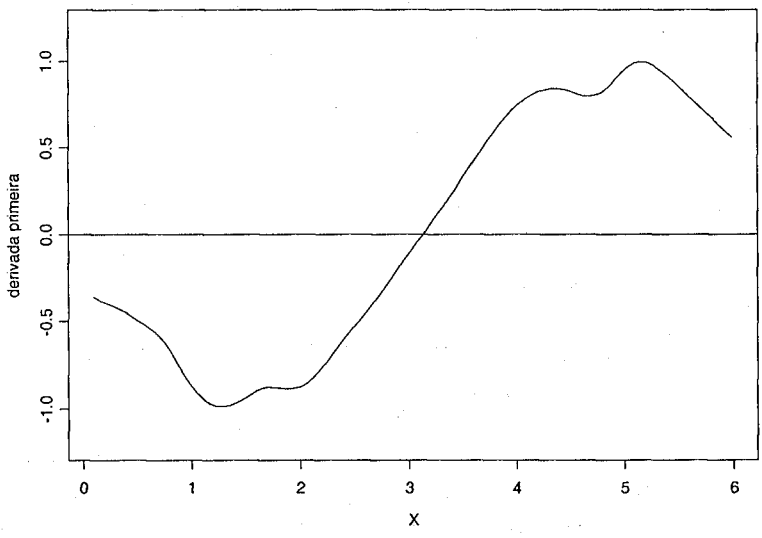

Figura 4: A derivada primeira da relação entre *X* e *Y*.

Cleveland, Devlin e Grosse (1988, p. 95-99) e Cleveland e Devlin (1988, p. 598-599) desenvolveram as propriedades estatísticas para regressão local em analogia aos procedimentos estatísticos utilizados no método dos mínimos quadrados ordinários. As hipóteses empregadas para derivar as propriedades são: (i) a função estimada, $\hat{g}$, aproxima-se em muito da verdadeira função, g, de modo que o viés é negligenciável; (ii) $\hat{g}$ é uma combinação linear de y_i e (iii) y possui distribuição normal com variância σ^2.

A utilização das duas primeiras hipóteses permite escrever:

$$\hat{y} = \sum_{j=1}^{n} l_{ij} y = Ly \qquad (11)$$

onde $\hat{y}$ representa o vetor das estimativas nos pontos de interesse, L representa a matriz de suavização cujas linhas são formadas pelos pesos, l_{ij}, empregados para estimar $\hat{y}$ e y representa o valores da variável dependente nos pontos de interesse. Por sua vez, o vetor dos resíduos, e, é estimado através de $(I - L)y$, onde I é a matrix identidade. Com a utilização da terceira hipótese obtém-se:

$$V(\hat{y}) = LV(y)L' = \sigma^2 \varepsilon LL' \qquad (13)$$

Logo, é necessário um estimador da variância do erro, σ^2_ε, para calcular a variância da estimativa de regressão local. Uma estimativa de σ^2_ε é obtida através de:

$$\hat{\sigma}^2 \varepsilon = \sum (yi - \hat{y}i) / gl_{erro} = SQR/gl_{erro} \qquad (14)$$

onde gl_{erro} representa os graus de liberdade da soma dos resíduos ao quadrado. Em analogia ao método de regressão linear, os graus de liberdade dos resíduos são determinados pelo traço da matrix $(I_n - L)$, que é igual ao número de observações, n, menos o traço de L. Este último é chamado de "número equivalente de parâmetros" na estimativa de regressão local. Empregando estes resultados, é possível calcular o intervalo de confiança pontual (*pointwise*) para a regressão local utilizando a distribuição t. O intervalo de confiança pontual é calculado para cada ponto de interesse da variável independente, sendo diferente e mais simples de ser obtido do que o intervalo de confiança global, o qual é calculado para todo o conjunto de observações da variável independente. Para uma apresentação dos dois intervalos de confiança, ver Hastie e Tibshrani (1990, cap. 3).

A Figura 5 apresenta a estimativa da regressão local com o intervalo de confiança pontual de 95% para a relação entre a taxa de fertilidade e a escolaridade média da população feminina. Como pode ser observado, o intervalo de confiança é mais estreito no centro da estimativa, alargando-se nas observações que se encontram próximas às pontas da distribuição. Este comportamento é muito semelhante ao da análise de regressão linear.

Outra importante aplicação de inferência estatística é na realização de testes de hipóteses entre modelos aninhados. Dois modelos são ditos aninhados quando um deles pode ser escrito como um caso especial do modelo mais geral. Entre os testes possíveis de serem realizados estão a comparação entre regressão linear e re-

gressão local, entre regressões locais com diferentes bandas e testes de significância das variáveis em modelos de regressão local multivariados. Para testar estas hipóteses, utiliza-se o teste F aproximado, que foi estudado, entre outros, por Cleveland, Devlin e Grosse (1988), Cleveland e Devlin (1988) e Hastie e Tibshrani (1990). Cleveland, Devlin e Grosse (1988) utilizam estudos de Monte Carlo para mostrarem que o teste F aproximado para diferentes tamanhos da amostra, bandas e grau da polinomial utilizado em regressão local tem uma distribuição muito próxima a distribuição F.

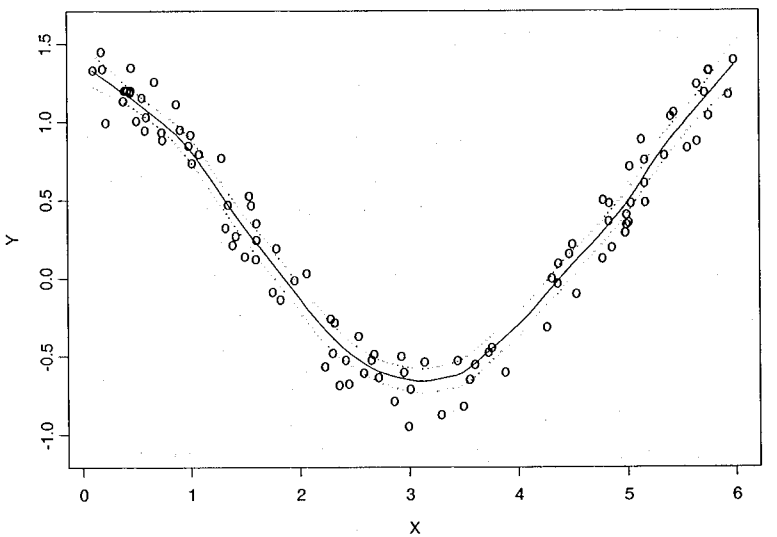

Figura 5: A estimativa de regressão local com intervalo de confiança pontual de 95% para a relação entre *X* e *Y*.

O teste F aproximado também é construído por analogia aos testes para os modelos de regressão linear. Por exemplo, para testar se a relação entre duas variáveis é linear ou não linear, o modelo restrito toma a forma de uma regressão linear simples, $Y_i = \alpha + \beta X_i + \varepsilon_i$, com $n - 2$ graus de liberdade e soma do quadrado dos resíduos igual a SQR_0. O modelo não linear é representado por uma regressão local cuja soma do quadrado dos resíduos é igual a SQR_1. O teste da hipótese nula de linearidade toma a forma:

$$F = \frac{SQR_0 - SQR_1 / traço(L) - 2}{SQR_1 / n - traço(L)} \qquad (15)$$

com graus de liberdade igual a *traço(L) - 2* e *n - traço(L)*. Na próxima seção é realizado um exemplo de aplicação do teste F aproximado.

7 Aplicações de regressão local

Aplicações de regressão local em ciências sociais ainda são em número reduzido, em particular, frente as potencialidades do método. No presente trabalho três aplicações para a área de desenvolvimento econômico são apresentadas com o objetivo de mostrar as potencialidades do método.

A Figura 6 mostra a estimativa de regressão local para a relação entre as variáveis taxa de fertilidade por mulher (Marquetti, 1997) e escolaridade média da população feminina acima de 25 anos (Barro e Lee, 2000). Foram utilizadas observações para 63 países em 1985. Como pode ser observado ocorre uma mudança de relação entre escolaridade e fertilidade em torno dos oito anos de educação. A taxa de natalidade mantêm-se constante, ligeiramente abaixo de dois filhos, na população feminina com 8 ou mais anos de escolaridade média.

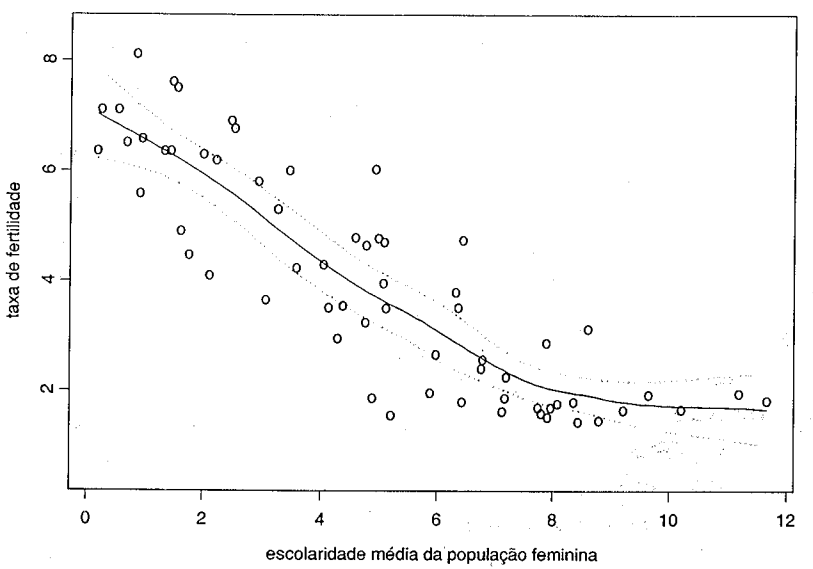

Figura 6: A estimativa de regressão local com intervalo de confiança pontual de 95% para a relação entre a escolaridade média da população feminina e a taxa de fertilidade. Fonte: Marquetti (1997) e Barro e Lee (2000).

É possível testar se a relação entre a taxa de fertilidade e a escolaridade média da população feminina é não linear. Para tal é necessário considerar o modelo linear como aninhado ao modelo estimado de regressão local. Utilizando o teste F aproximado para comparar as duas estimativas obtêm-se

$$F = \frac{64{,}99 - 53{,}97 / 4{,}25 - 2}{53{,}97 / 63 - 4{,}25} = 5{,}33 \qquad (16)$$

com 2,25 e 58,75 graus de liberdade. Portanto, a hipótese nula de que a relação entre a taxa de fertilidade e a escolaridade média da população feminina é linear é rejeitada a um teste de 5% de grau de confiança.

A Figura 7 mostra a estimativa de regressão local, linha cheia, e a estimativa do modelo de regressão linear, linha pontilhada, para a relação entre as variáveis em estudo. Observa-se que o modelo não linear é capaz de captar a mudança na fertilidade média que ocorre em torno de oito anos de educação formal, enquanto o modelo linear não capta esta mudança. Além disso, o modelo linear é viesado, a estimativa do intercepto e da declividade é afetada pelas observações com elevada educação média da população feminina e reduzida taxa de natalidade. O processo de visualização confirma o resultado do teste F aproximado de que a regressão local produz uma estimativa de melhor qualidade do que o modelo de regressão linear.

O emprego de múltiplas variáveis explicativas dificulta o processo de visualização. A Figura 8 mostra a superfície estimada por regressão local entre as variáveis taxa de fertilidade, escolaridade média da população feminina e produtividade do trabalho. Essa última informação é medida em dólares internacionais de 1985 e foi em Summer e Heston (1991).

O gráfico mostra que não há uma mudança significativa da relação entre a taxa de fertilidade e a escolaridade média da população feminina para os países com a produtividade do trabalho abaixo de US$ 15.000. Para estes países, a taxa média de fertilidade declina até atingir dois filhos a medida em que a escolaridade média da população feminina se aproxima dos oito anos, quando permanece relativamente constante. O aumento da produtividade do trabalho, aparentemente, não tem efeito sobre a taxa de fertilidade para países com escolaridade média acima de oito anos. Por sua vez, o efeito do aumento da produtividade do

trabalho é maior para os países com uma produtividade média acima de US$ 15.000. Também é possível observar que nos países com elevada produtividade do trabalho, o aumento da escolaridade não possui efeito sobre a taxa de fertilidade.

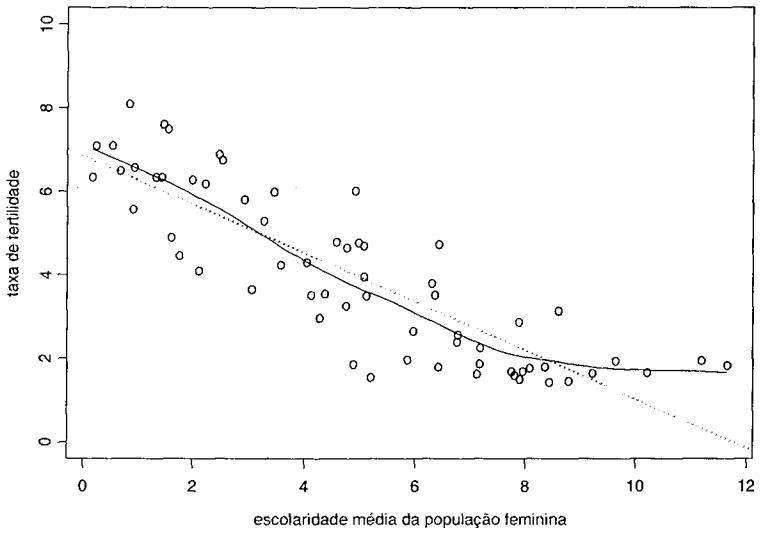

Figura 7: Comparação entre a estimativa de regressão local (linha cheia) e a estimativa de regressão linear (linha pontilhada) para a relação entre a escolaridade média da população feminina e a taxa de fertilidade. Fonte: Marquetti (1997) e Barro e Lee (2000).

A Figura 9 apresenta a estimativa de regressão local para a relação entre o rendimento nominal médio mensal em reais das pessoas residentes com 10 anos ou mais de idade e a taxa de alfabetização da população com 10 anos ou mais de idade para os municípios brasileiros em 2000. Os municípios de Campos de Júlio e Santana de Parnaíba são *outlyiers* e não foram considerados na análise. Observa-se uma relação positiva e não linear entre o rendimento nominal médio mensal e a taxa de alfabetização. A relação entre as variáveis em análise é côncava, o acréscimo do rendimento nominal médio nos municípios pobres ocasiona um aumento na taxa de alfabetização maior do que o mesmo acréscimo do rendimento nos municípios mais ricos.

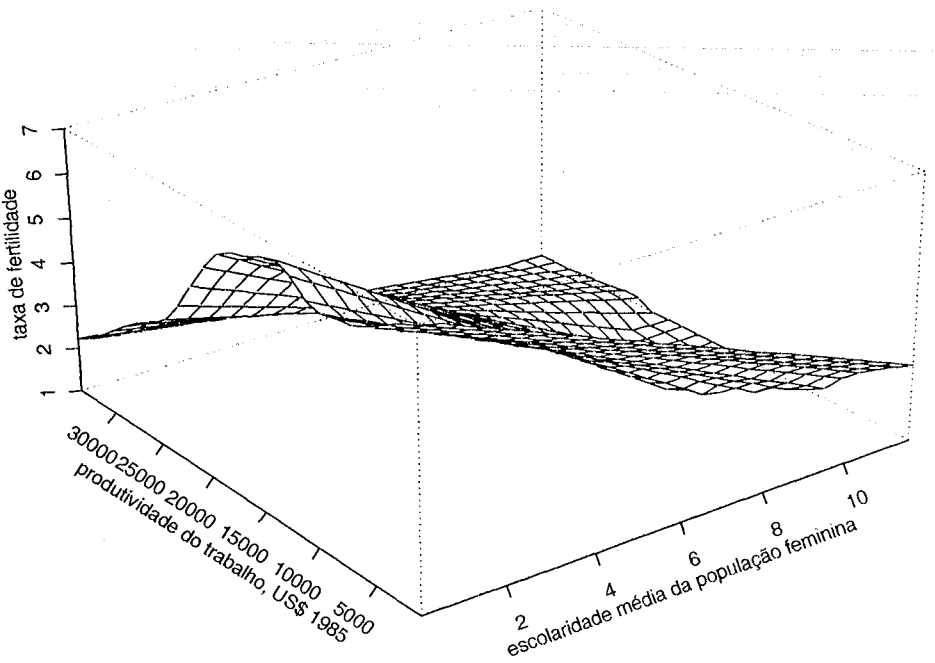

Figura 8: Estimativa de regressão local em um modelo multivariado entre para a relação entre a escolaridade média da população feminina, a produtividade do trabalho e a taxa de fertilidade. Fonte: Marquetti (1997), Barro e Lee (2000) e Summer e Heston (1991).

É importante ressaltar que o modelo linear muitas vezes não é capaz de captar a verdadeira relação entre as variáveis em análise mesmo no caso em que o modelo toma a forma semilog ou log-linear. A Figura 10 apresenta a estimativa de regressão local, linha cheia, e a do modelo de regressão linear, linha pontilhada, para a relação entre o logaritmo do rendimento nominal médio mensal e a taxa de alfabetismo. A estimativa do modelo de regressão linear é viesada para os municípios com maior rendimento nominal médio mensal.

Por sua vez, a Figura 11 mostra a estimativa de regressão local para a relação entre o PIB per capita e o Índice de Desenvolvimento Socioeconômico para a educação (IDESE-Educação) nos municípios do Rio Grande do Sul em 2000. Para a metodologia de construção do IDESE-Educação e para a estimativa do PIB municipal

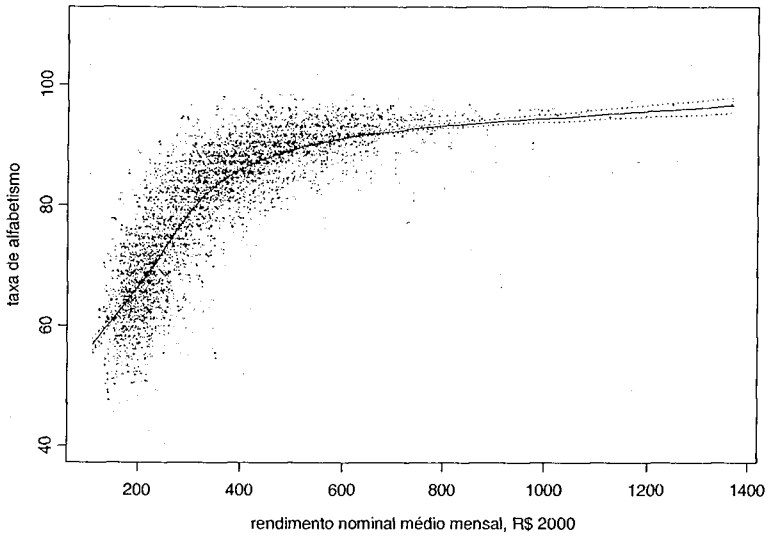

Figura 9: A estimativa de regressão local com intervalo de confiança pontual de 95% para a relação entre a escolaridade média da população feminina e a taxa de fertilidade. Fonte: Marquetti (1997) e Barro e Lee (2000).

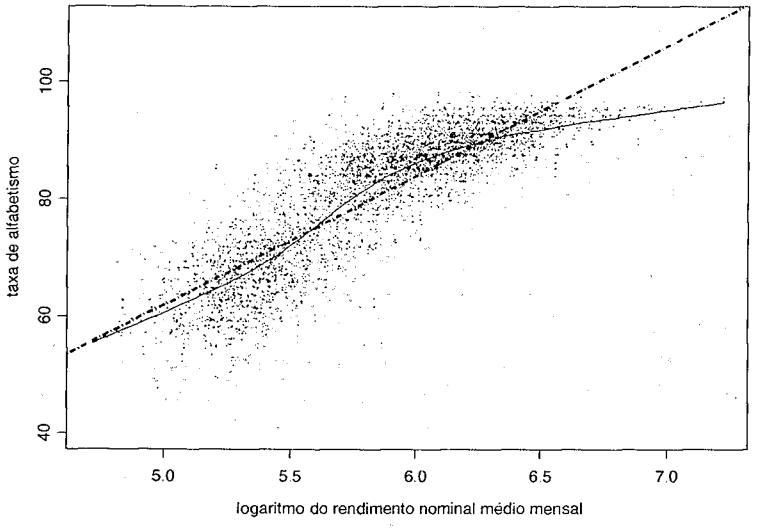

Figura 10: Comparação da estimativa de regressão local (linha cheia) com a de regressão linear (linha pontilhada), para a relação entre o logaritmo do rendimento nominal médio mensal e a taxa de analfabetismo. Fonte: IBGE (2003).

ver, respectivamente, FEE (2003a) e FEE (2003b). O município de Triunfo é um *outlyier*, não sendo considerado na análise.

Como pode ser observado, o IDESE-Educação aumenta nas cidades com renda per capita até R$ 10.000 a preços de 2000, permanecendo relativamente constante nas que possuem uma renda per capita superior a esse montante. Logo, ocorre uma mudança na relação entre as variáveis em estudo quando a renda per capita municipal atinge R$ 10.000 a preços de 2000. Regressão local permite detectar facilmente essas mudanças, o que não ocorre com o método de regressão linear.

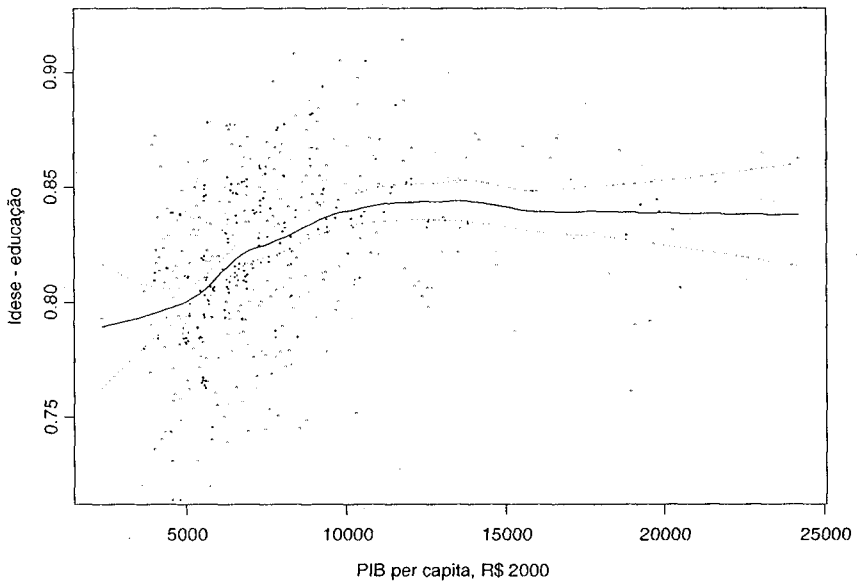

Figura 11: A estimativa de regressão local com intervalo de confiança pontual de 95% para a relação entre o PIB per capita e o IDESE-Educação para os municípios do Rio Grande do Sul em 2000. Fonte: FEE (2003a e 2003b).

8 Conclusão

Regressão local é um poderoso instrumento de análise de regressão. Sua vantagem em relação ao método de mínimos quadrados é sua flexibilidade, ao não considerar uma forma funcional

prévia não somente elimina-se a hipótese de linearidade, bem como permite que os dados "falem por si próprios". Sua vantagem em relação aos demais métodos não paramétricos é justamente sua proximidade com o método de mínimos quadrados. Como os demais métodos não-paramétricos, suas desvantagens em relação aos métodos tradicionais de regressão são os problemas de dimensionalidade e de visualização em análises com mais de três variáveis. Contudo, é possível contornar estes problemas ao assumir certa estrutura dos dados, por exemplo, que as variáveis independentes tenham efeito aditivo sobre a variável dependente (Hastie e Tibshirani, 1990).

Regressão local pode ser empregada no estudo de relações binárias e multivariadas entre variáveis, na descoberta da forma funcional entre as variáveis, na análise das hipóteses dos modelos de regressão, na realização de testes de hipóteses, entre outros. Enfim, ela pode ser visto como um método disponível para modelar a estrutura funcional entre as variáveis em estudo. Contudo, é importante salientar que regressão local não é um método competitivo à análise de regressão tradicional. Os métodos paramétricos e não-paramétricos possuem vantagens e desvantagens particulares, devendo serem vistos como métodos complementares e que devem fazer parte do conhecimento dos que trabalham com a análise de informações.

Por fim, deve-se ressaltar que este texto apresenta somente os princípios básicos de regressão local, bem três aplicações ilustrativas do método. De maneira similar aos outros métodos de regressão não-paramétrica, regressão local tem sido desenvolvido e incorpora uma série de avanços e aplicações aqui não considerados. Loader (1999) apresenta muito destes avanços. Contudo, mesmo este autor, um dos principais pesquisadores do tema, não aborda em seu livro os desenvolvimentos em séries temporais (Masry e Fan, 1997) e em aprendizagem de robôs (*robot learning*) (Atkeson, Moore e Schaal, 1997). Os analistas de dados em engenharia, economia, sociologia, ciência política, medicina, estatística, entre outras áreas, tem muito a ganhar ao utilizarem regressão local.

9 Referências bibliográficas

ATKESON, Christopher,; MOORE, Andrew; SCHAAL, Stefan. Locally WWeighted LLearning for CControl. *Artificial Intelligence Review*, v. 11, p. 75-113, 1997.

BARRO, R.; LEE, J. *International Data on Educational Attainment Updates and Implications*, NBER Working Paper, September, 2000.

BOWMAN, Adrian; AZZALINI, Adelchi. *Applied Smoothing Techniques for Data Analysis*: The Kernel Approach with S-Plus Illustrations. Oxford: Oxford University Press, 1997.

CLEVELAND, W. S.; DEVLIN, Susan J.; GROASSE, Eric. Regression by LLocal FFitting: MMethods, PProperties, and CComputational AAlgorithms. *Journal of Econometrics*. New York: Elsevier Science. v. 37, n. 1, p. 87-114, January 1988.

CLEVELAND, W. S.; LOADER, Clive R. Smoothing by local Regression: Principles and Methods). In: W. Hardle; M. G. Schimek (eds). *Statistical Theory and Computational Aspects of Smoothing* Heidelberg (Germany): Physica-Verlag, p. 10-49, 1996a.

CLEVELAND, W. S.; LOADER, Clive R.. Rejoinder to Discussion of Smoothing by local Regression: Principles and Methods. In: W. Hardle; eand M. G. Schimek (editors), *Statistical Theory and Computational Aspects of Smoothing*, Heidelberg (Germany): Physica-Verlag, p. 113-120, 1996b.

CLEVELAND, William. Robust Locally Weighted Regression and Smoothing Scatterplots, *Journal of the* American *Statistical Association*, v. 74, p. 829-836, 1979.

CLEVELAND, William. Robust Locally Weighted Regression and Smoothing Scatterplots, *Journal of the American Statistical Association*, v. 74, p. 829-836, 1979.

CLEVELAND, William. *Visualizing Data.* Summit: Hobart Press, 1993.

FAN, J.; GIJBELS, Irene. *Local Polynomial Modeling and Its Applications*. London: Chapman and Hall, 1996.

FAN, Jianqing. Prospects of nonparametric modeling. *Journal of American* Statistical *Association*. Alexandria (VA): ASA (American Statistical Association). V. 05, n. 452, p. 1296-1300, 2000.

FAN, Jianqing. Design Adaptive Nonparameteric Regression, *Journal of the American Statistical Association*, v. 87, p. 998-1004, 1992.

FEE. *Índice de Desenvolvimento Socieconômico do RS*. Porto Alegre: FEE, 2003a. 1 CD-ROM.

FEE. *PIB-Municipal do RS 1985-01*. Porto Alegre: FEE, 2003b. 1 CD-ROM.

HASTIE, Trevor; LOADER, Clive. Local Regression: Automatic Kernel Carpentry. 1993. *Statistical Science*,. v. 8, p. 120-143, 1993.

HASTIE, Trevor; TIBSHIRANI, Robert. *Generalized Additive Models*. London: Chapman and Hall, 1990.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. *The Elements of Statistical Learaning*: data mining, inference e prediction. New York: Springer-Verlag, 2001.

IBGE. *Base de informações municipais 4*. Rio de Janeiro: IBGE, 2003. 1 CD-ROM.

JACOBY, William. Statistical Graphics for Visualizing Multivariate Data. Thousand Oaks: Sage Publications, 1998.

LOADER, Clive. *Local Regression and Likelihood.* New York: Springer-Verlag, 1999.

LOADER, Clive. *Old Faithful Erupts: Bandwidth Selection Reviewed.* Working paper, AT&T Bell Laboratory, 1995.

MARQUETTI, Adalmir. *Extended Penn World Tables*: economic growth data on 118 countries. New York, New School University, (http://homepage.newschool.edu/ ~foleyd/epwt/).

MASRY, Elias; FAN, Jianqing. Local Polinomial Estimation of Regression Functions for Mixing Processes. *Scandinavian Journal of Statistics*, v. 24, p. 165-179, 1997.

NADARAYA, E. A. On Estimating Regression. *Theory of Probability and ilts Applications*, v. 9, p. 141-142, 1964.

PAGAN, Adrian; ULAH, Aman. *Nonparametric Econometrics*. Cambridge: Cambridge University Press, 1999.

STONE, C. J. Consistent Nonparametric Regression, with discussion. *The Annals of Statistics*, 5, p. 549-645, 1977.

STONE, C. J. Optimal Rates of Convergence for Nonparametric Estimators. *The Annals of Statistics*, v. 8, p. 1348-1360, 1980.

SUMMERS, R; HESTON, A.. The Penn World Table (Mark 5): An Expanded Sset of International Comparisons, 1950-1988, *Quarterly Journal of Economics*, v. 106, p. 327-368, 1991.

WATSON, G. S. Smooth Rregression Aanalysis, *Sankhya, Series. A*, v. 26, p. 359-372, 1964.

Apêndice

Diversos softwares estatísticos comerciais e livres possibilitam a aplicação de regressão local, contudo nem todos oferecem a possibilidade de inferência estatística. Entre os softwares que possibilitam análise de inferência estatística estão o S-Plus (http://www.splus.mathsoft.com), o R que é uma implementação livre da linguagem S (http://www.r-project.org), o programa XploRe (http://www.xplore-stat.de), a versão 8 dos SAS (http://www.sas.com) e o programa Gauss (http://www.aptech.com). Entre os pacotes que possibilitam a estimativa de regressão local, mas sem a possibilidade de inferência estatística estão o E-views (http://www.eviews.com) e o SPSS (http://www.spss.com). Contudo, é importante ter presente que novas versões de softwares surgem rapidamente, e que a informação aqui apresentada pode já estar desatualizada, bem como regressão local pode estar disponível em programas não listados acima.