ÁREA DE INTERESSE: TEORIA ECONÔMICA E MÉTODOS QUANTITATIVOS

TÍTULO: ECONOMETRIA NÃO PARAMÉTRICA E EXPECTATIVA DE VIDA NOS MUNICÍPIOS DO NORDESTE: UMA APLICAÇÃO DO ESTIMADOR DE NADARAYA-WATSON.

Palavras-Chaves: **Regressão Não-Paramétrica, Estimador de Nadaraya-Watson; Expectativa de Vida; Região Nordeste.**

AUTORES:

ANDREI GOMES SIMONASSI
DOUTORANDO EPGE/FGV-RJ
Endereços:
simonassi@fgvmail.br

Escola de Pós-Graduação em Economia
Fundação Getulio Vargas
Praia de Botafogo 190, 11º andar
CEP: 22.250-900
Fax: (55-21) 25524898 - tel.: (55-21) 2559-5860

JOSÉ OSWALDO CÂNDIDO JÚNIOR
DOUTORANDO EPGE/FGV-RJ
ECONOMISTA DO IPEA

Endereços Eletrônicos:
candido@fgvmail.br
j.oswaldo@uol.com.br

Endereço Residencial:
Rua Tonelero, 68 – Apto.603 – Copacabana – Rio de Janeiro-RJ
CEP: 22030-000
Tel.: (021) 3208-4371

Escola de Pós-Graduação em Economia
Fundação Getulio Vargas
Praia de Botafogo 190, 11º andar
CEP: 22.250-900
Fax: (55-21) 25524898 - tel.: (55-21) 2559-5860

RESUMO:
Os métodos de Econometria Não-Paramétrica ainda são pouco explorados como ferramenta de análise econômica. Esse artigo traz uma breve discussão sobre regressão não-paramétrica, em particular, sobre o estimador de Nadaraya-Watson. Aplica-se tal técnica para investigar a relação entre a esperança de vida ao nascer e as condições sócio-econômicas dos municípios nordestinos a partir de variáveis como renda per capita, proporção de domicílios com água canalizada e proporção de domicílios com acesso a rede de esgotos. Identifica-se que as variáveis renda per capita, proporção de domicílios com água encanada e com acesso à rede geral de esgotos apresentam uma relação positiva com a esperança de vida ao nascer em regressões com apenas um regressor. Entretanto, com exceção da variável água encanada, a relação entre as variáveis não obedece uma relação de linearidade e a aplicação de uma técnica não-paramétrica mostra-se superior aos modelos de regressão paramétrica tradicionais. Os resultados também apontam que os benefícios sociais na expectativa de vida são mais elevados nos municípios menos desenvolvidos.

ABSTRACT
The methods of Nonparametric Econometric are yet to be more exploited as an instrument of economic analysis. This article relies on a short discussion about nonparametric regression, focusing primarily the Nadaraya-Watson estimator. The method is applied to a relationship between life expectancy and the social economic environment in the Brazil northeastern municipalities, taking the following set of explanatory variables: per capita income, ratio of residences that has piped water and ratio of proportion of residences provided by sewerage system. The empirical evidences identify a positive correlation between each explanatory variable and life expectancy. That is no evidence toward linearity in the models, what demonstrate the superiority of a nonparametric technique over a parametric one. The results also point out that social benefits in life expectancy are higher for the less development municipalities.

ECONOMETRIA NÃO PARAMÉTRICA E EXPECTATIVA DE VIDA NOS MUNICÍPIOS DO NORDESTE: UMA APLICAÇÃO DO ESTIMADOR DE NADARAYA-WATSON

Andrei Gomes Simonassi
José Oswaldo Cândido Júnior

1. INTRODUÇÃO

A diferença básica entre um modelo de regressão paramétrico e um outro não paramétrico envolve a quantidade de suposições que o econometrista estaria disposto a fazer e qual peso ele pretende dar aos dados por si mesmos.

Dado um conjunto de pontos $(X_i, Y_i)_{i=1}^{n}$, amostra aleatória de duas variáveis X e Y, o interesse consiste em investigar uma relação funcional entre as duas variáveis na forma:

$$Y_i = g(X_i) + e_i \tag{1.1}$$

Um modelo paramétrico assume que $g(X)$ é uma função desconhecida em um número finito de parâmetros, de forma que a tarefa consistiria em estimar os parâmetros desconhecidos, por exemplo por mínimos quadrados. Em um modelo não paramétrico a relação funcional entre as duas variáveis vive num espaço de funções muito mais amplo: assumimos apenas que $g(X)$ esta num espaço de funções seguindo algumas restrições convenientes e buscamos uma combinação linear de funções desse espaço que aproximem bem $g(X)$.

Dentre os métodos utilizados para regressão não paramétrica destacam-se o método de *Kernel* e o via *Splines*. Este primeiro será objeto deste trabalho, que consistirá em uma aplicação simples do estimador de núcleo conhecido como estimador de Nadaraya-Watson.

Neste contexto este artigo pretende investigar, a partir de um contexto não paramétrico, a relação entre a esperança de vida ao nascer nos municípios nordestinos e suas condições sócio-econômicas. A priori, espera-se que essas condições determinem as taxas de mortalidade por idade e conseqüentemente as expectativas de vida nos municípios. Para representar essas condições sócio-econômicas foram utilizados um indicador de renda, que determina a capacidade das famílias terem acesso aos bens que preservariam suas condições de vida, tais como acesso a médicos, remédios, alimentação adequada, lazer, dentre outros e dois indicadores representativos das condições sociais dos municípios: o percentual de domicílios com acesso à água canalizada e o percentual de domicílios com rede geral de instalações sanitárias, ou seja, com acesso a uma rede geral de esgoto.

O IBGE, segundo dados do Censo 2000, identificou 1.159 municípios brasileiros com taxas de mortalidade infantil acima de 40 óbitos por mil nascidos vivos. Desses municípios, 93,7% estão na Região Nordeste, o que certamente irá afetar a variável esperança de vida ao nascer, que é determinada pelo conjunto das taxas de mortalidade

por idade. Por outro lado, ainda segundo o Censo de 2000, dos 10,4 milhões de domicílios brasileiros com esgotamento sanitário inadequado, cerca de 4 milhões estão na região Nordeste. Além disso, em 2000 a taxa nacional de mortalidade de crianças menores de cinco anos que moram em domicílios adequados (aqueles com água e esgoto) foi de 26,1 por mil. Já para as crianças que residem em domicílios inadequados, a taxa atingiu a 44,8 por mil, alcançando o patamar de até 66,8 por mil no Nordeste.

Portanto, esses dados sugerem que os indicadores de acesso à água canalizada e esgotamento sanitário têm importância crucial sobre as expectativas de vida da população, sobretudo da infantil. E esse grau de importância se acentua da região Nordeste que apresentou os mais elevados indicadores de mortalidade infantil.

2. ASPECTOS METODOLÓGICOS

O método de *Kernel* é um método não paramétrico para estimação de curvas de densidade onde cada observação é ponderada pela distância em relação a um valor central, o núcleo. A idéia é centrar em cada observação x, onde se queira estimar a densidade, uma janela b, que define a vizinhança de x e os pontos que pertencem a estimação. Tal método pode ser generalizado, entretanto, para o caso de uma regressão, bastando para isso observarmos que no modelo descrito em (1.1), o que procuramos é uma estimativa $g(X)$ de $E(Y|X=x)$. Pela definição de esperança condicional temos:

$$E(Y|X=x)=\int y\, f_{Y/X}(y|x)dy=\frac{\int y\, f(x,y)dy}{f(x)} \qquad (1.2)$$

Pelo método do *Kernel* para estimação de densidades, já temos uma estimativa para o denominador da expressão indicada em (1.2):

$$\hat{f}(x)=\frac{1}{Nb}\sum_{i=1}^{N}K\left(\frac{X_i-x}{b}\right) \qquad (1.3)$$

onde: b – Janela ou parâmetro de suavização;
N : Número de observações
K: Núcleo.

Para estimar o numerador dessa expressão note que podemos estimar a densidade conjunta através de um núcleo multiplicativo, de forma a obter:

$$\hat{f}_b(x,y)=\frac{1}{N}\sum_{i=1}^{N}K_{b_1}(X_i-x)K_{b2}(Y_i-y)$$

Para obter o valor esperado de Y, a regressão não-paramétrica pondera os valores observados de Y, os Y_i, pela distância de cada X_i em relação à x. Ademais, regressão de Nadaraya-Watson utiliza uma ponderação normalizada para Y_i que soma um e que atribui maior peso sobre a observação Y_i, se o correspondente regressor X_i está mais próximo de x . Substituindo a densidade conjunta por essa estimativa no numerador da equação (1.2) e fazendo algumas manipulações algébricas temos o estimador de Nadaraya-Watson:

$$\hat{f}(x) = \frac{1}{N} \frac{\sum_{i=1}^{N} Y_i K\left(\frac{X_i - x}{b}\right)}{\frac{1}{N}\sum_{i=1}^{N} K\left(\frac{X_i - x}{b}\right)} = \sum_{i=1}^{N} W_{bi}(x).Y_i \qquad (1.4)$$

onde $W_{bi}(x)$ é o peso sobre a observação Y_i tal que:

$$W_{bi} = \frac{K_b(X_i - x)}{\sum_{i=1}^{N} K_b(X_i - x)}$$

Tomando algumas suposições não muito restritivas sobre a distribuição dos erros e_i e sobre a derivabilidade e continuidade das funções de densidade de *X* e Y, mostra-se que o estimador proposto em (1.3) converge em probabilidade para a função $g(X)$ do modelo descrito em (1.1), desde que $b \to 0$ *e* $Nb \to \infty$ *quando* $N \to \infty$.

Ademais, destaca-se que existem várias funções núcleo, dentre as quais a utilizada no processo de estimação pelo programa EasyReg (Bierens, 2004) é a Gaussiana (Normal) descrita abaixo:

$$\hat{f}_{N,i}(x) = \left(\left(Nb\sqrt{2p}\right)^k \sqrt{\det\left(\hat{\Omega}\right)}\right)^{-1} .\exp\left[\frac{-1}{2Nb^2}(X_i - x)' \hat{\Omega}^{-1}(X_i - x)\right];$$

com;

$$\hat{\Omega} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})(X_i - \overline{X})'$$

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

e onde k refere – se à dimensão do vetor X.

2.1 Seleção da Janela

Vale ressaltar que um aspecto delicado desse método de regressão é a escolha do parâmetro de suavização *b*. Assim como no método de Kernel para estimação de densidades, o parâmetro de suavização tem uma influência decisiva na estimativa obtida da função de regressão. Não há, entretanto, nenhum método ótimo para a escolha de b, sendo comum na literatura a utilização de métodos de validação cruzada ou de minimização do erro de predição.

Caso se conheça a verdadeira relação entre Y e X, ou seja, a verdadeira função $g(X)$, então se pode escolher a janela ótima tal que minimize o erro quadrático médio ou o erro quadrático médio integrado. A escolha de b depende do tamanho da amostra N e do que denominamos constante de proporcionalidade "c", então o valor ótimo de b que minimiza o erro quadrático médio integrado é dado por:

$$b_N = cN^{-1/k+4}$$

No programa EasyReg (Bierens,2004) para o caso de densidades univariadas (k=1) ou bivariadas (k=2) a regra ótima determina c=1.06 para k=1 e c=1 para k=2. Então, caso não se explicite será usado c=1, que é o valor de *default* do programa e está adequado para as regressões não-paramétricas realizadas neste trabalho.

3. BASE DE DADOS

O processo de estimação será conduzido a partir de dados *cross section* do ipeadata (IPEA) para os municípios do Nordeste no ano de 2000. As variáveis são assim definidas :

Esperança de Vida ao Nascer
Número de anos de vida que uma pessoa nascida hoje esperaria viver, se todas as taxas de mortalidade por idade se mantivessem idênticas ao que são hoje.

Domicílios com água canalizada rede geral:
Porcentagem da população que vive em domicílios com água canalizada.

Domicílios com instalações sanitárias rede geral (esgoto)
Porcentagem da população que vive em domicílios com instalações adequadas de esgoto, ou seja, com instalações sanitárias não compartilhadas com outro domicílio e com escoamento através de fossa séptica ou rede geral de esgoto.

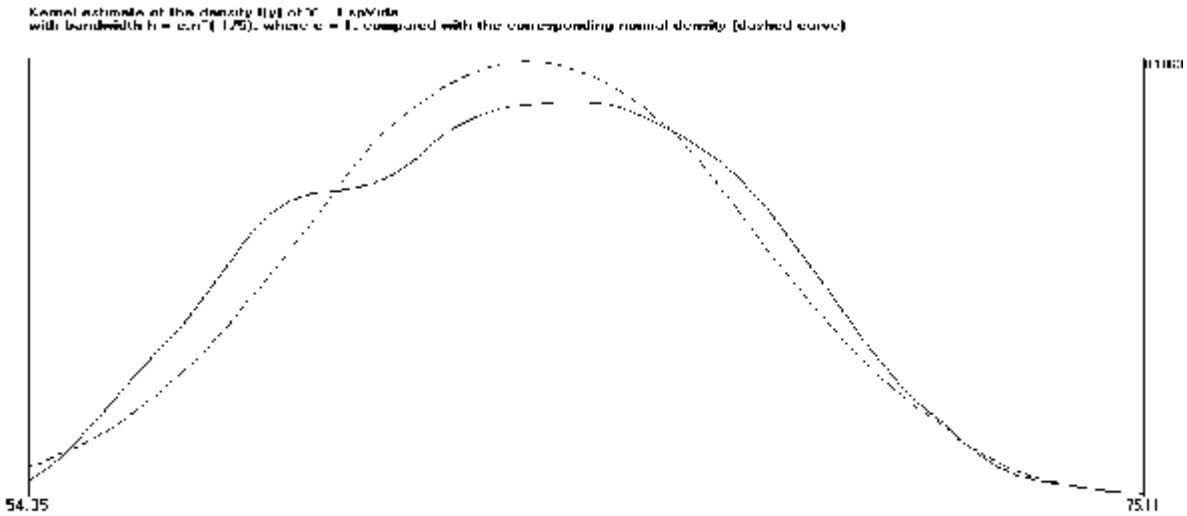
Renda per Capita
Razão entre o somatório da renda per capita de todos os indivíduos e o número total desses indivíduos. A renda per capita de cada indivíduo é definida como a razão entre a soma da renda de todos os membros da família e o número de membros da mesma. Valores expressos em reais de 1º de agosto de 2000.

4. RESULTADOS

4.1 Estimação da Densidade da Variável Esperança de Vida ao Nascer

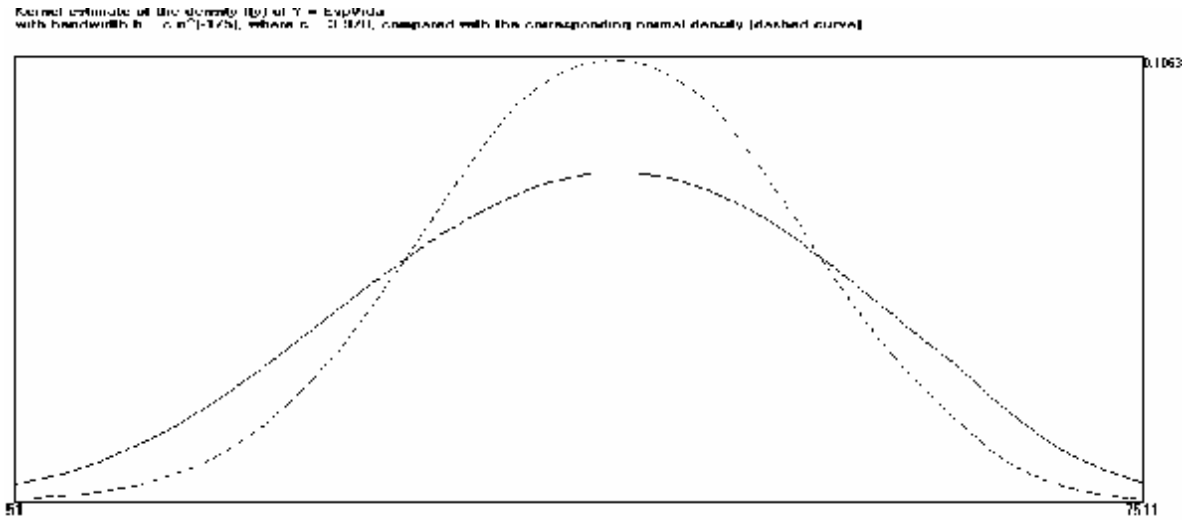
A estimação da densidade da variável Esperança de Vida ao Nascer (EVN) usando a função núcleo normal é mostrada na figura abaixo. A estimação da densidade não-paramétrica é comparada com uma densidade normal. A constante de proporcionalidade da janela escolhida foi igual a 1 e o formato da densidade sugere uma assimetria à direita, sugerindo que observações próximas a 70 anos apresentam baixa freqüência na amostra, dado que a média do nono quantil é de 67,92. A distribuição ainda apresenta dois picos de freqüência em torno de 59 anos e 63 anos.

FIGURA 1
ESTIMATIVAS DE DENSIDADE DA ESPERANÇA DE VIDA AO NASCER



Na figura 2 a densidade da EVN é estimada alterando-se a regra de seleção da janela em que a constante de proporcionalidade sobe para 3,978, segundo uma “regra de bolso”, que leva em consideração o desvio-padrão da amostra. A densidade não-paramétrica passa a apresentar um formato simétrico e em relação à normal e gera-se uma curtose, oriunda de caudas grossas. Isto sugere uma concentração de valores (em relação à distribuição normal) nos extremos de baixa e alta esperança de vida ao nascer.

FIGURA 2
ESTIMATIVAS DA DENSIDADE DA ESPERANÇA DE VIDA AO NASCER – JANELA MAIS AMPLA



4.2 Regressões não Paramétricas

a) Com um regressor

Nesta seção serão implementadas regressões não-paramétricas da variável Esperança de Vida ao Nascer (EVN) usando apenas um regressor (variável econômica ou social). A seleção da janela utilizará o método de validação cruzada. Todas as regressões seram comparadas com modelo linear estimado por Mínimos Quadrados Ordinários.

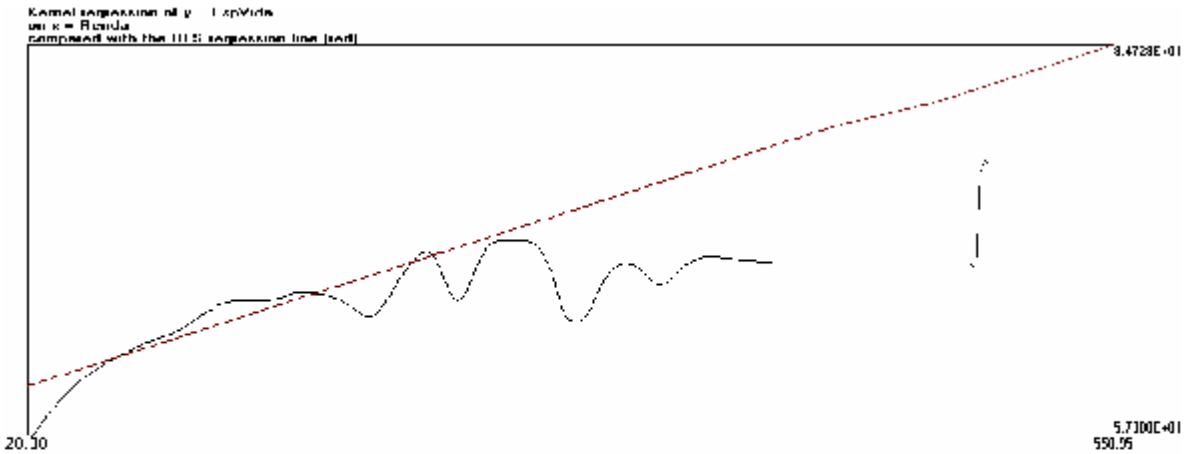
A tabela 1 resume as estatísticas de ajustamento das três regressões do modelo não- paramétrico e do modelo linear (OLS). Em termos de R-Quadrado, o modelo não-paramétrico apresentou um ajustamento superior ao de OLS, quando a variável independente é a renda per capita. Para demais as regressões, os dois modelos geraram valores próximos em termos de R-Quadrado.

Tabela 1
Municípios da Região Nordeste - Ano: 2000

Esperança de Vida ao Nascer - Variável Dependente			
Variável Independente	Renda Per-Capita	% Domícilios com Água Encanada	% Domícilios com Rede de Esgoto
1. Modelo Não Paramétrico			
Constante de proporc. da Janela	1	1	1
Desvio-Padrão	1,07E+01	1,29E+01	1,33E+01
R-Quadrado	0,238	0,0861	0,0567
2. Modelo Linear - OLS			
Desvio-Padrão	3,37E+00	3,58E+00	3,65E+00
R-Quadrado	0,1866	0,0924	0,0547

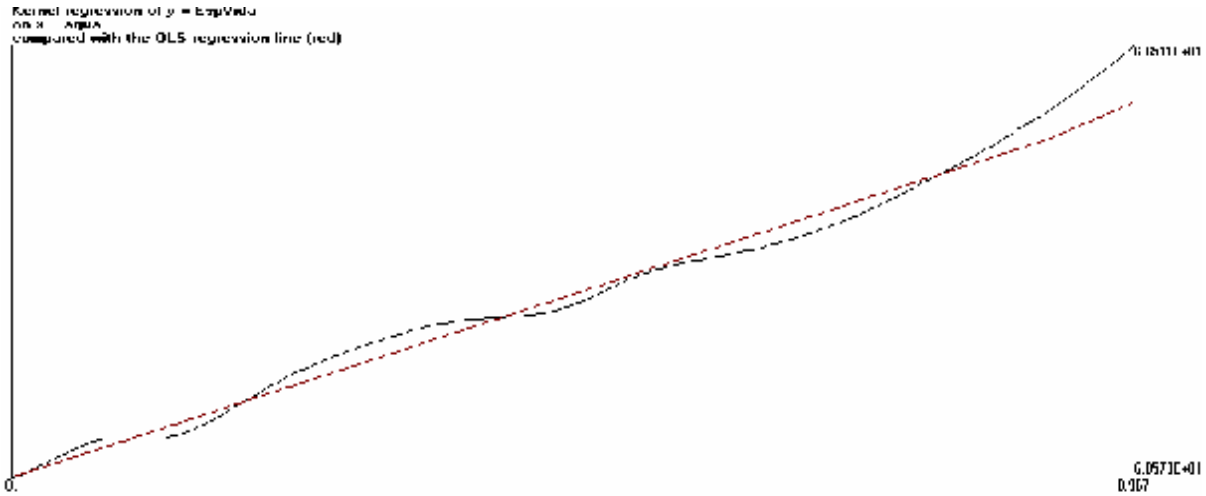
As próximas três figuras comparam graficamente as linhas de regressão do modelo não-paramétrico com o modelo linear de OLS. A regressão não-paramétrica sugere que os efeitos da renda per capita sobre a variável EVN apresentam um comportamento não-linear, sobretudo para os níveis intermediários e elevados de renda. Para níveis de renda mais baixos, a regressão não paramétrica sugere que o efeito da renda é positivo, mas marginalmente decrescente. O modelo linear aproximaria razoavelmente bem nestes casos, no entanto para níveis elevados de renda o ajustamento do modelo linear é precário.

FIGURA 3
ESPERANÇA DE VIDA AO NASCER x RENDA PER CAPITA



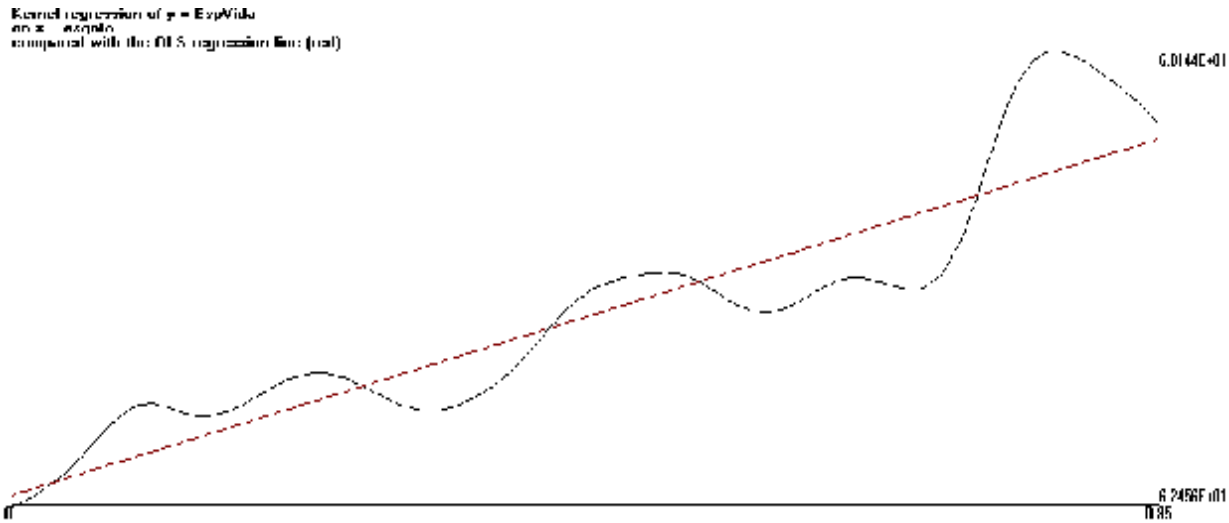
A regressão de EVN contra a proporção de domicílios com água encanada em termos não-paramétricos tem comportamento extremamente próximo daquele previsto pelo modelo linear. Ambos modelos sugerem um impacto positivo desse indicador sobre a variável EVN, ao longo de toda amostra (figura 3). Já na regressão de EVN contra a proporção de domicílios com rede de esgoto, a estimativa não-paramétrica é próxima da linearidade quando essa proporção varia de 0 a 50%, ou seja, espera-se um crescimento da EVN se há um aumento da proporção de domicílios com rede geral de esgotos. Por outro, lado, a partir de valores acima de 50% o comportamento não-paramétrico sugere uma função oscilatória para representar a relação entre as duas variáveis.

FIGURA 4
ESPERANÇA DE VIDA AO NASCER x DOMICÍLIOS COM ÀGUA



ENCANADA

FIGURA 5
ESPERANÇA DE VIDA AO NASCER x DOMICÍLIOS COM REDE GERAL DE ESGOTOS



b) Com dois regressores

À medida que se aumenta a dimensionalidade das variáveis explicativas, é necessário que se aumente exponencialmente o número de observações para que as regressões não-paramétricas desempenhem uma análise estatisticamente significante ("maldição da dimensionalidade"). No entanto, para o caso de dois regressores e com um número de observações (1787) relativamente elevado, pode-se ainda utilizar esse método de estimação. Com dois regressores, geram-se gráficos de superfícies que irão mostrar os efeitos das variáveis renda per capita e proporção de domicílios com água encanada (Água) ou rede de esgoto (Esgoto) sobre a variável EVN.

A primeira regressão reúne renda per capita e a variável Água, como variáveis explicativas para EVN. O gráfico de superfície permite algumas interpretações interessantes. Mantida fixa a variável Água, o efeito da renda per capita sobre o EVN é positivo e próximo do linear, para níveis de renda mais baixos. A partir de níveis intermediários e elevados de renda o efeito da renda é aproximada por uma função "degrau". Portanto, o efeito marginal da renda é nulo para trechos de valores intermediários e de valores elevados para variável Água.

Similarmente à variável renda per capita, a variável Água tem efeito positivo e próximo da linearidade sobre EVN nos seus níveis mais baixos. Posteriormente, a EVN irá crescer rapidamente com valores elevados de Água. Além disso, para níveis intermediários de renda o efeito marginal de Água sobre EVN é decrescente, voltando a ser crescente nos níveis mais elevados de renda per capita.

Conclusivamente, é possível aumentar a Esperança de Vida ao Nascer quando se parte de níveis muito baixos para níveis intermediários de renda per capita e da proporção de domicílios com água encanada. No entanto, os ganhos em EVN novamente serão significativos para níveis mais elevados de renda e com a quase universalização dos domicílios com acesso à água canalizada.

O valor médio da EVN cresce exponencialmente para aumento nos níveis mais baixos de renda per capita e da proporção de domicílios com rede geral de esgoto. Esse efeito é similar ao encontrado na regressão anterior. Por outro lado, os efeitos da variável Esgoto sobre EVN permanecem positivos para níveis intermediários dessa variável. No entanto, para níveis mais elevados de renda, o efeito marginal da variável esgoto é nulo. Além disso, para trechos intermediários e elevados da variável esgoto, o efeito marginal da renda sobre a EVN é nulo, apresentando degraus entre esses níveis.

Portanto, a Esperança de Vida ao Nascer nos municípios nordestinos pode crescer substancialmente com elevação da proporção de domicílios com acesso a rede de esgoto, e esse efeito somente irá cessar nos municípios com níveis mais elevados de renda per capita, onde provavelmente o ganho com EVN está associada a outras variáveis não tratadas neste trabalho, como a violência e doenças típicas dos centros urbanos maiores, que na amostra detém maior renda per capita.

FIGURA 6
ESTIMATIVAS DA MÉDIA DA ESPERANÇA DE VIDA AO NASCER COMO FUNÇÃO DA RENDA PER CAPITA E DA PROPORÇÃO DE DOMICÍLIOS COM ÁGUA ENCANADA.

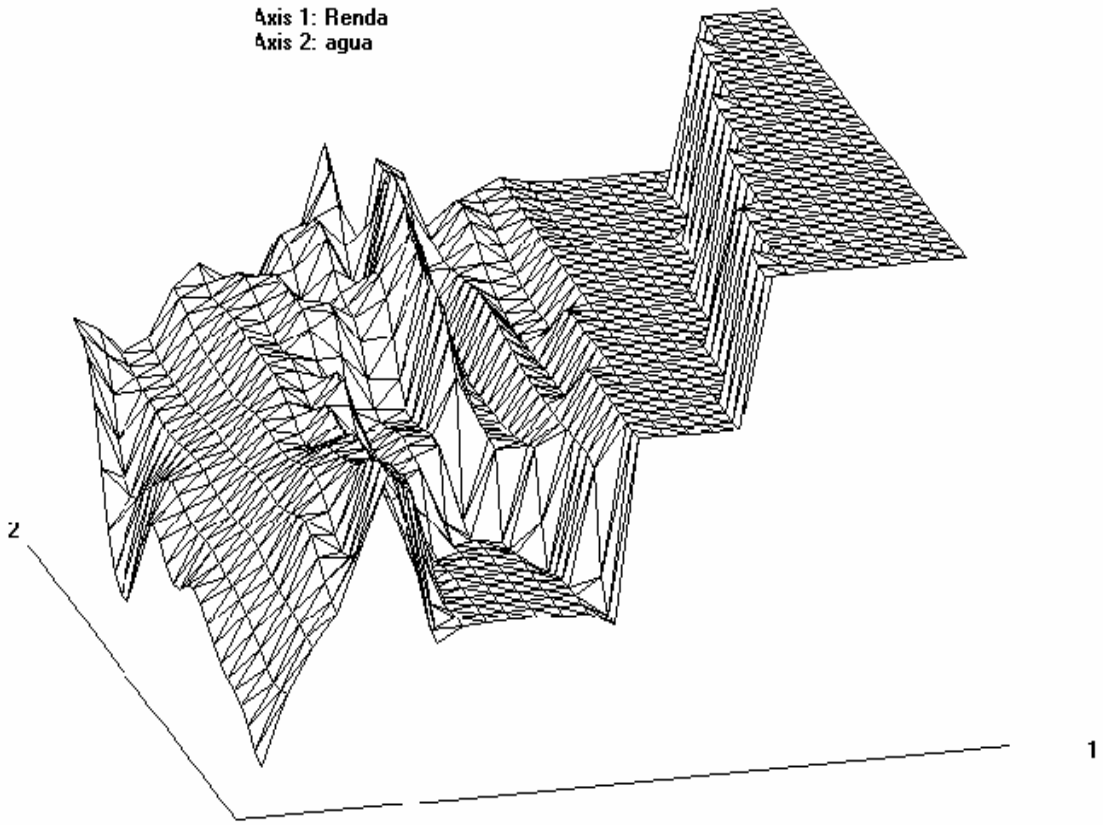
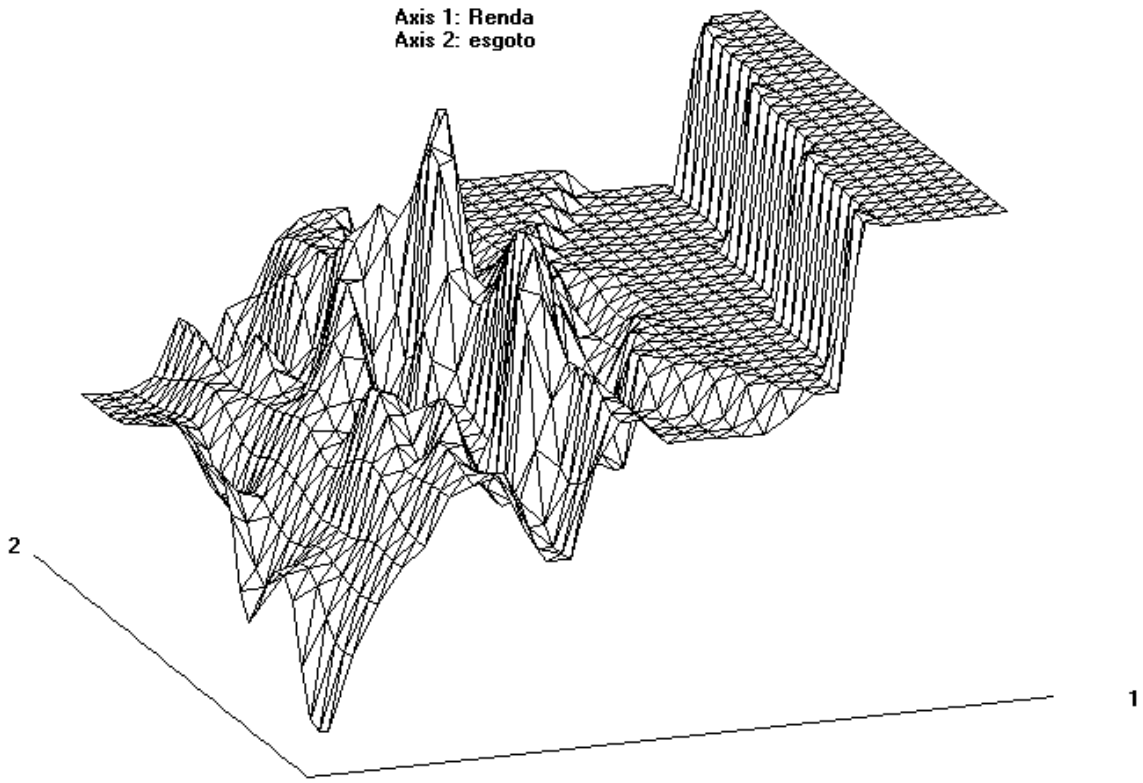


FIGURA 7
ESTIMATIVAS DA MÉDIA DA ESPERANÇA DE VIDA AO NASCER COMO FUNÇÃO DA RENDA PER CAPITA E DA PROPORÇÃO DE DOMICÍLIOS COM REDE GERAL DE ESGOTO.



5. CONCLUSÕES

O método de regressão não-paramétrica é instrumental útil para investigações empíricas sobretudo nos casos em que não se tem uma definição teórica precisa sobre as relações entre as variáveis. Portanto, nestes casos, assumir uma especificação linear e/ou hipóteses comportamentais simplificadoras sobre o comportamento das variáveis pode ser extremamente restritivo. Na econometria não-paramétrica os "dados devem falar por eles mesmos" e com isso ganha-se em termos de flexibilidade funcional e poder preditivo.

Segundo o IBGE, as maiores taxas de mortalidade infantil se concentram nos municípios da região Nordeste, o que influencia diretamente a esperança de vida ao nascer. No exercício investigado neste trabalho identifica-se que as variáveis renda per capita, proporção de domicílios com água encanada e com acesso à rede geral de esgotos apresentam uma relação positiva com a esperança de vida ao nascer em regressões com apenas um regressor para os municípios nordestinos. No entanto, com exceção da variável água encanada a relação entre as variáveis não obedece uma relação de linearidade ao longo da amostra, o que sugere um ajustamento da regressão não-

paramétrica superior aos modelos de de regressão linear de Mínimos Quadrados Ordinários. Portanto, os efeitos de aumento de renda, de acesso à rede de esgoto e de água canalizada podem produzir ganhos significativos na esperança de vida ao nascer nos municípios onde há uma carência maior desses serviços. As estimações sugerem que políticas públicas destinadas a realizar investimentos nesses municípios mais carentes devem obter um retorno social elevado em termos de expectativa de vida.

Quando utilizam-se dois regressores, os resultados continuam, de um modo geral, sugerindo que ganhos na esperança de vida ao nascer ocorrem para níveis baixos e intermediários das variáveis explicativas. Os gráficos de superfície permitem comparar os ganhos marginais de uma variável explicativa, enquanto se mantém a outra constante, o que em termos de política pública é um indicativo de como os recursos podem ser combinados de forma ótima para alcançar o objetivo de aumento de esperança de vida dos municípios nordestinos. Por exemplo, a expectativa de vida nos municípios nordestinos pode crescer substancialmente com elevação da proporção de domicílios com acesso à rede de esgoto, e esse efeito somente irá cessar nos municípios com níveis mais elevados de renda per capita.

É provável que os ganhos marginais para expectativa de vida em municípios de mais elevada renda per capita estejam associados a outras variáveis não tratadas neste trabalho, tais com a violência nos centros urbanos ou incidência de doenças típicas dessas cidades. Desse modo, as políticas públicas devem ter um foco diferenciado para grupos de municípios com níveis de renda per capita diferenciados (baixos x elevados, por exemplo), quando se trata de elevar a esperança de vida ao nascer. Portanto, essas são possíveis extensões do trabalho e que foram identificadas graças ao método de estimação não-paramétrica.

6. REFERÊNCIAS

Bierens, H. J. (1994), Topics in Advanced Econometrics, Cambridge, UK, Cambridge University Press.

Bierens, H. J. (2004), "EasyReg International", Department of Economics, Pennsylvania State University, University Park, PA

Epanechnikov, V. A. (1969) “Nonparametric Estimates of a Multivariate Probability Density.” Theory of Probability and its Applications 14 153-158.

Flôres, R. G. (2005) Notas de Aula do Curso de Econometria II.

IBGE. Censo Demográfico de 2000.

IBGE. (2005). “Pesquisa de Informações Básicas Municipais – MUNIC- Suplemento de Meio Ambiente”.

Ipea Data: “Dados Macroeconômicos e Regionais”. www.ipeadata.gov.br.

Kweon, Y. e Kockelman, K. (2004). “Nonparametric Regression Estimation of Household VMT”. Paper Submitted for Presentation & Publication at the 2004 Annual Meeting of the Transportation Research Board.

Nadaraya, E. (1964) “On Estimating Regression”. Theory of Probability and its Applications, 9:141-42.

Silverman, B. W. (1986). “Density Estimation For Statistics And Data Analysis”. Published in Monographs on Statistics and Applied Probability, London: Chapman and Hall.