UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

Regressão Semiparamétrica Utilizando Modelos de Índice Único

Autora: Márcia Helena Barbian
Orientador: Professor Dr. Flávio Augusto Ziegelmann

Porto Alegre, Dezembro de 2006.

Universidade Federal do Rio Grande do Sul
Instituto de Matemática
Departamento de Estatística

Regressão Semiparamétrica Utilizando Modelos de Índice Único

Autora: Márcia Helena Barbian

Monografia apresentada para obtenção
do grau de Bacharel em Estatística.

Banca Examinadora:
Professor Dr. Flávio Augusto Ziegelmann (orientador)
Professor Dr. Álvaro Vigo (convidado do Departamento de Estatística da UFRGS)

Porto Alegre, Dezembro de 2006.

SUMÁRIO

RESUMO

O interesse em modelagem não paramétrica e semiparamétrica tem crescido significativamente na última década. Uma importante razão para isso é o aumento da capacidade computacional, visto que os cálculos dos estimadores são complexos e necessitam de muito processamento. Baseado nisso, muitos métodos e técnicas têm sido propostos e estudados.

A regressão semiparamétrica é um método que pode ser utilizado em muitas situações. As principais motivações para o seu uso são: o desconhecimento sobre o tipo de relação entre variáveis regressoras e uma variável dependente, e quando possuímos um modelo com um número não pequeno de covariáveis.

Este trabalho irá tratar da regressão semiparamétrica, utilizando-se o Modelo de Índice Único. Alguns aspectos teóricos do estimador serão discutidos, bem como seu método de estimação. Extensivos estudos de simulação serão feitos para verificar o desempenho do modelo em diversas situações - além de compará-lo em alguns casos ao método paramétrico. Além disso, uma série financeira empírica de taxa de câmbio dólar/real será investigada e modelada semiparametricamente.

Também abordar-se-á, em uma breve introdução, a regressão não paramétrica e a estimação de funções densidade de probabilidade utilizando-se *kernels*.

AGRADECIMENTOS

Agradeço aos meus pais Deonisio e Alice, pelo apoio e incentivo, principalmente nos momentos mais difíceis que enfrentei, se fazendo sempre presentes. Além de todo amor, carinho e compreenção que me foram dados durante toda a minha vida.

Aos meus irmãos, Eduardo e Jackson, pela compreeenssão e paciência, aliás, muita paciência.

Aos professores do curso de estatística, pelo conhecimento transmitido durante a condução do curso.

Ao meu orientador, Professor Flávio, pela paciência, compreensão e muita ajuda na realização deste trabalho, sempre se mostrando disposto a ajudar.

Aos colegas de curso e amigos, em especial a Ana e Manu, que desde o ínicio do curso estiveram comigo, pela amizade, pelo apoio nos momentos difíceis (muitos momentos difíceis) e pelas muitas risadas que demos juntas.

E a todos que de uma forma direta ou indirerta contribuíram para a realização deste trabalho

1. INTRODUÇÃO

Como sabemos, os modelos de regressão são utilizados em diversas áreas com o objetivo de descrever o relacionamento entre variáveis exploratórias *Xs* e uma variável resposta *Y*. A regressão paramétrica é o método mais utilizado para definir esse comportamento. Ela supõe que a função de regressão possui alguma forma funcional específica. Além disso, faz várias suposições com relação aos dados, como normalidade do erro aleatório, por exemplo.

Todavia, tal abordagem pode ser inadequada em muitas situações práticas, visto que muitas vezes os dados não satisfazem a todas as suposições impostas pela regressão paramétrica. Uma alternativa encontrada para estes casos é utilizar a regressão não paramétrica, a qual permite maior flexibilidade na possível forma da função desconhecida, sendo que nenhuma suposição *a priori* é necessária - com exceção de um certo grau de suavidade da curva desconhecida - para fazer a análise estatística. Sua idéia principal é permitir que "os dados falem por si mesmos".

Em vista dessas características, é natural concluir que a modelagem não paramétrica é mais flexível e, portanto, pode se adaptar a uma classe grande de problemas. Mas, um preço é pago por tamanha flexibilidade, como baixas taxas de convergência, pouco poder de extrapolação e, na regressão multivariada, o *curse of dimensionality*.

O *curse of dimensionality* faz com que a regressão não paramétrica - que tem como princípio estimar curvas usando somente informação sobre os vizinhos locais – não tenha informação suficiente sobre a função a ser estimada.

Dadas as características acima, a regressão semiparamétrica foi escolhida como assunto a ser abordado nessa monografia. Apesar de não ser tão flexível como a regressão não paramétrica, tende a evitar sua inconveniência de falta de vizinhança em altas dimensões e possui mais flexibilidade que a "amarrada" regressão paramétrica.

Vários estudos têm sido realizados e muitos modelos e métodos sugeridos para descrever e estimar a função de regressão semiparametricamente. Entre estes podemos

citar: modelos aditivos, os modelos de redes neurais, modelos de coeficientes variáveis, modelos de índice único e modelos multiplicativos, dentre outros. Nesta monografia abordaremos somente a estimação utilizando o modelo de índice único.

O modelo de índice único (do inglês *single-index*) transforma um vetor de d variáveis regressoras em um índice linear e, a partir desse índice, calcula uma regressão não paramétrica univariada, evitando-se o *curse of dimensionality*. O método escolhido para a estimação desse modelo será o semiparamétrico de mínimos quadrados (SMQ), desenvolvido por Ichimura (1993), utlizando-se *kernel* gaussiano.

A ênfase dessa monografia é a aplicação do modelo de índice único em vários problemas estatísticos. Para isso serão simuladas regressões de diferentes formas para avaliar o desempenho desse estimador em diferentes situações, além de aplicações a uma série financeira real. O programa utilizado para a execução das análises foi o software livre **R** versão 2.4.0. Também serão abordadas propriedades teóricas do modelo bem como seu método de estimação. O restante da monografia é organizado da seguinte maneira, além desta introdução que constitui o capítulo 1:

Capítulo 2: Falará sobre a estimação de funções densidade de probabilidade não parametricamente utilizando-se *kernels*, bem como suas propriedades e técnicas para a escolha do parâmetro de suavização. Uma breve introdução sobre estimação de regressão não paramétrica também será vista, com enfoque sobre a regressão polinomial local. Por último, será apresentada uma breve discussão sobre o conhecido problema enfrentado pela regressão não paramétrica multidimensional, o *curse of dimensionality*.

Capítulo 3: Abordará o modelo de índice único e falará sobre algumas de suas propriedades. O método de estimação do modelo utilizando MQS será explicitado, e uma modificação desse método, utilizada quando os erros são heteroscedásticos, será citada.

Capítulo 4: Capítulo reservado a simulações, verificando a adequabilidade da regressão utilizando modelos de índice único em diferentes casos. Além disso, uma breve comparação entre o método e a modelagem paramétrica será apresentada.

Capítulo 5: Utilização da modelagem do índice único em uma série financeira empírica.

Capítulo 6: Finaliza este trabalho, comentando sobre os resultados encontrados e sobre o desempenho do estimador. Também serão feitas as considerações finais.

2. REGRESSÃO NÃO PARAMÉTRICA UTILIZANDO *KERNEL*

Os modelos de regressão são largamente utilizados em diversas áreas do conhecimento, tais como: economia, computação, administração, engenharias, biologia, agronomia, saúde, sociologia, etc. Seu principal objetivo é descrever o relacionamento entre uma ou várias variáveis exploratórias *Xs* e uma variável resposta *Y*. A forma da função de regressão pode ser útil para estimar, por exemplo, os valores esperados de *Y* dado certos valores de *X*, ou se existe algum tipo de dependência especial entre as duas variáveis.

Se n observações $\{(X_i, Y_i)\}_{i=1}^{n}$ forem coletadas, a regressão pode ser modelada como

$$Y_i = m(x_i) + \varepsilon_i, \qquad i = 1,...,n, \tag{2.1}$$

sendo m a função de regressão desconhecida e ε_i o erro aleatório da observação i, com $E[\varepsilon_i / x_i] = 0$. O objetivo da análise de regressão é produzir uma boa estimativa da função desconhecida m, ou seja, estimar a dependência média de *Y* em relação a *X*.

O método mais freqüentemente usado para estimar a função média é a regressão paramétrica, que assume que a curva m possui alguma forma funcional específica; por exemplo, uma reta com o intercepto e o nível de inclinação desconhecidos, além de uma família específica para a distribuição de probabilidades dos erros aleatórios. Os modelos paramétricos são completamente determinados por um número finito de parâmetros. O ajuste do modelo é facilmente interpretável e estimado com acuracidade. Mas, se as suposições do modelo são violadas as estimativas tornam-se inconsistentes.

Todavia, se não temos conhecimento sobre o comportamento das variáveis, isto é, se elas respeitam ou não as suposições impostas pela regressão paramétrica, a abordagem não paramétrica pode ser uma escolha alternativa.

A sua principal idéia é permitir que "os dados falem por si mesmos", o que indica que nenhuma suposição *a priori* é necessária para fazer a análise estatística, a não ser que a curva desconhecida tenha certo grau de suavidade. Distintas metodologias podem ser aplicadas para estimar a estrutura subjacente dos dados. Dentre elas estão: suavização via *kernel*, *splines* e *wavelets*, por exemplo. Nesta monografia utilizaremos *kernels* como o método padrão de estimação. Para maiores detalhes sobre suavização utilizando *splines* ver Green e Silverman (1994), já para *wavelets* Efromovich (1999).

Antes de começar a tratar sobre regressão não paramétrica, seria interessante entender o processo de estimação de densidades. Logo, o restante desse capítulo será focado na suavização utilizando *kernel*, que é apresentada no contexto de estimação de funções densidade univariadas e depois estendida para a análise de regressão.

2.1 Estimação de Densidades Não Parametricamente

Considere uma variável aleatória contínua e sua função densidade de probabilidade (fdp). A fdp descreve a distribuição da variável aleatória contínua e, através dela, calculamos não somente a média e a variância, mas também a probabilidade de ocorrência de um valor em um determinado intervalo.

Além da estimação da fdp ser muito útil como indicadora do comportamento de uma dada variável, ela também pode ajudar no entendimento sobre a estimação da regressão não paramétrica.

2.1.1 Histograma

O histograma é um método muito utilizado para indicar a forma da fdp de uma variável. Através dele é possível, por exemplo, sugerir se a distribuição de dada variável é unimodal ou bimodal, simétrica ou assimétrica, etc.

O primeiro passo para a construção de um histograma é dividir o suporte da variável em intervalos. Cada observação é colocada no intervalo apropriado. Feito isso,

calcula-se a proporção da amostra contida naquele intervalo e divide-se pela largura do intervalo. O histograma pode ser escrito como:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} I(x - \tilde{x}_i; h/2), \tag{2.2}$$

onde n é o tamanho amostral, h é a largura do intervalo, $\tilde{x}_i$ é o centro do intervalo onde x_i está localizado, e $I(x - \tilde{x}_i; h/2)$ é a função indicadora, isto é, assume valor 1 se $x_i - \tilde{x}_i$ está contido no intervalo $(-h/2; h/2)$ e 0 caso contrário.

Apesar de ser uma técnica fácil de implementar, indicando de maneira razoável o comportamento de uma dada variável, o histograma possui algumas falhas como:

1. Há perda de informação, pois a observação x_i é substituída pelo ponto central de seu intervalo;

2. Uma variável contínua é discretizada, pois o estimador não é suavizado;

3. O comportamento do estimador depende do valor inicial $\tilde{x}_1$ e do tamanho de seus intervalos.

Para maiores detalhes sobre o histograma ver Härdle *et al* (2004).

2.1.2 Estimação de Densidades Utilizando *Kernel*

O método possui a mesma lógica do histograma, mas em vez de uma "caixa" ser usada para a construção dos blocos, utiliza-se uma função de *kernel* suavizada. O estimador de *kernel* assume a seguinte forma:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\{(x - x_i)/h\}, \tag{2.3}$$

onde K é uma função densidade de probabilidade, chamada de função *kernel*, cuja variância é controlada pelo parâmetro h. Geralmente K é uma função simétrica com

média zero e variância um, tal que $\int K(x)dx = 1$. Por simplicidade teórica e de aplicação, comumente K é uma distribuição normal padronizada. Utilizaremos o *kernel* normal como padrão nesta monografia, que em comparação com o *kernel* ótimo perde pouca eficiência. Para maiores detalhes sobre diferentes *kernels* e suas propriedades ver Härdle (1994).

Antes de discutirmos sobre o funcionamento desse estimador é essencial conhecer alguns de seus componentes. Considere a expressão:

$$\hat{f}_h(g_t) = \frac{1}{nh}\sum_{i=1}^{n} K\{(g_t - x_i)/h\}$$

em que

$x_i \rightarrow$ é o i-ésimo valor observado da variável i;

$h \rightarrow$ janela ou parâmetro de suavização;

$n \rightarrow$ tamanho da amostra;

$g_t \rightarrow$ t-ésimo ponto no qual a função será estimada (grade t).

O cálculo da densidade estimada pode ser indicado como:

$$\hat{f}(g_1) = \frac{1}{nh}\left[\phi\left(\frac{g_1 - x_1}{h}\right) + \phi\left(\frac{g_1 - x_2}{h}\right) + \ldots + \phi\left(\frac{g_1 - x_n}{h}\right)\right]$$

$$\hat{f}(g_2) = \frac{1}{nh}\left[\phi\left(\frac{g_2 - x_1}{h}\right) + \phi\left(\frac{g_2 - x_2}{h}\right) + \ldots + \phi\left(\frac{g_2 - x_n}{h}\right)\right]$$

$$\vdots$$

$$\hat{f}(g_t) = \frac{1}{nh}\left[\phi\left(\frac{g_t - x_1}{h}\right) + \phi\left(\frac{g_t - x_2}{h}\right) + \ldots + \phi\left(\frac{g_t - x_n}{h}\right)\right].$$

Note que quanto mais afastada a observação amostral estiver do particular ponto da grade, onde se deseja estimar a fdp, menor será o seu peso na estimativa daquele ponto e, conseqüentemente, menos suavizada será a estimativa. Por outro lado, quanto maior for o valor de h, maior será a semelhança de pesos entre as unidades amostrais na estimação da grade, o que faz com que a estimativa tenha valores muito parecidos independentemente da localização. Logo, a função será muito suavizada.

2.1.2.1 Propriedades Assintóticas do Estimador de Densidades *Kernel*

Para verificarmos a eficiência da densidade estimada utilizando-se *kernel* é necessário especificar um critério de erro apropriado. Um critério muito utilizado da estatística paramétrica clássica é o erro quadrático médio (EQM), que mede a proximidade de um estimador $\hat{\theta}$ de seu verdadeiro parâmetro θ.

$$EQM(\hat{\theta}) = E(\hat{\theta} - \theta)^2 \tag{2.4}$$

$$= Var(\hat{\theta}) + (E\hat{\theta} - \theta)^2 \tag{2.5}$$

onde (2.5) decompõem o EQM de $\hat{\theta}$ em duas partes: a primeira é sua variância, e a segunda seu vício ao quadrado. Se estimarmos f sobre os reais, pode-se usar o erro quadrático médio integrado (EQMI), que no caso unidimensional é

$$EQMI(\hat{f}_h) = E\left\{\int [\hat{f}(x) - f(x)]^2 dx\right\} \tag{2.6}$$

$$= \int \left[E\left\{\hat{f}(x)\right\} - f(x)\right]^2 dx + \int Var\left\{\hat{f}(x)\right\} dx. \tag{2.7}$$

Após alguns cálculos podemos chegar a

$$EQMI(\hat{f}_h) = \frac{1}{nh}\|K\|_2^2 + \frac{h^4}{4}\{\mu_2(K)\}^2 \|f''\|_2^2 + o\left(\frac{1}{nh}\right) + oh^4, \tag{2.8}$$

onde $\|K\|_2^2 = \int K(z)^2 dz$ e $\mu_r(K) = \int z^r K(z) dz$. Ignorando os termos de alta ordem, uma fórmula aproximada via expansão de Taylor para (2.7) é o erro quadrático médio integrado assintótico (EQMIA), dado por

$$EQMIA(\hat{f}_h) = \frac{1}{nh}\|K\|_2^2 + \frac{h^4}{4}\{\mu_2(K)\}^2\|f''\|_2^2. \quad (2.9)$$

O EQMIA é muito informativo, pois podemos analisar como a janela se relaciona com o componente do vício e da variância. Quando h aumenta, aumenta o vício, enquanto a variância diminui. Quando h diminui acontece o efeito contrário.

2.1.2.2 Escolhendo o Parâmetro de Suavização

Para realizarmos a estimação é necessário escolher o valor para o parâmetro de suavização h. Como visto na seção anterior, a janela é o fator fundamental na estimação da densidade. É através dela que se decide como a estimativa será: mais suavizada, menos suavizada, de maior variância ou de menor variância.

Para visualizarmos de maneira mais clara o efeito que o valor da janela tem sobre a estimativa, simulamos 250 observações da seguinte distribuição: $Y \sim N(12,2)$. A Figura 2.1 mostra as estimativas utilizando *kernels* de diferentes janelas.

PLUG-IN

Otimizar a expressão acima do EQMIA em relação à h, nos permite obter uma expressão aproximada para a janela ótima:

$$h_{opt} = \left(\frac{\|K\|_2^2}{\|f''\|_2^2 \{\mu_2(K)\}^2 n} \right)^{1/5}. \quad (2.10)$$

O problema da estimação do h ótimo ainda não foi resolvido completamente, pois h_{opt} ainda depende de $\|f''\|_2^2$ que é um valor desconhecido. O método de *plug-in* tem como princípio substituir os termos desconhecidos da expressão acima por suas estimativas, o que permite calcular o h_{opt}.

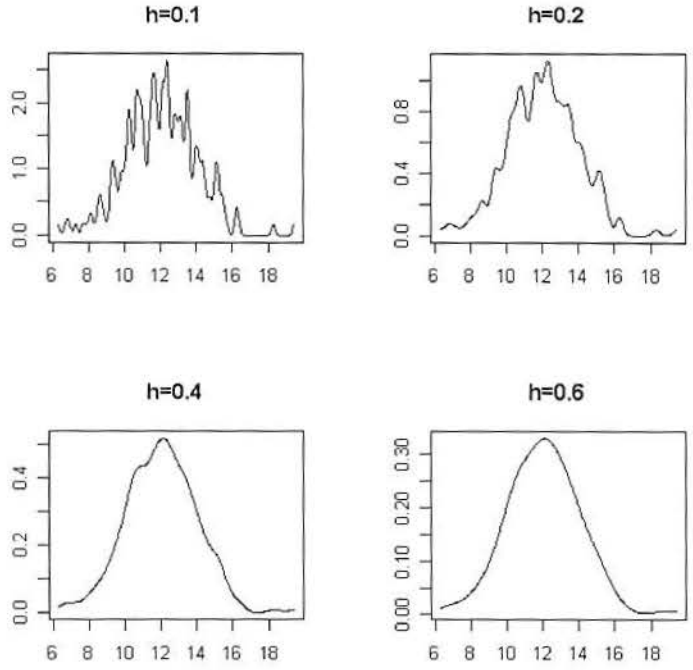

Figura 2.1 - Valores estimados da função $N(12,2)$ utilizando kernels de diferentes janelas.

Método de Silverman (*Rule of Thumb)*: Suponha que nós sabemos ou assumimos que a densidade desconhecida f pertence à família de distribuições normais com média μ e variância σ^2. Considerando o *kernel* normal nós teremos a seguinte janela:

$$\hat{h}_{rot} = \left(\frac{4\hat{\sigma}^5}{3n} \right)^{1/5}. \qquad (2.11)$$

O que nós buscamos, assumindo-se normalidade, é uma maior maleabilidade. Na realidade não sabemos se X é normalmente distribuída. Se for, então $\hat{h}_{rot}$ em (2.11) dará a janela ótima. No caso em que X não tiver uma distribuição muito diferente da distribuição normal, o $\hat{h}_{rot}$ em (2.11) fornecerá uma janela muito próxima da ótima. Todavia, se a verdadeira densidade diferenciar-se substancialmente da curva de uma distribuição normal (por ser multimodal, por exemplo), a estimativa para o $\hat{h}_{opt}$ poderá

estar consideravelmente equivocada. Para maiores detalhes sobre este estimador ver Härdle *et al* (2004).

Método de Park & Marron: tem como proposta estimar $\|f''\|_2^2$ usando uma estimativa não paramétrica para f e substituir a segunda derivada desta estimativa na função (2.10), resultando em

$$\hat{h}_{pm} = \left(\frac{\|K\|_2^2}{\|\hat{f}''\|_2^2 \{\mu_2(K)\}^2 n} \right)^{1/5} . \tag{2.12}$$

Park e Marron (1990) mostraram que $\hat{h}_{pm}$ possui um ótimo nível de convergência. A desvantagem é que para pequenos valores da janela, o estimador de $\|\hat{f}''\|_2^2$ pode resultar em valores negativos.

CROSS-VALIDATION

No contexto de estimação de funções densidade de probabilidade, Rudemo (1982) e Bowman (1984) aplicaram validação cruzada de mínimos quadrados (VCMQ) para o problema de escolha do parâmetro de suavização. Considere o EQMI de $\hat{f}_h(x)$ dado pela expressão (2.6). Através de um pequeno desenvolvimento a função pode ser reescrita como

$$EQMI\{\hat{f}_h(\cdot)\} = E\int \hat{f}_h(x)^2 dx - 2E\int \hat{f}_h(x) f(x) dx + \int f(x)^2 dx , \tag{2.13}$$

ainda obtendo

$$EQMI\{\hat{f}_h(\cdot)\} - \int f(x)^2 dx = E\int \hat{f}_h(x)^2 dx - 2E\int \hat{f}_h(x) f(x) dx . \tag{2.14}$$

Considerando somente os termos que dependem da janela, eles podem ser estimados pela seguinte expressão

$$VCMQ(h) = \frac{1}{n}\sum_{i=1}^{n} \int \hat{f}_{-i,h}(x)^2 \, dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i,h}(x_i), \tag{2.15}$$

onde

$$\hat{f}_{-i,h}(x) = (1-n)^{-1} \sum_{u \neq i}^{n} K_h(x - x_u) \tag{2.16}$$

é a densidade estimada para o ponto x_i baseada na amostra sem a observação x_i. A idéia da validação cruzada é usar a informação sobre parte da amostra para estimar a outra parte, no sentido de previsão fora da amostra. Minimizar a função (2.15) em relação à h é o mesmo que minimizar (2.16). Portanto, o h estimado de $\hat{f}_{-i,h}(x)$ será a estimativa do parâmetro de suavização ótimo. O princípio do método de validação cruzada pode ser aplicado a outros estimadores não se restringindo ao método de *kernel*. Para maiores detalhes sobre validação cruzada ver Härdle (1994).

2.2 Estimação Não Paramétrica de Curvas de Regressão

A idéia principal da estimação da função de regressão não paramétrica utilizando kernel é ajustar uma curva de regressão local em cada ponto x, onde as observações mais próximas do ponto onde se deseja estimar a curva receberão um peso maior e as mais afastadas um peso menor. O parâmetro de suavização determinará quão local será a estimação ou, em outras palavras, o quanto as observações podem estar afastadas de x e ainda contribuir para a estimativa.

2.2.1 Regressão Polinomial Local

Os estimadores polinomiais locais estimam a função de regressão em um ponto particular através do ajuste local de um polinômio de ordem p. Stone (1977) e Cleveland (1979) foram os primeiros a estudar sistematicamente estes estimadores.

Considere $\{(X_i, Y_i)\}_{i=1}^{n}$ uma amostra aleatória de tamanho n, sendo $m(x) = E(Y / X = x)$ e $\varepsilon_1, ..., \varepsilon_n$ os erros aleatórios. Nós podemos formular o seguinte modelo:

$$Y_i = m(X_i) + \varepsilon_i, \tag{2.17}$$

onde $E[\varepsilon_i / X_i] = 0$.

Além disso, sejam $m'(z), m''(z), ..., m^{(p)}(z)$ as derivadas da função de regressão $m(z)$ e x vizinho de x_0. Através da expansão da série de Taylor, $m(x)$ pode ser escrita como

$$m(x) \approx m(x_0) + m'(x_0)(x - x_0) + \frac{m''(x_0)}{2}(x - x_0)^2 + ... + \frac{m^{(p)}(x_0)}{p!}(x - x_0)^p. \tag{2.18}$$

A função acima indica que podemos aproximar a função de regressão desconhecida $m(X_i)$ usando (2.18). Então, se nós considerarmos um ponto de interesse x no domínio da variável aleatória X, pode-se definir o estimador polinomial local de $m(x)$ como $\hat{m}_p(x) = \hat{\beta}_0$, com $\hat{\beta}_0$ dado pela solução do seguinte problema de mínimos quadrados:

$$(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_n) = \arg \min_{\beta_0, \beta_1, ..., \beta_p} \sum_{i=1}^{n} \{Y_i - \sum_{u=0}^{p} \beta_i (X_i - x)^u\}^2 K_h(X_i - x). \tag{2.19}$$

É importante notar que, em contraste com o estimador de mínimos quadrados paramétrico, este estimador varia de acordo com x. Também note que a estimação polinomial local fornece como subproduto estimativas das derivadas da função $m(x)$.

ESTIMADORES DE NADARAYA-WATSON E LINEAR LOCAL

Os dois estimadores apresentados nesta seção são casos especiais do estimador polinomial local. Uma aproximação simples é construir o estimador da média local, também conhecido como estimador de Nadaraya-Watson, dado a seguir:

$$\hat{m}_{nw}(x) = \frac{\sum_{i=1}^{n} K_h(X_i - x)Y_i}{\sum_{i=1}^{n} K_h(X_i - x)}. \tag{2.20}$$

Uma desvantagem desse estimador é que ele possui uma desagradável forma de vício, particularmente em regiões onde as derivadas da função de regressão são elevadas. Wand e Jones (1995) fornecem maiores discussões sobre essas características.

Do ponto de vista da função aproximada, o estimador de Nadaraya-Watson usa aproximações locais constantes. Todavia, se incluirmos um termo linear da expansão da série de Taylor, deriva-se outro caso muito especial de estimador polinomial local. Ele é o estimador linear local, que corresponde a $p=1$ em (2.19). A fórmula para esse estimador é dada por

$$\hat{m}_l(x) = \frac{S_{n,2}(x)T_{n,0}(x) - S_{n,1}(x)T_{n,1}(x)}{S_{n,2}(x)S_{n,0}(x) - S_{n,1}^2(x)}, \tag{2.21}$$

onde

$$S_{n,j}(x) = \sum_{i=1}^{n} K_h(X_i - x)(X_i - x)^j \tag{2.22}$$

e

$$T_{n,l}(x) = \sum_{i=1}^{n} K_h(X_i - x)(X_i - x)^l Y_i. \tag{2.23}$$

A principal vantagem do estimador linear local (e polinomial em geral) é sua adaptação a várias formas de densidades. Além disso, o estimador linear local possui um melhor desempenho nos limites da função estimada que o estimador de Nadaraya-Watson. Para maiores detalhes ver Fan (1992) e Härdle *et al* (2004). O estimador linear local, com *kernel* normal, será o método padrão adotado nessa monografia.

2.2.2 Escolha da Janela na Regressão Não Paramétrica

Para obter-se o estimador polinomial local, é necessário escolher a ordem p, a janela h e o *kernel* K. Como usaremos p=1 e o *kernel* normal como padrão, será a janela h que indicará a principal regra em determinar o ajuste do modelo.

CROSS-VALIDATION

Como já mencionado anteriormente, a validação cruzada fornece um método útil de escolha do parâmetro de suavização. A validação cruzada no contexto de regressão é definida como

$$VC(h) = n^{-1}\sum_{i=1}^{n}\left\{Y_i - \hat{m}_{h,-i}(X_i)\right\}^2, \tag{2.24}$$

onde $\hat{m}_{h,-i}(X_i)$ é o estimador de $m(x) = E(Y/X = x)$ com a observação (X_i, Y_i) omitida. Pode-se mostrar que

$$E\{VC(h)\} = n^{-1}\sum E\left\{\hat{m}_{h,i}(X_i) - m(X_i)\right\}^2 + n^{-1}\sum \sigma^2(X_i). \tag{2.25}$$

Como o segundo termo $n^{-1}\sum\sigma^2(X_i)$ não depende de h, (2.24) fornece um simples estimador do EQMI(h) definido em (2.13). Portanto, o estimador de h utilizando validação cruzada, denotado por h_{VC}, é aquele que minimiza a expressão (2.24). Para maiores detalhes ver Allen (1974), Stone (1974) e Bowman (1997).

2.2.3 *Curse of Dimensionality*

Teoricamente a regressão não paramétrica pode ser utilizada com preditores multidimensionais. Segundo Härdle (1994), o procedimento da média local ainda será assintoticamente consistente sobre a superfície da regressão. Todavia, alguns problemas podem surgir quando estimamos uma regressão não paramétrica com diversas covariáveis.

Em primeiro lugar, a função de regressão $m(x)$ está em uma superfície multidimensional e sua forma não pode ser mostrada em dimensões superiores a dois. Logo, a curva estimada não fornecerá uma descrição geométrica do relacionamento entre X e Y. Segundo, o elemento básico da suavização não paramétrica, mediante o calculo da média sobre os vizinhos locais, freqüentemente será aplicado a um conjunto de pontos relativamente ínfimo porque, até mesmo em amostras de tamanho superior a 1000, as observações poderão estar esparsamente distribuídas no espaço multidimensional. O seguinte exemplo deve ilustrar de maneira mais clara a lógica por trás do *curse of dimensionality.*

Pelo primeiro gráfico (Figura 2.2) podemos perceber que os pontos estão muito próximos e a distância entre eles é pequena, não sendo maior que três. Agora imagine que consideraremos os mesmos pontos, só que em vez de um espaço bidimensional acrescentamos mais uma variável no modelo, transformando o espaço em tridimensional (Figura 2.3). Os pontos que anteriormente estavam muito próximos agora estão separados por grandes distâncias. Logo, estimar uma regressão utilizando os dados no segundo modelo seria mais difícil, pois não há vizinhos locais e, sem estes, não há informação sobre a função a ser estimada. Já a Figura 2.4 mostra, no espaço tridimensional, como os pontos do primeiro gráfico estariam distribuídos se todos tivessem o mesmo valor para a terceira variável.

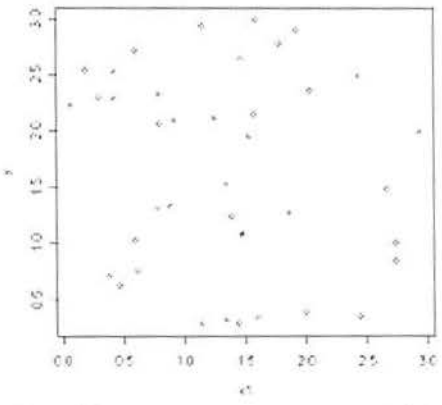

Figura 2.2 – Regressão utilizando somente uma variável regressora $Y = X_1$.

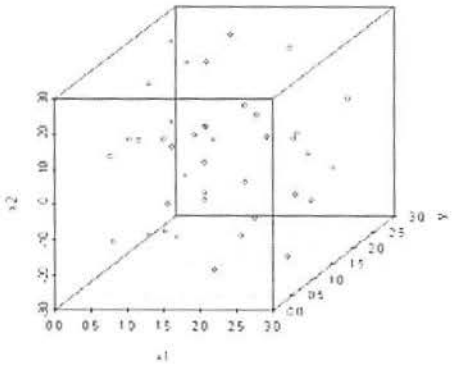

Figura 2.3 – Regressão utilizando duas variáveis regressoras $Y = X_1 + X_2$.

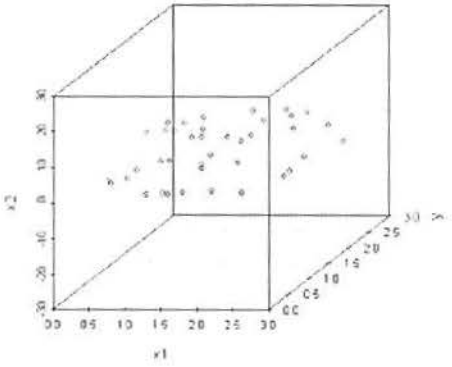

Figura 2.4 – Regressão utilizando uma variável regressora $Y = X_1$, mostrada no espaço tridimensional.

3. REGRESSÃO SEMIPARAMÉTRICA

Métodos de estimação não paramétrica são muito flexíveis, visto que podem ser aplicados a uma gama ampla de dados e não fazem nenhuma suposição em relação ao tipo de relação entre as variáveis. Entretanto, suas propriedades estatísticas deterioram quando o número de variáveis exploratórias no modelo aumenta. Este problema é conhecido como *curse of dimensionality* e já foi tratado no capítulo anterior.

A pergunta a fazer é a seguinte: como se pode resolver esse problema, visto que a modelagem paramétrica muitas vezes não é adequada e a as técnicas puramente não paramétricas enfrentam dificuldades. Como resposta surge a regressão semiparamétrica, que une a flexibilidade do modelo não paramétrico com as altas taxas de convergência da estimação paramétrica.

Vários estudos têm sido realizados e muitos modelos e métodos sugeridos para modelar e estimar a função de regressão semiparametricamente. Entre os modelos semiparamétricos podemos citar os seguintes: modelos aditivos, modelos de redes neurais, modelos de coeficientes variáveis, modelos de índice único e modelos multiplicativos, dentre outros. Para maiores detalhes sobre modelos aditivos, modelos de coeficientes variáveis e modelos multiplicativos ver Ziegelmann (2002). Em nosso estudo será considerado o modelo de índice único.

3.1 O Modelo de Índice Único

O modelo de índice único (do inglês *single-index*) é um modelo semiparamétrico de idéia bastante intuitiva. Ele transforma um vetor de d variáveis regressoras em um índice linear e a partir desse índice estima uma regressão não paramétrica univariada. Logo uma regressão multidimensional é estimada unidimensionalmente na sua parte não paramétrica.

Tal modelo é muito atraente, pois a função de regressão pode envolver múltiplas variáveis exploratórias (através do índice linear), enquanto retém a flexibilidade não paramétrica (através da função g) sem sucumbir ao *curse of dimensionality*.

O modelo de índice único é dado por

$$Y_i = g(\alpha_0 + \beta' X_i) + \varepsilon_i \tag{3.1}$$

onde g é uma função univariada com certo grau de suavidade, α_0 é uma constante, $\beta = (\beta_1, \beta_2, ..., \beta_d)$ é um vetor de d constantes que indica a direção de suavização no espaço $\Re^d$, $X_1, ..., X_n$ representam vetores aleatórios d-dimensionais e $\varepsilon_1, ..., \varepsilon_n$ são variáveis aleatórias com média zero e variância finita. O modelo supõe ainda que (ε_i / X_i) são independentes e identicamente distribuídas, para $i = 1, ..., n$.

O primeiro ponto a considerar é a identificabilidade do modelo (3.1) que é provada por Ichimura (1993). O problema de identificabilidade surge quando o modelo estatístico não está completamente especificado. O modelo é identificável para um dado conjunto de observações se, para aquele conjunto, pode-se encontrar uma única estimativa do parâmetro desconhecido que minimize o critério de estimação escolhido. Horowitz (1998) apresenta uma ótima discussão sobre esses aspectos, para modelos de índice único.

As condições impostas a este modelo para assegurar sua identificabilidade são as seguintes:

(1) $g(\cdot)$ é diferenciável e não constante no suporte de $\beta' X$;

(2) X_i são variáveis aleatórias contínuas que possuem uma função densidade de probabilidade conjunta;

(3) O suporte de X_i não está contido em nenhum subespaço de $\Re^d$;

(4) $\alpha_0 = 0$ e $\beta_1 = 0$.

3.2 Estimando o Modelo de Índice Único

A estimação do modelo é basicamente dividida em duas etapas: a estimação do vetor β (índice) e a estimação da função $g(\cdot)$. Dentre os métodos propostos para estimar o modelo de índice único, podemos citar: estimação da derivada média, investigado por Härdle e Stoker (1989); regressão inversa fatiada, proposto por Li (1991); estimação da variância média mínima, proposto por Xia *et al* (2002) e Xia e Härdle (2006); e semiparamétrico de mínimos quadrados (SMQ), desenvolvido por Ichimura (1993), o qual será o método abordado nesta monografia.

O procedimento de estimação utilizando SMQ pode ser esquematizado da seguinte maneira:

(1° Passo) Estimar β por $\hat{\beta}$;

(2° Passo) Calcular o valor do índice $z_i = \hat{\beta}' X_i$;

(3° Passo) Estimar a função $g(\cdot)$ usando um método não paramétrico univariado da regressão de Y por z.

O que torna a estimação do modelo não trivial é que para estimar β é necessário utilizar a função $g(\cdot)$ e, pelo que vimos acima, a função $g(\cdot)$ é uma função de β.

Para solucionar o problema de estimação suponha que dispomos de certo vetor β. Assim, através dele obtemos o estimador linear local (2.21) de $g(z)$ via a minimização do seguinte problema de mínimos quadrados com respeito a a e b:

$$\sum_{i=1}^{n} \{Y_i - [a + b(\beta' X_i - z)]\}^2 K_h(\beta' X_i - z) \tag{3.2}$$

onde $K_h(\cdot)$ é a função *kernel* e h é a janela.

Segundo Ichimura (1993) há pelo menos duas vantagens em utilizar um estimador via *kernel*: a primeira é que a função objetivo é diferenciável com

probabilidade aproximada de 1 se uma função *kernel* diferenciável é usada, enquanto a segunda é que, quando um estimador via *kernel* é utilizado, a derivada da função objetivo converge para a da distribuição limite.

Voltando ao problema de estimação, o estimador linear local de $g(z)$ é dado por

$$\hat{g}(z) = \hat{a} . \tag{3.3}$$

O método utilizado para obter as estimativas para o vetor de suavização condicional desconhecido β será baseado na minimização da seguinte função de validação cruzada:

$$S_{T_C}(\beta, h) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{g}_{-i}(\beta' X_i)\}^2 \tag{3.4}$$

onde $\hat{g}_{-i}(\beta' X_i)$ é dado por (3.3) com a observação (X_i, Y_i) removida. O vetor β e a janela h são estimados conjuntamente via a minimização de $S_{T_C}(\beta, h)$.

Não é claro, *a priori*, se o mesmo h pode ser usado para construir bons estimadores tanto de β quanto de $g(\cdot)$. Hall (1989) sugere que duas diferentes janelas possam ser necessárias – a primeira para construir um estimador preliminar de $g(\cdot)$ para que β possa ser estimado e, a segunda, para construir um estimador final de $g(\cdot)$. Já Härdle, Hall e Ichimura (1993) demonstraram que no contexto de SMQ, o mesmo h pode ser utilizado para a estimação ótima tanto de β quanto de $g(\cdot)$.

Seguiremos esta abordagem para a escolha da janela. Logo, depois de estimados, $\hat{\beta}$ e $\hat{h}$ serão utilizados na função (3.3) que resultará no estimador final de $g(\cdot)$.

Ichimura (1993) provou que, sob algumas condições de regularidade, a estimativa de β obtida desse procedimento é assintoticamente normal e consistente a uma taxa de $\sqrt{n}$, que é a taxa típica encontrada por estimadores paramétricos. Além disso, esse nível de convergência é o mesmo que seria encontrado se $g(\cdot)$ fosse

conhecido. Como resultado, o estimador de $g(\cdot)$ é assintoticamente não afetado pela incerteza sobre o vetor β.

3.3 O Estimador Semiparamétrico de Mínimos Quadrados Ponderados (SMQP)

O estimador de β utilizando SMQ é consistente até mesmo quando os erros são heteroscedásticos. Todavia, no caso heteroscedástico a eficiência do estimador pode ser melhorada introduzindo-se uma função peso apropriada ω. Estes pesos podem, por exemplo, ser proporcionais à variância do erro. Ichimura (1990) estuda este caso usando um parâmetro de suavização determinístico e Härdle, Hall e Ichimura (1993) apresentam um método de estimação conjunto da função peso e da janela.

A estimação do modelo utilizando o conhecido SMQP é idêntica ao método SMQ. A única diferença é que uma função peso é incorporada ao método, resultando em

$$\sum_{i=1}^{n}\{Y_i-[a+b(\beta' X_i-z)]\}^2 K_h(\beta' X_i-z)\omega(\beta' X_i), \tag{3.5}$$

que é a função (3.2) ponderada por $\omega(\beta' X_i)$, e

$$S_{T_C}(\beta,h)=\frac{1}{n}\sum_{i=1}^{n}\{Y_i-\hat{g}_{-i}(\beta' X_i)\}^2 \omega(\beta' X_i), \tag{3.6}$$

representando a função de validação cruzada para o SMQP.

4. SIMULAÇÕES

O objetivo deste capítulo é verificar o desempenho do modelo de índice único, apresentado no Capítulo 3, sob diversas situações. Para tanto, serão utilizados o *kernel* gaussiano e o método da validação cruzada. A função de validação cruzada será minimizada utilizando o algoritmo simplex apresentado por Nelder e Mead (1965) e implementado no software R.

4.1 Exemplos

Exemplo 1: Em primeiro lugar, faremos uma simulação de um modelo que se encaixa no padrão de um índice único, isto é, $Y = g(\beta' X)$. O exemplo é definido por:

$$Y_i = 10.\cos(0.6 + 0.4X_1 + 1.2X_2) + \varepsilon_i \qquad (4.1)$$

onde $\varepsilon_i \sim N(0,1)$, $X_1 \sim U(-5;5)$ e $X_2 \sim U(-5;5)$.

Abaixo, a Figura 4.1 ilustra a verdadeira curva, representando a esperança condicional da função Y definida em (4.1), ou seja, $\mathrm{E}(Y_i / X_1, X_2) = 10.\cos(0.6 + 0.4X_1 + 1.2X_2)$.

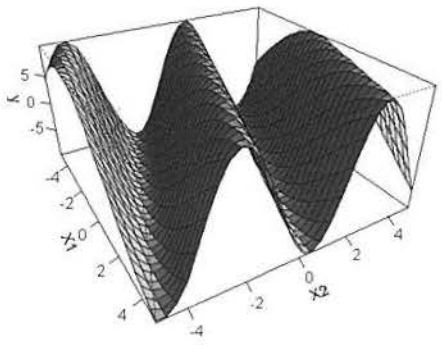

Figura 4.1 - Curva da esperança condicional da função (4.1).

As funções foram estimadas para 30^2 pontos no intervalo de $[-5,5]^2 \in \Re^2$. Para fins de exemplificação, foram utilizados diferentes tamanhos de amostra, sendo que para cada tamanho amostral a função foi executada somente uma vez.

Na Figura 4.2 são mostradas as regressões estimadas da função (4.1) utilizando-se amostras de tamanho 50, 100, 200, e 500.

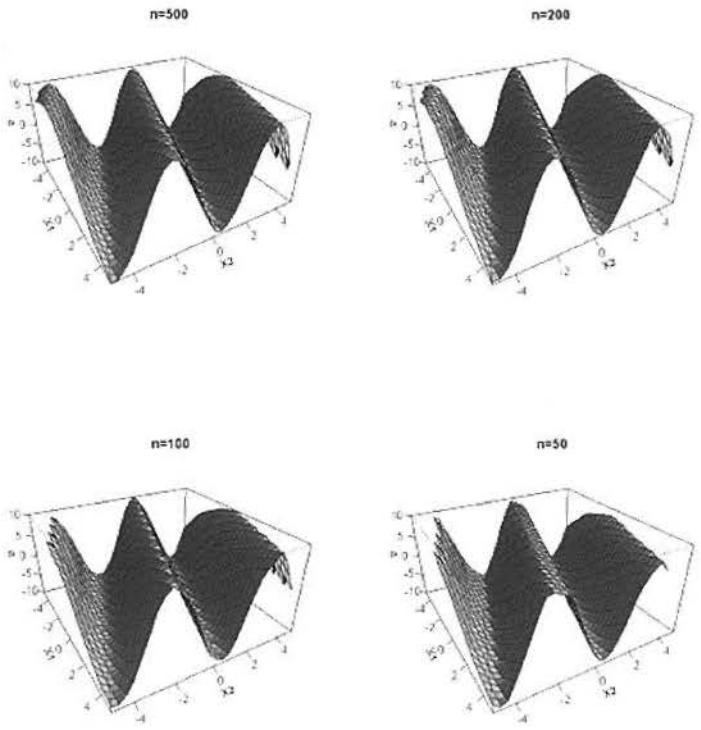

Figura 4.2 - Estimativas da função (4.1) utilizando amostras de diferentes tamanhos.

Podemos perceber pela Figura 4.2 que, independentemente do tamanho amostral, as curvas de regressão estimaram adequadamente a verdadeira função, sendo que nas partes centrais não é possível diferenciá-las visivelmente. Entretanto, é aparente a dificuldade de estimação especialmente nos limites negativos dos suportes das duas variáveis (nos limites da curva) para tamanhos de amostra 50 e 100, onde não há valores estimados.

Uma maneira de resolver essa dificuldade seria a utilização de diferentes janelas na estimação da função de regressão, lembrando que estamos trabalhando com h fixo.

Essa técnica permite que a janela leve em consideração o esparsamento dos dados, assumindo valores altos em regiões onde há poucas observações. Isso permite obter mais informação em regiões limites, onde, geralmente, há poucos valores observados. Por outro lado, em regiões onde há uma maior concentração de observações, a janela assumiria pequenos valores, devido à grande quantidade de vizinhos. Para maiores detalhes sobre o estimador utilizando janelas variáveis ver Ziegelmann (2002).

Para que possamos comparar as funções estimadas - com amostras de diferentes tamanhos - utilizaremos o erro quadrático médio (EQM) como medida de erro entre a função verdadeira e as funções estimadas. Abaixo a definição do EQM com relação aos pontos estimados.

$$EQM = \frac{1}{I.J}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\hat{g}(\beta_1 x_i + \hat{\beta}_2 x_j) - m(X_i, X_j)\right)^2$$

onde $\beta_1 = 1$ (necessário para a identificabilidade do modelo), x_i é o i-ésimo valor da variável X_1, $\hat{\beta}_2$ é o valor estimado para o coeficiente da variável X_2, x_j é o j-ésimo valor da variável X_2 e $m(X_i, X_j)$ é o valor da função no ponto (x_i, x_j). Todos os cálculos do EQM desse capítulo terão $I = 30$ e $J = 30$. Assim, o erro será estimado em 900 pontos.

Como as funções estimadas para tamanhos de amostra 50 e 100 não possuem estimativas nos limites da curva de regressão, não seria sensato calcular o EQM no suporte de [-5,5], pois o erro, para as amostras de tamanho 50 e 100, seria subestimado. Para solucionar esse prolema calculamos o EQM com relação aos 30^2 pontos do intervalo de $[-4,4]^2 \in \Re^2$ em vez de $[-5,5]^2 \in \Re^2$. Esse intervalo foi escolhido, pois apresentava estimativas para todos os pontos calculados, independentemente do tamanho amostral.

O Quadro 4.1 fornece a relação do EQM com relação aos 30^2 pontos do intervalo de $[-4,4]^2 \in \Re^2$, em relação aos diferentes tamanhos amostrais. Os valores abaixo indicam que com o aumento da amostra o erro tende a diminuir, além de sugerir a

adequabilidade da estimativa com relação à verdadeira regressão, visto que o EQM assume valores baixos.

amostra	n=500	n=200	n=100	n=50
EQM	0.09437237	0.09834525	0.2919258	0.5430704

Quadro 4.1 - EQM das estimativas da função (4.1) para amostras de diferentes tamanhos.

Analisando a Figura 4.3, podemos perceber que a partir de uma amostra de tamanho 200 a diferença entre os erros é pequena. Isso pode indicar que quando a regressão estimada se comporta de acordo com um índice, não é necessária uma amostra muito grande para se obter uma boa estimativa.

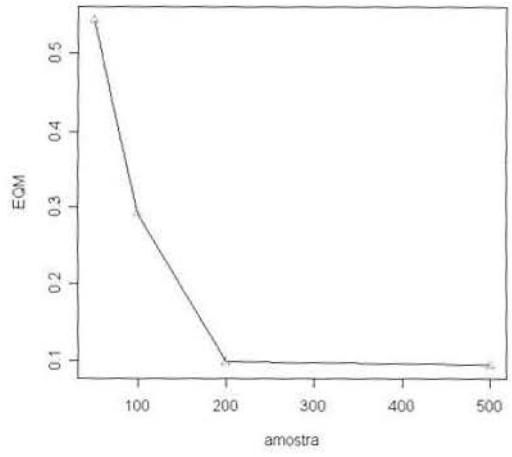

Figura 4.3 - Gráfico indicando o valor do EQM em relação ao tamanho amostral.

Uma dúvida muito comum em Regressão Paramétrica - onde se utiliza a ANOVA como ferramenta para testar a suposição de significância das variáveis regressoras - é quanto ao número de covariáveis que devem ser incorporadas ao modelo.

Em Regressão Semiparamétrica a dúvida também ocorre. Afinal, além do particular referencial teórico para determinado problema, como definir quais variáveis irão compor o modelo dentre um conjunto de candidatas? Para resolver essa questão

utilizaremos o critério de validação cruzada. Assim, o modelo que obtiver o menor valor nesse critério será o escolhido.

Para demonstrar a eficácia desse método foram realizadas 50 simulações da função Y, utilizando uma amostra de tamanho 50; isto é, obtivemos 50 conjuntos de dados de tamanho 50. Para a estimação da função $Y_i = 10.\cos(0.6 + 0.4X_1 + 1.2X_2) + \varepsilon_i$ foram utilizados três modelos:

1. $\hat{Y}_i = \hat{g}(X_1)$;

2. $\hat{Y}_i = \hat{g}(X_1 + \hat{\beta}_2 X_2)$;

3. $\hat{Y}_i = \hat{g}(X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3)$, onde $X_3 \sim U(-5;5)$.

No primeiro procedimento a covariável X_2 foi retirada, no segundo foram utilizadas as covariáveis corretas e no terceiro uma terceira variável regressora X_3 foi adicionada, cujos valores foram gerados independentemente das variáveis X_1 e X_2, os quais pertencem a uma variável aleatória de distribuição uniforme no intervalo [-5;5]. Essa variável X_3 não faz parte da função Y, sendo colocada artificialmente no cálculo da estimação da verdadeira regressão.

O valor encontrado na validação cruzada foi comparado entre os três modelos. Abaixo o Quadro 4.2 indica o número de vezes que cada um dos modelos foi escolhido, a saber, com uma, duas ou três variáveis regressoras.

n° de covariáveis no modelo	1	2	3
n° de vezes escolhido	0	26	24

Quadro 4.2 - Número de escolhas por cada modelo nas 50 simulações.

Apesar de a verdadeira função possuir somente duas covariáveis, o modelo que possui três variáveis regressoras foi escolhido diversas vezes. Isso pode nos levar a pensar que o método utilizado para a escolha do número de variáveis explicativas do

modelo não foi adequado. Porém, deve-se analisar o Box-Plot do critério de validação cruzada na Figura 4.4 para então tirar conclusões.

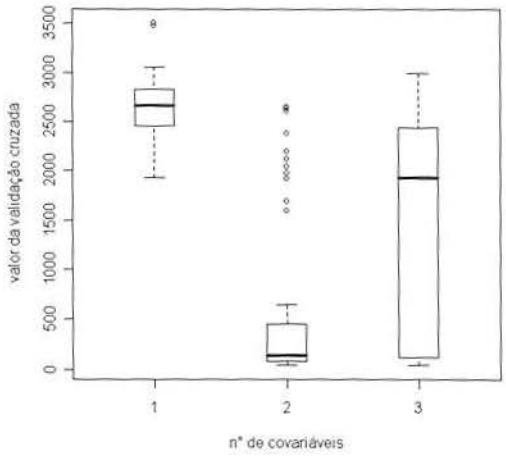

Figura 4.4 - Box-Plot do critério de validação cruzada das estimativas da função (4.1).

Analisando o Box-Plot vemos que o modelo que possui a menor mediana para o valor da função de validação cruzada foi o de duas covariáveis. O modelo de três covariáveis possui uma variância muito grande, assumindo valores tanto altos quanto baixos. Isso explica o porquê dele ter sido escolhido diversas vezes.

Calculando a média - nas 26 simulações em que se escolheram três variáveis explicativas para o modelo - do $\hat{\beta}_2$ para o modelo com duas covariáveis e comparando com a média de $\hat{\beta}_2$ do modelo de três variáveis temos os seguintes valores: $\overline{\hat{\beta}_2}$ (2variáveis)= 2,211 e $\overline{\hat{\beta}_2}$ (3variáveis)= 2.655.

Isso indica que, apesar do modelo escolher três covariáveis, ele estima de maneira muito parecida o valor de $\hat{\beta}_2$ em relação ao modelo que utiliza duas covariáveis. Além disso, calculando a média do $\hat{\beta}_3$ nas 24 simulações em que se escolheram três variáveis explicativas temos que $\overline{\hat{\beta}_3}$ =0.1331238. Comparando-se este valor com os coeficientes estimados para β_2 observa-se que a terceira variável possui pouca influência na estimativa do modelo (note que as variáveis estão na mesma escala).

Portanto, esse comportamento do critério de validação cruzada pode ser explicado pelo fato de a variável X_3 não influenciar de maneira decisiva na estimação da regressão, e do $\hat{\beta}_2$ ser praticamente o mesmo nos dois casos. Logo, os valores estimados entre os dois modelos serão muito parecidos.

Apesar disso, deve-se ressaltar que o critério parece parametrizar em excesso o modelo.

No exemplo acima $\varepsilon_i \sim N(0,1)$. No entanto, se este é um estimador semiparamétrico ele deve ser robusto, independentemente da distribuição do erro. Para verificar o desempenho desse estimador em relação a possíveis especificações do erro, estudamos a mesma função média do exemplo anterior mas, agora, com $\varepsilon_i \sim N(0;3)$ (somente aumentando a variância) e $\varepsilon_i \sim [\exp(1)-1]$ (introduzindo assimetria no erro).

Nas Figuras 4.5 e 4.6 seguem as regressões estimadas para a média condional em (4.1) com $\varepsilon_i \sim N(0;3)$ e $\varepsilon_i \sim [\exp(1)-1]$, respectivamente. Novamente as funções estimadas comportam-se de maneira muita parecida com a forma da verdadeira regressão, sendo que na parte central não é possível diferenciá-las visivelmente. Novamente ocorreram problemas na estimação das bordas, onde não há estimativas.

Para comparar as curvas estimadas utilizando diferentes distribuições para o erro aleatório calculamos o EQM com relação aos 30^2 pontos do intervalo de $[-4,4]^2 \in \Re^2$, este intervalo foi novamente utilizado para excluir a área onde ocorreram problemas de estimação, logo poderemos comparar o EQM dos diferentes tamanhos amostrais. Abaixo, segue o Quadro 4.3 contendo o EQM das regressões estimadas.

amostra	$\varepsilon_i \sim N(0;1)$	$\varepsilon_i \sim N(0;3)$	$\varepsilon_i \sim [\exp(1)-1]$
n=500	0.09437237	0.6412783	0.03893523
n=200	0.09834525	0.6345203	0.05038906
n=100	0.2919258	1.923616	0.2153295
n=50	0.5430704	2.90799	0.485508

Quadro 4.3 - EQM da função (4.1) para diferentes distribuições do erro.

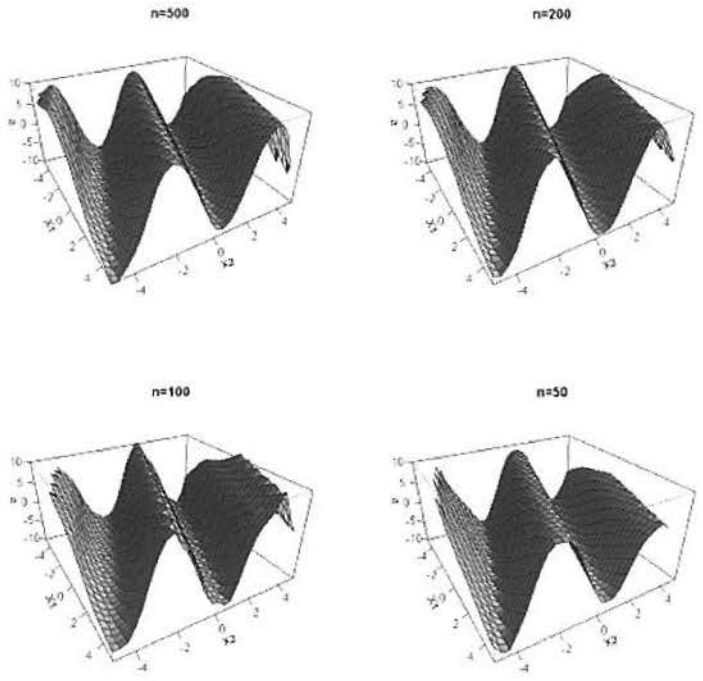

Figura 4.5 - Estimativas de (4.1) quando $\varepsilon_i \sim N(0;3)$, para amostras de diferentes tamanhos.

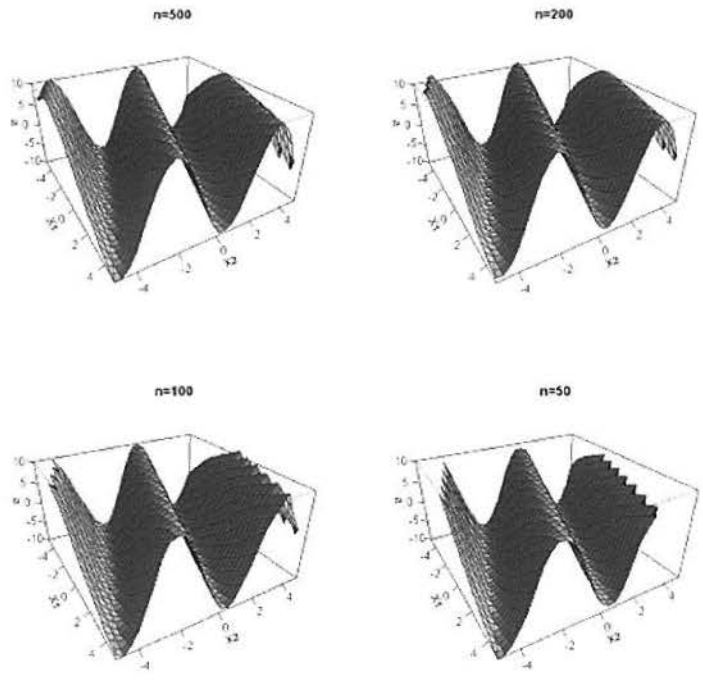

Figura 4.6 - Estimativas de (4.1) quando $\varepsilon_i \sim e[xp(1)-1]$, para amostras de diferentes tamanhos.

A Figura 4.7 mostra de maneira mais clara como o modelo de índice único se comporta com relação à amostra. Nela plotamos o valor do índice com relação à sua estimativa para os três casos de erros. Para estes gráficos utilizamos a amostra de tamanho 200 e grade de tamanho 600. Os pontos pretos indicam as observações amostrais, os pontos vermelhos são os valores estimados e a curva azul é a verdadeira função de regressão.

Pode-se perceber que as estimativas não são fortemente afetadas pela distribuição do erro aleatório, logo poderíamos supor que o modelo de índice único deve ser robusto independentemente do tipo de ruído populacional.

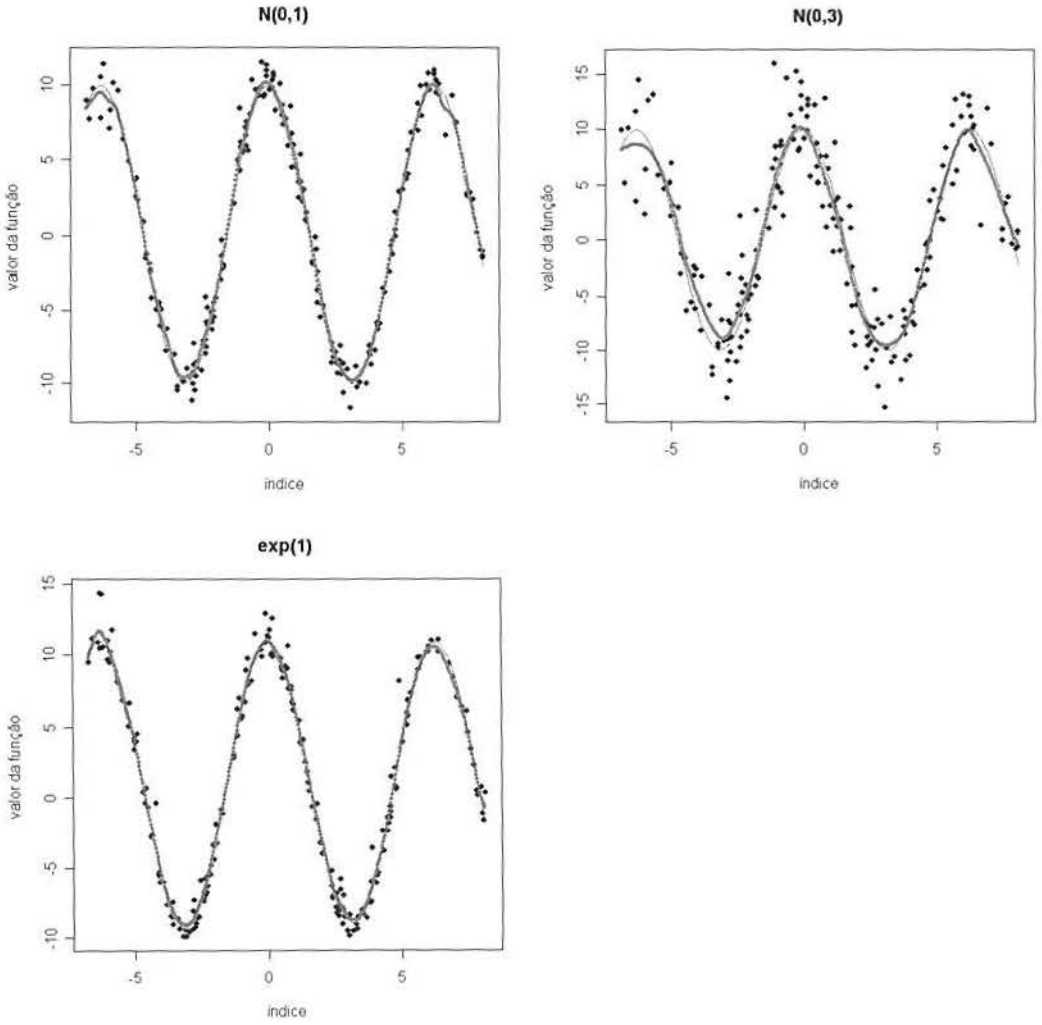

Figura 4.7 - Valor do índice com relação à estimativa (vermelho), a amostra (preto) e a verdadeira função de regressão (azul).

Exemplo 2: Agora vamos estimar o exemplo de um modelo que não se encaixa no padrão de índice único em relação as variáveis originais X_1 e X_2, isto é, $Y = g(X_1) + g(X_2)$. O exemplo é definido por:

$$Y_i = 1.3 + 2.3(X_1{}^3) + 1.2(X_2{}^3) + \varepsilon_i \qquad (4.2)$$

onde $\varepsilon_i \sim N(0,1.5)$, $X_1 \sim U(-10;10)$ e $X_2 \sim U(-10;10)$.

A Figura 4.8 mostra como é o comportamento da esperança condicional da função Y definida em (4.2), ou seja, $\mathrm{E}(Y_i / X_1, X_2) = 1.3 + 2.3(X_1{}^3) + 1.2(X_2{}^3)$.

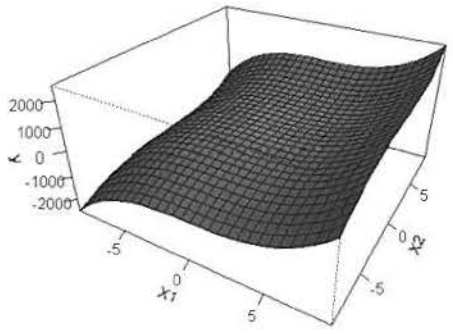

Figura 4.8 - Curva da esperança condicional da função (4.2).

As funções foram estimadas para 30^2 pontos no intervalo de $[-10,10]^2 \in \Re^2$. A Figura 4.9 mostra as estimativas de uma replicação da regressão descrita acima, utilizando diferentes tamanhos de amostra. Novamente o modelo de índice único parece seguir adequadamente a forma da função de regressão, sendo que sua eficácia parece aumentar de acordo com o maior número de observações. Observa-se que, mais uma vez, há problemas de estimação no limite da função, onde alguns pontos não possuem estimativas. Entretanto, essa característica é menos proeminente que no exemplo anterior.

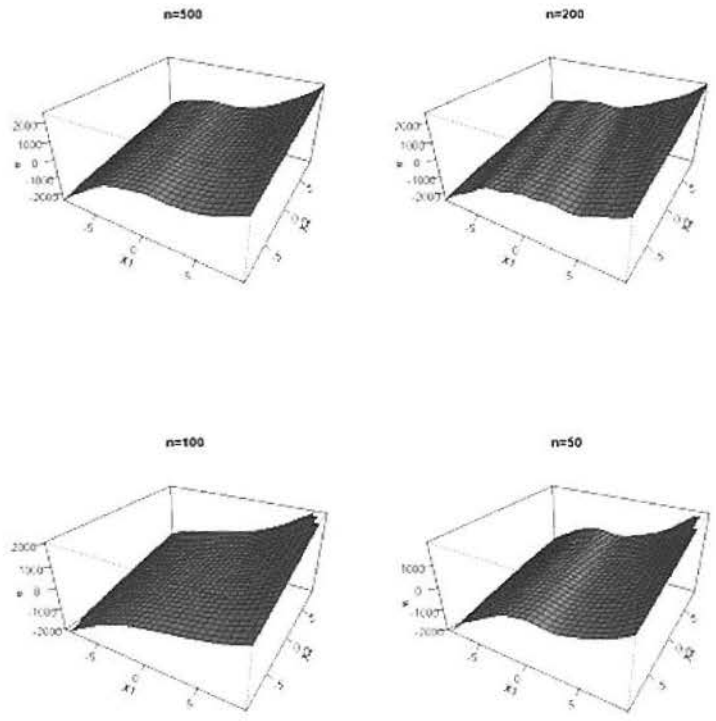

Figura 4.9 - Estimativas da função (4.2) para amostras de diferentes tamanhos.

Utilizaremos o EQM para comparação dos erros das estimativas em relação à verdadeira função de regressão. No Quadro 4.4 se encontra o EQM com relação aos 30^2 pontos no intervalo de $[-9,9]^2 \in \Re^2$. Novamente o EQM foi calculado utilizando um intervalo com tamanho um pouco menor que o do verdadeiro suporte amostral, essa medida foi adotada para que o EQM não fosse subestimado nas amostras de tamanho 50 e 100.

Amostra	n=500	n=200	n=100	n=50
EQM	67113.24	73415.03	81616.94	75508.97

Quadro 4.4: EQM das estimativas da função (4.2) para amostras de diferentes tamanhos.

A Figura 4.10 mostra o comportamento do EQM com relação ao tamanho amostral. Ao contrário do esperado o EQM da amostra de tamanho 100 é superior ao erro encontrado na estimação contendo 50 observações – não obstante esse

comportamento pode estar relacionado com flutuações amostrais. Também podemos observar que o EQM da simulação contendo 200 observações é muito superior ao daquela contendo 500 observações. Lembrando que no Exemplo 1 o EQM também diminui quando o tamanho amostral aumenta, mas a diferença entre os erros não é tão visível. Logo, poderíamos sugerir que quando um modelo não se encaixar no padrão (não estiver corretamente especificado) de um índice único, o tamanho da amostra terá uma maior influência na estimativa.

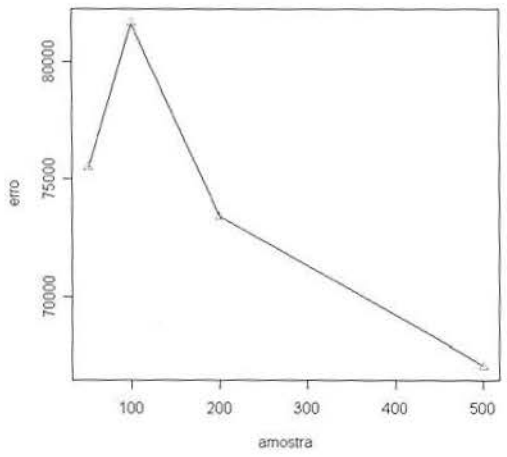

Figura 4.10 - Valor do EQM em relação ao tamanho amostral.

Já havíamos falado no Exemplo 1 que o número de variáveis regressoras será escolhido a partir do valor da validação cruzada. Novamente procedemos à simulação de 50 regressões, utilizando em cada uma delas uma amostra de tamanho 50.

Para a estimação, foram utilizados novamente três modelos. No primeiro procedimento a covariável X_2 foi retirada, no segundo foram utilizadas as covariáveis corretas e no terceiro uma terceira variável $X_3 \sim U(-10;10)$ foi incorporada no cálculo da estimação. É importante frisar que a variável X_3 não faz parte da função de regressão Y_i e foi gerada independentemente de X_1 e X_2, isto é, representa uma variável que não possui nenhuma influência sobre Y_i.

Abaixo o Quadro 4.5 indica o número de vezes que cada um dos modelos foi escolhido, a saber, com uma, duas ou três variáveis regressoras.

n° de covariáveis no modelo	1	2	3
n° de vezes escolhido	2	20	28

Quadro 4.5 - Número de escolhas por cada modelo nas 50 simulações.

Desta vez o modelo com três variáveis foi mais vezes escolhido. Para descrever o comportamento do critério neste caso, veja a Figura 4.11 contendo os Box- Plots da função de validação cruzada para os diferentes modelos.

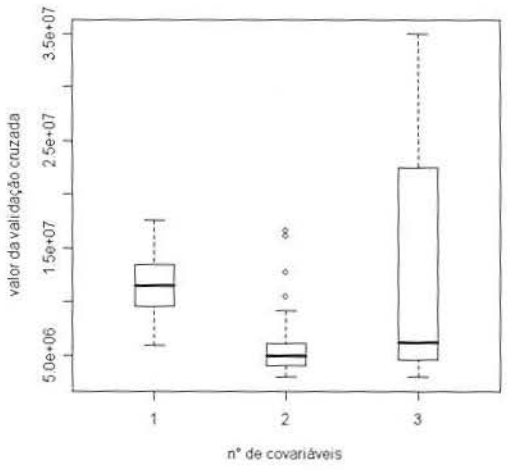

Figura 4.11 - Box-Plot do critério de validação cruzada das estimativas da função (4.2).

Pelo Box-Plot vemos que a escolha de covariáveis (utilizando o valor da validação cruzada) em um modelo que não tem a forma do índice único se comporta de maneira similar ao exemplo com especificação correta.

Calculando a média do $\hat{\beta}_2$ nas 28 simulações em que se escolheram três variáveis explicativas para o modelo temos os seguintes valores: $\overline{\hat{\beta}_2}$ (2variáveis)= 0,335 e $\overline{\hat{\beta}_2}$ (3variáveis)= 0.3106. Além disso, calculando a média do $\hat{\beta}_3$ nas 28 simulações em que se escolheram três variáveis explicativas temos que $\overline{\hat{\beta}_3}$ =0.0037.

Novamente, apesar do modelo escolher três covariáveis os valores estimados são muito parecidos. Pois, as estimativas de β_2 são semelhantes nos modelos com três e duas variáveis regressoras. Além disso, o valor de $\overline{\hat{\beta}_3}$ é pequeno em comparação com o $\overline{\hat{\beta}_2}$, exercendo quase nenhuma influência na estimativa da regressão. Essa comparação é possível porque a escala de X_3 é a mesma que das covariáveis X_1 e X_2.

Exemplo 3: Para comparar o desempenho do modelo de índice único em relação a outros métodos de estimação procedemos ao seguinte exemplo. Considere a função a seguir:

$$Y_i = 0.2 + 2(X_1^2) + 2.3(X_2^2) + \varepsilon_i \tag{4.3}$$

onde $\varepsilon_i \sim N(0,1)$, $X_1 \sim U(-5;5)$ e $X_2 \sim U(-5;5)$.

Fez-se a estimativa da regressão acima usando três diferentes modelos. O 1° é o de uma regressão paramétrica simples. O 2° é uma regressão paramétrica utilizando-se X_1^2 e X_2^2, e o 3° utiliza o modelo de índice único para as variáveis originais.

Para comparar esses três procedimentos simulamos 50 conjuntos de observações da função acima para amostras de tamanhos 50, 100, 200 e 500. As funções foram estimadas para 30^2 pontos no intervalo de [-5;5] $\in \Re^2$. Abaixo, as Figuras 4.12 a 4.15 mostram a forma da regressão verdadeira e as curvas estimadas - para uma das simulações - dos três diferentes métodos para amostras de tamanho 50, 100, 200 e 500 respectivamente.

Pelas Figuras 4.12 a 4.15 torna-se claro que a estimação paramétrica utilizando-se as covariáveis ao quadrado foi o método que mais se aproximou da forma verdadeira da função de regressão, seguido pelo modelo de índice único. O modelo paramétrico com as covariáveis originais foi o que apresentou o pior desempenho.

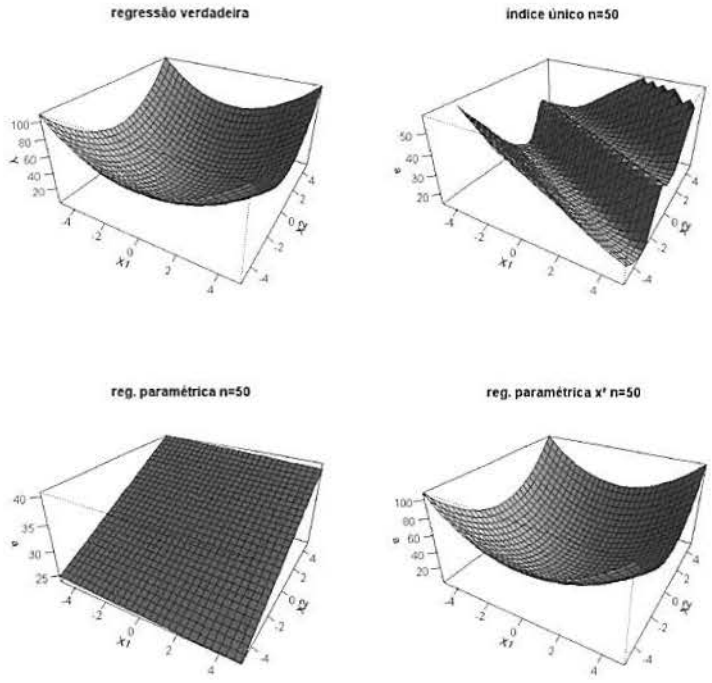

Figura 4.12 - Regressão verdadeira da função (4.3) e seus valores estimados com diferentes métodos para uma amostra de tamanho 50.

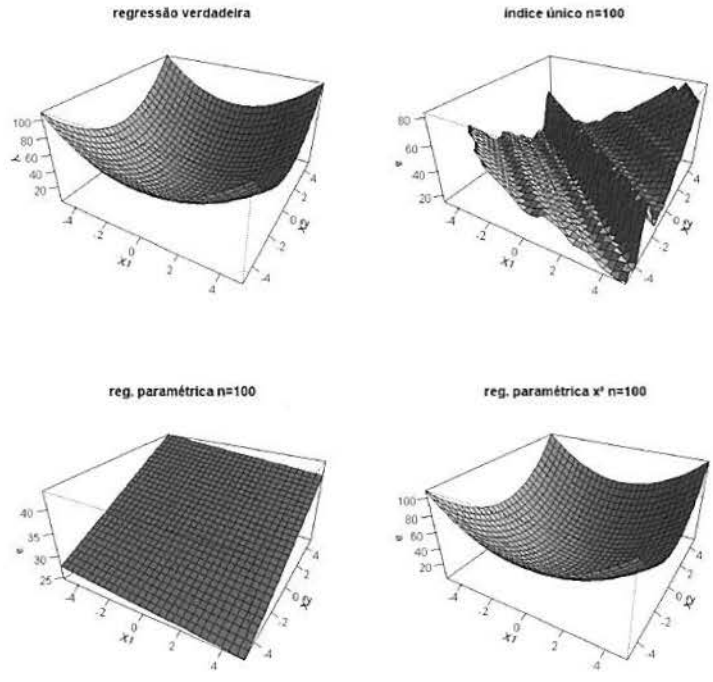

Figura 4.13 - Regressão verdadeira da função (4.3) e seus valores estimados de diferentes métodos para uma amostra de tamanho 100.

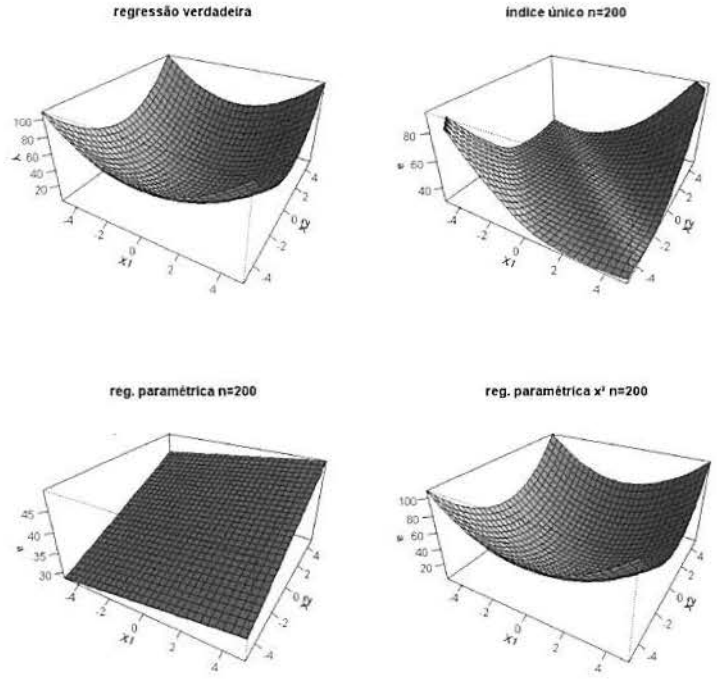

Figura 4.14 - Regressão verdadeira da função (4.3) e seus valores estimados de diferentes métodos para uma amostra de tamanho 200.

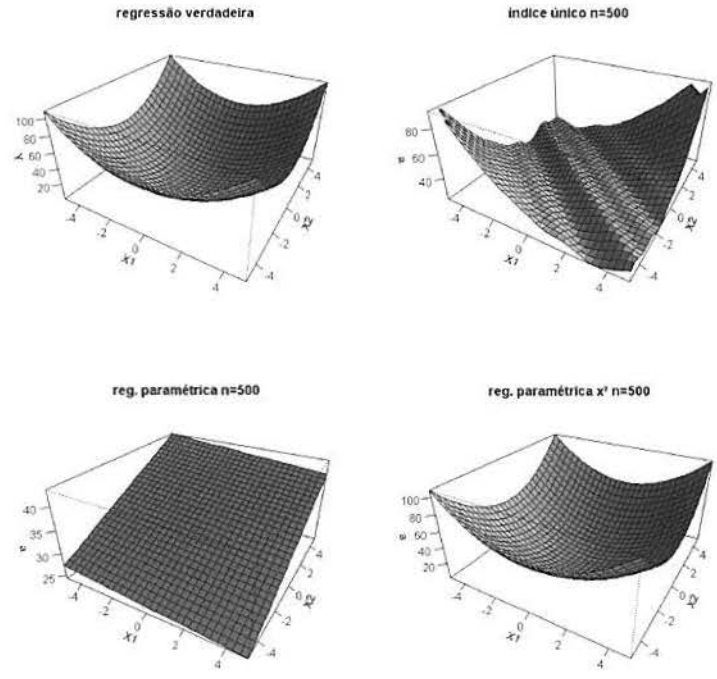

Figura 4.15 - Regressão verdadeira da função (4.3) e seus valores estimados de diferentes métodos para uma amostra de tamanho 500.

Também, é visível que o modelo de índice único apresentou dificuldades na estimação dos limites da verdadeira função, principalmente nas amostras de tamanho 50 e 100.

Para comparação dos diferentes métodos calculamos o EQM de cada estimativa em relação aos 30^2 pontos do intervalo [-4,4], novamente utilizamos um intervalo menor para tentar "escapar" dos casos onde a regressão não possui estimativas.

Como simulamos várias regressões, não observamos as estimativas individualmente, logo em alguns casos poderia haver áreas, localizadas nos limites das funções estimadas, que não teriam estimativas. Nesses casos consideramos o erro do ponto estimado com relação ao verdadeiro como sendo zero, subestimando o valor do EQM em alguns casos. É importante frisar que eram raros esses fatos, visto que já utilizamos um menor intervalo para o cálculo do EQM. A Figura 4.16 compara os três estimadores através do Box-Plot do EQM das 50 replicações.

É importante salientar que o modelo paramétrico com as covariáveis ao quadrado teve muito mais facilidade na estimação. Visto que ele já especificava a priori o comportamento não linear da função (4.5), estimando somente os coeficientes das variáveis. O modelo de índice único teve um trabalho mais árduo, pois além de estimar os coeficientes da função havia o comportamento não linear de X_1^2 e X_2^2 que não se encaixa perfeitamente no padrão do modelo índice único (não formam um índice).

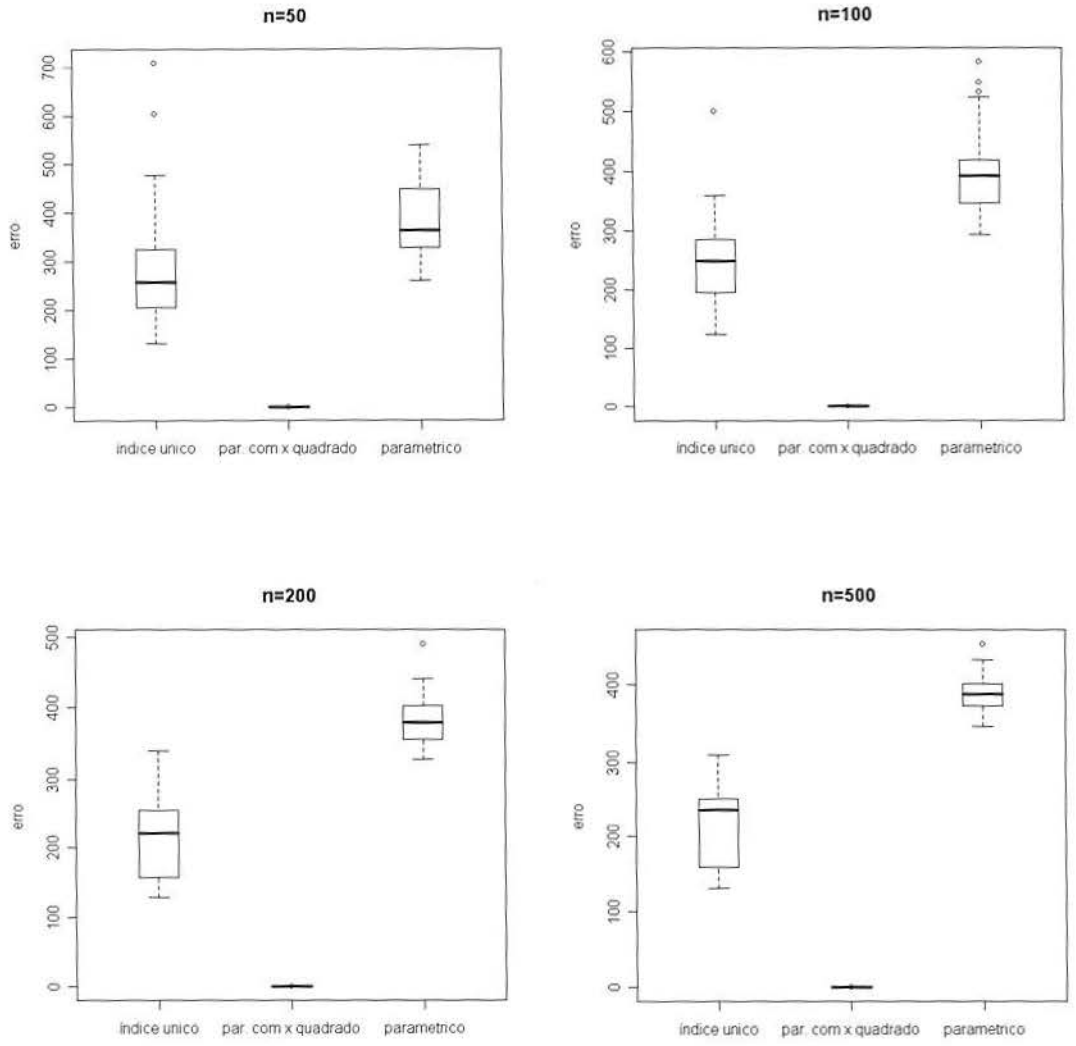

Figura 4.16 - Box-Plot do EQM para amostras de diferentes tamanhos das estimativas da função (4.3) utilizando diferentes métodos de estimação.

5. APLICAÇÃO PRÁTICA

Risco e incerteza são características comuns de dados financeiros. Uma aproximação muito utilizada para obter medidas desses fatores é a volatividade dos retornos. Os retornos podem ser definidos por

$$Y_t = \log X_t - \log X_{t-1}, \tag{5.1}$$

onde X_t é o preço de uma ação no tempo t. O termo volatividade representa qualquer medida de variabilidade, tal como variância ou desvio padrão.

Sejam $m(x) = E(Y / X = x)$ e $\sigma^2(x) = Var(Y / X = x) > 0$. Considere o seguinte modelo:

$$Y_t = m(X_t) + \sigma(X_t)\varepsilon_t, \tag{5.2}$$

onde $E(\varepsilon_t / X_t) = 0$, $Var(\varepsilon_t / X_t) = 1$ e X_t representa um vetor de variáveis defasadas de Y_t, isto é, $X_t = (Y_{t-1}, Y_{t-2}, Y_{t-3}, ...)$.

No exemplo a seguir o estimador de $m(x)$ será obtido através do ajuste de um modelo ARMA paramétrico em um primeiro estágio. Para obter um estimador da volatividade dos retornos nós reescreveremos o modelo (5.2) como

$$\hat{r}_t^2 = [Y_t - m(X_t)]^2 = \hat{\sigma}^2(X_t)\varepsilon_t^2. \tag{5.3}$$

Considerando a esperança condicional dos resíduos ao quadrado $\hat{r}_t^2$, nós temos que

$$E[\hat{r}_t^2 / X_t] = \sigma^2(X_t). \tag{5.4}$$

Portanto, estimaremos $\sigma^2(X_t)$ da regressão dos resíduos ao quadrado estimados, $\hat{r}_t^2$, pelas defasagens de Y_t. Tal estimador será chamado de estimador residual e a partir dele iremos estimar a volatividade. Para maiores detalhes ver Ziegelmann (2002).

5.1 Estimação da Série Financeira Câmbio Dólar/Real

No contexto de aplicação do modelo de índice único, pretendemos estimar a volatividade dos retornos de uma série financeira real. A série a ser modelada é a da cotação diária do preço de venda do dólar comercial. A série possui 1727 observações, sendo que inicia em 3/1/2000 e termina em 14/11/2006, considerando somente os dias de negociação. Esses dados estão disponíveis no site https://www3.bcb.gov.br/sgspub/consultarvalores/consultarValoresSeries.paint?method=consultarValoresSeries (acessado em 16/11/2006). Esta série temporal é apresentada na Figura 5.1. Há uma clara sugestão de uma tendência estocástica, o que indica que ela é não estacionária. Além disso, ela parece apresentar um comportamento heteroscedástico, onde as primeiras observações apresentam uma variância inferior.

Figura 5.1 - Série temporal da cotação diária do dólar.

Já na Figura 5.2, temos a série de retornos definida por (5.1), cuja volatividade desejamos estimadar. Note os diferentes níveis de volatilidade ao longo do tempo.

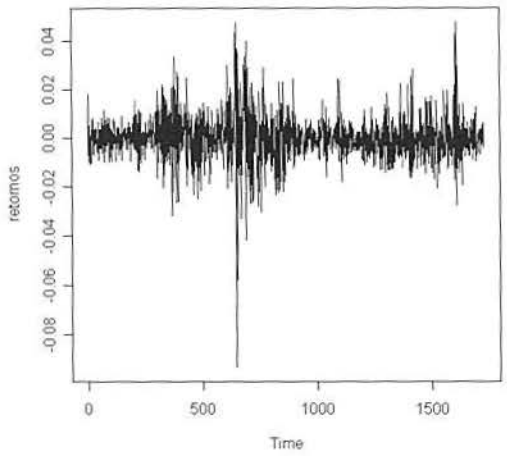

Figura 5.2 - Série dos retornos da cotação diária do dólar comercial.

Primeiramente obtemos $\hat{m}(x)$ da série de retornos. O modelo ajustado foi um AR(2). Através de (5.3) obtemos o estimador da volatividade dos retornos, o qual será obtido através dos quadrados dos resíduos estimados abaixo:

$$\hat{r}_t^2 = [Y_t - (0.1783Y_{t-1} - 0.1247Y_{t-2})]^2 .$$

A Figura 5.3 descreve o comportamento da série dos $\hat{r}_t^2$, obtida conforme acima.

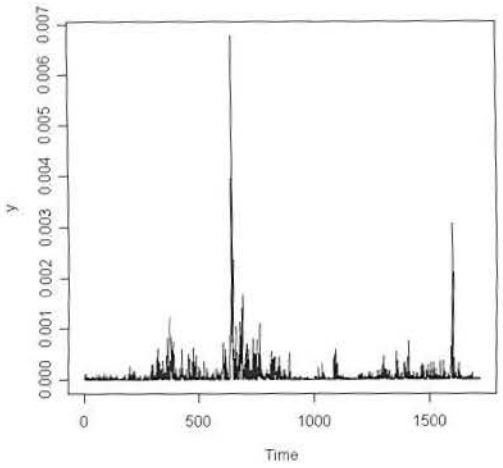

Figura 5.3 - Série dos $\hat{r}_t^2$.

Primeiramente foi escolhido o número de variáveis defasadas que devem ser utilizadas no processo de estimação. Para isso calculamos o valor da função de validação cruzada para modelos contendo 1, 2, 3 e 4 variáveis regressoras.

Abaixo segue o Quadro 5.1 com os coeficientes estimados e o valor da validação cruzada dos modelos contendo 1, 2, 3 e 4 variáveis regressoras para as últimas 1000 observações da série.

n° de covariáveis	1	2	3	4
validação cruzada	1.75037e-05	1.638722e-05	1.628829e-05	1.628715e-05
$\hat{\beta}_1$	1	1	1	1
$\hat{\beta}_2$	0	0.6327133	0.703880	0.69153477
$\hat{\beta}_3$	0	0	0.252762	0.25279091
$\hat{\beta}_4$	0	0	0	0.03139233
$\hat{h}$	1.1	1.3835850	1.624461	2.04405610

Quadro 5.1 - Coeficientes estimados e o valor da validação cruzada dos modelos contendo 1, 2, 3 e 4 variáveis regressoras.

O modelo que apresentou o menor valor para o critério da validação cruzada foi o de quatro covariáveis. Apesar disso, utilizaremos o modelo com 3 variáveis defasadas. Tomou-se essa decisão baseando-se em três motivos: primeiro, o valor da validação cruzada do modelo com 3 covariáveis é muito próximo do com 4; segundo $\hat{\beta}_4$ tem um valor muito pequeno, indicando que Y_{t-4} tem quase nenhuma influência na estimação da função de volatividade residual e terceiro pelo resultado das simulações feitas no Capítulo 4 sabemos que o modelo de índice único tende a parametrizar em excesso a função de estimação.

A Figura 5.4 mostra os valores estimados para a série da volatividade. A curva preta é a série dos $\hat{r}_t^2$, e a curva azul representa a série estimada. Para uma melhor visualização dos resultados dividimos a série em 6 partes.

Antes de analisar a Figura 5.4 é importante frisar que as estimativas plotadas abaixo possuem diferentes escalas, visto que cada parte da figura corresponde a uma parte da série temporal. Além disso, é importante considerar o comportamento da série como um todo e não estabelecer comparativos entre as diferentes partes estimadas.

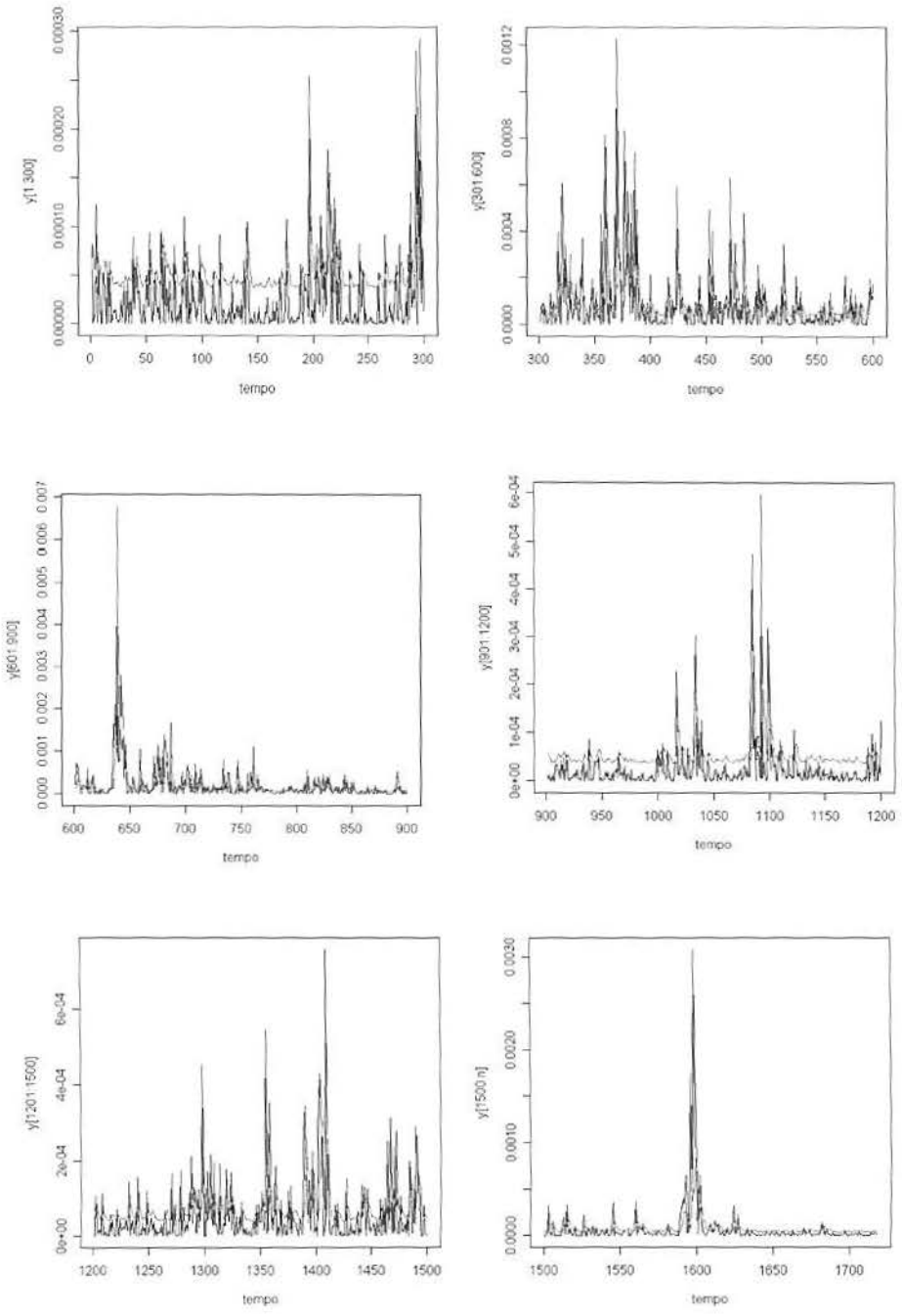

Figura 5.4 - Série do dos $\hat{r}_t^2$, real (preto) e estimada (azul).

As recomendações citadas acima são importantes, pois, se simplesmente olhássemos os gráficos plotados diríamos que a série estimada foi inadequada, já que a parte inicial e central (observações 1-300 e observações 900-1200) foi superestimada. Mas, esse não é o caso, visto as diferenças de escala. Além disso, a estimativa possui um comportamento praticamente idêntico à série verdadeira.

Para tornar mais evidente esse comportamento, plotamos na Figura 5.5, abaixo, a série estimada e verdadeira contendo todas as observações.

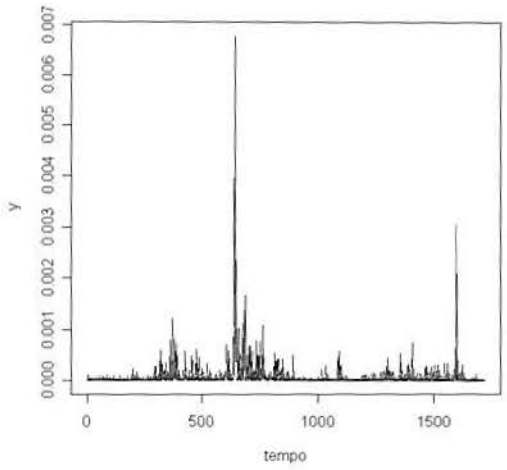

Figura 5.5 - Série real (preto) e estimada (azul).

A fim de visualizar o comportamento da volatividade em relação aos retornos, isto é, se a volatividade estimada se comporta de maneira a indicar a tendencia de maior variabilidade dos retornos em algumas partes da série, plotamos os retornos e abaixo deles a série da volatividade estimada, para uma melhor visualização da adequabilidade da estimativa.

As Figuras 5.6 e 5.8 mostram a série de retornos, já as Figuras 5.7 e 5.9 indicam a série da volatividade estimada.

Pelas figuras percebe-se que a volatividade estimada parece indicar de maneira razoável a variabilidade da série, observa-se que, quando os retornos mudam bruscamente a volatividade aumenta da mesma forma, indicando essa tendência.

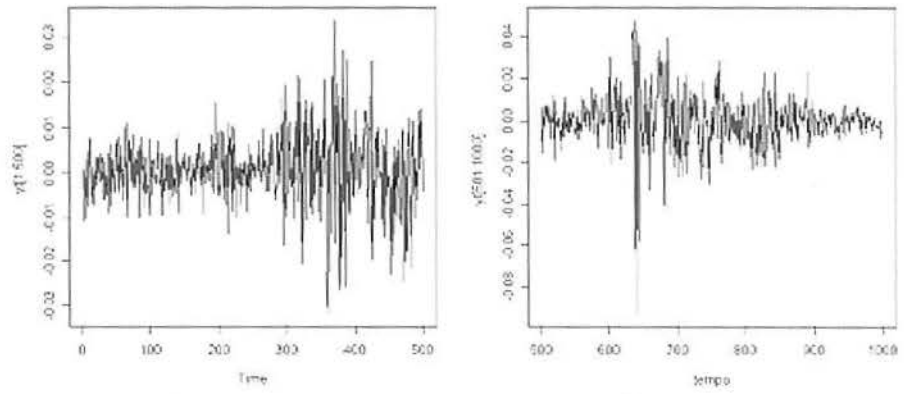

Figura 5.6 - Série dos retornos, da observalçao 1 à 500 e da observação 501 à 1000.

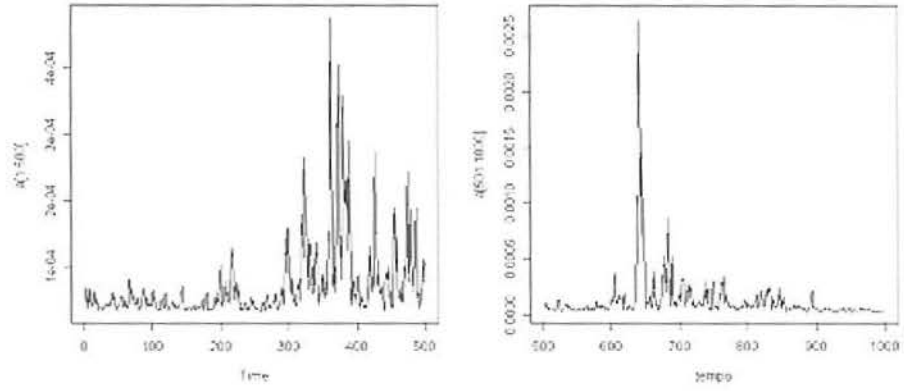

Figura 5.7 – Série da volatividade estimada, da observalçao 1 à 500 e da observação 501 à 1000.

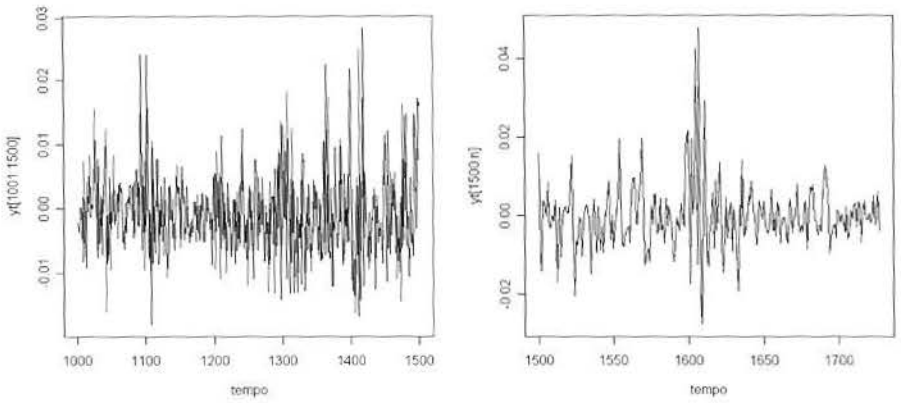

Figura 5.8 - Série dos retornos, da observalçao 1001 à 1500 e da observação 1501 à 1728.

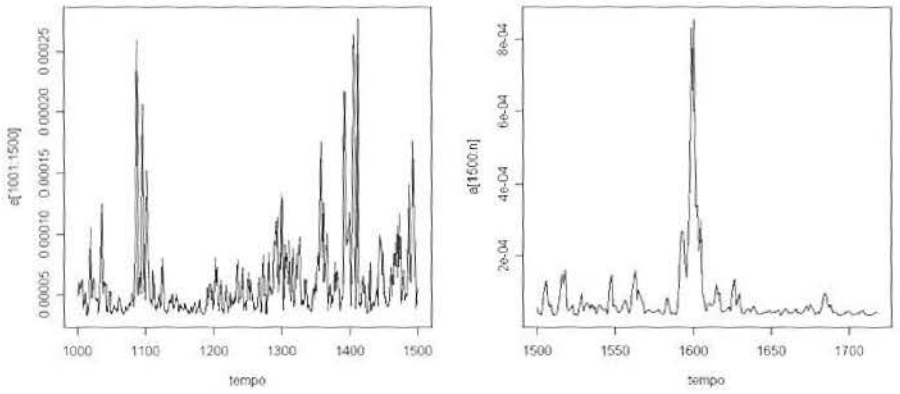

Figura 5.9 – Série da volatividade estimada, da observalçao 1001 à 1500 e da observação 1501 à 1728.

6. CONSIDERAÇÕES FINAIS

Como já foi abordado nos capítulos anteriores sabemos que a regressão pode ser estimada não parametricamente, parametricamente ou semiparametricamente. O problema encontrado na regressão paramétrica é sua forte restrição, enquanto a regressão não paramétrica - apesar de sua grande aplicabilidade e ausência de restrições - infelizmente não se "comporta" muito bem quando há muitas variáveis regressoras envolvidas.

Então tentamos, além de tornar mais conhecido esse tipo de modelagem, convencer o leitor que a regressão semiparamétrica estima de maneira adequada relações entre variáveis em muitos tipos de dados. Previamente, porém, foram discutidos alguns aspectos de estimação de funções densidade de probabilidade e regressão não paramétrica, que auxiliam na compreensão dos modelos semiparamétricos e consequentemente no método de estimação do modelo de índice único.

Pelas simulações verificamos que a modelagem é muito boa quando a função se comporta de acordo com um índice de covariáveis, isto é, $g(\beta' X_i)$. Onde havia esse tipo de comportamento as estimativas também se apresentaram robustas com relação aos diferentes tipos de erros aleatórios.

Além disso, descobrimos que o método de validação cruzada tende a parametrizar em excesso o modelo, escolhendo mais variáveis explicativas que o necessário. Entretanto é importante salientar que apesar dessa tendência as estimativas com variáveis a mais são muito parecidas com as estimativas contendo o número correto de variáveis regressoras.

Nos casos em que a função não se encaixa no padrão de um índice único, isto é, $g(\beta X_1) + g(\beta X_2)$, as estimativas se comportam de maneira diferente, tendo ou não bom desempenhos dependendo da forma da função de regressão. Logo poderíamos sugerir que em alguns casos seria melhor utilizar as covariáveis transformadas (X^2 por exemplo) na estimação de algumas regressões, o que provavelmente, dependendo da transformação escolhida, melhoraria a eficácia da estimativa.

Na análise da série financeira empírica, estimamos a volatividade dos retornos da série da cotação do dólar comercial de 3/1/2000 a 14/11/2006 . O modelo neste caso pareceu estimar adequadamente o comportamento da volatividade, como pudemos perceber nos gráficos apresentados.

REFERÊNCIAS BIBLIOGRÁFICAS

Allen, D. M. (1974). The relationship between variable and data argmentation and method of prediction. Technometrics, **16**, 125-127.

Bowman, A. W. (1984). An altenative method of cross-validation for the smoothing density estimates. *Biometrika*, **71**, 353-360.

Bowman, A. W. e Azzalini A. (1997). *Applied Smoothing Techniques for Data Analysis*. New York: Oxford University Press.

Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of American Statistical Association*, **83**, 596-610.

Efromovich, S. (1999). *Nonparametric Curve Estimation: Methods, Theory, and Applications*, New York: Springer-Verlag

Fan, J (1992). Design-adaptive nonparametric regression. *Journal of the American Statistics Association*, **87**, 998-1004.

Green, P. J. e Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. London: Chapman and Hall.

Hall, P (1989). On projection.pursuit regression. *Annals of Statistics*, **17**, 573-588.

Härdle W., Hall P. e Ichimura H. (1993). Optimal smoothing in single-index models. *Annals of Statistics*, **21**, 157-178.

Härdle W. (1994). *Applied Nonparametric Regression*. Disponível em http://www.quantlet.com/mdstat/scripts/anr/pdf/anrpdf.pdf (acessado em 30/10/2006).

Härdle W, Müller M., Sperlich S. e Werwatz A. (2004). *Nonparametric and Semiparametric Models*. Disponível em http://www.quantlet.com/mdstat/scripts/spm/html/spmhtml.html (acessado em 30/10/2006).

Härdle, W. e Stoker, T. (1989). Investigating Smooth Multiple Regression by the Method of Average Derivatives. *Journal of the American Statistical Association*, **84**.

Horowitz J. L.(1998). *Semiparametric Methods in Econometrics*. New York. Springer-Verlag.

Ichimura H. (1990). Semiparametric weighted least squares estimation of single-index models. Artigo não publicado.

Ichimura H. (1993). Semiparametric least-squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, **58**, 71-120.

Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, **86**, 316-342.

Nelder, J. A. e Mead, R. (1965). A simplex algorithm for function minimization. *Computer Journal*, **7**, 308–313.

Park, B. U. e Marron, J. S. (1990). Comparasion of data-driven bandwidth selectors. *Journal of the American Statistical Association*, **85**, 66-72.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, **9**, 65-78.

Stone, C. J. (1977). Consistent nonparametric regression. *Annals of Statistics*, **5**, 595-620.

Stone, M. A. (1974). Cross-validatory choice and assessment of statistical prediction. *Journal of the Royal Statistical Society*, Series B, **36**, 111-147.

Wand, M. P. e Jones, M. C. (1995). *Kernel Smoothing*. London: Chapman and Hall.

Xia Y., Tong H., Li W. K e Zhu Li-Xing (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society*, Série B (Statistical Methodology), **64**, 363-410.

Xia Y., Härdle W., Linton O. (2006). Optimal Smoothing for a Computationally and Statistically Efficient Single-Index Estimator. Disponível em http://personal.lse.ac.uk /lintono/downloads/XHLsihoe.pdf (acessado em 29/10/2006).

Ziegelmann F. A. (2002). *Estimation of Volatility Functions: Nonparametrics and Semi-Parametric Methods*.Tese. Institute of Mathematics and Statistics, University of Kent at Canterbury.

ANEXO 1

Comandos do software R utilizados no exemplo 5.3

```
###################### FUNÇÃO SIMULADA
nerro=50
erro=rep(0,nerro)
for(b in 1:nerro){
set.seed(171+b)
n=200
s <- 1
fi1 <- 2
fi2 <- 2.3
c <- 0.2
e <- rnorm(n,0,s)
y <- rep(0,n)
x1=runif(n,-5,5)
x2=runif(n,-5,5)
y= c+(fi1*(x1^2))+(fi2*(x2^2))+e

####################CÁLCULO DA FUNÇÃO DE VALIDAÇÃO CRUZADA
fr=function(b){
s2 = matrix(data = NA, nrow = n, ncol = n)
t0 = matrix(data = NA, nrow = n, ncol = n)
s1 = matrix(data = NA, nrow = n, ncol = n)
t1 = matrix(data = NA, nrow = n, ncol = n)
s0 = matrix(data = NA, nrow = n, ncol = n)

v=rep(0,n)
for(i in 1:n)
{
 for (j in 1:n){
s2[j,i] = (1-0^abs(i-j))*dnorm(1*x1[j]+b[1]*x2[j],1*x1[i]+b[1]*x2[i],b[2])*
        ((1*x1[j]+b[1]*x2[j]-(1*x1[i]+b[1]*x2[i]))^2)


t0[j,i] = (1-0^abs(i-j))*dnorm(1*x1[j]+b[1]*x2[j],1*x1[i]+b[1]*x2[i],b[2])*y[j]
```

```
s1[j,i] = (1-0^abs(i-j))*dnorm(1*x1[j]+b[1]*x2[j],1*x1[i]+b[1]*x2[i],b[2])*
        (1*x1[j]+b[1]*x2[j]-(1*x1[i]+b[1]*x2[i]))

t1[j,i] = (1-0^abs(i-j))*dnorm(1*x1[j]+b[1]*x2[j],1*x1[i]+b[1]*x2[i],b[2])*
        (1*x1[j]+b[1]*x2[j]-(1*x1[i]+b[1]*x2[i]))*(y[j])

s0[j,i] = (1-0^abs(i-j))*dnorm(1*x1[j]+b[1]*x2[j],1*x1[i]+b[1]*x2[i],b[2])
        }

v[i]= {y[i]-{((sum(s2[,i])*sum(t0[,i])) - (sum(s1[,i]))*sum(t1[,i])))/
            (sum(s2[,i])*sum(s0[,i])-(sum(s1[,i]))^2)}}^2
}
        sum(v)
}

z=optim(c(1,1),fr)
vetor=z$par

########CALCULO DA ESTIMATIVA (G)
a = rep(0,n)
b1=1
b2=vetor[1]
h=vetor[2]
s2 = matrix(data = NA, nrow = n, ncol = n)
t0 = matrix(data = NA, nrow = n, ncol = n)
s1 = matrix(data = NA, nrow = n, ncol = n)
t1 = matrix(data = NA, nrow = n, ncol = n)
s0 = matrix(data = NA, nrow = n, ncol = n)

for(i in 1:n)
{
 for (j in 1:n){
 s2[j,i] = dnorm(b1*x1[j]+b2*x2[j],b1*x1[i]+b2*x2[i],h)*
        ((b1*x1[j]+b2*x2[j]-(b1*x1[i]+b2*x2[i]))^2)

t0[j,i] = dnorm(b1*x1[j]+b2*x2[j],b1*x1[i]+b2*x2[i],h)*y[j]
```

```
s1[j,i] = dnorm(b1*x1[j]+b2*x2[j],b1*x1[i]+b2*x2[i],h)*
        (b1*x1[j]+b2*x2[j]-(b1*x1[i]+b2*x2[i]))

t1[j,i] = dnorm(b1*x1[j]+b2*x2[j],b1*x1[i]+b2*x2[i],h)*
        (b1*x1[j]+b2*x2[j]-b1*x1[i]-b2*x2[i])*(y[j])

s0[j,i] = dnorm(b1*x1[j]+b2*x2[j],b1*x1[i]+b2*x2[i],h)
        }
a[i]= ((sum(s2[,i])*sum(t0[,i])) - (sum(s1[,i])*sum(t1[,i])))/(sum(s2[,i])*sum(s0[,i])-(sum(s1[,i])^2))
}
verificando=cbind(y,a)

##COMPARANDO POPULAÇÃO E AMOSTRA UTILIZANDO A FUNÇÃO PERSP
zi=1*x1+vetor[1]*x2
xp <- seq(-4,4, length= 30)
yp <- seq(-4,4, length= 30)
fa <- approxfun(zi, a)
fa2 <- function(xp,yp) { r <-1*(xp)+vetor[1]*(yp); fa(r) }
za <- outer(xp, yp,fa2)
fy2 <- function(xp,yp) { r= c+(fi1*(xp^2))+(fi2*(yp^2)) }
zy <- outer(xp, yp,fy2)

##############################CALCULO DO EQM
em=matrix(data = NA, nrow = 30, ncol = 30)

for (i in 1:30)
{
for (j in 1:30){
          em[i,j]=(za[i,j]-zy[i,j])^2
         if (is.na(em[i,j])) em[i,j]=0
    }
}
erro[b]=(sum(em))/900

#######################PLOT DAS FUNÇÕES ESTIMADAS, E VERDADEIRA
xp <- seq(-10, 10, length= 30)
yp=xp
```

```
fa <- approxfun(zi, a)
fa2 <- function(xp,yp) { r <-1*xp+vetor[1]*yp; fa(r) }
za <- outer(xp, yp,fa2)
persp(xp, yp, za, theta = 30, phi = 30, expand = 0.5, col = "lightblue",ltheta = 120, shade = 0.75, ticktype = "detailed", xlab = "X", ylab = "Y", zlab = "a")

fy2 <- function(xp,yp) { r <-c+fi1*(xp^2)+fi2*(yp^2); }
zy <- outer(xp, yp,fy2)
persp(xp, yp, zy, theta = 30, phi = 30, expand = 0.5, col = "lightblue",ltheta = 120, shade = 0.75, ticktype = "detailed", xlab = "X", ylab = "Y", zlab = "y")
```