Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Programa de Pós-Graduação em Estatística

Função Variância em Modelos de Regressão Não-Paramétrica: Estimação e Usos

Mestranda: Fernanda Nogueira de Assis
Orientador: Gregorio Saravia Atuncar

Belo Horizonte, Dezembro de 2010

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Programa de Pós-Graduação em Estatística

Função Variância em Modelos de Regressão Não-Paramétrica: Estimação e Usos

Dissertação apresentada ao Programa de Pós-Graduação do Departamento de Estatística da Universidade Federal de Minas Gerais, como requisito para obtenção do título de Mestre em Estatística.

Mestranda: Fernanda Nogueira de Assis
Orientador: Gregorio Saravia Atuncar

Belo Horizonte, Dezembro de 2010

Aos meus pais, Nilton e Líbia, pelo amor e incentivo dedicados sempre.
Por abdicarem de tantas coisas para si próprios para sucesso dos filhos.
Por serem meus exemplos de vida, o exemplo que quero seguir.
E por acreditarem em mim em todos os momentos.

AGRADECIMENTOS

Em primeiro lugar, agradeço a Deus. Obrigada Senhor por me permitir chegar até aqui, pelo dom da perseverança que me foi dado, por ter colocado pessoas tão maravilhosas no meu caminho, simplesmente anjos que foram fundamentais para que eu ingressasse, permanecesse e caminhasse nesta fase.

Agradeço aos meus pais, a quem não há palavras que possam expressar minha gratidão. Obrigada pela dedicação, amor, incentivo, renúncias... Obrigada por serem MEUS PAIS. Me orgulho muito de vocês.

À minha irmã, Danielle, por ter sido alguém em quem pude me espelhar. Por ter tanto me incentivado e me mostrado o que eu poderia alcançar com dedicação, esforço e autoconfiança. Ao meu irmão, Fabrício, que sei que à sua maneira está sempre na torcida pelo meu sucesso. Amo vocês!

Ao meu noivo, Cristiano, por ser meu refúgio. Pela paciência, pelo amor, por entender muitas vezes minha ausência. Por me mostrar, com sua tranquilidade, a força do pensamento positivo. Obrigada por existir na minha vida.

Ao meu orientador, Gregorio, por ser mais que um orientador. Por ser um motivador, por me transmitir tanta sabedoria, por sempre acreditar em mim. Ver seu amor por aquilo que faz, é algo realmente estimulante.

Aos meus amigos do mestrado, sem os quais com toda certeza eu não chegaria até aqui. Em especial, Cris, Fábio, “Carois”, amigos presentes nos estudos, nas descontrações, na torcida. E em particular, Thaize, minha companheira nos fins de semana, fosse para estudos para a prova do mestrado, durante o mestrado, para jogar conversa fora. Amiga, você foi fundamental nesta etapa!

Aos meus grandes amigos, por estarem na minha vida. Malu, obrigada por ser essencial na minha decisão, pelos conhecimentos transmitidos, por ser um exemplo de determinação. Ju e Janete, obrigada pela torcida incondicional! Helen e Lu, obrigada pela companhia nos momentos difíceis. Glauco, Ana, Reje, Renata... não posso citar todos que eu gostaria, mas deixo aqui o meu muito obrigada a todos os amigos e familiares que sempre torceram por mim.

Também não posso deixar de agradecer a Globaltech, que me liberou para as aulas do mestrado e reuniões, e contribuiu muito para meu crescimento profissional. Ao Banco BMG por também me dar condições para que eu finalizasse esta etapa.

Se você quiser alguém em quem confiar,
confie em si mesmo.
Quem acredita sempre alcança!
(Renato Russo)

RESUMO

A técnica de análise de regressão já é bem difundida e utilizada em várias áreas de atuação. Porém, quando as suposições associadas ao modelo de regressão usual não são válidas, ou ainda quando o relacionamento entre as variáveis sob estudo não é linear, muitos se vêm perante uma grande dificuldade. É então que surge a necessidade de abordagens não-paramétricas. Neste trabalho avaliamos com detalhes um método de regressão não-paramétrico e focamos no caso em que a variância não é constante ao longo dos valores observados (modelos heteroscedásticos). Concentramos em um estimador para a variância que foi proposto por Chen, Cheng & Peng (2009), a fim de entender melhor suas propriedades e aplicações. Também estudamos uma medida de adequação do modelo no caso da regressão não-paramétrica que é o coeficiente de determinação proposto em Huang & Chen (2008).

ABSTRACT

The technique of linear regression analysis is widely used in real problems. However, when the assumptions necessary for the model are not satisfied, or when the relationship between the predictor and the dependent variable is not linear, we have serious difficulties in using the model. In such cases the use of nonparametric approaches is more appropriate. In this work, we evaluate the performance of the technique of nonparametric regression using kernel estimator and we study specially the case on which the variance is not constant (heteroscedastic model). We focus on the studying of the estimator for the variance proposed by Chen, Cheng e Peng (2009). We also study a measure to evaluate the quality of the kernel estimator of the regression function. This measure, which is the coefficient of determination, was introduced by Huang e Cheng (2008).

SUMÁRIO

LISTA DE FIGURAS

1. INTRODUÇÃO

Uma análise frequentemente utilizada é o estudo do relacionamento entre duas ou mais variáveis. A técnica conhecida como análise de regressão é empregada com este foco, e através de sua aplicação é encontrada uma equação que representa, de forma funcional e estatística, a relação existente entre as variáveis sob estudo.

Denominamos por variável resposta ou variável dependente, aquela na qual queremos avaliar se outras variáveis, denominadas variáveis preditoras, independentes ou explicativas, têm algum tipo de efeito sobre ela. Representamos por Y a variável reposta e X_j a *j-ésima* variável preditora. A análise de regressão, além de testar a hipótese da existência de relação entre as variáveis, estabelece a direção, o grau e o tipo de relacionamento, e ainda permite a realização de previsões para a variável resposta a partir do conhecimento das variáveis preditoras.

Os modelos usuais de regressão linear são descritos do seguinte modo:

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi} + \varepsilon_i, \qquad i = 1,2,...,n \tag{1.1}$$

sendo β_0 o intercepto e β_j, $j = 1, ..., p$, a variação esperada em Y para cada unidade de variação em X_j, considerando as outras variáveis independentes fixas. Considere aqui o modelo de efeitos aleatórios, ou seja, $X_{1i}, X_{2i}, ..., X_{pi}$, $i = 1,2,...,n$, são variáveis aleatórias com densidade $f(\cdot)$. ε_i é denominado o *i-ésimo* termo de erro, ou seja, a quantidade não explicada pelo modelo.

Estes modelos de regressão são extremamente úteis, mas baseiam-se nas suposições de que os erros do modelo possuem média zero e variância constante, são independentes e seguem a distribuição normal. A regressão não-paramétrica vem como uma alternativa para os casos em que há violação dessas suposições e ainda pode ser utilizada no intuito de descrever qualquer tipo de relação entre as variáveis, e não apenas linear.

Considere dados bivariados $(X_1, Y_1), ..., (X_n, Y_n)$ que formam uma amostra aleatória da população (X, Y). O modelo de regressão não-paramétrica pode ser escrito da seguinte maneira:

$$Y_i = m(X_i) + \sigma(X_i)\varepsilon_i, \tag{1.2}$$

onde $E(\varepsilon_i \mid X_i) = 0$, $Var(\varepsilon_i \mid X_i) = 1$ e X e ε são independentes. A função de regressão e a variância condicional estão representadas, respectivamente, por $m(x) = E(Y \mid X = x)$ e $\sigma^2(x) = Var(Y \mid X = x)$.

Neste trabalho, o interesse principal está nos modelos heteroscedásticos, ou seja, quando a função variância $\sigma^2(X_i)$ não é constante para todos os valores X_i's.

Fan & Yao (1998) propuseram estimar $\sigma^2(x)$ por $\hat{\sigma}^2(x) = \hat{\alpha}_1$ sendo:

$$(\hat{\alpha}_1, \hat{\beta}_1) = \underset{(\alpha_1, \beta_1)}{arg\,min} \sum_{i=1}^{n} \{\hat{r}_i - \alpha_1 - \beta_1 (X_i - x)\}^2 K\left(\frac{X_i - x}{h}\right) \tag{1.3}$$

onde $\hat{r}_i = (Y_i - \hat{m}(X_i))^2$, K é uma função densidade e h , $h > 0$ é chamado de janela, parâmetro que controla a suavização a ser feita no modelo.

O inconveniente deste estimador de variância condicional é que ele não é sempre positivo. Assim, Yu & Jones (2004) sugerem como alternativa trabalhar com $log\,(\sigma^2(x))$ ao invés de $\sigma^2(x)$. Seguindo este raciocínio, Chen, Cheng & Peng (2009) propõem um estimador para a variância $\sigma^2(x)$. Seja a quantidade:

$$log\,(r_i) = \upsilon(X_i) + log(\varepsilon_i^2 / d) \tag{1.4}$$

onde $r_i = (Y_i - m(X_i))^2$, $\upsilon(x) = log(d\sigma^2(x))$ e d satisfaz $E\{log(\varepsilon_i^2 / d)\} = 0$. Baseado nestas equações, estima-se $\upsilon(x)$ por $\hat{\upsilon}(x) = \hat{\alpha}_2$:

$$(\hat{\alpha}_2, \hat{\beta}_2) = \underset{(\alpha_2, \beta_2)}{arg\,min} \sum_{i=1}^{n} \{log(\hat{r}_i + n^{-1}) - \alpha_2 - \beta_2 (X_i - x)\}^2 K\left(\frac{X_i - x}{h}\right) \tag{1.5}$$

O emprego de $log(\hat{r}_i + n^{-1})$ ao invés de $log(\hat{r}_i)$ é para evitar $log(0)$. Uma vez que $E(\varepsilon_i^2 \mid X_i) = 1$ e $r_i = exp\{\upsilon(X_i)\}\varepsilon_i^2 / d$, estima-se d por:

$$\hat{d} = \left[\frac{1}{n}\sum_{i=1}^{n} \hat{r}_i\, exp\{-\hat{\upsilon}(X_i)\}\right]^{-1} \tag{1.6}$$

Veremos posteriormente algumas considerações com respeito a esta constante.

Portanto, o estimador para $\sigma^2(x)$ proposto por Chen, Cheng & Peng (2009) é definido como:

$$\hat{\sigma}^2(x) = exp\{\hat{v}(x)\}/\hat{d} \tag{1.7}$$

Este será o estimador no qual iremos nos concentrar neste trabalho, avaliando suas propriedades, sua adequação e seu emprego em situações reais.

Também será mostrado neste trabalho uma medida de adequação do modelo a ser utilizado, isto é, o coeficiente de determinação no caso da utilização da regressão não-paramétrica. Esta medida foi apresentada em Huang & Chen (2008).

Na Seção 2 serão apresentadas as definições de alguns termos importantes utilizados neste trabalho. Na Seção 3 será definido o procedimento empregado para estimação da função de regressão, $m(x)$, e na Seção 4 o método de estimação da janela. Na Seção 5 abordaremos o coeficiente de determinação no caso da regressão não-paramétrica. Na Seção 6 mostraremos os resultados das simulações feitas e na Seção 7 os resultados das aplicações a dados reais. Na Seção 8 apresentaremos as conclusões com este trabalho e propostas para trabalhos futuros.

2. DEFINIÇÕES PRELIMINARES

Nesta seção apresentamos algumas definições necessárias para acompanhar a metodologia que será descrita no decorrer deste trabalho.

2.1. Notação de ordem

A notação de ordem O e o é definida para função de valores reais em geral. Aqui, iremos restringir nossa atenção apenas para sequências de valores reais e usaremos a notação descrita em Wand & Jones (1995).

Seja a_n e b_n sequências de números reais. Dizemos que a_n é de ordem b_n quando $n \to \infty$, e escrevemos

$$a_n = O(b_n) \text{ quando } n \to \infty, \text{ se e somente se } lim\ sup_{n\to\infty} |a_n / b_n| < \infty .$$

Em outras palavras, $a_n = O(b_n)$ se $|a_n / b_n|$ permanece limitado quando $n \to \infty$.

Dizemos que a_n é de ordem pequena em relação a b_n, e escrevemos

$$a_n = o(b_n) \text{ quando } n \to \infty, \text{ se e somente se } lim_{\ n\to\infty} |a_n / b_n| = 0 .$$

Ainda, se A_n e B_n são duas sequências de valores aleatórios reais, dizemos que:

$A_n = O_P(B_n)$ se para todo $\varepsilon > 0$ existe um λ e M tal que $P(|A_n / B_n| > \lambda) < \varepsilon$, para todo $n > M$.

$A_n = o_P(B_n)$ se para todo $\varepsilon > 0$, $lim_{n\to\infty} P(|A_n / B_n| > \varepsilon) = 0$

2.2. Expansão de Taylor

Uma ferramenta matemática vital para obter aproximações assintóticas quando estamos lidando com núcleo-estimadores é a expansão de Taylor. Lima (1976), por exemplo,

apresenta o teorema a seguir. De acordo com a notação utilizada por Lima (1976), dizemos que $f : I \to \Re$ é p vezes derivável no ponto $x \in I$ quando houver um intervalo aberto J contendo x, tal que f é p-1 vezes derivável em $I \cap J$ e, além disso, existir $f^{(p)}(x)$.

Teorema 1: Seja $f : I \to \Re$ p vezes derivável no ponto $x \in I$. Então, para todo a tal que $x + a \in I$, tem-se

$$f(x+a) = f(x) + f'(x)a + \frac{f''(x)}{2!}a^2 + \ldots + \frac{f^{(p)}(x)}{p!}a^p + o(a^p)$$

O teorema de Taylor permite aproximar os valores de uma função em uma vizinhança de um dado ponto em termos de derivadas de ordem superior naquele ponto, desde que a função seja "suave" o bastante para que estas derivadas existam.

2.3. Estimador $\sqrt{n}$-consistente

A seguinte definição foi extraída de Lehmann & Casella (1998).

Uma sequência de estimadores δ_n é $\sqrt{n}$-consistente para θ se $\sqrt{n}(\delta_n - \theta)$ é limitado em probabilidade, ou seja,

$$\delta_n - \theta = O_P(1/\sqrt{n})$$

3. ESTIMAÇÃO DA FUNÇÃO DE REGRESSÃO

O método adotado para o ajuste da função de regressão é o ajuste polinomial local apresentado em Fan & Gijbels (1995). De acordo com este método, a função de regressão desconhecida $m(x)$ é aproximada localmente por um polinômio de ordem p, e ao contrário dos métodos paramétricos que estimam a função globalmente, a regressão local estima a função $m(x)$ na vizinhança de cada ponto de interesse $x=x_0$. Suponha que a *(p+1)-ésima* derivada de $m(x)$ no ponto x_0 existe. Usando a expansão de Taylor, para x em uma vizinhança de x_0, temos:

$$m(x) \approx m(x_0) + m'(x_0)(x - x_0) + \frac{m''(x_0)}{2!}(x - x_0)^2 + \ldots + \frac{m^{(p)}(x_0)}{p!}(x - x_0)^p \qquad (3.1)$$

Este polinômio é então ajustado localmente pelo método de regressão de mínimos quadrados ponderados e denotamos a solução desse problema por $\hat{\beta}_j$, $j = 0, \ldots, p$:

$$\hat{\beta}_j = \underset{\beta_j}{arg\,mim} \sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j (X_i - x_0)^j \right\}^2 K_h(X_i - x_0) \qquad (3.2)$$

sendo h, a janela que controla o tamanho da vizinhança local e $K_h(\cdot) = \frac{K(\cdot / h)}{h}$, com K uma função núcleo atribuindo pesos a cada ponto. Assumimos que K é uma função densidade de probabilidade simétrica.

Usando a expansão de Taylor em (3.1), obtém-se que $\hat{m}_\nu(x_0) = \nu! \hat{\beta}_\nu$ é um estimador para a ν*-ésima* derivada de m avaliada em x_0, com $\nu = 0, 1, \ldots, p$.

Usando notação matricial, podemos escrever:

$$X = \begin{pmatrix} 1 & (X_1 - x_0) & \cdots & (X_1 - x_0)^p \\ \vdots & \vdots & & \vdots \\ 1 & (X_n - x_0) & \cdots & (X_n - x_0)^p \end{pmatrix} \qquad y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \qquad \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}$$

e denotar *W* a matriz diagonal *n* x *n* de pesos:

$$W = diag\{K_h(X_i - x_0)\} \tag{3.3}$$

O problema de mínimos quadrados ponderados em (3.2) pode então ser escrito da seguinte forma:

$$\min_{\beta}(y - X\beta)^T W(y - X\beta)$$

com $\beta = (\beta_0,...,\beta_p)^T$. O vetor solução é dado por:

$$\hat{\beta} = (X^T WX)^{-1} X^T Wy \tag{3.4}$$

Assim, o estimador $\hat{m}_0(x)$ pode ser expresso como:

$$\hat{m}_0(x) = e_1^T (X^T WX)^{-1} X^T Wy \tag{3.5}$$

onde $e_1 = (1,0,...,0)^T$ é um vetor $(p+1)$ x 1 com o valor 1 na primeira posição.

Pela equação (3.5), é possível observar que para encontrar a estimativa de $m^{(v)}(x)$, é necessário conhecer a matriz *W*, que depende do núcleo *K* e da janela *h*. A escolha da função núcleo *K*, apesar de desempenhar papel fundamental na estimação não-paramétrica, é uma tarefa de menor dimensão que a escolha de *h*. Estudos demonstram que a função núcleo dita ótima, leva a pequenas melhorias em relação à maioria das funções núcleo utilizadas. Consequentemente, a simplicidade e o custo computacional frequentemente determinam a escolha de *K*. Silverman (1986) apresenta uma medida de eficiência para comparar alguns núcleos que são comumente utilizados com o núcleo Epanechnikov, que é o considerado ótimo. O núcleo gaussiano tem uma eficiência em torno de 0,95, sendo que $0 < Eficiência \leq 1$, ou seja, um valor bem razoável. Para maiores detalhes veja Silverman (1986). Assim, durante todo o nosso trabalho adotamos a função núcleo como a função densidade da distribuição normal padrão.

Outra questão relevante no ajuste polinomial local é a escolha da ordem do polinômio. Uma vez que as modelagens do vício e variância estão inicialmente controladas pela janela, esta questão se torna menos crucial. Para uma dada janela *h*, um valor grande de p reduziria o vício, mas resultaria em alta variância e um custo computacional considerável. Estudos demonstram que na estimação de $m^{(\nu)}(x)$, a diferença entre a ordem do polinômio, p, e a ordem da derivada, ν, tem papel importante. Se a diferença $p-\nu$ for par, ou seja, $p=\nu+2k$ e passarmos para uma diferença ímpar $p=\nu+2k+1$, não haverá aumento na variabilidade, mas se estivermos em um caso em que $p-\nu$ for ímpar, ou seja, $p=\nu+2k+1$, e passarmos para um consecutivo caso par, $p=\nu+2k+2$, haverá um preço a ser pago em termos de aumento de variabilidade. Assim, ajustes em que $p-\nu$ sejam de ordem par não são recomendados, e durante todo nosso projeto trabalhamos com ajustes $p-\nu$ de ordem ímpar. De acordo à literatura, recomenda-se o uso da menor ordem ímpar, isto é, $p=\nu+1$, ou ocasionalmente $p=\nu+3$. Deste modo, se, por exemplo, estivermos interessados na estimação de $m^{(0)}(x)$, é conveniente utilizarmos $p=1$ ou $p=3$.

Por conseguinte, nosso problema se concentra na escolha da janela *h*. O parâmetro de suavização possui um efeito fundamental na regressão local, pois possui papel determinante na variabilidade e no vício da estimativa. Se o *h* escolhido for pequeno, a estimativa terá um vício reduzido, mas uma variabilidade elevada. Uma janela $h=0$ resulta basicamente em interpolar os dados (modelo mais complexo). Por outro lado, se o *h* escolhido for grande, a estimativa terá um vício elevado, mas pequena variabilidade. Se $h=+\infty$, a modelagem local se torna uma modelagem global e é o mesmo que a estimativa da regressão linear, o modelo mais simples. Ou seja, a janela controla a complexidade do modelo. Este será um dos pontos iniciais a ser avaliado. Iremos nos basear no método para estimação da janela ótima descrito em Fan & Gijbels (1995), que será descrito na Seção 4.

4. MÉTODO DE ESTIMAÇÃO DA JANELA ÓTIMA

Como já mencionado anteriormente, um bom ajuste da função de regressão $m(x)$ depende crucialmente da janela *h*. Vários métodos para escolha automática da janela já foram estudados criteriosamente. Fan & Gijbels (1995) propuseram um método que se baseia em dois estágios. No primeiro, estima-se uma janela piloto e no segundo estima-se uma janela mais refinada. Esta foi a técnica adotada neste trabalho e que será apresentada com detalhes nesta seção.

4.1. Vício e Variância

Quando estamos lidando com problemas de estimação da janela, um fator chave é ter uma boa percepção do vício e variância dos estimadores, uma vez que um balanceamento entre estas duas quantidades forma o centro de muitos critérios de estimação da janela. O vício e a variância do estimador $\hat{\beta}$ vêm diretamente da expressão em (3.4):

$$\begin{aligned} E\left(\hat{\beta} \mid X_1,...,X_n\right) &= \left(X^T W X\right)^{-1} X^T W m \\ &= \beta + \left(X^T W X\right)^{-1} X^T W r \\ Var\left(\hat{\beta} \mid X_1,...,X_n\right) &= \left(X^T W X\right)^{-1}\left(X^T \Sigma X\right)\left(X^T W X\right)^{-1} \end{aligned} \qquad (4.1)$$

onde: $m = \left\{m(X_1),...,m(X_n)\right\}^T$

$\beta = \left\{m(x_0),...,m^{(p)}(x_0)/\ p!\right\}^T$

$r = m - X\beta$

$\Sigma = diag\left\{K_h^2(X_i - x_0)\sigma^2(X_i)\right\}$

As expressões exatas do vício e variância não são diretamente utilizáveis, pois dependem de quantidades desconhecidas: o resíduo *r* e a matriz diagonal Σ. Assim, há a necessidade de aproximações para o vício e variância.

O teorema seguinte, cujos resultados foram obtidos por Ruppert & Wand (1994), apresenta expansões assintóticas de primeira ordem para o vício e variância do estimador $\hat{m}_\nu(x_0) = \nu!\,\hat{\beta}_\nu$.

Teorema 2: Assuma que $f(x_0) > 0$ e que $f(\cdot)$, $m^{(p+1)}(.)$ e $\sigma^2(\cdot)$ são contínuas em uma vizinhança de x_0. Além disso, assuma que $h \to 0$ e $nh \to \infty$, quando $n \to \infty$. Então, a variância condicional de $\hat{m}_\nu(x_0)$ é dada por:

$$Var\{\hat{m}_\nu(x_0) \mid X_1,...,X_n\} = e_{\nu+1}^T S^{-1} S^* S^{-1} e_{\nu+1} \frac{\nu!^2\,\sigma^2(x_0)}{f(x_0) n h^{1+2\nu}} + o_P\left(\frac{1}{n h^{1+2\nu}}\right) \tag{4.2}$$

O vício condicional assintótico para $p-\nu$ ímpar é dado por:

$$Vício\{\hat{m}_\nu(x_0) \mid X_1,...,X_n\} = e_{\nu+1}^T S^{-1} c_p \frac{\nu!}{(p+1)!} m^{(p+1)}(x_0) h^{p+1-\nu} + o_P\left(h^{p+1-\nu}\right) \tag{4.3}$$

O vício condicional assintótico para $p-\nu$ par é dado por:

$$\begin{aligned} Vício\{\hat{m}_\nu(x_0) \mid X_1,...,X_n\} &= e_{\nu+1}^T S^{-1} \tilde{c}_p \frac{\nu!}{(p+2)!} m^{(p+2)}(x_0) + \\ &= (p+2) m^{(p+1)}(x_0) \frac{f'(x_0)}{f(x_0)} h^{p+2-\nu} + o_P\left(h^{p+2-\nu}\right) \end{aligned} \tag{4.4}$$

onde: $e_{\nu+1} = (0,...,0,1,0,...,0)^T$, com o valor 1 na $(\nu+1)-ésima$ posição

$$\mu_j = \int u^j K(u)\,du$$

$$\lambda_j = \int u^j K^2(u)\,du$$

$$S = \left(\mu_{j+l}\right)_{0 \le j,l \le p} = \begin{bmatrix} \mu_0 & \mu_1 & \cdots & \mu_p \\ \mu_1 & \mu_2 & \cdots & \mu_{p+1} \\ \vdots & \vdots & & \vdots \\ \mu_p & \mu_{p+1} & \cdots & \mu_{2p} \end{bmatrix}$$

$$S^* = \left(\lambda_{j+l}\right)_{0\le j,l\le p} = \begin{bmatrix} \lambda_0 & \lambda_1 & \cdots & \lambda_p \\ \lambda_1 & \lambda_2 & \cdots & \lambda_{p+1} \\ \vdots & \vdots & & \vdots \\ \lambda_p & \lambda_{p+1} & \cdots & \lambda_{2p} \end{bmatrix}$$

$$c_p = \left(\mu_{p+1},...,\mu_{2p+1}\right)^T$$

$$\tilde{c}_p = \left(\mu_{p+2},...,\mu_{2p+2}\right)^T$$

Ressaltamos que as condições impostas neste teorema, bem como em alguns teoremas que serão vistos posteriormente, de $h \to 0$ e $nh \to \infty$ quando $n \to \infty$, vem do fato de que a janela ótima teórica, como veremos na próxima subseção, é da ordem $n^{-1/(2p+3)}$. Assim, ao multiplicar por n e considerar $n \to \infty$, teremos $nh \to \infty$.

Os detalhes da prova do Teorema 2 podem ser vistos em Gomes (2010).

Do resultado apresentado acima, é evidente a diferença teórica entre os casos $p-\nu$ par e $p-\nu$ ímpar. Para $p-\nu$ ímpar, a expressão para o vício assintótico tem uma estrutura mais simples e não envolve o termo $f'(x_0)$. Como já mencionado anteriormente, focaremos apenas no caso de $p-\nu$ ímpar.

4.2. Estimação Ideal da Janela

Um dos critérios utilizados para encontrar a janela local ótima para estimar $m^{(\nu)}(x_0)$, é a minimização do Erro Quadrático Médio (MSE) condicional que é dado por:

$$MSE = \left[Vício\left\{\hat{m}_\nu(x_0) \mid X_1,...,X_n\right\}\right]^2 + Var\left\{\hat{m}_\nu(x_0) \mid X_1,...,X_n\right\}$$

Esta escolha ideal da janela local pode ser aproximada pela janela local ótima assintótica, ou seja, a janela que minimiza o MSE assintótico. Desta forma, utilizando as expressões (4.2) e (4.3), chega-se a seguinte janela local ótima:

$$h_{opt}(x_0) = C_{\nu,p}(K)\left[\frac{\sigma^2(x_0)}{\left\{m^{(p+1)}(x_0)\right\}^2 f(x_0)}\right]^{1/(2p+3)} n^{-1/(2p+3)} \qquad (4.5)$$

sendo: $C_{\nu,p}(K)=\left[\frac{(p+1)!^2(2\nu+1)\int K^2(t)dt}{2(p+1-\nu)\left\{\int t^{p+1}K(t)dt\right\}^2}\right]^{1/(2p+3)}$

Uma simples medida, comumente usada, de perda global é o Erro Quadrático Integrado Médio (MISE) ponderado:

$$MISE=\int\left(\left[\,Vício\left\{\hat{m}_\nu(x_0)\mid X_1,\ldots X_n\right\}\right]^2+Var\left\{\hat{m}_\nu(x_0)\mid X_1,\ldots X_n\right\}\right)w(x)\,dx$$

com w≥0 alguma função peso.

Minimizando essa expressão, chegamos a uma janela teórica constante ótima, e usando novamente as expressões em (4.2) e (4.3), encontramos uma janela constante ótima assintoticamente:

$$h_{opt}=C_{\nu,p}(K)\left[\frac{\int\sigma^2(x)w(x)/\,f(x)\,dx}{\int\left\{m^{(p+1)}(x)\right\}^2w(x)\,dx}\right]^{1/(2p+3)}n^{-1/(2p+3)} \qquad (4.6)$$

Estas janelas ótimas assintóticas dependem de quantidades desconhecidas, como $f(\cdot)$, a variância condicional $\sigma^2(\cdot)$ e a função derivada $m^{(p+1)}(.)$. Desta forma, é preciso buscar procedimentos práticos de estimação da janela. Uma abordagem possível é substituir as quantidades desconhecidas por estimadores pilotos, método conhecido por plug-in. Outra alternativa, recorrendo ao lado prático, é a que será apresentada nas duas próximas subseções.

4.3. Vício e Variância Estimados

Este procedimento de seleção não conta com a complexidade das expressões assintóticas de vício e variância, e seus resultados ficam “próximos” das expressões exatas para o vício e variância.

O vício condicional $\left(X^T WX\right)^{-1} X^T Wr$ em (4.1) contém a quantidade desconhecida $r = m - X\beta$. Usando expansão de Taylor de ordem *a*, este vício condicional pode ser aproximado por $\left(X^T WX\right)^{-1} X^T W\tau$, onde τ é um vetor *n x 1* com o *i-ésimo* elemento:

$$\tau_i = \beta_{p+1}\left(X_i - x_0\right)^{p+1} + ... + \beta_{p+a}\left(X_i - x_0\right)^{p+a}$$

A escolha de *a* tem uma influência na performance do seletor de janela resultante. Suponha que foi utilizado o ajuste linear local (*p = 1*). Neste caso, se *a* for igual a 3, o seletor de janela resultante é $\sqrt{n}$-consistente. Entretanto, a escolha de *a = 2* reduz os custos computacionais e resulta em um procedimento de estimação da janela que não é muito longe de ser $\sqrt{n}$ - consistente. Assim, neste trabalho foi adotado *a = 2*.

Desta forma, o vício aproximado pode ser escrito como:

$$Vício\left(\hat{\beta} \mid X_1,...,X_n\right) = \left(X^T WX\right)^{-1} X^T W\tau = S_n^{-1}\begin{pmatrix} \beta_{p+1}S_{n,p+1} + \beta_{p+2}S_{n,p+2} \\ \vdots \\ \beta_{p+1}S_{n,2p+1} + \beta_{p+2}S_{n,2p+2} \end{pmatrix} \qquad (4.7)$$

e pode ser estimado por:

$$\hat{Vício}\left(\hat{\beta} \mid X_1,...,X_n\right) = S_n^{-1}\begin{pmatrix} \hat{\beta}_{p+1}S_{n,p+1} + \hat{\beta}_{p+2}S_{n,p+2} \\ \vdots \\ \hat{\beta}_{p+1}S_{n,2p+1} + \hat{\beta}_{p+2}S_{n,2p+2} \end{pmatrix} \qquad (4.8)$$

onde $S_{n,j} = \sum_{i=1}^{n} K_h\left(X_i - x_0\right)\left(X_i - x_0\right)^j$, $S_n \equiv X^T WX$ é a matriz *(p+1)* x *(p+1)* $\left(S_{n,j+l}\right)_{0\le j,l\le p}$ e $\hat{\beta}_{p+1}$ e $\hat{\beta}_{p+2}$ são os coeficientes de regressão estimados obtidos pelo ajuste de um polinômio de grau *p+2*, localmente.

Agora, aproxima-se a variância condicional em (4.1) usando homocedasticidade local:

$$Var\left(\hat{\beta} \mid X_1,...,X_n\right) = \left(X^T WX\right)^{-1}\left(X^T W^2 X\right)\left(X^T WX\right)^{-1}\sigma^2\left(x_0\right) \qquad (4.9)$$

A quantidade desconhecida $\sigma^2(x_0)$ pode ser estimada pela soma de resíduos quadrados ponderada pelo núcleo e normalizada, a partir do ajuste polinomial de ordem ($p+2$) usando uma janela piloto h (a escolha dessa janela piloto será discutida na subseção 4.4):

$$\hat{\sigma}^2(x_0) = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 K_h(X_i - x_0)}{tr\left\{W - WX\left(X^T WX\right)^{-1} X^T W\right\}} \tag{4.10}$$

De acordo com Fan & Gijbels (1996), o vício deste estimador é da ordem h^{p+a+1}, mas este vício pode ser considerado desprezível quando a = 2 e a janela piloto não for muito grande.

Assim, um estimador para a variância condicional será dado por:

$$\hat{Var}\left(\hat{\beta} \mid X_1,...,X_n\right) = \left(X^T WX\right)^{-1}\left(X^T W^2 X\right)\left(X^T WX\right)^{-1}\hat{\sigma}^2(x_0) \tag{4.11}$$

Com o vício e variância estimados, obtêm-se o seguinte estimador para o erro quadrático médio de $\hat{\beta}_\nu = \hat{m}_\nu(x_0)/\nu!$:

$$\hat{MSE}_{\nu,p}(x_0;h) = \hat{b}^2_{\nu,p}(x_0) + \hat{V}_{\nu,p}(x_0) \tag{4.12}$$

onde $\hat{b}^2_{\nu,p}(x_0)$ denota o $(\nu+1)$-*ésimo* elemento do vetor de vício estimado em (4.8) e $\hat{V}_{\nu,p}(x_0)$ denota o $(\nu+1)$-*ésimo* elemento da diagonal da matriz em (4.11).

4.4. Critério Residual Quadrático

O objetivo é chegar a uma estatística cuja minimização leva a um estimador para a janela ótima. Considere o Critério Residual Quadrático (RSC) definido do seguinte modo:

$$RSC(x_0;h) = \hat{\sigma}^2(x_0)\left\{1 + (p+1)V\right\} \tag{4.13}$$

onde $\hat{\sigma}^2(\cdot)$ é dado pela expressão (4.10), e *V* é primeiro elemento da diagonal da matriz $(X^T W X)^{-1}(X^T W^2 X)(X^T W X)^{-1}$.

Intuitivamente podemos entender o RSC da seguinte maneira: quando a janela *h* é muito grande, o polinômio não se ajusta bem, o vício é grande e então a soma de resíduos quadrados $\hat{\sigma}^2(x_0)$ também é. Quando a janela *h* é muito pequena, a variância do ajuste será grande e consequentemente *V* também será. Como ambos os fatores $\hat{\sigma}^2(x_0)$ e *V* estão incorporados no RSC, esta quantidade irá prevenir ambas as escolhas extremas da janela.

O Teorema 3, estabelecido em Fan & Gijbels (1995), fornece uma justificativa teórica para a quantidade RSC.

Teorema 3: Suponha que $\sigma^2(x) = \sigma^2(x_0)$ em uma vizinhança de x_0. Se $h \to 0$ e $nh \to \infty$ quando $n \to \infty$, então,

$$E\{RSC(x_0;h) \mid X_1,...,X_n\} = \sigma^2(x_0) + C_p \beta_{p+1}^2 h_n^{2p+2} + (p+1)\, a_0 \frac{\sigma^2(x_0)}{n h_n f(x_0)} + o_P\left\{h_n^{2p+2} + (nh_n)^{-1}\right\}$$

onde a_0 denota o primeiro elemento da diagonal da matriz $S^{-1} S^* S^{-1}$ e $C_p = \mu_{2p+2} - c_p^T S^{-1} c_p$ com $c_p = (\mu_{p+1},...,\mu_{2p+1})^T$.

A prova deste teorema está disponível no Apêndice deste trabalho. Em Fan & Gijbels (1996) algumas sugestões para a realização desta prova foram fornecidas, mas seu desenvolvimento como um todo foi feito no decorrer deste trabalho.

O resultado do Teorema 3 revela que a minimização da $E\{RSC(x_0;h) \mid X_1,...,X_n\}$ é aproximadamente igual a:

$$h_0(x_0) = \left\{ \frac{a_0 \sigma^2(x_0)}{2C_p \beta_{p+1}^2 n f(x_0)} \right\}^{1/(2p+3)} \tag{4.14}$$

Comparando com a janela ótima assintótica em (4.5), encontra-se a relação:

$$h_{opt}(x_0) = adj_{\nu,p} h_0(x_0) \tag{4.15}$$

sendo:

$$adj_{\nu,p} = \left[\frac{(2\nu+1) C_p \int K^2(t)\,dt}{(p+1-\nu)\left\{\int t^{p+1} K(t)\,dt\right\}^2 \int K^2(t)dt} \right]^{1/(2p+3)} \tag{4.16}$$

Note que esta constante depende somente da função núcleo *K*.

Suponha que nosso interesse seja estimar $m^{(\nu)}(.)$ no intervalo [a,b] usando um ajuste polinomial de ordem *p*. Assim, encontra-se a janela $\hat{h}$ minimizando a versão integrada do RSC:

$$IRSC(h) = \int_{[a,b]} RSC(y;h)dy \tag{4.17}$$

e obtêm-se o seletor de janela constante RSC:

$$\hat{h}_{\nu,p}^{RSC} = adj_{\nu,p}\hat{h} \tag{4.18}$$

A utilização desta janela apresenta boa performance, mas uma taxa de convergência lenta. Assim sendo, Fan & Gijbels (1995) propõem utilizar esta janela em um primeiro estágio e refiná-la em um segundo estágio (para ver alguns exemplos de comparação, veja Miranda (2007)). O procedimento final para a estimação da janela é então feito da seguinte forma:

Estimação piloto: Neste 1º estágio, ajustamos um polinômio de ordem *p+2*, usamos a janela $\hat{h}_{p+1,p+2}^{RSC}$ em (4.18) como a janela piloto e obtemos as estimativas $\hat{\beta}_{p+1}$, $\hat{\beta}_{p+2}$ e $\hat{\sigma}^2(x_0)$.

Seletor da janela: Este é o 2º estágio, no qual encontramos a janela que minimiza o erro quadrático médio integrado estimado:

$$\hat{h}_{\nu,p}^{R} = arg\min_{h} \int_{[a,b]} M\hat{S}E_{\nu,p}(y;h)dy$$

e usamos este seletor de janela refinado para ajustar um polinômio de ordem *p*, onde $M\hat{S}E_{\nu,p}(y;h)$ é dado como em (4.12).

Ressaltamos que neste trabalho utilizamos $p=1$.

5. COEFICIENTE DE DETERMINAÇÃO

Como já mencionado anteriormente, métodos de regressão não-paramétrica, como o ajuste polinomial local, são amplamente utilizados para explorar tendências desconhecidas na análise de dados. Dada esta popularidade, ferramentas de análise de variância, análogas às da regressão linear, seriam de grande utilidade para melhor interpretar as curvas não-paramétricas. Baseado nisso, Huang & Chen (2008) desenvolveram inferências de análise de variância para a regressão não-paramétrica. Um dos pontos abordados, que consideramos neste trabalho, foi o coeficiente de determinação, que é uma medida da proporção da variabilidade na variável resposta que é explicada pela sua relação com a(s) variável(is) preditora(s).

Os autores propuseram, além de um coeficiente de determinação global, um coeficiente local, que avalia pontualmente quanto da variação na resposta está sendo explicada pelo modelo. A idéia do coeficiente de determinação é comparar quanto o modelo de regressão está explicando com a variação total que se tem no modelo. Se esta razão for alta, lembrando sua faixa de variação que é de 0 a 1, indica que aquele modelo está adequado no sentido de responder por grande parte da variação que se tem na variável resposta.

Desta forma, para a construção do coeficiente de determinação, denotado por R^2, é necessário definir algumas quantidades. A mensuração da variabilidade dos dados ajustados pelo modelo é feita pela Soma de Quadrados devida à Regressão (SSR), definida da seguinte forma:

$$SSR_p(x;h) = \frac{n^{-1}\sum_{i=1}^{n}\left[\left(\sum_{j=0}^{p}\beta_j(X_i - x)^j\right) - \bar{Y}\right]^2 K_h(X_i - x)}{\hat{f}(x;h)} \qquad (5.1)$$

sendo: $\bar{Y} = \sum_{i=1}^{n} Y_i \Big/ n$

A variação total do modelo, representada pela Soma de Quadrados Total (SST), é medida pela soma de quadrados exata para os $Y_i's$:

$$SST(x;h) = \frac{n^{-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 K_h(X_i - x)}{\hat{f}(x;h)} \tag{5.2}$$

E no caso da regressão não-paramétrica, a seguinte decomposição é válida a partir do ajuste polinomial local em um ponto x_0 na amplitude de variação dos X_i's:

$$SST(x;h) = SSE_p(x;h) + SSR_p(x;h) \tag{5.3}$$

onde a Soma de Quadrados devida ao Erro (SSE) é uma medida da variabilidade que não foi explicada pelo modelo de regressão, e é definida do seguinte modo:

$$SSE_p(x;h) = \frac{n^{-1}\sum_{i=1}^{n}\left(Y_i - \sum_{j=0}^{p}\hat{\beta}_j (X_i - x)^j\right)^2 K_h(X_i - x)}{\hat{f}(x;h)} \tag{5.4}$$

Baseado em (5.3), define-se o R^2 local:

$$R_p^2(x;h) = 1 - \frac{SSE_p(x;h)}{SST(x;h)} = \frac{SSR_p(x;h)}{SST(x;h)} \tag{5.5}$$

O R^2 global é definido a partir da integração das correspondentes expressões em (5.3):

$$\begin{aligned} SST(h) &= \int SST(x;h)\hat{f}(x;h)dx \\ SSE_p(h) &= \int SSE_p(x;h)\hat{f}(x;h)dx \\ SSR_p(h) &= \int SSR_p(x;h)\hat{f}(x;h)dx \end{aligned} \tag{5.6}$$

E assim:

$$R_p^2(h) = 1 - \frac{SSE_p(h)}{SST(h)} = \frac{SSR_p(h)}{SST(h)} \tag{5.7}$$

6. SIMULAÇÕES

Nesta seção iremos apresentar as simulações realizadas durante este trabalho. Na subseção 6.1 foram feitas simulações para avaliar o estimador da variância proposto. Na subseção 6.2 tentamos introduzir uma contribuição da variância, nos casos heteroscedásticos, na matriz de pesos utilizada na estimação de β, objetivando aperfeiçoar a estimação da função de regressão. Na subseção 6.3, realizamos simulações a fim de entender melhor e avaliar o desempenho do coeficiente de determinação proposto para a regressão não-paramétrica.

6.1. Estimador da Variância

Nosso objetivo nesta subseção é avaliar o estimador de variância apresentado em (1.7). Considere o seguinte modelo de regressão, o mesmo utilizado no estudo numérico de Chen, Cheng & Peng (2009):

$$Y_i = 0{,}5\left\{X_i + 2\exp\left(-16X_i^2\right)\right\} + \sigma(X_i)\varepsilon_i \tag{6.1}$$

sendo:

$$\sigma(X_i) = 0{,}4exp\left(-2X_i^2\right) + 0{,}2 \tag{6.2}$$

Foi considerado X_i ~*Uniforme [-2,2]*. ε_i é independente de X_i e segue a distribuição *Normal (0,1).*

A título de ilustração, apresentamos na Figura 6.1 o gráfico de x versus y gerados a partir de uma simulação da função (6.1) com a curva da função de regressão teórica e a estimativa da função de regressão, conforme método descrito na Seção 3. Na Figura 6.2 exibimos o gráfico do desvio-padrão apresentado em (6.2), com sua estimativa conforme a expressão (1.7).

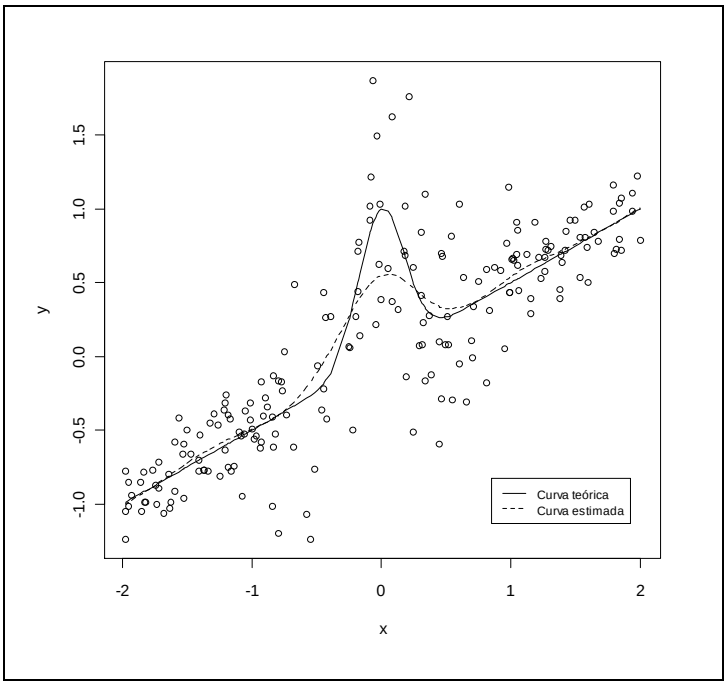


Figura 6.1: Valores de x versus y gerados a partir de uma simulação da função (6.1) com curva teórica e estimada

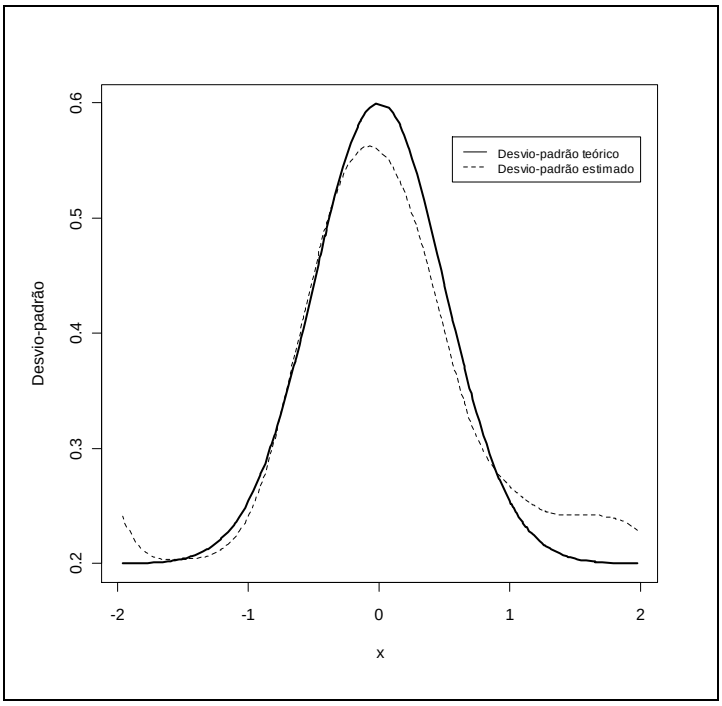


Figura 6.2: Função desvio-padrão em (6.2) com estimativa

Para avaliarmos melhor a performance do estimador $\hat{\sigma}(\cdot)$ de $\sigma(\cdot)$, foram geradas 1000 amostras de tamanho *n* = 200 do modelo em (6.1). O método para a estimação da janela utilizado foi o descrito na Seção 4. A performance foi medida pelo Erro Médio de Desvio

Absoluto (MADE) e pelo Erro Médio de Desvio Quadrático (MSDE):

$$MADE(\hat{\sigma}) = \frac{1}{n}\sum_{i=1}^{n} |\hat{\sigma}(x_i) - \sigma(x_i)| \qquad MSDE(\hat{\sigma}) = \frac{1}{n}\sum_{i=1}^{n} \{\hat{\sigma}(x_i) - \sigma(x_i)\}^2$$

A seguir são apresentados os digramas de caixa das quantidades MADE e MSDE resultantes das 1000 simulações.

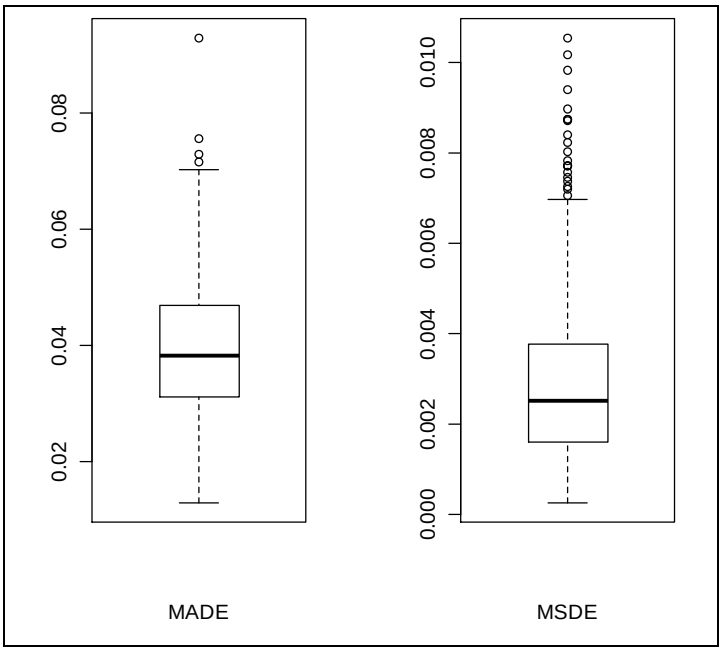


Figura 6.3: Diagramas de caixa das quantidades MADE e MSDE
com respeito ao modelo em (6.1)

Como podemos observar na Figura 6.3, o estimador de variância proposto em (1.7) apresenta resultados satisfatórios. Os valores do MADE estão concentrados principalmente em torno de 0,04 e os do MSDE em torno de 0,003. Isto fornece indícios de adequação desse estimador, sugerindo que podemos aplicá-lo a análise de dados reais.

Ainda em relação a este estimador, fizemos algumas análises a fim de compreender melhor a constante $\hat{d}$ apresentada em (1.6):

$$\hat{d} = \left[\frac{1}{n}\sum_{i=1}^{n} \hat{r}_i \, exp\{-\hat{v}(X_i)\}\right]^{-1}$$

Para o estimador da variância proposto, d é tal que $E\left\{log\left(\varepsilon_i^2 / d\right)\right\} = 0$. Ou seja,

teoricamente:

$$
\begin{aligned}
&E\left\{log\left(\varepsilon_i^2 / d\right)\right\} = 0 \\
&E\left\{log\left(\varepsilon_i^2\right) - log\left(d\right)\right\} = 0 \\
&E\left\{log\left(\varepsilon_i^2\right)\right\} = E\left\{log\left(d\right)\right\} \\
&E\left\{log\left(\varepsilon_i^2\right)\right\} = log\left(d\right)
\end{aligned}
$$

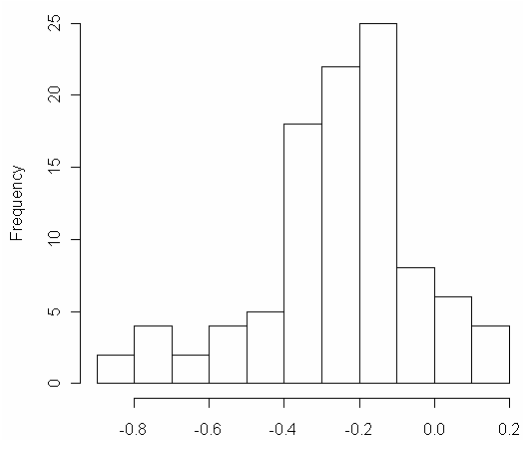
Ainda, uma vez que $E\left(\varepsilon_i^2 \mid X_i\right) = 1$ e $r_i = exp\{\upsilon(X_i)\}\varepsilon_i^2 / d$, podemos escrever:

$$
E\left(r_i \mid X_i\right) = \frac{exp\{\upsilon(X_i)\}E\left(\varepsilon_i^2 \mid X_i\right)}{d}
$$

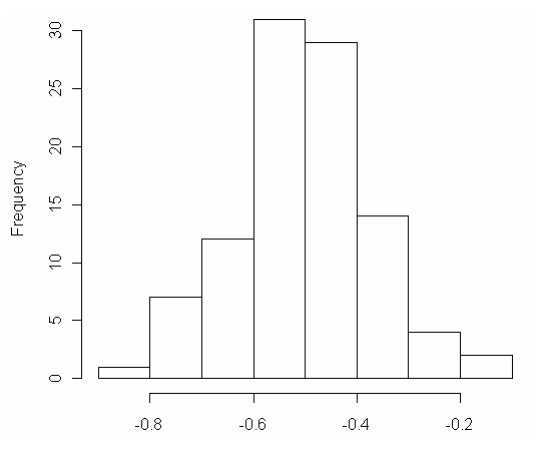
$$
r_i = \frac{exp\{\upsilon(X_i)\}}{d} \quad \Rightarrow \quad d = \frac{exp\{\upsilon(X_i)\}}{r_i}
$$

Percebemos então que a estimativa de *d* apresentada se baseia na média harmônica. Chen, Cheng & Peng (2009) declararam que foi devido a evidências de que dados financeiros apresentam caudas mais pesadas do que a distribuição normal que eles propuseram o estimador para a variância apresentado em (1.7). Assim, apesar de não termos encontrado nenhuma evidência teórica, associamos o fato da utilização da média harmônica às caudas pesadas. Tentamos utilizar as médias aritmética ou geométrica, mas estas não apresentaram bons resultados. Fizemos algumas simulações para $\hat{d}$ para diferentes tamanhos de amostras e utilizando o modelo em (6.1). Os histogramas a seguir apresentam os resultados obtidos para o $log\left(\hat{d}\right)$.

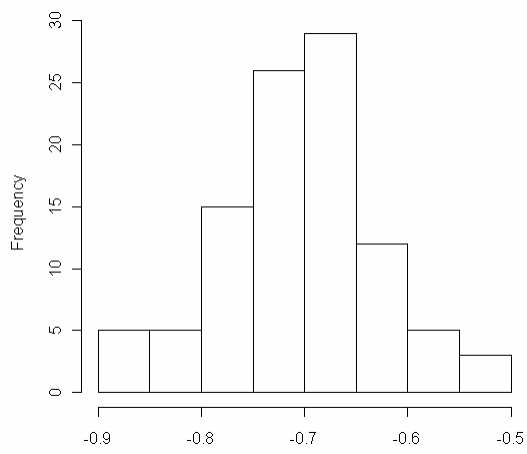
100 amostras de tamanho *n* = 50



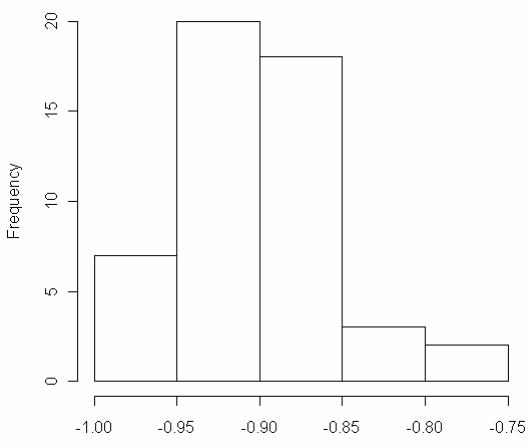
100 amostras de tamanho *n* = 100



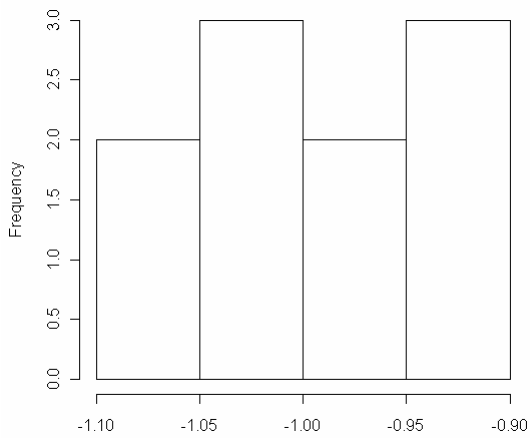
100 amostras de tamanho n = 200



50 amostras de tamanho n = 500



10 amostras de tamanho n = 1000*



*Foram feitas apenas 10 simulações devido ao custo computacional, mas ainda assim foi apresentado o histograma para visualização do deslocamento de $log(\hat{d})$.

Figura 6.4: Histogramas dos valores de $log(\hat{d})$ obtidos nas simulações

A partir da geração de uma amostra aleatória *Normal (0,1)* de tamanho 5000, elevando os valores ao quadrado, tomando o *log* e finalmente calculando a média destes valores, vimos que $E\{log(\varepsilon_i^2)\}$ é aproximadamente igual a -1,28. Pela análise da Figura 6.4, vemos que para amostras pequenas, os valores de $log(\hat{d})$ estão bem distantes desta média. À medida que aumentamos o tamanho da amostra, $log(\hat{d})$ foi se aproximando deste valor. Para um tamanho de amostra n = 5000, devido ao custo computacional, fizemos apenas duas

simulações, e os valores encontrados de $log(\hat{d})$ foram -1.154782 e -1.143393, valores bem mais próximos do real.

As conclusões que podemos chegar dessas simulações, é que a suposição para d só é válida assintoticamente. O que é curioso, é que mesmo sabendo desse resultado, vimos pela Figura 6.3 que para uma amostra de tamanho n = 200, o estimador final para a variância se comportou muito bem. Ou seja, aparentemente o desempenho de $\hat{\sigma}^2(x)$ não depende fortemente de $\hat{d}$.

6.2. Matriz de pesos para o caso heteroscedástico

Uma questão levantada no decorrer deste trabalho foi em relação à matriz de pesos utilizada na estimação de β. A matriz W, como vista em (3.3), depende apenas do núcleo K empregado. Foi questionado se, no caso dos modelos heteroscedásticos, essa matriz não deveria ter também uma contribuição da variância. Consideramos essa observação pertinente na tentativa de aperfeiçoar a estimação da função de regressão, e com este intuito, realizamos algumas simulações.

Inicialmente estudamos como deveríamos incorporar essa contribuição da variância no estimador. Nos baseamos no Método de Mínimos Quadrados Ponderados (Montgomery & Peck (1992)). A matriz de pesos será então calculada do seguinte modo:

$$W^* = \begin{bmatrix} \frac{K_h(X_1 - x_0)}{\sigma^2(x_0)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{K_h(X_n - x_0)}{\sigma^2(x_0)} \end{bmatrix} \tag{6.3}$$

Ou seja, dividiremos cada peso inicial pela variância local. Desta forma, observações com variância alta terão um peso menor que observações com variância baixa.

Fizemos algumas simulações utilizando o modelo apresentado em (6.1) e estimando a função de regressão de duas formas: com o β sendo estimado pelo modo usual e com o β sendo estimado usando a matriz de pesos W^* apresentada em (6.3). A Figura 6.5 apresenta este resultado.

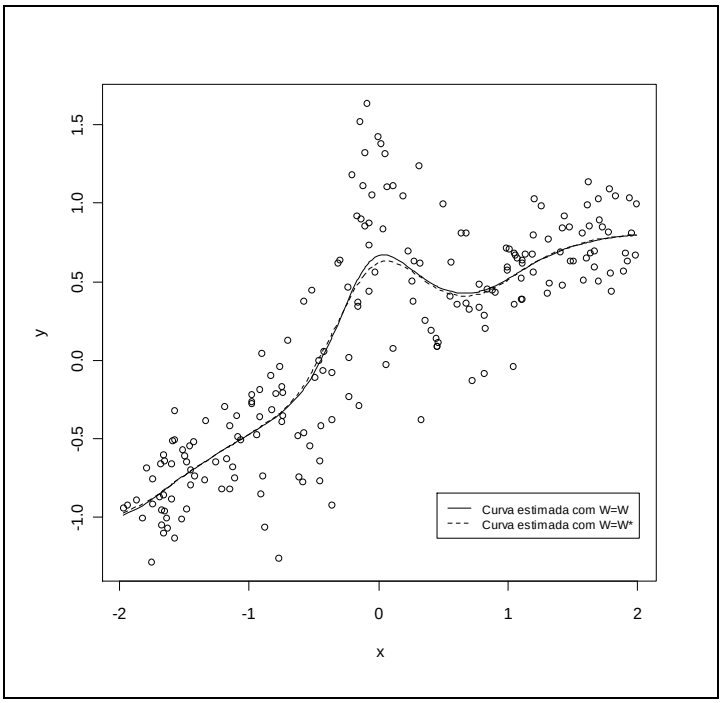


Figura 6.5: Valores de x versus y gerados a partir de uma simulação da função (6.1) com estimativas da função de regressão utilizando diferentes matrizes de pesos

Pela Figura 6.5, não é possível identificar nenhuma melhoria expressiva quando utilizamos a matriz W^*. Acreditamos que isso se deve ao fato de que o procedimento utilizado para estimação da janela já leva em consideração a heteroscedasticidade. Como observamos pela expressão (4.13) que é utilizada no 1º estágio, e (4.11), utilizada no 2º estágio, ambas levam em consideração a variância local.

6.3. Desempenho do Coeficiente de Determinação

Nosso objetivo nesta subseção é avaliar o desempenho do coeficiente de determinação exibido na Seção 5. Inicialmente utilizamos novamente o modelo em (6.1) e calculamos o coeficiente de determinação local e global para este caso. O coeficiente de determinação global foi de 60,63%. A Figura 6.6 exibe o coeficiente de determinação local.

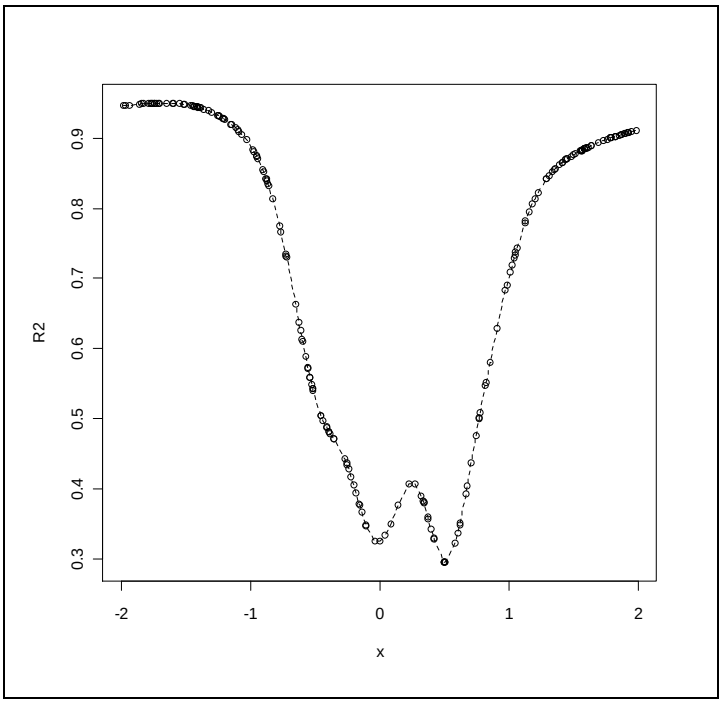


Figura 6.6: Coeficiente de determinação local para uma simulação do modelo (6.1)

Ao avaliarmos a Figura 6.6, notamos que o gráfico se encontra coerente, uma vez que se avaliarmos a Figura 6.1, vemos que onde há uma maior dispersão dos dados, ou seja, maior variância, é em torno do valor zero, o que realmente acarreta em um menor coeficiente de determinação. Porém, como este gráfico é construído ponto a ponto, o menor coeficiente de determinação deveria ser exatamente para o valor de x igual a zero, no entanto, não é isto que ocorre. Devido a esta inconsistência, construímos um gráfico do desvio-padrão versus o coeficiente de determinação para avaliar o comportamento. A Figura 6.7 ilustra esta situação.

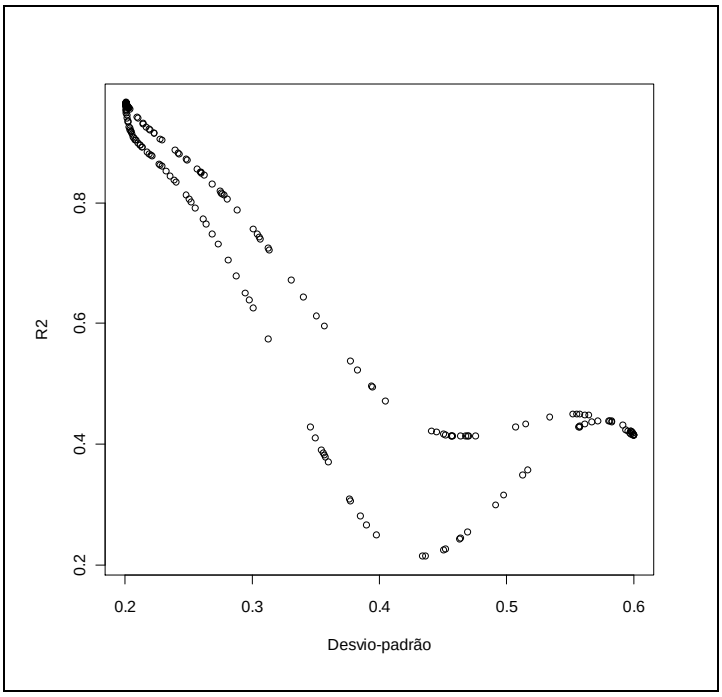


Figura 6.7: Desvio-padrão versus coeficiente de determinação local para uma simulação do modelo (6.1)

A partir da Figura 6.7, é possível observar duas curvas distintas neste gráfico. Elas apresentam o mesmo comportamento, porém há uma discrepância entre elas que se amplia à medida que o desvio-padrão aumenta, mas que torna a diminuir para valores altos de desvio-padrão. A identificação destas duas curvas se deve ao fato de que a função desvio-padrão que estamos utilizando (expressão (6.2)) é simétrica, como visualizada na Figura 6.2. Portanto, esta é a diferença na estimação dos dados de um lado e outro da curva a partir do zero. Novamente, onde o desvio-padrão é mais alto, há uma incoerência, pois é neste caso que o coeficiente de determinação deveria ser menor.

Devido aos resultados encontrados, utilizamos também um modelo homocedástico para avaliar como o coeficiente de determinação iria se comportar. Foi utilizado o seguinte modelo:

$$Y_i = 6\,sen(2X_i) + 2\,exp(-16X_i^2) + \sigma\varepsilon_i \qquad (6.4)$$

com $\sigma = 0{,}3$, X_i ~*Uniforme [-2,2] e* ε_i ~ *Normal (0,1)*.

A Figura 6.8 exibe a o gráfico de x versus y gerados a partir de uma simulação da função (6.4) com a estimativa da função de regressão e a Figura 6.9 o coeficiente de determinação local para este caso.

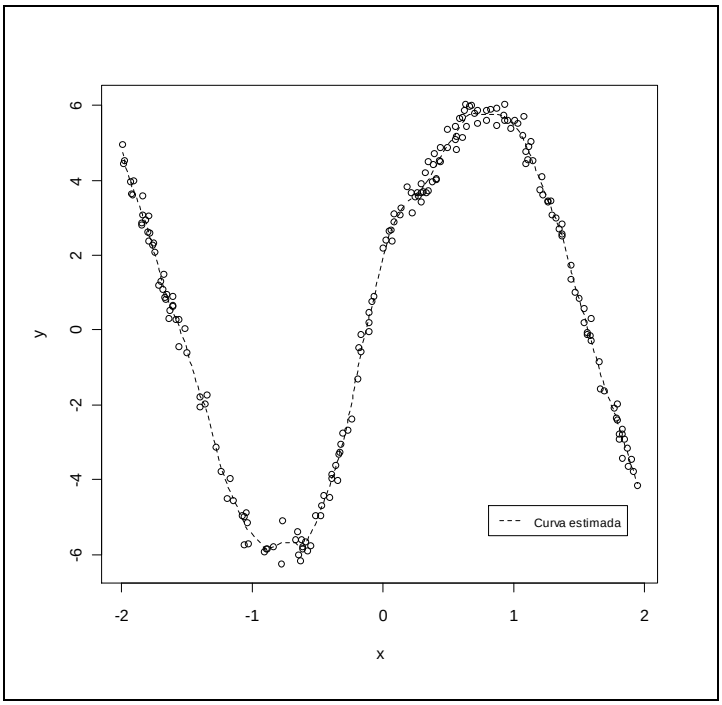


Figura 6.8: Valores de x versus y gerados a partir de uma simulação da função (6.4) com estimativa da função de regressão

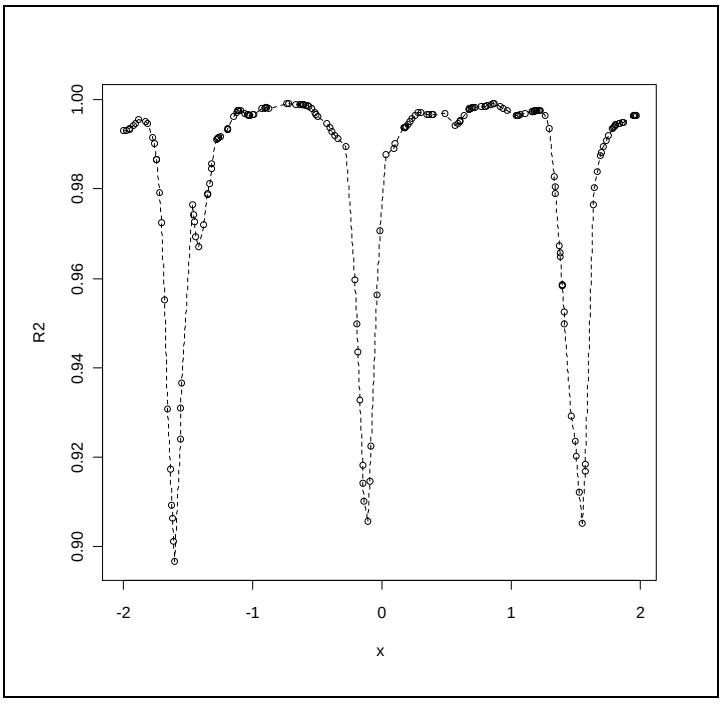


Figura 6.9: Coeficiente de determinação local para uma simulação do modelo (6.4)

Observando a Figura 6.9, podemos concluir que no caso homocedástico, o coeficiente de determinação se comporta muito bem. Por se tratar de um modelo homocedástico, era esperado que o valor de R^2 se mantivesse razoavelmente constante ao longo dos valores observados. E apesar da Figura 6.9 apresentar oscilações, se observarmos a escala, vemos que a variação é muito pequena, de 0,90 a 0,99. Talvez essas oscilações ocorram onde há uma frequência menor de dados.

O que podemos concluir com estas simulações é que o coeficiente de determinação local se comporta bem nos modelos homocedásticos, mas nos modelos heteroscedásticos existem algumas incoerências. Essas incoerências foram observadas principalmente onde a variância é maior, locais em que ocorrem algumas oscilações que não conseguimos explicar. O que suspeitamos fortemente é que a densidade conjunta $f(x,y)$ afeta de alguma forma o coeficiente de determinação. Visualizamos, nestas simulações, que em regiões de menor concentração de observações, eram onde ocorriam oscilações no R^2 local. Porém, neste trabalho não foi possível identificar de que forma isto acontece, e este é um ponto que deixamos em aberto para trabalhos futuros.

7. APLICAÇÕES

7.1. Exemplo 1

Neste primeiro exemplo de aplicação, apresentamos dados de Simonoff (1996) que estão disponíveis na página do autor na internet, onde ele disponibiliza os arquivos de dados utilizados em seus livros. Os dados são referentes a um vinhedo em Chateau. Este vinhedo é dividido em 52 linhas de produção, e as 52 observações no conjunto de dados correspondem ao rendimento das colheitas nos anos de 1989, 1990 e 1991, medido pelo número total de cestas. As cestas são utilizadas para transportar as uvas logo após a colheita, e contêm cerca de 14 quilos de uva. Os números das linhas estão ordenados, com o aumento do número da linha refletindo movimento do noroeste para o sudeste. Linhas de 31-52 são menores que linhas 1-30 (90 metros, aproximadamente, versus 110). O arquivo contém o número da linha de produção e o número de cestas para a safra de cada ano sob estudo.

Inicialmente fizemos um gráfico do rendimento total nos três anos, ou seja, o número total de cestas, como uma função da linha de produção. Logo após, ajustamos a curva de regressão não-paramétrica baseada no método descrito na Seção 3, com a janela tendo sido selecionada pelo procedimento descrito na Seção 4. A Figura 7.1 ilustra essa aplicação.

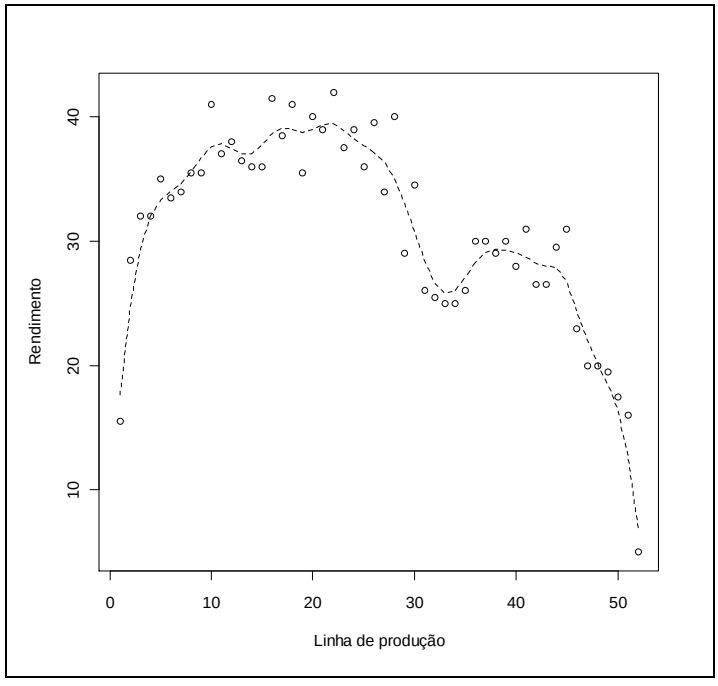


Figura 7.1: Gráfico de dispersão da linha de produção versus o rendimento de um vinhedo com a curva do ajuste polinomial local

Ao analisar a Figura 7.1, podemos verificar que a curva de regressão não-paramétrica se ajustou muito bem aos dados. Isto é uma evidência a favor tanto do método de regressão polinomial local, como do método de estimação da janela adotado.

A Figura 7.2 exibe o desvio-padrão estimado para estes dados e a Figura 7.3 apresenta o coeficiente de determinação local.

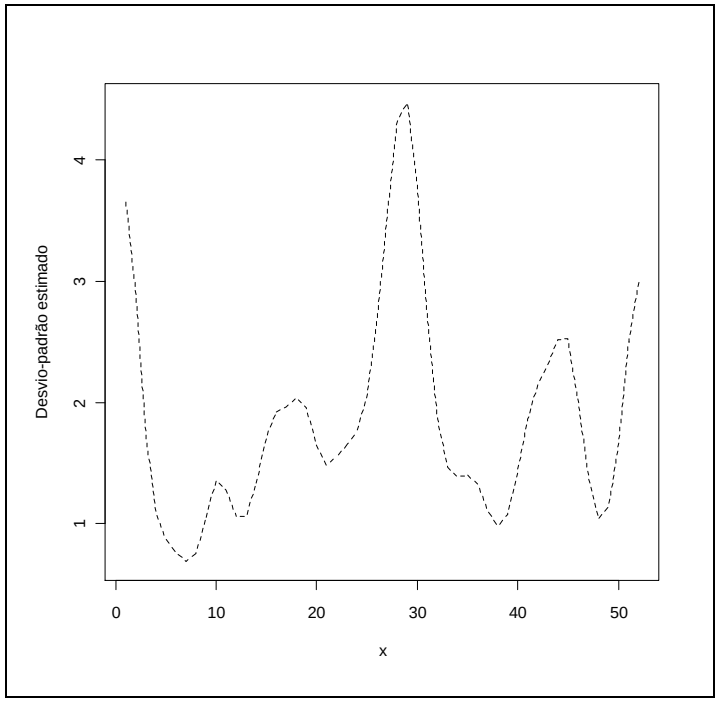


Figura 7.2: Desvio-padrão estimado para os dados do vinhedo

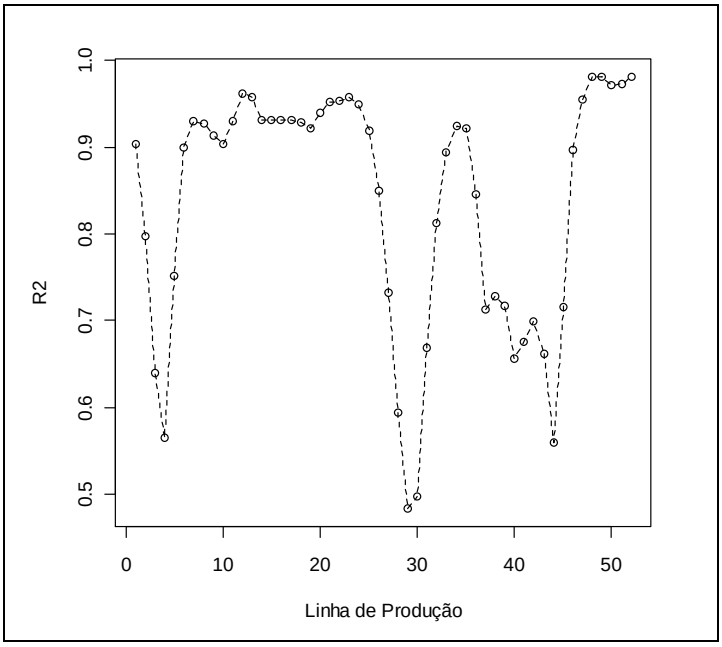


Figura 7.3: Coeficiente de determinação local para os dados do vinhedo

O coeficiente de determinação local mais baixo foi em torno de 50% para as linhas 29-30. Para muitas linhas, como de 7-25, o coeficiente de determinação foi acima de 90%. Isto reflete diferença na proporção de variação explicada pela regressão polinomial local. Se avaliarmos a Figura 7.3 em conjunto com a Figura 7.1, verificamos que, como levantado na subseção 6.3, neste exemplo também houve algumas contradições com o coeficiente de determinação local, no sentido de que avaliando ponto a ponto, há alguns que, por estarem mais afastados da curva de regressão estimada, deveriam apresentar menor coeficiente de determinação. Porém se analisarmos a Figura 7.3 em conjunto com a Figura 7.2, podemos concluir que este coeficiente local está bem coerente, pois onde o desvio-padrão é maior é onde o coeficiente é menor. O coeficiente de determinação global encontrado para este exemplo foi de 95,05%.

7.2. Exemplo 2

Abordamos nesta aplicação a análise de dados ambientais. Os dados foram retirados do site do Sistema de Organização Nacional de Dados Ambientais (SONDA). Segundo informações contidas no site, a rede SONDA de dados nasceu de um projeto do Instituto Nacional de Pesquisas Espaciais (INPE) para implementação de infra-estrutura física e de recursos humanos destinados a levantar e melhorar a base de dados dos recursos de energia solar e eólica no Brasil. A rede de estações SONDA conta com 13 estações próprias e 2 estações colaboradoras distribuídas por todo o território brasileiro. Os sensores existentes em cada estação determinam quais variáveis são medidas em cada caso. Neste trabalho analisamos alguns dados da estação de Brasília, e encontramos uma relação interessante, que se aplica a este trabalho, entre as variáveis: Radiação Global Horizontal e Radiação Difusa.

O fato da radiação solar incidente sobre a superfície terrestre constituir de um recurso não controlado pela natureza, gera um interesse maior pelo seu conhecimento como elemento meteorológico, pois, por exemplo, como cita Drechmer (2005), a radiação solar é um dos fatores determinantes da produção agrícola e seu conhecimento como elemento meteorológico é essencial na escolha adequada de culturas capazes de melhor se adaptarem às condições de cada região.

A radiação que incide na superfície horizontal é constituída de duas componentes. A componente resultante da interação da radiação solar com gases e partículas existentes na atmosfera é denominada radiação difusa, enquanto que a radiação que incide diretamente sobre a superfície do solo é chamada radiação direta (Drechmer (2005)).

A radiação global é medida por um aparelho denominado piranômetro que é posicionado horizontalmente em uma superfície horizontal (por isso se chama de radiação global horizontal). Para medir a radiação difusa, existem alguns métodos que podem se utilizados (para detalhes, veja (Drechmer & Ricieri (2006)). O método utilizado pela rede SONDA consiste em adicionar um disco de sombreamento ao piranômetro, o que impede a incidência da radiação direta. Ou seja, a radiação global é a radiação solar medida por um piranômetro sem sombreador e a radiação difusa por um piranômetro com sombreador.

As medições feitas pelos sensores em cada estação são registradas a cada minuto. Como ao longo de um dia há muita variação de incidência solar, consideramos adequado trabalhar com um determinado horário por dia, e o horário escolhido foi 12:00. Assim, nossas variáveis de interesse são:

Radiação Global: Média da radiação global em W/m^2 às 12:00
Radiação Difusa: Média da radiação difusa em W/m^2 às 12:00

Para o ano de 2010, na base de dados da rede SONDA, para a estação de Brasília, havia registro de dois meses: Janeiro e Fevereiro. Estes são os dados que abordamos neste trabalho, totalizando 59 observações.

A Figura 7.4 ilustra o relacionamento entre estas variáveis.

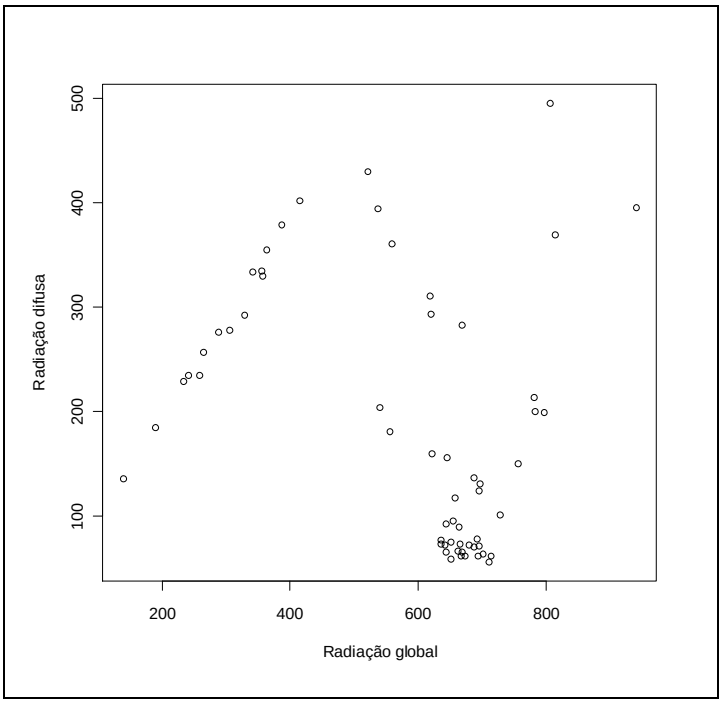


Figura 7.4: Radiação global versus radiação difusa

Analisando a Figura 7.4, identificamos que realmente há uma relação entre as variáveis sob estudo, o que torna conveniente a estimação da curva de regressão. Porém, na estimação da curva de regressão não-paramétrica, nos deparamos com uma dificuldade. Por haver alguns intervalos sem observações (veja no gráfico, por exemplo, os dados de radiação global entre 420 e 500, aproximadamente), o algoritmo não consegue estimar a janela em todos os intervalos, não sendo capaz, por conseguinte, de encontrar a janela ótima. Uma alternativa para contornar este problema, foi realizar uma transformação nos dados originais. A transformação aqui utilizada, e que proporcionou bons resultados, foi dividir os dados de radiação global e difusa por 1000. Ou seja, nossas novas variáveis são:

$$Radiação\ Global^{*} = Radiação\ Global\ /\ 1000$$
$$Radiação\ Difusa^{*}: Radiação\ Difusa\ /\ 1000$$

A Figura 7.5 exibe o gráfico de dispersão entre essas variáveis com a curva de regressão não-paramétrica estimada.

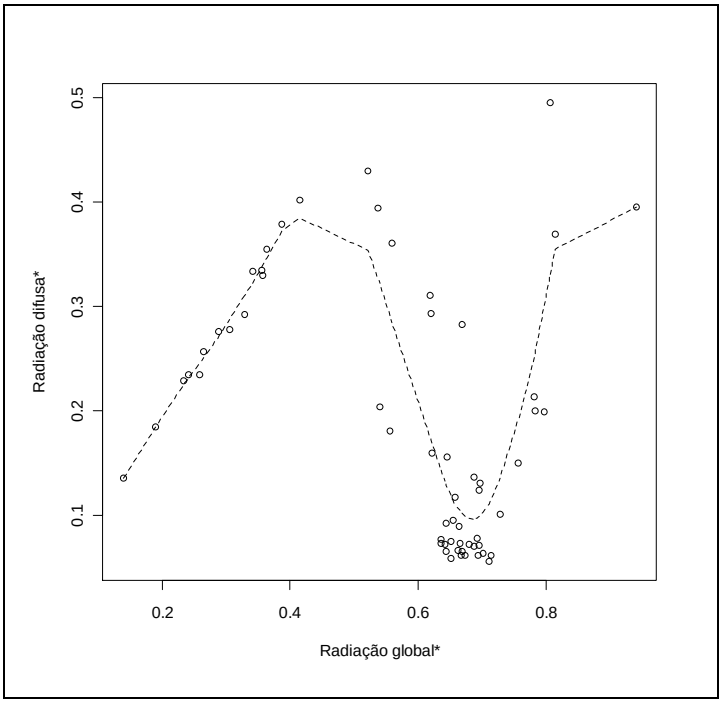


Figura 7.5: Radiação global* versus radiação difusa* com estimativa da curva de regressão não paramétrica

A Figura 7.6 exibe o gráfico do desvio-padrão estimado e a Figura 7.7 o coeficiente de determinação local estimado, ambos os gráficos feitos com os dados transformados.

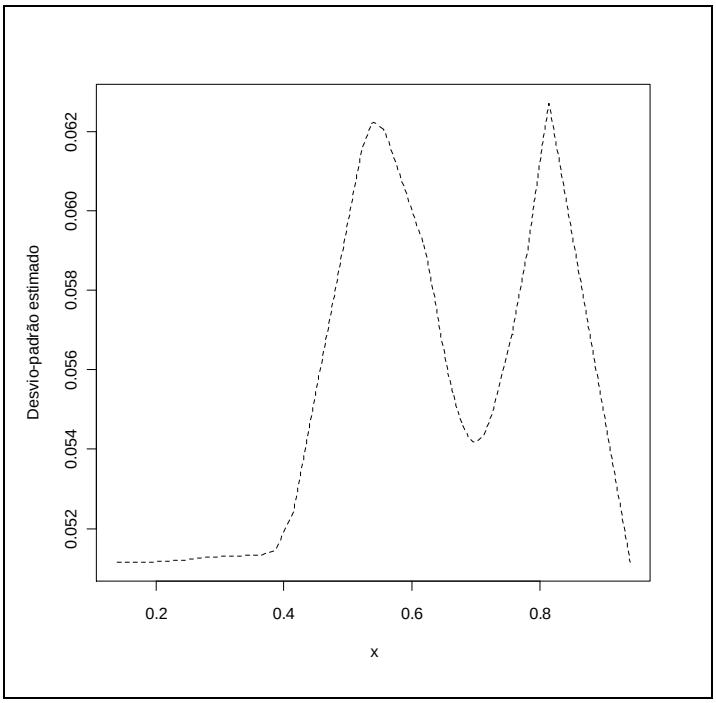


Figura 7.6: Desvio-padrão estimado para os dados de radiação

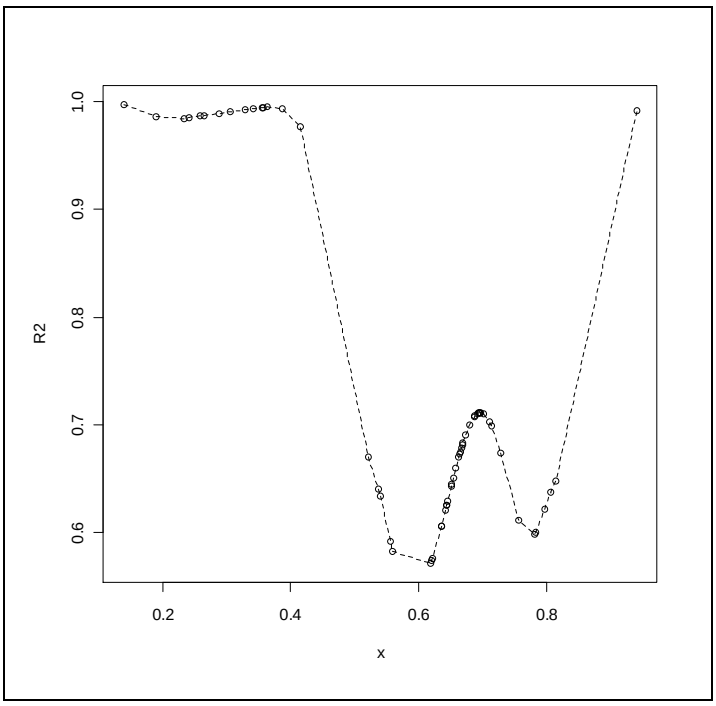


Figura 7.7: Coeficiente de determinação local para os dados de radiação

Avaliando a Figura 7.7 em conjunto com a Figura 7.6, podemos verificar que o coeficiente de determinação local está totalmente coerente com o desvio-padrão. Onde o desvio-padrão é maior, temos um coeficiente de determinação local menor. Os valores do coeficiente de determinação local foram bem altos, de 60%, aproximadamente, a 99%.

8. CONSIDERAÇÕES FINAIS

O método de estimação da função de regressão apresentado aqui, o ajuste polinomial local, já foi utilizado em vários trabalhos e estudado com detalhes. Vimos que realmente ele se comporta muito bem, além de ser de fácil implementação. O método de estimação da janela ótima, que é fundamental para um bom desempenho da estimação da função de regressão, foi apresentado neste trabalho com detalhes, incluindo os detalhes teóricos por trás dos teoremas que fundamentam o método, e apresentou excelente performance. Uma abordagem que não foi considerada é a construção de intervalos de confiança para a função $m^{(\nu)}(x_0)$. Fan & Gijbels (1996), por exemplo, apresentam estes intervalos, e uma avaliação destes pode ser feita em trabalhos futuros.

Em relação à estimação da função variância, apesar de não termos avaliado todos os detalhes teóricos que foram apresentados em Chen, Cheng & Peng (2009), o método exibido aqui apresentou resultados satisfatórios. Nas simulações as estimações se aproximaram bastante das funções reais. Um ponto que pode ser estudado posteriormente, não mencionado no artigo em questão, é se este método de estimação da variância pode sofrer efeitos de fronteira. Levantamos esta questão por verificar em alguns gráficos feitos (por exemplo, Figura 6.2) que nas fronteiras o desvio-padrão estimado apresentava oscilações, mas ressaltamos que o ajuste polinomial local, método de estimação da função regressão, não sofre estes efeitos de fronteira (Fan & Gijbels (1996)).

Nossas maiores dificuldades neste trabalho foram relacionadas ao coeficiente de determinação local apresentado em Huang & Chen (2008). Encontramos algumas incoerências com este coeficiente, principalmente nos casos heteroscedasticos, por acreditarmos que a avaliação ponto a ponto deveria apresentar menor coeficiente de determinação para um dado ponto mais afastado da curva de regressão estimada. Porém, vimos que em alguns casos não é exatamente isto que ocorre. Sabemos que mesmo a avaliação sendo local, toda a vizinhança de um dado ponto interfere no cálculo, e por isto cogitamos a idéia, com base apenas nas figuras que encontramos, de que a densidade conjunta $f(x,y)$ afeta fortemente o coeficiente local. Neste trabalho não foi possível identificar de que forma isto ocorre, e deixamos esta discussão para trabalhos futuros.

REFERÊNCIAS BIBLIOGRÁFICAS

Chen, L. H, Cheng, M. Y, Peng, L. (2009) *Conditional variance estimation in heteroscedastic regression models*. Journal of Statistical Planning and Inference, 139, 236-245.

Drechmer, P. A. O (2005) *Comportamento e correção da radiação solar difusa obtida com o anel de sombreamento*. Dissertação de Mestrado. Universidade Estadual do Oeste do Paraná, Cascavel, Paraná.

Drechmer, P. A. O., Ricieri, R. P. (2006) *Irradiação global, direta e difusa, para a região de Cascavel, Estado do Paraná*. Programa de Mestrado em Engenharia Agrícola. Universidade Estadual do Oeste do Paraná, Cascavel, Paraná.

Fan, J., Gijbels, I. (1995) *Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaptation*. Journal of the Royal Statistical Society, 57, 371-394.

Fan, J., Gijbels, I. (1996) *Local polynomial modelling and its applications*. Chapman & Hall, London.

Fan, J., Yao, Q. (1998) *Efficient estimation of conditional variance functions in stochastic regression*. Biometrika, 85, 645-660.

Gomes I. C. (2010) *Regressão polinomial multivariada – Estimação e Aplicações*. Dissertação de Mestrado. Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais.

Lehmann, E. L., Casella, G. (1998) *Theory of point estimation*. Springer Texts in Statistics, New York.

Lima, E. L. (1976) *Curso de Análise*. Volume 1. Instituto de Matemática Pura e Aplicada, Rio de Janeiro.

Huang, L. S., Chen, J. (2008) *Analysis of variance, coefficient of determination and F-test for local polynomial regression*. The Annals of Statistics, 63, 2085-2109.

Miranda, M. F. (2007) *Estimação dos coeficientes de um processo de difusão*. Dissertação de Mestrado. Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais.

Montgomery, D. C., Peck, E. A. (1992) *Introduction to linear regression analysis*. John Wiley & Sons, New York.

Ruppert, D., Wand, M. P. (1994) *Multivariate weighted least squares regression*. The Annals of Statistics, 22, 1346-1370.

Sistema de Organização Nacional de Dados Ambientais (SONDA). Disponível em: <*http://sonda.cptec.inpe.br/*>. Acesso em Setembro / Outubro de 2010.

Silverman, B. W. (1986) *Density estimation for statistics and data analysis*. Chapman and Hall, London.

Simonoff, J. S. (1996) *Smoothing methods in Statistics*. Springer Series in Statistics, New York. Arquivo de dados disponível em: <*http://pages.stern.nyu.edu/~jsimonof/ SmoothMeth/Data/ASCII*>. Acesso em Fevereiro de 2010.

Wand, M. P., Jones, M. C. (1995) *Kernel smoothing*. Chapman & Hall/CRC, London.

Yu, K., Jones, M. C. (2004) *Likelihood-based local linear estimation of the conditional variance function*. Journal of the American Statistical Association, 99, 139-144.

APÊNDICE – Prova do Teorema 3

Apresentamos a seguir a prova do Teorema 3 introduzido na subseção 4.4. Em Fan & Gijbels (1996) algumas sugestões para a realização desta prova foram fornecidas, mas seu desenvolvimento como um todo foi feito no decorrer deste trabalho.

De acordo com a expressão (4.13), temos que:

$$RSC(x_0;h) = \hat{\sigma}^2(x_0)\{\,1+(p+1)V\,\}$$

Portanto,

$$E\{\,RSC(x_0;h)\mid X_1,...,X_n\,\} = E\{\,\hat{\sigma}^2(x_0)\mid X_1,...,X_n\,\}\{\,1+(p+1)V\,\}$$

Seja $d_n = tr\left\{W - WX\left(X^T WX\right)^{-1} X^T W\right\}$ e $W_i = K_h(X_i - x_0)$. Pela expressão 4.10, a soma de quadrados residual ponderada normalizada, resultante de um ajuste polinomial local de ordem p e baseado em um janela h, é denotada do seguinte modo:

$$\hat{\sigma}^2(x_0) = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 K_h(X_i - x_0)}{tr\left\{W - WX\left(X^T WX\right)^{-1} X^T W\right\}}$$

Então, podemos escrever:

$$\hat{\sigma}^2(x_0) = d_n^{-1}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 W_i \tag{8.1}$$

e em forma matricial:

$$\hat{\sigma}^2(x_0) = d_n^{-1}\left(y - X\hat{\beta}\right)^T W\left(y - X\hat{\beta}\right) \tag{8.2}$$

Utilizando da simetria da matriz W e da estimativa $\hat{\beta}$ em (3.4), temos:

$$
\begin{aligned}
\hat{\sigma}^2(x_0) &= d_n^{-1}\left(y - X\hat{\beta}\right)^T W\left(y - X\hat{\beta}\right) \\
&= d_n^{-1}\left(y^T - \hat{\beta}^T X^T\right) W\left(y - X\hat{\beta}\right) \\
&= d_n^{-1}\left(y^T - \left(\left(X^T WX\right)^{-1} X^T W y\right)^T X^T\right) W\left(y - X\left(\left(X^T WX\right)^{-1} X^T W y\right)\right) \\
&= d_n^{-1}\left(y^T - \left(y^T WX\left(X^T WX\right)^{-1} X^T\right)\right) W\left(y - X\left(\left(X^T WX\right)^{-1} X^T W y\right)\right) \\
&= d_n^{-1}\left(y^T\left(W - WX\left(X^T WX\right)^{-1} X^T W\right)\right)\left(I - X\left(\left(X^T WX\right)^{-1} X^T W\right)\right) y \\
&= d_n^{-1}\left(y^T\left(W - WX\left(X^T WX\right)^{-1} X^T W - WX\left(X^T WX\right)^{-1} X^T W + \right.\right. \\
&\qquad \left.\left. + WX\left(X^T WX\right)^{-1} X^T WX\left(X^T WX\right)^{-1} X^T W\right)\right) y \\
&= d_n^{-1}\left(y^T\left(W - WX\left(X^T WX\right)^{-1} X^T W - WX\left(X^T WX\right)^{-1} X^T W + \right.\right. \\
&\qquad \left.\left. + WX\left(X^T WX\right)^{-1} X^T W\right)\right) y \\
&= d_n^{-1}\left(y^T\left(W - WX\left(X^T WX\right)^{-1} X^T W\right)\right) y
\end{aligned}
$$

Uma vez que $Y_i = m(X_i) + \sigma(X_i)\varepsilon_i$, e usando homocedasticidade local, na forma matricial podemos escrever $y = m + \sigma(x_0)\varepsilon$. Assim:

$$
\begin{aligned}
E\left\{\hat{\sigma}^2(x_0) \mid X_1,...,X_n\right\} &= d_n^{-1} E\left(\left(m + \sigma(x_0)\varepsilon\right)^T\left\{W - WX\left(X^T WX\right)^{-1} X^T W\right\}\left(m + \sigma(x_0)\varepsilon\right)\right) \\
&= d_n^{-1} E\left(\left(m^T + \varepsilon^T\sigma(x_0)^T\right)\left\{W - WX\left(X^T WX\right)^{-1} X^T W\right\}\left(m + \sigma(x_0)\varepsilon\right)\right) \\
&= d_n^{-1}\left[E\left(m^T\left\{W - WX\left(X^T WX\right)^{-1} X^T W\right\} m\right)\right. \\
&\qquad + E\left(m^T\left\{W - WX\left(X^T WX\right)^{-1} X^T W\right\}\sigma(x_0)\varepsilon\right) \\
&\qquad + E\left(\varepsilon^T\sigma(x_0)^T\left\{W - WX\left(X^T WX\right)^{-1} X^T W\right\} m\right) \\
&\qquad \left. + E\left(\varepsilon^T\sigma(x_0)^T\left\{W - WX\left(X^T WX\right)^{-1} X^T W\right\}\sigma(x_0)\varepsilon\right)\right] \\
&= d_n^{-1}\left[m^T\left\{W - WX\left(X^T WX\right)^{-1} X^T W\right\} m\right. \\
&\qquad + m^T\left\{W - WX\left(X^T WX\right)^{-1} X^T W\right\}\sigma(x_0) E(\varepsilon) \\
&\qquad + E\left(\varepsilon^T\right)\sigma(x_0)^T\left\{W - WX\left(X^T WX\right)^{-1} X^T W\right\} m \\
&\qquad \left. + E\left(\varepsilon^T\sigma(x_0)^T\left\{W - WX\left(X^T WX\right)^{-1} X^T W\right\}\sigma(x_0)\varepsilon\right)\right]
\end{aligned}
$$

Como $E(\varepsilon_i \mid X_i) = 0$,

$$E\left\{\hat{\sigma}^2(x_0)\mid X_1,...,X_n\right\}= d_n^{-1}\Big[m^T\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}m + E\left(\varepsilon^T\sigma(x_0)^T\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}\sigma(x_0)\varepsilon\right)\Big] \quad (8.3)$$

Primeiro, vamos trabalhar com o segundo termo desta equação. Uma vez que o valor desta esperança é um número real, podemos trabalhar com as propriedades de traço de uma matriz. Desde modo:

$$\begin{aligned}
&E\left(\varepsilon^T\sigma(x_0)^T\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}\sigma(x_0)\varepsilon\right)=\\
&=E\left(tr\left(\varepsilon^T\sigma(x_0)^T\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}\sigma(x_0)\varepsilon\right)\right)\\
&=E\left(tr\left(\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}\sigma(x_0)\varepsilon\,\varepsilon^T\sigma(x_0)^T\right)\right)\\
&=E\left(tr\left(\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}\sigma^2(x_0)\varepsilon^2\right)\right)\\
&=tr\left(\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}\sigma^2(x_0)\right)E\left(\varepsilon^2\right)
\end{aligned}$$

Por hipótese, sabemos que $E(\varepsilon_i \mid X_i)=0$, $Var(\varepsilon_i \mid X_i)=1$ e X e ε são independentes. Além disso, $Var(\varepsilon_i \mid X_i)=E(\varepsilon_i^2 \mid X_i)-(E(\varepsilon_i \mid X_i))^2 \Rightarrow E(\varepsilon_i^2 \mid X_i)=1$. Logo, voltando em (8.3):

$$\begin{aligned}
&E\left\{\hat{\sigma}^2(x_0)\mid X_1,...,X_n\right\}\\
&=d_n^{-1}\left[m^T\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}m+tr\left(\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}\sigma^2(x_0)\right)\right]
\end{aligned}$$

Substituindo a expressão de d_n:

$$\begin{aligned}
&E\left\{\hat{\sigma}^2(x_0)\mid X_1,...,X_n\right\}\\
&=\left(tr\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}\right)^{-1}\Big[m^T\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}m\\
&\qquad\qquad+\sigma^2(x_0)\;tr\left(\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}\right)\Big]\\
&=\left(tr\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}\right)^{-1}m^T\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}m+\sigma^2(x_0)\\
&=d_n^{-1}\,m^T\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}m+\sigma^2(x_0)
\end{aligned}$$

Uma vez que $r=m-X\beta \Rightarrow m=r+X\beta$,

$$
\begin{aligned}
E\left\{\hat{\sigma}^2(x_0)\mid X_1,...,X_n\right\} &= d_n^{-1}(r+X\beta)^T\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}(r+X\beta)+\sigma^2(x_0)\\
&= d_n^{-1}\left(r^T+\beta^TX^T\right)\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}(r+X\beta)+\sigma^2(x_0)\\
&= d_n^{-1}\Big(r^T\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}\\
&\qquad +\beta^TX^T\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}\Big)(r+X\beta)+\sigma^2(x_0)\\
&= d_n^{-1}\Big(r^T\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}r\\
&\qquad +r^T\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}X\beta\\
&\qquad +\beta^TX^T\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}r\\
&\qquad +\beta^TX^T\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}X\beta\Big)+\sigma^2(x_0)
\end{aligned}
$$

Notando que $\beta^TX^T\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}r$ é um escalar, sabemos que seu valor é igual a sua transposta $r^T\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}X\beta$. Logo,

$$
\begin{aligned}
E\left\{\hat{\sigma}^2(x_0)\mid X_1,...,X_n\right\} &= d_n^{-1}\Big(r^T\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}r\\
&\qquad +2\beta^TX^T\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}r\\
&\qquad +\beta^TX^T\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}X\beta\Big)+\sigma^2(x_0)\\
&= d_n^{-1}\Big(r^T\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}r\\
&\qquad +2\beta^T\left\{X^TW-X^TWX\left(X^TWX\right)^{-1}X^TW\right\}r\\
&\qquad +\beta^T\left\{X^TW-X^TWX\left(X^TWX\right)^{-1}X^TW\right\}X\beta\Big)+\sigma^2(x_0)\\
&= d_n^{-1}r^T\left\{W-WX\left(X^TWX\right)^{-1}X^TW\right\}r+\sigma^2(x_0)
\end{aligned}
\qquad (8.4)
$$

O *i-ésimo* elemento do vetor de resíduo r pode ser aproximado por:

$$
r_i = m(X_i)-\sum_{j=0}^{p}\beta_j(X_i-x_0)^j \qquad (8.5)
$$

De acordo com a expressão 3.1 e usando a expansão de Taylor na subseção 2.2., temos:

$$
\begin{aligned}
m(X_i) &= \sum_{j=0}^{\infty} \frac{(X_i - x_0)}{j!} m^{(j)}(x_0) \\
&= \sum_{j=0}^{p} \frac{(X_i - x_0)^j}{j!} m^{(j)}(x_0) + \frac{(X_i - x_0)^{p+1}}{(p+1)!} m^{(p+1)}(x_0) + O_P(X_i - x_0)^{p+2}
\end{aligned}
\tag{8.6}
$$

Assim, uma vez que $\beta_{p+1} = \dfrac{m^{(p+1)}(x_0)}{(p+1)!}$,

$$
\begin{aligned}
r_i &= m(X_i) - \textstyle\sum_{j=0}^{p} \beta_j (X_i - x_0)^j \\
&= \sum_{j=0}^{p} \beta_j (X_i - x_0)^j + \beta_{p+1}(X_i - x_0)^{p+1} + O_p(X_i - x_0)^{p+2} - \textstyle\sum_{j=0}^{p} \beta_j (X_i - x_0)^j \\
&= \beta_{p+1}(X_i - x_0)^{p+1}(1 + O_P(h))
\end{aligned}
\tag{8.7}
$$

Agora, vamos trabalhar com o termo $r^T \left\{ W - WX\left(X^T WX\right)^{-1} X^T W \right\} r$ da expressão (8.4). Como este resultado é um escalar, podemos trabalhar com as propriedades de traço:

$$
\begin{aligned}
r^T \left\{ W - WX\left(X^T WX\right)^{-1} X^T W \right\} r &= tr\left\{ r^T \left\{ W - WX\left(X^T WX\right)^{-1} X^T W \right\} r \right\} \\
&= tr\left\{ \left\{ W - WX\left(X^T WX\right)^{-1} X^T W \right\} rr^T \right\} \\
&= tr\left\{ Wrr^T \right\} - tr\left\{ WX\left(X^T WX\right)^{-1} X^T Wrr^T \right\}
\end{aligned}
\tag{8.8}
$$

Vamos desenvolver a quantidade rr^T usando o resultado da expressão (8.7):

$$
rr^T = \begin{bmatrix} \beta_{p+1}(X_1 - x_0)^{p+1}(1 + O_P(h)) \\ \vdots \\ \beta_{p+1}(X_n - x_0)^{p+1}(1 + O_P(h)) \end{bmatrix} \left[\beta_{p+1}(X_1 - x_0)^{p+1}(1 + O_P(h)) \cdots \beta_{p+1}(X_n - x_0)^{p+1}(1 + O_P(h)) \right]
$$

$$
= \begin{bmatrix} \beta_{p+1}^2 (X_1 - x_0)^{2p+2}(1 + O_P(h)) & \cdots & \beta_{p+1}^2 (X_1 - x_0)^{p+1}(X_n - x_0)^{p+1}(1 + O_P(h)) \\ \vdots & \ddots & \vdots \\ \beta_{p+1}^2 (X_1 - x_0)^{p+1}(X_n - x_0)^{p+1}(1 + O_P(h)) & \cdots & \beta_{p+1}^2 (X_n - x_0)^{2p+2}(1 + O_P(h)) \end{bmatrix}
$$

Agora, desenvolvendo o primeiro termo da expressão (8.8), temos:

$$tr\left\{Wrr^T\right\} = tr\left\{\begin{bmatrix} \beta_{p+1}^2 (X_1 - x_0)^{2p+2} K_h(X_1 - x_0)(1 + O_P(h)) & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & \beta_{p+1}^2 (X_n - x_0)^{2p+2} K_h(X_n - x_0)(1 + O_P(h)) \end{bmatrix}\right\}$$

$$= \sum_{i=1}^{n} \beta_{p+1}^2 (X_i - x_0)^{2p+2} K_h(X_i - x_0)(1 + O_P(h))$$

$$= \beta_{p+1}^2 \sum_{i=1}^{n} K_h(X_i - x_0)(X_i - x_0)^{2p+2}(1 + O_P(h))$$

Como já visto que $S_{n,j} = \sum_{i=1}^{n} K_h(X_i - x_0)(X_i - x_0)^j$, então,

$$tr\left\{Wrr^T\right\} = \beta_{p+1}^2 S_{n,2p+2}\left(1 + O_P(h)\right) \tag{8.9}$$

Trabalhando com o segundo termo da expressão (8.8), temos:

$$tr\left\{WX\left(X^TWX\right)^{-1}X^TWrr^T\right\} = tr\left\{r^TWX\left(X^TWX\right)^{-1}X^TWr\right\} \tag{8.10}$$

sendo:

$$X^TWr = \begin{bmatrix} 1 & \cdots & 1 \\ (X_1 - x_0) & \cdots & (X_n - x_0) \\ \vdots & \ddots & \vdots \\ (X_1 - x_0)^p & \cdots & (X_n - x_0)^p \end{bmatrix} \begin{bmatrix} K_h(X_1 - x_0) & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & K_h(X_n - x_0) \end{bmatrix} \begin{bmatrix} \beta_{p+1}(X_1 - x_0)^{p+1}(1 + O_P(h)) \\ \vdots \\ \beta_{p+1}(X_n - x_0)^{p+1}(1 + O_P(h)) \end{bmatrix}$$

$$= \begin{bmatrix} K_h(X_1 - x_0) & \cdots & K_h(X_n - x_0) \\ K_h(X_1 - x_0)(X_1 - x_0) & \cdots & K_h(X_n - x_0)(X_n - x_0) \\ \vdots & & \vdots \\ K_h(X_1 - x_0)(X_1 - x_0)^p & \cdots & K_h(X_n - x_0)(X_n - x_0)^p \end{bmatrix} \begin{bmatrix} \beta_{p+1}(X_1 - x_0)^{p+1}(1 + O_P(h)) \\ \vdots \\ \beta_{p+1}(X_n - x_0)^{p+1}(1 + O_P(h)) \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{n} \beta_{p+1} K_h(X_i - x_0)(X_i - x_0)^{p+1}(1 + O_P(h)) \\ \vdots \\ \sum_{i=1}^{n} \beta_{p+1} K_h(X_i - x_0)(X_i - x_0)^{2p+1}(1 + O_P(h)) \end{bmatrix}$$

$$X^T W r = \begin{bmatrix} \beta_{p+1} S_{n,p+1}(1+O_P(h)) \\ \vdots \\ \beta_{p+1} S_{n,2p+1}(1+O_P(h)) \end{bmatrix}$$
$$= \beta_{p+1} c_n (1+O_P(h))$$

onde $c_n = (S_{n,p+1},...,S_{n,2p+1})^T$.

Substituindo essa expressão em (8.10), e usando que $S_n = X^T W X$, encontramos:

$$tr\left\{ r^T W X (X^T W X)^{-1} X^T W r \right\} = \left(c_n^T \beta_{p+1}(1+O_P(h)) S_n^{-1} \left(\beta_{p+1} c_n (1+O_P(h)) \right) \right)$$
$$= \beta_{p+1}^2 c_n^T S_n^{-1} c_n (1+O_P(h)) \qquad (8.11)$$

Finalmente, substituindo (8.9) e (8.11) em (8.8):

$$r^T \left\{ W - WX(X^T W X)^{-1} X^T W \right\} r = tr\left\{ W r r^T \right\} - tr\left\{ WX(X^T W X)^{-1} X^T W r r^T \right\}$$
$$= \beta_{p+1}^2 S_{n,2p+2}(1+O_p(h)) - \beta_{p+1}^2 c_n^T S_n^{-1} c_n (1+O_P(h)) \qquad (8.12)$$
$$= (S_{n,2p+2} - c_n^T S_n^{-1} c_n) \beta_{p+1}^2 (1+O_P(h))$$

Lembrando que $S_n = X^T W X$, com $S_{n,j} = \sum_{i=1}^n K_h(X_i - x_0)(X_i - x_0)^j$ e adotando a notação $S_n^* = X^T \Sigma X$, com $S_{n,j}^* = \sum_{i=1}^n (X_i - x_0)^j K_h^2(X_i - x_0)\sigma^2(X_i)$, vamos encontrar aproximações para as matrizes S_n e S_n^*.

$$S_{n,j} = \sum_{i=1}^n K_h(X_i - x_0)(X_i - x_0)^j$$
$$E\left(\frac{1}{n} S_{n,j}\right) = \int \frac{1}{n} n \frac{1}{h} K\left(\frac{u - x_0}{h}\right)(u - x_0)^j f(u)\, du$$

Seja $v = \left(\frac{u - x_0}{h}\right) \Rightarrow \begin{matrix} u = x_0 + vh \\ du = h\, dv \end{matrix}$

Assim,

$$\frac{1}{n}E\left(S_{n,j}\right)=\int \frac{1}{h}K(v)(vh)^{j}\ f(x_0+vh)h\,dv$$
$$=\int K(v)(vh)^{j}\ f(x_0+vh)dv$$

Usando expansão de Taylor (Teorema 1):

$$\frac{1}{n}E\left(S_{n,j}\right)=\int K(v)v^{j}\,h^{j}\left(\ f(x_0)+hv\ f^{'}(x_0)+\frac{h^2v^2}{2}\ f^{''}(x_0)+o\left(h^2\right)\right)dv$$
$$=\int K(v)v^{j}\,h^{j}\,f(x_0)dv+\int\ K(v)v^{j+1}\,h^{j+1}\ f^{'}(x_0)dv+\int K(v)\frac{v^{j+2}\,h^{j+2}}{2}\ f^{''}(x_0)dv+o\left(h^2\right)$$
$$=h^{j}\,f(x_0)\int K(v)v^{j}\,dv+h^{j+1}\ f^{'}(x_0)\int K(v)v^{j+1}\,dv+\frac{h^{j+2}}{2}\ f^{''}(x_0)\int K(v)v^{j+2}\,dv+o\left(h^2\right)$$

Como já visto que $\mu_j=\int u^{j}K(u)\,du$ e $\lambda_j=\int u^{j}K^{2}(u)\,du$,

$$\frac{1}{n}E\left(S_{n,j}\right)=h^{j}\ f(x_0)\mu_j+h^{j+1}\ f^{'}(x_0)\mu_{j+1}+\frac{h^{j+2}}{2}\ f^{''}(x_0)\mu_{j+2}+o\left(h^2\right)$$
$$E\left(S_{n,j}\right)=n\left(h^{j}\ f(x_0)\mu_j+h^{j+1}\ f^{'}(x_0)\mu_{j+1}+\frac{h^{j+2}}{2}\ f^{''}(x_0)\mu_{j+2}+o\left(h^2\right)\right) \tag{8.13}$$

Fazendo $h\to 0$ e $nh\to\infty$, temos:

$$E\left(S_{n,j}\right)=n\,h^{j}\ f(x_0)\mu_j\left(1+o_P(1)\right)$$

E, portanto,

$$S_{n,j}=n\,h^{j}\ f(x_0)\mu_j\left(1+o_P(1)\right) \tag{8.14}$$

e,

$$S_n=n\ f(x_0)H\,S\,H\left(1+o_P(1)\right) \tag{8.15}$$

onde: $S = \begin{bmatrix} \mu_0 & \mu_1 & \cdots & \mu_p \\ \mu_1 & \mu_2 & \cdots & \mu_{p+1} \\ \vdots & \vdots & & \vdots \\ \mu_p & \mu_{p+1} & \cdots & \mu_{2p} \end{bmatrix}$ e $H = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & h & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & h^p \end{bmatrix}$

Trabalhando agora com a matriz S_n^*, temos:

$$S_{n,j}^* = \sum_{i=1}^{n} (X_i - x_0)^j K_h^2(X_i - x_0)\sigma^2(X_i)$$

$$E\left(\frac{1}{n}S_{n,j}^*\right) = \int \frac{1}{n} n (u - x_0)^j \frac{1}{h^2} K^2\left(\frac{u - x_0}{h}\right)\sigma^2(u) f(u)\, du$$

Fazendo $v = \left(\frac{u - x_0}{h}\right) \Rightarrow \begin{matrix} u = x_0 + vh \\ du = h dv \end{matrix}$

Então,

$$E\left(\frac{1}{n}S_{n,j}^*\right) = \int (vh)^j \frac{1}{h^2} K^2(v)\sigma^2(x_0 + vh) f(x_0 + vh) h\, dv$$

Usando expansão de Taylor (Teorema 1):

$$\begin{aligned} E\left(\frac{1}{n}S_{n,j}^*\right) &= \int v^j h^{j-1} K^2(v)\left(\sigma^2(x_0) + hv\sigma^{2'}(x_0) + \frac{h^2v^2}{2}\sigma^{2''}(x_0) + o(h^2)\right) \\ &\quad \times\left(f(x_0) + hv f'(x_0) + \frac{h^2v^2}{2} f''(x_0) + o(h^2)\right) dv \\ &= \int v^j h^{j-1} K^2(v)\sigma^2(x_0) f(x_0)\, dv + \int v^{j+1} h^j K^2(v)\sigma^2(x_0) f'(x_0)\, dv \\ &\quad + \int \frac{v^{j+2}h^{j+1}}{2} K^2(v)\sigma^2(x_0) f''(x_0)\, dv + \int v^{j+1} h^j K^2(v)\sigma^{2'}(x_0) f(x_0)\, dv \\ &\quad + \int v^{j+2} h^{j+1} K^2(v)\sigma^{2'}(x_0) f'(x_0)\, dv + \ldots + o(h^2) \\ &= h^{j-1}\sigma^2(x_0) f(x_0)\int v^j K^2(v)\, dv + h^j\sigma^2(x_0) f'(x_0)\int v^{j+1} K^2(v)\, dv \\ &\quad + \frac{h^{j+1}}{2}\sigma^2(x_0) f''(x_0)\int v^{j+2} K^2(v)\, dv + h^j \sigma^{2'}(x_0) f(x_0)\int v^{j+1} K^2(v)\, dv \\ &\quad + h^{j+1}\sigma^{2'}(x_0) f'(x_0)\int v^{j+2} K^2(v) dv + \ldots + o(h^2) \end{aligned}$$

$$E\left(\frac{1}{n}S^*_{n,j}\right) = h^{j-1}\sigma^2(x_0)f(x_0)\lambda^j + h^j\sigma^2(x_0)f'(x_0)\lambda^{j+1} + \frac{h^{j+1}}{2}\sigma^2(x_0)f''(x_0)\lambda^{j+2}$$
$$+ h^j\sigma^{2'}(x_0)f(x_0)\lambda^{j+1} + h^{j+1}\sigma^{2'}(x_0)f'(x_0)\lambda^{j+2} + \ldots + o\left(h^2\right)$$

E, portanto,

$$E\left(S^*_{n,j}\right) = n\left(h^{j-1}\sigma^2(x_0)f(x_0)\lambda^j + h^j\sigma^2(x_0)f'(x_0)\lambda^{j+1} + \frac{h^{j+1}}{2}\sigma^2(x_0)f''(x_0)\lambda^{j+2} \right.$$
$$\left. + h^j\sigma^{2'}(x_0)f(x_0)\lambda^{j+1} + h^{j+1}\sigma^{2'}(x_0)f'(x_0)\lambda^{j+2} + \ldots + o\left(h^2\right) \right) \tag{8.16}$$

Fazendo $h \to 0$ e $nh \to \infty$,

$$E\left(S^*_{n,j}\right) = n h^{j-1}\sigma^2(x_0)f(x_0)\lambda^j \left\{1 + o_P(1)\right\}$$

E, logo,

$$S^*_{n,j} = n h^{j-1}\sigma^2(x_0)f(x_0)\lambda^j \left\{1 + o_P(1)\right\} \tag{8.17}$$

e,

$$S^*_n = n h^{-1}\sigma^2(x_0)f(x_0)H\,S^*\,H\left\{1 + o_P(1)\right\} \tag{8.18}$$

onde: $S^* = \begin{bmatrix} \lambda_0 & \lambda_1 & \cdots & \lambda_p \\ \lambda_1 & \lambda_2 & \cdots & \lambda_{p+1} \\ \vdots & \vdots & & \vdots \\ \lambda_p & \lambda_{p+1} & \cdots & \lambda_{2p} \end{bmatrix}$

Agora, usando as aproximações (8.15) e (8.18), podemos escrever d_n da seguinte forma:

$$d_n = tr\left\{W - WX\left(X^T WX\right)^{-1}X^T W\right\}$$
$$= tr\left\{W\right\} - tr\left\{WX\left(X^T WX\right)^{-1}X^T W\right\}$$

$$
\begin{aligned}
d_n &= S_{n,0} - tr\left\{ \left(X^T W X \right)^{-1} X^T W^2 X \right\} \\
&= S_{n,0} - tr\left\{ S_n^{-1} S_n^* / \sigma^2(x_0) \right\} \\
&= S_{n,0} - tr\left\{ \frac{1}{n f(x_0)} H^{-1} S^{-1} H^{-1} \frac{n}{h} f(x_0) H S^* H \right\} \\
&= S_{n,0} - \frac{1}{h} tr\left\{ H^{-1} S^{-1} H^{-1} H S^* H \right\} \\
&= S_{n,0} - \frac{1}{h} tr\left\{ H^{-1} S^{-1} S^* H \right\} \\
&= S_{n,0} - \frac{1}{h} tr\left\{ S^{-1} S^* \right\}
\end{aligned}
\tag{8.19}
$$

Vamos avaliar o termo $S_{n,0}$. Pela expressão (8.14), temos que assintoticamente:

$$
\begin{aligned}
S_{n,j} &= n h^j f(x_0) \mu_j \left(1 + o_P(1)\right) \\
S_{n,0} &= n f(x_0) \mu_0 \left(1 + o_P(1)\right)
\end{aligned}
$$

Uma vez que $\mu_0 = 1$,

$$
S_{n,0} = n f(x_0)\left(1 + o_P(1)\right)
$$

Além disso, se $h \to 0$ e $nh \to \infty$, temos:

$$
\frac{1}{h} tr\left\{ S^{-1} S^* \right\} = O_P\left(h^{-1}\right)
$$

Logo,

$$
\begin{aligned}
d_n &= n f(x_0)\left(1 + o_P(1)\right) - O_P\left(h^{-1}\right) \\
&= n f(x_0) + O_P\left(h^{-1}\right)
\end{aligned}
\tag{8.20}
$$

Voltando na expressão da esperança em (8.4) e substituindo os resultados encontrados em (8.12), (8.14) e (8.20) temos:

$$E\left\{\hat{\sigma}^2(x_0)\mid X_1,...,X_n\right\} = d_n^{-1}\, r^T\left\{W - WX\left(X^T WX\right)^{-1} X^T W\right\} r + \sigma^2(x_0)$$

$$= \left(n\, f(x_0) + O_P\left(h^{-1}\right)\right)^{-1}\left(S_{n,2p+2} - c_n^T\, S_n^{-1}\, c_n\right)\beta_{p+1}^2\left(1 + O_P(h)\right) + \sigma^2(x_0)$$

$$= \left(\frac{1}{n\, f(x_0)}\right)\left(n h^{2p+2}\, f(x_0)\mu_{2p+2} - c_n^T\, S_n^{-1}\, c_n\right)\beta_{p+1}^2 + \sigma^2(x_0) + o_P\left(h^{2p+2}\right)$$

$$= \left(h^{2p+2}\mu_{2p+2} - \frac{c_n^T\, S_n^{-1}\, c_n}{n\, f(x_0)}\right)\beta_{p+1}^2 + \sigma^2(x_0) + o_P\left(h^{2p+2}\right)$$

Agora, lembrando que $c_n = \left(S_{n,p+1},...,S_{n,2p+1}\right)^T$ e usando os resultados de (8.14) e (8.15), podemos reescrever o termo $\frac{c_n^T\, S_n^{-1}\, c_n}{n\, f(x_0)}$ da seguinte forma:

$$\frac{c_n^T\, S_n^{-1}\, c_n}{n\, f(x_0)} = \left(\frac{\left(S_{n,p+1},...,S_{n,2p+1}\right)\left(n\, f(x_0) HSH\right)^{-1}\left(S_{n,p+1},...,S_{n,2p+1}\right)^T}{n\, f(x_0)}\right)$$

$$= \left(\frac{\left(n h^{p+1} f(x_0)\mu_{p+1},...,n h^{2p+1} f(x_0)\mu_{2p+1}\right)\left(H^{-1}S^{-1}H^{-1}\right)\left(n h^{p+1} f(x_0)\mu_{p+1},...,n h^{2p+1} f(x_0)\mu_{2p+1}\right)^T}{n\, f(x_0)\, n\, f(x_0)}\right)$$

$$= \left(\frac{n\, f(x_0)\left(h^{p+1}\mu_{p+1},...,h^{2p+1}\mu_{2p+1}\right)\left(H^{-1}S^{-1}H^{-1}\right) n\, f(x_0)\left(h^{p+1}\mu_{p+1},...,h^{2p+1}\mu_{2p+1}\right)^T}{n\, f(x_0)\, n\, f(x_0)}\right)$$

$$= \left(h^{p+1}\mu_{p+1},...,h^{2p+1}\mu_{2p+1}\right)\left(H^{-1}S^{-1}H^{-1}\right)\left(h^{p+1}\mu_{p+1},...,h^{2p+1}\mu_{2p+1}\right)^T$$

Como $H = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & h & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & h^p \end{bmatrix}$ e logo $H^{-1} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1/h & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1/h^p \end{bmatrix}$, temos:

$$\frac{c_n^T\, S_n^{-1}\, c_n}{n\, f(x_0)} = h^{p+1}\left(\mu_{p+1},...,\mu_{2p+1}\right) S^{-1}\, h^{p+1}\left(\mu_{p+1},...,\mu_{2p+1}\right)^T$$

$$= h^{2p+2}\, c_p^T\, S^{-1}\, c_p$$

onde: $c_p = \left(\mu_{p+1},...,\mu_{2p+1}\right)^T$

Agora, voltando na expressão da esperança:

$$\begin{aligned} E\left\{\hat{\sigma}^2(x_0) \mid X_1,...,X_n\right\} &= \left(h^{2p+2}\mu_{2p+2} - h^{2p+2}\, c_p^T\, S^{-1}\, c_p\right)\beta_{p+1}^2 + \sigma^2(x_0) + o_P\left(h^{2p+2}\right) \\ &= \left(\mu_{2p+2} - c_p^T\, S^{-1}\, c_p\right)h^{2p+2}\beta_{p+1}^2 + \sigma^2(x_0) + o_P\left(h^{2p+2}\right) \\ &= C_p\, \beta_{p+1}^2\, h^{2p+2} + \sigma^2(x_0) + o_P\left(h^{2p+2}\right) \end{aligned} \tag{8.21}$$

onde $C_p = \mu_{2p+2} - c_p^T S^{-1} c_p$ e $S = \left(\mu_{j+l}\right)_{0\le j,l\le p}$.

Vimos que o termo *V* que aparece na expressão (4.13) denota o primeiro elemento da diagonal da matriz $\left(X^T WX\right)^{-1}\left(X^T W^2 X\right)\left(X^T WX\right)^{-1}$, ou seja,

$$V = e_1^T\, S_n^{-1}\left(S_n^* / \sigma^2(x_0)\right)S_n^{-1}\, e_1$$

Usando (8.15) e (8.18) temos:

$$\begin{aligned} V &= e_1^T\, \frac{1}{n\, f(x_0)} H^{-1}S^{-1}H^{-1}\, \frac{n}{h}\, f(x_0) HS^* H\, \frac{1}{n\, f(x_0)} H^{-1}S^{-1}H^{-1}e_1\left(1 + o_P(1)\right) \\ V &= \frac{1}{n\, h\, f(x_0)} e_1^T S^{-1}H^{-1}HS^*HH^{-1}S^{-1}e_1\left(1 + o_P(1)\right) \\ V &= \frac{1}{n\, h\, f(x_0)} e_1^T S^{-1}S^*S^{-1}e_1 + o_P\left((nh)^{-1}\right) \\ V &= \frac{a_0}{n\, h\, f(x_0)} + o_P\left((nh)^{-1}\right) \end{aligned} \tag{8.22}$$

onde a_0 denota o primeiro elemento da diagonal da matriz $S^{-1}S^*S^{-1}$.

Finalmente, combinando (8.21) e (8.22):

$$\begin{aligned} E\left\{RSC(x_0;h) \mid X_1,...,X_n\right\} &= E\left\{\hat{\sigma}^2(x_0) \mid X_1,...,X_n\right\}\left\{1 + (p+1)V\right\} \\ &= \sigma^2(x_0) + C_p\, \beta_{p+1}^2\, h^{2p+2} + (p+1)\, a_0\, \frac{\sigma^2(x_0)}{n\, h\, f(x_0)} \\ &\quad + o_P\left\{h^{2p+2} + (nh)^{-1}\right\} \end{aligned}$$

que é exatamente a expressão no Teorema 3.