Universidade Federal de Minas Gerais

Instituto de Ciências Exatas

Departamento de Estatística

Análise de Covariância Não-paramétrica

Dissertação de Mestrado

Gabriel Vinícius Araújo Fonseca

Orientador: Gregorio Saravia Atuncar

Agradecimentos

Agradeço primeiramente a Deus pela ajuda espiritual durante este trabalho.

Ao professor Gregorio que soube me orientar nas horas em que mais precisei.

Aos meus amigos e familiares que de forma direta e indireta me apoiaram nas horas mais difíceis.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio financeiro.

E por fim à minha namorada Bárbara que por dias e horas soube esperar e apoiar sempre e sempre.

Sumário

Resumo

Neste trabalho, apresentaremos uma estatística de teste sobre a igualdade entre curvas f_i de regressão não-paramétrica, quando temos tanto ruído homogêneo quanto não homogêneo e ruído heterocedástico, ou seja, quando a variância depende do regressor e é diferente para cada grupo. Tal teste é desenvolvido por Munk, Neumeyer e Scholz (2006).

A abordagem que será apresentada é muito natural, pois ela substitui as estatísticas de máxima verossimilhança de uma análise de covariância paramétrica por estatísticas não-paramétricas. Neste caso, é usado o núcleo-estimador para substituir essas estatísticas.

Para finalidades práticas, uma variação *bootstrap* é sugerida, mais conhecida como "*wild bootstrap"*. Essa técnica visa uma melhor estimativa da distribuição dos erros e assim obter um valor mais coerente da estatística de teste. Foram feitas simulações para verificar o nível e o poder do teste para alguns modelos. Faremos ainda uma comparação entre o teste de Hall e Hart (1990) e o de Munk, Neumeyer e Scholz (2006), através do nível e o poder do teste, a título de curiosidade. Por fim, algumas aplicações com dados reais serão descritas.

1. Introdução e Revisão da Literatura

Ao longo dos últimos anos, a estatística não-paramétrica vem ocupando um lugar de destaque na área científica da estatística. A estimação funcional pelo método do núcleo ou simplesmente núcleo-estimador é uma opção para estimativas não-paramétricas. Essa técnica vem sendo aplicada em funções de densidade, intensidade, distribuição e também de regressão.

Uma questão crucial na aplicação deste método é a determinação do parâmetro de suavização ou janela, normalmente denotado por h, que controla a quantidade de suavização a ser feita. Se h é muito pequeno, admite-se demasiado ruído amostral, e se h é muito grande, omite-se as características da curva. Existem vários métodos para a estimação desse parâmetro, como a validação cruzada e o *plug-in*, os mais difundidos.

No caso da função de regressão, temos diferentes métodos de estimação, sendo os mais difundidos na atualidade o de Nadaraya-Watson (Nadaraya, 1964 e Watson, 1964) e a regressão polinomial local. Contudo a regressão polinomial local possui uma forma direta (automática) para a seleção da janela ótima, ou seja, existe um critério de erro de estimação envolvido na escolha da janela, como erro quadrado médio (MSE, em inglês). Assim apresentaremos dois métodos: o primeiro utiliza o *plug-in* para obter a estimativa da janela ótima assintótica, desenvolvido por Ruppert, Sheather e Wand (1995) e o segundo encontra a estimativa da janela ótima através da minimização da soma do quadrado dos resíduos integrado, o qual foi desenvolvido por Fan e Gijbels (1995).

Existe ainda uma forma de se obter a estimativa da função regressão para efeitos fixos (igualmente espaçados), dado por Priestley e Chao (1972), uma variação do Nadaraya-Watson (1964). Porém, mesmo com a suposição que (X,Y) são vetores aleatórios, a regressão estimada através dos métodos polinomial local e Nadaraya-Watson estimam muito bem para o caso em que temos efeitos fixos.

No estudo de análise de covariância, Hall e Hart (1990) desenvolveram um teste *bootstrap* para comparar duas curvas de regressão. Porém, o estudo realizado por eles foi apenas para amostras de tamanhos iguais $(n_1 = n_2)$ e para os mesmos pontos para a variável preditora, ou seja, são dados da seguinte forma $\{(x_i, Y_i, Z_i), 1 \le i \le n\}$.

Em Dette e Neumeyer (2001) foram discutidos três testes estatísticos usando estimadores não-paramétricos. Dado o modelo

$$Y_{ij} = f_i\left(t_{ij}\right) + \sigma_i\left(t_{ij}\right)\varepsilon_{ij}, \qquad i = 1,\ldots,k \text{ e } j = 1,\ldots,n_i;$$

as hipóteses para o teste são:

$$H_0 : f_1 = \ldots = f_k \text{ vs. } H_1 = f_i \neq f_j, \text{ para algum } i \neq j.$$

As estatísticas de teste são:

$$T_N^{(1)} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \hat{f}\left(t_{ij}\right)\right)^2 - \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \hat{f}_i\left(t_{ij}\right)\right)^2$$

$$T_N^{(2)} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\hat{f}\left(t_{ij}\right) - \hat{f}_i\left(t_{ij}\right)\right)^2$$

$$T_N^{(3)} = \sum_{i=1}^{k}\sum_{j=1}^{i-1}\int_0^1\left(\hat{f}_i\left(t\right) - \hat{f}_j\left(t\right)\right)^2 w_{ij}\left(t\right)dt$$

onde $\hat{f}$ é o estimador de regressão das amostras em conjunto (sob H_0), $\hat{f}_i$ é estimador da regressão para cada conjunto (sob H_1), $w_{ij}\left(\cdot\right)$ são funções peso positivos que satisfazem $w_{ij} = w_{ji}$, $1 \leq j < i \leq k$ e t_{ij} são valores fixos e igualmente espaçados dado pela amostra. Porém, todas as três estatísticas de teste não incorporam a função $\sigma_i\left(t_{ij}\right)$, ou seja, não possuem boa aplicabilidade quando temos dependência da variância com a variável preditora e variâncias diferentes entre os grupos. Note que neste caso não há mais as suposições feita em Hall e Hart (1990) e assim garantindo melhor aplicação em dados reais.

Contudo, em Munk, Neumeyer e Scholz (2006), há uma proposta de um novo teste, semelhante ao $T_N^{(1)}$ de Dette e Neumeyer (2001). Para isso, é usada a estatística de máxima verossimilhança de uma análise de covariância paramétrica heterocedástica para uma entrada e depois a transferindo para o contexto não-paramétrico. Mas em Munk, Neumeyer e Scholz (2006) foi utilizado o estimador de Nadaraya-Watson para a estimação da função de regressão, o qual não possui uma forma direta de se obter a janela ótima. Assim substituímos o estimador de Nadaraya-Watson pelo polinomial local e utilizamos uma das formas diretas para a escolha da janela.

Contudo a convergência desse método é lenta e o uso somente do teste estatístico não nos permite fazer uma boa decisão em cima das hipóteses. Assim é construída uma variação *bootstrap* para refinar os erros heterocedásticos e assim obter um valor crítico para a estatística de teste. Proposto, inicialmente, por Wu (1986) e Beran (1986) e finalizado posteriormente por Liu (1988), esse mecanismo é conhecido na literatura por "*Wild Bootstrap"*.

Assim, o objetivo central deste trabalho é a abordagem da análise de covariância não-paramétrica desenvolvida por Munk, Neumeyer e Scholz (2006), utilizando a regressão polinomial local e suas formas diretas de se obter a janela ótima.

Dessa forma, este trabalho compõe-se das seguintes partes: no segundo capítulo, descreveremos o procedimento da regressão não-paramétrica, tanto para efeitos fixos quanto para efeitos aleatórios. No capítulo 3, apresentamos os testes estatísticos de Munk, Neumeyer e Scholz (2006) e Hall e Hart (1990), sendo o primeiro para os casos com dois grupos ou mais, e para o segundo somente para o caso com dois grupos. O capítulo quatro contém os dois métodos automáticos para a escolha da janela ótima. No quinto capítulo é descrito o procedimento "*wild bootstrap"* e as condições para a sua aplicação. No capítulo 6, são apresentadas as simulações e duas aplicações em dados reais. A primeira aplicação é referente à produção de cebolas em duas regiões da Austrália. Já a segunda é referente ao fluxo líquido de carbono na atmosfera produzido através de mudanças no uso do solo, como desmatamento para uso agropecuário. No capítulo 7 são apresentadas as conclusões e a proposta de trabalhos futuros.

2. Regressão Não-paramétrica

Uma técnica estatística extensamente usada, em geral, é a regressão (linear). Os modelos de regressão são ferramentas poderosas para se modelar a variável Y como função da variável preditora X, permitindo a predição de valores futuros de Y e a construção de testes e estimativas intervalares para as predições e parâmetros.

Modelos de regressão são também suscetíveis a alguns problemas como em outros modelos paramétricos. Considere um simples modelo de regressão linear,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

com os erros ε_i geralmente dados como variáveis aleatórias identicamente e independentemente distribuídas de acordo com uma normal com média zero e variância σ^2. Se esse modelo é uma boa representação da realidade, as estimativas de mínimos quadrados (também chamados de Estimadores de Máxima Verossimilhança) de β podem ser calculadas e por fim utilizá-las para uma previsão depois da construção do modelo.

Mas, às vezes, ajustar uma relação funcional paramétrica envolvendo o modelo (1) pode trazer inferências equivocadas, quando os dados não seguem as suposições necessárias. Logo podemos utilizar os métodos de regressão não-paramétrica, baseados em núcleos-estimadores, para estimar o modelo. Assim, nas duas seções subseqüentes, iremos apresentar duas técnicas muito difundidas para se obter a estimativa do modelo não-paramétrico.

2.1. Método de Nadaraya-Watson (1964)

Uma alternativa mais geral para (1), é um modelo de regressão não-paramétrico

$$Y_i = m(X_i) + \varepsilon_i, \tag{2}$$

onde os ε_i's são $i.i.d.$. Se (X_i, Y_i), $i = 1, \ldots, n$ são observações de um vetor (X, Y), a curva de regressão $m(x)$ é a esperança condicional $m(x) = E(Y \mid X = x)$ com $E(\varepsilon \mid X = x) = 0$, e $V(\varepsilon \mid X = x) = \sigma^2(x)$ não necessariamente constante.

Note que, se

$$m(x) = E(Y \mid X = x), \tag{3}$$

tem-se que

$$\begin{aligned} m(x) &= \int y f(y \mid x) dy \\ &= \int y \frac{f(x,y)}{f_X(x)} dy, \end{aligned} \tag{4}$$

onde $f_X(x)$, $f(x,y)$, e $f(y \mid x)$ são a densidade marginal de X, a densidade conjunta de (X,Y), e a densidade condicional de Y dado $X = x$, respectivamente. Uma estimativa pelo método do núcleo para $f(x,y)$ é

$$\hat{f}(x,y) = \frac{1}{nh_x h_y} \sum_{i=1}^{n} K_x\left(\frac{x - X_i}{h_x}\right) K_y\left(\frac{y - Y_i}{h_y}\right),$$

enquanto uma estimativa para $f_X(x)$ é

$$\hat{f}_X(x) = \frac{1}{nh_x} \sum_{i=1}^{n} K_x\left(\frac{x - X_i}{h_x}\right).$$

Substituindo em (4), as estimativas acima, temos

$$\hat{m}(x) = \int \frac{y\left(nh_x h_y\right)^{-1} \sum_{i=1}^{n} K_x\left(\frac{x - X_i}{h_x}\right) K_y\left(\frac{y - Y_i}{h_y}\right)}{\left(nh_x\right)^{-1} \sum_{i=1}^{n} K_x\left(\frac{x - X_i}{h_x}\right)} dy.$$

$$= \frac{\int \frac{y}{h_y} \sum_{i=1}^{n} K_x\left(\frac{x - X_i}{h_x}\right) K_y\left(\frac{y - Y_i}{h_y}\right) dy}{\sum_{i=1}^{n} K_x\left(\frac{x - X_i}{h_x}\right)}.$$

Note que substituindo $u = \left(y - Y_i\right)/h_y$, $\int K_y(u)du = 1$, $\int uK_y(u)du = 0$ e assumindo que K é simétrica em torno do zero, temos

$$\hat{m}(x) = \frac{\int \sum_{i=1}^{n} \left(Y_i + uh_y\right) K_x\left(\frac{x - X_i}{h_x}\right) K_y(u) du}{\sum_{i=1}^{n} K_x\left(\frac{x - X_i}{h_x}\right)}$$

$$= \frac{\sum_{i=1}^{n} K_x\left(\frac{x - X_i}{h_x}\right) \int \left(Y_i + uh_y\right) K_y\left(u\right) du}{\sum_{i=1}^{n} K_x\left(\frac{x - X_i}{h_x}\right)}.$$

Logo, obtemos o núcleo-estimador de Nadaraya-Watson,

$$\hat{m}_{NW}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_x}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_x}\right)} \equiv \sum_{i=1}^{n} w_i Y_i, \tag{5}$$

uma função linear de Y com pesos

$$w_i = (nh)^{-1} \frac{K\left(\frac{x - X_i}{h}\right)}{\hat{f}_X(x)}.$$

O núcleo estimador de Nadaraya-Watson é mais natural para modelos de efeitos aleatórios, como em (2). Se f_X é conhecido, um peso alternativo óbvio é

$$w_i = (nh)^{-1} \frac{K\left(\frac{x - X_i}{h}\right)}{f_X(x)}.$$

Se os pontos da amostra $\left(x_1, ..., x_n\right)$ são fixos, devemos considerar uma forma diferente para o núcleo-estimador da função de densidade, pois a intuição de (4) estaria perdida. Assim uma forma diferente para a "função densidade" seria

$$\hat{f}_X(x_i) = \frac{1}{n\left(x_i - x_{i-1}\right)}.$$

Neste caso, temos o estimador de Priestley-Chao (Priestley e Chao, 1972) para a função de regressão (trocando $f_X\left(x\right)$ por $f_X\left(x_i\right)$) será

$$\hat{m}_{PC}(x) = h^{-1} \sum_{i=1}^{n} \left(x_i - x_{i-1}\right) K\left(\frac{x - x_i}{h}\right) Y_i. \tag{6}$$

Existem outros estimadores na literatura para o caso de efeitos fixos, como o de Gasser e Müller (1984), que não serão adotados aqui, pois fogem do objetivo principal do trabalho. Mais informações podem ser obtidas em Simonoff (1996) e Bowman e Azzalini (1997).

2.2. Regressão Polinomial Local

Supondo o mesmo modelo (2), agora o interesse é estimar a função de regressão $m(x) = E(Y \mid X = x)$ e suas derivadas usando uma amostra aleatória de (X, Y). Essa forma de se obter a estimativa da função de regressão é a polinomial local. Então, se as p primeiras derivadas de $m(x)$ no ponto x_0 existem, nós podemos aproximar a função $m(x)$ por um polinômio de ordem p, dado por:

$$m(x) \approx m(x_0) + m'(x_0)(x - x_0) + \ldots + m^{(p)}(x_0)(x - x_0)^p \,/\, p!$$

onde x_0 é um ponto vizinho a x. Assim devemos obter o valor de $\hat{m}(x; h, p) = \hat{\beta}_0$ através dos mínimos quadrados de:

$$(\hat{\beta}_0, \ldots, \hat{\beta}_p)^T = \arg\min_{\beta} \sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j \left(X_i - x_0 \right)^j \right\}^2 K\left(\frac{X_i - x_0}{h} \right), \qquad (7)$$

onde K é a função densidade simétrica conhecida como função núcleo e h é a janela.

Seja $\mathbf{X}$ uma matriz dada da seguinte forma:

$$X = \begin{bmatrix} 1 & X_1 - x & \cdots & (X_1 - x)^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_n - x & \cdots & (X_n - x)^p \end{bmatrix}$$

assim como $Y = (Y_1, \ldots, Y_n)^T$ e $W = diag\left[K\{(X_1 - x)/h\}, \ldots, K\{(X_n - x)/h\} \right]$. Logo temos que

$$(\hat{\beta}_0, \ldots, \hat{\beta}_p)^T = \left(X^T W X \right)^{-1} X^T W Y$$

é a solução dos mínimos quadrados para (7) e assim definimos $\hat{m}(x) = \hat{\beta}_0$.

3. Análise de Covariância Não-paramétrica

3.1. Comparação de Dois Grupos

Um tema clássico na análise estatística é a comparação de dois ou mais grupos. Para simplificar a notação, iremos restringir, por um momento, ao caso de dois grupos. A extensão para três ou mais grupos será apresentada em 3.2.

No contexto de regressões, considere o seguinte modelo

$$Y_{ij} = f_i\left(t_{ij}\right) + \sigma_i\left(t_{ij}\right)\varepsilon_{ij}, \qquad j = 1,\ldots,n_i, \;\; i = 1,2, \tag{8}$$

onde t_{ij} são valores fixos das medidas, f_i denotam as funções de regressões desconhecidas, $f_i\left(t_{ij}\right) = E\left[Y_{ij}\right]$ e σ_i^2 as funções de variância desconhecidas, $\sigma_i^2\left(t_{ij}\right) = Var\left[Y_{ij}\right]$ da observação j-ésima $\left(j = 1,\ldots,n_i\right)$ no i-ésimo grupo $\left(i = 1,2\right)$. Os erros ε_{ij} são assumidos como variáveis independentes e identicamente distribuídos com média 0 e variância 1. Nosso objetivo é testar a igualdade das funções de regressão f_1 e f_2.

Sob a suposição paramétrica sobre os erros ε_{ij} e as funções f_i e σ_i^2, temos a comum análise de covariância. Sem essas suposições, em particular quando a forma da função f_i não é especificada, temos a análise de covariância não-paramétrica, que vem recebendo bastante atenção na literatura.

Muitos testes para

$$H_0 : f_1 = f_2 \text{ versus } H_1 : f_1 \neq f_2 \tag{9}$$

não podem ser aplicados para o modelo em geral (8), pois assumem que os grupos têm o mesmo tamanho amostral, os regressores seguem a mesma distribuição entre as populações, ou que existe um erro homocedástico, por exemplo, as variâncias σ_i^2 são independentes do regressor t.

O teste apresentado neste trabalho é baseado na idéia de comparar um estimador de "mínimos quadrados" ponderado sob a suposição de igualdade das curvas de regressão com um estimador que é baseado nos estimadores não-paramétricos $\hat{f}_i$ para f_i (sob a hipótese alternativa), exatamente como na análise de covariância paramétrica.

Para motivar o procedimento assuma, por um momento, que as funções de regressão são constantes $f_i(t) \equiv \mu_i$, as funções de variância são constantes e conhecidas $\sigma_i^2(t) = \sigma_i^2$ e os erros ε_{ij} são normalmente distribuídos. Em outras palavras, considere o teste de igualdade de médias $H_0 : \mu_1 = \mu_2$ para duas amostras

$$Y_{ij} \overset{i.i.d}{\sim} N\left(\mu_i, \sigma_i^2\right), \quad j = 1, \ldots, n_i, \; i = 1, 2.$$

O método de máxima verossimilhança nos leva a estimar $\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$ na amostra individual $\left(i = 1, 2\right)$, e

$$\hat{\mu} = a\hat{\mu}_1 + \left(1 - a\right)\hat{\mu}_2, \qquad \text{onde } a = \frac{\sigma_1^{-2} n_1}{\sigma_1^{-2} n_1 + \sigma_2^{-2} n_2},$$

na amostra conjunta (sob H_0). Considerando $\tilde{Y}$ como toda a amostra, no teste da razão de verossimilhança temos

$$\begin{aligned}
\lambda\left(\tilde{Y}\right) &= \frac{L\left(\hat{\mu} \mid \tilde{Y}\right)}{L\left(\hat{\mu}_1, \hat{\mu}_2 \mid \tilde{Y}\right)} \\
&= \frac{\prod_{i=1}^{2} \prod_{j=1}^{n_i} \left(2\pi\sigma_i^2\right)^{-1/2} \exp\left\{-\frac{\left(y_{ij} - \hat{\mu}\right)}{2\sigma_i^2}\right\}}{\prod_{i=1}^{2} \prod_{j=1}^{n_i} \left(2\pi\sigma_i^2\right)^{-1/2} \exp\left\{-\frac{\left(y_{ij} - \hat{\mu}_i\right)}{2\sigma_i^2}\right\}} \\
\lambda\left(\tilde{Y}\right) &= \prod_{i=1}^{2} \prod_{j=1}^{n_i} \exp\left\{-\frac{\left(y_{ij} - \hat{\mu}\right)}{2\sigma_i^2} + \frac{\left(y_{ij} - \hat{\mu}_i\right)}{2\sigma_i^2}\right\}.
\end{aligned}$$

O logaritmo da razão de verossimilhança tem a seguinte forma

$$-\frac{2\ln\left(\lambda\left(\tilde{Y}\right)\right)}{N} = \frac{1}{N} \sum_{i=1}^{2} \sum_{j=1}^{n_i} \left(Y_{ij} - \hat{\mu}\right)^2 \sigma_i^{-2} - \frac{1}{N} \sum_{i=1}^{2} \sum_{j=1}^{n_i} \left(Y_{ij} - \hat{\mu}_i\right)^2 \sigma_i^{-2}, \tag{10}$$

onde $N = n_1 + n_2$ denota o tamanho total da amostra.

Considere no modelo de regressão a classe dos estimadores comuns

$$\tilde{f}(x) = a(x)\,\hat{f}_1(x) + \big(1 - a(x)\big)\,\hat{f}_2(x), \tag{11}$$

onde $\hat{f}_i$ denota o estimador pelo método do núcleo da função de regressão f_i $(i = 1,2)$. Nessa classe, é minimizado o erro quadrado médio assintótico de $\tilde{f}$

$$AMSE\left[\tilde{f}\right] = \left(\frac{a^2(x)\sigma_1^2(x)}{n_1 h r_1(x)} + \frac{\big(1-a(x)\big)^2\sigma_2^2(x)}{n_2 h r_2(x)}\right)\int K^2(u)du,$$

onde h denota o parâmetro de suavização e K a função núcleo. Logo encontramos o peso

$$a(x) = \frac{\sigma_1^{-2}(x)n_1 r_1(x)}{\sigma_1^{-2}(x)n_1 r_1(x) + \sigma_2^{-2}(x)n_2 r_2(x)},$$

que minimiza o $AMSE\left[\tilde{f}\right]$, onde r_i denota a função densidade da i-ésima amostra. Agora, substituímos σ_i^2 e r_i pelos apropriados núcleo-estimadores $\hat{\sigma}_i^2$, $\hat{r}_i$ para $i = 1,2$, e denote por $\hat{f}$ o resultado do estimador conjunto de $\tilde{f}$. Assim, como a estatística de teste para a hipótese (9), analogamente a (10), temos:

$$T_N = \frac{1}{N}\sum_{i=1}^{2}\sum_{j=1}^{n_i}\Big(Y_{ij} - \hat{f}\left(t_{ij}\right)\Big)^2\hat{\sigma}_i^{-2}\left(t_{ij}\right) - \frac{1}{N}\sum_{i=1}^{2}\sum_{j=1}^{n_i}\Big(Y_{ij} - \hat{f}_i\left(t_{ij}\right)\Big)^2\hat{\sigma}_i^{-2}\left(t_{ij}\right). \tag{12}$$

Todos os estimadores serão apresentados na próxima seção, onde reapresentaremos a estatística de teste (12) para o caso geral. Em Munk, Neumeyer e Scholz (2006), foi mostrado que sob a hipótese nula a estatística de teste padronizada

$$N\sqrt{h}\left(T_N - \frac{C}{Nh}\right)$$

é assintoticamente normal com média zero e variância assintótica τ^2, onde C e τ^2 dependem somente da função núcleo K, e são definidos por

$$C = 2K(0) - \int K^2(u)du \quad \text{e}$$

$$\tau^2 = 2\int\big(2K - K * K\big)^2(u)du,$$

onde $*$ denota a convolução de K com ela mesma. Feito os cálculos para $K \sim N(0,1)$, encontramos $C = 0,516$ e $\tau^2 = 0,813$. A seguir iremos apresentar o teste para *k* grupos e apresentaremos as propriedades e suposições para a convergência da estatística de teste.

3.2. Comparação de k Grupos

Nesta seção, iremos apresentar a extensão da estatística T_N definida em (12) para o caso de k amostras, ou seja, trabalharemos com o seguinte modelo

$$Y_{ij} = f_i(t_{ij}) + \sigma_i(t_{ij})\varepsilon_{ij}, \quad j = 1,\ldots,n_i, \quad i = 1,\ldots,k, \tag{13}$$

e as hipóteses seriam

$$H_0 : f_1 = \ldots = f_k \text{ versus } H_1 : f_i \neq f_j \text{ para algum } i \neq j. \tag{14}$$

Adicionalmente algumas notações e suposições devem ser levadas em considerações para garantir a convergência da estatística de teste. Logo assuma que os tamanhos amostrais são

$$\frac{n_i}{N} = \kappa_i + O\left(\frac{1}{N}\right), \quad i = 1,\ldots,k, \tag{15}$$

onde $\kappa_i \in (0,1)$ e $N = \sum_{i=1}^{k} n_i$ é o tamanho total amostral, ou seja, tamanho das amostras muito discrepantes entre uma e outra pode comprometer os resultados. Os pontos fixados t_{ij} podem ser modelados por uma densidade r_i tal que

$$\int_0^{t_{ij}} r_i(t)dt = \frac{j}{n_i}, \quad j = 1,\ldots,n_i, \quad i = 1,\ldots,k. \tag{16}$$

Isso garante que a probabilidade de termos um ponto em cada intervalo entre os t_{ij} são iguais. Detalhes podem ser vistos em Sacks e Ylvisaker (1970). Posteriormente, assumiremos que as densidades r_i e as funções de variância σ_i^2 poderão ser limitadas acima de zero, ou seja,

$$\inf_{t\in[0,1]} r_i(t) > 0, \quad \inf_{t\in[0,1]} \sigma_i^2(t) > 0, \quad i = 1,\ldots,k. \tag{17}$$

Asssume-se que as funções de densidade, regressão e variância são *d*-vezes continuamente diferenciáveis, isto é,

$$r_i, f_i, \sigma_i \in C^d(0,1), \quad i = 1, \ldots, k, \tag{18}$$

onde $d \geq 2$.

Como mencionado na seção anterior, o teste estatístico é baseado nos núcleo-estimadores de f_i e σ_i. Para esse fim, é necessário um núcleo simétrico $K : \mathbb{R} \to \mathbb{R}$, com um suporte compacto e de ordem *d* (mais detalhes em Gasser *et al.*, 1985), ou seja,

$$\frac{(-1)^j}{j!} \int K(u) u^j du = \begin{cases} 1, & j = 0 \\ 0, & 1 \leq j \leq d-1, \\ k_d \neq 0, & j = d \end{cases} \quad \int K^2(u) du < \infty. \tag{19}$$

Assuma que a janela $h = h_N$ satisfaz as seguintes condições

$$Nh^{2d} \to 0 \quad \text{e} \quad Nh^2 / (\log h)^2 \to \infty \ \text{ para } \ N \to \infty. \tag{20}$$

A seguir, apresentaremos os estimadores para r_i, f_i e σ_i, onde $\hat{f}_i$ e $\hat{\sigma}_i$ são baseados nos estimadores de Nadaraya-Watson ou pelo estimador de regressão polinomial local. Para se estimar as densidades r_i usaremos

$$\hat{r}_i(x) = \frac{1}{n_i h} \sum_{j=1}^{n_i} K\left(\frac{x - t_{ij}}{h} \right). \tag{21}$$

O estimador de f_i é definido por

$$\hat{f}_i(x) = \frac{1}{n_i h} \sum_{j=1}^{n_i} K\left(\frac{x - t_{ij}}{h} \right) Y_{ij} \frac{1}{\hat{r}_i(x)}, \quad i = 1, \ldots, k. \tag{22}$$

Seguindo a mesma idéia da seção anterior, tem-se a generalização da estatística de teste (12) para k amostras

$$T_N = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \hat{f}\left(t_{ij}\right) \right)^2 \hat{\sigma}_i^{-2}\left(t_{ij}\right) - \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \hat{f}_i\left(t_{ij}\right) \right)^2 \hat{\sigma}_i^{-2}\left(t_{ij}\right), \tag{23}$$

onde o estimador comum de f é obtido como (sob a hipótese nula),

$$\hat{f}(x) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} K\left(\frac{x-t_{ij}}{h}\right) Y_{ij}\hat{\sigma}_i^{-2}\left(t_{ij}\right)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i} K\left(\frac{x-t_{ij}}{h}\right)\hat{\sigma}_i^{-2}\left(t_{ij}\right)}. \tag{24}$$

Finalmente, as variâncias σ_i^2 têm que ser estimadas por um método não-paramétrico em geral. Esse estimador foi proposto com um espírito similar aos estimadores de Ruppert *et al.* (1997), Fan e Yao (1998) e Härdle & Tsybakov (1997). Assim, neste contexto, define-se

$$\hat{\sigma}_i^2(x) = \frac{1}{n_i h}\sum_{j=1}^{n_i} K\left(\frac{x-t_{ij}}{h}\right)\left(Y_{ij} - \hat{f}_i(t_{ij})\right)^2 \frac{1}{\hat{r}_i(x)}, \quad i = 1,\ldots,k. \tag{25}$$

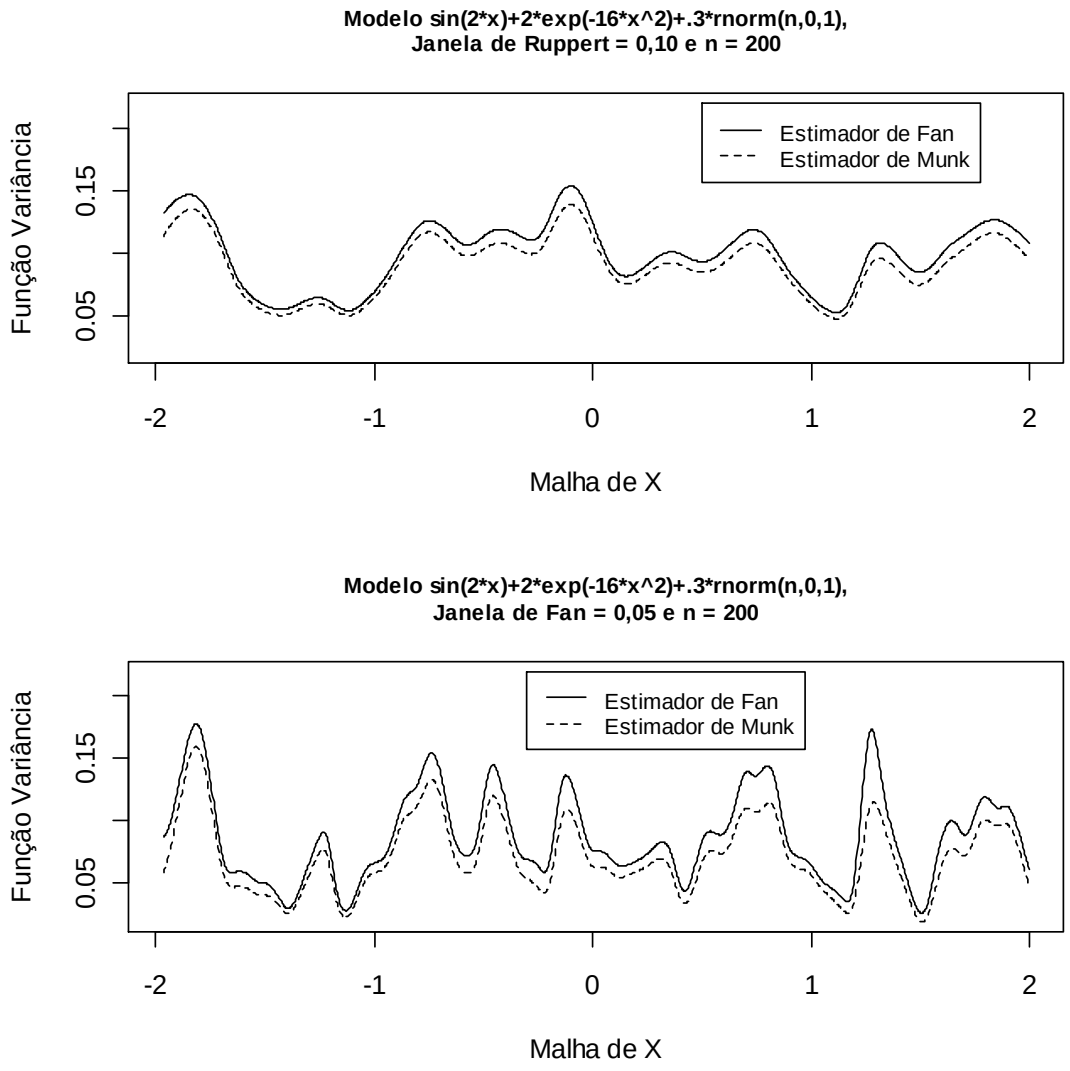

Figura 1: Gráfico para a função variância.

Outro estimador é o de Fan e Gijbels (1995) que será apresentado na seção 4.2 em (38). Ambos apresentam valores bem próximos, sendo que a variância dada em (38) é sempre maior que em (25), como podemos ver na figura 1. Isso devido ao denominador da função variância de Fan ser menor que a de Munk em (25). O comportamento diferente das funções variâncias para cada gráfico se deve ao valor da janela ótima. O primeiro foi escolhido pelo método de Ruppert (seção 4.1) onde o valor encontrado é igual a 0,11 e o segundo pelo de Fan (seção 4.2) igual a 0,05. Logo, com uma janela menor, a curva tem o comportamento mais suavizado. Outros estimadores de $\hat{\sigma}_i^2$ estão sendo estudados por Atuncar (2009).

O Teorema 1 fornece a distribuição assintótica da estatística de teste T_N.

Teorema 1

Assuma o modelo (13), onde os erros ε_{ij} *são variáveis aleatórias com variância* $\mathrm{var}\left(\varepsilon_{ij}\right)=1$ *e* $E\left[\varepsilon_{ij}^4\right]\le M<\infty\ \forall i,j$. *Então sob as suposições (15)-(20) e sob a hipótese nula, com* T_N *definida em (24), temos que*

$$U_N = N\sqrt{h}\left(T_N - \frac{C}{Nh}\right) \xrightarrow[N\to\infty]{D} \mathrm{N}\left(0,\tau^2\right),$$

onde

$$\tau^2 = 2(k-1)\int \left(2K - K*K\right)^2(u)du,$$

e $*$ *denota a convolução de K com ela mesma. A constante C é definida como* $C = 2K(0) - \int K^2(u)du$.

Para o teste de hipótese (14), rejeitamos H_0 com um nível α quando

$$\frac{N\sqrt{h}\left(T_N - \frac{C}{Nh}\right)}{\tau} > u_{1-\alpha}, \qquad (26)$$

onde $u_{1-\alpha} = \Phi^{-1}(1-\alpha)$ denota o quantil $\left(1-\alpha\right)$ de uma distribuição normal padrão. Note que *C* e τ^2 são constantes conhecidas. A consistência do procedimento de teste (26) para a hipótese alternativa H_1 segue do próximo resultado.

Teorema 2

Assuma que $f_i \neq f_j$*, sob* H_1*, em um conjunto de probabilidade positivo na reta para algum* i *e* j *em* $\{1,\ldots,k\}$*. Sob as suposições do Teorema 1, temos*

$$R_N = \sqrt{N}\left(T_N - \mu\right) \xrightarrow[N\to\infty]{D} \mathrm{N}\left(0, \gamma^2\right), \tag{27}$$

onde as constantes são definidas como

$$\mu = \sum_{j=1}^{k} \sum_{\substack{l=1 \\ l<j}}^{k} \int \left(f_j - f_l\right)^2 (x) \frac{\sigma_l^{-2}(x)\kappa_l r_l(x)\sigma_j^{-2}(x)\kappa_j r_j(x)}{\sum_{l=1}^{k}\sigma_l^{-2}(x)\kappa_l r_l(x)} dx \quad \text{e} \quad \gamma^2 = 4\mu. \tag{28}$$

As provas dos Teoremas 1 e 2 podem ser encontradas no apêndice do artigo de Munk, Neumeyer e Scholz (2006).

O Teorema 2 pode ser utilizado em vários caminhos. Primeiro, uma aproximação do poder pode ser obtida via

$$\begin{aligned} P_{H_1}\left(\frac{N\sqrt{h}\left(T_N - \frac{C}{Nh}\right)}{\tau} > u_{1-\alpha}\right) &= \Phi\left(\frac{\mu\sqrt{N}}{\gamma} - \frac{\tau u_{1-\alpha}}{\gamma\sqrt{Nh}} - \frac{C}{\gamma\sqrt{N}h}\right) + o(1) \\ &= \Phi\left(\frac{\mu}{\gamma}\sqrt{N}\right) + o(1). \end{aligned} \tag{29}$$

Segundo, um simples intervalo de confiança $(1-\alpha)$ unilateral para a medida de discrepância, μ em (28) entre as funções f_i $(i = 1,\ldots,k)$ é obtida como

$$CI_{1-\alpha} = \left[0, T_N + \sqrt{T_N c + \frac{c^2}{4}} + \frac{c}{2}\right] \tag{30}$$

onde $c = 4u^2_{1-(\alpha/2)} \big/ N$ e para $\alpha : 0 < \alpha < 0,5$.

A convergência de U_N para a distribuição normal é lenta para tamanhos amostrais finitos, assim Munk, Neumeyer e Scholz (2006) fazem um estudo usando "*Wild Bootstrap*". Tal método será descrito no capítulo 5.

Na próxima seção descreveremos o procedimento *bootstrap* adotado por Hall a Hart (1990) para a comparação entre duas médias em um foco não-paramétrico.

3.3. Comparação com o teste de Hall e Hart (1990)

Antes de trabalharmos com Dette e Neumeyer (2001) e Munk, Neumeyer e Scholz (2006), trabalhamos com teste de comparação entre médias em regressão não-paramétrica, mais simples que o apresentado na seção anterior. Assim, iremos apresentar somente o teste para duas amostras e mais detalhes sobre o caso geral pode ser encontrado em Hall e Hart (1990). Assuma que os dados seguem a seguinte forma $\{(x_i, Y_i, Z_i), 1 \le i \le n\}$, e são estruturados de acordo com os modelos

$$Y_i = f(x_i) + \varepsilon_i, \quad Z_i = g(x_i) + \eta_i, \quad 1 \le i \le n, \tag{31}$$

onde $\varepsilon_1, \ldots, \varepsilon_n$ e $\eta_1, \ldots, \eta_n$ denotam os erros aleatórios independentes, sendo que os $\varepsilon_i's$ possuem uma distribuição e os $\eta_i's$ possuem outra. As hipóteses do teste são

$$H_0 : f = g \quad \text{versus} \quad H_1 : f \neq g. \tag{32}$$

Seja $\sigma^2 = \text{var}(\varepsilon_i)$ e $\tau^2 = \text{var}(\eta_i)$. Neste caso os pontos $x_i's$ são fixos, e sem perda de generalidade, definidos em um intervalo $(0,1)$, embora essa suposição possa ser removida.

Defina $D_i = Y_i - Z_i$ para $1 \le i \le n$, e seja $D_i = D_{i-n}$ para $n+1 \le i \le n+m$, onde $m = [np]$, a parte inteira de np, com $0 < p < 1$ fixo. Assim a estatística de teste é dada como

$$S_n = \left[\sum_{j=0}^{n-1} \left(\sum_{i=j+1}^{j+m} D_i \right)^2 \right] \left[n \sum_{i=1}^{n-1} \frac{(D_{i+1} - D_i)^2}{2} \right]^{-1}. \tag{33}$$

O valor de S_n tende a ser moderado quando a hipótese nula em (31) é verdadeira, e grande quando ela é falsa. O quanto "grande" será o valor da estatística de teste será determinado ou por uma aproximação assintótica ou por uma aproximação *bootstrap*, os quais ambos serão apresentados a seguir.

Considere que $\{W(t), 0 \le t \le 1\}$ denota um Movimento Browniano, e estenda W para o intervalo $(0,2)$ definindo $W(t) = W(t-1) + W(1)$ para $1 < t < 2$. A distribuição assintótica da estatística S_n, sobre a hipótese nula, é dada por

$$S_n \to S = \int_0^1 \left[W(t+p) - W(t) \right]^2 dt \tag{34}$$

sendo a convergência em distribuição ($n \to \infty$) e relembrando que p é fixado no intervalo $(0,1)$.

Para explicar o comportamento de S_n sob a hipótese alternativa H_1, defina, para cada n e $i = 1, \ldots, n$, x_i como o quantil i/n de uma distribuição com densidade r. Estendendo f para o intervalo $[0,2)$, e o mesmo para g e r, e se f e g são limitados e contínuos no intervalo $(0,1)$, temos que

$$n^{-1} S_n \to s(f,g) \Big/ \left(\sigma^2 + \tau^2 \right), \tag{35}$$

quase certamente, onde

$$n^{-3} \sum_{j=0}^{n-1} \left(\sum_{i=j+1}^{j+m} D_i \right)^2 \to s(f,g) = \int_0^1 \left(\int_t^{t+p} \left[f(u) - g(u) \right] r(u)\, du \right)^2 dt, \tag{36}$$

com probabilidade 1 quando $n \to \infty$. Mais detalhes podem ser encontrados em Hall e Hart (1990).

A fórmula (34) descreve o limite S de S_n sob a hipótese nula. Em princípio, um nível assintótico α do teste pode ser obtido usando esse resultado. Por integração numérica ou métodos Monte Carlo, calcule u tal que $\Pr(S > u) = \alpha$, e rejeita-se H_0 quando $S_n > u$. O erro do nível para esse teste é da ordem de n^{-1}. Entretanto, o *bootstrap* fornece um erro para o nível do teste da ordem de n^{-2} e não é mais trabalhoso para ser aplicado do que o teste assintótico. O teste *bootstrap* possui uma aproximação exata do quantil u_n tal que, sob a hipótese nula, $\Pr(S_n > u_n) = \alpha$.

O método *bootstrap* funciona da seguinte maneira. Seja $\bar{Y}-\bar{Z}=n^{-1}\sum_{i=1}^{n}\left(Y_i-Z_i\right)$ e $d_i=Y_i-Z_i-\left(\bar{Y}-\bar{Z}\right)$. Do conjunto $\left\{d_1,\ldots,d_n\right\}$, retiramos aleatoriamente e com reposição uma reamostra $\left\{d_1^*,\ldots,d_n^*\right\}$. Logo defina

$$S_n^*=\left[\sum_{j=0}^{n-1}\left(\sum_{i=j+1}^{j+m}d_i^*\right)^2\right]\left[n\sum_{i=1}^{n-1}\frac{\left(d_{i+1}^*-d_i^*\right)^2}{2}\right]^{-1} \tag{37}$$

o qual é idêntico ao definido em (33), exceto que D_i é substituído por d_i^*. Repetindo a reamostragem, calcule $\hat{u}_n$ tal que $\Pr\left(S_n^*>\hat{u}_n\mid\Omega\right)=\alpha$, onde $\Omega=\left\{\left(x_i,Y_i,Z_i\right),1\le i\le n\right\}$ denota a amostra. O teste *bootstrap* rejeita H_0 em favor de H_1 se $S_n>\hat{u}_n$.

Quando H_0 é verdadeira, a distribuição empírica $\hat{F}$ de $\left\{d_1,\ldots,d_n\right\}$ é uma boa aproximação para a distribuição nula de D_i, e assim o teste *bootstrap* é uma boa aproximação para o teste exato. Se H_1 é verdadeira, os dados $\left\{d_1,\ldots,d_n\right\}$ não garantirão que $\hat{F}$ se aproxima bem da distribuição nula de D_i, mas garantirão que o *bootstrap* proporcione uma boa aproximação para o teste em grandes amostras. Mais detalhes sobre as provas e a generalização do teste para mais de duas regressões, podem ser obtidas em Hall e Hart (1990).

4. Métodos de Seleção Automática da Janela

4.1. Método 1 – Ruppert, Sheather e Wand (1995)

Seja $\hat{m}(\cdot; h, p)$ o estimador de $m(x)$ dado por uma determinada janela h e por uma regressão polinomial de ordem p. Assim temos que o erro quadrado médio integrado de $\hat{m}(\cdot; h, p)$ dada a amostra $X_1, \ldots, X_n$ é igual a:

$$\mathrm{MISE}\left\{\hat{m}(\cdot; h, p) \mid X_1, \ldots, X_n\right\} = E\left[\int \left\{\hat{m}(x; b) - m(x)\right\}^2 f(x)\, dx \mid X_1, \ldots, X_n\right].$$

Usando a seguinte notação $\mu_l(L) = \int u^l L(u)\, du$ e $R(L) = \int L(u)^2\, du$, onde L é uma função qualquer e que ambas integrais convergem, podemos reescrever o erro quadrado médio integrado de $m(x)$ da seguinte forma, para p ímpar (ver o teorema 4.1 de Ruppert e Wand, 1994):

$$\begin{aligned}\mathrm{MISE}\left\{\hat{m}(x; h, p) \mid X_1, \ldots, X_n\right\} = {} & n^{-1}h^{-1}R(K_p)\int_S \sigma^2(x)\, dx \\ & + h^{2p+2}\left\{\mu_{p+1}(K)/(p+1)!\right\}^2 \int m^{(p+1)}(x)^2 f(x)\, dx \\ & + o_p\left(n^{-1}h^{-1} + h^{2p+2}\right),\end{aligned}$$

onde $S \subset \Re$ é suporte da variável aleatória X e K_p é função núcleo de ordem $p+1$, onde $K_{0,p} = K_p$. Para isso definimos $K_{r,p}(u) = r!\left\{\left|M_{r,p}(u)\right| / \left|N_p\right|\right\} K(u)$ onde N_p é uma matriz $(p+1)\times(p+1)$ tendo os elementos (i, j) iguais a $\int u^{i+j-2} K(u) du$ e $M_{r,p}(u)$ igual a N_p, exceto na linha $(r+1)$ que é substituída por $(1, u, \ldots, u^p)$.

A aproximação do $\mathrm{MISE}\left\{\hat{m}(x; h, p)\right\}$ quando p não é ímpar, contém complicações nas contas para a janela ótima, por isso optamos por considerar o valor de p ímpar. Assim o valor assintótico da janela ótima aproximada é igual a:

$$h_{MISE} \simeq \left[\frac{(p+1)(p!)^2 R(K_p)\int_S \sigma^2(x) dx}{2n\mu_{p+1}(K_p)^2 \int_S m^{(p+1)}(x)^2 f(x) dx}\right]^{1/(2p+3)}.$$

A estratégia adotada por Ruppert, Sheather e Wand (1995) é utilizar a técnica plug-in para substituir as integrais desconhecidas. Assim iremos restringir os valores de $p=3$ e $r=2$ ($s=2$, o qual aparecerá mais a frente) e no caso de $\sigma^2(x)=\sigma^2$, ou seja, erros homocedásticos. Por simplicidade, definimos o suporte de X igual a $S=[a,b]$. Toda parte assintótica será omitida, pois foge do objetivo do nosso trabalho e para não deixá-lo muito extenso. Assim mais detalhes podem ser encontrados em Ruppert, Sheather e Wand (1995).

Considere que, a janela ótima assintótica, pode ser escrita como:

$$h_{AMISE}=C_1(K)\left[\frac{\sigma^2(b-a)}{\theta_{22}n}\right]^{1/5},$$

onde $C_1(K)$ é uma constante que depende somente da função núcleo e que

$$\theta_{rs}=\int m^{(r)}(x)m^{(s)}(x)\,f(x)dx, \qquad r,s\geq 0, \quad r+s \text{ par.}$$

Um estimador do tipo núcleo de θ_{rs} é dado por:

$$\hat{\theta}_{rs}(g)=n^{-1}\sum_{i=1}^{n}\hat{m}^{(r)}(X_i;g)\hat{m}^{(s)}(X_i;g),$$

porém necessitamos do valor de g. Assim, um valor assintótico para ele será

$$g_{AMSE}=C_2(K)\left[\frac{\sigma^2(b-a)}{|\theta_{24}|n}\right]^{1/7},$$

onde $C_2(K)$ é também uma constante que depende somente da função núcleo. Em h_{AMISE}, temos que substituir σ^2 por um estimador, igual a

$$\hat{\sigma}_p^2(\lambda)=\nu^{-1}\sum_{i=1}^{n}\left\{Y_i-\hat{m}\left(X_i;\lambda,p\right)\right\}^2,$$

sendo $\nu=n-2\sum_i w_{ii}+\sum\sum_{ij}w_{ij}^2$ e $w_{ij}=e_1^T\left(X^TWX\right)^{-1}X^TWe_j$, onde e_j é um vetor contento 1 na posição j e zeros nas demais posições. Contudo, temos que encontrar o valor de λ, o qual pode ser aproximado assintoticamente por

$$\lambda_{AMSE} = C_3(K)\left[\frac{\sigma^4(b-a)}{\theta^2_{p+1,p+1}n^2}\right]^{1/(4p+5)}.$$

Para $\hat{\sigma}^2_p(\lambda)$ devemos usar um valor de p menor ou igual a três (o valor escolhido antes) e também ímpar, ou seja, $p=1$. Assim temos que

$$\lambda_{AMSE} = C_3(K)\left[\frac{\sigma^4(b-a)}{\theta^2_{22}n^2}\right]^{1/9},$$

onde $C_3(K)$ é uma constante.

Note que tanto em g_{AMSE} quanto em λ_{AMSE} aparecem valores também desconhecidos (σ^2 e θ_{rs}). Se continuarmos a utilizar a mesma regra que usamos em h_{AMISE}, o processo ficará indefinido. Então, seguindo Härdle e Marron (1993), dividi-se a amplitude de X em M blocos e ajusta-se um modelo para cada bloco. Essa divisão pode ser em blocos de mesmo tamanho ou em blocos igualmente espaçados. A opção usada neste trabalho será a primeira, pois tem a vantagem de adaptar melhor os dados de forma não uniforme e diminuir a chance de superestimação.

Seja M o número de sub-amostras e seja χ_j a j-ésima sub-amostra dos X_i's. Se M divide n e $t = n/M$, então $\chi_j = \left\{X_{((j-1)t+1)}, \ldots, X_{(jt)}\right\}$. Agora seja $\hat{m}^Q_j(x)$ o estimador de mínimos quadrados quártico obtido através dos valores de X_i da sub-amostra χ_j. Assim temos o estimador para θ_{rs} igual a

$$\hat{\theta}^Q_{rs}(M) = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{M}\left(\hat{m}^Q_j\right)^{(r)}(X_i)\left(\hat{m}^Q_j\right)^{(s)}(X_i)1_{\{X_i\in\chi_j\}}.$$

Similarmente, o estimador para σ^2 é dado por

$$\hat{\sigma}^2_Q(M) = (n-5M)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{M}\left\{Y_i - \hat{m}^Q_j(X_i)\right\}^2 1_{\{X_i\in\chi_j\}}.$$

Esses estimadores requerem uma regra para a escolha de M. A regra de C_p de Mallows (1973) foi adequada para resolver o problema. A escolha de $\hat{M}$ vem do conjunto $\{1,2,\ldots,M_{\max}\}$, o qual deve minimizar

$$C_p(M) = RSS(M) \Big/ \left\{ RSS(M_{\max}) \Big/ \left(n - 5M_{\max}\right) \right\} - \left(n - 10M\right),$$

onde $RSS(M)$ é a soma do quadrado dos resíduos baseado no ajuste dos blocos sobre M blocos. Para reduzir as chances de superestimação, defini-se $M_{\max}$ da seguinte forma

$$M_{\max} = \max\left\{ \min\left([n/20], M^* \right), 1 \right\}$$

para algum M^* inteiro positivo. A escolha de M^* para funções de regressão não influi muito nos resultados e baseado em estudos anteriores, opta-se por $M^* = 5$.

Uma forma opcional para obter θ_{rs} é o truncamento dos dados em $100\alpha\%$ nas fronteiras, para α pequeno. A razão para isso é que a regressão polinomial local com derivadas de alta ordem pode conter estimativas que variam muito, próximo das fronteiras. Logo, no caso em que o suporte dos X_i's está no intervalo $[a,b]$, temos que $\hat{\theta}_{rs}(g)$ pode ser substituído por

$$\hat{\theta}_{rs}^{\alpha}(g) = n^{-1} \sum_{i=1}^{n} m^{(r)}(X_i) m^{(s)}(X_i) 1_{\left\{ (1-\alpha)a + \alpha b < X_i < \alpha a + (1-\alpha)b \right\}}.$$

Assim, a seleção da janela automática através do plug-in direto segue a seguinte ordem:

1. Encontre os valores para $\hat{\theta}_{24}^{Q}(\hat{M})$ e $\hat{\sigma}_{Q}^{2}(\hat{M})$ baseados nos ajustes dos blocos quárticos onde $\hat{M}$ é encontrado através da técnica de Mallows (1973).
2. Estime θ_{22} usando $\hat{\theta}_{22}^{0,05}\left(\hat{g}_{AMSE}\right)$, onde

$$\hat{g}_{AMSE} = C_2(K) \left[\frac{\hat{\sigma}_{Q}^{2}(\hat{M})(b-a)}{\left| \hat{\theta}_{24}^{Q}(\hat{M}) \right| n} \right]^{1/7}$$

   e estime $\hat{\sigma}^2$ usando $\hat{\sigma}_{1}^{2}\left(\hat{\lambda}_{AMSE}\right)$, onde

$$\hat{\lambda}_{AMSE} = C_3(K) \left[\frac{\hat{\sigma}_{Q}^{4}(\hat{M})(b-a)}{\hat{\theta}_{22}^{0,05}\left(\hat{g}_{AMSE}\right)^2 n^2} \right]^{1/9}$$

3. A janela selecionada é

$$\hat{h}_{DPI} = C_1(K)\left[\frac{\hat{\sigma}_1^2\left(\hat{\lambda}_{AMSE}\right)(b-a)}{\hat{\theta}_{22}^{0,05}\left(\hat{g}_{AMSE}\right)n}\right]^{1/5}.$$

São fornecidos na Tabela 1 os valores das constantes $C_1(K)$, $C_2(K)$ e $C_3(K)$ para as funções núcleo Normal padrão, Epanechnikov e Biponderada. A constante $C_2(K)$ é igual a $C_2^I(K)$, quando $\theta_{24} < 0$ e igual a $C_2^{II}(K)$, quando $\theta_{24} > 0$. As distribuições de Epanechnikov e Biponderada são dadas por, respectivamente:

$$K(x) = 0,75\left(1-x^2\right)I_{[-1,1]}(x),$$

$$K(x) = \frac{15}{16}\left(1-x^2\right)^2 I_{\{|x|\leq 1\}}(x).$$

Tabela 1: Valores das constantes que dependem da função núcleo.

Constantes	Normal Padrão	Epanechnikov	Biponderada
$C_1(K)$	$\left\{1/\left(2\sqrt{\pi}\right)\right\}^{1/5}$	$15^{1/5}$	$35^{1/5}$
$C_2^I(K)$	$\left\{3/\left(8\sqrt{\pi}\right)\right\}^{1/7}$	$315^{1/7}$	$(8505/13)^{1/7}$
$C_2^{II}(K)$	$\left\{15/\left(16\sqrt{\pi}\right)\right\}^{1/7}$	$(1575/2)^{1/7}$	$(42525/26)^{1/7}$
$C_3(K)$	$\left\{4\left(1/2+2\sqrt{2}-(4/3)\sqrt{3}\right)/\sqrt{2\pi}\right\}^{1/9}$	$4725^{1/9}$	$(322665/32)^{1/9}$

O mais interessante desse método é o gasto computacional que é pequeno e, além disso, ele pode ser encontrado nos pacotes do programa R®, chamado de "KernSmooth". Um método alternativo é apresentado em Fan e Gijbels (1995), descrito a seguir.

4.2. Método 2 – Fan e Gijbels (1995)

Para o método apresentado por Fan e Gijbels (1995), usaremos o MSE (erro quadrado médio) da estimativa de $m(x)$. Logo o valor da janela que minimiza o MSE é dado por

$$h_{MSE}\left(x_0\right) = \left\{\frac{a_0\sigma^2\left(x_0\right)}{2\left(p+1\right)b_0^2\beta_{p+1}^2 n f_X\left(x_0\right)}\right\}^{1/(2p+3)},$$

onde a_0 é o primeiro elemento da diagonal da matriz $S^{-1}S^*S^{-1}$, onde S e S^* são matrizes $(p+1)\times(p+1)$ com os elementos (i,j) iguais a s_{i+j-2} e v_{i+j-2}, respectivamente, onde $s_j = \int u^j K(u)\,du$ e $v_j = \int u^j K^2(u)\,du$. Agora, b_0 é o primeiro elemento do vetor $p+1$ dado por $S^{-1}\left(s_{p+1},\ldots,s_{2p+1}\right)^T$.

Como no caso em Ruppert, Wand e Sheather (1995), existem valores que são desconhecidos e usando o método plug-in para substituir esses valores, temos uma estimativa semelhante ao primeiro caso apresentado. Contudo, usaremos um método de minimização da função de soma do quadrado dos resíduos dada por:

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

onde V é o primeiro elemento da diagonal da matriz $S_n^{-1}S_n^*S_n^{-1}$, com $S_n = X^TWX$ e $S_n^* = X^TW^2X$. Já o estimador para $\sigma^2(x)$ é dado por:

$$\hat{\sigma}^2(x_0) = \frac{1}{tr\left\{W - WX\left(X^TWX\right)^{-1}X^TW\right\}}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 K\left(\frac{X_i - x_0}{h}\right), \qquad (38)$$

com $\left(\hat{Y}_1,\ldots,\hat{Y}_n\right) = X\hat{\beta}$. A Intuição por trás da função RSC é que quando h é muito pequeno, o valor de V é alto, pois representa a variância. Já quando h é muito grande, tanto o vício quanto a soma dos quadrados dos resíduos de $\hat{\sigma}^2(x_0)$ também terá o seu valor alto. Logo, o RSC sofre alterações bruscas para ambos os extremos de h.

Assim, a janela ótima, dada a amostra $(X_1,\ldots X_n)$ no intervalo $[c,d]$, é o valor que minimiza a versão integrada do RSC:

$$IRSC(h) = \int_{[c,d]} RSC(y;h)\,dy.$$

Em simulações realizadas anteriormente por Fan e Gijbels (1995), é mostrado que esse procedimento é bom, porém a sua taxa de convergência é lenta. Então é feito um processo com dois estágios, sendo que o primeiro é para a escolha da janela piloto e o segundo para a janela ótima. Contudo, optamos por fazer somente o primeiro estágio devido ao seu alto custo computacional, pois agregando os dois passos, demandaríamos de mais tempo para este

trabalho. Em projetos no futuro, pretendemos utilizar os dois estágios e tentar de alguma forma minimizar o custo computacional.

5. *Wild Bootstrap*

Nesta seção será apresentado um método *boostrap* para a estatística de teste T_N (23). O uso do teste assintoticamente (U_N) não trás bons resultados nem para pequenas amostras quanto para grandes amostras, como pode ser observado na figura 2, em 1000 simulações numa amostra de tamanho $n=100$. Isso ocorre devido à convergência lenta e assim um procedimento *boostrap* é realizado para se obter a distribuição da estatística T_N.

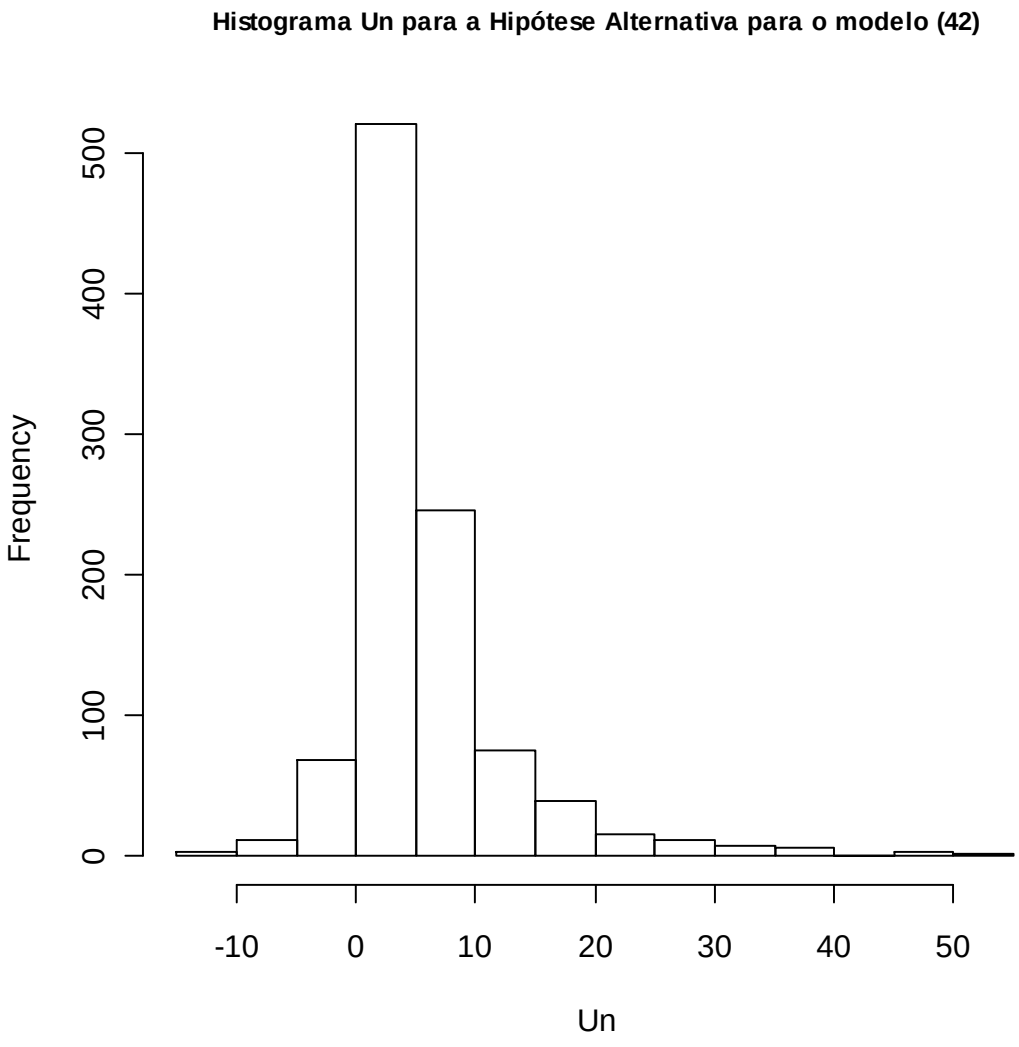

Figura 2: Estudo da convergência da estatística de teste U_N.

Conhecido na literatura como *wild bootstrap*, desenvolvido por Liu (1988) seguindo uma sugestão de Wu (1986) e Beran (1986), é um método bem difundido e que promove um

refinamento dos erros heterocedásticos em modelos de regressão. Sob a hipótese nula, considere o seguinte erro estimado:

$$\hat{\varepsilon}_{ij} = Y_{ij} - \hat{f}\left(t_{ij}\right), \quad j = 1,...,n_i, \;\; i = 1,2,$$

onde $\hat{f}$ é o estimador da regressão comum (assumindo que $f_1 = f_2$). Agora defina uma variável aleatória V_{ij} $i.i.d.$ e independente da amostra $\left\{Y_{ij}\right\}$, com as seguintes esperanças:

$$E\left[V_{ij}\right] = 0, \tag{39}$$

$$E\left[V_{ij}^2\right] = 1 \text{ e} \tag{40}$$

$$E\left[V_{ij}^3\right] = 1. \tag{41}$$

Somente as condições (39) e (40) já garantem a consistência do *bootstrap* para os estimadores (para mais detalhes vide Liu, 1988 e Mammem, 1993). Mas usando a condição chave (41), garantimos as propriedades de segunda ordem do *bootstrap* desenvolvido por Wu (1986). Então o terceiro momento de U_N e todos três primeiros momentos de $N\sqrt{h}\left(T_N - \frac{C}{Nh}\right)\Big/\tau$ serão estimados corretamente para um $O\left(N^{-1}\right)$ por esse *bootstrap*.

Uma das formas possíveis de se encontrar a distribuição da variável V_{ij} é assumirmos que é dada da seguinte forma:

$$V_{ij} = \begin{cases} a, & \text{com probabilidade } p \\ b, & \text{com probabilidade } 1-p. \end{cases}$$

Outras formas de se obter a distribuição de V_{ij} pode ser encontrada em Liu (1998). Logo, teremos um sistema de equações com três equações e três incógnitas, dadas por

$$\begin{cases} ap + b(1-p) = 0 \\ a^2 p + b^2 (1-p) = 1. \\ a^3 p + b^3 (1-p) = 1 \end{cases}$$

Resolvendo esse sistema, obtemos a seguinte distribuição para V_{ij}

$$V_{ij} = \begin{cases} \left(1-\sqrt{5}\right)/2, & \text{com probabilidade } \left(\sqrt{5}+1\right)/2\sqrt{5} \\ \left(1+\sqrt{5}\right)/2, & \text{com probabilidade } \left(\sqrt{5}-1\right)/2\sqrt{5} \end{cases}$$

Assim define-se a observação *bootstrap* como

$$Y_{ij}^{*} = \hat{f}\left(t_{ij}\right) + V_{ij}\hat{\varepsilon}_{ij}, \quad j = 1,...,n_i, \quad i = 1,2,$$

e denote por T_N^{*} a estatística de teste definida em (23), mas baseada na amostra *bootstrap* $\left\{Y_{ij}^{*}\right\}$. Um teste assintótico com nível α rejeitará a hipótese nula sempre que a estatística T_N (baseada na amostra original $\left\{Y_{ij}\right\}$) é maior que o quantil $\left(1-\alpha\right)$ da distribuição de T_N^{*} condicionada à amostra $\left\{Y_{ij}\right\}$. Mais detalhes sobre esse procedimento pode ser obtido em Mammen (1993) e Härdle e Mammen (1990).

6. Análise Prática

Neste capítulo apresentaremos alguns resultados das simulações feitas para o nível e o poder da estatística de teste T_N. Baseados nos artigos que trabalhamos, foram geradas 1000 amostras de tamanhos 10, 20, 50 e 100 e para cada amostra, foram geradas 200 reamostras *bootstrap.* Para efeito de comparação, fizemos simulações para o teste de Munk, Neumeyer e Scholz (2006) e para o teste *bootstrap* de Hall e Hart (1990). Salientamos que essa comparação se deve somente a título de motivação, ou seja, como já havíamos trabalhado com a estatística S_n anteriormente, optamos por compará-la com a estatística T_N.

Ainda foi feita uma re-análise de um experimento retirado de Ratkowsky (1983), que consiste em obter uma relação entre rendimento por cebola (peso por cebola) e a densidade da plantação (cebolas por m^2). Em Houghton (1999), retiramos as informações de um estudo sobre o fluxo líquido de carbono na atmosfera devido às mudanças no solo para a utilização na agropecuária e comparamos esse fluxo entre as diferentes regiões.

6.1. Simulações

Nesta seção faremos uma avaliação do teste aqui apresentado, usando a idéia principal de Munk, Neumeyer e Scholz (2006), utilizando a escolha da janela de Ruppert, Sheather e Wand (1995) e a regressão polinomial local. Em simulações, percebemos que o gasto computacional com o método do Fan e Gijbels (1995) era muito grande e que a variância do estimador da janela ótima também era maior que no método de Ruppert, Sheather e Wand (1995), na maioria dos casos. Lembrando que utilizamos apenas o primeiro estágio. Mas reafirmamos que o gasto computacional é muito grande, logo se usarmos mais de um estágio para a escolha de janela, esse custo será maior ainda.

Para avaliarmos o poder e o nível do teste, simulamos 1000 amostras e em cada uma delas foram reamostradas 200 vezes para a utilização da técnica *wild boostrap*. Optamos por utilizar modelos que vinham sendo utilizados na literatura aqui trabalhada. Além disso, fixamos o valor do desvio padrão, devido ao uso do método de Ruppert, Sheather e Wand (1995) para a estimação da janela ótima, apesar de que, a nossa intenção era de fazer o uso de modelos não homogêneos. Assim, os modelos escolhidos foram:

$$\text{Caso 1: } f_1(x) = 0,04x + 1 = f_2(x), \qquad \sigma = 0,015 \tag{42}$$

Caso 2: $f_1(x) = seno(2x) + 2\exp(-16x^2) = f_2(x), \qquad \sigma = 0,3$ (43)

Caso 3: $f_1(x) = seno(2\pi x) = f_2(x), \qquad \sigma = 0,1$ (44)

Caso 4: $f_1(x) = \exp(x) = f_2(x), \qquad \sigma = 0,05$ (45)

Os gráficos dessas funções podem ser vistos na figura 3. Em (42) e (43), x foi gerado aleatoriamente através de uma Uniforme no intervalo (-2,2) e em (44) e (45) através de uma Uniforme no intervalo (0,1). Em todos os casos, os erros aleatórios foram gerados através de um Normal padrão. Na tabela 2 temos os resultados dessas simulações para o nível e percebemos que o teste possui um nível bom, próximo do $\alpha = 0,05$.

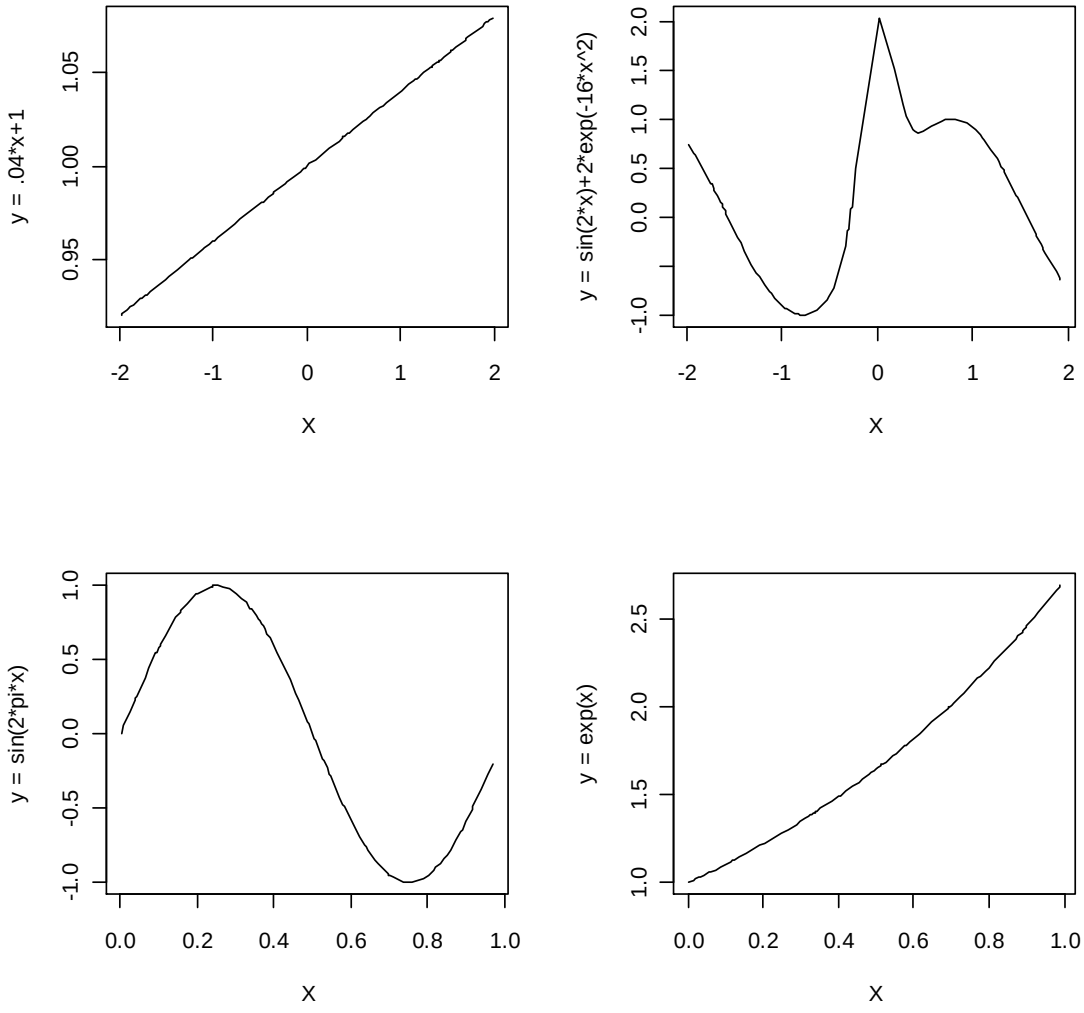

Figura 3: Modelos para as simulações para o teste T_N.

(n_1, n_2)	Caso 1 (42)	Caso 2 (43)	Caso 3 (44)	Caso 4 (45)
(10,10)	0,040	0,052	0,039	0,048
(20,20)	0,040	0,033	0,021	0,036
(25,50)	0,042	0,056	0,035	0,045
(50,50)	0,058	0,059	0,045	0,065
(100,100)	0,068	0,060	0,060	0,051

Tabela 2: Avaliação do nível do teste T_N.

Agora as funções para o poder do teste foram as seguintes:

$$\text{Caso 5: } f_1(x) = 0{,}1 + 0{,}04x, \qquad f_2(x) = 0{,}04x, \qquad \sigma = 0{,}015 \tag{46}$$

$$\text{Caso 6: } f_1(x) = seno(2x) + 2\exp(-16x^2), \qquad f_2(x) = seno(2x), \qquad \sigma = 0{,}3 \tag{47}$$

$$\text{Caso 7: } f_1(x) = seno(2\pi x), \qquad f_2(x) = seno(2\pi x) - x, \qquad \sigma = 0{,}1 \tag{48}$$

$$\text{Caso 8: } f_1(x) = \exp(x) + seno(4\pi x), \qquad f_2(x) = \exp(x), \qquad \sigma = 0{,}05 \tag{49}$$

Em (46) e (47), x foi gerado aleatoriamente através de uma Uniforme no intervalo (-2,2) e em (48) e (49) através de uma Uniforme no intervalo (0,1) e o α foi fixado em 0,05. Na Tabela 3 temos os resultados das simulações onde percebemos que o poder do teste cresce conforme aumentamos o tamanho das amostras e em amostras pequenas, temos um poder razoavelmente bom. Os tempos para cada simulação variaram de 3 horas ($n_1 = n_2 = 10$) a dois dias ($n_1 = n_2 = 100$), dependendo também do conjunto de dados e modelo gerado.

(n_1, n_2)	Caso 5 (46)	Caso 6 (47)	Caso 7 (48)	Caso 8 (49)
(10,10)	0,061	0,048	0,152	0,224
(20,20)	0,760	0,215	0,800	0,845
(25,50)	0,976	0,631	0,972	0,981
(50,50)	0,997	0,896	0,998	0,999

Tabela 3: Avaliação do poder do teste T_N.

6.2. Comparação entre as estatísticas T_N e S_N

Em primeiro lugar, no caso do procedimento de Hall e Hart (1990), o parâmetro de suavização é o m como pode ser visto em (33). Para a escolha desse parâmetro, foi usada a função risco de Rice (1984), dada por

$$R(k)=\sum_{i=1}^{n}E\left[\frac{(\hat{\mu}_{ki}-\mu_i)^2}{n}\right]$$

onde $R(k)$ é uma aproximação discreta do erro quadrado médio integrado. Assim, considere $u_1,\ldots,u_n$ as observações do modelo $U_i=(\mu_i+\gamma_i),\ i=1,\ldots,n$, onde os $\mu_i's$ são constantes e os $\gamma_i's$ são variáveis aleatórias não correlacionadas com $E(\gamma_i)=0$ e $\text{var}(\gamma_i)=\sigma_\gamma^2,\ (i=1,\ldots,n)$. Logo defina

$$\hat{\mu}_{ki}=\begin{cases}\sum_{j=1}^{i+k}u_j/(i+k), & 1\le i\le k,\\ \sum_{j=i-k}^{i+k}u_j/(2k+1), & k+1\le i\le n-k,\\ \sum_{j=i-k}^{n}u_j/(n-i+k+1), & n-k+1\le i\le n,\end{cases} \tag{50}$$

e

$$R(k;u_1,\ldots,u_n)=\frac{1}{n}\sum_{i=1}^{n}(u_i-\hat{\mu}_{ki})^2+\hat{\sigma}_\gamma^2\left[\left(1-\frac{2k}{n}\right)(1+2k)^{-1}+\frac{2}{n}T_k\right],$$

onde $\hat{\sigma}_\gamma^2=\sum_{i=2}^{n-1}(u_{i+1}-2u_i+u_{i-1})^2/6n$ e $T_k=\sum_{i=1}^{k}(k+i)^{-1}$. A função $R(k;u_1,\ldots u_n)$ estima a função $R(k)$ e então o valor de $\hat{k}$ é o que minimiza a função risco definida acima.

Assim, considere $D_1,\ldots,D_n$ uma amostra de tamanho n. Então uma estimativa $\hat{m}$ é dada por $2\hat{k}+1$, onde $\hat{k}$ é valor que minimiza a função $R(k;D_1,\ldots,D_n)$ e que $\hat{m}$ é usado para construir a estatística S_n. Já no teste *bootstrap*, optou-se por usar uma nova estimativa de m para cada reamostra *bootstrap*, ou seja, para cada uma das 500 simulações, teremos um valor $\hat{k}^*$, onde o valor final $\hat{m}^*=2\hat{k}^*+1$, sendo esta usada na estimativa da estatística S_n^*.

Para avaliar o poder do teste estatístico, geramos 1000 amostras de tamanho 15 e 30 com as seguintes funções

$$f_1(x)=\exp(x),\qquad f_2(x)=\exp(x)+cx, \tag{51}$$

onde c é uma constante pré-definida (ver Tabela 4). Neste caso $x_i's$ são fixos e definimos $x_i=i/n,\ i=1,\ldots,n$. Foi utilizado a regressão polinomial local e o método de Ruppert para

estimarmos a função regressão e a janela ótima, respectivamente, para cada simulação e fixamos $\alpha = 0,05$.

Na Tabela 4, percebemos que S_n possui um poder maior quando temos uma pequena diferença entre as funções (c = 1). Porém o teste T_N é mais poderoso quando a diferença é maior (c = 5), para amostras de tamanho igual a 30.

	$T_N(\alpha = 0,05)$		$S_n(\alpha = 0,05)$	
	$n_1 = n_2 = 15$	$n_1 = n_2 = 30$	$n_1 = n_2 = 15$	$n_1 = n_2 = 30$
$c = 1$	0,086	0,224	0,184	0,236
$c = 2$	0,179	0,732	0,464	0,452
$c = 5$	0,771	1,000	0,968	0,926

Tabela 4: Análise do poder das estatísticas de teste T_N e S_n.

Agora, para avaliarmos o nível do teste, geramos amostras de tamanhos 10, 20 e 30 com funções

$$f_1(x) = \exp(x) = f_2(x) \tag{52}$$

sendo x definido igual ao caso (51) e os mesmos estimadores utilizados para avaliarmos o poder da estatística de teste. Foram avaliados os níveis para $\alpha = 0,05$ e para $\alpha = 0,10$, tanto para a estatística T_N quanto para S_n, como pode ser visto na Tabela 5. Notamos que o nível para T_N está muito próximo da estatística de teste S_n e que encontramos uma boa aproximação do nível para as duas estatísticas de teste.

	T_N		S_n	
	$\alpha = 0,05$	$\alpha = 0,10$	$\alpha = 0,05$	$\alpha = 0,10$
$n_1 = n_2 = 10$	0,048	0,086	0,048	0,095
$n_1 = n_2 = 20$	0,048	0,102	0,055	0,103
$n_1 = n_2 = 30$	0,063	0,110	0,053	0,098

Tabela 5: Análise do nível das estatísticas de teste T_N e S_n.

6.3. Aplicações

Nesta seção iremos trabalhar com dois bancos de dados. O primeiro deles contém observações fornecidas por Ian Rogers do Ministério da Agricultura Sul Australiano, onde temos duas regiões diferentes para a plantação de cebolas do tipo Imperial Espanhola Branca. As localidades são Purnong Landing e Virginia, e para cada um temos a variável X, densidade (plantas/m^2) e a variável Y, rendimento (g/planta). Mais detalhes sobre os dados ver em Ratkowsky (1983).

Na Figura 4 temos os pontos para as duas localidades e suas respectivas regressões, feitas através do estimador Polinomial Linear Local. Para a estimação da janela h para cada grupo, foi usado o método de Ruppert, Sheather e Wand (1995). Num total de 42 observações para cada localidade, antes tivemos que retirar a primeira observação da região de Virginia, pois era um dado discrepante. Sem a retirada dessa observação, não era possível obter as estimativas da regressão polinomial. O valor da estatística de teste T_N foi igual a 26090,06 e o quantil de 95% do *wild bootstrap*, realizado com 200 reamostras, ficou em 34,11. Logo, rejeitamos a hipótese nula e dizemos que $f_1 \neq f_2$, ou seja, há uma diferença entre as duas localidades e cada uma deve ser analisada separadamente. Neste caso, não foi possível fazer o uso da estatística S_n, pois os dados possuem amostras de tamanhos diferentes e a variável preditora assume valores diferentes para cada localidade.

Para a próxima análise, os dados foram retirados do artigo de Houghton (1999). Trata-se do fluxo líquido anual de carbono na atmosfera derivada das mudanças geradas pelo uso da terra de 1850 a 1990 (141 observações). Essas mudanças são provocadas pelo desmatamento de florestas para o uso em plantações e entre outros motivos. Esses dados estão divididos em nove grandes regiões do mundo. Podemos ver na Tabela 6 que somente a América do Sul e Central juntamente com o Sul e Sudeste da Ásia englobam 56,09% da quantidade de carbono no mundo inteiro, sendo que toda a região tropical do mundo engloba 63,8%. Isto também pode ser percebido na Figura 5.

Assim vamos testar se a quantidade de carbono na região da América do Sul e Central e a região do Sul e Sudeste da Ásia são iguais. Na Figura 6 temos o gráfico com somente as duas regiões.

Região	Fluxo Líquido Total 1850-1990	Percentual (%)
Sul e Sudeste da Ásia	38,6	31,33
América do Sul e Central	30,5	24,76
África Tropical	9,5	7,71
Subtotal Tropical	78,6	63,80
América do Norte	12,7	10,31
Europa	4,9	3,98
Antiga União Soviética	10,4	8,44
China	9,4	7,63
Região desenvolvida do Pacífico	4,1	3,33
Norte e Médio Oriente da África	3,1	2,51
Subtotal Não Tropical	44,6	36,20
Total Global	123,2	100,00

Tabela 6: Fluxo líquido total de carbono na atmosfera das alterações na terra para o cultivo.

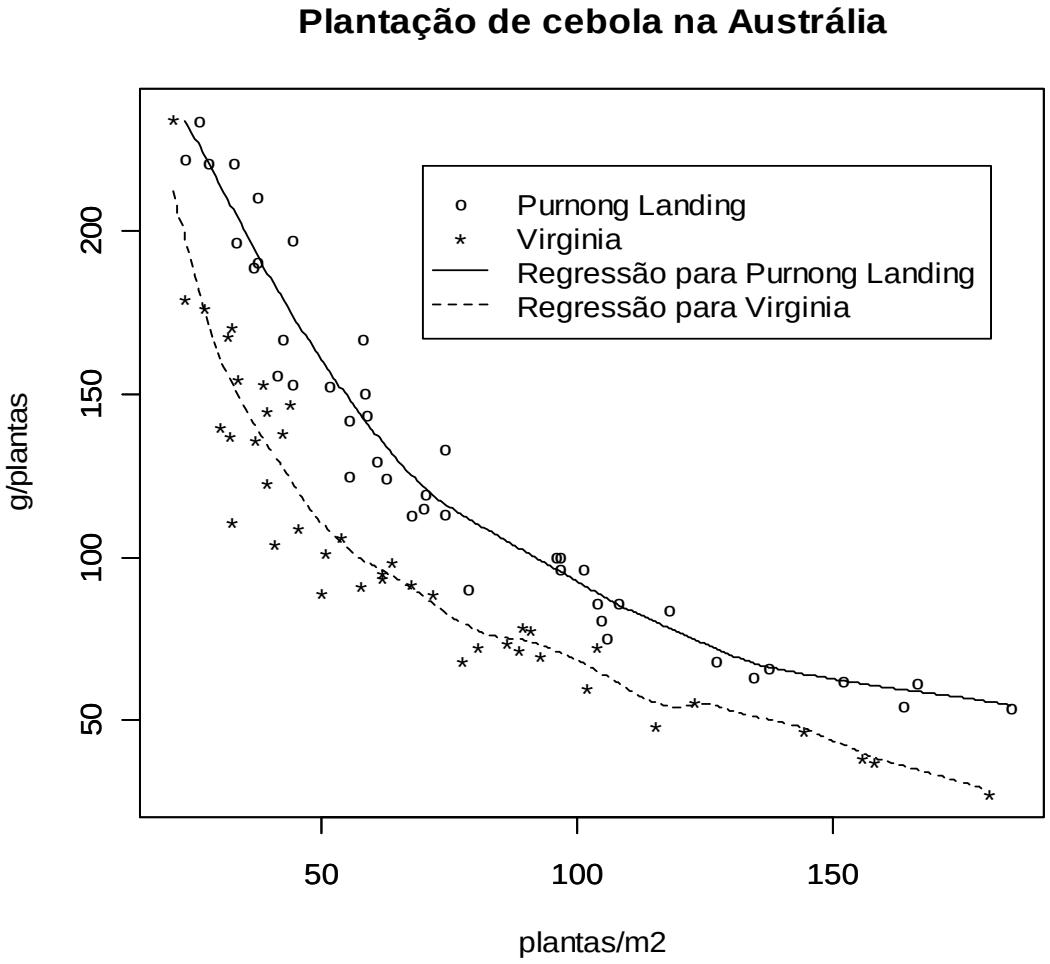

Figura 4: Regressões envolvendo rendimento de plantações de cebola para as duas localidades do sul da Austrália.

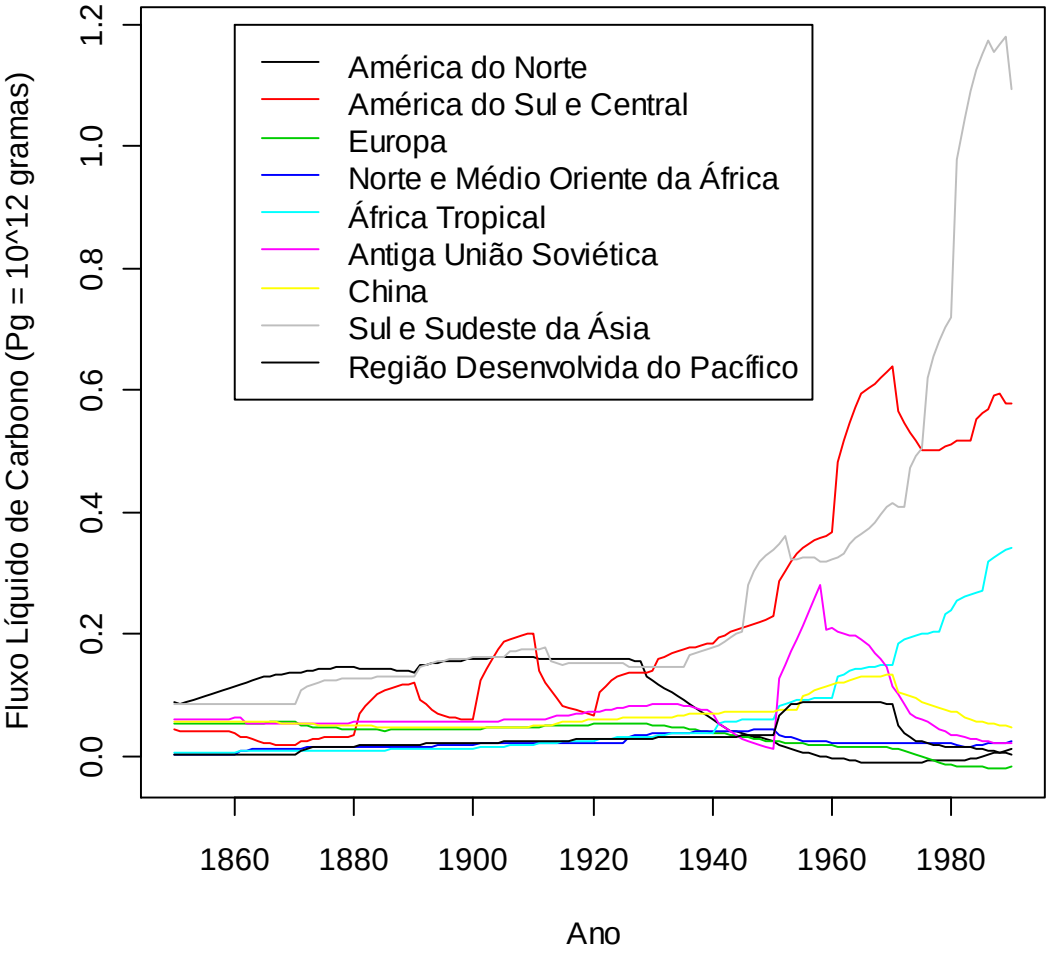

Figura 5: Fluxo líquido total de carbono na atmosfera, derivada das alterações na terra para o cultivo, em diferentes regiões do mundo.

Para os dados da América do Sul e Central, a janela foi igual a 1,375 e para Sul e Sudeste da Ásia igual a 1,858. Assim temos que o teste T_N, dada a amostra, ficou em 101315 e o ponto crítico dado pelo quantil de 95% das 200 reamostras *bootstrap* ficou em $T^*_{N[0,95]} = 0,423$. Logo rejeitamos a hipótese nula e dizemos que as duas regiões possuem comportamentos diferentes.

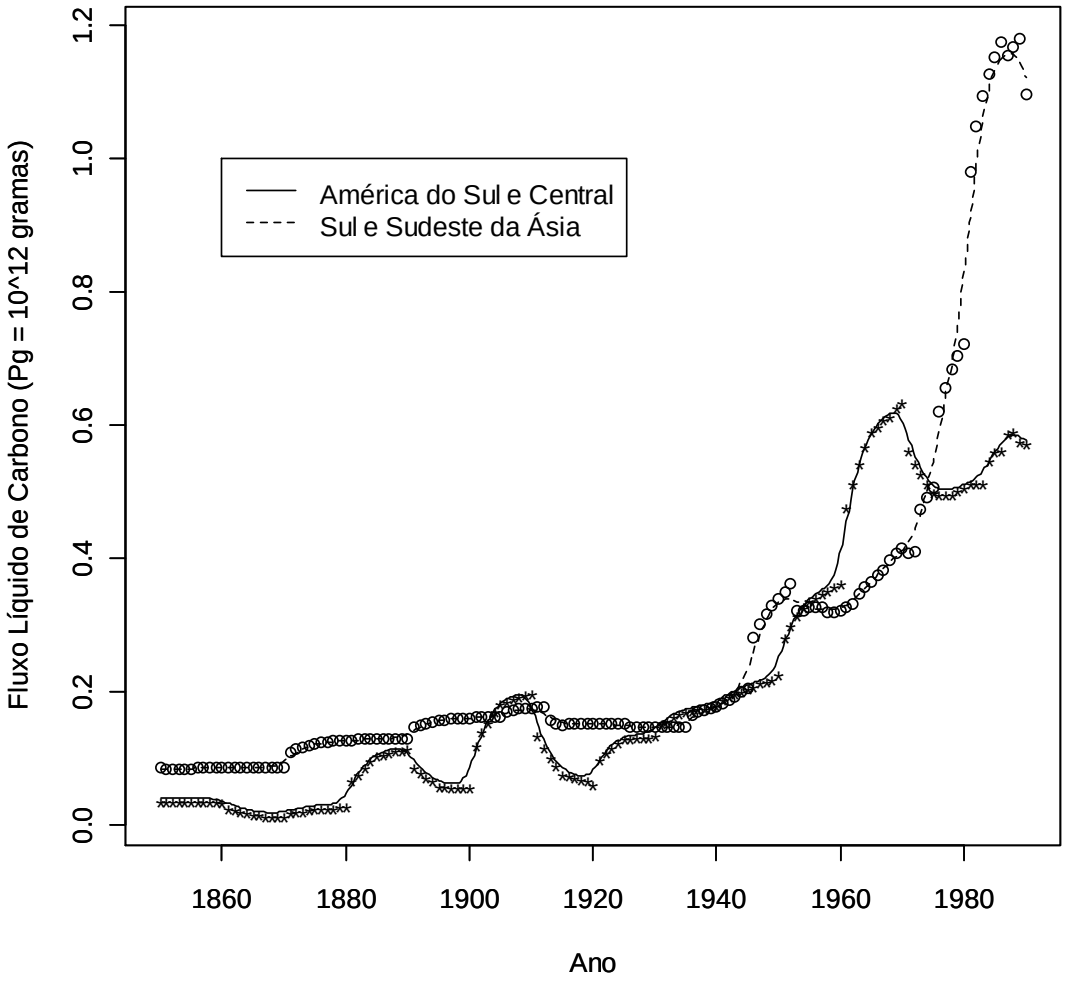

Figura 6: Fluxo líquido total de carbono na atmosfera, derivada das alterações na terra para o cultivo, nas regiões da América do Sul e Central e do Sul e Sudeste da Ásia.

Como há um aumento súbito, a partir de 1960, do fluxo líquido de carbono no Sul e Sudeste da Ásia, resolvemos refazer o teste de comparação entre as regiões Sul e Sudeste da Ásia com a América do Sul e Central, porém entre os anos de 1850 a 1960. Na Figura 7 temos a regressão para essas duas regiões. O valor da estatística de teste T_N igual a 466761,2 e $T^*_{N[0,95]} = 0,432$. Logo rejeitamos a hipótese nula e dizemos que há diferença entre as duas funções de regressão, ou seja, o fluxo líquido de carbono entre os anos de 1850 e 1960 para a América do Sul e Central é diferente da região Sul e Sudeste da Ásia.

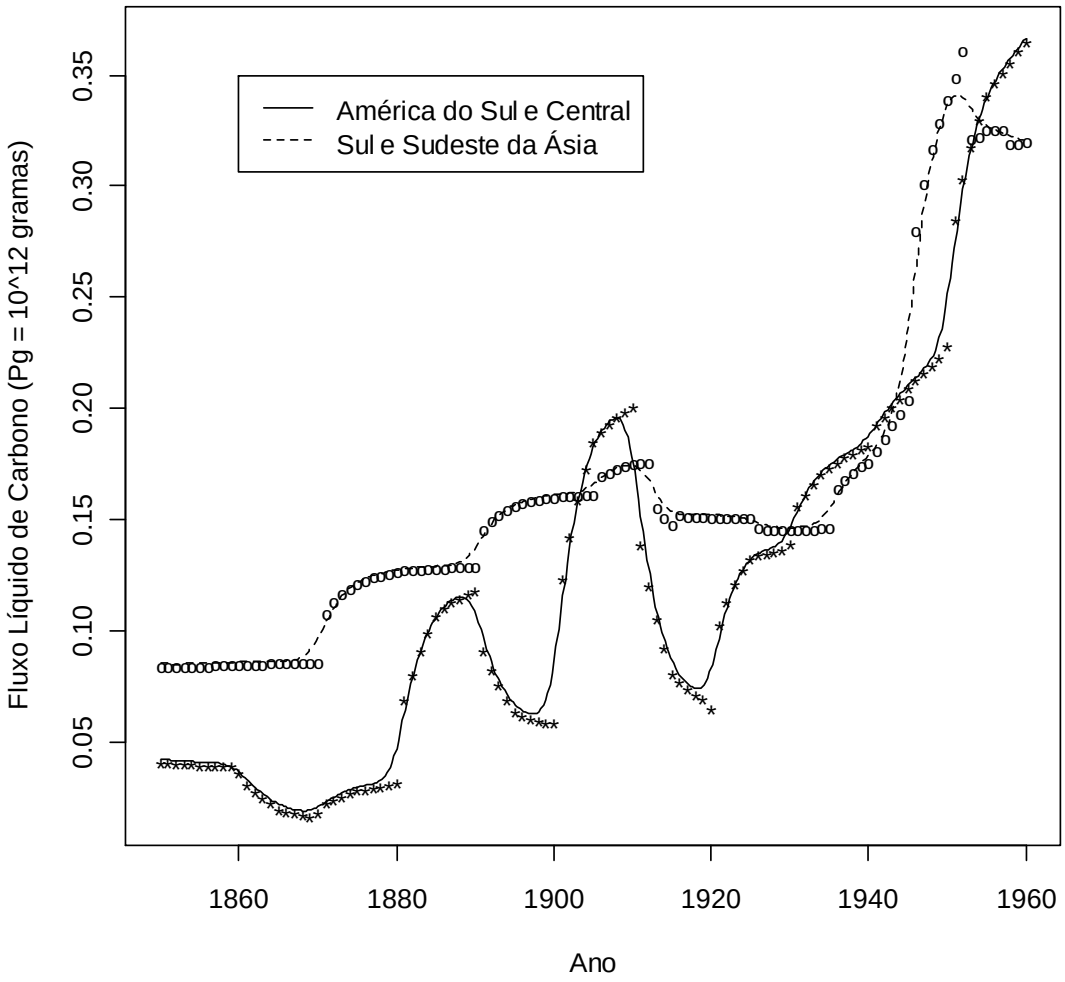

Figura 7: Fluxo líquido total de carbono na atmosfera, derivada das alterações na terra para o cultivo, nas regiões da América do Sul e Central e do Sul e Sudeste da Ásia até 1960.

7. Conclusões e Trabalhos Futuros

Neste trabalho avaliamos uma proposta para análise de covariância envolvendo métodos não-paramétricos. Vimos que a sua generalização é natural, vinda do procedimento paramétrico. A convergência assintótica da estatística envolvida para uma distribuição normal é lenta, porém as constantes C e τ^2, da distribuição assintótica, dependem somente da função núcleo K. Vimos também que o método "*wild bootstrap*" melhora o desempenho do teste estatístico, corrigindo esse problema.

Em Munk, Neumeyer e Scholz (2006), é mostrada a influência da janela no nível e poder da estatística de teste T_N. Para o nível há pouca alteração, porém para o poder, percebe-se uma redução dos valores, quando aumentamos o valor da janela. Logo concluímos que a escolha da janela é de fundamental importância para se obter uma boa decisão no teste e que tanto o método 1 (Ruppert, Sheather e Wand, 1995) quanto o método 2 (Fan e Gijbels, 1995) são muito bons.

Contudo salientamos que o método 1 é mais rápido que o método 2, pois não envolve a minimização de uma função. Mas, apesar de não termos simulado para erros não homogêneos, sabemos que o método 1 não garante sua boa aplicação nestes casos. Ruppert, Sheather e Wand (1995) propõem a substituição do termo $\sigma^2(b-a)$ por $\int_a^b v(x)dx$, onde $v(x)$ seria estimado por $\hat{v}(\cdot;\lambda)$, dada a janela λ. Neste caso usaríamos o núcleo-estimador para encontrar $\hat{v}(\cdot;\lambda)$ e a mesma regra para encontrar a estimativa de λ_{AMSE}. Outro ponto a ser levado em consideração é que o método 1 já possui a sua implementação no programa R®, versão gratuita do S-plus®.

Na parte das simulações, comparamos a estatística T_N com o teste de Hall e Hart (1990), e observamos que o seu desempenho é melhor quando temos amostras de tamanhos maiores e também maiores diferenças entre as funções. Ainda verificamos que o nível do teste ficou próximo do valor fixado, o que também foi observado para o teste S_n. Mas salientamos que o teste *boostrap* S_n limita-se ao seu uso caso tenhamos amostras de mesmo tamanho e da seguinte forma $\{(x_i, Y_i, Z_i), 1 \le i \le n\}$. Contudo no procedimento de Munk, Neumeyer e Scholz (2006), não há essa limitação na forma do conjunto amostral e também pode ser aplicado quando temos amostras de tamanhos diferentes.

Percebemos ainda que o nível da estatística de teste T_N é bom, mesmo quando temos amostras de tamanhos diferentes, o que não é possível de se obter com a estatística S_N. Já em relação ao poder do teste, obtivemos valores muito bons, sendo que já com amostras de tamanho $n = 50$, o poder já estava próximo de um em todos os modelos avaliados.

Para projetos no futuro, mais estudos envolvendo simulações para o poder e o nível do teste aqui apresentado, como para amostras de tamanhos diferentes e erros não homogêneos. Avaliaremos situações diferentes para as funções de regressão, onde os erros poderão ser correlacionados. Aplicar novos métodos para a estimação da função variância, pois com isso podemos ter melhores resultados para a estatística de teste em modelos com alta heterocedasticidade.

Referências

Atuncar, G. S. (2009). Estimadores da Variância do Núcleo Estimador de uma Função de Regressão (em andamento).

Beran, R. (1986). Comentários em "Jackkinfe, bootstrap and other resampling methods in regression analysis" por C. F. J. Wu. *Ann. Stat.*, **14**, 1295-1298.

Bowman, A. W. e Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis*. Oxford Statistical Science Series, **18**.

Dette, H. e Neumeyer, N. (2001). Nonparametric analysis of covariance. *Ann. Stat.*, **20**, 2071-2086.

Fan, J. e Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Roy. Stat. Soc.*, *Ser. B*. **57**, 371-394.

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

Fan, J., Zhang, C. e Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Stat.,* **29**, 153-193.

Gasse, T. e Müller, H.-G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. Stat.*, **11**, 171-185.

Gasser, T., Müller, H.-G. e Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. *J. Roy. Stat. Ser. B*, **47**, 238-252.

Hall, P. e Hart, J. W. (1990). Bootstrap test for difference between means in nonparametric regression. *J. Amer. Stat. Assoc.*, **85**, 1039-1049.

Härdle, W. e Mammen, E. (1990). Comparing nonparametric versus parametric regression fits. Pré-impresso SFB 123, Univ. Heidelberg.

Härdle, W. e Marron, J. S. (1993). Fast and simple scatterplot smoothing. CORE discussion paper No. 9143, Univ. Catholique de Louvain.

Härdle, W. e Tsybakov, A. (1997). Local polynomial estimators of the volatility function in nonparametric auto regression. *J. Econometrics*, **81**, 223-242.

Houghton, R. A. (1999). The annual net flux of carbon to the atmosphere from changes in land use 1850-1990. *Tellus*, **51B**, 298-313.

Liu, R. Y. (1988). Bootstrap procedures under some non i.i.d. models. *Ann. Stat.*, **16**, 1696-1708.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, **15**, 661-675.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Ann. Stat.*, **21**, 255-285.

Munk A., Neumeyer, N. e Scholz, A. (2006). Nonparametric analysis of covariance – The case of inhomogeneous and heteroscedastic Noise. *Scand. J. Stat.*, **34**, 511-534.

Nadaraya, E. A. (1964). On estimating regression. *Theory Probab. Appl.*, **10**, 186-190.

Priestley, M. B. e Chao, M. T. (1972). Nonparametric function fitting. *J. Roy. Stat. Soc.*, *Ser. B*, **34**, 385-392.

Ratkowsky, D. A. (1983). *Nonlinear regression modeling*. Dekker, New York.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Stat.*, **12**, 1215-1230.

Ruppert, D. e Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Stat.*, **22**, 1346-1370.

Ruppert, D., Wand, M. P., Holst, U. e Hössjer, O. (1997). Local polynomial variance-function estimation. *Technometrics*, **39**, 262-273.

Ruppert, D., Sheather, S. J. e Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Stat. Assoc.*, **90**, 1257-1270.

Sacks, J. e Ylvisaker, D. (1970). Designs for regression problems for correlated errors. *Ann. Math. Sat.*, **41**, 2057-2074.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. Springer Series in Statistics.

Watson G. S. (1964). Smooth regression analysis. *Sankhyā*, Ser. A, **26**, 359-372.

Wu, C. F. J. (1986). Jackkinfe, bootstrap and other resampling methods in regression analysis. *Ann. Stat.*, **14**, 1261-1295.