UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Junio 2014, Vol. 7, No. 1, pp. 49–66

Estudio de simulación para comparar varios estimadores de varianza en el marco de la regresión no paramétrica

A simulation study for the comparison of several variance estimators in the nonparametric regression framework

Alvaro José Flórez[a]
alvaro.florez@correounivalle.edu.co

Javier Olaya[b]
avier.olaya@correounivalle.edu.co

Resumen

En el presente trabajo se prueban varios estimadores de varianza basados en diferencias, en el marco de la regresión no paramétrica. Dichos estimadores tienen la principal ventaja de no depender de los parámetros de suavización, además de que son poco exigentes en términos computacionales. Se usan principalmente estimadores basados en diferencias ordinarias y basados en las diferencias óptimas de Hall. Se crean escenarios utilizando diferentes funciones de regresión, tamaños de muestra y distribuciones de los errores y se introduce el uso de la distribución semi-normal para probar los estimadores de varianza, en casos de distribuciones asimétricas de los errores. Los resultados parecen apoyar la idea de que los estimadores basados en diferencias óptimas de Hall no son mejores en todos los escenarios planteados.

***Palabras clave*:** estimadores basados en diferencias, diferencias ordinarias, diferencias óptimas, distribución semi-normal.

Abstract

We test several difference-based variance estimators in the nonparametric regression model. These estimators have the main advantage of not depending on the smoothing parameters. Furthermore, they also show low computational demand. We mainly use estimators based on ordinary differences, along with estimators based on Hall's optimal differences. We set scenarios using some regression functions, some sample sizes, and some error distributions. In particular we bring in the use of the half-normal distribution to test the variance estimators under some

[a]Profesor auxiliar. Escuela de Estadística, Universidad del Valle, Colombia.
[b]Profesor titular. Escuela de Estadística, Universidad del Valle, Colombia.

asymmetric error distributions. Results seem to support the idea that the Hall's optimal differences estimators not perform better than the others on all sets of scenarios.

***Keywords*:** Difference-based estimators, ordinary differences, optimal differences, half-normal distribution.

1. Introducción

La estimación de una función f poblacional por medio de modelos de regresión ha sido ampliamente estudiada durante mucho tiempo y presenta una gran variedad de herramientas estadísticas, de las cuales la modelación paramétrica es la que más ha sido desarrollada y entendida (Draper & Smith 1966, Draper & Smith 1998). Sin embargo, hay muchos casos donde estos tipos de modelos no son recomendables, ya sea por el incumplimiento de uno o más de los supuestos, o por la falta de información que se tenga sobre la relación funcional de los datos. Lo anterior hace que la utilización de métodos de regresión no paramétrica sean una buena opción para la estimación de la función f, puesto que estos métodos son menos exigentes, especialmente en los supuestos, que su contraparte paramétrica (Eubank 1998, Altman 1992, Cleveland 1979).

Dentro del estudio de la regresión no paramétrica se han presentado grandes avances en las últimas décadas, debido principalmente a los enormes progresos tecnológicos que han cubierto la gran demanda computacional que dichos métodos exigen, donde se han propuesto una variedad de herramientas y técnicas para el modelamiento de f, así como también un número considerable de estimadores de varianza. Puesto que este parámetro no puede ser estimado de la misma forma como se hace en la regresión paramétrica, debido a que las técnicas de suavización producen estimaciones sesgadas de las respuestas, pues lo anterior llevaría a una sobreestimación de la varianza (Hall et al. 1990, Hall & Marron 1990, Gasser et al. 1986, Dette et al. 1998, Seifert & Gasser 1993, Buckley et al. 1988).

Dada la importancia de la estimación la varianza de los errores, este trabajo busca documentar algunos de los estimadores de varianza que se han desarrollado, y que se usan con más frecuencia, para los modelos de regresión no paramétrica. También se pretende mostrar el comportamiento que presentan los estimadores estudiados bajo situaciones diferentes, y así poder identificar en qué casos es más conveniente el uso un estimador sobre los demás. Así, los escenarios donde se ponen a prueba los estimadores resultan de combinar distintas funciones de regresión con diferentes distribuciones de los errores y diferentes tamaños de muestra. De otro lado, los autores que proponen los estimadores han conducido sus propias simulaciones para comparar los que están proponiendo con los demás. Sin embargo, persisten diferencias de opinión sobre cuáles son mejores y en qué casos. Este estudio se propone como meta conducir un estudio de simulación en el cual los investigadores (Gasser et al. 1986, Hall et al. 1990, Carter & Eagleson 1992, Brown & Levine 2007) que han propuesto los estimadores que se comparan no intervienen

en la construcción de los escenarios de simulación, ni en la formulación de las conclusiones. Se trata entonces de un estudio independiente que busca nuevas luces sobre el uso de los estimadores bajo diferentes escenarios.

2. Antecedentes

Siempre que se hace una propuesta para un estimador de varianza, en el modelo de regresión no paramétrico, es natural pensar que es necesario ponerlo a prueba de alguna forma para que se puedan ver sus ventajas y desventajas frente a los otros estimadores que han sido previamente desarrollados. Este tipo de comparaciones se conducen generalmente estudiando las propiedades teóricas de los estimadores y evaluándolas por simulación. A continuación se presentan algunos artículos donde se ha hecho algún tipo de comparación, ya sea teórica o práctica, de los estimadores que se utilizaron en este estudio: el estimador de Rice (Rice 1984) y los estimadores basados en diferencias ordinarias y en diferencias óptimas de Hall, Kay y Titterington (Hall et al. 1990).

La primera comparación de los estimadores basados en diferencias fue hecha en la presentación del estimador de Gasser, Sroka y Jennen-Steinmetz (estimador GSJS) (Gasser et al. 1986), quienes compararon el estimador de Rice con un estimador propuesto por Wahba (1978) y con el estimador GSJS. Dicha comparación fue hecha por medio de simulaciones, teniendo en cuenta cambios en la función poblacional, el tamaño de muestra y la varianza de los errores; allí se encontró que el sesgo en todos los casos es siempre positivo y es proporcionalmente más grande para tamaños de muestra y varianza pequeños. Además, de acuerdo con sus autores, el sesgo del estimador GSJS es mucho más pequeño que el de los otros dos estimadores.

Hall et al. (1990) presentan el estimador de varianza basado en diferencias en forma general y además se hace referencia a tres métodos de asignación para las diferencias, llamados asignación ordinaria, *spike* y óptima de Hall, siendo estas dos últimas propuestas por Hall et al. (1990). A fin de hacer las comparaciones Hall propone un Error Cuadrático Medio (ECM) asintótico para cada uno de estos estimadores, el cual solo depende del factor de la varianza, mientras que el componente del sesgo se considera insignificante. En consecuencia el ECM asintótico es independiente de la función de regresión f.

Luego de encontrar el ECM de cada uno, se procedió a calcular la eficiencia teórica de estos estimadores, de orden 2 al 5. Se encontró que los estimadores basados en diferencias óptimas de Hall y los basados en diferencias *spike* presentaban incrementos en la eficiencia al aumentar el orden de los estimadores. Ocurrió lo contrario con el estimador basado en diferencias ordinarias, siendo el primero de estos estimadores el que presentaba la mayor eficiencia en todos los casos (Hall et al. 1990, p. 525).

Dette et al. (1998) redefinen el ECM de estos estimadores y muestran que el ECM depende no solamente del componente de la varianza, sino también de la compo-

nente del sesgo. En esta investigación se hizo una comparación bajo simulaciones de los ECM teóricos de los estimadores basados en diferencias ordinarias y óptimas de Hall bajo funciones de regresión distintas (Dette et al. 1998, p.759-763). De acuerdo con Dette et al. (1998), en pocos casos los estimadores basados en diferencias óptimas de Hall presentaban ECM inferiores a los de los estimadores basados en diferencias ordinarias, conclusiones que contradicen las formuladas en Hall et al. (1990).

3. Modelo de regresión no paramétrico

Un modelo de regresión, sea paramétrico o no paramétrico, pretende estimar una función poblacional tomando información de n pares de observaciones de una variable Y y de una variable X (en nuestro caso ambas continuas), entre las cuales se presume la existencia de cierta relación, tal como se expresa en la ecuación (1).

$$y_i = f(x_i) + \varepsilon_i, \qquad i = 1, 2, \cdots, n \tag{1}$$

Donde Y se conoce como variable respuesta y X como variable predictora, explicativa o covariable. Los pares (x_i, y_i) son un conjunto de n observaciones de X y Y. Al conjunto de valores de X se le conoce habitualmente como puntos del diseño. f es la función de regresión o curva de regresión y los ε_i son los llamados errores, que son variables aleatorias no observables que se asumen independientes y que satisfacen que $E(\varepsilon_i) = 0$ y $V(\varepsilon_i) = \sigma^2 < \infty$. Este artículo se refiere a la estimación de σ^2, en el caso en que la función de regresión f se estime por métodos de suavización.

La principal diferencia que existe entre la regresión paramétrica y la no paramétrica, radica en que en la regresión paramétrica, el investigador debe suponer la forma de la función de regresión y solamente desconoce los valores de los parámetros que componen la función. Mientras que, en el ámbito no paramétrico, no se supone *a priori*, un comportamiento de la función de regresión f poblacional. En cambio, la forma de la función estimada se crea a partir del comportamiento de los mismos datos. Por lo tanto, la regresión no paramétrica se considera como una colección de técnicas para ajustar curvas donde se tiene poco conocimiento *a priori* de su forma de f.

Dentro de la teoría de la regresión no paramétrica, se debe asumir que f es suave, lo que quiere decir, que si se desea estimar la función f en un punto x, se espera que las observaciones y_i asociadas a los x_i cercanos a x, posean información de f en x. Lo cual indica que es posible promediar de alguna forma las respuestas y_i más cercanas al punto donde se estime $f(x)$. En el marco de la regresión no paramétrica esto es presentado por Eubank (1998) como suavización.

Formalmente, se asume que f es una función cuadrado integrable que tiene dos derivadas continuas. Si se denota W_2^2 al espacio de todas las funciones que satisfacen estas condiciones, se dice que f es suave si pertenece a W_2^2.

Para la estimación de f se encuentran muchos métodos de regresión no paramétrica en la literatura, donde los suavizadores usados comúnmente son los estimadores lineales para regresiones simples, es decir con una sola covariable. Los estimadores del tipo lineal de f tienen la forma dada por la ecuación (2).

$$\hat{f}(x_i) = \sum_{i=1}^{n} K(x, x_i; \lambda) y_i \qquad i = 1, 2, \cdots, n \tag{2}$$

Donde $K(x, x_i; \lambda)$ es una colección de pesos que dependen de los puntos del diseño x_i y de un $\lambda > 0$, el cual es denominado parámetro de suavización o ancho de banda, y determina el grado de suavización a los datos, el cual es definido por el usuario (Eubank 1998, Levine 2006, Olaya 2012). Se consideran lineales porque para un λ dado, los estimadores resultan ser funciones lineales de las respuestas y_i. Dentro de los métodos de suavización en modelos de regresión con una sola variable de predicción se encuentran: la suavización *kernel*, la regresión LOESS y la suavización por *splines*.

4. Estimación de la varianza en un modelo de regresión no paramétrico

En un modelo lineal la suma de cuadrados de los errores brinda las bases para la estimación de la varianza de los errores, por lo cual en un enfoque no paramétrico se puede pensar que la estimación se podría hacer de forma análoga. No obstante, realizar la estimación de esta forma no es válido debido a la presencia del sesgo de $\hat{f}$ (Bowman & Azzalini 1997), el cual tendrá el efecto de aumentar el valor de la suma de cuadrados de los errores y por lo tanto sobreestimar el parámetro de varianza.

Por esta razón, dentro del contexto de la regresión no paramétrica existe un número considerable de estimadores de σ^2, los cuales pueden ser considerados por separado en dos grupos. En el primer grupo se encuentran los estimadores que dependen del parámetro de suavización, los cuales realizan la estimación de la varianza basándose en la suma de cuadrados de los errores de un ajuste no paramétrico de f, por medio de un método de suavización como Kernel o Splines. Algunos de estos estimadores son el estimador de Hall & Marron (1990), que está basado en suavización *Kernel*, y los estimadores de Wahba (1978) y de Buckley et al. (1988) que están basados en suavización Spline.

El segundo grupo está conformado por los estimadores basados en diferencias, los cuales se apoyan en las respuestas y_i asociadas a una vecindad predeterminada de x, estos estimadores tienen la ventaja de no depender explícitamente del parámetro de suavización. En este tipo de estimadores se asume el modelo de regresión de la ecuación 1, donde f es una función desconocida y los errores ε_i se asumen independientes e idénticamente distribuidos con media 0 y varianza σ^2. Además, el diseño se encuentra ordenado de la siguiente forma $0 \leq x_1 \leq x_2 \leq \cdots \leq x_n \leq 1$.

Este tipo de estimadores no requieren ningún parámetro de suavización. El orden de los estimadores de diferencias viene dado por el número de observaciones que se relacionan para calcular el residual local.

El más simple de estos estimadores fue propuesto por Rice (3) en 1984. Dicho estimador puede presentar algunos problemas debido a que la diferencia $(y_i?y_{i-1})$ puede ser influenciada por las fluctuaciones bruscas que puede presentar la función de regresión f, y por lo tanto la estimación de la varianza puede inflarse.

$$\hat{\sigma}_R^2 = \frac{1}{2(n-1)} \sum_{i=2}^{n} (y_i - y_{i-1})^2 \tag{3}$$

Gasser et al. (1986) proponen el estimador GSJS, basado en interpolación lineal, el cual contrarrestaría el problema del estimador de Rice. Con este propósito los autores proponen unos seudo-residuales, los cuales se obtienen tomando una tripleta consecutiva de puntos de diseño x_{i-1}, x_i, x_{i+1}, a fin de calcular la diferencia que hay entre la línea recta que une las observaciones límites (x_{i-1}, y_{i-1}) y (x_{i+1}, y_{i+1}) y la observación central (x_i, y_i), de la siguiente manera:

$$\begin{aligned} \tilde{\varepsilon}_i &= \frac{x_{i+1} - x_i}{x_{i+1} - x_{i-1}} y_{i-1} + \frac{x_i - x_{i-1}}{x_{i+1} - x_{i-1}} y_{i+1} - y_i \\ &= a_i y_{i-1} + b_i y_{i+1} - y_i \end{aligned} \tag{4}$$

El estimador GSJS está definido de la siguiente forma:

$$\sigma_{GSJ}^2 = \frac{1}{n-2} \sum_{i=3}^{n} c_i^2 \tilde{\varepsilon}_i^2 \quad \text{donde } c_i^2 = (a_i^2 + b_i^2 + 1)^{-1} \tag{5}$$

Hall et al. (1990) introdujeron los estimadores basados en diferencias en forma general. Una diferencia se define como una sucesión de números que cumplen con las siguientes condiciones:

$$\sum d_i = 0, \ \sum d_j^2 = 1, \qquad \text{donde } d_j \neq 0 \tag{6}$$

Se asume que $d_j = 0$ para $j < -m_1$ y $j > m_2$, donde los valores $m_1, m_2 \geq 0$ y $d_{-m_1} d_{m_2} \neq 0$. El orden de la sucesión viene dado por $m = m_1 + m_2$. Por conveniencia en los cálculos se toma $m_1 = 0$ y $m_2 = m$. Entonces el estimador de σ^2 basado en estas diferencias tiene la forma dada por la ecuación (7):

$$\sigma_{HKT}^2 = \frac{1}{n-m} \sum_{k=m_1+1}^{n-m_2} \left(\sum_{j=0}^{m_2} d_j y_{j+k} \right)^2 \tag{7}$$

Para la diferencia de primer orden, solamente se tiene un resultado válido para (d_0, d_1) el cual es $d_0 = \frac{1}{\sqrt{2}}$ y $d_1 = -d_0$, que se define como la primera diferencia $\triangle Y = \frac{y_i - y_{i-1}}{\sqrt{2}}$, cuyo estimador coincide con el estimador de Rice (3).

Cuando se tienen diferencias de órdenes superiores, se obtiene más de una solución para cada orden, por lo cual se tendrán infinitos estimadores de varianza de la forma (7) por cada orden. Por lo cual determinar el orden m del estimador, así como la escogencia de las diferencias es de gran importancia. Una forma de realizar dicha asignación, es por medio de una diferencia ordinaria que se usa comúnmente:

$$d_j = \begin{cases} \binom{2m}{m}^{-1/2}\binom{m}{j}(-1)^j & \text{para } 0 \leq j \leq m, \\ 0 & \text{en otro caso} \end{cases} \tag{8}$$

El estimador de la ecuación (7) con la asignación de la ecuación (8) se conoce como estimador de diferencias ordinarias. Cuando se obtiene dicho estimador con una diferencia de segundo orden, este coincide con el estimador GSJS (ver ecuación (5)), cuando se tiene un diseño equidistante, los valores de x_i se encuentran igualmente espaciados.

Hall et al. (1990) proponen una asignación distinta, a la cual denominan diferencias óptimas de Hall, la cual está basada en una definición que se propone del ECM de este estimador y la varianza asintótica (descripción formal del teorema en Hall et al. (1990, apéndice 1)), los cuales son ambos iguales a $n^{-1}\tau^2$, donde τ^2 se define en la ecuación (9), en la que kx— denota la kurtosis de ε/σ.

$$\tau^2 = var(\varepsilon^2) + 2\sigma^4 \sum_{k\neq 0}\left(\sum_j d_j d_{j+k}\right)^2 = \sigma^4\left(k + 2\sum_k\left(\sum_j d_j d_{j+k}\right)^2\right) \tag{9}$$

Teniendo en cuenta la definición del ECM del estimador de Hall, se observa que este valor solamente depende de los valores de d_j, además de la distribución de los errores. En esta definición se asume que la función f tiene un efecto insignificante sobre el error cuadrático medio, ya que la función f se considera suave y los puntos de diseño x_i adyacentes se encuentran cada vez más juntos, a medida que el tamaño de la muestra aumenta.

Se sigue que la asignación óptima de los d_j se obtiene minimizando la siguiente expresión:

$$\delta = \sum_{k\neq 0}\left(\sum_j d_j d_{j+k}\right)^2 \tag{10}$$

Para el m-ésimo orden la diferenciación sucesiva óptima y con δ, se tiene que $\delta = (2m)^{-1}$, por lo tanto:

$$\sum_{j=1}^{m} d_j d_{j+k} = -(2m)^{-1} \qquad (1 \leq |k| \leq m) \tag{11}$$

Por lo cual la varianza asintótica mínima que se puede obtener utilizando una diferencia sucesiva de m-ésimo orden es de $n^{-1}\tau^2$, donde:

$$\tau^2 = var(\varepsilon^2) + m^{-1}\sigma^4 \tag{12}$$

En Hall et al. (1990, apéndice 3), se plantea el siguiente cálculo para encontrar las diferencias óptimas:

Para un m, se observa que:

$$D(d_0, \cdots, d_m) = \frac{1}{2} \sum_{k=1}^{m} \left(\sum_{j=0}^{k} d_j d_{j+k} \right)^2 \tag{13}$$

$$D(d_0, \cdots, d_m) = (d_0 d_m)^2 + (d_0 d_{m-1} + d_1 d_m) + \cdots + (d_0 d_1 + \cdots + d_{m-1} d_m)^2 \tag{14}$$

Además se asume que: $s_1 = -(d_0 + d_m)$, $s_2^2 = 1 - (d_0^2 + d_m^2)$, $t_1 = (\frac{1}{2} - \frac{1}{4} s_1^2 - \frac{1}{2} s_2^2)^{\frac{1}{2}}$. Por lo tanto, $d_0 = -\frac{1}{2} s_1 + t_1$, $d_m = -\frac{1}{2} s_1 - t_1$. Usando estas fórmulas para d_0 y d_m, además de tomar $s_1 = d_1 + \cdots + d_{m-1}$ y $s_2^2 = d_1^2 + \cdots + d_{m-1}^2$ y sustituyendo d_0 y d_m en la ecuación (14), se obtiene una función que involucra solamente los valores $d_1, \cdots, d_m$. A estas expresiones se les incorpora las restricciones de las diferencias (ver ecuación (6)) y se puede obtener los valores que minimizan la ecuación (13) por medio de un método de optimización.

En la Tabla 1 se pueden observar las diferencias óptimas para estimadores de orden $1 \leq m \leq 5$:

Tabla 1: *Diferencias óptimas de Hall para estimadores de orden* $1 \leq m \leq 5$. *Fuente: Hall et al. 1990.*

m	$(d_0, \cdots, d_m)$
1	$(0.7071, -0.7071)$
2	$(0.8090, -0.5, -0.309)$
3	$(0.1942, 0.2809, 0.3832, -0.8582)$
4	$(0.2708, -0.0142, -0.6909, -0.4858, -0.4617)$
5	$(0.9064, -0.26, -0.2167, -0.1774, -0.142, -0.1103)$

En la Tabla 1 se observa que a medida que el orden aumenta uno de los valores d_j tiende a acercarse a la unidad mientras que los otros convergen a 0; también se observa que este pico se encuentra en el medio de la diferenciación cuando el orden es par y en un extremo cuando el orden es impar.

Teniendo en cuenta la observación anterior, Hall, Kay y Titterington realizaron una asignación forzando al d_j central de la sucesión a asumir valores cercanos a la unidad, mientras que a los otros los acercan a 0. Esta asignación fue llamada *spike* (pico). La asignación de este tipo se hace de la siguiente forma:

Si el orden es par, $v = \frac{m}{2}$

$$d_j = \begin{cases} \left(\dfrac{2v}{2v+1} \right)^{1/2} & \text{para } j = v \\ -\left[2v(2v+1) \right]^{-1/2} & \text{para } 0 \leq j \leq v-1 \text{ ó } v+1 \leq j \leq 2v \\ 0 & \text{en otros casos} \end{cases} \tag{15}$$

Si el orden es impar, $v = \frac{m-1}{2}$

$$d_j = \begin{cases} \left(\dfrac{2v+1}{2v}\right)^{1/2} & \text{para } j = v \\ -\left[2v(2v-1)\right]^{-1/2} & \text{para } 0 \leq j \leq v-1 \text{ ó } v+1 \leq j \leq 2v \\ 0 & \text{en otros casos} \end{cases} \tag{16}$$

5. Metodología

Para el proceso de simulación se planteó el siguiente modelo de regresión:

$$y_i = f(x_i) + \varepsilon_i, \qquad i = 1, 2, \cdots, n \tag{17}$$

Donde los valores y_i representan las respuestas, f la función de regresión poblacional, x_i la covariable y ε_i los errores aleatorios. Además se deben cumplir en todas las simulaciones las siguientes condiciones:

- El diseño es equidistante, y los valores x_i se encuentran ordenados en el intervalo $[0, 1]$, además no se tienen medidas repetidas en ningún valor de x_i.
- Los valores ε_i son independientes e idénticamente distribuidos con $E(\varepsilon_i) = 0$ y $var(\varepsilon_i) = \sigma^2$.
- La función f es continua y doblemente diferenciable.

El proceso de simulación se realizó en distintos escenarios, los cuales presentan diferencias en la función de regresión, distribución de los errores y tamaños de muestra. A fin de obtener observaciones suficientes para realizar las comparaciones, se consideraron 1000 repeticiones para cada simulación.

Los diferentes cambios en cada uno de estos factores son los siguientes:

Función poblacional f:

- $8\sin(0.5\pi x_i) - 4$ sin oscilaciones.
- $4\sin(3\pi x_i)$ número de oscilaciones bajo.
- $4\sin(7\pi x_i)$ número de oscilaciones alto.

Varianza de los errores: $\sigma^2 = 0.5$ (variación baja), $\sigma^2 = 1$(variación alta).

Tamaño de muestra: $n = 50, 100, 300$.

Distribución de ε_i:

- $N(0, \sigma^2)$, distribución simétrica.

- $|N(0,1)| - \left(\frac{2}{\pi}\right)^{1/2}$, distribución asimétrica a la derecha
- $\left(\frac{2}{\pi}\right)^{1/2} - |N(0,1)|$, distribución asimétrica a la izquierda

Las distribuciones asimétricas se definen a partir de una variable que se distribuye semi-normal, definida como el valor absoluto de una variable que se distribuye normal estándar (Olmos et al. 2012). En ambos casos el procedimiento empleado centra las distribuciones semi-normales en 0, pero mantiene una de ellas asimétrica a la derecha y la otra a la izquierda (ver Figura 1).

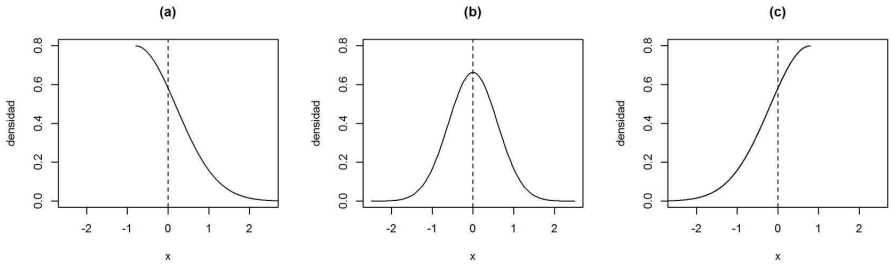

Figura 1: *Distribución de los errores (a) distribución asimétrica a la derecha, (b) distribución simétrica, (c) distribución asimétrica a la izquierda. Fuente: elaboración propia.*

Los estimadores que se seleccionaron para realizar las comparaciones son los siguientes: el estimador de Rice (Ri), los estimadores HKT basados en diferencias óptimas de Hall, de orden 2 al 5 (Op2, Op3, Op4, Op5) y los estimadores basados en diferencias ordinarias orden 2 al 5 (Or2, Or3, Or4, Or5). Hay que tener en cuenta que el estimador de orden 1 de los dos métodos de asignación seleccionados coincide con el estimador de Rice. Además, cuando se tienen diseños equidistantes, como en este caso, el estimador basado en diferencias ordinarias de orden 2 es igual al estimador GSJS.

Como indicador para la comparación de los estimadores se utiliza el error cuadrático medio (ECM) empírico, calculado de la siguiente forma:

$$ECM(\hat{\sigma}_j^2) = \frac{1}{1000}\sum_{i=1}^{1000}(\hat{\sigma}_{ji}^2 - \sigma^2)^2 \qquad j = 1, \cdots, 9 \tag{18}$$

donde $\hat{\sigma}_{ji}^2$ es la estimación de la varianza por medio del estimador j en la simulación i. Este es un indicador que tienen en cuenta no solamente el sesgo del estimador sino también su variabilidad.

Las simulaciones se llevaron a cabo usando el software estadísticos R siguiendo estos pasos:

1. Se generan los valores de ε_i teniendo en cuenta la distribución de los errores, el valor de varianza y el tamaño de muestra propuesto.

2. Se generan los valores de y_i siguiendo el modelo de la ecuación (17), teniendo en cuenta cada una de las funciones poblacionales f propuestas. Donde $x_i = \frac{i-0.5}{n}$, $i = 1, \cdots, n$.

3. Luego de generar los valores de y_i, se procede a estimar la varianza por medio de cada uno de los estimadores seleccionados.

4. Los pasos anteriores se repiten 1000 veces para obtener la distribución empírica de cada estimador y así poder calcular el sesgo y el error cuadrático medio (ECM) de cada uno.

Teniendo en cuenta las diferentes distribuciones de los errores, tamaños de muestra y funciones de regresión se tienen 54 escenarios de simulación, en los cuales se realizaron las estimaciones de varianza por medio de los 9 estimadores propuestos.

6. Resultados

En cada una de las situaciones planteadas se estimó la varianza con cada uno de los estimadores seleccionados para el estudio, luego se encontró el sesgo y el ECM de cada uno y se observaron sus distribuciones de forma gráfica por medio de diagramas de cajas y alambres; todo ello a fin de realizar las comparaciones y determinar en qué casos es más recomendable el uso de uno de estos estimadores sobre los demás. A continuación se muestran diagramas de cajas y alambres y tablas del ECM de las tres funciones que se simularon, bajo diferentes cambios en los tamaños de muestra y varianzas.

Cuando se comparan todos los estimadores bajo la función que no presenta oscilaciones (Figura 2) se observa que los estimadores basados en diferencias óptimas de Hall (Op2, Op3, Op4 y Op5) presentan menor dispersión que los estimadores basados en diferencias ordinarias (Or2, Or3, Or4 y Or5) en todas las simulaciones, pero los últimos presentan mejor manejo del sesgo cuando se tienen muestras pequeñas (n=50). Si comparamos el ECM de los estimadores (Tabla 2) se puede observar que los valores para cada uno son muy parecidos, aunque los menores valores se observan para los estimadores Op2, Ri y Or2.

Tabla 2: *Error cuadrático medio de los estimadores bajo la función* $8\sin(0.5\pi x) - 4$ *para tamaños de muestra 50 y 300, y varianza igual a 0.5 y 1. Fuente: elaboración propia.*

Caso	Ri	Or2	Or3	Or4	Or5	Op2	Op3	Op4	Op5
$n = 50,\ \sigma^2 = 0.5$	0.014	0.018	0.023	0.027	0.030	0.013	0.017	0.025	0.044
$n = 50,\ \sigma^2 = 1$	0.065	0.087	0.105	0.121	0.136	0.054	0.054	0.061	0.079
$n = 300,\ \sigma^2 = 0.5$	0.002	0.003	0.004	0.004	0.005	0.002	0.002	0.002	0.002
$n = 300,\ \sigma^2 = 1$	0.010	0.013	0.015	0.017	0.019	0.008	0.008	0.007	0.007

En la Figura 3 y la Tabla 3 se puede observar el comportamiento de los estimadores bajo los escenarios que tienen la función de regresión que presenta pocas

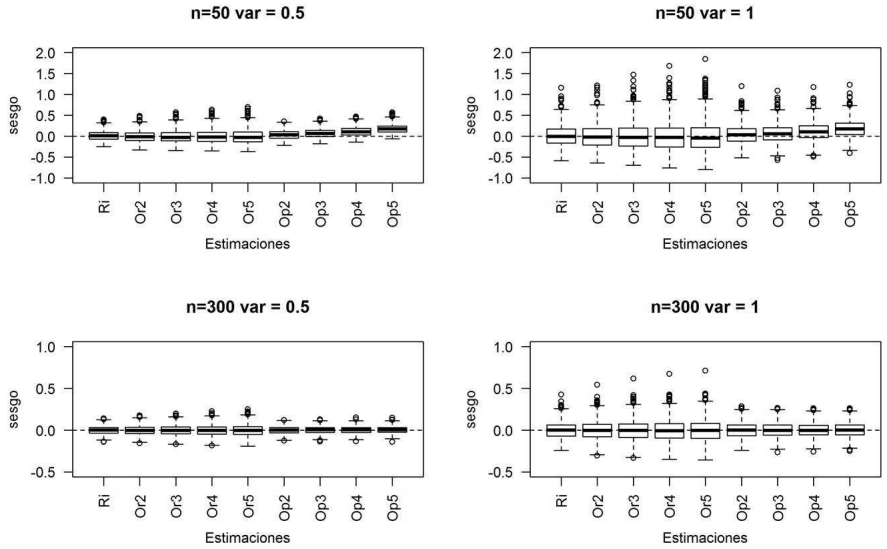

Figura 2: *Diagrama de cajas del sesgo de los estimadores bajo la función* $8\sin(0.5\pi x) - 4$ *para tamaños de muestra 50 y 300, y varianza igual a 0.5 y 1. Fuente: elaboración propia.*

oscilaciones. En la Figura 3 se aprecia que los estimadores basados en diferencias ordinarias tienen buen control sobre el sesgo sin importar los tamaños de muestra o el orden usado, a diferencia de los estimadores basados en diferencias óptimas que presentan estimaciones con sesgos positivos, especialmente para tamaños de muestra de 50, además el sesgo es mayor a medida que aumenta el orden. Es importante tener presente que estos sesgos disminuyen considerablemente cuando el tamaño de muestra es de 300, además presentan una variabilidad inferior que los estimadores basados en diferencias ordinarias.

Al observar el ECM (Tabla 3) se aprecia que para tamaños de muestra pequeños los estimadores basados en diferencias ordinarios son mejores que los estimadores basados en diferencias óptimas, pero para las situaciones con tamaño de muestra de 300 estos últimos tienen mejor comportamiento.

Tabla 3: *Error cuadrático medio de los estimadores bajo la función* $4\sin(3\pi x)$ *para tamaños de muestra 50 y 300, y varianza igual a 0.5 y 1. Fuente: elaboración propia.*

Caso	Ri	Or2	Or3	Or4	Or5	Op2	Op3	Op4	Op5
$n = 50,\ \sigma^2 = 0.5$	0.034	0.019	0.023	0.026	0.029	0.128	0.387	0.916	1.850
$n = 50,\ \sigma^2 = 1$	0.082	0.080	0.098	0.114	0.130	0.172	0.439	0.975	1.906
$n = 300,\ \sigma^2 = 0.5$	0.003	0.003	0.004	0.005	0.005	0.002	0.002	0.002	0.004
$n = 300,\ \sigma^2 = 1$	0.010	0.013	0.016	0.018	0.021	0.008	0.008	0.008	0.009

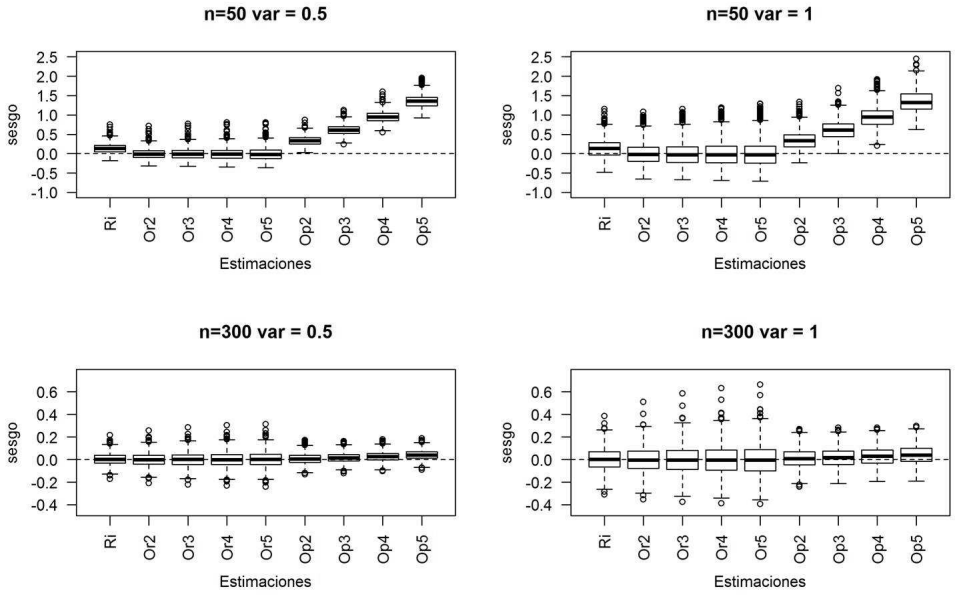

Figura 3: *Diagrama de cajas y alambres del sesgo de los estimadores bajo la función* $4\sin(3\pi x)$ *para tamaños de muestra 50 y 300, y varianza igual a 0.5 y 1. Fuente: elaboración propia.*

Cuando se tienen modelos con la función de regresión que presenta mayor número de oscilaciones, se puede observar en la Figura 4 que los estimadores basados en diferencias óptimas de Hall presentan estimaciones extremadamente sesgadas, aunque las medianas del sesgo parecen acercarse a 0 cuando el tamaño de muestra es de 300, al igual que en las anteriores simulaciones, pero bajo esta función estos estimadores parece que necesitan tamaños de muestra muchos más grandes para que se tenga buen control sobre el sesgo. Los estimadores basados en diferencias ordinarias presentan comportamientos más estables, puesto que todas las medianas estuvieron próximas a 0.

En la Tabla 4 se puede observar los ECM de los estimadores, donde se puede apreciar que para tamaños de muestra pequeños los estimadores basados en diferencias ordinarias presentan mejor comportamiento que los basados en diferencias óptimas, pero cuando el tamaño de muestra aumenta a 300, estos últimos mejoran considerablemente, aunque solamente el de orden 2 tiene un resultado similar al de los estimadores basados en diferencias ordinarias.

Al contrario que ocurre con los anteriores funciones poblacionales planteadas, donde en algunos casos los estimadores basados en diferencias óptimas de Hall son una buena alternativa, los estimadores basados en diferencias ordinarias son los únicos estimadores que presentan el comportamiento deseado para la estimación de la

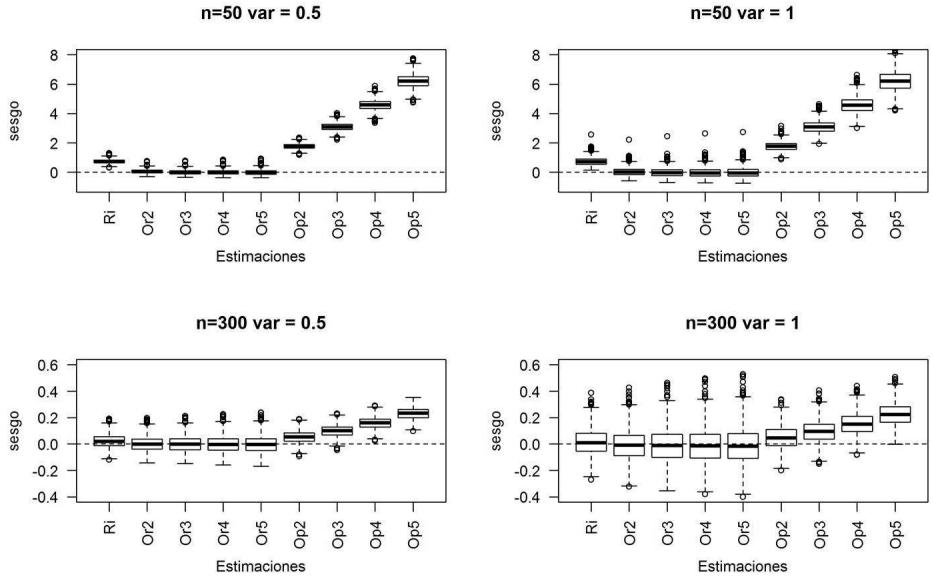

Figura 4: *Diagrama de cajas y alambres del sesgo de los estimadores bajo la función* $4\sin(7\pi x)$ *para tamaños de muestra 50 y 300, y varianza igual a 0.5 y 1. Fuente: elaboración propia.*

Tabla 4: *Error cuadrático medio de los estimadores bajo la función* $4\sin(7\pi x)$ *para tamaños de muestra 50 y 300, y varianza igual a 0.5 y 1. Fuente: elaboración propia.*

Caso	Ri	Or2	Or3	Or4	Or5	Op2	Op3	Op4	Op5
$n = 50,\ \sigma^2 = 0.5$	0.583	0.025	0.026	0.030	0.034	3.161	9.608	21.065	38.634
$n = 50,\ \sigma^2 = 1$	0.622	0.084	0.097	0.110	0.122	3.230	9.708	21.286	39.117
$n = 300,\ \sigma^2 = 0.5$	0.003	0.003	0.004	0.004	0.005	0.005	0.012	0.028	0.056
$n = 300,\ \sigma^2 = 1$	0.011	0.014	0.016	0.019	0.021	0.011	0.017	0.032	0.059

varianza, siendo estos los estimadores recomendados cuando se tengan situaciones similares a esta última.

Puesto que todas las situaciones simuladas anteriormente se hicieron bajo una distribución de los errores simétrica $N(0, \sigma^2)$, por lo tanto hace falta observar si existen diferencias en las estimaciones cuando se tiene una distribución asimétrica de los errores. Para esto se presenta en la Figura 5, distribución de los sesgos de tres estimadores de varianza (el estimador de Rice, el basado en diferencia ordinarias de orden 2 y el basado en diferencia óptimas de Hall de orden 2) por medio de diagrama de cajas y alambres bajo tres condiciones distintas de distribución de los errores (asimétrica a la derecha, simétrica, asimétrica a la izquierda).

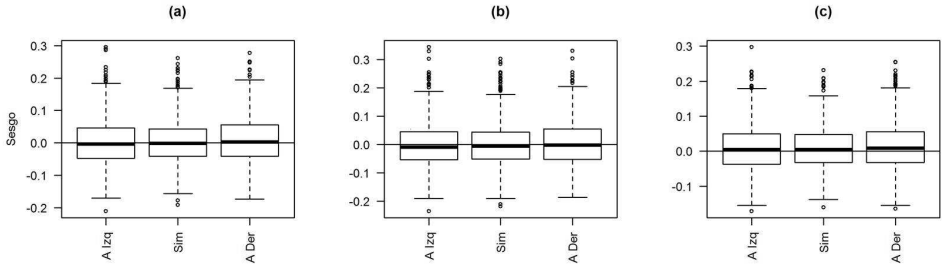

Figura 5: *Diagrama de cajas y alambres de los estimadores bajo las tres condiciones de los errores con la función* $8\sin(0.5\pi x) - 4$ *y* $n = 100$*. (a) estimador de Rice (b) estimador de diferencias ordinarias de orden 2 (c) estimador de diferencias óptimas de orden 2. Fuente: elaboración propia.*

En la Figura 5 se observa que los diagramas de cajas y alambres del sesgo del estimador de Rice bajo las tres distribuciones de los errores no presentan diferencias en sus comportamientos, al igual que ocurre con los otros dos estimadores evaluados, lo que nos indica que estos estimadores de varianza no se ven afectados por la distribución de los errores; ocurre lo contrario con los estimadores que están basados en los diferentes métodos de suavización, los cuales deben asumir normalidad de los errores. Cuando se simulan los resultados bajo los demás escenarios planteados se tienen las mismas observaciones mencionadas anteriormente, como también ocurre con los demás estimadores que se tienen en cuenta en este estudio.

7. Conclusiones

De los estimadores basados en diferencias óptimas de Hall, el estimador de orden 2 es el que presenta mejor comportamiento, puesto que en las simulaciones planteadas se observó que se producían estimaciones cada vez más sesgadas y con mayor ECM cuando el orden de este estimador aumentaba. Por lo cual no es recomendable el uso de los estimadores basados en diferencias óptimas de Hall de órdenes superiores a 2.

En ninguna de las situaciones simuladas se encontró diferencias en las distribuciones de los sesgos de los estimadores basados en diferencias ordinarias, por lo cual el uso de cualquiera de estos estimadores, sin importar el orden, produce estimaciones muy similares bajo situaciones parecidas a los escenarios propuestos.

El estimador de Rice presentó buen comportamiento en algunas de las situaciones planteadas, aunque en ninguno de estos escenarios presentó el mejor comportamiento sobre los demás estimadores. Es decir, el uso del estimador de Rice no se recomienda en ninguno de los casos.

Cuando se tienen funciones sin cambios u oscilaciones el estimador basado en

diferencias óptimas de Hall de orden 2 presenta mejor comportamiento que los estimadores basados en diferencias ordinarias, ya que este estimador presenta menor dispersión que los otros estimadores y tiene buen manejo del sesgo; lo anterior se ve reflejado en los menores valores del ECM. Al tener funciones con oscilaciones es necesario que se tenga un tamaño de muestra grande, para que este estimador tenga mejor comportamiento que los estimadores basados en diferencias ordinarias.

Los estimadores basados en diferencias ordinarias tienen un buen manejo del sesgo en todos los escenarios que se plantearon con las diferentes funciones poblacionales. Pero tiene mejor comportamiento que los estimadores basados en diferencias óptimas de Hall cuando se tienen funciones con oscilaciones, y además el tamaño de muestra es pequeño.

No se encontraron diferencias significativas en ninguna de las distribuciones de los sesgos de los estimadores cuando se plantean diferentes distribuciones de los errores, por lo cual no es necesario asumir ningún comportamiento de los errores para el uso de alguno de estos estimadores de varianza.

8. Trabajo futuro

Dado que en este trabajo se usaron diseños equidistantes, una posible extensión sería estudiar las diferencias que se presentan entre estos estimadores cuando se tienen diseños aleatorios o diseños no equidistantes. De esta forma también se pueden plantear diferencias entre el estimador basado en diferencias ordinarias y el estimador GSJS, puesto que en caso de un diseño equidistante estos estimadores son iguales.

Se podría además proponer algún criterio para establecer el tipo de estimador basado en diferencias que se debe usar dependiendo de la situación que se tenga, como el tamaño de muestra (puesto que se observó que el estimador basado en diferencias ordinarias presentó mejor comportamiento cuando n es pequeño, pero cuando n es grande los estimadores óptimos presentaron mejor comportamiento) y el tipo de función.

Recibido: 22 de noviembre de 2013
Aceptado: 20 de marzo de 2014

Referencias

Altman, N. S. (1992), 'An introduction to kernel and nearest-neighbor nonparametric regression', *The American Statistician* **46**(3), 175–185.

Bowman, A. W. & Azzalini, A. (1997), *Applied Smoothing Techniques for Data Analysis. The Kernel Approach with S-plus Illustrations*, Oxford University Press.

Brown, L. D. & Levine, M. (2007), 'Variance estimation in nonparametric regression via the difference sequence method', *Annals of Statistics* **35**(5), 2219–2232.

Buckley, M. J., Eagleson, G. K. & Silverman, B. W. (1988), 'The estimation of residual variance in nonparametric regression', *Biometrika* **75**(2), 189–199.

Carter, C. K. & Eagleson, G. K. (1992), 'A Comparison of Variance Estimators in Nonparametric Regression', *Journal of the Royal Statistical Society, Series B* **54**(3), 773–780.

Cleveland, W. S. (1979), 'Robust Locally Weighted Regression and Smoothing Scatterplots', *Journal of the American Statistical Association* **74**(368), 829–836.

Dette, H., Munk, A. & Wagner, T. (1998), 'Estimating the Variance in Nonparametric Regression. What is a Reasonable Choice?', *Journal of the Royal Statistical Society, Series, B* **60**(4), 751–764.

Draper, N. R. & Smith, H. (1966), *Applied Regression Analysis*, John Wiley & Sons, New York.

Draper, N. R. & Smith, H. (1998), *Applied Regression Analysis*, 3 edn, John Wiley & Sons, New York.

Eubank, R. L. (1998), *Nonparametric Regression and Spline Smoothing*, 2 edn, Marcel Dekker, New York.

Gasser, T., Sroka, L. & Jennen-Steinmetz, C. (1986), 'Residual variance and residual pattern in nonlinear regression', *Biometrika* **73**(3), 625–633.

Hall, P., Kay, J. W. & Titterington, D. M. (1990), 'Asymptotically optimal difference-based estimation of variance in nonparametric regression', *Biometrika* **77**(3), 521–528.

Hall, P. & Marron, J. S. (1990), 'On variance estimation in nonparametric regression', *Biometrika* **77**(2), 415–419.

Levine, M. (2006), 'Bandwidth selection for a class of difference-based variance estimators in the nonparametric regression: A possible approach', *Journal Computational Statistics & Data Analysis* **50**(12), 3405–3431.

Olaya, J. (2012), *Métodos de regresión no paramétrica*, Programa Editorial Universidad del Valle, Colombia.

Olmos, N. M., Varela, H., Gómez, H. W. & Bolfarine, H. (2012), 'An extension of the half-normal distribution', *Statistical Papers* **53**(4), 875–886.

Rice, J. (1984), 'Bandwidth choice for nonparametric regression', *Annals of Statistics* **12**(4), 1215–1230.

Seifert, B. & Gasser, T. (1993), 'Nonparametric estimation of residual variance revisited', *Biometrika* **80**(2), 373–383.

Wahba, G. (1978), 'Improper priors, spline smoothing, and the problem of guarding against model errors in regression', *Journal of the Royal Statistical Society, Series, B* **40**(3), 364–372.