UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE AGRONOMÍA
COMISIÓN DE ESTUDIOS DE POSTGRADO
POSTGRADO EN ESTADÍSTICA

ESTIMACIÓN DE PARÁMETROS DE UN MODELO DE REGRESIÓN LINEAL SIMPLE DISCONTINUO A TRAVÉS DE LA ESTIMACIÓN KERNEL

Prof. Juan B. Mejía G
Tutor: Dr. Miguel Balza
Tutores Consejeros: Dra. Harú Martínez de Cordero
M.Sc. Jorge Flores

Maracay, Marzo de 2013

Trabajo Especial de Grado presentado ante los honorables miembros del Comité Académico del Postgrado de Estadística de la Facultad de Agronomía de la Universidad Central de Venezuela como requisito para optar al Título de Magister Scientiarum en Estadística.

DEDICATORIAS

A dios todo poderoso

A mis padres que me dieron la vida

AGRADECIMIENTOS

A mi madre que en paz descanse, quien siempre me apoyo.

A mi esposa quien me acompaña en este camino de vida.

Al Profesor Luis Pérez quien con su incondicional apoyo me fortaleció en la realización de este proyecto.

A mi tutor profesor Miguel Balza quien siempre confió en mi y mis tutores consejeros quienes me brindaron su apoyo.

TABLA DE CONTENIDOS

TABLA DE CUADROS

TABLAS DE FIGURAS

ESTIMACIÓN DE PARÁMETROS DE UN MODELO DE REGRESIÓN LINEAL SIMPLE DISCONTINUO A TRAVÉS DE LA ESTIMACIÓN KERNEL

Prof. Juan B. Mejia G
Tutor: Dr. Miguel Balza

RESUMEN

El principal objetivo es estimar el modelo de regresión lineal simple discontinuo por la metodología de regresión kernel donde se evaluará el ajuste del modelo con respecto al clásico.

Este estudio se enmarca dentro de la investigación exploratoria, descriptiva y explicativa pues por medio del análisis, comparación y descripción de los modelos clásicos y el de regresión kernel se puede establecer las diferencias entre ambos métodos.

Los datos estadísticos que sustentan esta investigación vienen por la simulación de estos, en tres escenarios donde las funciones del antes y después del punto de discontinuidad tienen igual pendiente en un caso y pendientes diferentes en los otros dos casos. La data fue construida por el programa Excel 2007 y procesados en el xlstat 2011 programa complementario del Excel. Para la comparación de los modelos se utilizó el coeficiente de determinación R^2 ajustado y para la metodología de regresión kernel, el estimador de Nadaraya-Watson con polinomios locales de grado 1.

El resultado obtenido es que la estimación de los parámetros no es única para un modelo de regresión simple discontinuo, pues se realiza una estimación de parámetros para cada punto x_i estimado.

Finalmente se concluye que la eficiencia del modelo kernel es superior o igual al modelo de regresión lineal clásico y que la regresión kernel permite hallar una sola curva estimada ajustada a los datos y suavizada en la discontinuidad.

Palabras claves: Estimador Nadaraya-Watson, Polinomios Locales, Regresión Lineal Simple Discontinua, Regresión Kernel, Estimación de Parámetros.

ESTIMATION OF PARAMETERS A DISCONTINUOUS LINEAR REGRESSION SIMPLE MODEL ACROSS THE ESTIMATION KERNEL

Prof. Juan B. Mejia G
Tutor: Dr. Miguel Balza

ABSTRACT

The main objective is to estimate the simple linear regression model for discontinuous kernel regression methodology which evaluate the fit of the model with respect to the classic.

This study is part of exploratory research, descriptive and explanatory because through the analysis, comparison and description of classical models and kernel regression can establish the differences between the two methods.

The statistics that support this research arise for the simulation of these, in three scenarios where the functions before and after the discontinuity point have the same slope in a pending case and different in the other two cases. The data was built by Excel 2007 and processed in the supplementary program XLSTAT 2011 Excel. For comparison of the models used the adjusted coefficient of determination R^2 and kernel regression methodology, the Nadaraya-Watson estimator with local polynomials of degree 1.

The result is that the estimation of the parameters is not unique for a simple regression model was discontinued as an estimation of parameters estimated for each point x_i.

Finally we conclude that the kernel model efficiency is not less than classical linear regression model and kernel regression allows one to find estimated curve fitted to the data and smoothed discontinuity.

Key words: Estimator Nadaraya-Watson, Local Polynomials, Linear Simple Discontinuous Regression, Regression Kernel, Estimation Parameter

INTRODUCCIÓN

El modelo de regresión lineal relaciona la variable dependiente Y con las variables regresoras X_i (*con* $i=1, 2, \ldots, n$) con un error aleatorio ε_i que toma aquellos factores de la realidad no controlables u observables, donde se desea estimar los parámetros β_i *(con* $i=0,1,2,\ldots,n$*)* las cuales miden las influencias de las variables independientes en función de la variable dependiente. Para que dicho modelo tenga validez, previamente se debe probar los supuestos de homocedasticidad, normalidad de los errores, independencia de los errores y no colinealidad.

Los estudios realizados en campos psicológicos, educativos, económicos, entre otros, generó situaciones tales, en las cuales la variable dependiente o independiente tenga una discontinuidad o salto en su comportamiento. Tal situación generó el hecho que al construir un modelo de regresión lineal por los métodos clásicos se obtuvo que la validez del modelo no fuese la idónea porque incumplía con los supuestos del modelo, creando que el análisis y el ajuste no fueran idóneos. En esta investigación la variable dependiente es quien contiene la discontinuidad.

El diseño de regresión lineal discontinuo fue sugerido por primera vez por Thistlethwaite y Campbell (1960). Los autores realizaban un modelo lineal clásico donde se estudiaba la variable dependiente como un antes y después del salto o discontinuidad, realizando luego una comparación entre ambas regresiones construidas. El hecho de escoger una variable y considerarla como dos efectos individuales no es lo más apropiado pues la variable solo sufre una variación no controlada.

Diversos autores que se mencionan en el marco teórico, han presentado soluciones al hecho de discontinuidad en la variable dependiente en un modelo de regresión lineal a través de modelos paramétricos o no paramétricos.

El presente trabajo, propone realizar la estimación del modelo de regresión lineal simple discontinuo a través de la regresión kernel como solución a la estimación de un modelo único que explique el comportamiento de la variable dependiente suavizando la discontinuidad o salto y ajustando a una única curva el modelo. Estimado el modelo con la regresión kernel se realizará una valoración de la eficiencia contra el modelo clásico el cual estudia la variable Y como un antes y después del salto, tal comparación será con respecto al coeficiente de determinación R^2 que genere cada modelo.

ANTECEDENTES

Entre los autores que han desarrollado el tema de la regresión discontinua se encuentran, Thistlethwaite y Campbell (1960)*;* en su trabajo realizado, ellos implementan por primera vez y de manera matemáticamente bien desarrollado el tema de análisis de regresión discontinua, pues se utilizó el carácter discontinuo como un tratamiento más. Aparte los efectos, se anexaba al modelo un tratamiento que era justamente donde ocurría la discontinuidad; adicionalmente, los autores *in comento* realizaban una comparación entre grupos que se generaban a partir de la discontinuidad. A partir de estas comparaciones, verificaban si los conjuntos de datos en discontinuidad tenían tendencia de comportamientos iguales con respecto a las variables en estudio.

La regresión discontinua fue tema entre distintos autores, Campbell y Stanley (1963) presentan una monografía en la que siguen realizando un análisis de regresión antes y después pero donde deseaban elevar este análisis a ser un diseño, pues los autores consideraban que el modelo aleatorizado que presentaban Thistlethwaite y Campbell (1960) no era justificable en un entorno dado pues el diseño de regresión discontinua era un diseño cuasi-experimental. Ellos reconocían que la regresión discontinua era de corte *sharp* o *fuzzy*, propusieron usar estudios de correlación antes y después del punto de corte, estimar el punto de salto para poder relacionar el comportamiento de las regresiones generadas antes y después de la discontinuidad.

Campbell (1969) realizó una investigación donde definió el ámbito de los puntos de corte y la agudeza o difusión de los datos en estos puntos. Aparte, retomó la discusión del tema de sí era un análisis o un diseño. Además, empleó pruebas del antes – después para los datos discontinuos y utilizó un criterio de medida para los puntos de corte.

Boruch (1973) desarrolló un análisis estadístico que detalló y distinguió la consideración de las variables aleatorias mixtas y en las que cada grupo

separado por la discontinuidad representaba grupos intactos para el análisis correspondiente. Propuso un análisis de regresión simple con los grupos construidos, asignándoles un valor *dummy* a las variables, generando así un estudio individual de cada grupo. Con ello se realizaba un análisis de regresión individual, para luego a través de una prueba *t*, poder comparar en el diseño las diferencias y similitudes en los análisis de regresión entre los grupos formados por los puntos de discontinuidad.

Boruch y DeGracie (1975) presentan una evaluación de compensación en el programa educativo de los Estados Unidos presentado al congreso donde usan la regresión kernel, siendo su aporte más importante el de especificar mejor el modelo y los efectos curvilíneos en el análisis estadístico.

En la década de los setenta a los noventa, se desarrollaron análisis de regresión discontinuas, pero solamente aplicados a un problema particular, tal como lo hicieron Cook y Campbell (1979), Judd y Kenny (1981), Kidder (1981), Berk y Rauma (1983), entre otros; quienes desarrollaron el análisis de regresión para el estudio de situaciones, como la psicología, comportamiento de los estudiantes, comportamiento de presos, comportamiento de la economía, para el análisis de producción en empresas. Otros autores realizaron investigación acerca de la metodología y la importancia de nombrar al análisis de regresión discontinua como diseño de regresión, a título de ejemplos: Golberger (1972), Sacks y Ylvisaker (1976), Rubin (1977), Trochim (1980), entre otros.

A principios de los noventa, Branden y Bryant (1990), propusieron un modelo de regresión múltiple con análisis de correlación pero las variables codificadas de manera dummy. Por otro lado, Fan y Gijbels (1996) realizan un aporte importante con respecto al análisis de regresión discontinua como lo es la estimación no paramétrica al campo del análisis de regresión discontinua para poder unificar en una sola estimación la curva de regresión. Härdle (2004)

propone algunos modelos semi-paramétricos y no paramétricos para la construcción del modelo de regresión; además, un óptimo ancho de banda, lo que permite que Lorca Federico (2006) introduzca la definición del estimador de Nadaraya-Watson, para construir un modelo de regresión no paramétrica. Lee Howard (2008) propone la prueba de Chow para el análisis de una regresión lineal discontinua donde la prueba permite hacer una comparativa entre variables diferentes con comportamientos similares. El autor, basándose en investigaciones anteriores, refuerza el estudio del antes y después del salto o discontinuidad. Imbems y kalyanaraman (2009) investigaron sobre el problema de optimizar el parámetro de suavizado (ancho de banda) para la regresión de una estimación discontinua. Como se observa, en los antecedentes anteriormente esbozados, en ningún momento se persigue eliminar la discontinuidad, más bien los puntos estimados sobreviven con ella.

JUSTIFICACIÓN

Las investigaciones indicadas anteriormente, acerca de la estimación del modelo de regresión discontinua son el estudio para lograr sobreponerse al hecho del salto que tienen los datos en su realidad y llegar a explicación de la variable *Y* con respecto a la variable independiente, en ningún momento evita la discontinuidad; sólo se la apropian y trabajan en función de ella. Algunos autores consideran esa discontinuidad como un tratamiento; otros simplemente generan un modelo de un antes y un después.

Este trabajo busca presentar una vía para resolver el problema de la discontinuidad en el modelo de regresión lineal simple; no obstante, la finalidad es el de obtener una estimación de la regresión lineal, donde se propone usar una metodología de regresión kernel. Por otra parte, para la comprobación del procedimiento, se realiza una comparación con respecto al coeficiente de determinación R^2.

Téngase en cuenta que todo problema suscitado debe tener varias alternativas para poder solventar, aparte se debe tomar la mejor metodología que se adapte a la situación en estudio. Por ello, se plantea la estimación kernel como una posible solución; pues, esta función lo que realiza es suavizar el hecho de existencia de discontinuidad en los datos de un modelo de regresión, tal función se puede adaptar al comportamiento de los datos tratando de eliminar la discontinuidad a la hora de poder formalizar un modelo de regresión que se ajuste y explique los datos en su realidad, obteniendo así otra alternativa que permita a los investigadores a futuro, escoger la opción conveniente a su estudio a realizar.

OBJETIVO GENERAL

Proponer un procedimiento de estimación kernel para la estimación de parámetros de un modelo de regresión lineal simple discontinuo.

OBJETIVOS ESPECÍFICOS

Generar una función Kernel de mejor ajuste al modelo de regresión lineal simple discontinuo.

Evaluar el ajuste de la función Kernel utilizada en el modelo de regresión simple discontinuo.

Estimar los parámetros de la regresión lineal simple con la metodología de estimación Kernel.

Verificar si el modelo de estimación es apropiado en comparación con el modelo clásico

MARCO TEÓRICO

Modelo Lineal Simple

El modelo de regresión lineal simple estudia la relación lineal entre la variable respuesta (*Y*) y la variable regresora (*X*), a partir de una muestra $\{(x_i, Y)\}_{i=1,2,\dots,n}$ que sigue el modelo:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1,2,3,\dots,n$$

Por tanto, es un modelo de regresión paramétrico de diseño fijo. En forma matricial

$$\vec{y} = \beta_0 \vec{1} + \beta_1 \vec{x} + \vec{\varepsilon_i}$$

donde $\vec{y}^t = (y_1, y_2, \dots, y_n)$, $\vec{1}^t = (1,1,\dots,1)$, $\vec{x}^t = (x_1, x_2, \dots, x_n)$,

$\vec{\varepsilon}^t = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$,

Se supone que se verifican las siguientes hipótesis:

1. La función de regresión es lineal,

   $$m(x_i) = \beta_0 + \beta_1 x_i, \qquad i = 1,2,3,\dots,n$$

   o, equivalentemente que la esperanza de los errores aleatorios es cero $E(\varepsilon_i) = 0$, *i = 1,...,n.*

2. La varianza es constante (homocedasticidad),

   es decir, *Var* $(\varepsilon_i) = \sigma^2$, *i = 1,...,n.*

3. La distribución es normal,

   eso es, $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1,...,n$.

4. Las observaciones Y_i son independientes. Bajo las hipótesis de normalidad, esto equivale a que la $Cov(Y_i, Y_j) = 0$, si $i \neq j$
5. Esta hipótesis, en función de los errores sería "los ε_i son independientes", donde bajo normalidad, equivale a que $Cov(\varepsilon_i, \varepsilon_j) = 0$, si $i \neq j$

Al hecho del incumplimiento de estos supuestos por la discontinuidad de los datos experimentales se utilizará la metodología kernel.

Modelo Lineal General

Los modelos clásicamente usados en Estadística son paramétricos y la suposición es que la muestra de observaciones proviene de una familia paramétrica conocida. En estos casos, el problema es estimar los parámetros desconocidos, realizar pruebas de hipótesis u obtener intervalos de confianza para los mismos. Esta suposición puede ser relativamente fuerte porque el modelo paramétrico supuesto puede no ser el correcto, si existe alguno (los datos pueden ser tales que no exista una familia paramétrica adecuada que de un buen ajuste). Por otra parte, los métodos estadísticos desarrollados para un modelo paramétrico particular pueden llevar a conclusiones erróneas cuando se aplican a un modelo ligeramente perturbado (falta de robustez respecto del modelo). Estos problemas llevaron a la tendencia de desarrollar métodos no paramétricos o semi-paramétricos para analizar los datos.

Por lo cual, se plantea un modelo general y a partir de este se construirá el modelo adecuado. Sea (x_i, y_i) y $\varepsilon_i \sim (0, \sigma^2)$

$$y_i = m(x_i) + \varepsilon_i \qquad donde \ \ i = 1,2,\dots,n$$

m es la función a estimar y a través del kernel se construirá la expresión definitiva de la estimación.

Estimación a través de la función Kernel

Según Acuña, Edgar (2007) si las funciones de densidades condicionales $f_i(x)$ de la clase C_i son desconocidas pueden ser estimadas usando la muestra tomada y aplicando métodos de estimación, uno de ellos es el método de Kernel.

En el caso univariado, el estimador Kernel de la función de densidad *f(x)* se obtiene de la siguiente manera. Considérese que $x_1, x_2, \dots, x_n$ es una variable aleatoria *X* con función de densidad *f(x)*, defínase la función de distribución empírica por

$$F_n(x) = \frac{n^{\underline{o}}\ obs \leq \ x}{n}$$

donde $n^{\underline{o}}\ obs \leq \ x$ es el numero de observaciones menores o iguales al *x* seleccionado y *n* el número total de datos, el cual es un estimador de la función de distribución acumulada *F(x)* de *X*.

Considerando que la función de densidad *f(x)* es la derivada de la función de distribución *F* y usando aproximación Rosenblatt (1956) propone como estimador:

$$\hat{f}(x) = \frac{n^{\underline{o}}\{X_i : X_i \in (x-h, x+h)\}}{2hn} = \frac{F_n(x+h) - F_n(x-h)}{2h}$$

recordando que

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x-h < X < x+h)$$

por otro lado

$$E(F_n(x)) = F_x$$

$$E\left(\hat{f}(x)\right) = \frac{1}{2h}(F(x+h) - F(x-h))$$

$$\lim_{h\to 0} E\left(\hat{f}(x)\right) = f(x)$$

además el numerador de $\hat{f}(x)$ sigue una distribución binomial

$$B(n, F(x+h) - F(x-h))$$

siendo la varianza

$$Var\hat{f}(x) = \frac{1}{4h^2n^2} n[F(x+h) - F(x-h)][1 - (F(x+h) - F(x-h))]$$

por tanto

$$\lim_{h\to 0} Var\hat{f}(x) = \frac{f(x)}{2hn}$$

De esto se tiene que la estimación es consistente por el siguiente Teorema 1 que dice: "Dado $x \in B_i$ y si f es *Lipshitz* en B_i, entonces $\hat{f}(x)$ es consistente en media cuadrática si $n \to \infty \Rightarrow h \to 0 \;\; y \;\; nh \to \infty$". De acuerdo a Rosenblatt (1956) y Balza (1980) puede escribirse el Error Cuadratico Medio Integrado Aproximado (ECMIA) de esta función como:

$$ECMIA = \frac{1}{2hn} + \frac{h^4}{36} R(f'')$$

Y dicho valor es minimizado por

$$h^* = \left(\frac{9}{2R(f'')n}\right)^{1/5}$$

y el ECMIA resultante

$$ECMIA = \frac{5}{4.2^{4/5}.9^{1/5}.n^{4/5}} (R(f''))^{1/5}$$

Se debe tener en cuenta que el estimador $\hat{f}(x)$ es discontinuo en $X_i \pm h$ y es constante entre esos valores, por lo tanto ese estimador puede escribirse de la forma

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right)$$

donde *h* es un valor positivo cercano a cero y la función peso *K* está definida por

$$K(u) = \begin{cases} 0 & si \quad u \in [-1, 1] \\ 1/2 & en\ caso\ contrario \end{cases}$$

este es llamado la función Kernel Uniforme.

Función Kernel

Tomada la muestra de n observaciones reales $x_1, x_2, \ldots, x_n$ se define según Lorca, F (2006) la estimación de la función núcleo *K* como:

$$\hat{f}_n(x) = \frac{1}{nh_n}\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right)$$

donde *K(x)* es una función, denominada función Kernel, que satisface ciertas condiciones de regularidad, generalmente es una función de densidad simétrica como por ejemplo la de la distribución normal, y $\{h_n\}$ es un conjunto de constantes positivas conocidas como parámetro de suavización o *bandwith*.

El estimador kernel puede interpretarse como una suma de saltos. La función *K* determina la forma de los saltos mientras que el parámetro h_n determina su anchura. También puede interpretarse como una transformación continua de la función de distribución empírica, de acuerdo a la función *K(x)*

que se encarga de redistribuir la masa de probabilidad *1/n* en la vecindad de cada punto muestral.

Un inconveniente de la estimación Kernel es que al ser el parámetro de ventana fijo a lo largo de toda la muestra, existe la tendencia a presentarse distorsiones en las colas de la estimación.

Entre los kernels más usados están:

a) El kernel Rectangular o Uniforme es definido por

$$K(u) = \begin{cases} 0 & si \quad |u| > 1 \\ 1/2 & si \quad |u| \leq 1 \end{cases}$$

En este caso cualquier punto en el intervalo (*x-h, x+h*) contribuye *1/2nh* al estimado de *f(x)* en el punto *x*, y cualquier punto fuera de ese intervalo no contribuye en nada.

b) El Kernel Gaussiano definido por

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \qquad |u| < \infty$$

En este caso el Kernel representa una función peso más suave donde todos los puntos contribuyen al estimado de *f(x)* en *x*.

c) El Kernel Triangular definido por

$$K(u) = \begin{cases} 1 - |u| & para\ |u| < 1 \\ 0 & en\ otros\ casos \end{cases}$$

d) El Kernel "Bipeso" definido por

$$K(u) = \begin{cases} {}^{15}/_{16}\,(1 - u^2)^2 & para\ |u| < 1 \\ 0 & en\ otro\ caso \end{cases}$$

e) El kernel Epanechnikov definido por

$$K(u) = \begin{cases} \frac{3}{4}(1 - u^2) & para\ |u| < 1 \\ 0 & en\ otro\ caso \end{cases}$$

f) El kernel “Tripeso” definido por

$$K(u) = \begin{cases} \frac{35}{32}(1-u^2)^3 & para\ |u| < 1 \\ \\ 0 & en\ otros\ casos \end{cases}$$

g) El kernel Arco coseno definido por

$$K(u) = \begin{cases} \frac{\pi}{4} cos\left(\frac{\pi}{2}u\right) & para\ |u| < 1 \\ \\ 0 & en\ otros\ casos \end{cases}$$

Propiedades Estadísticas

Consistencia

Las condiciones que se presentan a continuación estudian el sesgo y la consistencia en un punto *x* para las estimaciones de la función kernel, donde tal función es simétrica, basadas bajo el Teorema de Bochner (1955) que se menciona más adelante.

1. $\int_{-\infty}^{\infty} K(u)du = 1$
2. $\lim_{n\to\infty} h(n) = 0$
3. $\int_{-\infty}^{\infty} |K(u)du| < \infty$

Teorema 2, Bochner (1955) “Sea *K(y)* una función acotada que satisface las condiciones $\int_{-\infty}^{\infty} |K(u)du| < \infty$ y $\lim_{n\to\infty} h(n) = 0$. Sea $g \in l$. Sea

$$g_n(x) = \frac{1}{h_n}\int_{-\infty}^{\infty} K\left(\frac{y}{h}\right) g(x-y)dy$$

donde $\{h_n\}$ son un conjunto de constantes positivas que satisface $\lim_{n\to\infty} h(n) = 0$. Entonces si *x* es un punto de continuidad de *g*,

$\lim_{n\to\infty} g_n(x) = g(x)\int_{-\infty}^{\infty} K(y)dy$”

Del teorema anterior, se tiene un Corolario 2.1 el cual dice "La estimación definida de $\hat{f}(x)$ es asintóticamente insesgada en todos los puntos *x* en los cuales la función de densidad de probabilidad es continua si las constantes de h_n satisfacen $\lim_{n\to\infty} h(n) = 0$ y si la función *K(y)* satisface $\int_{-\infty}^{\infty} K(u)du = 1$ y $\int_{-\infty}^{\infty} |K(u)du| < \infty$"

Teorema 3 "El estimador definido $\hat{f}_n(x)$ es consistente, es decir el error cuadrático medio $ECM\left[\hat{f}_n(x)\right] \to 0 \quad \forall x \in R$ cuando $n \to \infty$, se añade la condición adicional de que $\lim_{n\to\infty} n\, h(n) = \infty$"

Del teorema 3, se tiene que una convergencia rápida a 0 del parámetro h_n provoca una disminución del sesgo, pero la varianza aumentaría de forma considerable, es de esta manera que la elección del ancho de banda ideal debe de converger a cero pero a un ritmo más lento que *1/n*. Para mejor detalles de estos teoremas y Corolario, ver Balza (1980).

El parámetro *h* es llamado el ancho de banda. Si *h* es muy pequeño entonces el estimador de densidad por Kernel degenera en una colección de *n* picos cada uno de ellos localizado en cada punto muestral. Si *h* es demasiado grande entonces el estimado se sobre-suaviza y se obtiene casi una distribución uniforme. El valor de *h* también depende del tamaño de la muestra, con muestras pequeñas se debe escoger un *h* grande y con muestras grandes se puede escoger un *h* pequeño.

Rosenblat (1956) demostró que el estimador Kernel es consistente y asintóticamente Normal.

Se debe tomar en cuenta que cuando la muestra es grande es más conveniente elegir *h* pequeño. A continuación se listan algunas elecciones de *h*:

a) $h = \frac{rango\ (X)}{2(1+\log_2 n)}$

b) $h = 1{,}06\,.\ min\left(\hat{\sigma}, {}^{R}/_{1{,}34}\right) n^{-1/5}$

donde $\hat{\sigma}$ es la desviación estándar estimada del conjunto de datos y R representa el rango intercuartílico, las constantes provienen de asumir que la densidad desconocida es Normal y un Kernel Gaussiano. Este es básicamente el método usado por *SAS/INSIGHT* para estimar la curvatura.

c) $h = 1{,}44\,.\hat{\sigma} n^{-1/5}$

Otros métodos más sofisticados son:

d) El método de Sheather y Jones (1991) que propone estimar la curvatura usando también el método del Kernel, pero con un ancho de banda *g* distinto al que se usa para estimar la densidad.

e) Usando validación cruzada, propiamente el método "dejando uno afuera". Aquí el *h* es considerado como un parámetro que debe ser estimado. Hay dos alternativas, usando mínimos cuadrados (aquí se obtiene un estimador insesgado), o máxima verosimilitud (aquí se obtiene un estimador sesgado). Ver Bowman y Azzalini (1997), para una implementación en *S-Plus*.

f) Usando "*Bootstrapping*", en este caso se encuentra un estimado del Error Cuadrático Medio Integrado usando muestras con reemplazamiento y se minimiza con respecto a *h*.

Cao, Cuevas y González (1994) hacen una comparación de varios métodos de elegir el ancho de banda *h* y llegan a la conclusión de que, sin considerar el "*boostrapping*", el método *d* es el de mejor rendimiento.

La viabilidad de este trabajo se presenta de una manera muy sencilla, pues se genera una función Kernel adaptada a la regresión discontinua y tratar de sobreponer ese salto y así poder estimar la mejor función que se ajuste a los datos.

El método kernel se ha utilizado para el mejoramiento de la estimación de una regresión lineal discontinua, tal como se refiere a continuación:

El método más simple de suavizamiento es la regresion Kernel. Un punto *x* se fija en el soporte de la función *m(.)* y un ancho de banda es definida alrededor de *x*.

La estimación Kernel es un promedio ponderado de las observaciones dentro de la ventana de suavizamiento

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)}$$

donde *K(.)* es la función Kernel de ponderación. La función Kernel es escogida de tal forma que las observaciones más próximas a *x* reciben mayor peso. Una función frecuentemente utilizada es la bicuadrática:

$$K(x) = \begin{cases} (1 - x^2)^2, & -1 \leq x \leq 1 \\ 0, & x > 1, \ x < -1 \end{cases}$$

Sin embargo, otro tipo de funciones de peso son utilizadas, tal como la gaussiana, $K(x) = \frac{1}{2\sqrt{\pi}} e^{-x^2/2}$ y la familia beta simétrica

$$K(x) = \frac{1}{Beta(0{,}5, \gamma + 1)} (1 - x^2)^{\gamma + 1} \qquad \gamma = 0{,}1{,}2{,}3, \ldots$$

Note que cuando se escoge γ = 0, 1, 2 y 3 genera las funciones Kernel uniforme (*Box*), de Epanechnikov, la bipeso y la tripeso, respectivamente.

La estimación Kernel es llamada la estimación de Nadaraya - Watson, en honor a sus creadores. Su simplicidad lo hace de fácil comprensión e implementación; no obstante, se debe tener en cuenta que los ajustes en los extremos son sesgados.

Estimador de Nadaraya-Watson

Dado un conjunto de *n* observaciones de *m* de la forma $(X_{1,}Y_1), \dots, (X_n, Y_n)$ el cual se denomina muestra, para Lorca, F (2006) el problema reside en estimar *m(x)* para un *x* no necesariamente incluida en la muestra. Para esto, se debe estimar la esperanza condicional:

$$E\{Y|X = x\} = m(x)$$

Sean *g(x)* la densidad marginal de *X*, f*(x,y)* la densidad conjunta y $\varphi(x)=\int yf(x,y)dy$ entonces:

$$E\{Y|X = x\} = \int yf(x|y)dy$$

$$E\{Y|X = x\} = \int y\frac{f(x,y)}{g(x)}dy$$

$$E\{Y|X = x\} = \frac{1}{g(x)}\int yf(x,y)dy$$

$$E\{Y|X = x\} = \frac{\varphi(x)}{g(x)}$$

La idea del estimador Nadaraya-Watson es justamente sustituir *g(x)* y *φ(x)* por sus respectivas estimaciones, como ocurre mas adelante cuando se sustituyan en función de la estimación por kernel.

Sea (*x*,*y*) un punto a estimar de función empírica (***X,Y***) = {(x_1,y_1), (x_2,y_2), …, (x_n,y_x)} = {(x_i,y_i)} con *i=1,2,…,n*. Utilizando la regresión kernel y el estimador de Nadaraya-Watson, se obtiene un estimador:

$$\widehat{m} = \frac{\sum_{i=1}^{n} w_i\, y_i}{\sum_{i=1}^{n} w_i}$$

donde $w_i = \frac{1}{h}K(u_i)$ $\quad y \quad$ $u_i = \frac{x_i - x}{h}$, lo que lleva a

$$\widehat{m} = \frac{\sum_{i=1}^{n} \frac{1}{h} K(u_i)\, y_i}{\sum_{i=1}^{n} \frac{1}{h} K(u_i)}$$

$$= \frac{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)}$$

que es un estimador natural de la esperanza condicional. Se puede observar que este estimador es un promedio "pesado" (local) de la variable respuesta y_i. Por otro lado, se debe determinar el valor *h* pues este determina el grado de suavidad del estimador.

Polinomios Locales

El estimador de Nadaraya -Watson puede ser visto como un caso especial de una clase amplia de estimadores de regresión basado en núcleos. Este estimador corresponde a ajustar localmente por mínimos cuadrados una constante a las respuestas. Para motivar ajustes localmente lineales y ajustes localmente polinomiales de mayor orden, se considera la expansión de Taylor de la función *m*, que se supondrá es lo suficientemente suave para que el desarrollo sea válido, alrededor de *x*

$$m(t) \approx m(x) + m'(x)(t - x) + \cdots + m^{(p)}(x)(t - x)^{(p)} \frac{1}{p!}$$

para todo punto *t* en un entorno del punto *x*. Esto sugiere una regresión polinomial local, que busca ajustar un polinomio de grado *p* en un entorno de *x*, incluyendo pesos basados en núcleos en el problema de minimización para penalizar observaciones correspondientes a valores de las covariables alejadas

de *x* y dar mayor preponderancia a las respuestas *yi* correspondientes a valores de *xi* cercanos a *x*. Así, el estimador local de grado 1 es

$$\begin{bmatrix}\hat{\alpha}(x)\\ \hat{\beta}(x)\end{bmatrix} = \left[\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) z_i\, z_i'\right]^{-1} . \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) z_i\, y_i$$

donde

$$z_i = \begin{bmatrix}1\\ x_i - x\end{bmatrix} \quad y \quad z_i z_i' = \begin{bmatrix}1 & x_i - x\\ x_i - x & (x_i - x)^2\end{bmatrix}$$

y la función estimadora local es

$$\widehat{y_i} = \hat{\alpha}(x) + \hat{\beta}(x).(x_i - x)$$

Es importante notar que, en contraste con la versión paramétrica de mínimos cuadrados, este estimador varía con *x*. Por lo tanto es una regresión local en un punto *x*. Cuando en la estimación de un punto, si sus vecinos no están tan cercanos y además se escoge un h muy pequeño puede ocurrir que solo existe un solo punto vecino o lo que es lo mismo *p=0*. Así, el estimador de Nadaraya-Watson es:

$$\widehat{y_i} = \frac{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)}$$

En otro caso, no existirá estimación en el entorno de *x*, si no hay al menos un punto vecino cercano, en este caso la función no estimará el valor y_i.

Coeficiente de Determinación

Una vez ajustada la recta de regresión a la nube de observaciones es importante disponer de un valor que mida la bondad del ajuste realizado y que permita decidir si el ajuste lineal es suficiente o se deben buscar modelos

alternativos. Como medida de bondad del ajuste se utiliza el coeficiente de determinación, definido como sigue

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Como la suma de cuadrados del error es menor o igual a la suma de cuadrados del modelo es decir $scE \leq scG$, se verifica que $0 \leq R^2 \leq 1$.

El coeficiente de determinación mide la proporción de variabilidad total de la variable (Y) respecto a su media que es explicada por el modelo de regresión. Es usual expresar esta medida en tanto por ciento, multiplicándola por cien.

Por otra parte, teniendo en cuenta que $\hat{y}_i - \bar{y} = \beta_1(x_i - \bar{x})$, se obtiene

$$R^2 = \frac{s_{XY}^2}{s_X^2 s_Y^2}$$

Dadas dos variables aleatorias cualesquiera X e Y, una medida de la relación lineal que hay entre ambas variables es el coeficiente de correlación definido por:

$$\rho = \frac{Cov(X,Y)}{\sigma(X)\,\sigma(Y)}$$

donde $\sigma(X)$ representa la desviación típica de la variable X (análogamente para $\sigma(Y)$). Un buen estimador de este parámetro es el coeficiente de correlación lineal muestral (o coeficiente de correlación de Pearson), definido por

$$r = \frac{S_{XY}}{S_X S_Y} = signo(\hat{\sigma}_1)\sqrt{R^2}$$

Por tanto, $r \in [-1,1]$. Este coeficiente es una buena medida de la bondad del ajuste de la recta de regresión.

MARCO METODOLÓGICO

Para evaluar el comportamiento de la regresión kernel en la estimación del modelo de regresión lineal simple discontinuo se procede de la siguiente manera:

Se utilizó como función de estimación kernel a

$$\widehat{f_n}(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

donde *K* representa la función kernel a utilizar. Se consideraron 8 funciones kernel; a saber, Uniforme, Epanechikov, Triangular, Gaussiana, Bipeso, Tripeso, Coseno y la Cuartica, pues son todos los que contiene el paquete estadístico utilizado.

La metodología de la estimación se hizo a través del estimador de Nadaraya-Watson y polinomios locales de grado 1.

Se consideraron 3 tomas de muestras; a saber, *n=20*, *n=50* y *n=100*.

El ancho de banda *h* para una primera etapa de investigación son

$$h_1 = \frac{\Delta x}{2(1 + log_2 n)}$$

$$h_2 = 1{,}06.min\left(\hat{\sigma}, \frac{R}{1{,}34}\right) n^{-1/5}$$

$$h_3 = 1{,}44.\hat{\sigma}.n^{-1/5}$$

para la segunda etapa de la investigación se elige un ancho de banda de tal forma que el valor será la mayor distancia que existe entre dos puntos

consecutivos de la variable *x* ya que si se toma la menor distancia pueda ocurrir que un punto *x* dentro de la muestra quede sin vecinos en su entorno.

Utilizando el generador de números aleatorios del software Excel 2007 se construyeron varios conjuntos de datos (x_i, y_i) para los modelos de regresión lineal discontinua, para los cuales se generaron 3 escenarios a continuación. Tales escenarios se generan en vista a la gran variedad de situaciones que se generan en economía, agronomía, medicina entre otras; por lo tanto se tomaron estos escenarios en función de verificar situaciones que son pertinentes en función de la investigación.

Escenario 1

El mismo trata de un modelo de regresión lineal discontinuo, definido en el intervalo (0,40) con un punto de discontinuidad en $x^{\circ} = 20$, siendo la longitud del salto de 4 unidades. Este escenario contempla un modelo de regresión lineal simple discontinuo en el cual se conserva la misma pendiente en la recta de regresión antes y después del salto como se observa en la figura 1.

Figura 1. Gráfico de escenario 1

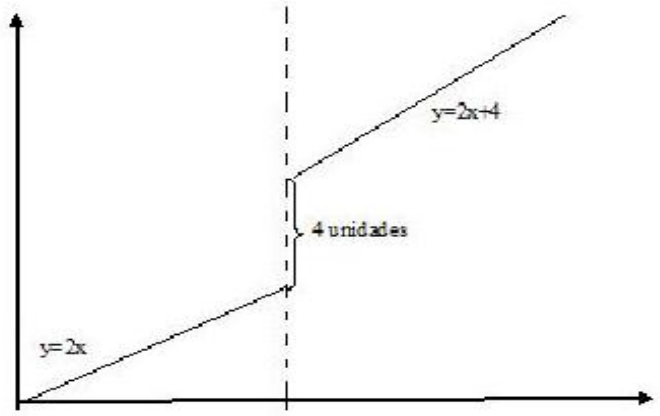
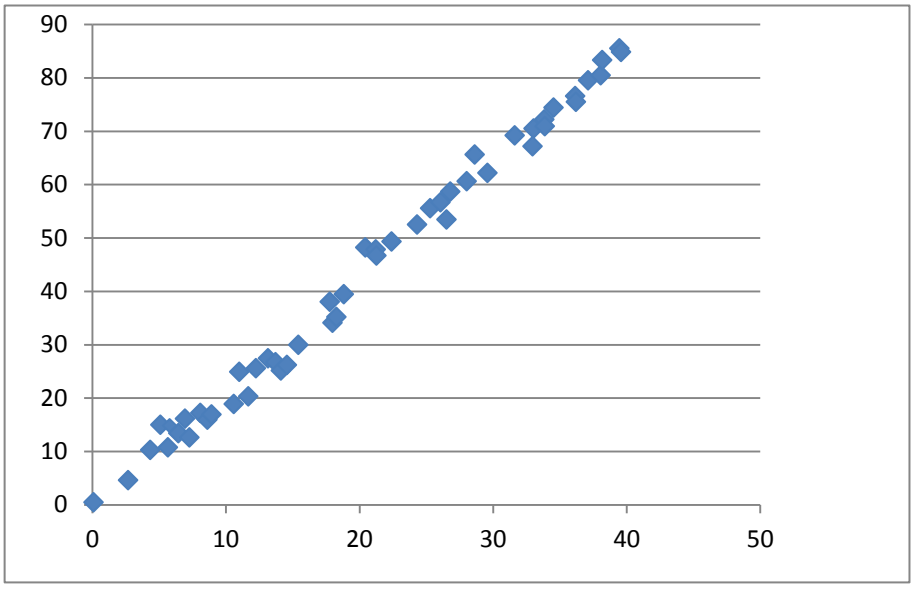


Figura 2. Gráfico de puntos generados por el Excel 2007 para el escenario 1 con n=50 y s=2



para el intervalo (0,20), el modelo viene dado por la curva *y=2x* y en el intervalo (20,40) por la ecuación *y=2x+4*.

La nube de puntos sobre el cual se construye la regresión kernel, se obtuvo adecuando a las ecuaciones anteriores un efecto aleatorio o error aleatorio, el cual sigue la distribución normal. Así, se generaron varios conjuntos de datos con las siguientes características

n=20; $\widehat{\sigma_1} = 2$; $\varepsilon_1 \sim (0; 4)$

n=20; $\widehat{\sigma_2} = 4$; $\varepsilon_2 \sim (0; 16)$

n=50; $\widehat{\sigma_1} = 2$; $\varepsilon_1 \sim (0; 4)$

n=50; $\widehat{\sigma_2} = 4$; $\varepsilon_2 \sim (0; 16)$

n=100; $\widehat{\sigma_1} = 2$; $\varepsilon_1 \sim (0; 4)$

n=100; $\widehat{\sigma_2} = 4$; $\varepsilon_2 \sim (0; 16)$

Escenario 2

El siguiente escenario es un modelo de regresión lineal discontinuo, definido en el intervalo (0,40) con un punto de discontinuidad en $x^{\circ} = 20$, siendo la longitud del salto de 4 unidades pero la pendiente de la ecuación después del salto sufre una ligera inclinación como se observa en la figura 2.

Figura 3. Gráfico del escenario 2

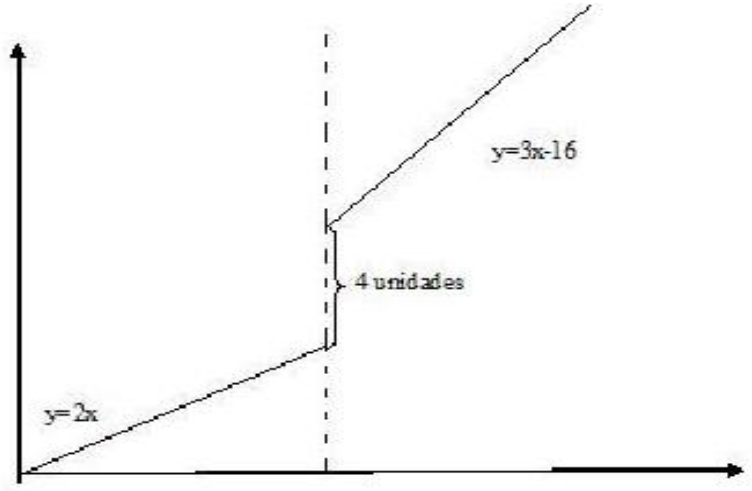
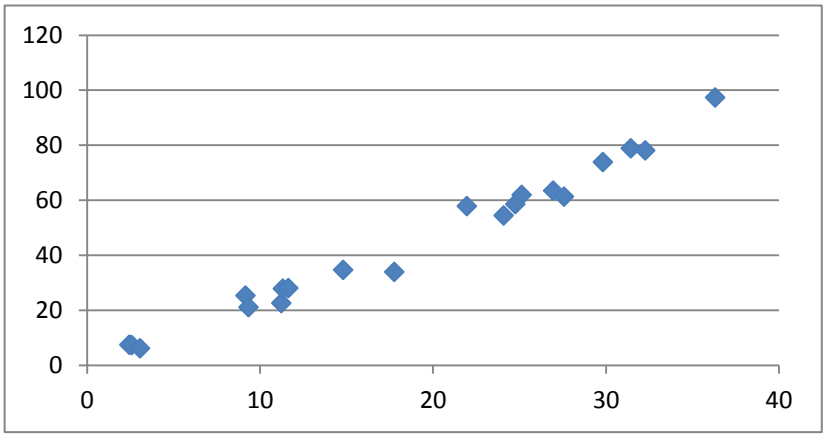


Figura 4. Gráfico de puntos generados por el Excel 2007 para el escenario 2 con n=20 y s=4



para el intervalo (0,20), el modelo viene dado por la curva *y=2x* y en el intervalo (20,40) por la ecuación *y=3x-16*.

Al igual que en el escenario anterior la nube de puntos sobre el cual construye la regresión kernel, se generó adecuando a las ecuaciones anteriores un efecto aleatorio o error aleatorio el cual sigue la distribución normal. Se tomaran las mismas características del escenario 1.

Escenario 3

El siguiente escenario es un modelo de regresión lineal discontinuo, definido en el intervalo (0,40) con un punto de discontinuidad en $x^{\circ} = 20$, siendo la longitud del salto de 4 unidades pero la pendiente de la ecuación después del salto cambia de pendiente y es opuesta a la ecuación antes del salto como se observa en la figura 3.

Figura 5. Gráfico de escenario 3

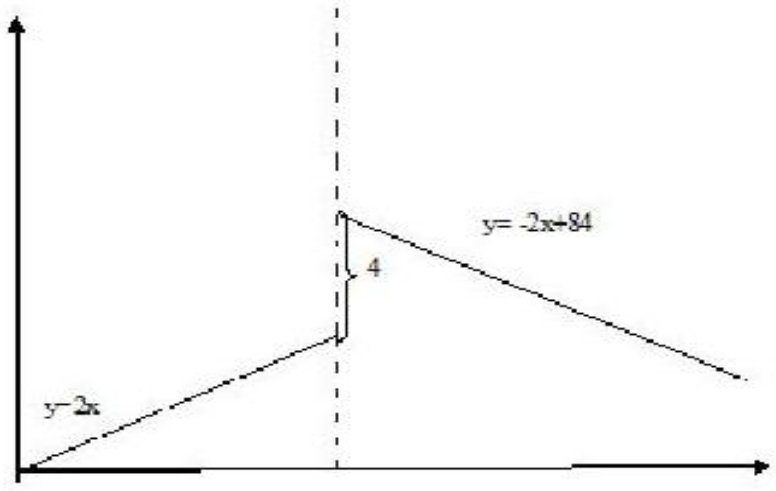
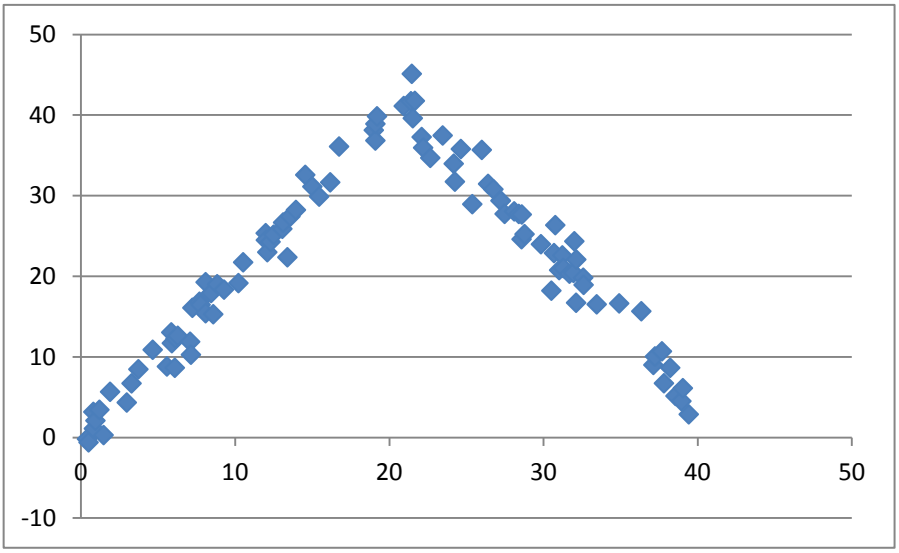


Figura 6. Gráfico de puntos generados por el Excel 2007 para el escenario 3 con n=100 y s=2



para el intervalo (0,20), el modelo viene dado por la curva *y=2x* y en el intervalo (20,40) por la ecuación *y=-2x+84*.

Al igual que en el escenarios anteriores la nube de puntos sobre el cual se construye la regresión kernel, se generó adecuando a las ecuaciones anteriores un efecto aleatorio o error aleatorio el cual sigue la distribución normal. Se tomara las mismas 6 características del escenario 1.

Sobre la nube de puntos generados se calcularon las estimaciones de los puntos $\widehat{y_i}$ utilizando las 8 regresiones kernel descrita anteriormente. Así mismo se calcularon las regresiones clásicas a las ecuaciones del antes y después del punto de discontinuidad, en ambos casos se determinó el coeficiente de ajuste R^2.

Como medida de comparación se utiliza la diferencia $\Delta_r = R_{rk}^2 - R_{rc}^2$ donde R_{rk}^2 es el coeficiente de ajuste de la regresión kernel y R_{rc}^2 es el coeficiente de ajuste de la regresión clásica ponderada de las ecuaciones antes y después. Todos aquellos $\Delta_r \epsilon (-5{,}5\%, 5{,}5\%)$, se dice que es un ajuste igual

entre la regresión kernel y la regresión ponderada clásica. Si está por encima de 5,5%, se dice que es un ajuste mejor el kernel que la ponderada y si es menor a -5,5% se dice que el ajuste del kernel no es el mejor respecto al ponderado. Esta medida se toma en función a los casos de ajuste que deben tener las regresiones en medicina, microbiología entre otras, las cuales deben de ser muy pequeña su variación y un ajuste alto. Por otro lado tenemos los casos de economía los cuales las tasas de ajuste y variación pueden ser altas por su naturaleza de comportamiento o como casos de agronomía que su tasa de ajuste y variación son intermedias. Debido a esto, se decide tomar un intervalo -5,5% y 5,5%, el cual es un ajuste intermedio entre los más minuciosos como los casos en medicina y entre los menos rigurosos como son los casos de economía.

La estimación de las regresiones kernel y regresiones clásicas se realizan en el *software xlstat* complemento del Excel de Windows.

DISCUSIÓN DE RESULTADOS

A continuación se mostrará el desarrollo manual que se realiza para calcular un $\hat{y}_i$ particular a través del estimador Nadaraya-Watson:

x_j	y_j	x_i	h
2,429	3,316	17,753	1,6358
2,564	3,779	17,753	1,6358
3,048	7,234	17,753	1,6358
9,147	19,132	17,753	1,6358
9,322	17,450	17,753	1,6358
11,219	20,130	17,753	1,6358
11,308	24,852	17,753	1,6358
11,625	21,549	17,753	1,6358
14,797	30,767	17,753	1,6358
17,753	35,664	17,753	1,6358
21,950	49,017	17,753	1,6358
24,077	52,542	17,753	1,6358
24,751	53,719	17,753	1,6358
25,121	51,875	17,753	1,6358
26,942	63,537	17,753	1,6358
27,567	62,734	17,753	1,6358
29,818	63,603	17,753	1,6358
31,432	66,325	17,753	1,6358
32,267	68,034	17,753	1,6358
36,310	75,125	17,753	1,6358

Para este ejemplo en particular, se tomo el escenario 1, n=20, s=2, el x_j es el valor generado aleatoriamente en el programa excel, y_j es el obtenido aplicada las funciones del escenario determinado, el x_i es el valor al cual se le va estimar su respectivo $\hat{y}_i$ (en cual está en negrilla en la tabla, diferenciado respecto a los demás) y h es el ancho de banda h_1.

Esta es la continuación de los datos anteriores

Uj	*K(uj)*	*K(uj)*Yj*	*K(uj)*1*
9,368238	3,4932E-20	1,158E-19	3,493E-20
9,285403	7,5641E-20	2,858E-19	7,564E-20
8,989882	1,1259E-18	8,145E-18	1,126E-18
5,261169	3,8931E-07	7,448E-06	3,893E-07
5,15408	6,7998E-07	1,187E-05	6,8E-07
3,994757	0,00013666	0,0027511	0,0001367
3,940279	0,00016964	0,0042159	0,0001696
3,746624	0,00035709	0,0076947	0,0003571
1,807081	0,07794833	2,398232	0,0779483
0	0	0	0
-2,56529	0,01485699	0,7282391	0,014857
-3,86603	0,00022667	0,0119097	0,0002267
-4,27796	4,2355E-05	0,0022752	4,235E-05
-4,50408	1,5693E-05	0,0008141	1,569E-05
-5,61714	5,616E-08	3,568E-06	5,616E-08
-5,9996	6,0906E-09	3,821E-07	6,091E-09
-7,37534	6,1525E-13	3,913E-11	6,153E-13
-8,3619	2,6162E-16	1,735E-14	2,616E-16
-8,87272	3,206E-18	2,181E-16	3,206E-18
-11,3443	4,5218E-29	3,397E-27	4,522E-29
	Σ=0,09375456	Σ=3,156155	Σ=0,0937546

Recordemos que $u_j = \frac{x_j - X_i}{h}$, en este ejemplo se utilizó el kernel Gaussiano, recordemos el estimador Nadaraya-Watson:

$$\begin{bmatrix}\hat{\alpha}(x)\\ \hat{\beta}(x)\end{bmatrix} = \left[\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) z_i\, z_i'\right]^{-1} .\ \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) z_i\, y_i$$

$$z_i = \begin{bmatrix}1\\ x_i - x\end{bmatrix} \quad y \quad z_i z_i' = \begin{bmatrix}1 & x_i - x\\ x_i - x & (x_i - x)^2\end{bmatrix}$$

$$\begin{bmatrix}\hat{\alpha}(x)\\ \hat{\beta}(x)\end{bmatrix} = A^{-1} . B$$

K(uj)(Xj-Xi)*	*K(uj)*(Xj-Xi)^2*	*K(uj)*yj*	*K(uj)*yj(Xj-xi)*
-5,3532E-19	8,2035E-18	1,2E-19	-1,77494E-18
-1,1489E-18	1,7451E-17	2,9E-19	-4,34131E-18
-1,6557E-17	2,4349E-16	8,1E-18	-1,19775E-16
-3,3505E-06	2,8835E-05	7,4E-06	-6,41004E-05
-5,7329E-06	4,8334E-05	1,2E-05	-0,00010004
-0,00089305	0,00583576	0,00275	-0,017977161
-0,00109341	0,00704757	0,00422	-0,027173424
-0,00218848	0,01341262	0,00769	-0,047158782
-0,23041704	0,68111804	2,39823	-7,089228632
0	0	0	0
0,0623443	0,26161509	0,72824	3,055906011
0,00143347	0,00906532	0,01191	0,075317472
0,00029639	0,00207414	0,00228	0,015921906
0,00011562	0,00085186	0,00081	0,005997762
5,1603E-07	4,7416E-06	3,6E-06	3,27872E-05
5,9774E-08	5,8663E-07	3,8E-07	3,74985E-06
7,4228E-12	8,9553E-11	3,9E-11	4,72111E-10
3,5786E-15	4,8949E-14	1,7E-14	2,37351E-13
4,6532E-17	6,7537E-16	2,2E-16	3,16578E-15
8,3911E-28	1,5572E-26	3,4E-27	6,30384E-26
Σ=-0,17041071	Σ=0,9811029	Σ=3,15616	Σ=-4,028522452

$$A = \begin{vmatrix} 0{,}09375456 & -0{,}17041071 \\ -0{,}17041071 & 0{,}9811029 \end{vmatrix} \quad B = \begin{vmatrix} 3{,}156155 \\ -4{,}02852245 \end{vmatrix}$$

$$\text{Det}(A) = 0{,}06294306$$

$$A^{-1} = \begin{vmatrix} 15{,}5871507 & 2{,}70737905 \\ 2{,}70737905 & 1{,}48951388 \end{vmatrix}$$

$$A^{-1} * B = \begin{vmatrix} 38{,}2887263 \\ 2{,}54436781 \end{vmatrix}$$

Por consiguiente

$$\widehat{y_i} = \hat{\alpha}(x) + \hat{\beta}(x).(x_i - x)$$

que será para este ejemplo

$$\hat{y}_i = 38{,}28 + 2{,}54(x_i - x)$$

Aplicado todo el procedimiento metodológico se generaron las siguientes cuadros comparativos:

CUADRO 1. R^2 de los distintos kernels con los respectivos h primera fase y donde (*) significa que la regresión kernel estimó todos los $\hat{y}_i$ en el escenario 1.

		s=2			*s=4*		
Kernel	*h*	*n=20*	*n=50*	*n=100*	*n=20*	*n=50*	*n=100*
Coseno	*1*	0,9904	0,9501	0,9861	0,9517	0,7870	0,8443
	2	0,9529	0,9521	0,9872	0,9462	0,9376	0,9637(*)
	3	0,9525	0,9249	0,9754	0,9536	0,9682(*)	0,9719(*)
Epanechnikov	*1*	0,9903	0,9501	0,9861	0,9516	0,7872	0,8446
	2	0,9529	0,9522	0,9872	0,9461	0,9380	0,9638(*)
	3	0,9525	0,925	0,9755	0,9535	0,9683(*)	0,9720(*)
Gaussiano	*1*	0,9907(*)	0,9914(*)	0,9913(*)	0,9632(*)	0,9678(*)	0,9615(*)
	2	0,9924(*)	0,9911(*)	0,9930(*)	0,9442(*)	0,9697(*)	0,9725(*)
	3	0,9913(*)	0,9919(*)	0,9932(*)	0,9624(*)	0,9718(*)	0,9729(*)
Cuartico	*1*	0,9909	0,9504	0,9859	0,9527	0,7844	0,8420
	2	0,9529	0,9519	0,9869	0,9474	0,9334	0,9628(*)
	3	0,9525	0,9245	0,9744	0,9541	0,9663(*)	0,9712(*)
Triangular	*1*	0,9906	0,9501	0,9860	0,9520	0,7868	0,8437
	2	0,9529	0,9520	0,9871	0,9463	0,9371	0,9633(*)
	3	0,9525	0,9249	0,9753	0,9539	0,9691(*)	0,9716(*)
Tricubo	*1*	0,9855	0,9520	0,9858	0,9413	0,7857	0,8429
	2	0,9529	0,9518	0,9837	0,9476	0,9283	0,9637(*)
	3	0,9525	0,9254	0,9668	0,9535	0,9660(*)	0,9717(*)
Tripeso	*1*	0,9909	0,9503	0,9854	0,9533	0,7837	0,8382
	2	0,9529	0,9516	0,9862	0,9486	0,9237	0,9620(*)
	3	0,9525	0,9240	0,9717	0,9547	0,9617(*)	0,9702(*)
Uniforme	*1*	0,9863	0,9503	0,9866	0,945	0,7897	0,8470
	2	0,9529	0,9524	0,9874	0,9455	0,9423	0,9646(*)
	3	0,9525	0,9251	0,9717	0,9526	0,9693(*)	0,9719(*)

CUADRO 2. R^2 de los distintos kernels con los respectivos h primera fase y donde (*) significa que la regresión kernel estimó todos los $\hat{y}_i$ en el escenario 2.

		s=2			*s=4*		
Kernel	*h*	*n=20*	*n=50*	*n=100*	*n=20*	*n=50*	*n=100*
Coseno	*1*	0,9933	0,9680	0,9907	0,9668	0,8614	0,896
	2	0,9666	0,9693	0.9915	0,9645	0,9594	0,9757(*)
	3	0,9658	0,9520	0,9836	0,9681	0,9793(*)	0,9812(*)
Epanechnikov	*1*	0,9932	0,9680	0,9907	0,9667	0,8616	0,8962
	2	0,9666	0,9694	0,9916	0,9645	0,9597	0,9758(*)
	3	0,9658	0,9520	0,9837	0,9680	0,9794(*)	0,9813(*)
Gaussiano	*1*	0,9931(*)	0,9946(*)	0,9942(*)	0,9744(*)	0,9791(*)	0,9743(*)
	2	0,9944(*)	0,9944(*)	0,9953(*)	0,9615(*)	0,9803(*)	0,9816(*)
	3	0,9937(*)	0,9949(*)	0,9955(*)	0,9739(*)	0,9816(*)	0,9818(*)
Cuartico	*1*	0,9936	0,9682	0,9906	0,9675	0,8597	0,8944
	2	0,9666	0,9692	0,9913	0,9654	0,9567	0,9752(*)
	3	0,9658	0,9518	0,983	0,9684	0,9781(*)	0,9807(*)
Triangular	*1*	0,9934	0,9680	0,9906	0,967	0,8613	0,8956
	2	0,9666	0,9693	0,9915	0,9646	0,9591	0,9755(*)
	3	0,9658	0,9520	0,9835	0,9683	0,9792(*)	0,9810(*)
Tricubo	*1*	0,9898	0,9680	0,9905	0,9596	0,8605	0,8950
	2	0,9666	0,9691	0,9892	0,9655	0,9534	0,9756(*)
	3	0,9658	0,9523	0,9779	0,9680	0,9779(*)	0,9811(*)
Tripeso	*1*	0,9936	0,9681	0,9902	0,9679	0,8593	0,8919
	2	0,9666	0,9690	0,9909	0,9662	0,9504	0,9746(*)
	3	0,9658	0,9514	0,9811	0,9689	0,9751(*)	0,982(*)
Uniforme	*1*	0,9904	0,9673	0,9910	0,9621	0,8631	0,8978
	2	0,9666	0,9695	0,9917	0,9640	0,9625	0,9754(*)
	3	0,9658	0,9521	0,9839	0,9674	0,9801(*)	0,9812(*)

CUADRO 3. R^2 de los distintos kernels con los respectivos h primera fase y donde (*) significa que la regresión kernel estimó todos los $\hat{y}_i$ en el escenario 3.

		s=2			*s=4*		
Kernel	*h*	*n=20*	*n=50*	*n=100*	*n=20*	*n=50*	*n=100*
Coseno	*1*	0,9584	0,7422	0,9316	0,7933	0	0,2882
	2	0,7869	0,751	0,9332	0,6843	0,6758	0,8338(*)
	3	0,7823	0,6104	0,8799	0,8015	0,8341(*)	0,8721(*)
Epanechnikov	*1*	0,9578	0,742	0,9318	0,7927	0	0,2896
	2	0,7869	0,7512	0,9334	0,6839	0,6781	0,8345(*)
	3	0,7823	0,6105	0,8802	0,8009	0,8348(*)	0,8724(*)
Gaussiano	*1*	0,9394(*)	0,9531(*)	0,9579(*)	0,8192(*)	0,8325(*)	0,8239(*)
	2	0,9660(*)	0,9517(*)	0,9657(*)	0,7491(*)	0,8405(*)	0,8738(*)
	3	0,9553(*)	0,9550(*)	0,9666(*)	0,8174(*)	0,8474(*)	0,8749(*)
Cuartico	*1*	0,9606	0,7436	0,9307	0,7977	0	0,2777
	2	0,7869	0,7497	0,9316	0,6894	0,6540	0,8299(*)
	3	0,7823	0,6082	0,8751	0,8035	0,8247(*)	0,8692(*)
Triangular	*1*	0,9592	0,742	0,9312	0,7944	0	0,2858
	2	0,7869	0,7506	0,9328	0,6846	0,6731	0,8322(*)
	3	0,7823	0,61	0,8794	0,8025	0,8337(*)	0,8708(*)
Tricubo	*1*	0,9372	0,7424	0,9304	0,7486	0	0,2819
	2	0,7869	0,7493	0,9164	0,6901	0,6282	0,8338(*)
	3	0,7823	0,6127	0,8377	0,8010	0,8229(*)	0,8715(*)
Tripeso	*1*	0,9607	0,7429	0,9280	0,8003	0	0,2604
	2	0,7869	0,7484	0,9286	0,6945	0,6039	0,8262(*)
	3	0,7823	0,6056	0,8617	0,8063	0,8009(*)	0,8646(*)
Uniforme	*1*	0,9404	0,7363	0,9340	0,7644	0	0,3009
	2	0,7869	0,7522	0,9343	0,6811	0,6097	0,8382(*)
	3	0,7823	0,6110	0,8817	0,7971	0,8398(*)	0,8710(*)

Se observa de los tres cuadros anteriores, se tienen 144 opciones para cada escenario generado, solo hubo 39 donde se estimaron todos los $\hat{y}_i$ denotado con (*) y 105 que dejaron de estimar aunque sea un $\hat{y}_i$, hubo casos donde no fue la mejor estimación y no se obtuvo un R^2, tales caso se presentaron en el escenario 3. Obsérvese que en los tres escenarios se tiene un comportamiento igual por lo cual se puede concluir que los *h* propuestos por la literatura no fueron los mejores.

	R^2 de la Regresion Lineal Simple del antes y después del punto de dicontinuiddad por Escenarios, promedio de ambos R^2		
Escenario	Indicador en punto de corte	R^2	Promedio
E1 s=2 n=20	Antes	0,98088	~~0,953835~~
	Despues	0,92679	
E1 s=2 n=50	Antes	0,95783	~~0,96917~~
	Despues	0,98051	
E1 s=2 n=100	Antes	0,97818	~~0,9718~~
	Despues	0,96542	
E1 s=4 n=20	Antes	0,93672	~~0,883675~~
	Despues	0,83063	
E1 s=4 n=50	Antes	0,8448	~~0,893475~~
	Despues	0,94215	
E1 s=4 n=100	Antes	0,90777	~~0,87887~~
	Despues	0,84997	
E2 s=2 n=20	Antes	0,98088	~~0,974335~~
	Despues	0,96779	
E2 s=2 n=50	Antes	0,95783	~~0,974515~~
	Despues	0,9912	
E2 s=2 n=100	Antes	0,97818	~~0,9811~~
	Despues	0,98402	
E2 s=4 n=20	Antes	0,93672	~~0,9282~~
	Despues	0,91968	
E2 s=4 n=50	Antes	0,8448	~~0,9085~~
	Despues	0,9722	
E2 s=4 n=100	Antes	0,90777	~~0,91802~~
	Despues	0,92827	
E3 s=2 n=20	Antes	0,98088	~~0,962855~~
	Despues	0,94483	
E3 s=2 n=50	Antes	0,95783	~~0,968935~~
	Despues	0,98004	
E3 s=2 n=100	Antes	0,97818	~~0,96967~~
	Despues	0,96116	
E3 s=4 n=20	Antes	0,93672	~~0,897595~~
	Despues	0,85847	
E3 s=4 n=50	Antes	0,8448	~~0,88422~~
	Despues	0,92364	
E3 s=4 n=100	Antes	0,90777	~~0,88439~~
	Despues	0,86101	

En el cuadro anterior se mostraron los resultados de las regresiones lineales simples de manera clásica y donde se promediaran las R^2 de los modelos para después comparar con la estimación kernel y su R^2.

CUADRO 4. Comportamiento de los distintos kernels con los respectivos h de la primera fase en comparación con el modelo clásico de regresión en el escenario 1 y donde se estimaron todos los $\hat{y}_i$.

		s=2			*s=4*		
Kernel	*h*	*n=20*	*n=50*	*n=100*	*n=20*	*n=50*	*n=100*
Coseno	*1*						
comparación	*2*						Superior
Regresión	*3*					Superior	Superior
Epanechnikov	*1*						
comparación	*2*						Superior
Regresión	*3*					Superior	Superior
Gaussiano	*1*	Igual	Igual	Igual	Superior	Superior	Superior
comparación	*2*	Igual	Igual	Igual	Superior	Superior	Superior
Regresión	*3*	Igual	Igual	Igual	Superior	Superior	Superior
Cuartico	*1*						
comparación	*2*						Superior
Regresión	*3*					Superior	Superior
Triangular	*1*						
comparación	*2*						Superior
Regresión	*3*					Superior	Superior
Tricubo	*1*						
comparación	*2*						Superior
Regresión	*3*					Superior	Superior
Tripeso	*1*						
comparación	*2*						Superior
Regresión	*3*					Superior	Superior
Uniforme	*1*						
comparación	*2*						Superior
Regresión	*3*					Superior	Superior

CUADRO 5. Comportamiento de los distintos kernels con los respectivos h en comparación con el modelo clásico de regresión en el escenario 2 y donde se estimaron todos los $\hat{y}_{i.}$

		s=2			*s=4*		
Kernel	*h*	*n=20*	*n=50*	*n=100*	*n=20*	*n=50*	*n=100*
Coseno	*1*						
comparación	*2*						Superior
Regresión	*3*					Superior	Superior
Epanechnikov	*1*						
comparación	*2*						Superior
Regresión	*3*					Superior	Superior
Gaussiano	*1*	Igual	Igual	Igual	Igual	Superior	Superior
comparación	*2*	Igual	Igual	Igual	Igual	Superior	Superior
Regresión	*3*	Igual	Igual	Igual	Igual	Superior	Superior
Cuartico	*1*						
comparación	*2*						Superior
Regresión	*3*					Superior	Superior
Triangular	*1*						
comparación	*2*						Superior
Regresión	*3*					Superior	Superior
Tricubo	*1*						
comparación	*2*						Superior
Regresión	*3*					Superior	Superior
Tripeso	*1*						
comparación	*2*						Superior
Regresión	*3*					Superior	Superior
Uniforme	*1*						
comparación	*2*						Superior
Regresión	*3*					Superior	Superior

CUADRO 6. Comportamiento de los distintos kernels con los respectivos h en comparación con el modelo clásico de regresión en el escenario 3 y donde se estimaron todos los $\hat{y}_i$.

		s=2			*s=4*		
Kernel	*h*	*n=20*	*n=50*	*n=100*	*n=20*	*n=50*	*n=100*
Coseno	*1*						
comparación	*2*						Igual
Regresión	*3*					Igual	Igual
Epanechnikov	*1*						
comparación	*2*						Igual
Regresión	*3*					Igual	Igual
Gaussiano	*1*	Igual	Igual	Igual	Inferior	Igual	Inferior
comparación	*2*	Igual	Igual	Igual	Inferior	Igual	Igual
Regresión	*3*	Igual	Igual	Igual	Inferior	Igual	Igual
Cuartico	*1*						
comparación	*2*						Igual
Regresión	*3*					Inferior	Igual
Triangular	*1*						
comparación	*2*						Igual
Regresión	*3*					Igual	Igual
Tricubo	*1*						
comparación	*2*						Igual
Regresión	*3*					Inferior	Igual
Tripeso	*1*						
comparación	*2*						Igual
Regresión	*3*					Inferior	Igual
Uniforme	*1*						
comparación	*2*						Igual
Regresión	*3*					Igual	Igual

La comparación entre los distintos kernels y la regresión lineal simple se define como Superior, Igual e Inferior lo que significa que: "será Superior cuando la diferencia del kernel – Regresión Lineal Clásica promediada con respecto al R^2 está por encima del 5,5% de diferencia", "Igual cuando la diferencia del kernel – Regresión Lineal Clásica promediada con respecto al R^2

este entre $\pm$5,5%" y "será Inferior cuando la diferencia del kernel – Regresión Lineal Clásica promediada con respecto al R^2 esté por debajo del -5,5%". Se observa en la cualidad Superior del Kernel se tiene 57 casos, en la cualidad Igual 53 casos, en la cualidad inferior 7 casos de los 117 casos posibles.

Como se observa en los cuadros 4, 5, 6 fueron las comparaciones de los kernels que estimaron todos los $\hat{y}_i$ con respecto al modelo clásico de regresión lineal, casi el 50% de los kernel se comportan mejor que el modelo clásico, el 45% se comporta igual que el modelo clásico y solo el 6% está por debajo del comportamiento del modelo clásico.

El comportamiento de los *h* seleccionado de la literatura tienen debilidades a la hora de la estimación de todos los $\hat{y}_i$, se debe recordar que el *h* es la amplitud del entorno de un x_i a la hora de estimar, por lo cual si el ancho de banda es muy pequeño puede ocurrir que en ese entorno el único punto que se encuentre sea el mismo, generando esto que no exista la estimación del punto $\hat{y}_i$, por lo cual hay que garantizar que el ancho de banda por lo menos tome un punto distinto al x_i que se está tomando como punto de estimación.

Ahora, se presentará la estimación de los kernels con el *h* propuesto para la segunda fase:

CUADRO 7. Comportamiento de los R^2 de los distintos kernels con la selección de *h* segunda fase propuesta en comparación con el modelo clásico de regresión en el escenario 1

	s=2			*s=4*		
h	*6,01*	*2,6*	*1,5*	*6,01*	*2,6*	*1,5*
Kernel	n=20	n=50	n=100	n=20	n=50	n=100
Coseno	0,99	0,992	0,992	0,906	0,968	0,963
comparación	Igual	Igual	Igual	Igual	Superior	Superior
Regresión						
Epanechnikov	0,99	0,992	0,992	0,906	0,968	0,963
comparación	Igual	Igual	Igual	Igual	Superior	Superior
Regresión						
Gaussiano	0,991	0,993	0,993	0,97	0,972	0,972
comparación	Igual	Igual	Igual	Superior	Superior	Superior
Regresión						
Cuartico	0,99	0,991	0,992	0,905	0,965	0,961
comparación	Igual	Igual	Igual	Igual	Superior	Superior
Regresión						
Triangular	0,99	0,992	0,992	0,906	0,968	0,962
comparación	Igual	Igual	Igual	Igual	Superior	Superior
Regresión						
Tricubo	0,989	0,991	0,992	0,912	0,965	0,962
comparación	Igual	Igual	Igual	Igual	Superior	Superior
Regresión						
Tripeso	0,991	0,99	0,991	0,905	0,957	0,96
comparación	Igual	Igual	Igual	Igual	Superior	Superior
Regresión						
Uniforme	0,99	0,992	0,992	0,905	0,969	0,965
comparación	Igual	Igual	Igual	Igual	Superior	Superior
Regresión						

CUADRO 8. Comportamiento de los R^2 de los distintos kernels con la selección de *h* de la segunda fase propuesta en comparación con el modelo clásico de regresión en el escenario 2

	s=2			*s=4*		
h	*6,01*	*2,6*	*1,5*	*6,01*	*2,6*	*1,5*
Kernel	n=20	n=50	n=100	n=20	n=50	n=100
Coseno	0,993	0,995	0,995	0,935	0,979	0,975
comparación	Igual	Igual	Igual	Igual	Superior	Superior
Regresión						
Epanechnikov	0,993	0,995	0,995	0,935	0,979	0,975
comparación	Igual	Igual	Igual	Igual	Superior	Superior
Regresión						
Gaussiano	0,993	0,996	0,996	0,979	0,981	0,98
comparación	Igual	Igual	Igual	Igual	Superior	Superior
Regresión						
Cuartico	0,993	0,994	0,994	0,935	0,977	0,974
comparación	Igual	Igual	Igual	Igual	Superior	Superior
Regresión						
Triangular	0,993	0,995	0,995	0,935	0,979	0,975
comparación	Igual	Igual	Igual	Igual	Superior	Superior
Regresión						
Tricubo	0,992	0,994	0,995	0,939	0,978	0,974
comparación	Igual	Igual	Igual	Igual	Superior	Superior
Regresión						
Triweight	0,993	0,994	0,994	0,934	0,972	0,973
comparación	Igual	Igual	Igual	Igual	Superior	Superior
Regresión						
Uniforme	0,993	0,995	0,995	0,935	0,98	0,976
comparación	Igual	Igual	Igual	Igual	Superior	Superior
Regresión						

CUADRO 9. Comportamiento de los R^2 de los distintos kernels con la selección de *h* de la segunda fase propuesta en comparación con el modelo clásico de regresión en el escenario 3

	s=2			*s=4*		
h	*6,01*	*2,6*	*1,5*	*6,01*	*2,6*	*1,5*
Kernel	n=20	n=50	n=100	n=20	n=50	n=100
Coseno	0,948	0,955	0,961	0,583	0,832	0,83
Comparación	Igual	Igual	Igual	Inferior	Igual	Igual
Regresión						
Epanechnikov	0,947	0,956	0,961	0,583	0,833	0,83
comparación	Igual	Igual	Igual	Inferior	Igual	Igual
Regresión						
Gaussiano	0,871	0,952	0,967	0,748	0,847	0,873
comparación	Inferior	Igual	Igual	Inferior	Igual	Igual
Regresión						
Quartico	0,949	0,951	0,959	0,581	0,816	0,823
comparación	Igual	Igual	Igual	Inferior	Igual	Inferior
Regresión						
Triangular	0,949	0,954	0,96	0,583	0,832	0,823
comparación	Igual	Igual	Igual	Inferior	Igual	Inferior
Regresión						
Tricubo	0,947	0,951	0,961	0,632	0,82	0,825
comparación	Igual	Igual	Igual	Inferior	Inferior	Inferior
Regresión						
Triweight	0,951	0,946	0,955	0,578	0,776	0,817
comparación	Igual	Igual	Igual	Inferior	Inferior	Inferior
Regresión						
Uniforme	0,946	0,953	0,963	0,576	0,839	0,838
comparación	Igual	Igual	Igual	Inferior	Igual	Igual
Regresión						

Se aplicará las mismas cualidades que se utilizaron en la primera fase las cuales son, Superior, Igual e Inferior. Como se observa en los cuadros 7, 8, 9 fueron las comparaciones de los kernels que se les fijó el *h* propuesto con

respecto al modelo clásico de regresión lineal, casi el 25% de los kernel se comportan mucho mejor que el modelo clásico, el 67% se comporta igual que el modelo clásico y solo el 15% está por debajo del comportamiento del modelo clásico.

Véase que en los primeros 2 escenarios el kernel tiene un comportamiento bueno con respecto al modelo clásico y mucho más cuando hay un mayor número de observaciones y en el escenario 3 es donde el comportamiento del modelo clásico inferior en algunos casos que la estimación kernel representando ese 15% de comportamiento inferior.

Como se observó en los cuadros presentados anteriormente, la estimación kernel para modelos de regresión lineal discontinuo tiene un comportamiento mejor o igual que la del modelo clásico del antes y después, aparte se debe recordar que el kernel hace el ajuste de una sola curva para los datos realizando un estudio único para todos los datos que se presentan. A continuación se colocará un ejemplo del gráfico ajustado en los 3 escenarios donde se escogieron al azar de los estimados algunos para que se observe el ajuste que realiza la estimación kernel.

En las gráficas a continuación se mostraran tres ejemplos en cada escenario estudiado donde se presenta los puntos sin estimar y los puntos ya suavizados con el kernel

Figura 7. Gráfica de estimación en el escenario 1, *s=2, n=50, h* propuesto con kernel de Epanechnikov

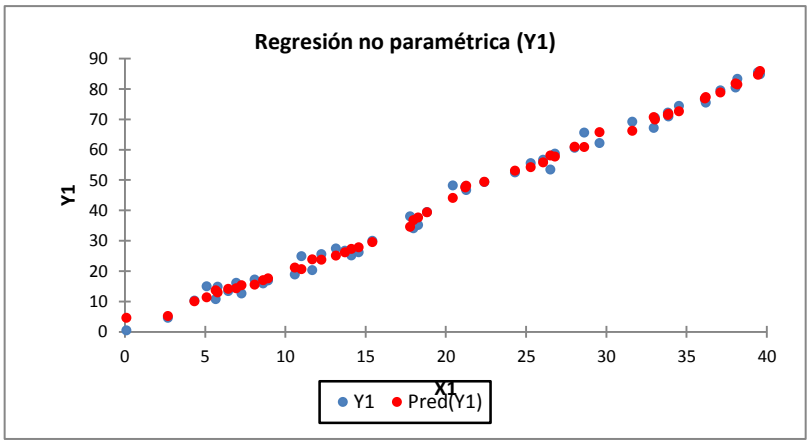


Figura 8. Gráfica de estimación en el escenario 2, *s=2, n=20, h* propuesto con kernel de Coseno

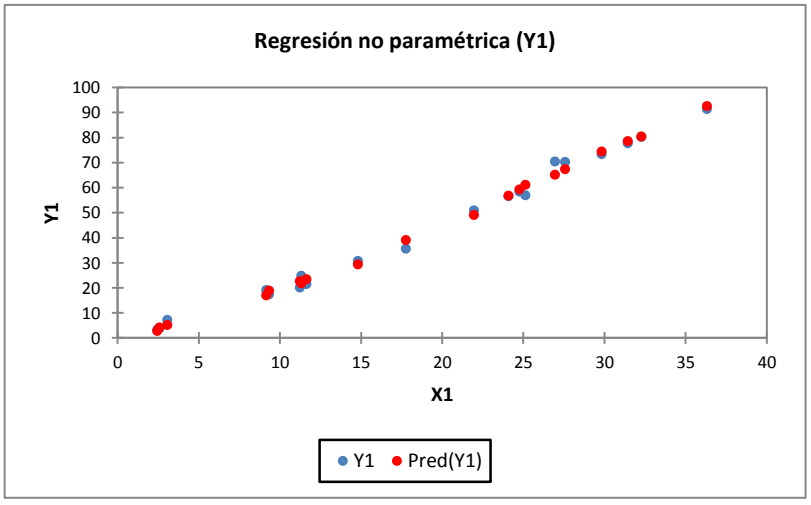
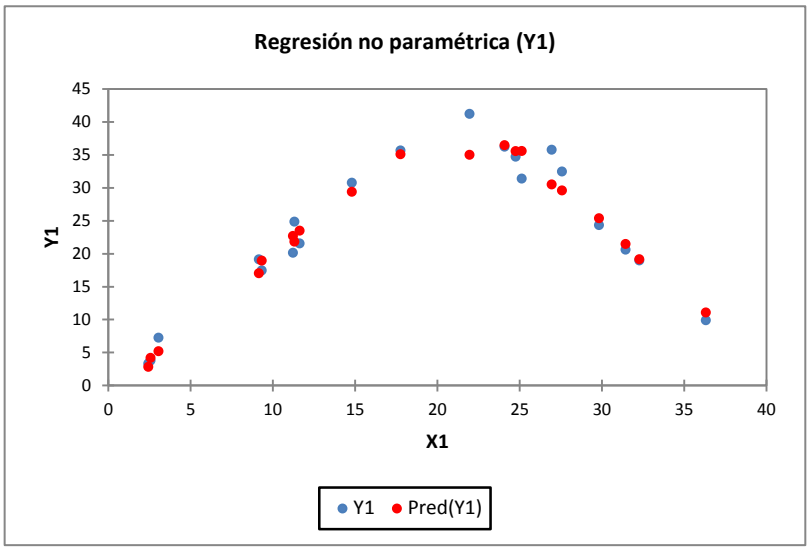


Figura 9. Gráfica de estimación en el escenario 3, *s=2, n=20, h* propuesto con kernel de Coseno



Los puntos azules son los datos sin estimar y los puntos rojos son los suavizados por la regresión kernel.

Se propone que al finalizar con la suavización de los puntos a través del kernel se realice una regresión lineal y medir su grado de ajuste a través del coeficiente de determinación. Se presenta en el ejemplo siguiente, donde se escogió el conjunto de datos del escenario 1 *n=20, s=2*:

X	Y	$\hat{y}_i$
2,42866298	4,85732597	3,024
2,56416517	5,12833033	4,240
3,04757836	6,09515671	4,839
9,14700766	18,2940153	17,094
9,3221839	18,6443678	19,353
11,2186041	22,4372082	22,623
11,3077181	22,6154363	21,133
11,6245003	23,2490005	23,399
14,7972045	29,594409	29,351
17,7532273	35,5064547	38,289
21,9495224	47,8990448	46,689
24,0772729	52,1545457	52,095
24,7511216	53,5022431	53,815
25,1210059	54,2420118	55,920
26,9417402	57,8834803	59,166
27,5673696	59,1347392	61,391
29,8178045	63,635609	65,064
31,4316233	66,8632466	66,608
32,2672201	68,5344401	67,746
36,3103122	76,6206244	76,054

donde los $\hat{y}_i$ son los suavizados por el kernel gaussiano y h_3. Aplicando la regresión lineal simple se obtiene:

Intervalo de confianza 95%

Modelizacion de la variable Y

Resumen de la variable dependiente

Variable	Núm. total de valores	Núm. de valores utilizados	Núm. de valores ignorados	Suma de los pesos	Media	Desviación típica
Y	19	19	0	19	41,309	22,791

Resumen de la Variable X

Variable	Media	Desviación típica
X	19,527	10,253

Coeficiente de ajuste

R (coeficiente de correlación)	0,998
R^2 (coeficiente de determinación)	0,996
R^2aj. (coeficiente de determinación ajustado)	0,996
SCR	33,966

Evaluación del valor de la información originado por las variables (H0 = Y=Moy(Y)):

Fuente	GDL	Suma los cuadrados	Cuadrado medio	F de Fisher	Pr > F
Modelo	1	9315,551	9315,551	4662,405	< 0,0001
Residuos	17	33,966	1,998		
Total	18	9349,517			

Parámetros del modelo

Parámetro	Valor	Desviación típica	t de Student	Pr > t	Límite inferior 95 %	Límite superior 95 %
Intersección	-2,019	0,713	-2,833	0,011	-3,522	-0,515
β	2,219	0,032	68,282	< 0,0001	2,150	2,287

La ecuación del modelo se escribe: $Y = -2,019 + 2,219*x_i$

Véase que el ajuste es del 99% lo que es un dato que nos permite decir que se puede realizar una buena estimación, aparte de poder estimar los parámetros a un grupo de datos con discontinuidad, en este caso que tengan un comportamiento parecido al escenario.

CONCLUSIONES Y RECOMENDACIONES

Conclusiones

Con la presentación de la siguiente investigación se pretende dar una vía alterna para poder estimar una función de regresión lineal simple discontinua que se ajuste a una respuesta certera, cercana a la realidad y poder sobreponerse a una discontinuidad de los datos sin tener que hacer un estudio para un antes-después del salto.

La estimación de los parámetros del modelo de regresión no se puede realizar como tal, por que el estimador varía con *x* entonces la regresión es local en un punto *x*, como lo afirma Lorca (2006). Obsérvese cuando se realiza la estimación de cada punto por polinomios locales, sí se realiza una estimación de parámetros individual para cada punto a estimar, pero como es único para cada punto, no se puede estimar un único parámetro para toda la regresión lineal simple discontinua. No obstante, se propone realizar una regresión lineal con los datos suavizados con el kernel hallándose una función de regresión estimada.

Dictaminar una única estimación kernel sería inadecuado, pues cada uno de los 8 kernels estudiado, tienen comportamientos iguales, pues su ajuste (R^2) y gráfica mantienen ajustes iguales. Obsérvese que existen algunos que se comportan levemente mejor dependiendo del escenario, pero en conclusión podría decirse que queda a elección del investigador cual desea utilizar.

La evaluación del ajuste de la estimación kernel arrojó que tiene un comportamiento mejor o igual que el modelo clásico; además se debe recordar que la estimación kernel permite hacer una única estimación ajustando los datos evitando la discontinuidad, haciendo que no se pierda información, ya

que a fin de cuentas es una sola variable en estudio con un comportamiento discontinuo en un punto.

La utilización de los anchos de banda que la literatura nos ofrece para la aplicación del kernel no son los mejores a la hora de la estimación de una regresión simple discontinua, por lo tanto se genero un nuevo ancho de banda en función de la mayor distancia entre dos puntos consecutivos para poder asegurar que todas las estimaciones $\hat{y}_i$ existiesen.

Concretando, para los casos considerados hay evidencia que la estimación kernel es un método que evita la discontinuidad de los datos en una regresión lineal simple, realizando una estimación y ajuste único de los datos, permitiendo evitar perdida de información de la variable, generando un estudio completo de la variable. Determinar si es mejor o no con respecto del modelo clásico, seria infructuoso, pues cada investigador le dará explicación para cada método clásico o no paramétrico, pero lo que arroja la investigación es una vía nueva y diferente que permite evitar la discontinuidad con un buen ajuste.

Recomendaciones

Ampliar el estudio del ancho de banda, pues en la literatura se encontró diversos ancho de banda ajustados a distintos comportamientos particulares pero no con respecto a la discontinuidad, lo que generaba que los h seleccionados al aplicarse al estudio no estimaba todos los puntos lo que generaba perdida de información y falta de ajuste del modelo.

A la elección del kernel sea a disposición del investigador y de la naturaleza de los datos, pues la diferencia entre uno u otro kernel no es significativa.

Abrir una investigación a la estimación de una función de regresión lineal discontinua en general.

REFERENCIAS BIBLIOGRÁFICAS

- ACUÑA, EDGAR (2007) *Análisis de Regresión.* Universidad de Puerto Rico.
- BALZA, MIGUEL (1980) *Estimación kernel de una función de densidad de probabilidad*. Universidad Central de Venezuela. Facultad de Ciencias, Caracas, Venezuela.

- BERK, R.A., y RAUMA, D. (1983) *Capitalizing on Nonrandom Assign-ment to Treatments: A Regression Discontinuity Evaluation of a Crime Control Program*. Journal of the American Statistical Association 78(381): 21-27, 1983.

- BOCHNER, S (1955) *Harmonic analysis and the theory of probability*. Press, Bercekley University of California.

- BORUCH, ROBERT F (1973) *Solutions to a problem in quasi-experimental evaluation*. Proceedings of the Annual Convention of the American Psychological Association, 1973, 9-10.

- BORUCH, R y DEGRACIE, J. (1975) *Regression-discontinuity analysis of the impact of title I-supported Reading programs.* Evaluation Research Evanston: Psychology Departament, Northwesterm University.

- BOWMAN, A.W., y AZZALINI, A. (1997), *Applied Smoothing Techniques for Data Analysis*. London: Oxford University Press.

- BRADEN, J. P. y BRYANT, T. J. (1990). *Regression discontinuity designs: Applications for school psychology*. School Psychology Review. 19(2), 232-239.

- CAMPBELL, D (1969) *Reforms as experiment*. American Psychologist. Northwestern University. 24, 409-429.

- CAMPBELL, D. T., STANLEY, J. C. (1963) *Experimental and cuasiexperimental design for research*. Chicago:Rand McNally (traducido al castellano en 1973)

- CAO, R., CUEVAS, A. y GONZALEZ, W.G. (1994). .*A comparative study of severalsmoothing methods in density estimation*. Comp. Statis. and Data Analysis 153-176.

- COOK, T. D. y CAMPBELL, D. T. (1979). *Quasi-experimention: Design and Analysis Issues for Field Settings*. Chicago, IL: Rand McNally.

- FAN, J. y GIJBELS, I. (1996). *Adaptive order polynomial fitting: Bandwidth robustification and bias reduction*. J. Comput. Graph. Statist. 4, 213-227.

- GOLBERGER, A.S. (1972) *Maximun likelihood estimation of regressions containing un observable independent variables*. International Economc Review. Vol. 13, pág. 1-15.

- HÄRDLE, W., (2004). *Nonparametric and Semiparametric Models*. Springer.

- IMBEMS, G, KALYANARAMAN, K (2009) *Optimal Bandwidth choice for the Regression Discontinuity Estimator*. Working Paper 14726. National Bureau Of Economic Research.

- JONES, M. y SHEATHER, S. J. (1991). *Progress in data based bandwidth selection for kernel density estimation*. Department of Statistics, University of North Carolina. Mimeo Series # 2088.

- JUDD, C.M. y KENNY, D.A. (1981). *Estimating the Effects of Social Interventions*. Cambridge University Press, Cambridge, MA.

- KIDDER, L. H. (1981). *Qualitative research and quasi-experimental frameworks. In M.B. Brewer & B.E. Collins (Eds.)* Scientific inquiry and the social sciences: A volume in honour of Donald T. Campbell (pp. 226-256). San Francisco: Jossey-Bass, pp.226-256.

- LEE, HOWARD (2008) *Using the Chow Test to Analyze Regression Discontinuities.* Departamento de Psicologia, Universidad de California. Vol 4, p.46-50.

- LORCA, F. (2006) Regresión no parametrica. Introducción al estimador de Nadaraya-Watson. Paper

- ROSENBLATT, M. (1956): Remarks on sorne nonparametric estimates of a density function. Annals of Mathematical Statistics 27, pp. 832-837.

- RUBIN, B (1977) *Assignment to treatment group on the basis of a covariate*. Journal of Educational Statistics. Vol. 2, No. 1., pp.1-26.

- SACKS, J y YLVISAKER, D (1976). *The best linear estimate of a regression function at a point.* (Personal communication)

- SILVERMAN, B.W. (1986): *Density estimation for statistics and data analysis*, Londres, Chapman and Hall.

- THISTLETHWAITE, D.L. y CAMPBELL, D. T. (1960) *Regression-discontinuity analysis: An alternative to the ex post facto experiment*. Journal of Educational Psychology, 51, 309-317.

- TROCHIM, W (1980) *The statiscal analysis of data from de regression-discontinuity design*. Versions paper

ANEXOS

Anexo 1. Estimación de un $\hat{y}_i$ del escenario 1 con el kernel Epanechikov con el estimador Nadaraya-Watson y polinomios locales de grado 1.

En este ejemplo, se tomo el escenario 1, n=20, s=2, el x_j es el valor generado aleatoriamente en el programa 59xcel, y_j es el obtenido aplicada las funciones del escenario determinado, el x_i es el valor al cual se le va estimar su respectivo $\hat{y}_i$ (en cual está en negrilla en la tabla, diferenciado respecto a los demás) y h es el ancho de banda en este caso fue h_1.

Xj	*Yj*	*Xi*	*h*
2,429	3,316	31,432	1,6358
2,564	3,779	31,432	1,6358
3,048	7,234	31,432	1,6358
9,147	19,132	31,432	1,6358
9,322	17,450	31,432	1,6358
11,219	20,130	31,432	1,6358
11,308	24,852	31,432	1,6358
11,625	21,549	31,432	1,6358
14,797	30,767	31,432	1,6358
17,753	35,664	31,432	1,6358
21,950	49,017	31,432	1,6358
24,077	52,542	31,432	1,6358
24,751	53,719	31,432	1,6358
25,121	51,875	31,432	1,6358
26,942	63,537	31,432	1,6358
27,567	62,734	31,432	1,6358
29,818	63,603	31,432	1,6358
31,432	66,325	31,432	1,6358
32,267	68,034	31,432	1,6358
36,310	75,125	31,432	1,6358

Continuación de la tabla anterior

Uj	*Uj*	*K(uj)*	*K(uj)*Yj*	*K(uj)*1*
17,7301	0	0	0	0
17,6473	0	0	0	0
17,3517	0	0	0	0
13,6230	0	0	0	0
13,5159	0	0	0	0
12,3566	0	0	0	0
12,3021	0	0	0	0
12,1085	0	0	0	0
10,1689	0	0	0	0
8,3618	0	0	0	0
5,7966	0	0	0	0
4,4958	0	0	0	0
4,0839	0	0	0	0
3,8578	0	0	0	0
2,7447	0	0	0	0
2,3623	0	0	0	0
0,9865	0,9865	0,0200	1,2733	0,0200
0	0	0	0	0
-0,5108	-0,5108	0,5542	37,7113	0,5542
-2,9824	0	0	0	0
		Σ=0,5743	Σ=38,9847	Σ=0,5743

Recordemos que $u_j = \frac{x_j - x_i}{h}$, en este ejemplo se utilizó el kernel Gaussiano, recordemos el estimador Nadaraya-Watson:

$$\begin{bmatrix} \hat{\alpha}(x) \\ \hat{\beta}(x) \end{bmatrix} = \left[\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) z_i \, z_i' \right]^{-1} . \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) z_i \, y_i$$

$$z_i = \begin{bmatrix} 1 \\ x_i - x \end{bmatrix} \quad y \quad z_i z_i' = \begin{bmatrix} 1 & x_i - x \\ x_i - x & (x_i - x)^2 \end{bmatrix}$$

$$\begin{bmatrix} \hat{\alpha}(x) \\ \hat{\beta}(x) \end{bmatrix} = A^{-1} . B$$

K(uj)(Xj-Xi)*	*K(uj)*(Xj-Xi)^2*	*K(uj)*yj*	*K(uj)*yj(Xj-xi)*
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
-0,03231021	0,05214283	1,27339239	-2,05502455
0	0	0	0
0,46316996	0,38702333	37,7113639	31,5114946
0	0	0	0
Σ=0,43085975	Σ=0,43916616	Σ=38,9847563	Σ=29,4564701

$$A=\begin{vmatrix} 0,57431936 & 0,43085975 \\ 0,43085975 & 0,43916616 \end{vmatrix} \qquad B=\begin{vmatrix} 38,9847563 \\ 29,4564701 \end{vmatrix}$$

Det(A)= 0,06658151

$$A^{-1}=\begin{vmatrix} 6,59591808 & -6,47116256 \\ -6,47116256 & 8,62580915 \end{vmatrix}$$

$$A^{-1}*B=\begin{vmatrix} 66,5226527 \\ 1,80919354 \end{vmatrix}$$

y_i= 66,5226527

Anexo 2. Estimación Kernel Triangular en el escenario 3, *n=20, s=2* donde se dejo de estimó algun $\hat{y}_i$, salida del XlStat 2012.

XLSTAT 2012.1.01 - Regresión no paramétrica - el 19/02/2012 a 20:16:10

Método: Polinomio

Grado del polinomio: 1

Muestra de aprendizaje: Todo

k vecinos más próximos: % = 100

Kernel: Triángulo

Ancho de banda: Constante (3,1832)

Estadísticas descriptivas:

Variable	Obs.	Datos perdidos	Datos Sin perder	Mínimo	Máximo	Media	Desviación Típica
Y1	20	0	20	3,316	41,219	23,472	11,343
X1	20	0	20	2,429	36,310	18,672	10,687

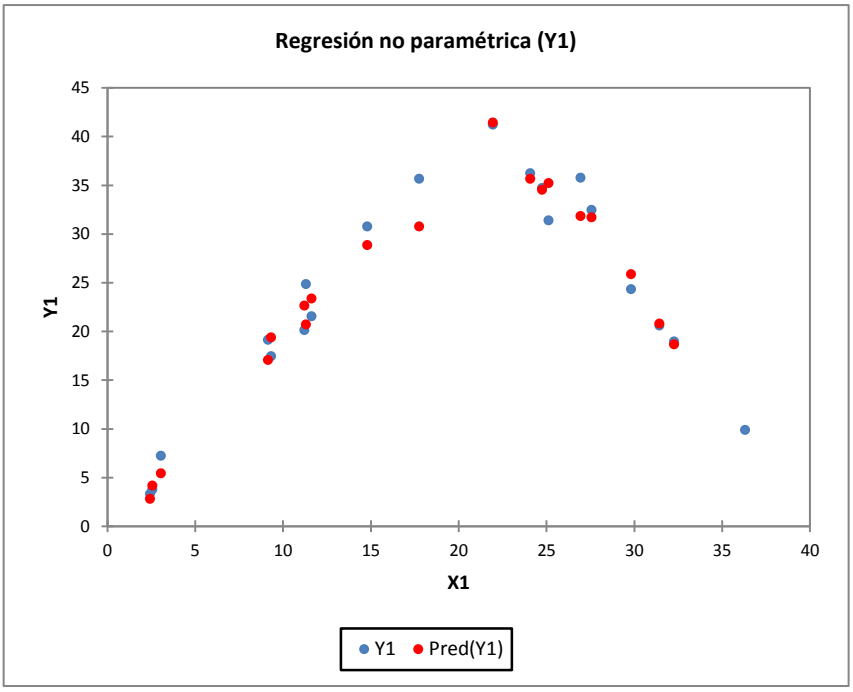
Regresión no paramétrica de la variable Y1:

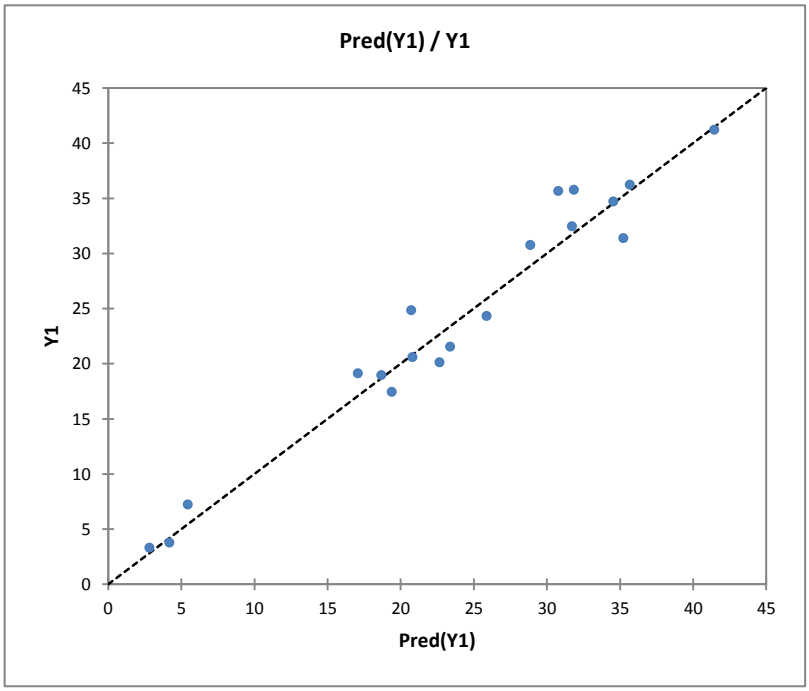
Coeficientes de ajuste:

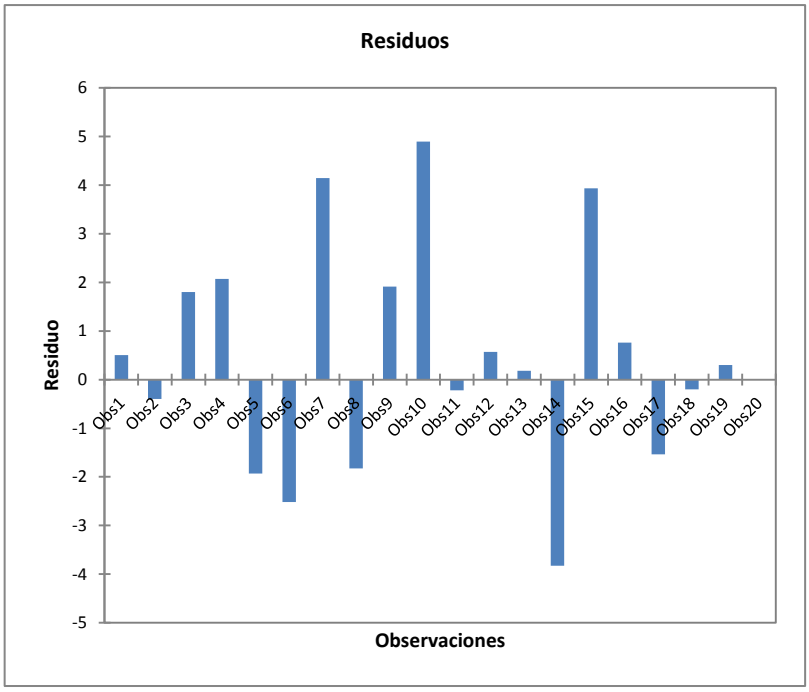
R^2	0,959
SEC	99,726
MEC	4,986
RMEC	2,233

Predicciones y residuos:

Observaciones	X1	Y1	Pred(Y1)	Residuos
Obs1	2,429	3,316	2,810	0,506
Obs2	2,564	3,779	4,174	-0,395
Obs3	3,048	7,234	5,430	1,804
Obs4	9,147	19,132	17,063	2,069
Obs5	9,322	17,450	19,382	-1,932
Obs6	11,219	20,130	22,645	-2,515
Obs7	11,308	24,852	20,709	4,143
Obs8	11,625	21,549	23,372	-1,823
Obs9	14,797	30,767	28,856	1,911
Obs10	17,753	35,664	30,767	4,897
Obs11	21,950	41,219	41,435	-0,216
Obs12	24,077	36,233	35,661	0,572
Obs13	24,751	34,714	34,533	0,181
Obs14	25,121	31,391	35,219	-3,828
Obs15	26,942	35,770	31,836	3,935
Obs16	27,567	32,465	31,705	0,760
Obs17	29,818	24,332	25,867	-1,535
Obs18	31,432	20,599	20,796	-0,197
Obs19	32,267	18,966	18,666	0,300
Obs20	36,310	9,884		


Regresión no paramétrica (Y1)
Y1
X1
Y1
Pred(Y1)


Pred(Y1) / Y1
Y1
Pred(Y1)


Residuos
Residuo
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
Obs1
Obs2
Obs3
Obs4
Obs5
Obs6
Obs7
Obs8
Obs9
Obs10
Obs11
Obs12
Obs13
Obs14
Obs15
Obs16
Obs17
Obs18
Obs19
Obs20
Observaciones

Anexo 3. Estimación Kernel Tripeso en el escenario 2, *n=50, s=4* donde se estimó todos los $\hat{y}_i$, salida del XlStat 2012.

XLSTAT 2012.1.01 - Regresión no paramétrica - el 19/02/2012 a 19:50:20

Método: Polinomio

Grado del polinomio: 1

Muestra de aprendizaje: Todo

k vecinos más próximos: % = 100

Kernel: Triweight

Ancho de banda: Constante (2,6944)

Estadísticas descriptivas:

Variable	Obs.	Datos perdidos	Datos Sin perder	Mínimo	Máximo	Media	Desviación Típica
Y1	50	0	50	-0,184	105,477	47,612	32,734
X1	50	0	50	0,084	39,572	20,478	11,666

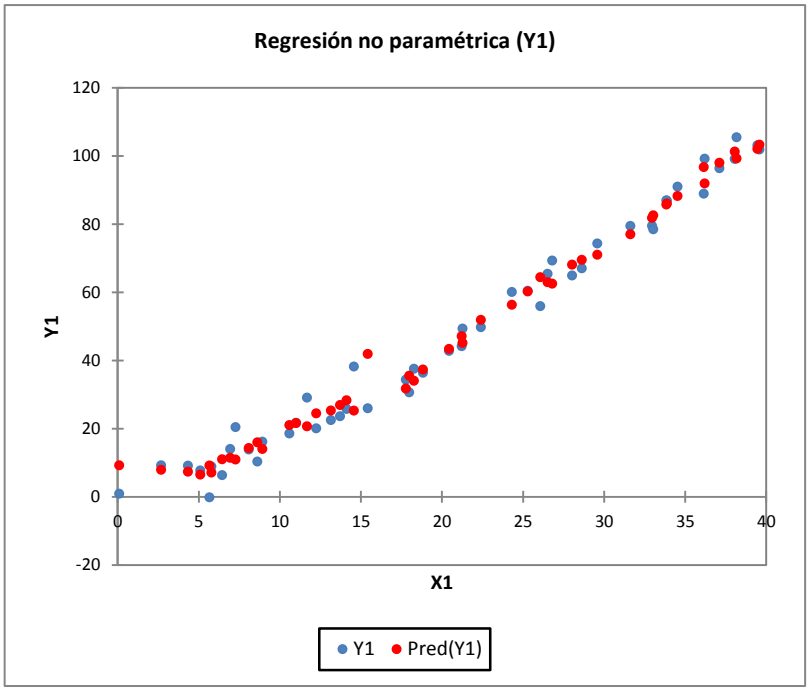
Regresión no paramétrica de la variable Y1:

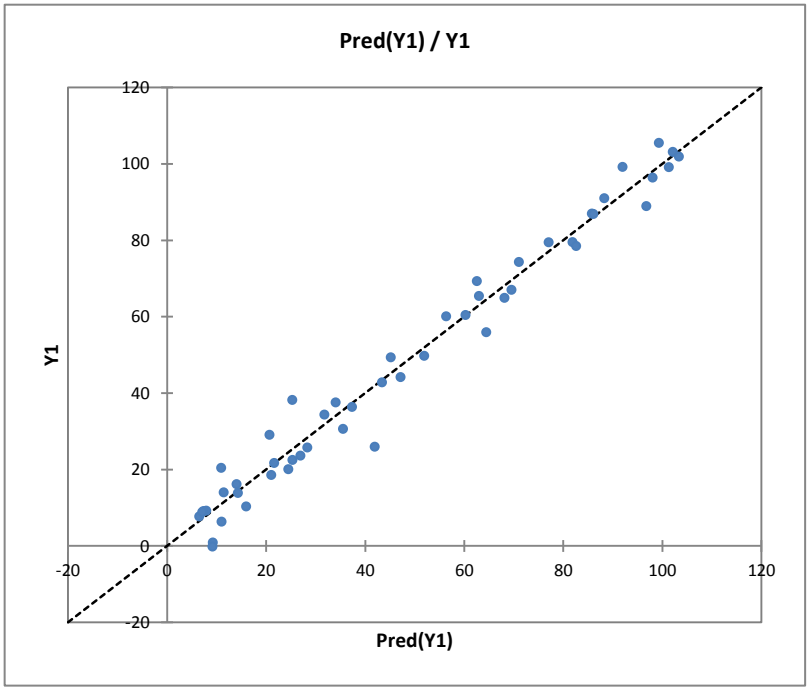
Coeficientes de ajuste:

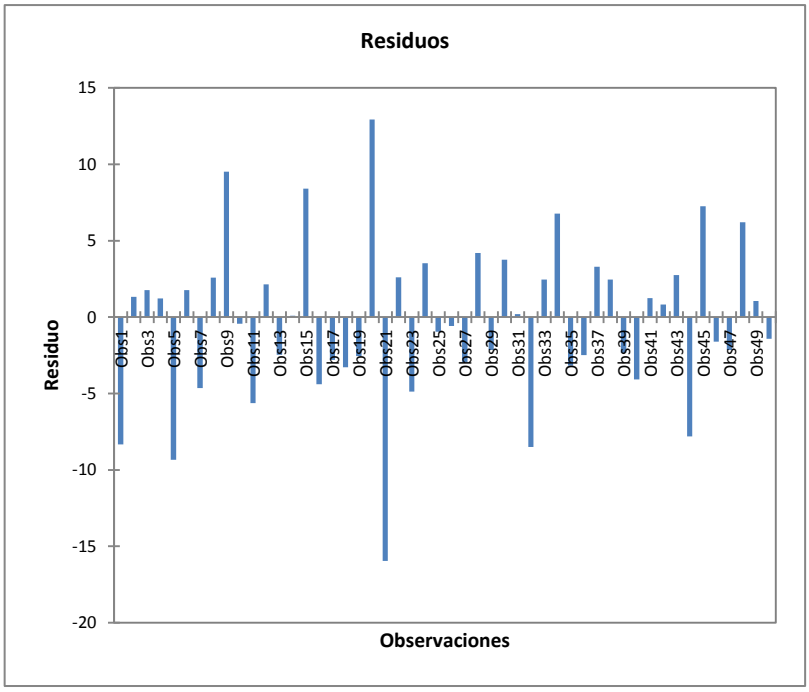
R^2	0,975
SEC	1305,701
MEC	26,114
RMEC	5,110

Predicciones y residuos:

Observaciones	X1	Y1	Pred(Y1)	Residuos
Obs1	0,084	0,874	9,212	-8,338
Obs2	2,670	9,212	7,894	1,319
Obs3	4,325	9,114	7,342	1,771
Obs4	5,086	7,699	6,479	1,220
Obs5	5,650	-0,184	9,147	-9,332
Obs6	5,772	8,855	7,087	1,768
Obs7	6,431	6,332	10,979	-4,647
Obs8	6,930	13,998	11,410	2,588
Obs9	7,259	20,417	10,908	9,509
Obs10	8,077	13,848	14,285	-0,437
Obs11	8,604	10,312	15,947	-5,635
Obs12	8,914	16,150	14,012	2,138
Obs13	10,581	18,552	21,016	-2,464
Obs14	10,993	21,695	21,600	0,095
Obs15	11,666	29,069	20,656	8,413
Obs16	12,236	20,059	24,459	-4,400
Obs17	13,142	22,487	25,296	-2,809
Obs18	13,705	23,613	26,900	-3,286
Obs19	14,106	25,742	28,292	-2,551
Obs20	14,559	38,195	25,258	12,937
Obs21	15,413	25,949	41,897	-15,948
Obs22	17,765	34,349	31,740	2,610
Obs23	17,972	30,617	35,499	-4,882
Obs24	18,263	37,545	34,014	3,531
Obs25	18,819	36,365	37,317	-0,952
Obs26	20,427	42,795	43,367	-0,572
Obs27	21,209	44,168	47,113	-2,945
Obs28	21,260	49,339	45,138	4,201
Obs29	22,391	49,743	51,885	-2,142
Obs30	24,302	60,078	56,321	3,758
Obs31	25,282	60,422	60,226	0,196
Obs32	26,056	55,916	64,414	-8,498
Obs33	26,507	65,410	62,944	2,466
Obs34	26,790	69,298	62,517	6,781
Obs35	28,018	64,913	68,089	-3,176
Obs36	28,617	67,016	69,505	-2,490
Obs37	29,571	74,305	71,005	3,300
Obs38	31,613	79,457	77,010	2,448
Obs39	32,949	79,495	81,826	-2,332
Obs40	33,024	78,480	82,562	-4,082
Obs41	33,832	86,967	85,732	1,235
Obs42	33,865	86,861	86,033	0,828
Obs43	34,521	90,986	88,238	2,748
Obs44	36,138	88,922	96,725	-7,803
Obs45	36,194	99,182	91,924	7,259
Obs46	37,104	96,386	97,995	-1,609
Obs47	38,050	99,123	101,272	-2,148
Obs48	38,161	105,477	99,262	6,215
Obs49	39,447	103,117	102,068	1,049
Obs50	39,572	101,886	103,301	-1,415


Regresión no paramétrica (Y1)
120
100
80
60
40
20
0
-20
Y1
0
5
10
15
20
25
30
35
40
X1
Y1
Pred(Y1)


Pred(Y1) / Y1
120
100
80
60
40
20
0
-20
Y1
-20
0
20
40
60
80
100
120
Pred(Y1)


Residuos
15
10
5
0
-5
-10
-15
-20
Residuo
Obs1
Obs3
Obs5
Obs7
Obs9
Obs11
Obs13
Obs15
Obs17
Obs19
Obs21
Obs23
Obs25
Obs27
Obs29
Obs31
Obs33
Obs35
Obs37
Obs39
Obs41
Obs43
Obs45
Obs47
Obs49
Observaciones