Suavización no paramétrica para análisis de datos

Book · September 2002

CITATIONS
18

READS
5,372

1 author:

Isaias Hazarmabeth Salgado-Ugarte
Universidad Nacional Autónoma de México
95 PUBLICATIONS **1,035** CITATIONS

SEE PROFILE

FACULTAD DE ESTUDIOS
SUPERIORES ZARAGOZA

PAPIME 192031
PAPIIT IN217596

Suavización no Paramétrica para Análisis de Datos

Isaías Hazarmabeth
Salgado Ugarte

Métodos Estadísticos No Paramétricos de Suavización para Análisis de Datos Biológicos

Métodos Estadísticos No Paramétricos
de Suavización para Análisis de Datos Biológicos

ISBN

IMPRESO EN MÉXICO PRINTED IN MEXICO

Esta obra se terminó de imprimir en

Se tiraron

Métodos Estadísticos No Paramétricos de Suavización para Análisis de Datos Biológicos

Isaías Hazarmabeth Salgado Ugarte

FACULTAD DE ESTUDIOS SUPERIORES ZARAGOZA

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Esta obra se realizó con apoyo de la UNAM a través de la DGAPA, PAPIIT, clave IN217596

Capítulo 1. Histogramas y polígonos de frecuencia

1.1. Introducción

Existen varios procedimientos estadísticos para mostrar la distribución de un lote de datos univariado. Como ejemplos clásicos podemos citar a los diagramas univariados de dispersión, los diagramas de tallo y hoja, los diagramas de caja y los histogramas. Otros procedimientos son los varios tipos de gráficos de cuantiles los cuales utilizan la función empírica de distribución acumulada (por ejemplo, los propuestos por Wilk y Gnanadesikan, 1968).

De acuerdo con Fox (1990) es más fácil entender a una distribución como función de densidad que como una función de distribución acumulada. Las densidades son los análogos continuos de las proporciones y a causa de que la función densidad es la derivada de la función de distribución acumulada, las áreas bajo la curva de la función densidad son probabilidades. Además, la función empírica de distribución acumulada es discontinua, lo cual no es muy propio para la estimación directa de densidades de variación gradual.

En este capítulo se introduce a los estimadores clásicos de densidad: histogramas y polígonos de frecuencia como antecedentes para después revisar estimadores más sofisticados.

Se utilizarán como auxiliares didácticos varios programas en lenguaje Pascal y en macros del paquete estadístico Stata (Stata Corp., 1997), los cuales se incluyen en los apéndices. Aunque varias interpretaciones más completas son posibles, se hará hincapié en el papel de histograma descriptivo y suavizador que poseen estos métodos. Desde un punto de vista exploratorio, las estimaciones de densidad son valiosas porque dan indicación del sesgo, multimodalidad y grosor de las colas de distribución de los datos, características que pueden ser investigadas adicionalmente en una etapa estadística confirmatoria (Silverman, 1986).

1.2. El histograma. Algunas notas acerca del estimador de densidad clásico

El método más ampliamente utilizado para representar la forma de una función de densidad de probabilidad es el histograma. Tarter y Kronmal (1976) hacen una revisión crítica de este procedimiento y destacan que el histograma es útil para propósitos descriptivos. Sin embargo, este procedimiento con frecuencia resulta en una estimación pobre de la función de densidad de la población. No obstante, el histograma, después de la correspondiente transformación de las unidades del eje de las abscisas (considerando intervalos de igual amplitud) proporciona una estimación cruda de la densidad (Chambers, *et al.*, 1986). Silverman (1986) y Fox (1990) dan definiciones formales para este estimador de densidad. Versiones más recientes se encuentran en Härdle (1991) y Scott (1992). Siguiendo a éste último autor tenemos, en general, la siguiente definición: Dividir la línea de los números reales en intervalos

$$B_j = [x_0 + (j-1)h, x_0 + jh) \;\; j \in Z$$

con amplitud $h > 0$, y **origen** en x_0.. Entonces el histograma es

$$\hat{f}_h(x) = (nh)^{-1} \sum_{i=1}^{n} \sum_j I(X_i \in B_j)\, I(x \in B_j)$$

Notar que en esta definición, las frecuencias han sido normalizadas para que el área bajo las barras del histograma sea igual a la unidad; una observación en particular que cae en el límite de un intervalo se coloca en el intervalo superior. Es común que la escala del eje vertical se escoja para que la altura de las barras represente a la frecuencia en número ó porcentaje (Fox, 1990; Chambers *et al.*, 1983).

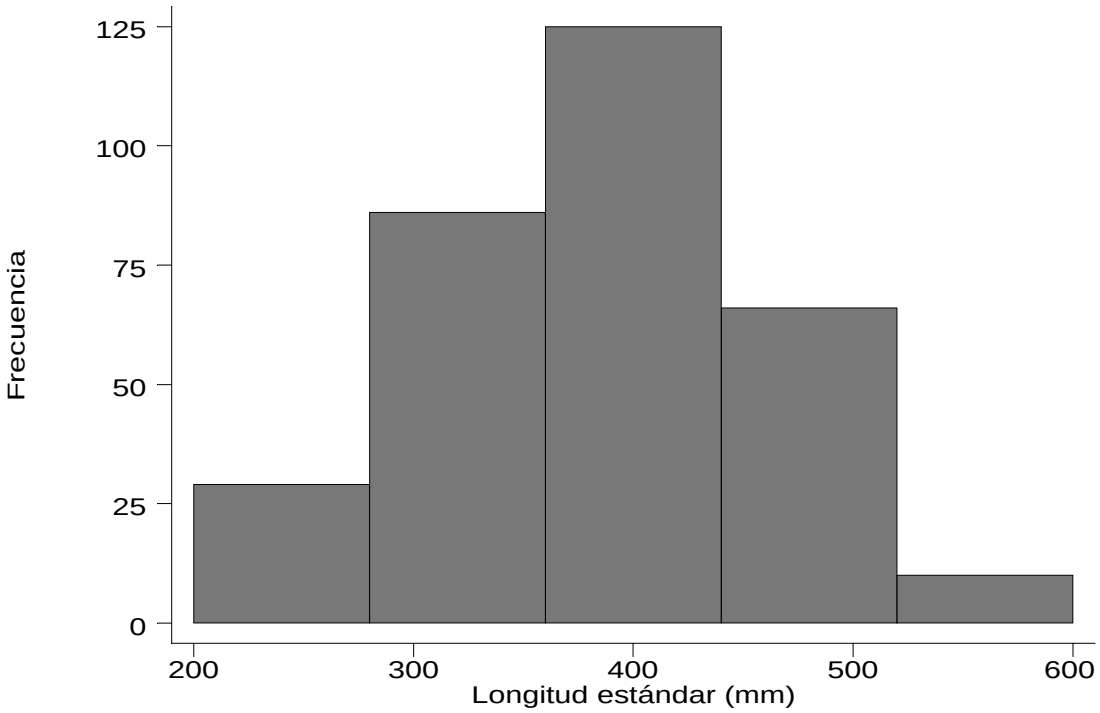

Figura 1.1 Histograma con cinco intervalos para los datos de longitud de la trucha coralina *Plectropomus leopardus* (Goeden, 1978)

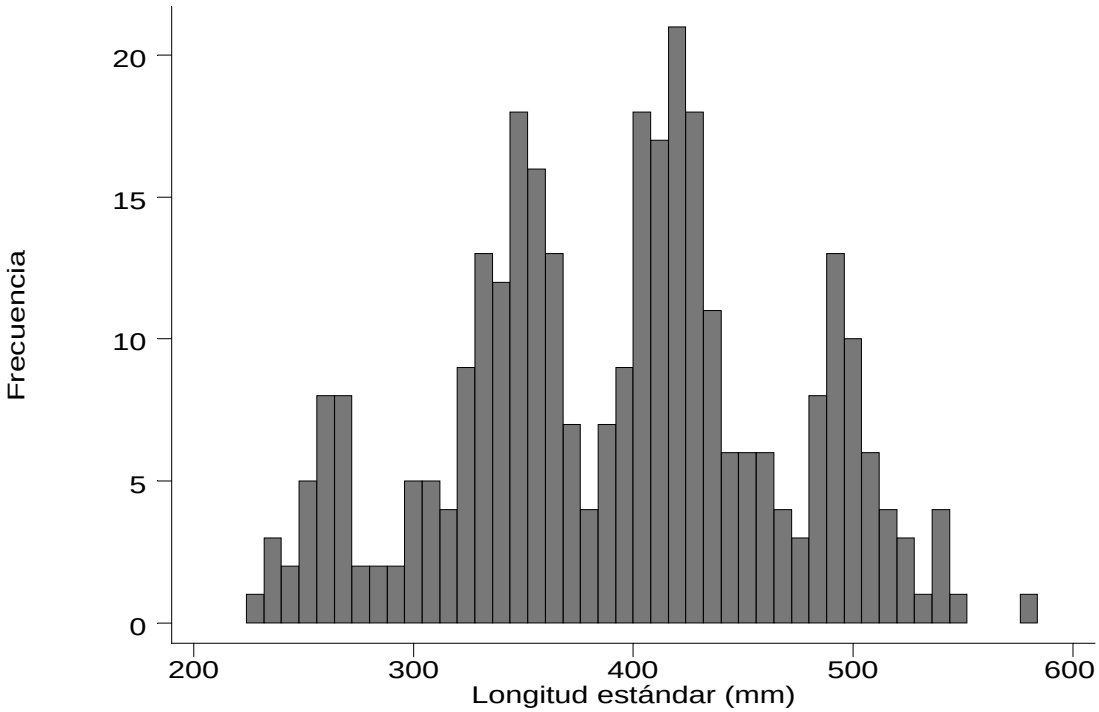

Figura 1.2. Histograma con 50 intervalos para los datos de longitud de la trucha coralina *Plectropomus leopardus* (Goeden, 1978)

Las figuras 1.1 y 1.2 muestran dos versiones de histograma. Los datos s han adaptado de Goeden (1978) y consisten de 316 observaciones de longitud de la trucha coralina *Plectropomus leopardus*. El primero es el histograma dibujado por omisión por el paquete estadístico Stata, el cual utiliza cinco intervalos. La segunda figura es el histograma para los mismos datos pero utilizando 50 intervalos. No resulta claro si la longitud de las truchas sigue una distribución multimodal o de si lo angosto de los intervalos destaca el ruido en los datos que sigan una distribución unimodal.

Estas dos gráficas dan impresiones muy diferentes de la distribución de longitudes. La Figura 1.1 muestra una distribución suave y al parecer unimodal, quizás una gaussiana o lognormal. La figura 1.2 en contraste, muestra una distribución muy irregular que contiene al menos 4 modas. La diferencia entre estas dos figuras ilustra algunos de los problemas en el uso de histogramas ya sea para estimar la función de densidad de probabilidad o para proporcionar una representación adecuada sencilla y descriptiva de la distribución de un lote de datos. Fox (1990) identifica cuatro problemas distintos en el histograma:

1.- El resultado es dependiente en el origen x_0.
Al construir un histograma, el investigador debe escoger la posición en la cual colocar el origen de los intervalos. Esta decisión se hace generalmente por conveniencia, con intervalos empezando en números centrados o redondeados. Este elemento de elección en la construcción del histograma puede interactuar con otras elecciones (como el número y amplitud de los intervalos) para producir resultados engañosos (Tarter y Kronmal, 1976). Silverman (1986) y Fox (1990) presentan algunos histogramas con varios origenes los cuales producen impresiones diferentes de la distribución de un sólo lote de datos (por ejemplo, distribuciones unimodales o bimodales dependiendo de la posición del origen). Para descubrir si este problema existe, deben dibujarse varios histogramas con diferentes orígenes, pero, esto puede conducir a se elija el preferido (intencionalmente o no) por el analista.

2.- El resultado depende de la amplitud y número de intervalos

Tarter y Kronmal (1976) señalaron que el número de intervalos debería determinarse con base en alguna función del tamaño de la muestra. Para lotes grandes, un gran número de intervalos debería dar una representación suave de la función de densidad desconocida. Al utilizar unos pocos intervalos, sin embargo, resultará en la pérdida de cualquier detalle de la distribución subyacente. Elegir un gran número de intervalos con un lote pequeño de datos producirá un diagrama univariado de dispersión. El considerar muy pocos intervalos producirá una imagen sin características. Frecuentemente, el número de intervalos y su amplitud se determinan arbitrariamente a pesar de su importancia, la cual radica en que la amplitud de intervalo determina el grado de suavidad del histograma resultante (Chambers, *et al.*, 1983; Silverman, 1986).

3.- El histograma es discontinuo, con saltos al final de los intervalos:

Las discontinuidades del histograma están principalmente en función de la localización arbitraria de los intervalos y la naturaleza discreta de los datas más que en función de la población que fue muestreada (Fox, 1990). El hecho es que para el histograma, la densidad local se calcula sólo en cada centro de clase de los intervalos y entonces las barras se dibujan suponiendo una densidad constante a lo largo de cada intervalo (Chambers, *et al.* 1983). Desde un punto de vista matemático, Silverman (1986) destacó que la discontinuidad de los histogramas causa dificultad extrema si se requieren derivadas de la densidad estimada.

4.- La amplitud fija de intervalo resulta en una representación desproporcionada de la densidad en el centro y las de las colas de la distribución.

Si los intervalos son lo suficientemente angostos para capturar detalles, típicamente en el centro de la distribución, estos pueden ser muy angostos para evitar ruido donde la densidad es baja, por lo común en las colas (Fox, 1990). En relación con este problema, el histograma puede generalizarse permitiendo que la amplitud de intervalo varíe (Silverman, 1986). Lo anterior se hace para el primer y el último intervalos del histograma, las cuales, generalmente se construyen para contener todos los puntos bajo un cierto valor y todos aquellos por encima de otro (Tarter y Kronmal, 1976). Sin embargo, los histogramas resultantes pueden malinterpretarse puesto que la altura y el área de las barras dejan de ser proporcionales (Fox, 1990).

De acuerdo con Tarter y Kronmal (1976), en consideración de algunos de los problemas descritos anteriormente, el histograma no es adecuado para estimar la distribución de la población subyacente, ni para propósitos inferenciales ni para comparar la distribución de varias poblaciones.

A pesar de sus inconvenientes, el valor del histograma como una importante y útil herramienta estadística es innegable. Sin embargo, vale la pena considerar como alternativas a las diferentes estimaciones de densidad disponibles (Wegman, 1972a,b; Silverman, 1986; Fox, 1990).

1.3. Polígonos de frecuencia

Como se afirmó anteriormente, las discontinuidades del histograma limitan su utilidad como un estimador de densidad. El **polígono de frecuencia** (PF) es un estimador de densidad continuo derivado por la interpolación lineal de los centros de clase del histograma. Scott (1985a) examinó las propiedades teóricas de polígonos de frecuencia univariados y bivariados y encontró que poseen mejoras considerables sobre los histogramas.

Vale la pena mencionar que Fisher (1932, 1958) desaprobó a los PF's por razones gráficas. El razonó que el PF conduce a una curva engañosa. De acuerdo con él, se debe tener cuidado de distinguir a la población hipotética continua de tamaño infinito de la cual se ha sacado la muestra de las observaciones finitas y discontinuas contenidas en dicha muestra. El trabajo de Scott ha probado que la objeción de Fisher hacia el uso de un estimador de densidad continuo no es válida, aunque su preocupación acerca del uso de técnicas que obscurecen el ruido estadístico con sofisticación matemática debe remarcarse (Scott, 1992).

El polígono de frecuencia (PF) univariado es el interpolante lineal de los centros de clase de un histograma con intervalos de la misma amplitud. El PF se extiende más allá del histograma hacia un intervalo vacío a cada extremo y se ha verificado que es una función de densidad *bona fide* (es positivo con integral igual a la unidad). Considerando que el PF conecta dos valores adyacentes del histograma $\hat{f}_0$ y $\hat{f}_1$, entre los centros de clase, el PF es descrito por la ecuación:

$$\hat{f}(x) = \left(\frac{1}{2} - \frac{x}{h}\right)\hat{f}_0 + \left(\frac{1}{2} + \frac{x}{h}\right)\hat{f}_1, \qquad -\frac{h}{2} \le x < \frac{h}{2}$$

La aleatoriedad en el polígono de frecuencia deriva enteramente de los intervalos del histograma.

Del trabajo de Scott(1985a) puede afirmarse que el PF en contraste con el histograma:

- aproxima mejor a densidades continuas por medio de interpolación lineal de intervalos más amplios.

- es menos eficiente cuando la densidad subyacente es discontinua.

- es más sensitiva con respecto a errores en la elección de la amplitud de intervalo, particularmente cuando $h < h_0$.

- requiere un error mayor en la amplitud de intervalo antes de que su error cuadrado medio integrado (ECMI) sea peor que el mejor ECMI de histograma.

- es mas eficiente para los datos en relación al histograma al crecer el tamaño de muestra.

Como ejemplo se incluyen dos polígonos de frecuencia para los datos de longitud de la trucha coralina. La Figura 1.3 muestra el PF con amplitud de intervalo de 8 como en el histograma de la Figura 1.2 pero con origen en 216. El PF de la Figura 1.4 tiene una amplitud de intervalo de 16 con un origen en 208. El programa para calcular estos PF utiliza centros de clase ligeramente diferentes por lo que no corresponden exactamente con los histogramas previos. Otra diferencia es que la escala vertical es densidad.

La suavidad de estas estimaciones en comparación con el histograma de la Figura 1.2 es evidente, pero también puede notarse el efecto de la posición del origen. La teoría asintótica establece que la elección del origen de los intervalos es despreciable. No obstante, en la práctica, como puede apreciarse en las figuras siguientes, se obtienen diferentes estimaciones con tan sólo variar ligeramente la posición del origen. Mas ejemplos se describirán en secciones posteriores.

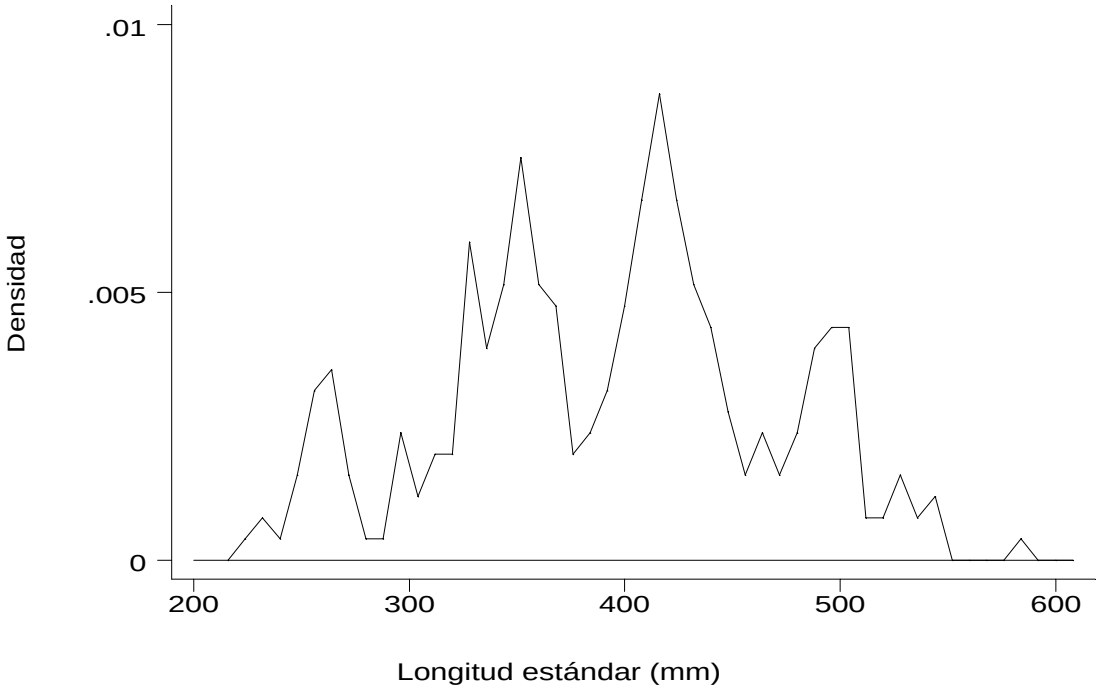

Figura 1.3. Polígono de frecuencia para los datos de longitud de la trucha coralina (Goeden, 1979). La amplitud de intervalo (8) es semejante al histograma de la Figura 1.1.2.2, pero los centros de clase difieren ligeramente (origen = 216).

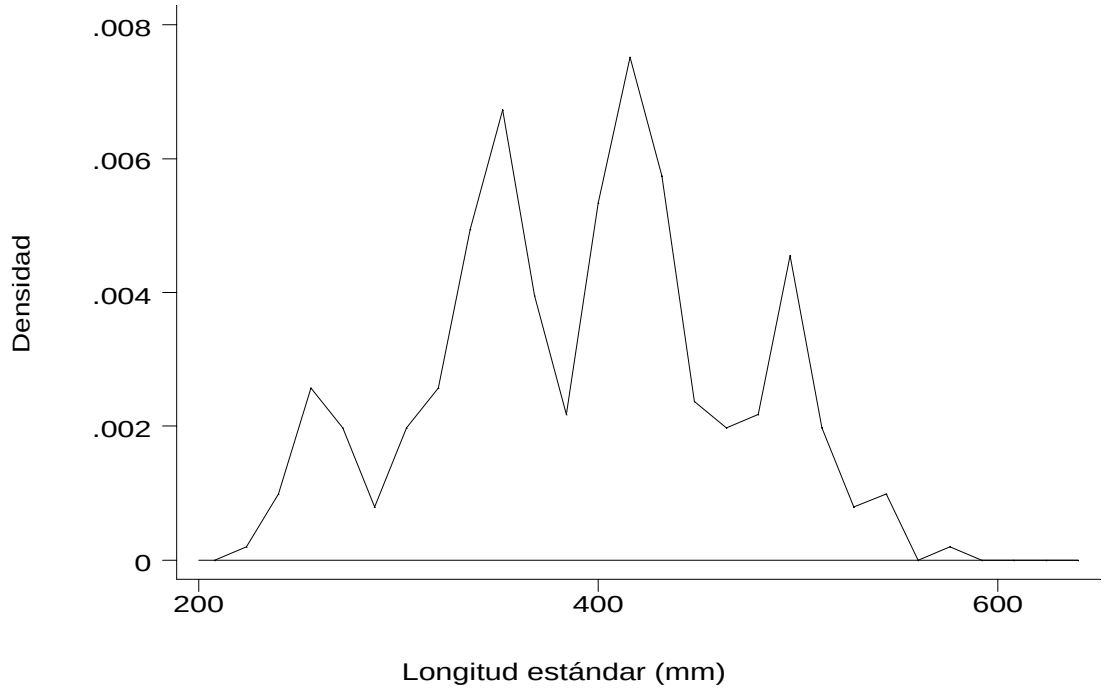

Figura 1.4 Polígono de frecuencia para los datos de longitud de la trucha coralina (Goeden, 1979) con amplitud de intervalo = 16 y origen = 208.

Capítulo 2. Estimadores de densidad por kernel

2.1 Trazas de densidad

Chambers *et al*. (1983) sugieren que para resolver la discontinuidad del histograma (problema citado en la sección 1.1) es posible calcular la densidad local en cada uno de los valores de *x*, y dejar que se traslapen los intervalos en los cuales se cuentan las observaciones. En esencia, se construyen intervalos de amplitud fija alrededor de cada observación considerando algún intervalo con amplitud *h*. Estos intervalos se traslapan donde las observaciones se concentran, por lo que se evita la apariencia discontinua del histograma. Estas densidades locales constituyen a las **trazas de densidad.**

Más formalmente, para cada observación *x*, calculamos:

$$\text{densidad local } en\ x = \frac{\text{numero de observaciones en } [x-h/2, x+h/2]}{h \times \text{ numero total de observaciones}}$$

Aplicado a los datos de longitud de la trucha coralina las trazas de densidad resultantes se muestran en las Figuras 2.1 y 2.2. El cálculo fue realizado con h = 40 y uniendo los valores de densidad con líneas y con líneas en pasos (lo cual probablemente es más adecuado debido a la naturaleza discontinua de la función de peso) respectivamente. Debido a esta implementación particular y debido a que cada observación en los datos queda incluida, es posible combinar esta gráfica bivariada con diagramas univariados de dispersión y diagramas de caja para cada eje. De esta forma es posible comparar la información de cada técnica de visualización de la distribución de los datos. La traza de densidad revela claramente cuatro modas, lo que confirma la impresión del histograma con 50 intervalos (Figura 1.2). Estas modas no son reveladas en absoluto por los diagramas de caja y son difíciles de encontrar en el diagrama univariado de dispersión.

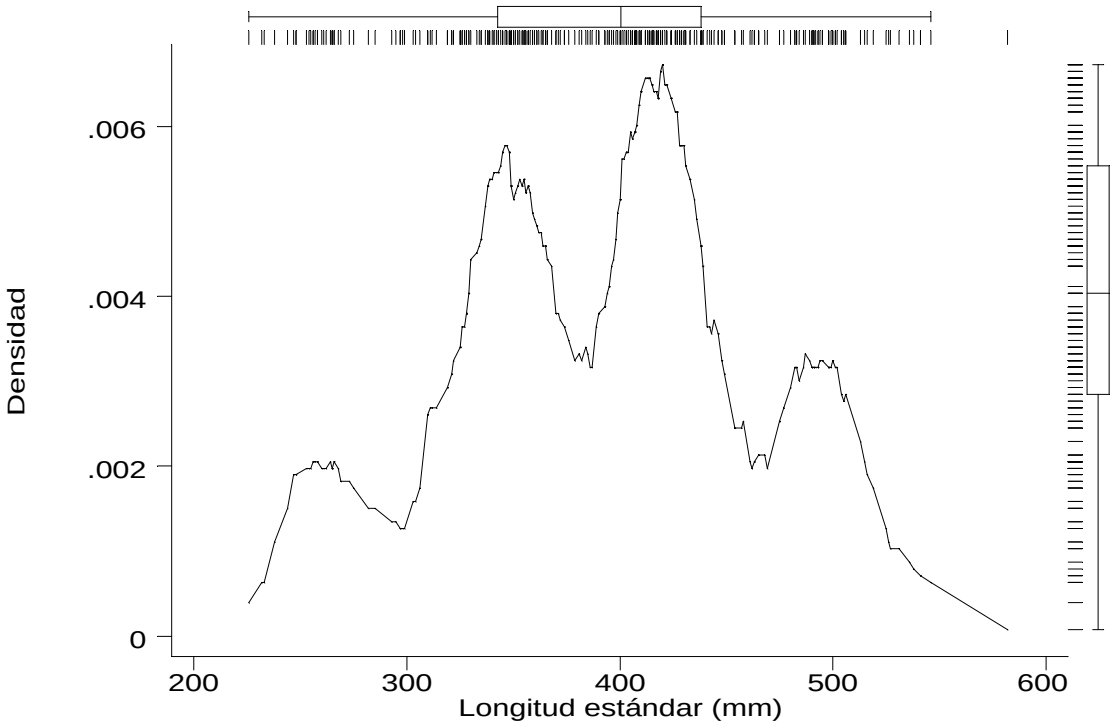

Figura 2.1 Traza de densidad para los datos de longitud de la trucha coralina con valores de densidad conectados por líneas rectas.

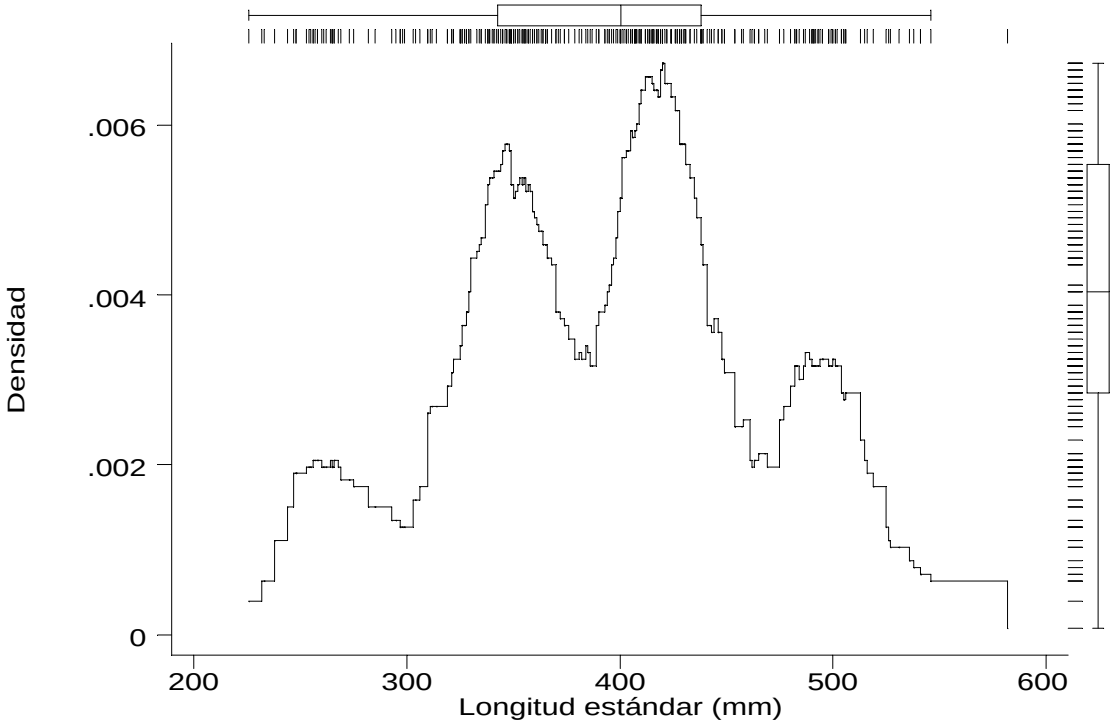

Figura 2.2 Traza de densidad para los datos de longitud de la trucha coralina con valores de densidad unidos por lineas a pasos.

Un inconveniente para el cálculo de este procedimiento es que realiza *n* resúmenes estadísticos básicos para obtener cada estimación de densidad de tal forma que toma un tiempo considerable para hacer la estimación con muestras grandes.

Un enfoque equivalente es considerar cada valor en la muestra no como un sólo punto sino como el centro de un intervalo extendido en *h* y definir una función de peso, por ejemplo una función ponderal rectangular:

$$W(u) = \begin{cases} 1, & \text{si } |u| \le 1/2; \\ 0, & \text{en cualquier otro caso} \end{cases}$$

Esta es una función de peso con forma rectangular. El extender en un intervalo la influencia de una observación es equivalente a reemplazar x_i por una función de x, $W[(x\text{-}x_i)/h]/h$, así que el valor de densidad en x, $f(x)$ es el promedio de todos los intervalos x_i en x, de acuerdo a la fórmula:

$$\hat{f}(x) = \frac{1}{hn} \sum_{i=1}^{n} W\left(\frac{x - x_i}{h}\right)$$

la cual corresponde exactamente a la regla descrita anteriormente.

Chambers, *et al.* (1983) sugiere que en la práctica, para economizar cálculos y tiempo, es suficiente considerar un cierto número de puntos igualmente espaciados sobre el recorrido de los datos e interpolar linealmente entre ellos.

La figura 2.3 despliega un ejemplo de este tipo de traza de densidad para los datos de longitud de la trucha coralina utilizando el mismo valor de *h* (40). Cada uno de los puntos de estimación es indicado por un signo "+" los cuales están conectados por líneas rectas.

Esta estimación de 50 puntos da la misma impresión general que las estimaciones de las Figuras 2.1 y 2.2, aunque parece un poco más suave como consecuencia del número reducido de puntos considerados para los cálculos.

Estos estimadores simples de la densidad eliminan la discontinuidad local de los intervalos en los histogramas, pero siguen teniendo algo de ruido. Una razón para esta rugosidad es la forma rectangular de la función de peso con forma cuadrada. Para suavizar aún más la estimación de densidad podemos considerar una función ponderal diferente, la cual varíe gradualmente a lo largo del intervalo *h*, por ejemplo la función ponderal coseno, definida como:

$$W(u) = \begin{cases} 1+\cos 2\pi\ u, & \text{si } |u| < 1/2; \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

La figura 2.4 despliega una traza de densidad con función coseno y 50 puntos para los datos de longitud de la trucha coralina. Cada punto está marcado con un círculo "o" y conectados con una curva suave (spline cúbico). El resultado es más suave y muestra la misma imagen de una distribución multimodal.

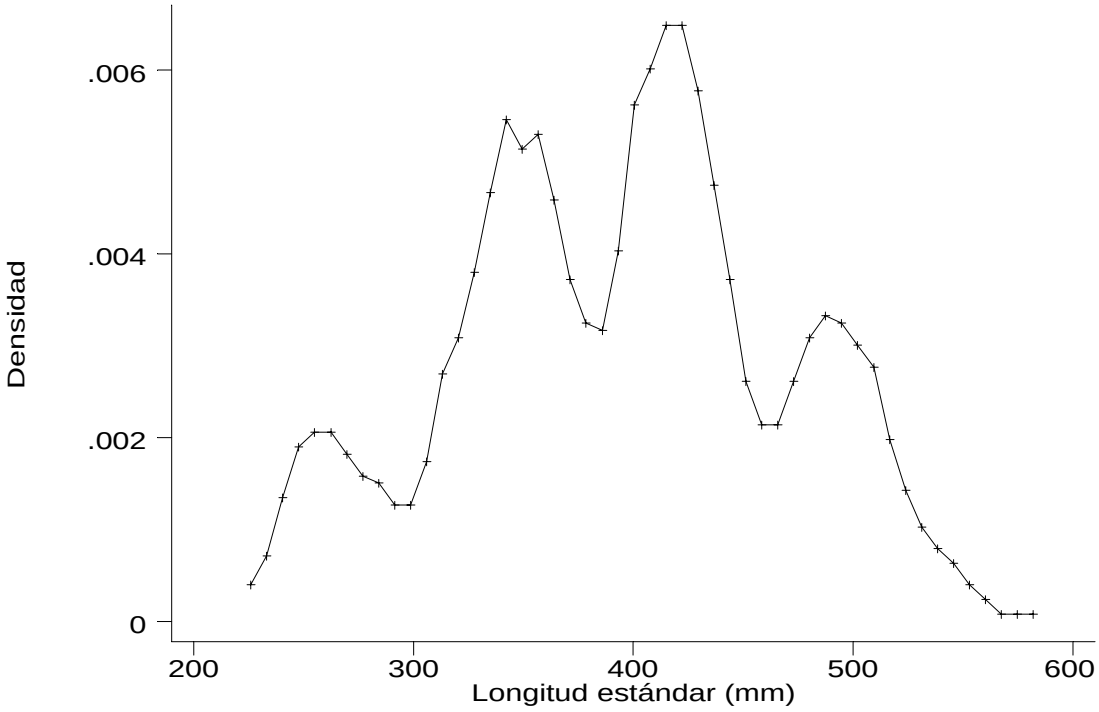

Figura 2.3 Traza de densidad para los datos de trucha coralina (estimación por 50 puntos), *h* = 40.

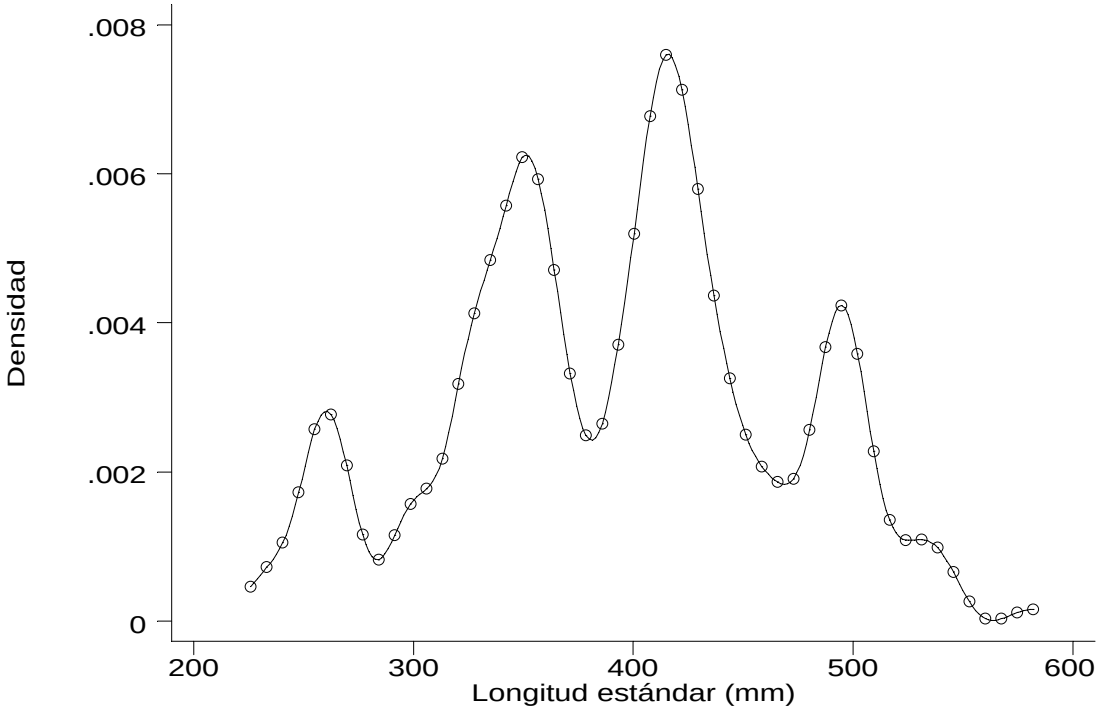

Figura 2.4 Traza de densidad por función coseno para los datos de trucha coralina (estimación de 50 puntos), *h* = 40.

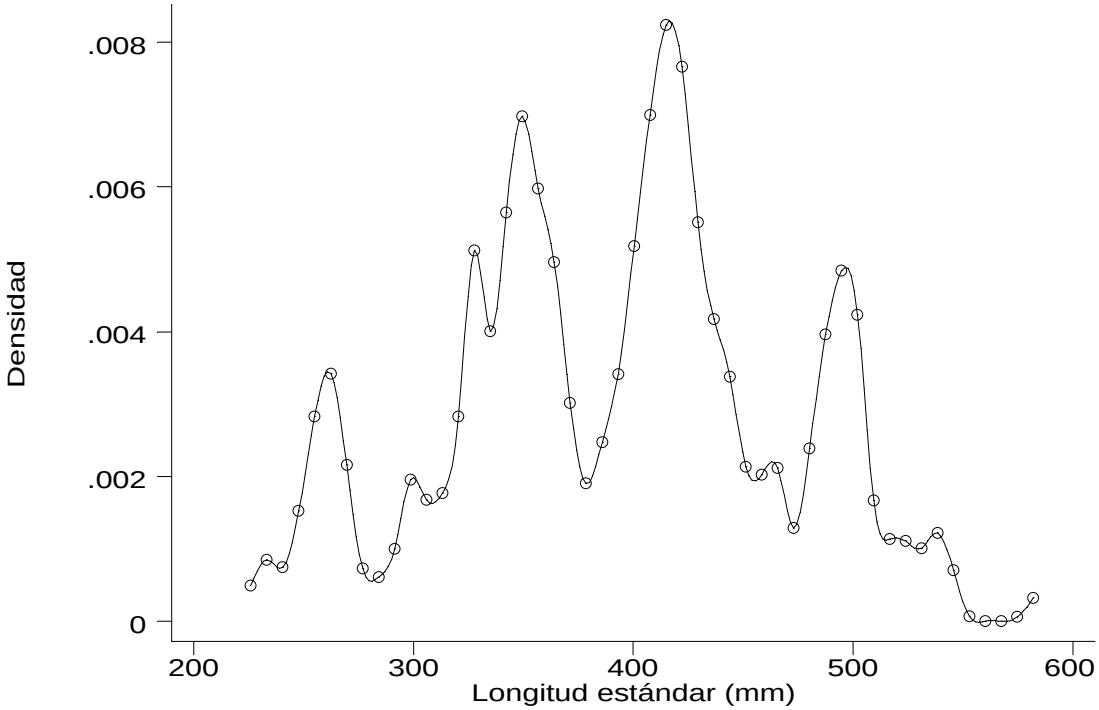

Figura 2.5 Traza de densidad por función coseno para los datos de longitud de la trucha coralina (estimación de 50 puntos) *h* = 20.

Es importante señalar que el valor de h determina que tan suave será la estimación de densidad. Al disminuir h $f(x)$ adquiere una apariencia menos suave como función de x. La figura 2.5 contiene la estimación de densidad por función coseno con $h = 20$.

2.2 El estimador de densidad simple

En una forma similar aunque un poco diferente (técnicamente), Fox (1990), siguiendo a Silverman (1986), describe un estimador de densidad simple considerando que la densidad puede pensarse como el límite de un histograma de barras centradas en *x*, al tender la amplitud de intervalo a cero:

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

Entonces, un estimador simple de densidad substituye a la proporción de la muestra en una región pequeña (llamada ventana) alrededor de *x* por la probabilidad escalando la estimación de tal forma que el área total bajo *f*(*x*) sea igual a la unidad:

$$\hat{f}(x) = \frac{1}{2h} \times \frac{\#[x - h < X_i < x + h]}{n}$$

Este estimador de densidad es semejante al histograma con intervalos de clase igual a 2*h* pero con cada punto observado en el centro del intervalo, esto es, sin origen fijo.

La función ponderal $S(\bullet)$ para este estimador simple esta dado por:

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

donde

$$S(u) = \begin{cases} 1/2, & \text{si } |u| < 1; \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Como en las funciones cuadradas y coseno descritas anteriormente, *u* representa el argumento de la función ponderal, la cual es semejante, pero ligeramente diferente y que de nuevo integra hasta la unidad.

En los programas escritos, el cálculo de este estimador simple de densidad se realiza considerando (para ahorro de tiempo y número de operaciones, como lo sugieren Chambers, *et al.*, 1983 y Fox, 1990) un número finito de puntos uniformemente espaciados (50). Tomando en cuenta que este estimador simple integra hasta uno entre $x_{(1)} - h$ y $x_{(2)} + h$, estos puntos están localizados (de manera convencional) de $x_{(1)} - h + (d/2)$ hasta $x_{(n)} + h + (d/2)$, donde *d* es el intervalo definido por el recorrido/50, donde el recorrido se extiende de $x_{(1)} - h$ a $x_{(2)} + h$. Esta convención es adoptada aquí para todos los programas correspondientes a esta sección (excepto para la implementación del estimador de densidad por kernel de amplitud de banda variable que utiliza a todas las observaciones). La conveniencia de esta definición es que los puntos de estimación (con excepción del último) pueden ser considerados como centros de clase de histogramas semejantes con fines comparativos y para determinación y caracterización de componentes gaussianos.

La estimación de densidad resultante para este procedimiento utilizando 50 puntos y considerando una amplitud de banda de 20 se muestra en la figura 2.6 Es virtualmente idéntica a la

obtenida con la traza de densidad cuadrada utilizando una amplitud de banda de 40 (Figura 2.3) mostrando (al menos) una distribución tetramodal. Esto se esperaba debido a que esta amplitud de banda se escogió para obtener versiones suavizadas aproximadamente equivalentes. Como antes, el valor de *h* controla la cantidad de suavización, y con valores mayores produciendo estimaciones más suaves.

Silverman (1986) recalcó que además de su carácter rugoso estéticamente no deseado, estas estimaciones algunas veces pudieran proporcionar una impresión incorrecta.

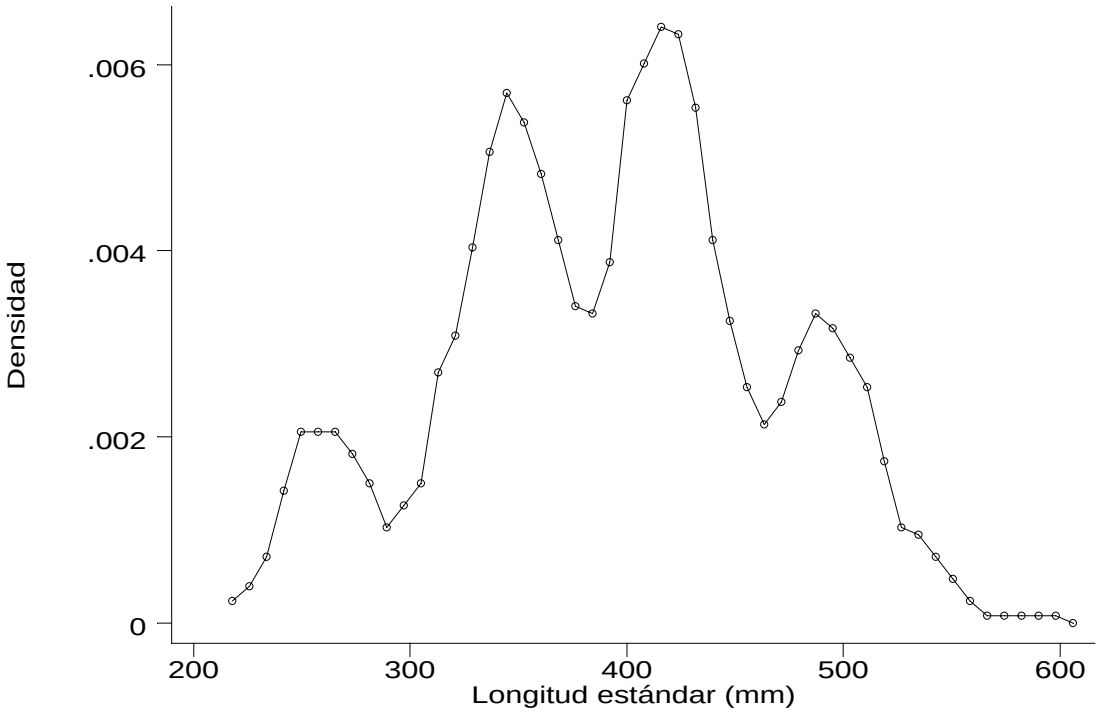

Figura 2.6 Estimación simple rectangular de densidad (*h* = 20 mm) para los datos de longitud de la trucha coralina.

2.3 Estimadores más sofisticados de densidad por kernel

Para resolver el problema de la discontinuidad causada por las esquinas cuadradas de la función ponderal rectangular que utiliza el estimador simple presentado con anterioridad (y por otras razones técnicas discutidas por Silverman, 1986), es conveniente considerar la generalización de este estimador de densidad simple. Esto se consigue reemplazando la función ponderal $S(\bullet)$ por otra produciendo figuras redondeadas en lugar de rectángulos. Para indicar esta distinción la nueva función ponderal se anota como función por kernel $K(\bullet)$ (Fox, 1990).

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

Aquí, la función kernel K es una densidad de probabilidad suave, simétrica y que integra a la unidad.

$$\int_{-\infty}^{+\infty}K[u]\,du=1$$

Algunas funciones por kernel se incluyen en el Cuadro 2.1, adaptado de Silverman (1986).

Cuadro 2.1. Algunos kerneles y sus eficiencias

Kernel	Ecuación	Eficiencia
Epanechnikov	$K(u)=\begin{cases}\frac{3}{4}(1-\frac{1}{5}u^2)/\sqrt{5}, & \text{si } \lvert u\rvert<\sqrt{5};\\ 0, & \text{en cualquier otro caso.}\end{cases}$	1
Biponderado	$K(u)=\begin{cases}\frac{15}{16}(1-u^2)^2, & \text{si } \lvert u\rvert<1;\\ 0, & \text{en cualquier otro caso.}\end{cases}$	≈ 0.9939
Triangular	$K(u)=\begin{cases}1-\lvert u\rvert, & \text{si } \lvert u\rvert<1;\\ 0, & \text{en cualquier otro caso.}\end{cases}$	≈ 0.9859
Gaussiano	$K(u)=\frac{1}{\sqrt{2\pi}}e^{-u^2/2}$	≈ 0.9512
Rectangular	$K(u)=\begin{cases}1/2, & \text{si } \lvert u\rvert<1;\\ 0, & \text{en cualquier otro caso.}\end{cases}$	≈ 0.9295

Como con otros suavizadores, para evaluar el desempeño de estos kerneles es necesario considerar el compromiso entre la varianza y el sesgo. La suma de la varianza y sesgo integrados es el error cuadrado integrado (total) medio (*ECIM*). Una buena función kernel deberá minimizar el sesgo por la asignación de un peso mayor a las observaciones cercanas al valor de x en el cual se estima la densidad (por ejemplo por medio del uso de una función gaussiana). Epanechnikov (1969), al minimizar el *ECIM* definió la función de eficiencia máxima que lleva su nombre. En el Cuadro 2.1 cada función kernel presenta su correspondiente eficiencia bajo este criterio. Como Silverman señaló, con base en el *ECIM*, hay muy poco que escoger entre los varios kerneles. Aún la función de kernel rectangular utilizada por el estimador simple tiene una eficiencia relativa del 93 %. Por tanto, es válido y deseable el escoger al kernel bajo otras consideraciones tales como el grado de diferenciabilidad o esfuerzo computacional (Silverman, 1986).

En esta sección se comentan algunas implementaciones de las funciones kernel Epanechnikov y Gaussianas. La estimación suave de la densidad con el kernel de Epanechnikov con una amplitud de banda de 20 se muestra en la Figura 2.7, con 50 puntos conectados con splines cúbicos. Parece una versión sobresuavizada, pero también sugiere una distribución polimodal. Al disminuir *h* a 10, se produce la densidad mostrada en la Figura 2.8, la cual da casi la misma impresión que el estimador simple de la Figura 2.6, con cuatro modas claramente definidas.

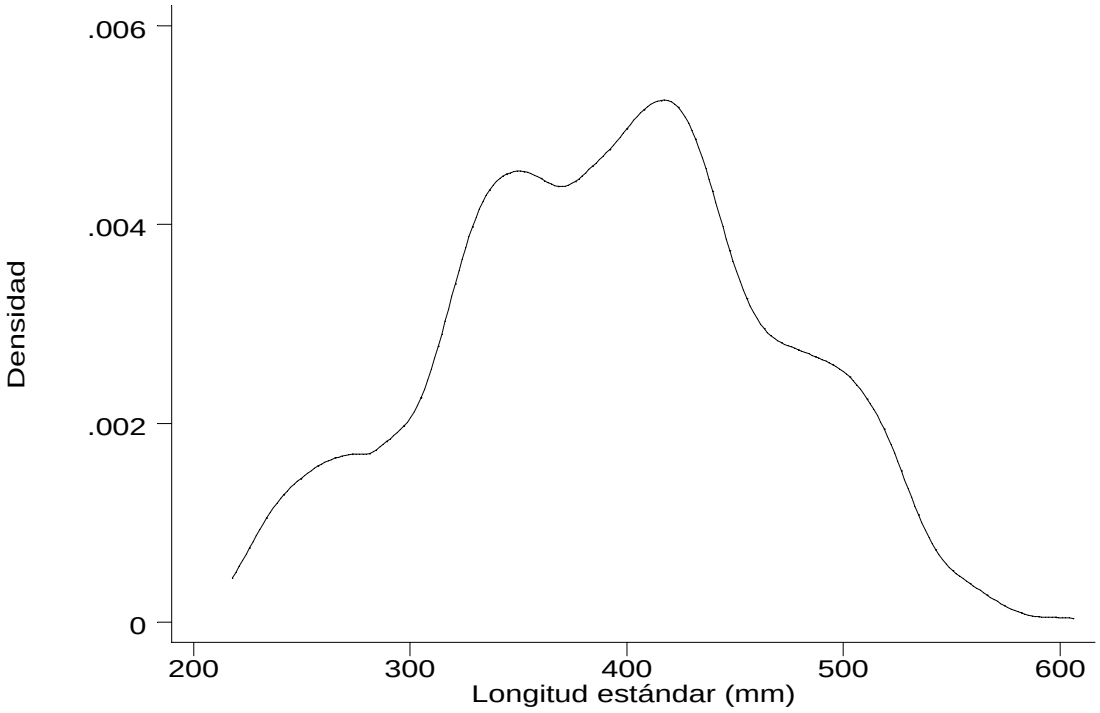

Figura 2.7. Estimación de densidad por kernel de Epanechnikov (*h* = 20 mm).

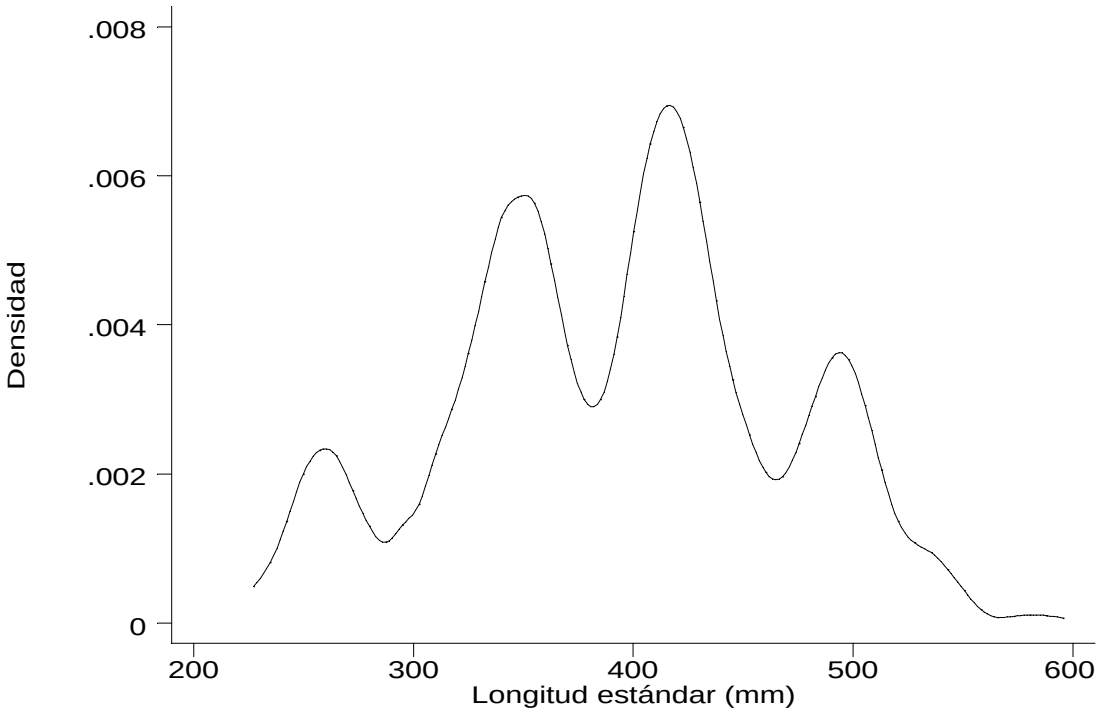

Figura 2.8 Estimación de densidad por kernel de Epanechnikov (*h* = 10 mm).

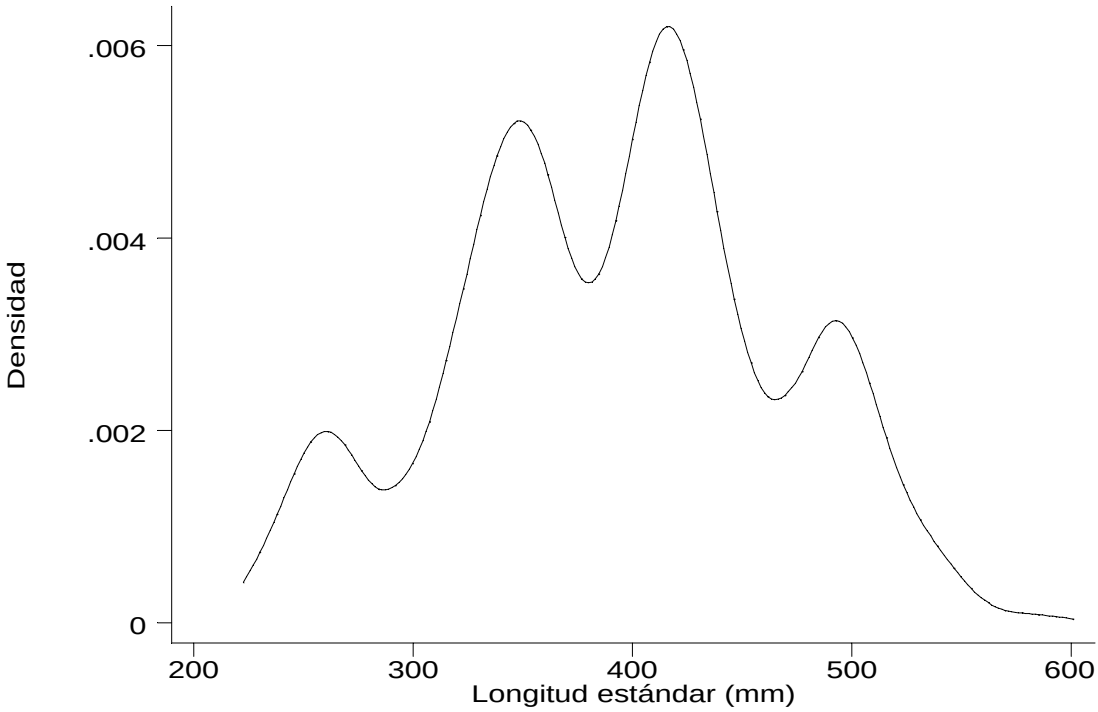

Figura 2.9. Estimación de densidad por kernel gaussiano (*h* = 15 mm).

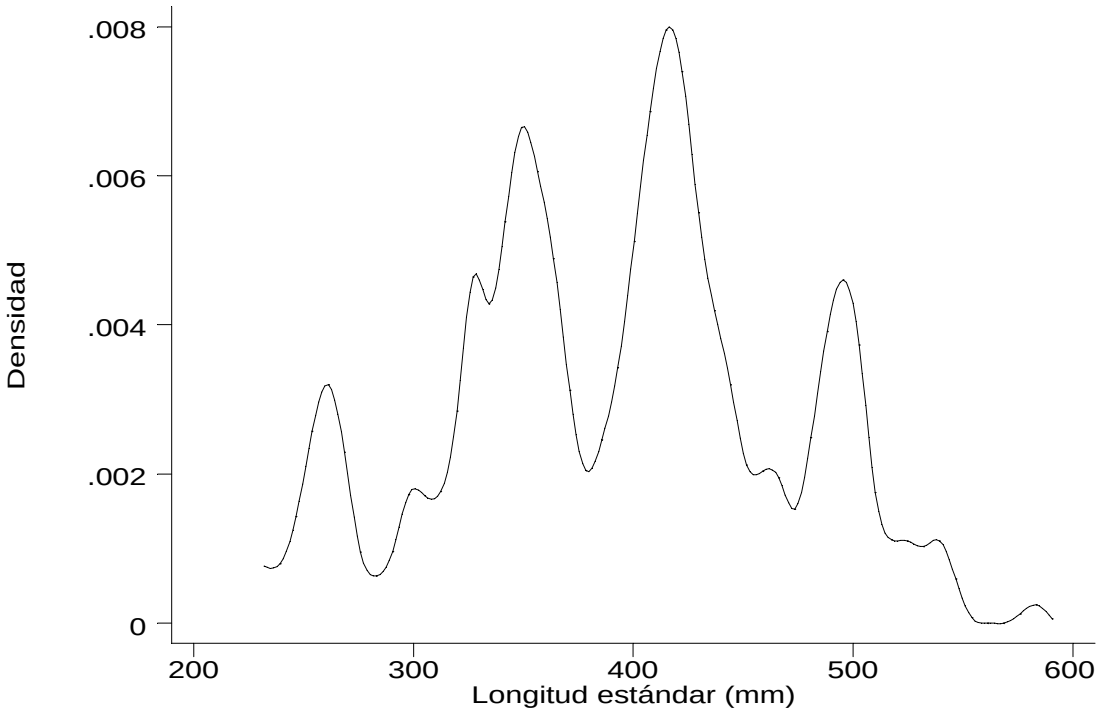

Figura 2.10 Estimación de densidad por kernel gaussiano (*h* = 5 mm).

Las figuras 2.9 y 2.10 muestran dos estimaciones por kernel gaussiano para la densidad de los datos de longitud de la trucha coralina. La figura 2.9 utiliza una *h* = 15 mm, mientras que la 2.10 la reduce a 5 mm. Tanto el ruido y el detalle se incrementa al disminuir *h*, y en esta última figura es posible distinguir la descomposición de la segunda moda y la aparición de una moda intermedia entre la tercera y cuarta con respecto a las de la figura 2.9. Algunas irregularidades aparecen en la cola superior de la distribución, incluyendo una prominencia alrededor de la observación de longitud máxima (un caso extraordinario moderado recordando los diagramas de caja marginales de las Figuras 2.1 o 2.2). Estas irregularidades pueden ser artefactos de la amplitud de ventana fija utilizada. No obstante, algunas de estas estructuras parecen tener sentido para este caso en particular, sugiriendo un grupo de edad traslapado ente las clases dominantes tercer y cuarta (las modas claramente definidas en la Figura 2.4) o grupos bimodales de edad debidos a un reclutamiento compuesto, característica también identificada por otros métodos que utilizan estimaciones de la edad y ecuaciones de crecimiento para este lote de datos (ver por ejemplo Pauly, 1988 o Sparre *et al.*, 1989).

2.4 Estimadores de densidad por kernel de amplitud variable

Todos los estimadores de la densidad por kernel descritos con anterioridad utilizan amplitud fija de ventana. Esta característica provoca que las estimaciones sean vulnerables al ruido en las colas o en cualquier intervalo con cuenta baja de observaciones en la distribución. Para resolver esta dificultad se han propuesto algunos métodos que utilizan una amplitud de banda variable. La descrita a continuación se tomó de Fox (1990) quien a su vez la adaptó de Silverman (1986). La idea general es el ajustar la amplitud de ventana de tal forma que sea más angosta a densidades altas y más amplia donde existen bajas densidades. El resultado de este procedimiento de amplitud variable es el retener detalle donde las observaciones se concentran y eliminar fluctuaciones ruidosas donde los

datos escasean. El algoritmo para calcular tales estimadores por kernel de amplitud variable se explica a continuación:

1. Calcular una estimación de densidad preliminar, por ejemplo la proporcionada por cualquiera de las funciones kernel con amplitud fija de ventana $f(x)$.
2. Con esta estimación inicial para cada observación de X, calcular factores locales para las ventanas w_i relacionados inversamente con la densidad:

$$w = \left[\frac{\tilde{f}_g}{\hat{f}_K(X_i)} \right]^{1/2}$$

donde

$$\tilde{f}_g = \left[\prod_{i=1}^{n} \hat{f}_K(X_i) \right]^{1/n}$$

es la media geométrica de $f(X_i)$, y por tanto los pesos tienen un producto y media geométrica de uno.

3. Utilizando los pesos, calcular el estimador de densidad por kernel de amplitud variable.

$$\tilde{f}_A(x) = \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{w_i} K\left[\frac{x - X_i}{w_i h} \right]$$

Notar que la aplicación de pesos a h produce ventanas de amplitud variable que resultan en una amplitud de banda media geométrica igual a h. El factor $1/w$ se requiere para asegurar que el área total bajo la estimación de densidad sea la unidad.

4. Iterar los pasos 2 y 3, utilizando $\tilde{f}_A$ en lugar de $\tilde{f}_K$ para recalcular los pesos. En la práctica, la iteración produce poco cambio en las estimaciones de densidad (Fox, 1990).

2.5 La elección de la amplitud de ventana

La elección de la amplitud de ventana en las estimaciones por kernel es equivalente a la selección de amplitud de intervalo en histogramas. Esta elección determina las características cualitativas de la densidad por kernel. Un enfoque, sugerido por Tarter y Kronmal (1976) es variar h hasta que resulte una figura satisfactoria (por lo general suave). Este procedimiento depende de la evaluación subjetiva del investigador, pero puede ser adecuada para propósitos exploratorios (Silverman, 1986). Ciertamente, es útil el comparar varios niveles de suavización, puesto que aspectos importantes de la densidad “aparecen” y “desaparecen” al cambiar la amplitud de ventana (Silverman, 1981a).

La teoría estadística proporciona algunos lineamientos para la elección de la amplitud. Desafortunadamente, como en el caso del histograma, por lo general no es posible optimizar la amplitud de ventana sin el conocimiento previo de la forma verdadera de la densidad. Por supuesto, si esta forma fuera conocida de antemano, no existiría problema de estimación alguno.

Siguiendo la guía de Tukey (1977), Scott (1979) y Silverman (1978, 1986), puede emplearse a la distribución gaussiana como un estándar de referencia. Entonces, aplicando un kernel gaussiano y minimizando al *ECIM* es posible utilizar las expresiones siguientes (Fox, 1990):

$$s=min \quad \left[\left(\frac{\sum (x_i - \bar{x})^2}{n-1} \right)^{1/2}, \frac{H\,\text{dispersión}}{1.349} \right]$$

Entonces *h* puede escogerse como:

$$h = \frac{0.9s}{n^{1/5}}$$

Donde *s* es la menor de dos estimaciones de parámetro σ de dispersión (escala) de la distribución gaussiana, es decir, la clásica desviación estándar insesgada y la robusta F-pseudosigma (*H* dispersión/ 1.349) basada en la dispersión de los cuartos (o en inglés conocida también como *hinge spread*) (Hoaglin, 1983), Fox, 1990). Este ajuste proporciona resistencia a colas potentes y funcionará bien para una amplia gama de densidades, pero, como lo indicó Silverman (1986), tiende a sobresuavizar distribuciones fuertemente sesgadas y multimodales. En este último caso, esta amplitud "óptima" de ventana puede considerarse como un buen punto de partida para afinación posterior.

2.6 Ejemplos del estimador de densidad por kernel de banda variable

La amplitud de ventana "óptima" para los datos de longitud de la trucha coralina se calculó de acuerdo a la fórmula dada arriba. Utilizando las funciones existentes del paquete estadístico Stata, se obtuvieron los siguientes valores *s* = 74.0629 y F-pseudosigma = 70.8814. Por tanto, *h* = 0.9×70.8814/316 = 20.18.

La implementación del estimador de densidad de amplitud variable inicia con una estimación por kernel Gaussiano con el valor anterior para *h*, calcula los pesos proporcionales a la densidad y finalmente los valores de densidad con amplitud variable para cada observación. Las densidades estimadas se muestran en la Figura 2.11, la cual presenta claramente dos modas en el centro y otras dos aparentemente sobresuavizadas en las colas. De forma similar en el uso del programa para estimación de trazas de densidad con función rectangular, al considerar a todos los valores de los datos para calcular las densidades, es posible desplegar los resultados en combinación con diagramas univariados de dispersión y diagramas de caja en los márgenes. Tomando en cuenta la naturaleza aparentemente multimodal de estos datos, podemos considerar esta ventana como lo suficientemente amplia para sobresuavizar (comparar por ejemplo con la Figura 2.8).

Debido a que se requiere de un gran número de operaciones, la implementación directa puede tomar un tiempo considerable para su cálculo. Considerar que es necesario determinar un peso proporcional a la densidad individual a partir de una estimación de densidad preliminar inicial.

Para ahorrar tiempo de cálculo, es posible desarrollar algoritmos que sólo consideren un número finito (por ejemplo) de puntos igualmente espaciados después de los pasos 1 y 2 del procedimiento del estimador de densidad de amplitud variable.

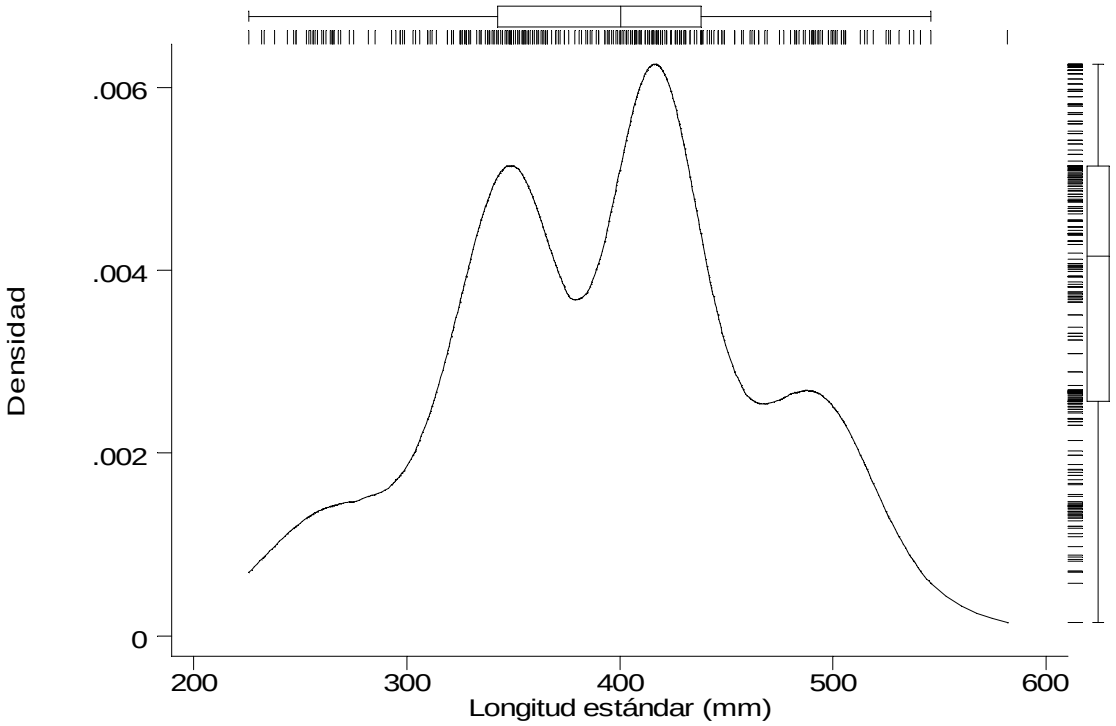

Figura 2.11. Estimación de densidad por kernel gaussiano de amplitud variable (*h* = 20.18 mm).

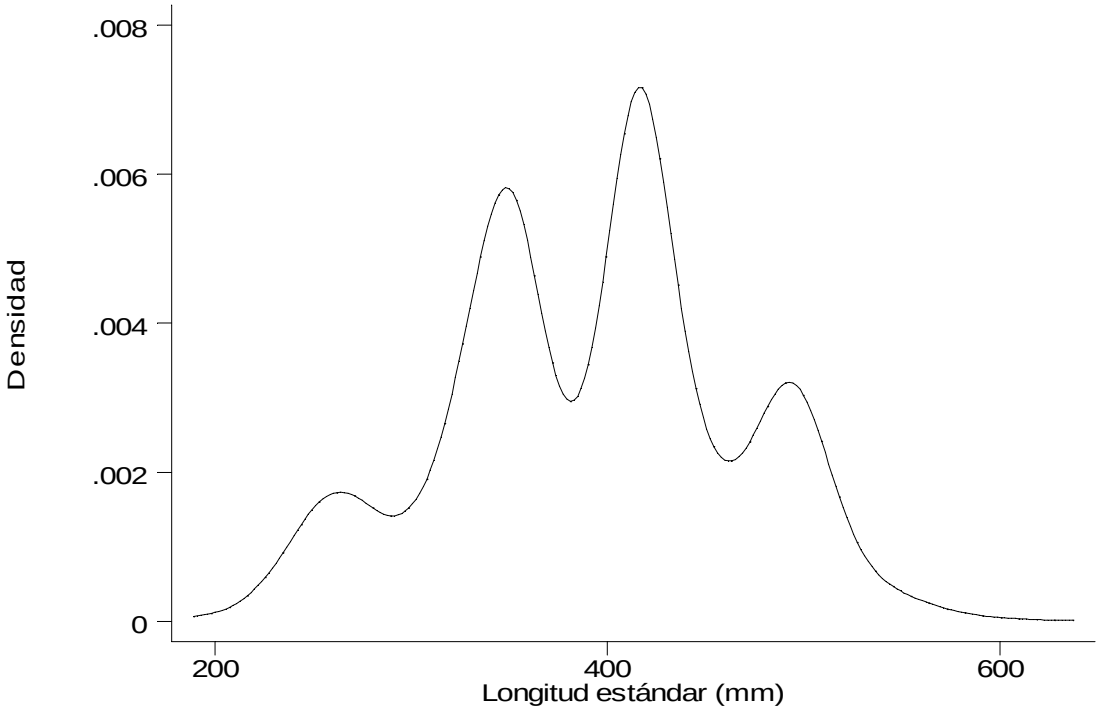

Figura 2.12 Estimación de densidad por kernel gaussiano de amplitud variable con 50 puntos y *h* = 15 mm.

Utilizando un programa con esta simplificación, se probó una amplitud de banda más pequeño. Los resultados para una $h = 15$ se muestran en la Figura 2.12; la separación y definición de las cuatro modas puede ser claramente visualizada. Este procedimiento es de gran interés en el análisis de la distribución de frecuencia de tallas porque ajusta la amplitud de ventana de acuerdo al número de observaciones que existen en un intervalo dado de talla. En el procedimiento tradicional, para suavizar irregularidades se varía la longitud del intervalo de clase del histograma, pero una vez que se ha elegido un intervalo, este intervalo fijo no evita la pérdida de detalle en intervalos con alta concentración de datos y la manifestación del ruido en intervalos con pocas observaciones. Otro enfoque exploratorio aplicando suavizadores no lineales resistentes a histogramas con agrupamiento mínimo se ha propuesto en Salgado-Ugarte (1992) y Salgado-Ugarte y Curts-García (1992, 1993).

En resumen, como un procedimiento general para analizar una distribución univariada, Fox (1990) sugiere empezar con la amplitud de ventana óptima y posteriormente ajustar (usualmente hacia abajo) este valor hasta seleccionar el valor menor de h que no produzca rugosidad inaceptable en la densidad estimada. Silverman (1986) proporciona sugerencias y orientación adicionales.

2.7 Un comentario final: ¿Cuántas modas existen?

Es bien conocido el carácter polimodal de los datos de frecuencia de tallas en Biología Pesquera y en Ecología. Esto indica por lo general varias distribuciones unimodales mezcladas. A este respecto, las estimaciones de densidad por kernel proporcionan varias formas para probar y evaluar la multimodalidad. Para detalles ver Silverman (1981b, 1983, 1986). Otros enfoques han sido sugeridos por Cox (1966) y Good y Gaskins (1980).

Las estimaciones de densidad presentadas pueden servir como un punto de inicio para intentar la estimación de los parámetros de cada uno de los componentes de edad subyacentes en la mezcla. Un ejemplo de la aplicación de estos estimadores se presenta en Salgado-Ugarte (1995).

Capítulo 3. Histogramas desplazados promedio (HDP), promedio ponderado de puntos redondeados (PPPR) y revisión de la estimación de densidad por kernel

3.1 Introducción

Los estimadores de densidad por kernel representan una colección importante de herramientas para explorar y analizar la distribución de los datos (ver Capítulo 2, Salgado-Ugarte *et al.*, 1993 y las referencias incluídas). Sin embargo, un problema importante en la aplicación de estos métodos es que tienen que realizar un gran número de cálculos lo que puede consumir una cantidad considerable de tiempo a menos que se utilicen procesadores rápidos y tamaños de muestra moderados. Scott (1985b) sugirió una forma para superar este problema: el **Histograma Desplazado Promedio** (**HDP**). Posteriormente, Härdle y Scott (1988) desarrollaron la estructura más general denominada **Promedio Ponderado de Puntos Redondeados** (**PPPR**).

Esta sección (basada en algunos capítulos de los libros de Härdle, 1991 y Scott, 1992) introduce muy brevemente a los histogramas desplazados promedio y al promedio ponderado de puntos redondeados para la estimación de densidad. Los programas en el lenguaje de Stata, Turbo Pascal y el ejecutable (Visual Basic Ver. 5.0) EDK 2000 para su cálculo se presentan en el Capítulo 7.

3.2 Histogramas desplazados promedio y promedio ponderado de puntos redondeados

Como se ha discutido anteriormente (Capítulo 2, Salgado-Ugarte, *et al.* 1993), el histograma se define al especificar dos parámetros: el origen (x_o) y la amplitud (h) de los intervalos. Se ha presentado suficiente evidencia (Silverman, 1986; Fox, 1990) que sugiere que la elección del origen puede tener una importante influencia en el histograma resultante, a pesar de algunos resultados teóricos que indican que debería de ser despreciable (Scott, 1992). Se presenta aquí un ejemplo, con los bien conocidos datos de precipitación de nieve en la ciudad de Búfalo, Nueva York (Parzen, 1979) durante 63 inviernos desde 1910/1911 hasta 1972/1973. La Figura 3.1 presenta cinco histogramas (con fracción como escala para el eje de las ordenadas) con la misma amplitud de intervalo (h = 10) pero diferente origen, cualquiera de los cuales es una estimación válida de la densidad (después de la transformación de escala correspondiente).

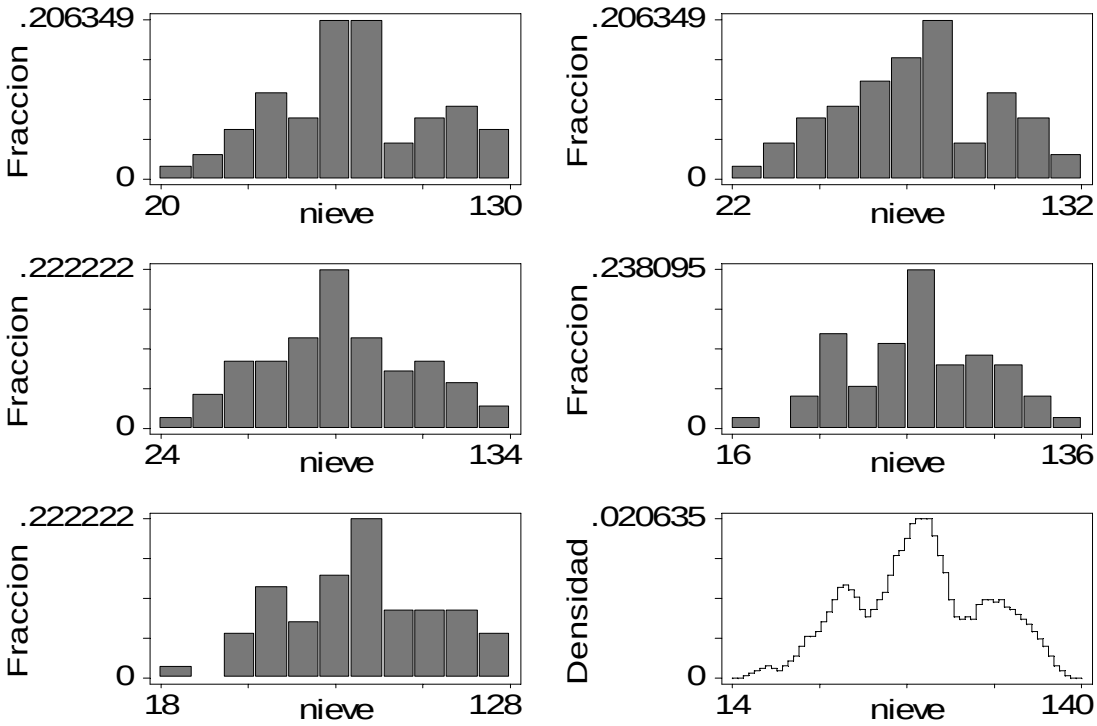

Figura 3.1 Histogramas para los datos de precipitación de nieve con diferentes orígenes, la misma amplitud de intervalo (h = 10) y el histograma desplazado promediado correspondiente.

Algunos de estos histogramas son unimodales, otros bimodales e inclusive trimodales. En todos ellos parece existir una moda alrededor de 80; en otros se despliegan modas secundarias alrededor de 50 o 100. Podemos escoger alguno de ellos para representar la distribución de los datos, pero la elección sería arbitraria. Para eliminar el efecto de la elección del origen, Scott (1985b) sugirió un ingenioso artificio: en lugar de escoger entre varios histogramas, Scott propuso promediar varios histogramas para obtener el **histograma desplazado promedio** (**HDP**). Considerando la definición de histograma como:

$$\hat{f}(x)=\frac{\upsilon_k}{nh}=\frac{1}{nh}\sum I_{[t_k,t_{k+1})}(x_i) \quad \text{para } x\in B_k \tag{1}$$

donde υ_k representa la frecuencia del *k* intervalo. Además, si consideramos una colección de *M* histogramas, $f_1, f_2, \ldots f_M$,, cada uno con amplitud de intervalo *h*, y orígenes en

$$t_0 = 0, \frac{h}{M}, \frac{2h}{M}, \ldots, \frac{(M-1)h}{M} \quad \text{.....(2)}$$

entonces el histograma desplazado promediado simple (no ponderado) se define como:

$$\hat{f}(\bullet) = \hat{f}_{\text{HDP}}(\bullet) = \frac{1}{M}\sum_{i=1}^{M} \hat{f}_i(\bullet) \qquad (3)$$

En este punto es conveniente referirse al intervalo más angosto $[k\delta, (k+1))\delta)$ donde $\delta = h/M$ como *Bk* y definir υk = cuenta del intervalo *Bk*, donde *Bk* = $[k\delta, (k+1))\delta)$.

Entonces, la expresión para el HDP (Scott, 1992) puede generalizarse como:

$$\begin{aligned} f(x;M) &= \frac{1}{M}\sum_{i=1-M}^{M-1} \frac{(M - |i|)\upsilon_{k+1}}{nh} \\ &= \frac{1}{nh}\sum_{i=1-M}^{M-1} \left(1 - \frac{|i|}{M}\right)\upsilon_{k+1} \quad \text{para } x \in B_k \end{aligned} \qquad (4)$$

Considerando que se calcula a partir de una serie de histogramas, el **HDP** tiene una apariencia de histograma (última gráfica de la Figura 3.1), aunque en la práctica puede hacerse continuo utilizando esquemas de interpolación lineal. El resultado en este caso es el **polígono de frecuencia del histograma desplazado promedio (PF-HDP)**. Esta operación de promedio hace que la estimación de densidad sea independiente de la elección del origen. El método del HDP es un caso especial del concepto más general desarrollado por Härdle y Scott (1988): el **promedio ponderado de puntos redondeados** ó **PPPR**. La expresión general para el **HDP**, y por tanto para el **PPPR** es (Scott, 1992; Härdle, 1991):

$$\hat{f}(x;M) = \frac{1}{nh}\sum_{|i|<M} w_M(i)\upsilon_{k+1} \quad \text{para } x \in B_k \qquad (5)$$

donde los pesos generales se definen por:

$$w_M(i) = M \times \frac{K(i/M)}{\sum_{j=1-M}^{M-1} K(j/M)} \qquad i = 1-M, \ldots, M-1 \qquad (6)$$

expresión en la cual *K* es una función continua definida en (-1,1). A menudo se escoge a *K* a una función de densidad tal como el kernel biponderado o cuártico (Scott, 1992).

De esta forma, el **PPPR** está basado en los intervalos más angostos (B_k) los cuales están definidos por *h*, la amplitud de banda y un nuevo parámetro *M*, que es el número de histogramas desplazados para promediar. Los puntos redondeados son la frecuencia de los intervalos en esta

trama de intervalos menores con longitud $\delta = h/M$. La operación de ponderación se simboliza por $w_{M()}$ y la especificación de esta función permite aproximar varios estimadores de densidad (por kernel).

Los cálculos del **PPPR** requieren de los siguientes pasos:

1. Agrupar a los datos
2. Crear los pesos
3. Ponderar los intervalos

En el primer paso se crea una trama de intervalos y se cuenta el número de observaciones que pertenecen a cada intervalo. La información acerca de los datos se reduce a un conjunto de cuentas de intervalos y sus correspondientes índices.

Posteriormente se calcula una función ponderal simétrica no negativa que sume la unidad. Por ejemplo, para el **HDP**:

$$w_{M(i)} = 1 - (|i| / M). \quad (7)$$

Y finalmente, alrededor de cada intervalo no vacío la estimación de densidad se incrementa por el producto de la frecuencia de ese intervalo y los pesos.

Esta es tan sólo una breve descripción del proceso. Para mayores detalles ver Härdle (1991) ó Scott (1992).

3.3 Implementaciones computarizadas de HDP-PPPR

La eficiencia de las estimaciones de densidad por **PPPR** reside en la reducción de los datos de un conjunto de observaciones a un conjunto de frecuencias e índices. Scott (1992) y Härdle (1991) presentan varios algoritmos, programas (FORTRAN y C) y macros (en lenguaje S) para calcular estimadores de densidad por **PPPR** y otros. Además, Brian Ripley ha mantenido una colección completa de programas en C y funciones en S del libro de Härdle, todas disponibles en la Internet de la colección de programas Statlib y recuperables por correo electrónico o FTP). Con base en estas rutinas, en este libro se incluyen varios programas en Turbo Pascal, Stata y un ejecutable a partir de código en Visual Basic para realizar estimación por **PPPR**. Estas rutinas se explican en el Capítulo 7.

Estos programas permiten explorar gráficamente a los datos utilizando las combinaciones deseadas de amplitud de ventanas (h) y número de histogramas a promediar (M). Es posible calcular estimaciones de densidad por **PPPR** y presentarlas gráficamente en forma de histograma (versión escalonada) o en forma de interpolación lineal para conectar los puntos estimados en forma similar a un polígono de frecuencia (versión polígono). Además del desplegado gráfico de la estimación de densidad es posible obtener los resultados numéricos.

La última gráfica de la Figura 3.1 es el **HDP** de los datos de precipitación de nieve con una amplitud de banda de 10, función ponderal triangular y promedio de cinco histogramas (combinación que produce 64 pares de valores de puntos de estimación y densidad. Esta estimación no depende más del origen y sugiere una estructura trimodal para la densidad desconocida.

En la Figura 3.2 se presentan estimaciones de densidad por **PPPR** utilizando 1, 2, 4, 8, 16 y 32 valores para M con $h = 13.6$. El histograma ordinario ($M = 1$) muestra dos modas. Cada uno de los **HDP**'s con $M > 2$ revela la presencia de tres modas. La apariencia de estas modas adicionales no es un artefacto del procedimiento **PPPR**, sino mas bien el resultado de una razón mejorada de señal/ruido obtenida por el promedio del origen. En lugar del parámetro origen x_0, se ha agregado otro: el número de histogramas desplazados promediados M. Esta adición se justifica por la mejora resultante sobre el histograma tradicional (Scott, 1992; Härdle, 1991). Además, en la misma figura notamos que desde $M \geq 4$, las estimaciones de densidad son esencialmente iguales.

3.4 Estimadores de densidad por kernel y por PPPR

Los estimadores de densidad por kernel pueden aproximarse por medio del **PPPR** especificando las funciones ponderales adecuadas. La función ponderal del procedimiento **PPPR** aproxima a la función kernel al aumentar *M*, el número de pasos (Härdle, 1991). En las Figuras 3.3 y 3.4 se incluyen las densidades estimadas por **PPPR** (versiones escalonada y poligonal respectivamente) junto con estimaciones de densidad por kernel. Claramente basta con escoger $M \geq 5$ para obtener un resultado que es esencialmente igual a la estimación por kernel (dependiendo de δ, el número de sub-intervalos). Las estimaciones que emplean interpolación lineal se aproximan más rápidamente a la estimación por kernel (Figura 3.4).

En este punto, para realizar las comparaciones entre las estimaciones por **PPPR** y kernel fue necesario ajustar algunos de los programas para EDK's presentados en Salgado-Ugarte, *et al.* (1993) y utilizándolos como base para escribir nuevas versiones. Estos programas ajustados y nuevos se describen en el Capítulo 7.

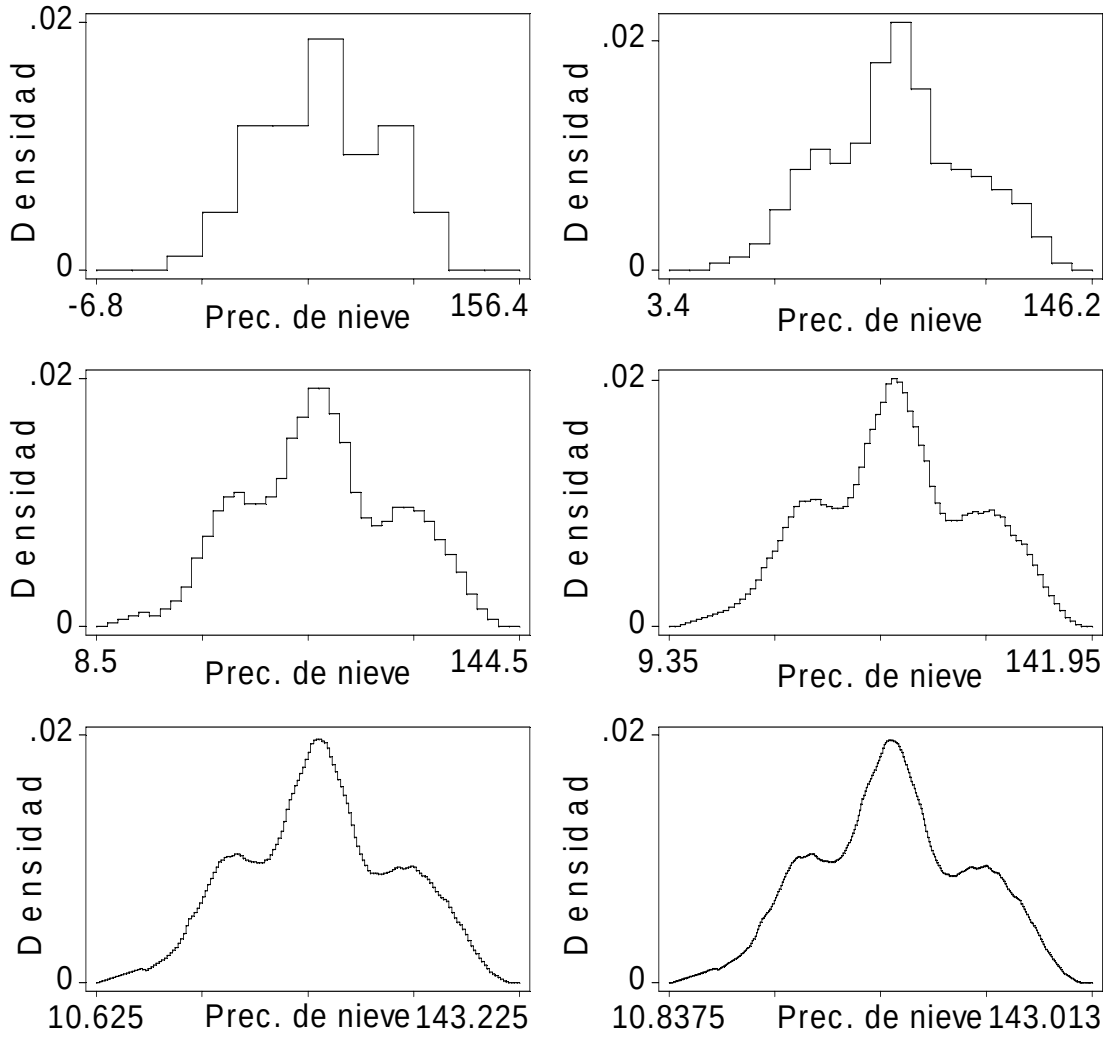

Figura 3.2 Histogramas desplazados promediados con valores crecientes de *M* (1, 2, 4, 6, 8, 16 y 32) con amplitud de intervalo (*h* = 13.6).

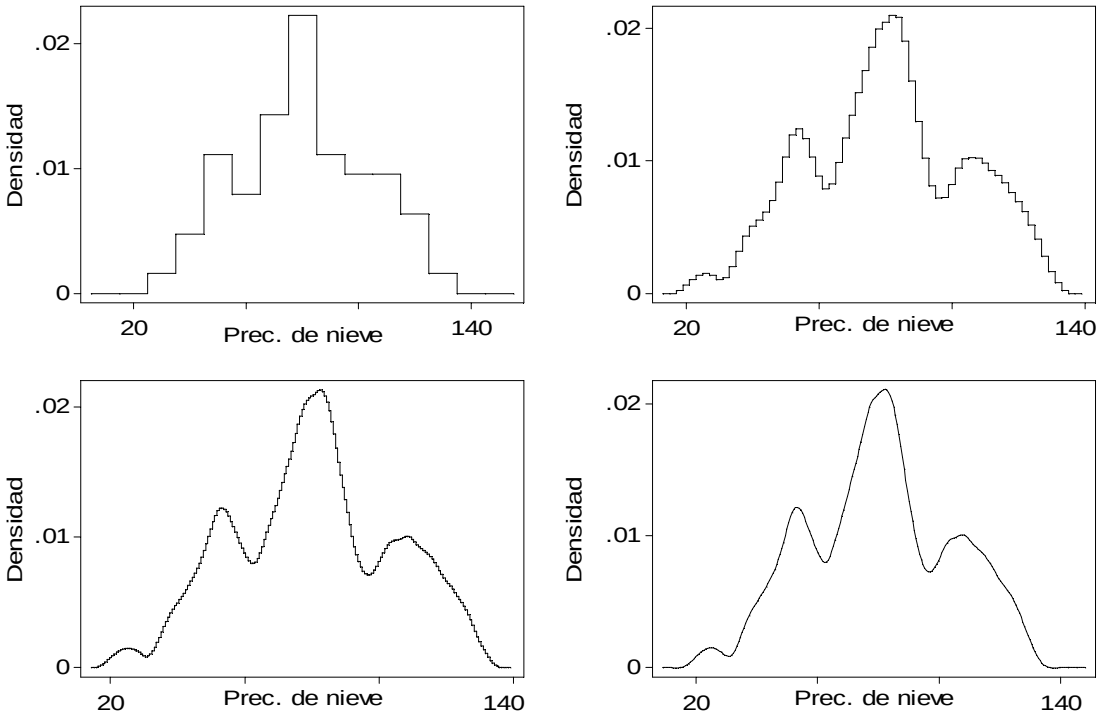

Figura 3.3 Comparación de la estimación de densidad por PPPR (versión escalonada) utilizando *M* = 1, 5, 15 y el estimador de densidad por kernel cuártico. La amplitud de banda es 10.

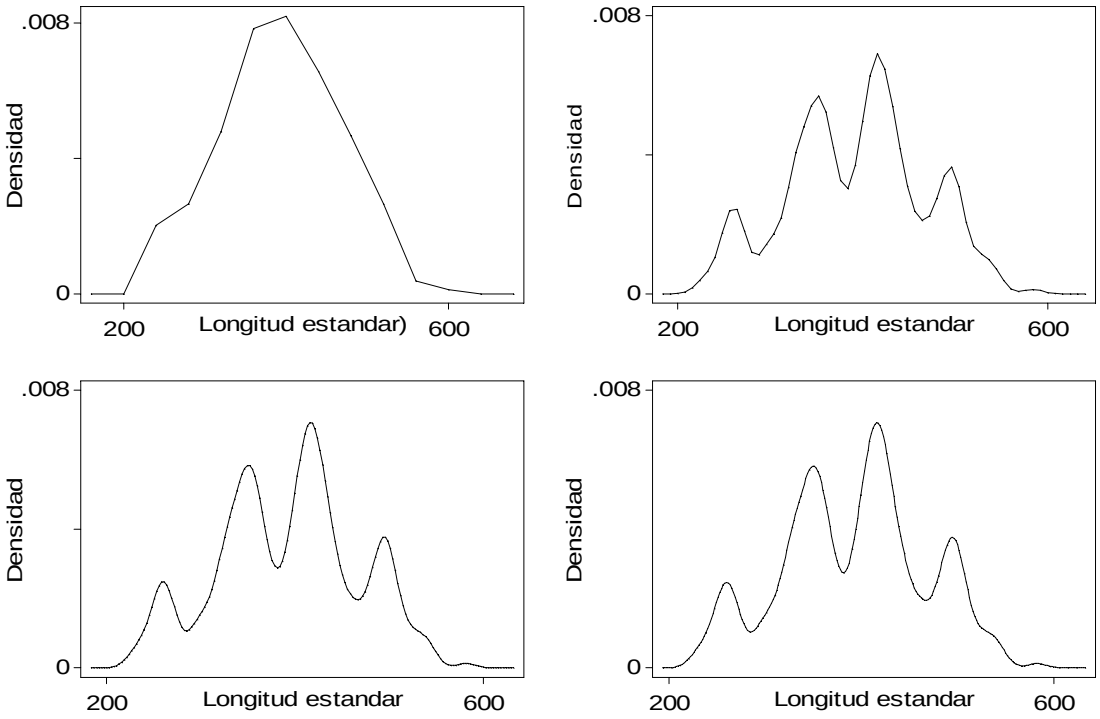

Figura 3.4 Comparación de la estimación de densidad por PPPR (versión poligonal) utilizando *M* = 1, 5, 15 y el estimador de densidad por kernel Gaussiano. La amplitud de banda es 10.

Estos programas calculan los kerneles uniforme, triangular, Epanechnikov, cuártico (biponderado), triponderado, Gaussiano y coseno. Las nuevas versiones consideran una trama de 50 puntos uniformemente espaciados desde $x_{(1)} - h -$ (recorrido×0.1) hasta $x_{(n)} + h +$ (recorrido×0.1). De acuerdo con Härdle (1991), quien sigue la sugerencia de Gasser *et al.* (1985), sólo se pueden comparar estimaciones de densidad por kernel correspondientes a funciones ponderales con el mismo soporte. Estos últimos autores sugirieron el uso del intervalo [-1,1] como un estándar. Tomando en cuenta estas consideraciones, las ecuaciones para los algoritmos arriba mencionados fueron modificados como se indica en el cuadro 3.1:

Cuadro 3.1 Expresiones kernel modificadas

Kernel	K(z)
Uniforme (kernsim.ado)	$\frac{1}{2}\, I(\|z\| \leq 1)$
Triangular (ASH)	$(1 - \|z\|)\, I(\|z\| \leq 1)$
Epanechnikov	$\frac{3}{4}(1 - z^2)\, I(\|z\| \leq 1)$
Cuártico	$(15/16)(1 - z^2)^2\, I(\|z\| \leq 1)$
Triponderadot	$(35/32)(1 - z^2)^3\, I(\|z\| \leq 1)$
Coseno	$(\pi/4)\cos((\pi/2)z)\, I(\|z\| \leq 1)$
Gaussiano	$(1/\sqrt{2\pi})\exp((-1/2)z^2)$

Para hacer comparaciones directas entre estimaciones se consideraron los datos de precipitación de nieve en Buffalo y los de longitud de trucha coralina previamente analizados en la Sección 1.2. En la Figura 3.3. la versión escalonada por **PPPR** con función ponderal cuártica (biponderada) con M = 1, 5, y 15 se compara con la estimación de densidad por kernel cuártico, ambas con amplitud de banda h = 10. En la Figura 3.4 se presenta la estimación de densidad por **PPPR** (versión poligonal) utilizando la función ponderal Gaussiana (M = 1, 5 y 15) comparada con el EDK Gaussiano ambas con h = 10. Podemos apreciar de estas dos gráficas que las estimaciones por **PPPR** con $M \geq 5$, son indistinguibles de la correspondiente estimación por kernel. En la Figura 3.4 es posible notar que la estimación por **PPPR** utilizando M = 1 no puede detectar la multimodalidad de los datos, pero una vez que el número de histogramas se incrementa las varias modas aparecen claramente.

3.5 Kerneles equivalentes

Cuando comparamos dos tipos diferentes de kernel que utilizan la misma amplitud de ventana encontramos que los resultados no son los mismos. Consideremos como ejemplo la Figura 3.5 que presenta estimaciones de densidad por kernel triponderado y Gaussiano de los datos de longitud para la trucha coralina utilizando una $h_G = h_T = 15$. La estimación triponderada es menos suave (sugiriendo un número mayor de modas) que la Gaussiana.

En general, la causa de esta diferencia es que a pesar de tener el mismo intervalo de soporte, cada uno de los kerneles tienen varianzas diferentes. Por esto, un enfoque para obtener un grado de suavización similar utilizando diferentes funciones ponderales es ajustar la amplitud de banda para obtener la misma cantidad de variación. A este respecto Scott (1976) dio las bases para los cálculos de factores para amplitudes de banda con suavización independiente.

El cuadro 3.1 (Modificado de Härdle, 1991 y Scott, 1992) da un resumen de tales factores de interconversión entre kerneles populares:

Cuadro 3.1 Algunos factores de conversión para kerneles comunes

A (fila) / desde (col.)	Uniforme	Triangular	Epanech.	Cuártico	Triponderado	Coseno	Gaussiano
Uniforme	1.000	0.715	0.786	0.663	0.584	0.761	1.740
Triangular	1.398	1.000	1.099	0.927	0.817	1.063	2.432
Epanechnikov	1.272	0.910	1.000	0.844	0.743	0.968	2.214
Cuártico	1.507	1.078	1.185	1.000	0.881	1.146	2.623
Triponderado	1.711	1.225	1.345	1.136	1.000	1.302	2.978
Coseno	1.315	0.941	1.033	0.872	0.768	1.000	2.288
Gaussiano	0.575	0.411	0.452	0.381	0.336	0.437	1.000

Al aplicar el factor correspondiente, podemos obtener aproximadamente el mismo grado de suavización con los kerneles triponderado y Gaussiano. Por lo tanto, de acuerdo al cuadro, se puede obtener casi la misma suavización con un kernel Gaussiano si multiplicamos por 2.978 la banda para kernel triponderado. La Figura 3.6 muestra al kernel Gaussiano con $h_G = 15$ y un kernel triponderado con $h_T = (h_G \times 2.978) = 15 \times 2.978 = 44.67$. Ahora, ambas estimadores de densidad muestran grados semejantes de suavización.

Otro enfoque cercanamente relacionado (más general) es la **transformación canónica de banda** propuesta por Marron y Nolan (1988) en la cual las funciones ponderales por sí mismas se escalan a su forma “canónica” permitiendo el uso de las mismas amplitudes de banda considerando como referencia al kernel Gaussiano (Scott, 1992; Marron y Nolan, 1988).

3.6 Algunas comparaciones del desempeño de programas para PPPR y EDK's

Como se documentó anteriormente (Capítulo 2 y Salgado-Ugarte, *et al.*, 1993), aún con computadoras veloces la estimación directa de los estimadores de densidad por kernel tiene un lento desempeño, siendo esta lentitud dependiente en *n*, el tamaño del lote de datos. Como se citó arriba, la técnica **PPPR** utiliza una discretización de los datos y por lo tanto, los cálculos dependen de *n* tan sólo en la etapa de agrupamiento. Puesto que los resultados de cualquier procedimiento computarizado tienen que ser discretizados de cualquier forma al ser desplegados en una pantalla o en una impresora, no existe pérdida de información (Härdle, 1991). Härdle y Scott (1988) presentan una discusión adicional acerca de la eficiencia computacional de las estimaciones de densidad por kernel y **PPPR.** A continuación se presentan algunos tiempos aproximados (medidos desde la entrada de la función kernel) de los programas presentados en esta obra efectuados en una computadora con procesador 486SX (25 MHz) con 6 Mb de RAM, sin coprocesador matemático y la versión regular del paquete estadístico Stata (Cuadro 3.2). El archivo de datos nombrado catfish.dta contiene los datos de longitud estándar del bagre estuarino *Cathorops melanopus* (Salgado-Ugarte, 1985).

Cuadro 3.2 Comparación de desempeño de estimadores de densidad por kernel discretizadas y HDP y PPPR

Archivo de datos	*n*	Programa	Número de puntos de estimación	Tiempo (en segundos)
trocolen.raw	316	Warpstep (*h*=20, *M*=15, Epanechnikov)	303	5
trocolen.raw	316	Warpstep (*h*=20, *M*=15, Gaussiano)	303	4
trocolen.raw	316	Warpstep (*h*=20, *M*=50, Epanechnikov)	1004	13
trocolen.dta	316	Kernsim	50	150
trocolen.dta	316	Kerngaus	50	170
catfish.raw	2439	Warpstep (*h*=20, *M*=5, Cuártico)	70	24
catfish.raw	2439	Warpstep (*h*=20, *M*=15, Cuártico)	204	26
catfish.dta	2439	Kernquar	50	1200

Como podemos observar en este cuadro, utilizando el **PPPR** se obtienen ahorros de tiempo impresionantes aún con *n* y número de intervalos en el orden de miles. Se recomienda un número de 10 a 15 intervalos para fines generales de exploración de la distribución de los datos.

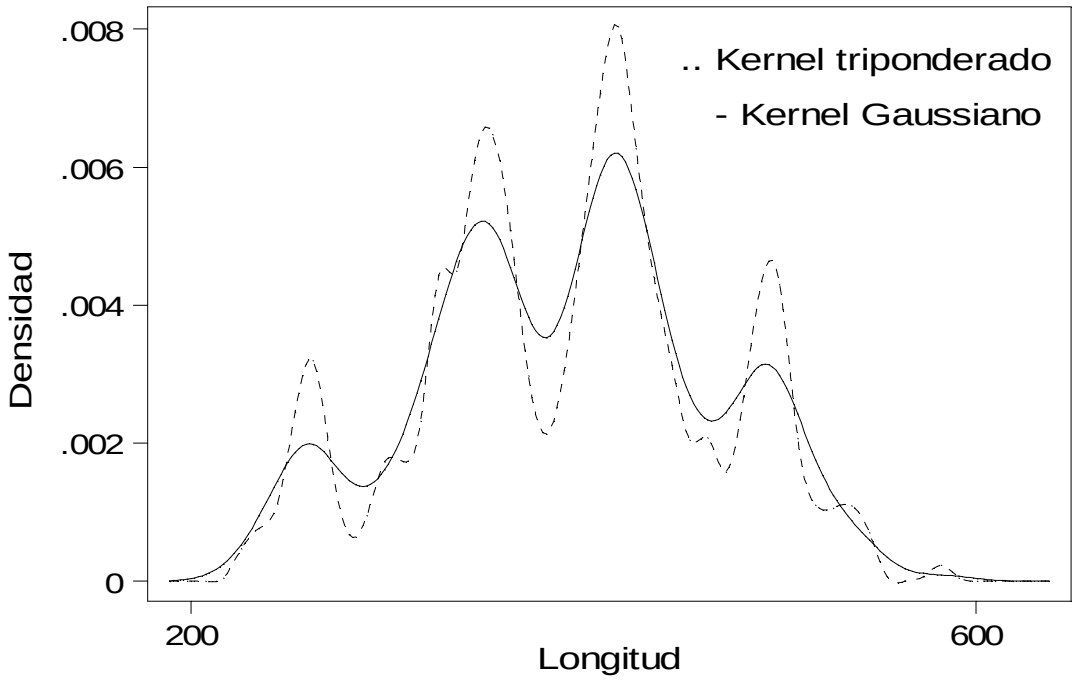

Figura 3.5. Estimación de densidad por kernel Gaussiano (línea continua) y triponderado (línea punteada) para los datos de longitud de la trucha coralina. La amplitud de banda $h_G = h_T = 15$.

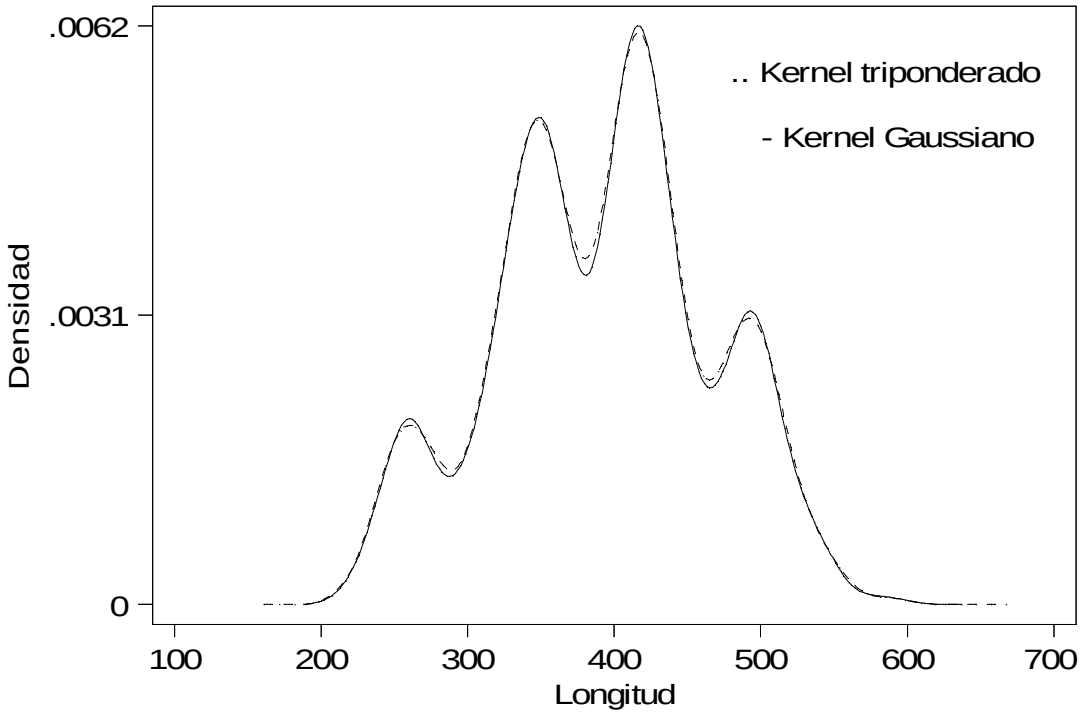

Figura 3.6 Estimación de densidad por kernel Gaussiano (línea continua), $h_G = 15$ y triponderado (línea punteada), $h_T = 44.67$, para los datos de longitud de la trucha coralina.

Capítulo 4. Reglas prácticas para selección de intervalo/banda en estimación univariada de la densidad

4.1. Introducción

La elección de la amplitud del intervalo/banda (parámetro de suavización) es uno de los problemas más relevantes de la estimación de la densidad. Como se discutió en los capítulos pasados, existen varios procedimientos para establecer un valor adecuado para este parámetro en histogramas, polígonos de frecuencia (**PF**), histogramas desplazados promedio (estimadores **HDP-PPPR**) y estimadores de densidad por kernel (**EDK**'s). Algunos de estos métodos de selección se enfocan al número "óptimo" de intervalos, mientras que otros producen aproximaciones a la amplitud óptima de intervalo al minimizar alguna medida del error estadístico bajo ciertas condiciones y suposiciones. En esta sección se presentan varias reglas que proporcionan algunos valores de referencia útiles para la elección del parámetro de suavización en el análisis de la densidad por histogramas, polígonos de frecuencia y estimadores de densidad por kernel, incluyendo a los histogramas desplazados promedio (**HDP**) y su generalización en el promedio ponderado de puntos redondeados (**PPPR**). Los programas para calcular estas reglas han sido incluidas en el Capítulo 7, donde además, contiene una nueva versión integrada de los programas para estimación de densidades univariadas (estimadores por kernel y **HDP-PPPR**).

4.2. Reglas para elección de número de intervalos en histogramas

Probablemente la regla más famosa para determinar el número de intervalos para la estimación de densidad del histograma fue propuesta por Sturges (1926). Al aplicar esta regla se espera que una variable normalmente distribuida pueda ser dividida apropiadamente de tal forma que las clases de frecuencia abarquen una serie binomial para todos los valores de *N* los cuales son potencias pares de dos (Doane, 1976). Es interesante notar que no obstante en su contribución original Sturges se refiere explicitamente a la elección del intervalo de clase, técnicamente su regla es mas bien un procedimiento para escoger el número de intervalos (Scott, 1992). De acuerdo a la sugerencia de Sturges, el número de intervalos puede calcularse como:

$$k = 1 + \log_2 n \qquad (1)$$

donde *k* es el número de intervalos, y *n* se refiere al número de observaciones.

La formulación de Sturges es ampliamente recomendada en textos introductorios de estadística. Se ha vuelto una guía para los investigadores y a menudo es utilizada como valor pre-establecido en programas estadísticos aún cuando en varios casos no es apropiada: esta regla no es aplicable a datos con distribución asimétrica, no Gaussiana o multimodal, los cuales ocurren la mayor parte del tiempo (Doane, 1976; Scott, 1992). Una forma de ajustar por sesgo es agregar intervalos en (1) de acuerdo a $\log_2(1+\hat{\gamma}\sqrt{n/6})$, $\hat{\gamma}$ es un estimador del coeficiente de sesgo estándar (Doane, 1976). Para fines exploratorios, Emerson y Hoaglin (1983) consideran que este ajuste adicional involucra muchos cálculos que podrían ser problemáticos sin una computadora. Una dificultad más seria es la no resistencia del coeficiente de sesgo.

4.3. Reglas para elección de amplitud de intervalo en histogramas

Como se ha citado en otros trabajos (Salgado-Ugarte, *et al.* 1993), Scott (1979) derivó una fórmula para calcular la amplitud óptima asintótica resultante en un error cuadrado integrado medio mínimo (***ECIM***) para histogramas. La fórmula de Scott involucra el conocimiento previo de la verdadera función de densidad, un evento más bien raro en el análisis de datos en la realidad. Por lo tanto, adoptando a la densidad Gaussiana como referencia, el propuso:

$$\hat{h} = 3.5\,\hat{\sigma}\; n^{-1/3} \qquad (2)$$

donde $\hat{h}$ es la amplitud de banda estimada y $\hat{\sigma}$ es una estimación de la desviación estándar de los datos.

En su contribución original, Scott, además analizó el desempeño de esta regla cuando se aplica a tres distribuciones no Gaussianas adoptadas como modelos de referencia.

- Sesgada (densidad log-normal)
- Colas potentes (*t* de Student)
- Bimodal (mezcla de dos distribuciones Gaussianas)

A partir de las simulaciones realizadas, este autor concluyó que la regla de referencia Gaussiana produce un parámetro de suavización que:

- Sobresuaviza a una distribución log-normal, pero para índices de sesgo tan grandes como uno, la diferencia con la amplitud de intervalo óptima verdadera es menos del 30 %.
- Es insensible a curtosis moderada.
- Sobresuaviza datos bimodales cuando la distancia entre las modas es mayor que 2. Con datos claramente bimodales, esta regla no es adecuada (Scott, 1979).

Recientemente, este autor ha proporcionado factores de corrección para *h* tomando en cuenta al sesgo y la curtosis (Scott, 1992).

Una regla más robusta ha sido propuesta por Freedman y Diaconis (1981a,b), en la que se reemplaza la estimación de la desviación estándar por un múltiplo del recorrido intercuartílico (***RIC***). La regla F-D se define como:

$$\hat{h} = 2\,(RIC)\,n^{-1/3} \qquad (3)$$

Varios autores han comparado el desempeño de las reglas (1), (2) y (3) (Emerson y Hoaglin, 1983; Scott, 1992; ver también la nota técnica en el Manual de Referencia del paquete estadístico Stata, 1995) llegando a las siguientes generalizaciones:

- Las reglas de Scott y Sturges concuerdan cercanamente para tamaños de muestra entre 50 y 500.
- Para tamaños de muestra mayores, la regla de Sturges da muy pocos intervalos (produciendo una sobresuavización que desperdicia mucha de la información de los datos)
- En general, las densidades no Gaussianas requieren de más intervalos.
- La regla F-D resulta en intervalos más angostos (35% más intervalos que la regla de Scott).
- Desde un punto de vista exploratorio, la característica más interesante de las reglas de Scott y F-D es que ambas dependen principalmente de $n^{-1/3}$. Considerando el correspondiente número de intervalos, se comportaría como $n^{1/3}$, una forma funcional intermedia entre log(*n*) y $\sqrt{n}$.

4.4. Reglas sobresuavizadas

Las reglas anteriormente descritas proporcionan un punto de partida simple y útil para su afinación posterior, si bien no son la respuesta última al problema. Recientemente la búsqueda de procedimientos basados en los datos para minimizar el ***ECIM*** (o cantidades relacionadas como el error cuadrado integrado medio asintótico ***ECIMA***) es el tema de considerables estudios. Los procedimientos descritos a continuación son algunos ejemplos de esta investigación (una explicación más amplia es dada por Scott, 1992).

Terrell y Scott (1985) mostraron que dependiendo del conocimiento de la escala de la densidad desconocida con base en los datos, existen límites superiores para la amplitud de intervalo de los histogramas. Por otra parte, en principio, no existe un límite inferior para *h* ya que la

densidad desconocida puede ser arbitrariamente ruidosa. Considerando este límite superior estos autores arribaron a la siguiente expresión:

$$h_{OS} \equiv \frac{b-a}{\sqrt[3]{2n}} \geq h_O \tag{4}$$

Donde h_O es la amplitud de intervalo óptima, h_{OS} es el límite superior para h (la amplitud sobresuavizada de intervalo) y $b - a$ es el recorrido de la muestra. Rearreglando (4) es posible obtener una regla para número sobresuavizado de intervalos:

$$\text{Número sobresuavizado de intevalos} = \frac{b-a}{h_O} \geq \frac{b-a}{h_{OS}} = \sqrt[3]{2n} \tag{5}$$

Eligiendo una amplitud de intervalo igual o mayor que h_{OS} producirá una estimación sobresuavizada así como utilizando un número de intervalos menor o igual al dado por (5).

Como concluyen Terrell y Scott (1985), las expresiones (4) y (5) dan resultados casi óptimos para una amplia variedad de densidades y proporcionan buenas estimaciones de densidad.

Estas reglas son muy simples y tienen valor mnemónico. Están encaminadas al hallazgo de un límite inferior para el número de intervalos. Sin embargo, en la estimación de la densidad el foco es encontrar al parámetro de suavización h_O mas bien que el número de intervalos. Una solución en este contexto fue propuesta por Terrell (1990) en su *Principio de Suavización Máxima*. Restringiendo a la varianza como fija su solución es:

$$h_{OS} \text{ homoscedástica } \equiv 3.729\sigma n^{-1/3} \geq h_O \tag{6}$$

La versión que utiliza al ***RIC*** es particularmente robusta, y por tanto especialmente útil en el caso de datos sesgados:

$$h_{OS} \text{ homoscedástica robusta } \equiv 2.603(IQR)n^{-1/3} \geq h_O \tag{7}$$

Las conservativas estimaciones sobresuavizadas proporcionadas por estas reglas pueden conducir a evitar aseveraciones prematuras sobre la estructura en las muestras. Por el contrario, la ocurrencia de características estructurales en estimaciones conservadoras representa una evidencia sólida de su existencia. Si tales estructuras no ocurren en estimaciones sobresuavizadas, entonces pruebas más especializadas o muestras adicionales se requieren para establecer su ocurrencia (Terrell, 1990).

4.5. Reglas para elección de número y amplitud de intervalos en polígonos de frecuencia

A pesar de comentarios iniciales (Fisher, 1932, 1958) desfavorables para el interpolante lineal de las marcas de clase de un histograma con amplitud uniforme de intervalo (es decir el polígono de frecuencia), el trabajo de Scott (1985a) acerca de las propiedades teóricas de polígonos de frecuencia (***PF***) uni y bivariados ha demostrado que produce estimaciones notablemente mejores en

comparación con el histograma (Scott, 1992). Los resultados mejorados producidos por el PF facilita substancialmente la localización de mínimo en ***ECIMA***, y por tanto en la estimación del número y amplitud de intervalo óptimas.

En contraste con el histograma puede afirmarse que el ***PF***:

- aproxima mejor a densidades continuas mediante interpolación lineal con intervalos más amplios.
- es menos eficiente cuando la densidad subyacente es discontinua.
- es más sensitivo con respecto a los errores en la elección de amplitud de intervalo, particularmente cuando $h > h_O$; por otra parte, se requiere un error considerablemente grande en la amplitud de intervalo del ***PF*** antes de que su ***ECIM*** sea peor que el del mejor histograma (Scott, 1992).

Estas diferencias se reflejan en la resultante regla de referencia Gaussiana para el ***PF***:

$$\text{Regla de referencia Gaussiana para PF:} \quad \hat{h} = 2.15\,\hat{\sigma}\, n^{-1/5} \tag{8}$$

La estimación de la desviación estándar puede ser una robusta, como ***RIC***/1.349 (la Pseudosigma). Esta regla puede modificarse también tomando en cuenta factores de ajuste por sesgo y curtosis (Scott, 1992).

Como en el caso de los histogramas es posible definir límites inferiores para la amplitud de intervalo o límites superiores para el número de clases. La expresión para número sobresuavizado de intervalos para ***PF*** equivalente a la (5) de histogramas es:

$$\text{Número sobresuavizado de intervalos para PF} \quad = \frac{b-a}{h_O} \geq \left(\frac{147}{2} n\right)^{1/5} \tag{9}$$

Una versión diferente para el problema de sobresuavización conduce a la correspondiente regla para amplitud de intervalo:

$$h_{OS} \equiv 2.33\,\sigma\, n^{-1/5} \geq h_O \tag{10}$$

La pequeña diferencia entre la regla sobresuavizada para PF y la regla óptima Gaussiana para ***PF*** sugiere que la regla sobresuavizada para ***PF*** puede ser utilizada en lugar de una regla óptima Gaussiana cuando sea difícil resolver explícitamente el problema variacional (Scott, 1992).

4.6. Reglas para elección de amplitud de banda en EDK's

En su monografía sobre estimación de densidad, Silverman (1986) discute varias reglas para elegir *h* al trabajar con estimadores de densidad por kernel (***EDK***). Uno es el *método gráfico de prueba* (Silverman, 1978) el cual consiste en dibujar la segunda derivada de la estimación de densidad ($\hat{f}$) para varios valores de *h* y después elegir la amplitud de banda correspondiente a la gráfica con "fluctuaciones rápidas bien definidas que no ocultan por completo la variación sistemática". Aunque algo de subjetividad está involucrada, en la práctica parece que las gráficas de prueba

amplifican la variación de la estimación de densidad y de tal forma la elección de una amplitud de banda apropiada no es muy difícil. No obstante, el mismo autor reconoce que la subjectividad de este procedimiento origina que su uso principal sea una forma de verificar los resultados con otros métodos.

Además al método gráfico bosquejado arriba, Silverman propuso utilizar una distribución estándar como referencia (de forma semejante a como hizo Scott, 1979 para los histogramas). Si se emplea un kernel Gaussiano, la amplitud óptima de banda es estimada por medio de:

$$\hat{h} = 1.06\,\hat{\sigma}\,n^{-1/5} \qquad (11)$$

Este autor compara el resultado de su expresión con distribuciones no Gaussianas arribando a conclusiones semejantes a las de Scott (1979): la amplitud de banda resultante de (11) sobresuaviza datos marcadamente sesgados, es notablemente insensible a la curtosis (usando a distribuciones log-normal, y de familia *t* como modelos), y sobresuaviza más y más al volverse la distribución más bimodal. El sugirió usar una medida robusta de la dispersión tal como el recorrido intercuartílico:

$$\hat{h} = 0.79 R n^{-1/5} \qquad (12)$$

Esta expresión da mejores resultados in distribuciones sesgadas y de cola larga pero para el caso bimodal sobresuaviza en forma adicional. Más adelante el sugiere utilizar una estimación adaptativa para la dispersión en lugar de $\hat{\sigma}$ en la ecuación (11):

$$A = \min(\hat{\sigma}, RIC/1.349) \qquad (13)$$

lo que resulta en:

$$\hat{h} = 1.06 A n^{-1/5} \qquad (14)$$

Härdle prefiere esta regla óptima adaptativa (su “regla mejorada del pulgar” en Härdle, 1991).

Puede valer la pena recordar que el ***RIC*** ≈ *Dispersión de los cuartos* (definida como en Tukey, 1977 y Hoaglin, 1983) dependiendo del método para el cálculo de los cuartiles (ver detalles en Frigge, *et al.*, 1989 ó Hamilton, 1992). En el paquete estadístico Stata, la diferencia entre el ***RIC*** y la *dispersión de los cuartos* disminuye gradualmente hasta desaparecer al aumentar el tamaño de muestra (Hamilton, 1992).

Silverman sugiere la aplicación de un ajuste adicional al reducir el factor 1.06 en la ecuación (7) hasta 0.9 de la forma siguiente:

$$\hat{h} = 0.9 A n^{-1/5} \qquad (15)$$

De acuerdo a sus simulaciones, para un kernel Gaussiano, esta regla proporcionará un ***ECIM*** dentro del 10% del valor óptimo para las distribuciones de cola larga, asimétricas y bimodales que usó como referencia (Silverman, 1976).

4.7. Reglas sobresuavizadas para EDK's

Scott (1992), basado en trabajo previo (Scott y Terrell, 1987; Terrell, 1990), y considerando a la variana como medidor de escala, derivó la siguiente regla de sobresuavización para los estimadores de densidad por kernel:

$$h_{OS} = 3\left[\frac{R(K)}{35\sigma_K^4}\right]^{1/5} \sigma n^{-1/5} \qquad (16)$$

Donde *R(K)* es la "rugosidad del kernel" y σ_K^4 es la varianza cuadrada del kernel los cuales son constantes y característicos de cada kernel. El cuadro 4.1 contiene los valores de rugosidad y varianza para algunos kerneles comunes.

Utilizando este cuadro es posible calcular la regla de sobresuavización para los kerneles incluídos. De esta forma, para un kernel biponderado (cuártico) es exactamente $h_{OS} = 3\sigma n^{-1/5}$; para el kernel Gaussiano, $h_{OS} = 1.144\sigma n^{-1/5}$. Esta regla es ligeramente más amplia (8%) que la regla de referencia óptima Gaussiana, (11) citada anteriormente la cual usa un factor de 1.06.

Cuadro 4.1. Valores de Rugosidad y varianza para kerneles comunes (adaptado de Scott, 1992)

Kernel	*R(K)*	σ_K^2
Uniforme	1/2	1/3
Triangular	2/3	1/6
Epanechnikov	3/5	1/5
Biponderado	5/7	1/7
Triponderado	350/429	1/9
Gaussiano	$0.5/\sqrt{\pi}$	1
Coseno	$\sigma^2/16$	$1-8/\pi^2$

Nota: los kerneles están soportados en [-1,1] excepto el Gaussiano, de acuerdo a las ecuaciones de Härdle, 1991 y Scott, 1992.

4.8. Validación cruzada por mínimos cuadrados

La validación cruzada (***VC***) es un procedimiento bien conocido para la elección automática del parámetro de suavización. Existen dos tipos de ***VC***: ***VC por máxima verosimilitud*** y ***VC por mínimos cuadrados*** (***VCMC***). Fue sugerida por Rudemo (1982) y Bowman (1984), y está basada en una idea muy simple. Considerando al **Error Cuadrado Integrado** (***ECI***) como una medida de distancia (*d*) entre el estimador $\hat{f}$ de la densidad *f* puede definirse como:

$$d_I(h) = \int (\hat{f}_h - f)^2(x)dx$$

$$= \int \hat{f}_h^2(x)dx - 2\int (\hat{f}_h f)(x)\,dx + \int f^2(x)\,dx \qquad (17)$$

Observando la expresión (17), podemos notar que el primer término puede calcularse de los datos, el último término no depende de la estimación ni del parámetro de suavización (*h*), y sólo el término intermedio tiene que ser estimado.

El principio de la ***VCMC*** es minimizar el primero y segundo términos con respecto a *h*. Para lograr esto, la ***VCMC*** toma ventaja de la esperanza calculada con respecto a una observación *X* adicional e independiente:

$$\int (\hat{f}_h f)(x)\,dx = E_X[\hat{f}_h(X)]$$

Como por lo general un conjunto de datos adicional no está disponible, se define a la *estimación dejando uno fuera* como:

$$E_X\left[\hat{f}_h(X)\right] = n^{-1}\sum_{i=1}^{n}\hat{f}_{h,i}(X_i)$$

Utilizando las expresiones equivalentes anteriores arribamos a la expresión para la validación cruzada por mínimos cuadrados:

$$VC(h) = \int \hat{f}_h^2(x)dx - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{h,i}(X_i) \qquad (18)$$

Analizando las propiedades de (18) Scott y Terrell (1987) mostraron que es un criterio insesgado de validación cruzada. Härdle (1991) da un algoritmo de computación directa para la ***VCMC***, notando que el número de iteraciones es cuadrático en el número de observaciones, un inconveniente que motiva la búsqueda de otro método más eficiente de cálculo. A este respecto, Silverman (1986) propuso el uso de un algoritmo basado en la transformación rápida de Fourier. Scott y Terrell (1987) utilizaron un procedimiento modificado de ***HDP***. Basado en la generalización del *HDP*, o sea el promedio ponderado de puntos redondeados (***PPPR***), método originalmente propuesto por Härdle y Scott (1988), Härdle (1991) presenta un algoritmo eficiente para cálculo que requiere un número lineal de iteraciones en el número de observaciones.

4.9. Validación cruzada sesgada

Tomando un enfoque diferente, Scott y Terrell (1987) consideraron estimar directamente el error cuadrado integrado medio asintótico (***ECIMA***). Estos autores llegaron a encontrar un sesgo en la estimación utilizando la norma L_2 y por lo tanto lo llamaron *validación cruzada sesgada* (***VCS***). La expresión general para la ***VCS*** es:

$$VCS(h)=\frac{1}{nh}\|K\|_2^2+\frac{h^4}{4}\mu_2^2(K)\left[\|\hat{f}_h''\|_2^2-\frac{1}{nh^5}\|K''\|_2^2\right] \qquad (19)$$

Härdle (1991) da un conjunto completo de expresiones y un algoritmo para su cálculo.

Scott y Terrell (1987) compararon a la ***VCMC*** (un estimador insesgado) y a la ***VCS*** aplicadas a datos simulados encontrando que:

- Para muestras pequeñas (n = 25), aproximadamente la mitad de las funciones ***VCS*** estimadas no tuvieron mínimos locales, aunque para n > 40 todas las estimaciones los tuvieron.
- La ***VCS*** tuvo menor desviación estándar conduciendo por tanto a una estimación adecuada del ***ECIMA***.
- Si la densidad subyacente es asimétrica o con colas gruesas (Cauchy, lognormal o mezclas Gaussianas), la ***VCS*** tiende a sobresuavizar mientras que la ***VCMC*** a pesar de su mayor dispersión, en promedio, produce mejores estimaciones.

Estas generalizaciones dan alguna guía en la elección del procedimiento: si la naturaleza de la verdadera densidad no es simétrica, entonces se recomienda el empleo de la ***VCMC***; de otra forma, se aconseja el uso de la ***VCS*** (pero ver abajo).

Scott (1992) concluye que, las reglas de sobresuavización, la ***VCS*** y la ***VCMC*** son un poderoso conjunto de herramientas para la elección de la amplitud de intervalo de histogramas y **PF**'s así como para la selección de amplitud de banda en estimadores de densidad por kernel. Este autor recomienda el uso de un gráfico con escala logarítmica log(h) y sugiere evaluar los resultados buscando fallas (falta de mínimos locales para la ***VCS*** o h = 0 degenerada para la ***VCMC***); si las amplitudes de banda resultantes difieren, entonces escoger aquella que conduce a la estimación con la menor cantidad de ruido local, especialmente cerca de los picos de densidad. Los tres procedimientos deberán examinarse simultáneamente aún con muestras de tamaño muy grande.

4.10. Ejemplos de aplicación

Como puede verse en las ecuaciones para las reglas arriba incluidas, excepto para los selectores de amplitud de banda por validación cruzada, es sencillo estructurar algoritmos computarizados para su cálculo ya sea utilizando lenguajes de programación de alto nivel o lenguajes de macros de paquetes estadísticos (como Stata). En esta obra se incluye un programa en Stata para cálculo de estas reglas el cual utiliza comandos de Stata en combinación con otras funciones disponibles y que se presenta de manera detallada en el Capítulo 7.

El resultado de este programa es un cuadro que incluye (número de ecuación incluido):

- Reglas para número óptimo de intervalos en histogramas (1), (5)
- Regla para número sobresuavizado de intervalos en ***PF***'s (9)
- Reglas para amplitud óptima de intervalo en histogramas (2), (3), (4), (6) y (7)
- Reglas para amplitud de intervalo en ***PF***'s (8) y (10)
- Reglas para amplitud óptima de banda para kernel Gaussiano (15) y (14)
- Regla para amplitud sobresuavizada de banda para kernel Gaussiano (16)

Para mostrar la aplicación de este programa se utilizan los datos de precipitación de nieve (Parzen, 1979; Härdle, 1991; Scott, 1992) y los datos de longitud estándar de bagres (Salgado-Ugarte, 1985) introducidos brevemente en la sección 3.6 (ver también Salgado-Ugarte, *et al.*, 1995a).

El cuadro 4.2 contiene los resultados para los datos de precipitación de nieve. En la Figura 4.1 se incluyen el histograma y el polígono de frecuencia con la regla de amplitud óptima de intervalo para distribución Gaussiana. Tanto el histograma como el ***PF*** se calcularon con programas revisados para estimación de densidad por **HDP-PPPR** (ver abajo y Capítulo 7). Debido a que en la estimación de densidad por HDP-PPPR el origen de la trama utilizada para el cálculo no es de importancia al incrementarse el parámetro *M* (número de histogramas desplazados), estos programas no permiten su especificación. Es posible probar otros histogramas con diferentes orígenes utilizando otros procedimientos de cálculo. Observando a los resultados con la amplitud de banda óptima de referencia Gaussiana, hay poca indicación de la multimodalidad de este conjunto de datos, aunque una prominencia adicional a la izquierda de la moda es ligeramente sugerida por ambos estimadores.

Cuadro 4.2 Algunas reglas prácticas para elección de número y amplitud de intervalo/banda para estimación de densidad por histogramas, polígonos de frecuencia y estimadores por kernel. Datos de precipitación de nieve

No. de intervalos de Sturges	6.9773
No. sobresuavizado de intervalos	5.0133
No. sobresuavizado de intervalos en PF	5.4091
Amplitud óptima Gaussiana de Scott	20.8641
Amplitud óptima robusta de Freedman-Diaconis	17.4413
Amplitud sobresuavizada de Terrell y Scott	20.2262
Amplitud sobresuavizada homoscedástica	22.2292
Amplitud sobresuavizada robusta	22.6999
Amplitud óptima Gaussiana en ***PF***	22.2680
Amplitud sobresuavizada en ***PF***	24.1323
Amplitud de banda óptima Gaussiana de Silverman	9.3215
Amplitud de banda óptima mejorada de Härdle	10.9787
Amplitud de banda sobresuavizada para kernel Gaussiano de Scott	11.8487

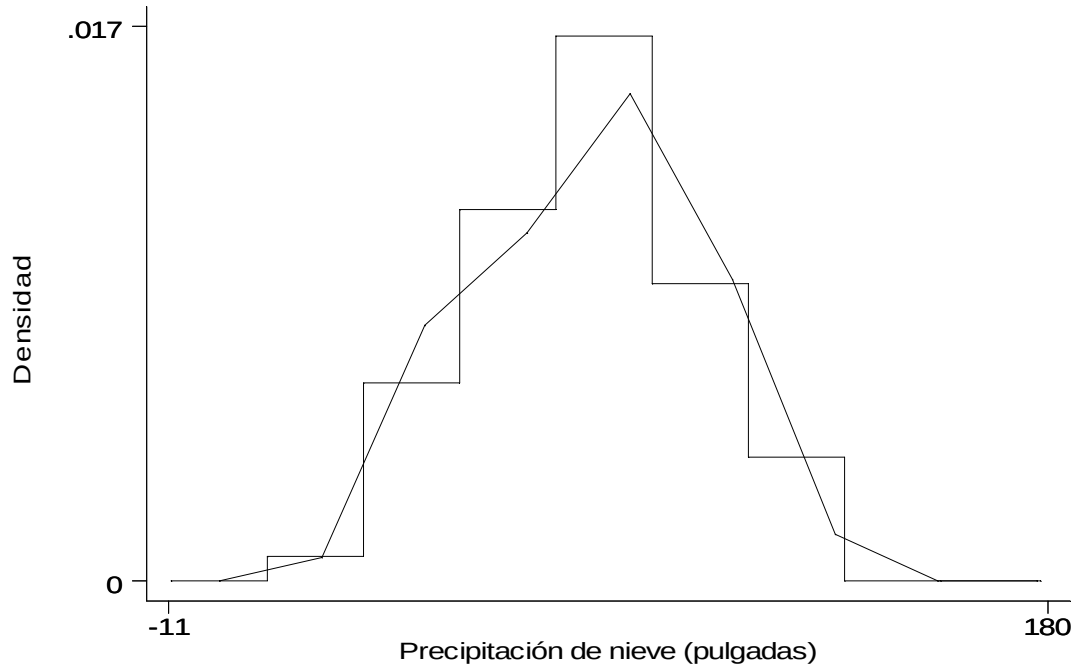

Figura 4.1 Histograma y polígono de frecuencia con la regla de referencia Gaussiana (20.9 y 22.3 respectivamente) para los datos de precipitación de nieve.

El histograma y el ***PF*** correspondientes a las amplitudes sobresuavizadas de banda se presentan en la Figura 4.2. Para el histograma se utilizó (de forma conservadora) la mayor amplitud de banda (sobresuavizada robusta). Como era de esperar, estas versiones muestran densidades muy suaves, pero en el ***PF*** sobresuavizado hay un indicio de la presencia de una prominencia a la derecha de la moda. Por lo tanto, valdrá la pena emplear otros métodos (por ejemplo el de Silverman, 1981b; 1983) para validar una estructura más compleja.

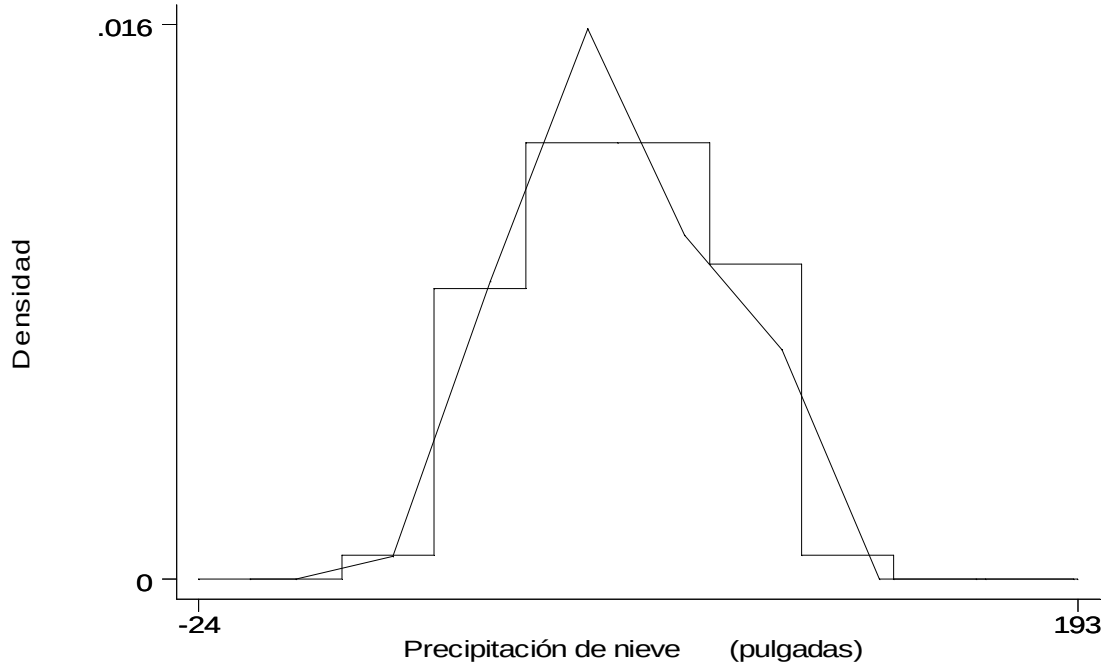

Figura 4.2 Histograma con intervalo de amplitud sobresuavizada robusta *h* = 22.7 y polígono de frecuencia con la amplitud sobresuavizada de intervalo *h* = 24, para los datos de precipitación de nieve.

El cuadro 4.3 contiene los parámetros de suavización sugeridos por las reglas prácticas para los datos de longitud de bagres.

Cuadro 4.3 Algunas reglas prácticas para elección de número y amplitud de intervalo/banda para estimación de densidad por histogramas, polígonos de frecuencia y estimadores por kernel. Datos de precipitación de nieve	
No. de intervalos de Sturges	12.2521
No. sobresuavizado de intervalos	16.9595
No. sobresuavizado de intervalos en PF	11.2383
Amplitud óptima Gaussiana de Scott	15.0376
Amplitud óptima robusta de Freedman-Diaconis	16.0466
Amplitud sobresuavizada de Terrell y Scott	13.2079
Amplitud sobresuavizada homocedástica	16.0215
Amplitud sobresuavizada robusta	20.8847
Amplitud óptima Gaussiana en PF	26.1322
Amplitud sobresuavizada en PF	28.3200
Amplitud de banda óptima Gaussiana de Silverman	10.9391
Amplitud de banda óptima mejorada de Härdle	12.8838
Amplitud de banda sobresuavizada para kernel Gaussiano de Scott	13.9048

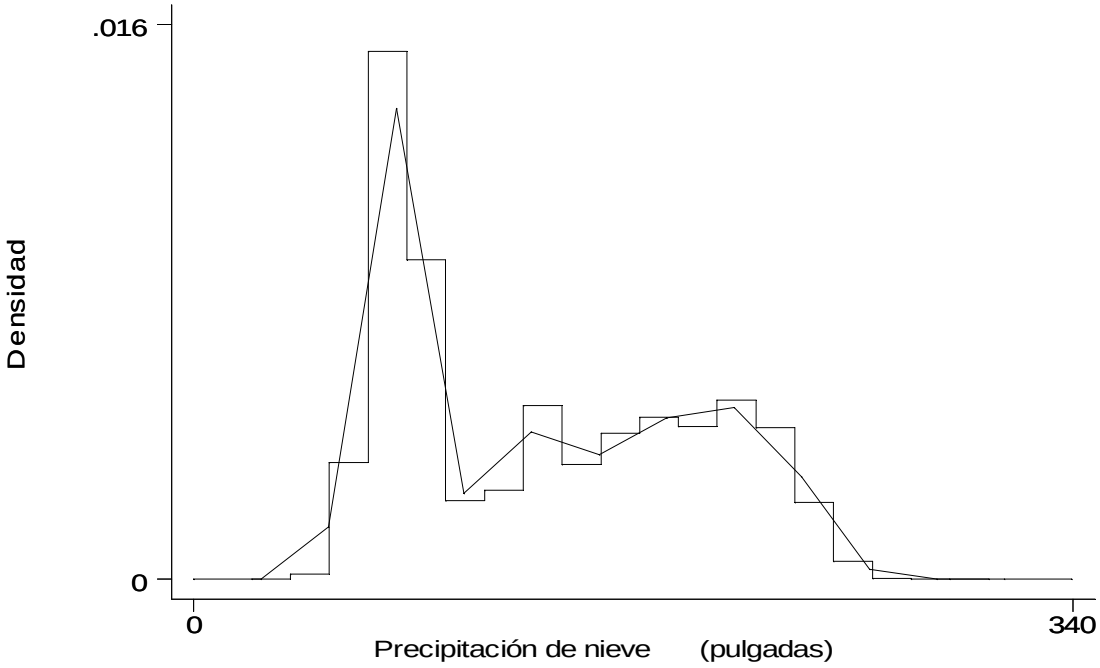

Figura 4.3 Histograma con amplitud óptima Gaussiana (*h* = 15) y polígono de frecuencia (*h* = 26.13) para los datos de longitud de bagres.

Tanto la regla óptima Gaussiana como la sobresuavizada robusta dan indicación de una estructura multimodal más compleja (Figuras 4.3 y 4.4 respectivamente). Al menos tres modas pueden identificarse fácilmente y los resultados sobresuavizados proporcionan un fuerte respaldo a su existencia. Por tanto, debe realizarse un análisis adicional enfocado a la caracterización de esta estructura multimodal (ver Izenman y Sommer, 1988 para un ejemplo de la estrategia a seguir; otras contribuciones recientes para la evaluación de la multimodalidad son Comparini y Gori, 1986, Roeder, 1990 y Müller y Sawitzki, 1991).

Como lo sugiere Terrell (1990) podemos utilizar las amplitudes sobresuavizadas de intervalo al buscar estimaciones conservativas tanto en histogramas o polígonos de frecuencia (ver más adelante la sección acerca de los programas revisados para la estimación univariada de densidad) ó podemos explorar las amplitudes de banda sugeridas para estimadores por kernel. Con los factores incluidos en el cuadro 4.1 y la ecuación 16, es posible calcular la amplitud de banda sobresuavizada para funciones ponderales no Gaussianas. Un enfoque adicional (y más sencillo) es convertir las amplitudes Gaussianas sobresuavizadas de banda a las amplitudes correspondientes para cualquiera de los kerneles incluidos en el Cuadro 3.2 del Capítulo 3 acerca de los estimadores ASH-WARP.

Siguiendo a Scott y Terrell (1987) y Härdle (1992) en esta obra se presentan modificaciones de los algoritmos propuestos por este último autor para estructurar una estimación basada en el ***PPPR*** para la ***VCMC*** y ***VCS*** aplicada a la estimación de densidad por kernel con los programas incluídos en el capítulo 7. Como estos autores discuten, el uso de ***PPPR*** es un método eficiente para el cálculo que nos permite localizar la amplitud de banda óptima y llevar a cabo simulaciones con un considerable número de observaciones y repeticiones si así se desea. Para lograr lo anterior se fija la amplitud de intervalo pequeño δ. Cómo $h = M \% \delta$, para encontrar el parámetro de

suavización “óptimo” es necesario tan sólo localizar el valor “óptimo” de *M*. Como en el caso de las estimaciones de densidad por ***PPPR*** presentadas anteriormente, estos programas utilizan archivos ejecutables en lenguaje Turbo Pascal los cuales realizan los cálculos principales y escriben los resultados en formato ASCII. Estos resultados son llamados por macros del paquete Stata, el cual a su vez lleva a cabo el proceso final de la estimación.

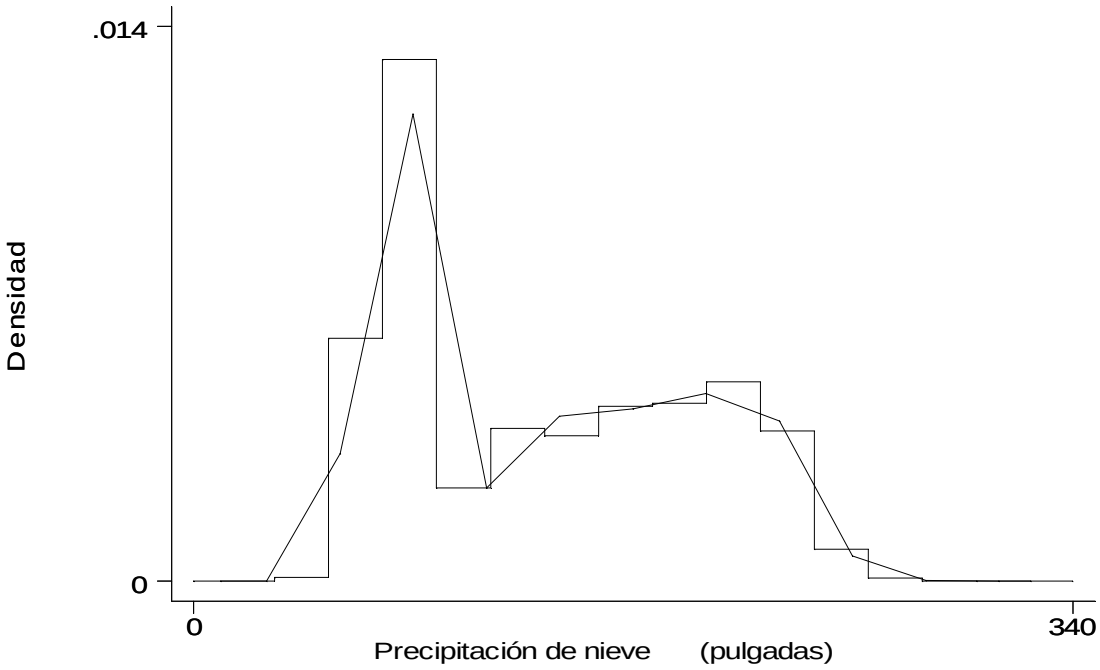

Figura 4.4 Amplitudes sobresuavizadas de intervalo para histograma (*h* = 20.88) y polígono de frecuencia (*h* = 28.32) calculados por el método PPPR.

Los valores del parámetro *delta* y el código para el kernel son indispensables para correr el programa de ***VCMC*** (si no se especifican este se detiene y despliega un mensaje de error). Es posible determinar el intervalo de *M* considerado para la búsqueda mínima. El valor pre-establecido inicial para *M* es 1; si no se especifica un valor final para *M* entonces se establece como aproximadamente un tercio del recorrido de las observaciones, valor que produce una primera aproximación de manera adecuada.

Este programa produce como resultado una gráfica del valor de la validación cruzada *vs. M* lo que permite localizar el intervalo con el mínimo si existe (recordar que $h = M \% \delta$). Además de la gráfica, el programa muestra como resultado un cuadro con los cinco valores de validación cruzada más bajos.

Considerando una de las recomendaciones explícitas de Scott (1992), y como se sugiere por las gráficas de Härdle (1991) el usuario deberá usar una escala logarítmica para los ejes de *M* y *h*. Además, Scott recomienda incluir una línea de referencia que represente la amplitud de intervalo sobresuavizada. Además de la observación gráfica, el cuadro con los cinco valores menores de ***VC*** permite obtener la amplitud “óptima” de intervalo. Si un valor mínimo no aparece en el intervalo de *M* utilizado, pero se sugiere por una tendencia monótona decreciente hacia alguno de los extremos,

es posible localizarlo mediante un proceso iterativo de ensayo y error, desplazando el intervalo al especificar valores inicial y final diferentes para la estimación.

Con este comando, es posible estimar el parámetro de suavización "óptimo" para una amplia gama de estimadores de densidad por kernel. Al especificar un kernel triangular es posible obtener la amplitud óptima de intervalo para el ***HDP***, y al aumentar el número de histogramas desplazados en la estimación de densidad ***PPPR*** (*M*) el estimador de densidad por kernel triangular correspondiente. Las otras funciones ponderales se aplican de acuerdo a sus respectivos estimadores de densidad.

Un ejemplo de la aplicación de este método se presenta a continuación, utilizando de nuevo los datos de precipitación de nieve. Después de una mirada preliminar al resultado utilizando un valor *delta* = 1, y el intervalo pre-establecido para *M* se encontró que el valor de ***VC*** mínimo se localiza en el intervalo $1 < M < 40$. Con un procedimiento de ensayo y error llegamos finalmente a nuestra elección final que se presenta en la Figura 4.5 que reproduce la Figura S.4.2 del texto de Härdle. Debe notarse, sin embargo, que la estimación de Härdle se obtuvo utilizando un algoritmo directo, no la función ***PPPR*** la cual no soporta al kernel Gaussiano en los programas de este autor. Esta diferencia explica la ligera discrepancia en los resultados (una amplitud de banda óptima de 9 propuesta por Härdle contra la óptima de 8 aquí presentada). Al disminuir *delta* se obtiene una aproximación ligeramente mayor, pero el costo computacional puede ser mayor. Además, del cuadro de resultados (Cuadro 4.4) podemos notar que hay poca diferencia en los valores de ***VC*** en el intervalo para valores de *h* de 7 a 10. Puede probar un valor de 0.5 para *delta*, con una *M* inicial de 10 y una *M* final de 40 por ejemplo, para obtener una amplitud de banda "óptima" de 9.5, la misma obtenida por Scott (1992) en su Figura 6.16, p. 172.

Se ha incluido adicionalmente una línea que representa la amplitud de banda sobresuavizada para kernel Gaussiano calculada previamente con el programa para las reglas prácticas de amplitud de intervalo/banda (**bandw.ado**) pero re-expresada como un valor de *M* aplicando h_{OS}/δ (en nuestra elección *delta* es igual a 1, y por tanto, el valor de *h* es igual a *M*). Los resultados obtenidos por los programas de Härdle (utilizando versiones actualizadas de las funciones del lenguaje estadístico *S* y los programas en *C* obtenidas del sitio FTP Statlib) y los resultados obtenidos con los programas presentados en la presente obra usando los mismos valores (*delta, kercode, mstart y mend)* fueron las mismas.

Cuadro 4.4 Validación cruzada por mínimos cuadrados para estimación de densidad por kernel Gaussiano. Datos de precipitación de nieve

Valor de VC	Valor de *M*	Amplitud de banda
-0.01113733	8	8.0000
-0.01113101	7	7.0000
-0.01111789	9	9.0000
-0.01108621	10	10.0000
-0.01107764	6	6.0000

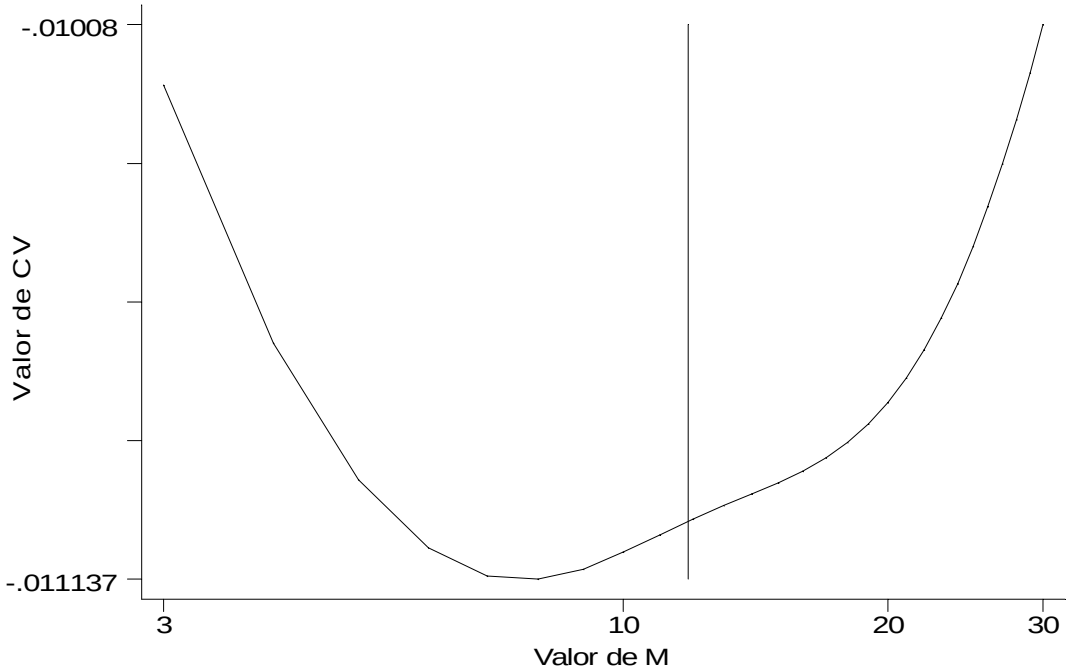

Figura 4.5 Función de valores de validación cruzada por mínimos cuadrados para estimación de densidad por kernel Gaussiano, para los datos de precipitación de nieve. La línea vertical representa la amplitud de banda sobresuavizada para kernel Gaussiano *h* = 11.85.

El programa para estimar al parámetro de suavización por medio de validación cruzada sesgada (***VCS***) es **bcvwarp.ado** y se presenta en el Capítulo 7. Es muy similar al programa para calcular la ***VCMC*** (**l2cvwarp.ado**) excepto que sólo dos tipos de funciones kernel están incluidas con los siguientes códigos:

1. Cuártica (biponderada)
2. Triponderada

De manera similar al programa de ***VCMC***, el resultado del programa es una gráfica del valor de validación cruzada sesgada contra el intervalo de *M* utilizado para los cálculos; esta gráfica ayuda a localizar el mínimo en la curva resultante. Después del desplegado gráfico aparece el cuadro con los cinco valores menores de los puntajes de ***VCS*** con sus correspondientes valores de *M* y *h*.

Como se mencionó anteriormente, por medio de las tablas incluidas en Härdle (1991) o Scott (1992), presentadas en forma ligeramente modificada en el Cuadro 3.2 del Capítulo 3 es posible transformar la amplitud de banda óptima obtenida con cualquiera de las dos funciones de arriba en cualquiera de los kerneles de dicho Cuadro para calcular la correspondiente estimación de densidad si fuera necesario. La aplicación para los datos de precipitación de nieve se presenta en el Cuadro 4.5 y en la Figura 4.6.

Cuadro 4.5 Validación cruzada sesgada para estimación de densidad por kernel Triponderado. Datos de precipitación de nieve		
Valor de VC sesgada	Valor de *M*	Amplitud de banda
0.00593182	43	43.0000
0.00593193	42	42.0000
0.00593229	44	44.0000
0.00593269	41	41.0000
0.00593329	45	45.0000

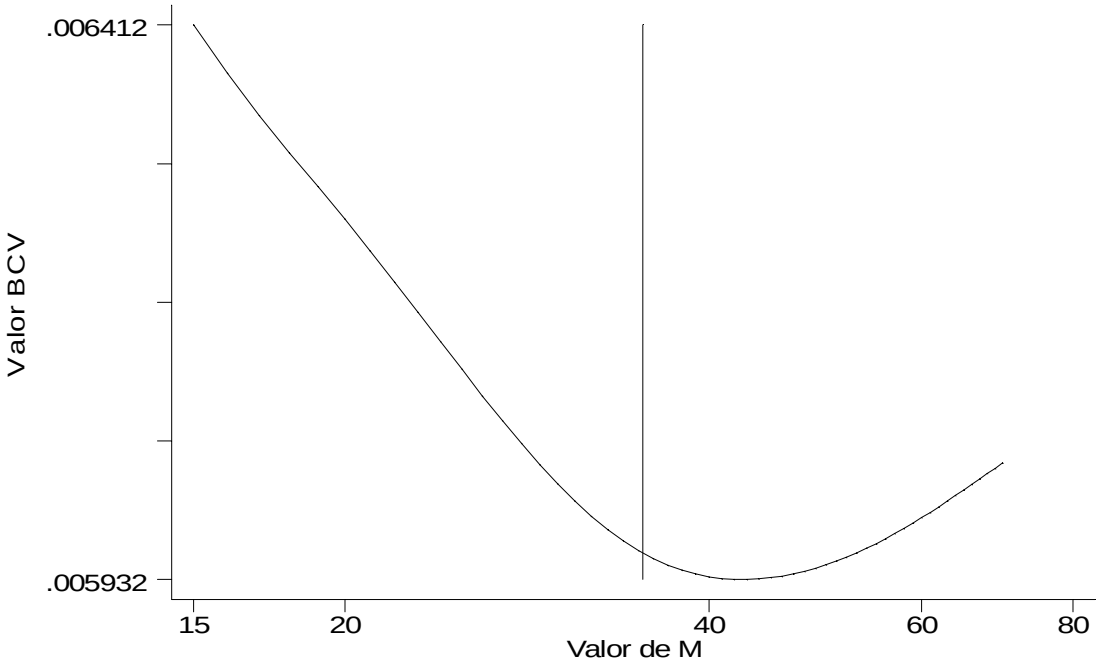

Figura 4.6 Valores de la función de validación cruzada sesgada con kernel triponderado, para los datos de precipitación de nieve. La línea vertical es la amplitud de banda sobresuavizada para kernel triponderado (h_{OS} = 35.29) obtenida por transformación del valor para kernel Gaussiano.

Notar que en la Figura 4.6 se ha transformado la amplitud sobresuavizada de banda para kernel Gaussiano proveniente del programa **bandw.ado** por medio del factor de conversión 2.978 para incluirla como la línea de referencia. La amplitud de banda para kernel triponderado sugerida es 43, la cual es mayor que el valor sobresuavizado. Para comparar con la amplitud de banda resultante de la ***VCMC***, el resultado (una amplitud de banda para kernel triponderado) se multiplica por 0.336 para obtener la amplitud de banda óptima Gaussiana = 14.5. Este es aproximadamente el mismo resultado de Scott (1992, Figura 6.16, p. 172) quien utilizó un algoritmo para ***VCS*** con kernel Gaussiano.

Las estimaciones de densidad correspondientes (calculadas con el programa **warpdenm.ado** presentado en el Capítulo 7) se incluyen en la Figura 4.7. El parámetro de suavización por ***VCMC*** da un indicio de la existencia de tres modas en estos datos. Por otro lado, la

amplitud de banda por ***VCS*** resulta en una representación muy suave sin ninguna evidencia de multimodalidad.

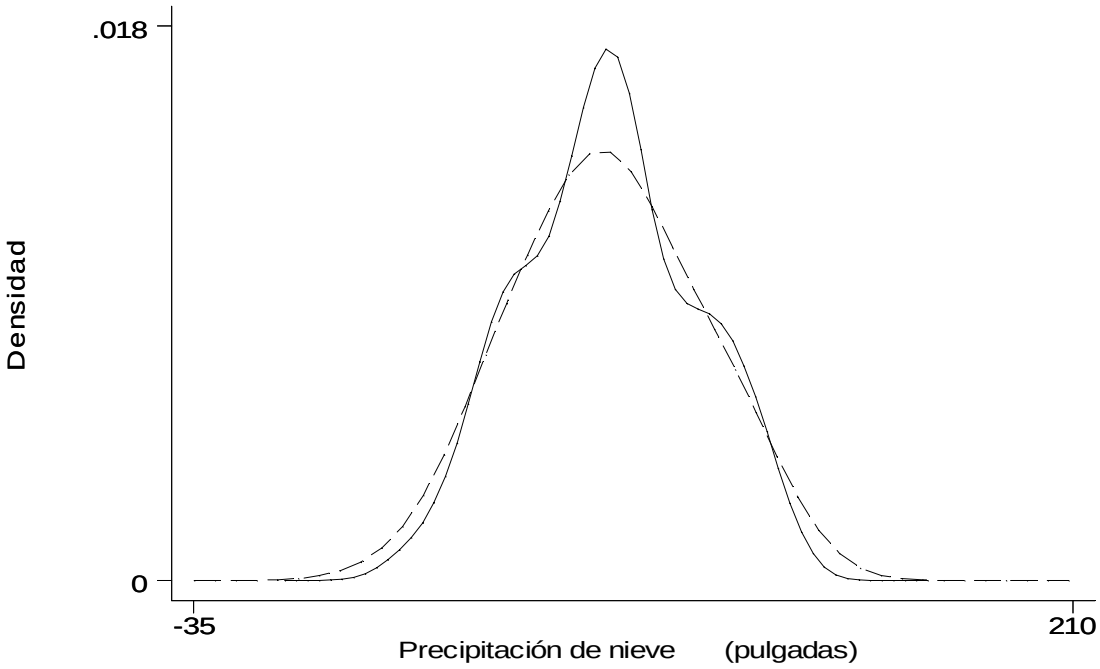

Figura 4.7 Estimaciones de densidad para los datos de precipitación de nieve; amplitud de banda por *VCMC* de 8 (línea continua); amplitud de banda por *VCS* de 14.5 (l

Para el segundo conjunto de datos (longitud de bagres, n = 2439) las amplitudes de banda por ***VCMC*** y ***VCS*** fueron calculadas utilizando un kernel cuártico (biponderado) para su comparación directa, y los resultados se presentan en los Cuadros 4.6 y 4.7.

Cuadro 4.6 Validación cruzada por mínimos cuadrados para estimación de densidad por kernel Gaussiano. Datos de longitud de bagres.

Valor de VC	Valor de *M*	Amplitud de banda
-0.00780885	8	8.0000
-0.00780763	9	9.0000
-0.00780166	7	7.0000
-0.00779905	10	10.0000
-0.00778844	11	11.0000

Cuadro 4.7 Validación cruzada sesgada para estimación de densidad por kernel Gaussiano. Datos de longitud de bagres.

Valor de VC sesgada	Valor de *M*	Amplitud de banda
0.00560286	8	8.0000
0.00560392	9	9.0000
0.00560413	7	7.0000
0.00560660	10	10.0000
0.00560979	6	6.0000

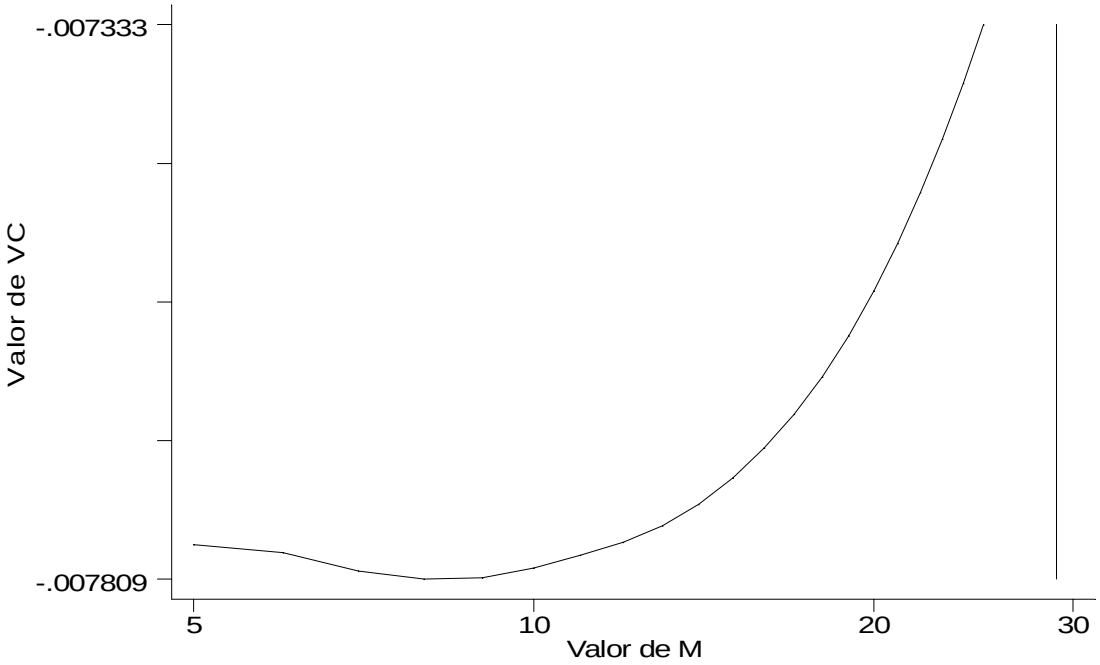

Figura 4.8 Valor de validación cruzada por mínimos cuadrados para los datos de longitud de bagres (*n* = 2439). La amplitud sobresuavizada de banda transformada es indicada por la línea en *M* = 29.

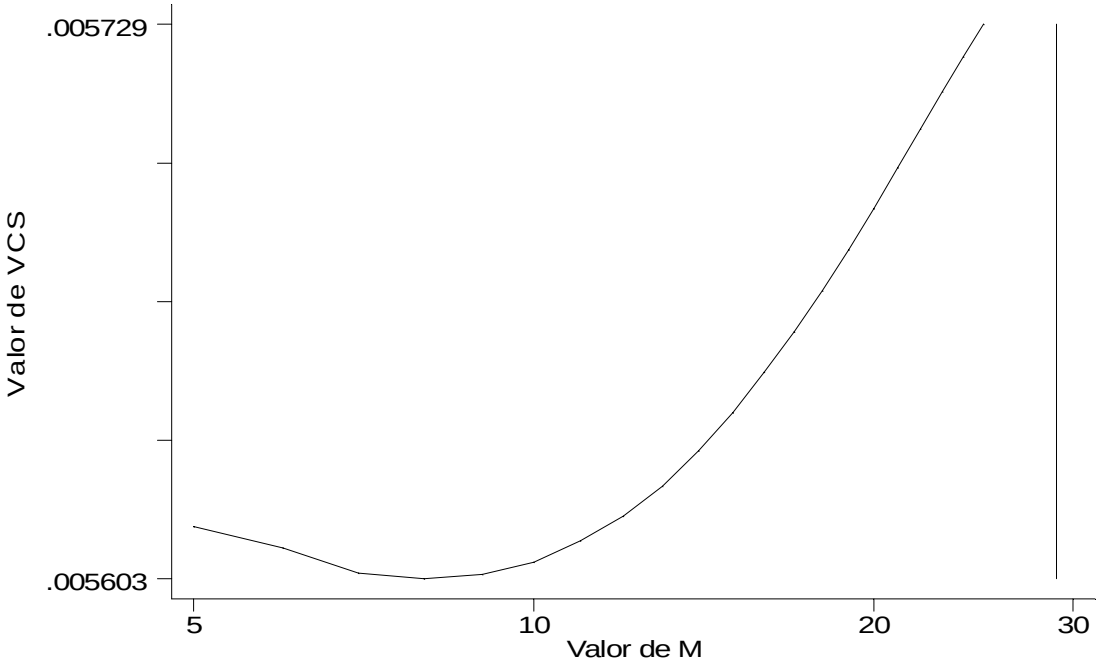

Figura 4.9 Valor de validación cruzada sesgada para los datos de longitud de bagres (*n* = 2439); la amplitud sobresuavizada de banda es indicada por la línea en *M* = 29.

Tanto la ***VCMC*** como la ***VCS*** producen la misma amplitud óptima de 8 (Figuras 4.8 y 4.9). Varias modas pueden notarse en la Figura 4.10 que presenta la estimación de densidad calculada con el procedimiento ***HDP-PPPR*** con esta amplitud de banda. Scott (1992) ha señalado que la concordancia entre estos criterios diferentes debiera tomarse en cuenta seriamente. En conclusión, existe una fuerte evidencia para una distribución multimodal de la longitud corporal estándar de los bagres.

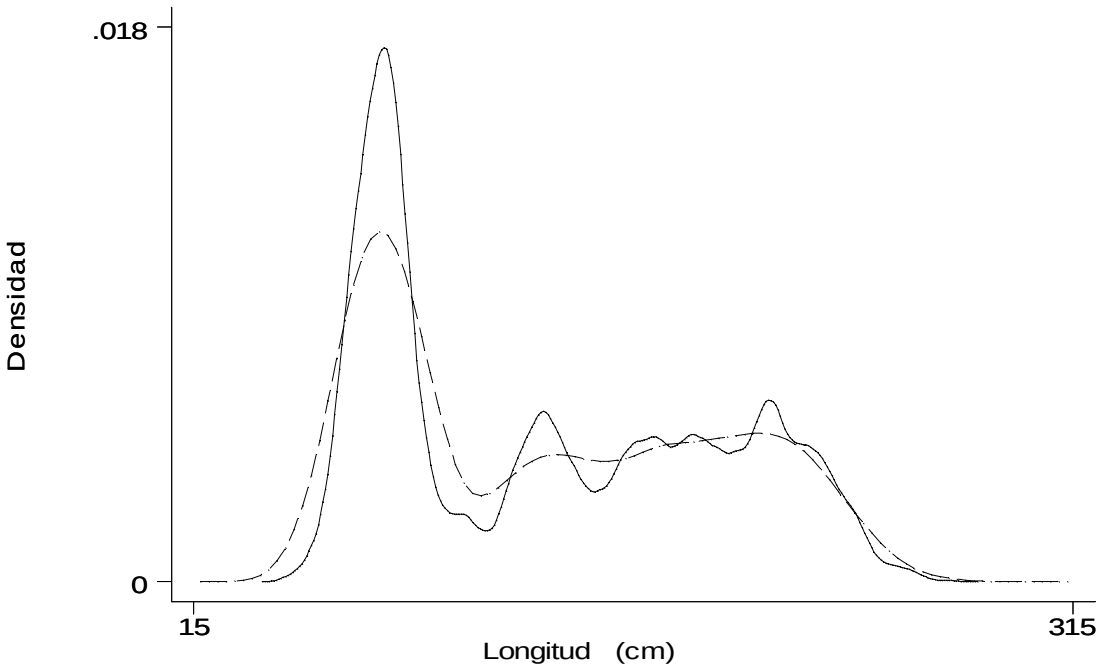

Figura 4.10 Estimadores de densidad por kernel para los datos de longitud de bagre con kernel cuártico (biponderado): amplitud de banda óptima por VCMC = amplitud de banda óptima por VCS = 8 (línea continua) y amplitud sobresuavizada de Scott transformada (línea punteada).

4.11. Algunos comentarios acerca de la elección del parámetro de suavización

Se han propuesto algunos otros procedimientos no discutidos en la presente obra. Como ejemplos podemos citar:

- Validación cruzada por muestreo repetitivo bootstrap (Taylor, 1989)
- Métodos Plug-In (Sheather y Jones, 1991; Hall, *et al.*, 1991)

Espero que esta obra motive el interés sobre la aplicación de estos útiles métodos para elegir el parámetro de suavización en la estimación de densidad. Mientras tanto, considero que la colección de reglas y métodos presentados arriba proporcionan una guía práctica para la selección de una estimación de densidad adecuada.

Es necesario reconocer que todos los procedimientos introducidos en este Capítulo tienen limitaciones, y como se ha afirmado por varios autores (Marron, 1986; Scott, 1992) la práctica analítica de dibujar varias estimaciones con diferentes parámetros de suavización no será reemplazada completamente por métodos automáticos de suavización. Desde un punto de vista exploratorio, cualquier elección de amplitud de banda produce estimaciones de utilidad; valores grandes de *h* permiten reconocer características estructurales generales tales como simetría, casos extraordinarios, modas y localización, mientras que valores pequeños de *h* revelan estructuras locales que pueden ser reales o artifácticas no presentes in la densidad “verdadera”.

No obstante, la búsqueda de un procedimiento para la elección completamente automática de la amplitud de banda o intervalo ha motivado el desarrollo de algoritmos novedosos como aquellos que se han mencionado en este apartado. En la actualidad este es un campo activo de investigación en Estadística (Scott, 1992).

Capítulo 5. Evaluación no paramétrica de la multimodalidad

5.1 Introducción

Los investigadores que trabajan con formas complejas de distribución han vuelto su atención en años recientes hacia las técnicas no-paramétricas tales como la estimación de densidad por kernel (Silverman, 1986) donde los componentes individuales mezclados pueden detectarse por la identificación de modas (máximos locales) en la distribución subyacente (Izenman y Sommer, 1988).

No hay garantía de que una mezcla de densidades unimodales resulten en una densidad multimodal con modas correspondiendo al número y localización de cada componente individual. Existe una dependencia sobre tanto el espaciamiento de las modas como de las formas relativas de las distribuciones de los componentes.

Sin embargo, en muchas instancias prácticas, la existencia de más de una sola moda sugiere la evidencia de una mezcla. Varias pruebas han sido propuestas para detectar la multimodalidad en una distribución (Hartigan y Hartigan, 1985). Por ejemplo, Good y Gaskins (1980) usaron el método de verosimilitud penalizada para estimación de densidad junto con ciertas técnicas quirúrgicas para "búsqueda de protuberancias", mientras que el método que presentaré posteriormente propuesto por Silverman (1981b) combina la estimación de densidad por kernel con un procedimiento de prueba jerárquico de muestreo repetitivo (bootstrap). Ambos métodos son no-paramétricos, basados en los datos y de cómputo intensivo.

Los contextos específicos desempeñan a menudo un papel prominente en la relación de las modas empíricas con la existencia de componentes mezclados. Así, Silverman (1978a, 1981a) en un estudio de mortalidad infantil súbita o "cot death" sugirió que una densidad de mortalidad esencialmente bimodal pudiera explicarse como una mezcla de la densidad de muerte en hospital (infantes que mueren en el hospital por causas conocidas) con una densidad contaminada y desplazada. De manera similar, Good y Gaskins (1980) atribuyeron a un ajuste trimodal de un conjunto de datos sobre condritas carbonosas a la existencia de (al menos) tres tipos diferentes de condritas. En el caso de mezclas filatélicas analizadas por Izenman y Sommer (1988), el contexto histórico desempeña el papel principal en la identificación de los componentes en la mezcla. La frecuencia de tamaños en peces puede representar diferentes grupos de edad.

En este Capítulo introduciré brevemente dos procedimientos paramétricos para la evaluación de la multimodalidad y daré algunos ejemplos de su aplicación en el análisis de la frecuencia de tallas en peces. El primero es el estadístico "dip", es tan sólo brevemente bosquejado. El segundo es el procedimiento de muestreo repetitivo suavizado sugerido por Silverman (1981b, 1983), y el cual es descrito en más detalle ya que utiliza estimadores de densidad por kernel, procedimientos revisados en capítulos anteriores.

5.2 La prueba "dip" de unimodalidad

Hartigan y Hartigan (1985) propusieron el "estadístico dip" como la máxima diferencia entre la función distribución empírica y la función distribución unimodal que minimiza la diferencia

máxima. Este estadístico puede calcularse por medio de un número de operaciones de orden *n*, para *n* observaciones, y es consistente para probar cualquier distribución unimodal contra cualquier distribución multimodal. Estos autores argumentan que la distribución nula apropiada es uniforme, al mostrar que el estadístico dip es asintóticamente mayor para la distribución uniforme que para cualquier otra distribución dentro de una amplia variedad de distribuciones unimodales como aquellas con colas que decrecen de manera exponencial. En su artículo, reportan la distribución asintótica del estadístico dip para el caso uniforme, especifican distribuciones derivadas empiricamente para algunas muestras finitas y llevan a cabo unos cuantos cálculos del poder. Además presentan el algoritmo para calcular este estadístico. En otra contribución (Hartigan, 1985) se introduce una rutina en lenguaje Fortran.

La distribución nula para la prueba dip es la distribución uniforme como el "peor de los casos" unimodal. Asintóticamente, $\oplus n$(dip) es positivo para la distribución uniforme y zero para distribuciones unimodales cuyas densidades decrecen exponencialmente a partir de la moda.

Existen varios estudios que han aplicado esta prueba, por ejemplo Izenman y Sommer (1988) quienes analizaron mezclas filatélicas, Roeder (1990) quien analizó datos de velocidades de galaxias. Recientemente, un método relacionado con la prueba dip pero más sofisticado ha sido publicado (Müller y Sawitzki, 1991).

5.3 Prueba de Silverman para multimodalidad

5.3.1 Descripción de la prueba

Esta prueba propuesta por Silverman (1981b) usa técnicas noparamétricas de estimación de densidad por kernel para determinar el número más probable de modas en la densidad subyacente.

Sea X_1, X_2,, … , X_n una muestra aleatoria de tamaño *n* de una distribución con densidad desconocida *f*. Definamos a $\hat{F}_n$ como la función empírica de distribución acumulada de la muestra obtenida al asignar masa de 1/*n* a cada X_i. Como una clase general de estimadores no paramétricos de densidad de *f*, Rosenblatt (1956) propuso

$$\hat{f}(x) = \int_{-\infty}^{+\infty} w_n(x,u)\, d\hat{F}_n(u) = n^{-1} \sum_{j=1}^{n} w_n(x, X_j), \qquad x \in \Re \tag{5.1}$$

donde la función ponderal positiva $w_n(x,u)$ satisface $\int_{-\infty}^{+\infty} w_n(x,u) = 1$. El estimador de densidad por kernel para *f* se obtiene estableciendo

$$w_n(x,u) = h^{-1} K((x-u)/h), \qquad x \in \Re \tag{5.2}$$

donde el kernel *K* es una función de densidad de probabilidad, y la amplitud de banda o amplitud de ventana $h = h(n) > 0$ del kernel converge a 0 al $n \rightarrow \equiv$. Por tanto, el estimador de densidad por kernel se escribe como

$$\hat{f}(x) = (nh)^{-1} \sum_{j=1}^{n} K((x - X_j)/h), \qquad x \in \Re \tag{5.3}$$

Como hemos visto previamente, la elección de *h* en 5.3 es un problema estadístico de importancia: Un valor demasiado pequeño de *h* produce una estimación de densidad demasiado dependiente en los valores de la muestra, mientras que un valor demasiado grande para *h* produce el efecto opuesto y sobresuaviza la estimación de densidad removiendo peculiaridades interesantes. En el Capítulo anterior se han presentado varios criterios óptimos para la elección de la amplitud de banda. Debe señalarse sin embargo, que la estimación de *h* óptima no es el interés principal aquí, puesto que estamos interesados en el conteo de modas, no en la suavización óptima (aunque ambos temas estén relacionados) (Izenman y Sommer, 1988).

La prueba de multimodalidad de Silverman puede describirse en los siguientes términos. La hipótesis nula H_o^k, establece que *f* posee cuando más *k* modas, mientras que la hipótesis alternativa, H_1^k, establece que *f* tiene más de *k* modas (*k* = 1, 2, …). Si establecemos que

$$N(f) = \#\{x: f'(x) = 0, f''(x) < 0\} \tag{5.4}$$

sea el número de modas en *f*, entonces $H_o^k: N(f) \le k$ and $H_1^k: N(f) > k$. Si $\hat{f}_h$ es el estimador de densidad por kernel de *f* con amplitud de banda *h*, entonces un estadístico de interes es $N(\hat{f}_h)$. Definiendo la *k*ésima amplitud crítica de banda como

$$h_{k,\text{crit}} = \inf\{h: N(\hat{f}_h) \le k\}, \tag{5.5}$$

la amplitud de banda menor que aún es compatible con H_o^k.

El método de Silverman es altamente dependiente de las propiedades del kernel Gaussiano, por lo que es necesario establecer

$$K(t) = (2\pi)^{-1/2} \exp(-t^2/2), \qquad t \in \Re, \tag{5.6}$$

como la función kernel en 5.3. Silverman mostró que para el kernel 5.6, $N(\hat{f}_h)$ es una función de *h* decreciente continua hacia la derecha y por tanto $N(\hat{f}_h) > k$ si y sólo si $h = h_{k,crit}$. Por lo tanto, para encontrar $h_{k,crit}$ contamos las modas en cada estimación de densidad $\hat{f}_h$ para diferentes valores de *h*. Cuando $h = h_{k,crit}$, $\hat{f}_h$ desplegará *k* modas más un "hombro" notable en su gráfica, y si *h* se reduce un poco más, brotará una moda adicional (*k + 1*) en el lugar de ese "hombro". Se denomina **densidad crítica** a cualquier densidad con un hombro, tal como $\hat{f}_{h_{k,crit}}$. De lo anterior se deriva

$$\Pr_f\{h_{k,\text{crit}} > h\} = \Pr\{N(\hat{f}_h) > k | X_1, X_2, \ldots, X_n \text{ es una muestra extraída de } f\} \tag{5.7}$$

Puesto que no es difícil sacar muestras de una densidad crítica, Silverman combinó la propiedad de monotonicidad de $N(\hat{f}_h)$ con el procedimiento bootstrap de muestreo repetitivo con reemplazamiento para construir una prueba funcional para la multimodalidad.

La significancia de $h_{k,crit}$ se evalúa a través del siguiente algoritmo:

1. Extraer *n* veces con reemplazamiento de la muestra aleatorio original X_1, X_2,, … , X_n, para obtener una muestra bootstrap $X_1^*, X_2^*, ..., X_n^*$.

2. Se obtiene una muestra bootstrap suavizada, $Y_1^*, Y_2^*, ..., Y_n^*$ por medio de calcular

$$Y_j^* = c_k \{ X_j^* + h_{k,\text{crit}} Z_j \}, \qquad j = 1, 2, \ldots, n, \tag{5.8}$$

donde Z_j es una desviación Gaussiana estándar independiente, y

$$c_k = (1 + [h_{k,\text{crit}} / s]^2)^{1/2} \tag{5.9}$$

se utiliza para restablecer la escala de tal forma que la varianza de Y_j^* sea igual a la varianza de la muestra s^2 de los datos originales (ver Efron, 1982). En otros términos, muestreamos de la convolución $c_k \{ \hat{F}_n * F_{norm} \}$ donde F_{norm} es la función distribución $N(0, h_{k,\,crit}^2)$. Esta convolución es más suave que $\hat{F}_n$ por sí sola.

3. Usar (5.8), (5.3) y (5.6) para formar una estimación de $\hat{f}^*_{h_{k,\,crit}}$, de *f*.
4. Repetir los pasos 1-3 un gran número *B* de veces. Definamos a $\hat{f}^{*b}_{h_{k,\,crit}}$ para denotar la estimación de densidad para la *b*ava muestra bootstrap suavizada.
5. Establecer

$$\begin{aligned} I_{k,b} &= 1, \ \text{si} N(\hat{f}^{*b}_{h_{k,\text{crit}}}) > k \\ &= 0, \ \text{en cualquier otro caso} \end{aligned} \tag{5.10}$$

Entonces,

$$P_k = B^{-1} \sum_{b=1}^{B} I_{k,b} \tag{5.11}$$

es el nivel de significancia estimado (ó valor *p*) de $h_{k,crit}$ y detenerse hasta obtener un valor lo suficientemente "grande".

Silverman no ofreció sugerencias de que tan "grande" pueda ser para seguir la regla de alto arriba citada. En su artículo aplicó el procedimiento a los datos de meteoritos condritas de Good y Gaskins (1980) (n = 22) y mostró que las amplitudes de banda críticas tuvieron valores *p:*

$P_1 = 0.08$, $P_2 = 0.05$, $P_3 = 0.79$ y $P_4 = 0.93$ parando a una k = 3 para una densidad trimodal.

En un artículo subsiguiente, Silverman (1983) mostró teóricamente que esta prueba bootstrap puede ser conservativo. Puesto que no se han publicado estudios de simulación de esta prueba, los investigadores la ven como una técnica analítica exploratorio de datos. Izenman y

Sommer (1988) sugieren que debe tomarse una actitud flexible en la aplicación de la prueba a datos con estructuras de distribución complejas y con colas largas. Estos autores comentan que no hay razón para esperar que la secuencia de valores *p* (5.11) sea monótonamente creciente; por cierto que el propio estudio de Silverman sobre las condritas ilustra este punto. Además, es posible, dependiendo de la posición de las modas, que se observen en la práctica grandes fluctuaciones en los valores de *p*. Basados en esta experiencia y en las anotaciones previas en relación con la naturaleza conservativa de la prueba bootstrap, los citados autores sugieren aplicar una regla de paro flexible con un valor nominal de *p* de 0.40 hasta que estudios detallados se lleven a cabo. Izenman y Sommer, no obstante, recomiendan fuertemente el estudio de las gráficas de las estimaciones de densidad cerca de cada banda crítica y las gráficas de las estimaciones de densidad de las propias bandas críticas durante el desarrollo de la prueba de Silverman.

5.3.2 Ejemplos de aplicación

Se presenta a continuación un ejemplo de la aplicación de los métodos presentados arriba. El conjunto de datos se ha presentado en apartados anteriores, pero aquí se incluye una breve descripción (del estudio original de Salgado-Ugarte, 1985). De julio, 1980 a agosto 1981 se colectó un total de 2436 individuos del bagre estuarino *Cathorops melanopus* en la Laguna de Tampamachoco localizada en la costa Noreste de México. Debido a que se encontró diferencia estadística significativa entre sexos se consideró conveniente analizar los datos por sexos separados. Los organismos de sexo indeterminado se incluyeron como una submuestra de aproximadamente 50% de su número total. En esta sección, presento el análisis de la longitud corporal para las hembras y los individuos de sexo indeterminado para evaluar su multimodalidad.

Un breve resumen del conjunto de datos se presenta en el Cuadro 5.1:

Cuadro 5.1 Número de hembras e individuos de sexo indeterminado de *Cathorops melanopus* considerados para evaluación de multimodalidad

Sexo	Frecuencia	Porcentaje	% Acumulado
Hembras	556	49.82	49.82
Indeterminado	560	50.18	100.00
Total	1116	100.00	

Para calcular las estimaciones de densidad por kernel se utilizó el procedimiento HDP-PPPR descrito en el Capítulo 3 por medio de los programas explicados en el Capítulo 7. Esta versión aceleró considerablemente el desempeño en todas las estimaciones.

5.3.3. Uso de reglas de amplitud de banda

El cuadro 5.2 contiene los resultados de las reglas para amplitud de intervalo/banda introducidas en el Capítulo 4 para el conjunto de datos aquí considerado. Enfocaré mi atención a los resultados en relación al kernel Gaussiano. La Figura 5.1 despliega la estimación de densidad por kernel Gaussiano que utiliza el valor de amplitud óptima del Cuadro 5.2. La estimación sobresuavizada se incluye en la Figura 5.2. Ambas figuras muestran un alto grado de multimodalidad y de acuerdo a las consideraciones de Terrell y Scott (1985) y Terrell (1990), estas modas son sugeridas sólidamente como estructuras reales del conjunto de datos. Esta multimodalidad merece un análisis adicional.

Cuadro 5.1 Algunas reglas prácticas para elección de número y amplitud de intervalo/banda para estimación de densidad por histogramas, polígonos de frecuencia y estimadores por kernel. Datos de longitud corporal de bagres	
No. de intervalos de Sturges	11.1241
No. sobresuavizado de intervalos	13.0687
No. sobresuavizado de intervalos en PF	9.6115
Amplitud óptima Gaussiana de Scott	18.7654
Amplitud óptima robusta de Freedman-Diaconis	18.3175
Amplitud sobresuavizada de Terrell y Scott	16.3750
Amplitud sobresuavizada homocedástica	19.9932
Amplitud sobresuavizada robusta	23.8402
Amplitud óptima Gaussiana en PF	29.3822
Amplitud sobresuavizada en PF	31.8421
Amplitud de banda óptima Gaussiana de Silverman	12.2995
Amplitud de banda óptima mejorada de Härdle	14.4861
Amplitud de banda sobresuavizada para kernel Gaussiano de Scott	15.6340

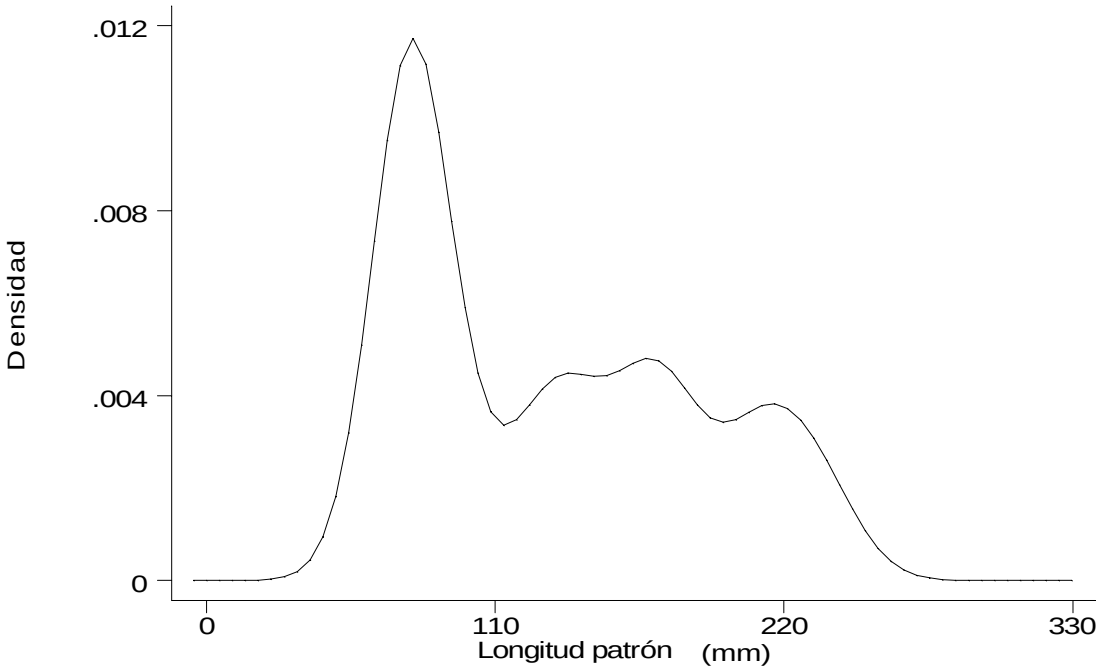

Figura 5.1 Estimación de densidad por kernel Gaussiano con amplitud de banda óptima *h* = 12.3

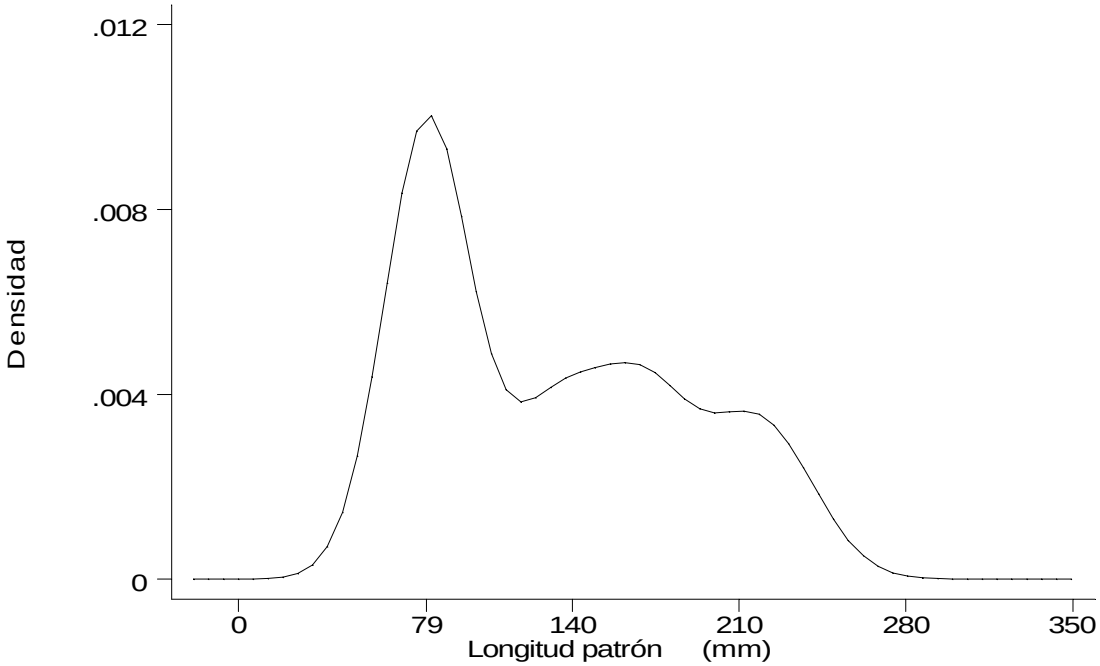

Figura 5.2 Estimación de densidad por kernel con la amplitud sobresuavizada de banda de Scott *h* = 15.6

Cuadro 5.2 Validación cruzada por mínimos cuadrados para estimación de densidad por kernel Gaussiano. Datos de longitud de bagres (*n* = 1116)		
Valor de VC	Valor de *M*	Amplitud de banda
-0.00841106	8	4.0000
-0.00841082	7	3.5000
-0.00840228	9	4.5000
-0.00840097	6	3.0000
-0.00838488	10	5.0000

Cuadro 5.3 Validación cruzada sesgada para estimación de densidad por kernel Triponderado. Datos de longitud de bagres (*n* = 1116)		
Valor de VC sesgada	Valor de *M*	Amplitud de banda
0.00564258	10	10.0000
0.00564327	11	11.0000
0.00564445	9	9.0000
0.00564636	12	12.0000
0.00564954	8	8.0000

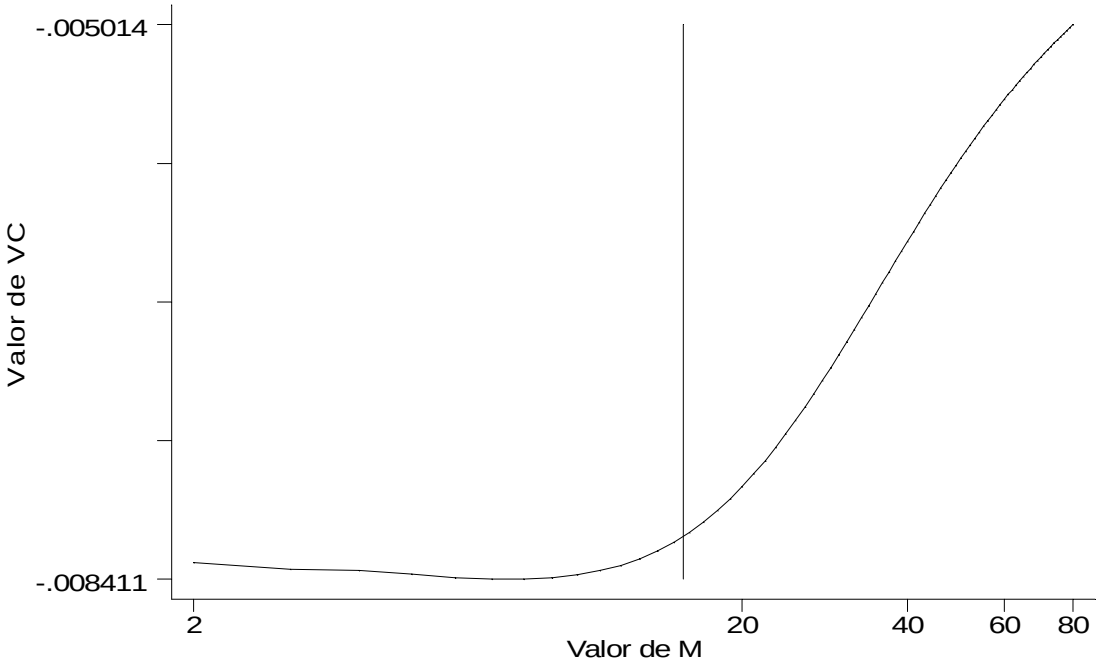

Figura 5.3 Función de validación cruzada por mínimos cuadrados para los datos de longitud de hembras-indeterminados de bagres (*n* = 1116). La línea vertical indica la amplitud de banda sobresuavizada en 15.6

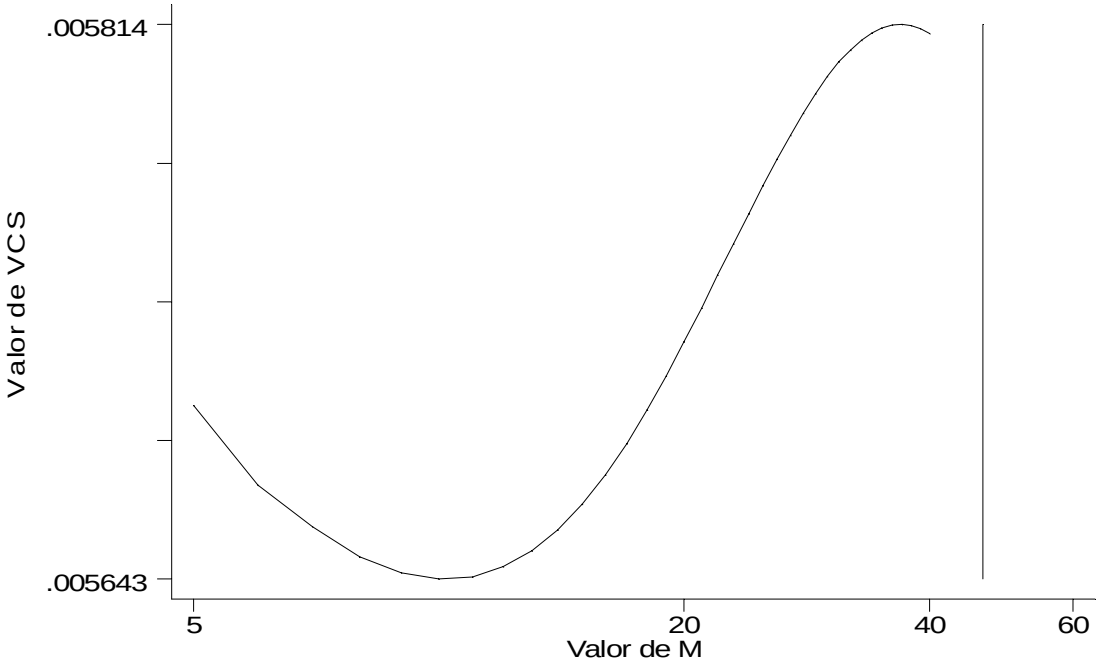

Figura 5.4 Función de validación cruzada sesgada para los datos de longitud de bagres hembras e indeterminados (*n* = 1116). La línea vertical indica la amplitud de banda sobresuavizada transformada a kernel Triponderado de 46.5

Al investigar la amplitud de banda óptima por medio de la validación cruzada se produjeron los resultados presentados en los cuadros 5.2 y 5.3. La función de validación cruzada por mínimos cuadrados con kernel Gaussiano se muestra en la Figura 5.3. Notar que esta función es casi plana cerca del mínimo. La figura 5.4 despliega la función de validación cruzada sesgada con kernel triponderado. El mínimo se distingue de manera más clara pero hay alguna evidencia de la existencia de un mínimo local adicional por encima de la amplitud de banda sobresuavizada.

Los valores de amplitud de banda sugeridos por VCMC y por VCS (transformada) fueron casi los mismos (4 y 3.36, respectivamente). La estimación de densidad con $h = 4$ se presenta en la Figura 5.5, en la cual la naturaleza multimodal de los datos se aprecia claramente.

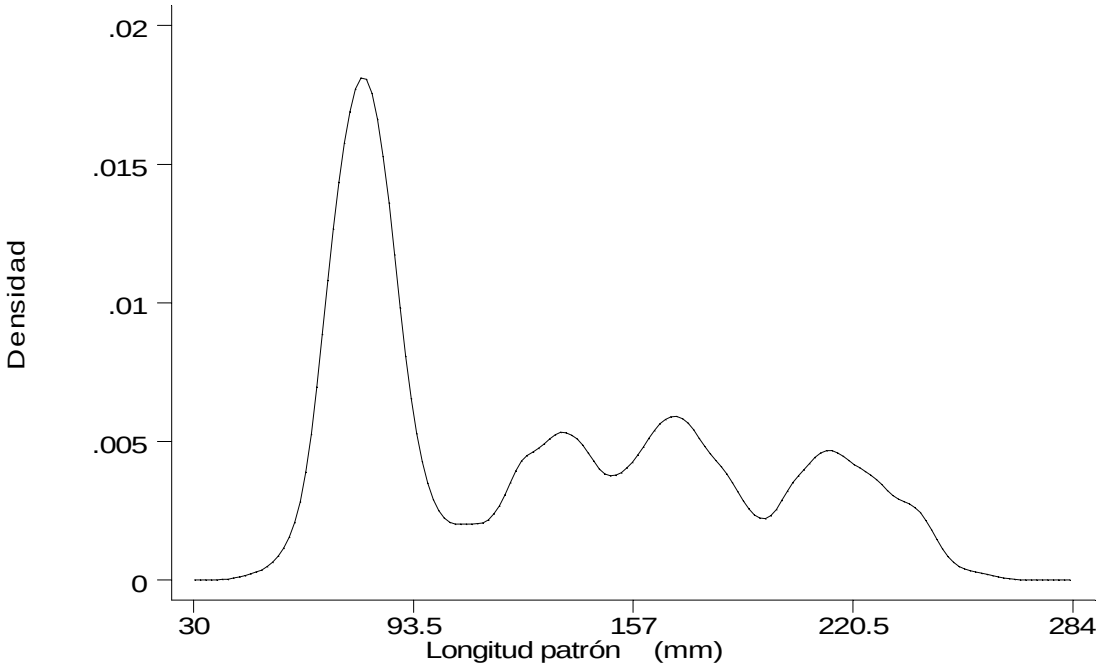

Figura 5.5 Estimación de densidad por kernel Gaussiano basada en la amplitud de banda sugerida por la validación cruzada *h* = 4 para los datos de longitud de bagres hembras e indeterminados (*n* =1116)

Recordando a Scott (1992) la concordancia en los valores de amplitud de banda óptima sugeridos por la VCMC y VCS debe tomarse seriamente. Todos los resultados anteriores proporcionan una evidencia sólida para una distribución multimodal en el conjunto de datos de longitud estándar de bagres. En la sección siguiente se presenta apoyo adicional a esta aseveración.

Al aplicar la prueba de unimodalidad del "estadístico dip" calculada por medio del programa en Fortran (Hartigan, 1985) obtenido de Statlib en la Internet y la función proporcionada por el Dr. Dario Ringach (comunicación personal) dio el valor de 0.02782 para los datos de longitud de bagre. Utilizando un argumento semejante al dado por Hartigan y Hartigan (1985) para su ejemplo, la hipótesis de unimodalidad es rechazada.

Es posible aplicar la prueba de Silverman. Los cálculos fueron efectuados con las macros y programas del paquete estadístico Stata (Statacorp, 1995) para estimación de densidad por HDP y PPPR presentados en los Capítulos 3 y 7. Para efectuar el muestreo repetitivo bootstrap se utilizaron programas adicionales para Stata para obtener $B = 100$ muestras repetitivas suavizadas de tamaño n para cada amplitud de banda crítica. El conjunto completo de programas se presenta en Salgado-Ugarte *et al.* (1997).

El Cuadro 5.4 muestra las amplitudes críticas de banda y sus correspondientes valores p para los datos de longitud de bagres. Estos resultados indican que la distribución de las longitudes es consistente con la ocurrencia de cuatro modas. De este cuadro puede observarse que la secuencia de valores de p no es estrictamente monótonamente creciente, pero después de cuatro modas no alcanza un valor menor a 0.57. Esto proporciona una evidencia sólida para apoyar la multimodalidad de los datos además de asegurar el número de las modas. Por otra parte, como se mencionó anteriormente, esta prueba se sabe que es más bien conservativa (Silverman, 1983) y por tanto puede subestimar el número de modas. Sin embargo, proporciona un valor mínimo confiable en el número de modas de un conjunto de datos (Roeder, 1990).

En los datos de longitud de bagres cuatro modas ocurren para h entre 12.10 y 3.65. Debido a la concordancia de los valores óptimos para h sugeridos por la VCMC y VCS, se prefirió no utilizar una amplitud de banda intermedia (por ejemplo $h = 8$), sino la amplitud de banda recomendada por la validación cruzada ($h = 4$).

Cuadro 5.4 Amplitudes críticas de banda y niveles estimados de significancia para los datos de longitud de bagres (hembras e indeterminados) $n = 1136$

Número de modas	Amplitud crítica de banda	Valores de p
1	24.46	0.00
2	15.88	0.00
3	12.10	0.00
4	3.65	0.70
5	3.39	0.57
6	2.75	0.86
7	2.63	0.81
8	2.62	0.57
9	2.35	0.71

NOTA: Los valores de p se obtuvieron de $B = 100$ muestras repetitivas bootstrap de tamaño 1116

5.3 Algunos comentarios adicionales

Se incluyen en este capítulo la aplicación de la prueba de multimodalidad a tres conjuntos de datos reportados en la literatura para evaluar su grado de modalidad con el procedimiento de Silverman: los datos de meteoritos condritas (tomados de Scott, 1992), los datos de grosor de timbres postales mexicanos antiguos (tomados de Izenman y Sommer, 1988) y los datos de velocidad de galaxias (de Roeder, 1990). Los programas presentados en la presente obra (Capítulo 7) fueron utilizados y los resultados se incluyen en los cuadros 5.5 a 5.7. Salvo algunas diferencias menores en los valores de amplitudes críticas de banda (debidas a diferencia en el procedimiento para el cálculo de la estimación de densidad por kernel) se obtuvieron las mismas conclusiones que las de los reportes originales. También se analizó la prueba de multimodalidad para los datos del géiser "Old faithful" del parque nacional de Yellowstone en los E:U.A. (adaptados de Härdle, 1991). Hay que notar que para este conjunto de datos el número de repeticiones bootstrap para cada banda crítica es de 600, lo

que representa el valor más alto reportado para la prueba de Silverman hasta el presente y sugiere que la duración de los períodos eruptivos del géiser se distribuye en forma bimodal.

Cuadro 5.5 Amplitudes críticas de banda y niveles de significancia estimados para los datos de condritas (Scott, 1992) *n* = 22

Número de modas	Amplitudes críticas	Valor de *p*
1	2.40	0.16
2	1.83	0.06
3	0.69	0.72
4	0.47	0.73

NOTA: Los valores de *p* se obtuvieron de *B* = 100 muestras repetitivas bootstrap de tamaño 22

Cuadro 5.6 Amplitudes críticas de banda y niveles de significancia estimados para los datos de grosor de timbres postales (Izenman & Sommer, 1988) *n* = 485

Número de modas	Amplitudes críticas	Valor de *p*
1	0.00667	0.00
2	0.00331	0.26
3	0.00300	0.05
4	0.00282	0.00
5	0.00253	0.01
6	0.00246	0.00
7	0.00148	0.52
8	0.00138	0.19
9	0.00105	0.62

NOTA: Los valores de *p* se obtuvieron de *B* = 100 muestras repetitivas bootstrap de tamaño 485

Cuadro 5.7 Amplitudes críticas de banda y niveles de significancia estimados para los datos de velocidad de galaxias (Roeder, 1990) *n* = 82

Número de modas	Amplitudes críticas	Valor de *p*
1	3037	0.000
2	2447	0.005
3	920	0.555
4	875	0.203
5	721	0.193
6	664	0.113
7	447	0.343

NOTA: Los valores de *p* se obtuvieron de *B* = 400 muestras repetitivas bootstrap de tamaño 82

Cuadro 5.8 Amplitudes críticas de banda y niveles de significancia estimados para los datos de duración de erupciones del géyser "Old faithful" (duración en minutos; Härdle, 1992) *n* = 272

Número de modas	Amplitudes críticas	Valor de *p*
1	0.830	0.000
2	0.127	0.495
3	0.084	0.948

NOTA: Los valores de *p* se obtuvieron de *B* = 600 muestras repetitivas bootstrap de tamaño 272

Capítulo 6. Regresión no paramétrica: estimadores por kernel, *ASH-WARP* y *k-NN*

6.1 Introducción

La regresión es sin duda el procedimiento estadístico más utilizado (Scott, 1992). Tiene numerosas variaciones en relación con la naturaleza de los datos analizados. En el paquete estadístico Stata existe una colección completa de comandos y archivos ado para ejecutar la mayor parte de los métodos de regresión existentes, incluyendo por ejemplo, regresión por cuantiles (mediana), regresión robusta y regresión logística en adición a la regresión ordinaria por mínimos cuadrados. Además, los programas recientemente propuestos para calcular Modelos Lineales Generalizados (denominados **glm** y **glmr**, Hilbe, 1993, 1994a; Royston, 1994a) permiten tener acceso a estas modernas y flexibles herramientas para análisis de datos basadas en la regresión con dos enfoques filosóficos diferentes (Royston, 1994b; Hilbe, 1994b).

Por otra parte, entre los procedimientos no paramétricos, los estimadores por kernel han sido reconocidos como una importante herramienta exploratoria y analítica. Härdle (1990) agrega que los estimadores por kernel poseen dos características que los distingue de otros estimadores no paramétricos: son de implementación simple y son entendibles a un nivel intuitivo. En capítulos anteriores se han presentado varios programas para calcular estimadores de densidad por kernel aplicados para el análisis de datos univariados por medio de algoritmos directos de cálculo intensivo y algoritmos discretizados o interpolados, además del eficiente procedimiento basado en los HDP-PPPR. En este Capítulo, relativo a los datos bivariados, se presenta una breve revisión de algunos estimadores de regresión no paramétrica. Los programas correspondientes para el paquete estadístico Stata en varios casos combinados con programas ejecutables en Turbo Pascal se incluyen en el Capítulo 7.

De acuerdo con Silverman (1985), el análisis de regresión proporciona un medio para explorar y presentar relaciones bivariadas, da predicciones y permite encontrar propiedades interesantes de la ecuación resultante. En este contexto, un estimador no paramétrico es deseable debido a que no obliga a que el modelo pertenezca a una clase rígidamente definida, sino que deja que los datos “hablen por sí mismos” en la elección del modelo. En algunos casos, el estimador no paramétrico puede sugerir un modelo paramétrico adecuado (tal como la regresión lineal simple); en otros, puede ser la forma de descubrir una función media muy compleja (no lineal). La figura 6.1 muestra la curva que resulta de utilizar uno de los programas mencionados aplicados a los datos sobre las erupciones del géiser “Old faithful” del parque de Yellowstone (E.U.A.). En este momento sólo se destaca el patrón de dos niveles sugerido por la curva ajustada y se dejan los detalles en su cálculo para secciones posteriores.

En la figura 6.2 se presenta una relación más complicada modelada de manera no paramétrica. Estas observaciones consisten de lecturas de acelerómetro a través del tiempo en un experimento sobre la eficacia de cascos durante choques de motocicleta simulados (datos discutidos en Silverman, 1985). La curva ajustada con uno de los programas aquí presentados da una clara indicación de la tendencia general de los datos a pesar de que no sigue de cerca la primera curvatura descrita por los datos y de que se vuelve más variable a valores posteriores de tiempo.

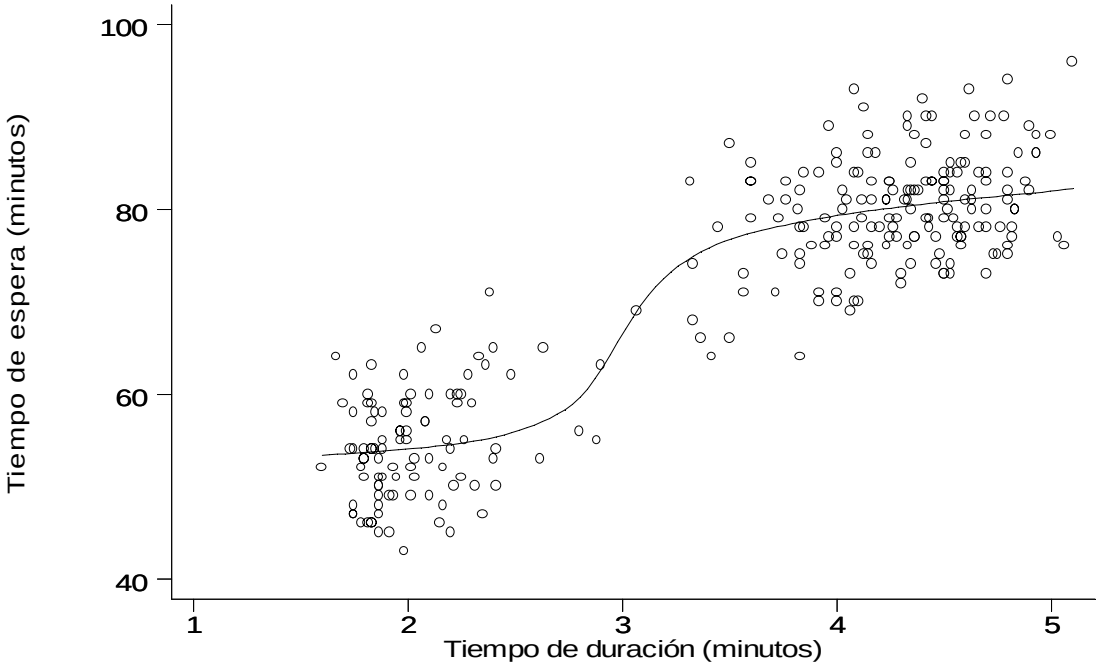

Figura 6.1 Regresión no paramétrica (estimación Nadaraya-Watson) para los datos del géiser "Old faithful", tiempo de espera para la siguiente erupción basado en el tiempo de duración de la última erupción, utilizando el programa kernreg.ado, con *h* = 0.8, kernel cuártico (biponderado) y 100 puntos para la estimación.

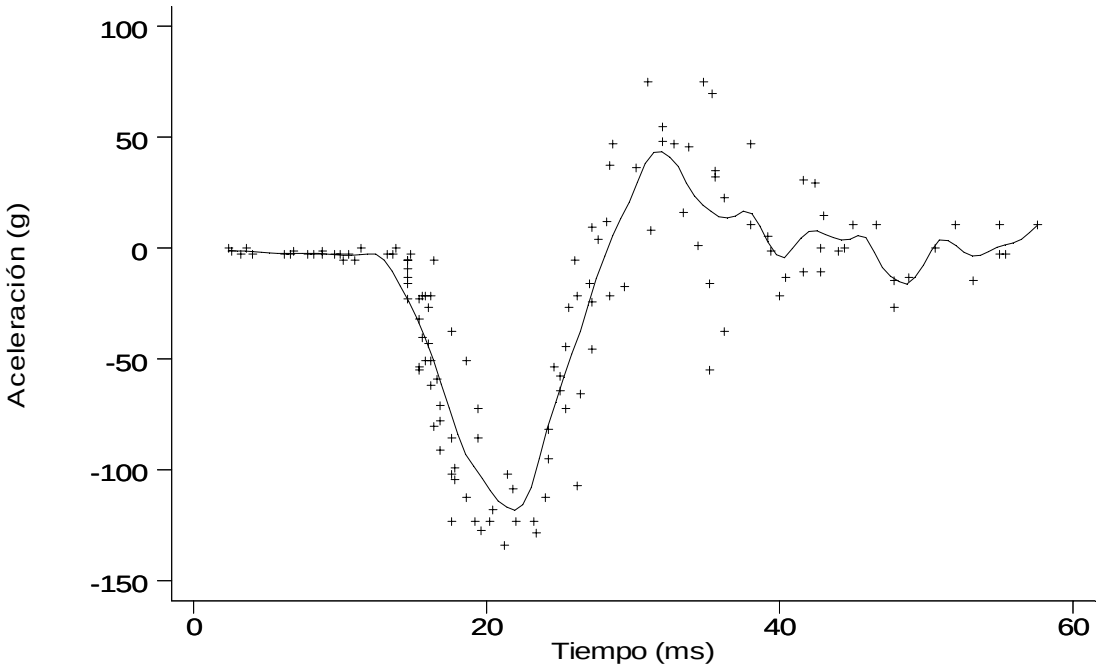

Figura 6.2 Regresión no paramétrica (estimación Nadaraya-Watson) para los datos de choques de motocicleta, aceleración dependiendo del tiempo, utilizando el programa kernreg.ado con *h* = 2.4, kernel cuártico (biponderado) y 100 puntos para estimación.

Finalmente se considera un conjunto de datos simulados generado por los siguientes comandos en Stata:

```
. set obs 200
. set seed 12345
. generate age=2.5*uniform()
. generate error= invnorm(uniform())*sqrt(100)
. sort age
. generate ngt=age*0.27
. generate tprime=age-ngt
. generate capq=2*_pi/(1-ngt)
.  generate  smaq=0.76*(tprime-0.1)+(0.76/capq)*(sin(capq*(tprime-
0.16))-sin(capq*(.1-.16)))
. generate ltfun=156*(1-exp(-smaq))
. generate ltsim=lt + error
. drop if ltsim<0 | lt < 0
```

Este conjunto de datos simulados se grafica en la figura 6.3. Las observaciones de respuesta se han generado por una función modificada para el crecimiento de von Bertalanffy (*FCVB*) la cual considera un cese estacional de crecimiento como lo describen Pauly *et al.* (1992):

$$L_t = L_\infty [1 - \exp(-q)]$$

en la cual

$$q = K(t' - t_0) + K/Q[\sin Q(t' - t_s) - \sin Q(t_0 - t_s)]$$

donde $Q = 2\pi/(1 - NGT)$. La variable *t'* se obtiene restando de la edad real (*t*) el tiempo total de no crecimiento que ocurre hasta la edad *t*. Podemos identificar en esta relación bivariada que la variable de respuesta es la longitud a la edad ($Y_i = L_{ti}$), y la variable independiente es la edad ajustada ($X_i = t'_i$).

Los valores de la variable independiente están uniformemente distribuídos en el intervalo [0,2.5] y los errores sumados son valores independientes distribuidos en forma Gaussiana con $\mu = 0$ y $\sigma^2 = 100$. El diagrama bivariado de dispersión para los valores simulados (función + error) y la función original se muestra en la Figura 6.3.

La serie de comandos del programa Stata se han incluido en una macro para automatizar la simulación de crecimiento (excepto la eliminación de los valores que proporcionan valores de longitud < 0). La macro está en el archivo **growsim.ado** descrito en el Capítulo 7.

La estimación de la regresión por kernel incluida en la Figura 6.3 es muy cercana a la respuesta funcional aunque podemos apreciar que se desvía donde el número de puntos de *X* es escaso y alguna influencia de valores extraordinarios de *Y*.

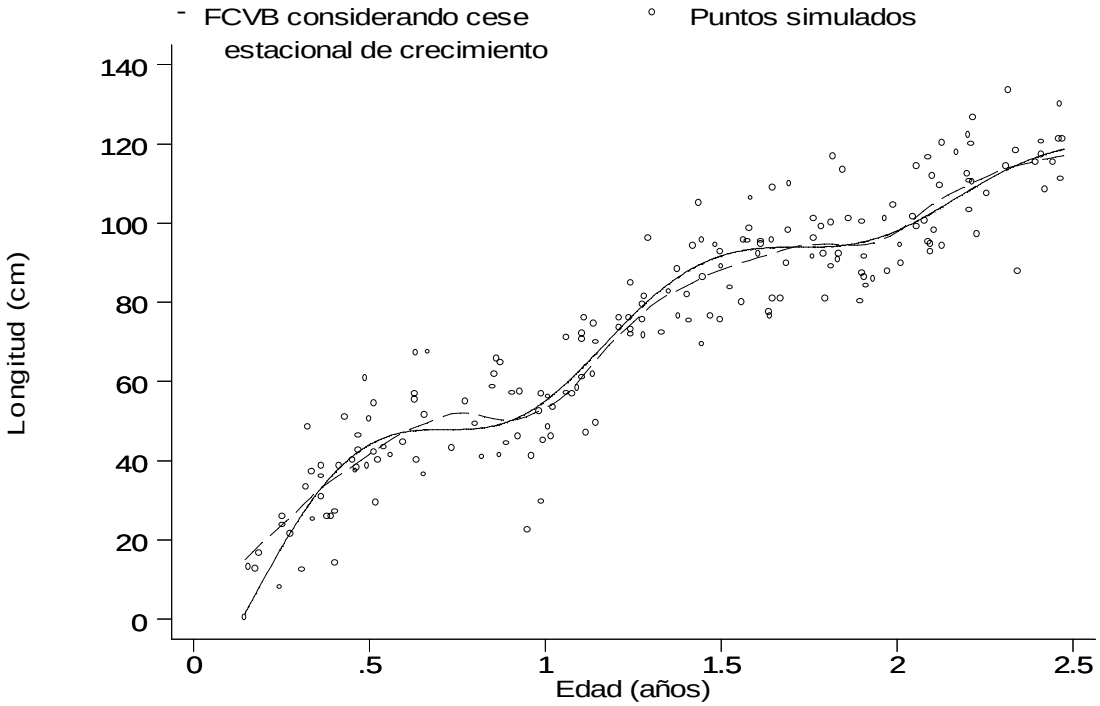

Figura 6.3 Función de crecimiento de von Bertalanffy (*FCVB*) considerando cese estacional de crecimiento de peces (línea continua), 186 pares edad-longitud simulados (círculos) y estimación de regresión por kernel (línea discontinua) utilizando una amplitud de banda de 0.22, kernel cuártico (biponderado) y 100 puntos de estimación

6.2 El enfoque tradicional

La regresión para *n* puntos $\{(X_i , Y_i)\}$ i = 1 a n está indicada por la siguiente expresión (Härdle, 1990, 1991; Scott, 1992):

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

donde m es la función regresión (desconocida) y ε es una variable aleatoria que denota la variación de Y alrededor de $m(X)$, que es la curva media de regresión $E[Y \mid X = x]$. Esta curva media condicional puede representarse como

$$E\,[Y\,|X = x] = \frac{\int yf(x,y)dy}{f(x)} \tag{2}$$

Donde $f(x,y)$ indica la densidad conjunta de (X,Y) y $f(x)$ es la densidad marginal de X. Cuando la distribución conjunta es Gaussiana con media de cero, la curva de regresión es lineal (Härdle, 1991).

La figura 6.1 muestra (sin considerar la curva ajustada) el diagrama bivariado de dispersión de los datos de tiempo duración contra el tiempo de espera en minutos para la siguiente erupción del

géiser “Old faithful” del Parque Nacional de Yellowstone (E.U.A.). Este conjunto de datos ha sido analizado por varios autores (Weisberg, 1980, Silverman, 1985, Härdle, 1991). En esta figura podemos distinguir una relación directamente proporcional entre las variables además de la existencia de dos grupos de concentración en los datos. Podemos intentar cuantificar la relación entre los tiempos de duración y espera aplicando el enfoque tradicional.

Siguiendo a Cook y Weisberg (1982), Silverman (1985) y Härdle (1991), se intenta en primer lugar un modelo lineal:

$$Y_i = \text{intercepto} + \alpha X_i + \varepsilon_i \qquad (3)$$

Utilizando el paquete estadístico Stata, introducimos los siguientes comandos:

```
. regress wait dura

Fuente   |       SC       gl       CM                  Nímero de obs =     272
---------+------------------------------               F(  1,   270) = 1162.17
  Modelo |   40644.437     1   40644.437               Prob > F      =  0.0000
Residual |  9442.68066   270  34.9728913               R^2           =  0.8115
---------+------------------------------               R^2 ajustada  =  0.8108
   Total |  50087.1176   271  184.823312               Error de est. =  5.9138

------------------------------------------------------------------------------
    wait |      Coef.  Err. estándar    t     P>|t|    [Intervalo de Conf. 95%]
---------+--------------------------------------------------------------------
    dura |   10.72958    .314737     34.091   0.000       10.10993    11.34923
   _cons |   33.47437   1.154822     28.987   0.000       31.20077    35.74798
------------------------------------------------------------------------------
```

Estos resultados indican una regresión estadísticamente significativa con un coeficiente de determinación alto. Los valores de *Y* estimados se muestran como la línea continua en la Figura 6.5. Cook y Weisberg notaron algunas inconsistencias con este modelo simple (como cierta curvatura en el gráfico de residuos) y sugirieron que a valores altos de *x* los valores lineales predichos fueron demasiado grandes. La curva no paramétrica de la figura 6.1 indica claramente esta desviación y además sugiere un modelo paramétrico alternativo que considera los cúmulos observados en los puntos. Este modelo alternativo puede especificarse como sigue:

$$Y_i = \text{intercepto} + \alpha X_i I(X_i \leq 3.3) + \beta X_i I(X_i > 3.3) + \gamma I(X_i > 3.3) + \varepsilon \qquad (4)$$

En este modelo se ajustan a los puntos dos relaciones lineales (una por cada cúmulo). Los cúmulos se definen como C_1 = (-∞, 3.3] y C_2 = (3.3, ∞). De esta forma el primer coeficiente es el intercepto para el ajuste lineal en C_1 con pendiente α; β es la pendiente del ajuste lineal en C_2 y γ representa la diferencia de interceptos entre los ajustes.

Esto puede lograrse en el paquete estadístico Stata. Una es crear las variables modificadas adecuadas (variables “dummy”) y posteriormente considerarlas como variables independientes en la regresión. Nombrando a C_1 como “dura1”, C_2 como “dura2” y definiendo “dura2d” como 1 para duraciones por encima de 3.3 y cero en cualquier otro caso, podemos introducir.

```
. generate dura1 = 0
. generate dura2 = 0
. generate dura2d = 0
. replace dura1 = dura if dura <= 3.3
. replace dura2 = dura if dura > 3.3
. replace dura2d = 1 if dura > 3.3
. regress wait dura1 dura2 dura2d
```

```
  Fuente |       SC       gl      CM                  Número de obs =     272
---------+------------------------------               F(  3,   268) =  442.34
  Modelo |  41671.3234     3  13890.4411               Prob > F      =  0.0000
Residual |  8415.79429   268  31.4022175               R^2           =  0.8320
---------+------------------------------               R^2 ajustada  =  0.8301
   Total |  50087.1176   271  184.823312               Error de est. =  5.6038

------------------------------------------------------------------------------
    wait |      Coef. Error estándar    t     P>|t|   [Intervalo de Conf. 95%]
---------+--------------------------------------------------------------------
   dura1 |   7.384168    1.99572      3.700   0.000       3.454882    11.31345
   dura2 |   5.247466   1.061537      4.943   0.000       3.157452     7.33748
  dura2d |    17.9808   6.167388      2.915   0.004         5.8381    30.12351
   _cons |   39.51536    4.12751      9.574   0.000       31.38888    47.64184
------------------------------------------------------------------------------
```

Terminamos de nuevo con una regresión altamente significativa, pero esta vez con dos líneas de menor pendiente comparadas con las del Modelo I. En la figura 5.4 se muestran los valores predichos de este modelo como dos líneas (una para cada cúmulo de puntos). La diferencia positiva entre los interceptos (coeficiente "dura2d" en el cuadro de resultados) sugiere una ordenada al origen mayor para el cúmulo C_2. Podemos probar la significancia de esta diferencia utilizando otro comando del paquete estadístico Stata:

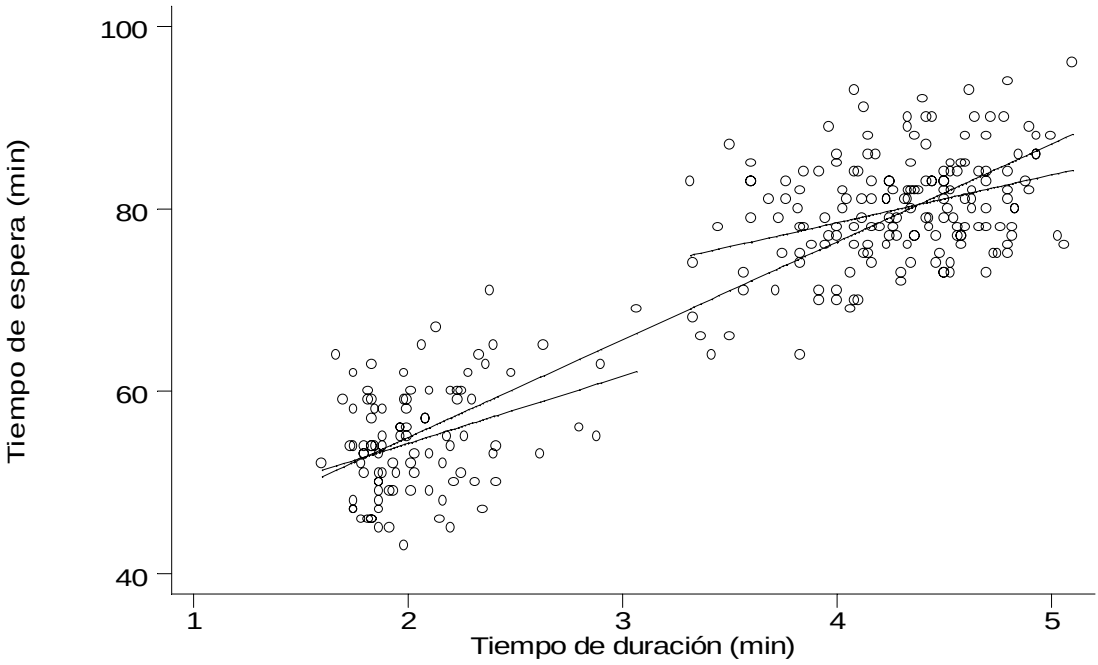

Figura 6.4 Dos modelos de regresión lineal por mínimos cuadrados ajustados a los datos del géiser "Old faithful". El modelo de regresión lineal simple con el total de puntos se representa por la línea continua y la regresión considerando los dos cúmulos de puntos está representada por las líneas separadas.

```
. test dura2d

(1)     dura2d = 0.0

       F(  1,   268) = 8.50
            Prob > F = 0.0039
```

Este resultado indica que los interceptos de ambos cúmulos son diferentes y sugiere que el modelo lineal simple no es correcto.

Como Silverman (1985) ha enfatizado antes en un ajuste relacionado a los mismos datos, la curva no paramétrica es un importante paso exploratorio hacia la elección del modelo final. Si esta curva no paramétrica no estuviera presente pudiera ser muy difícil apreciar una relación no lineal en el diagrama bivariado de dispersión.

En relación con los datos de impacto, se requeriría un modelo paramétrico mucho más complicado (un análisis detallado utilizaría procedimientos de series de tiempo). En esta obra no se profundiza al respecto. En el caso de los datos de crecimiento, varios pasos se requieren para determinar un modelo paramétrico. Se incluye una regresión lineal preliminar y dos ciclos de estimación no lineal para calcular los parámetros de la función de crecimiento de von Bertalanffy considerando cese estacional del crecimiento (ver Pauly *et al.*, 1992 para detalles).

Se presentará a continuación el enfoque no paramétrico.

6.3 El estimador Nadaraya-Watson (regresión por kernel)

El enfoque no paramétrico para la regresión se relaciona con la estimación de densidad. En la estimación de densidades a un valor dado de x, se consideran los puntos en un pequeño intervalo alrededor de x (ancho de intervalo en histogramas o amplitud de banda h para los estimadores de densidad por kernel). En la regresión, la variable de respuesta Y es ponderada en un cierto vecindario de x. Utilizando una notación simplificada es posible indicar la forma general para la regresión no paramétrica:

$$\hat{m}_h(x) = n^{-1} \sum_{i=1}^{n} W_{hi}(x) Y_i \qquad (5)$$

Esta expresión, aplicable a la mayor parte de los métodos no paramétricos de regresión, puede entenderse como un promedio ponderado de la variable de respuesta Y_i con pesos $W_{hi}(x)$ que depende del procedimiento y en la distancia entre x y X_i utilizando un parámetro de suavización h.

Bajo las suposiciones usuales, una notación alternativa para la función de regresión es:

$$m(x) = E(Y|X = x) = \frac{\int y f(x,y) dy}{f(x)} \qquad (6)$$

El denominador es una densidad que puede calcularse de manera no paramétrica por medio de un estimador de densidad por kernel. La densidad conjunta contenida en el numerador es posible de estimar utilizando al kernel multiplicativo (ver Härdle, 1991 para detalles). Posteriormente, considerando la misma amplitud de banda en ambos términos del cociente llegamos a la siguiente expresión general, propuesta independientemente por Nadaraya (1964) y Watson (1964):

$$\hat{m}_h(x) = n^{-1} \frac{\sum_{i=1}^{n} K_h(x - X_i) Y_i}{n^{-1} \sum_{j=1}^{n} K_h(x - X_j)} \tag{7}$$

De acuerdo con esta definición puede notarse que (Härdle, 1991):

- Los pesos dependen de toda la muestra X a través de la estimación de densidad por kernel $f_h(x)$
- Las observaciones Y_i obtienen mayor peso donde son escasas las correspondientes X_i
- Cuando el denominador es cero, entonces el numerador también es cero y la estimación se define como cero
- Al $h \rightarrow 0$, X_i converge a Y_i y la estimación se iguala a una interpolación lineal entre los datos
- Al $h \rightarrow \infty$, la función ponderada converge a 1 para todos los valores de x, y la estimación converge al valor medio de Y.
- Como en la estimación de densidad por kernel el valor de h determina la suavidad de la estimación.

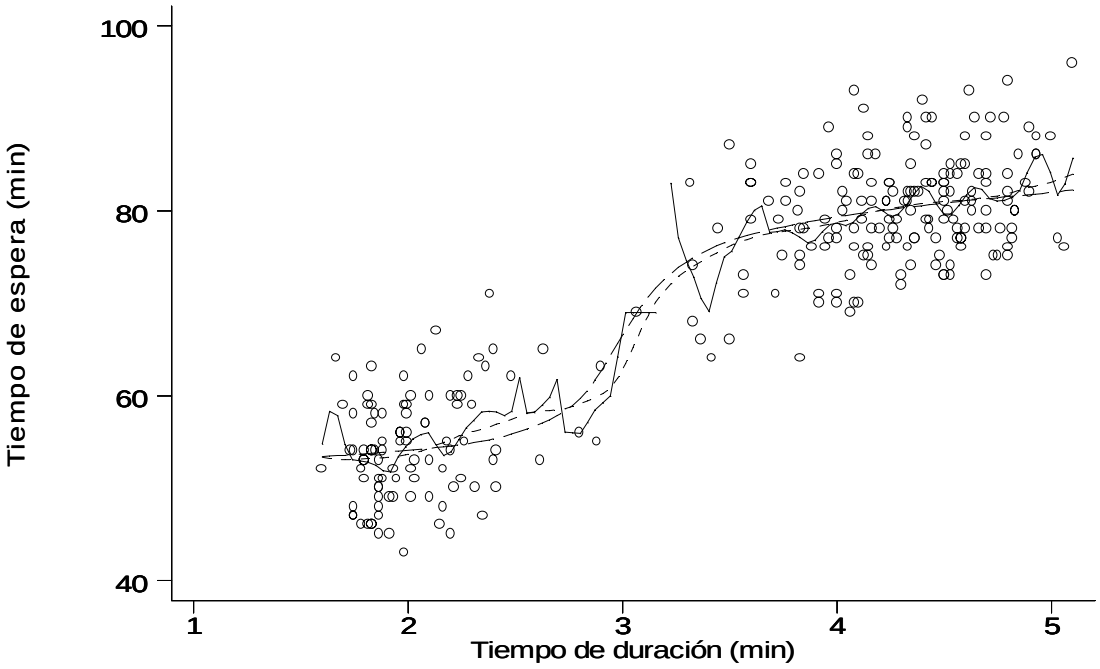

Figura 6.5 Regresión por kernel (estimador Nadaraya-Watson) para los datos del géiser "Old faithful", usando el programa kernreg.ado con el kernel cuártico (biponderado), 100 puntos para estimación y tres valores para *h*: 0.1 (líneas continuas), 0.4 (línea con puntuación corta) y 0.8 (línea con puntuación larga)

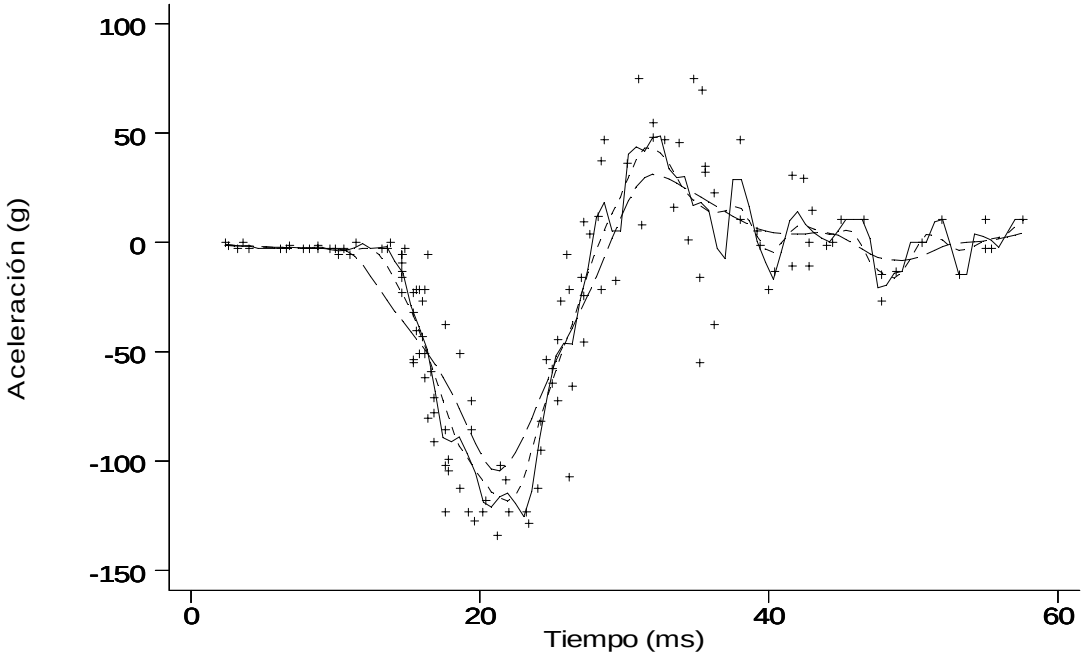

Figura 6.6 Regresión por kernel (estimador Nadaraya-Watson) para los datos de impactos simulados de motocicleta, usando el programa kernreg.ado con el kernel cuártico (biponderado), 100 puntos para estimación y tres valores para *h*: 1 (líneas continuas), 2.4 (línea con puntuación corta) y 5 (línea con puntuación larga)

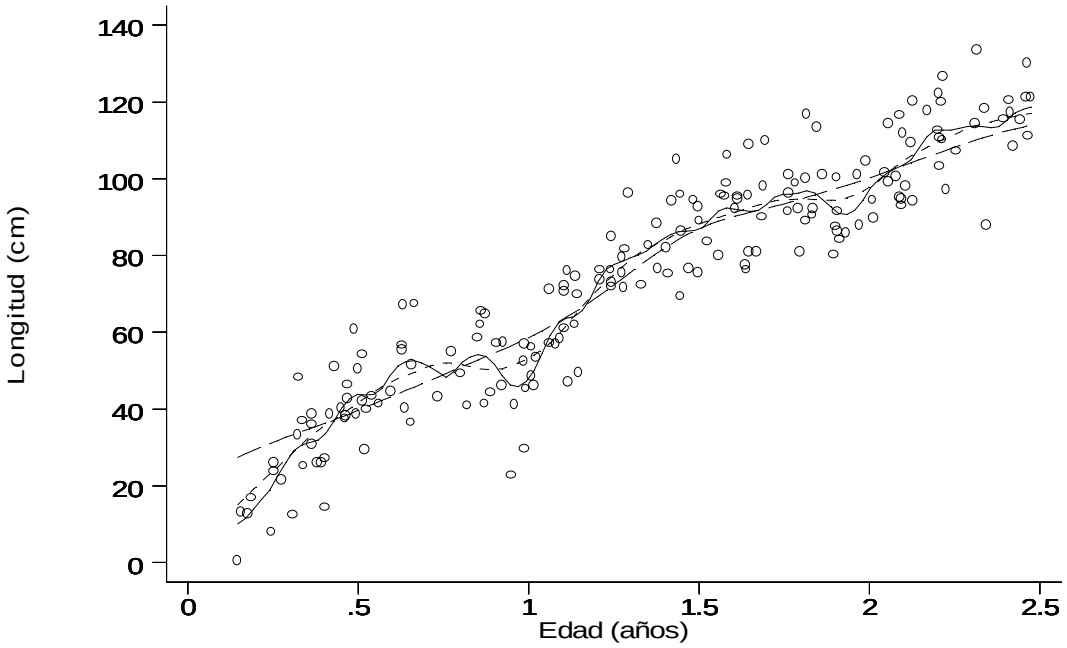

Figura 6.7 Regresión por kernel (estimador Nadaraya-Watson) para los datos simulados de crecimiento, usando el programa kernreg.ado con el kernel cuártico (biponderado), 100 puntos para estimación y tres valores para *h*: 0.1 (líneas continuas), 0.22 (línea con puntuación corta) y 0.5 (línea con puntuación larga)

En la figura 6.5 se presentan los resultados obtenidos con el programa para el paquete estadístico Stata **kernreg.ado** aplicado a los datos del géiser "Old Faithful" con tres diferentes amplitudes de banda (h = 0.1, 0.4 y 0.8), 100 puntos para estimación y el kernel cuártico (biponderado) reproduciendo los resultados de Härdle (1991). Vale la pena recordar que al utilizarse menos puntos para la estimación, el error de ésta se incrementa. Por lo tanto debe tenerse cuidado en la práctica para utilizar una trama de puntos lo suficientemente fina. El valor sugerido de 100 deberá ser adecuado en muchos casos, pero si el intervalo de valores de x es muy amplio podrá ser necesario probar otros números. El uso de la interpolación lineal para conectar a los puntos estimados conducirá a resultados más precisos (para una discusión completa sobre este tópico en relación a la estimación de densidad por kernel consultar a Jones, 1989). De manera semejante a la estimación de densidad univariada, al incrementarse el valor del parámetro de suavización, el ruido en la estimación disminuye. De hecho, usando h = 0.1 conduce a una estimación ruidosa con un valor faltante indicado por la interrupción en la línea continua. No obstante, todas las curvas muestran una respuesta promedio ligeramente creciente dentro de cada cúmulo de datos.

Se utilizó el programa **kernreg.ado** con los datos de simulación de impactos en motocicletas utilizando el mismo número de puntos de estimación y función ponderal (kernel cuártico) pero variando el valor de h de 1, 2.4 a 5. Los resultados se presentan en la Figura 6.6 en la cual podemos notar que 2.4 es una amplitud de banda razonable comparada con las curvas ruidosas y sobresuavizada.

Utilizando el mismo número de puntos y kernel, la figura 6.7 muestra los resultados correspondientes a los datos de simulación de crecimiento utilizando h = 0.1, 0.22 y 0.5. La línea con la amplitud de banda mayor es muy cercana a una curva común de la función de crecimiento de von Bertalanffy, mientras que la amplitud de banda intermedia (0.22) conduce a una curva con pendiente decreciente con varios ciclos (como la función original de la cual se generaron los datos). Considerando el largo y complicado algoritmo para ajustar una función de crecimiento de von Bertalanffy que tome en cuenta cese estacional de crecimiento, vale la pena destacar la forma relativamente sencilla para obtener una curva similar por medio de la regresión no paramétrica por kernel.

6.4 Utilizando el Promedio Ponderado de Puntos Redondeados (*PPPR*) para calcular la regresión por kernel (Nadaraya-Watson)

Como se ha descrito en otras publicaciones o secciones de esta obra (Salgado-Ugarte, *et al.*, 1993; Capítulo 2) el cómputo directo de estimaciones de densidad por kernel requiere una gran cantidad de cálculos y tiempo. Al notar que el estimador de regresión Nadaraya-Watson es una combinación de una densidades univariadas y recordando que los procedimientos ***HDS*** y ***PPPR*** pueden aproximar de manera eficiente densidades con kerneles soportados en [-1,1], es posible definir, en general, la aproximación por ***PPPR*** de la estimación Nadaraya-Watson (consultar detalles en Härdle, 1991 y Scott, 1992). Si se considera una función índice que proporcione el índice del intervalo pequeño al cual pertenece x:

$$l(x) = j \Leftrightarrow x \in [(j - 1/2)\delta,\ (j + 1/2)\delta)$$

y definiendo la suma de las observaciones de respuesta correspondiente a X_i en el intervalo B_z como:

$$Y_{\bullet z} = \sum_{i=1}^{n} Y_i \, I\,(X_i \in B_z)$$

obtenemos para las $x \in B_{l(x)}$:

$$\hat{m}_{M,K}(x) = \frac{\sum_{l=1-M}^{M-1} K(\frac{l}{M}) Y_{\bullet l(x)+l}}{\sum_{l=1-M}^{M-1} K(\frac{l}{M}) n_{l(x)+l}} \qquad (8)$$

De manera similar a la sección en la que se introdujeron los procedimientos ***HDP*** y ***PPPR***, en el Capítulo 7 se presentan varios programas que utilizan este método eficiente para estimación de regresión no paramétrica basados en los algoritmos y programas descritos en Härdle (1991) y Scott (1992).

La figura 6.8 muestra la aproximación utilizando una amplitud de banda de 0.4 con *M* de 10 y kernel cuártico. Para conectar los puntos estimados podemos usar interpolación lineal (versión poligonal), pero si se requiere es posible también utilizar una versión a pasos. Puesto que la correspondiente estimación por kernel se despliega como una función continua, es preferible utilizar la versión poligonal (Härdle, 1991). Además, considerando los mismos datos, la estimación interpolante es mas precisa que la versión discretizada a pasos (Jones, 1989).

Como sucede en la estimación de densidad por kernel utilizando los ***HDP*** o el ***PPPR***, al aumentar el valor de *M*, la estimación es más cercana a la estimación por kernel. En la figura 6.9 se incluyen aproximaciones por ***PPPR*** utilizando $M = 1$, 5 y 10 además de los resultados obtenidos con el programa **kernerg.ado** (con 100 puntos). Desde valores ≥ 5 no hay diferencia apreciable entre ambas estimaciones (para los datos del géiser, utilizando la aproximación por ***PPPR*** con $M =$ 10 da como resultado 88 puntos de estimación comparados con los 100 puntos que resultan de **kernreg.ado**, pero con un notable ahorro de tiempo).

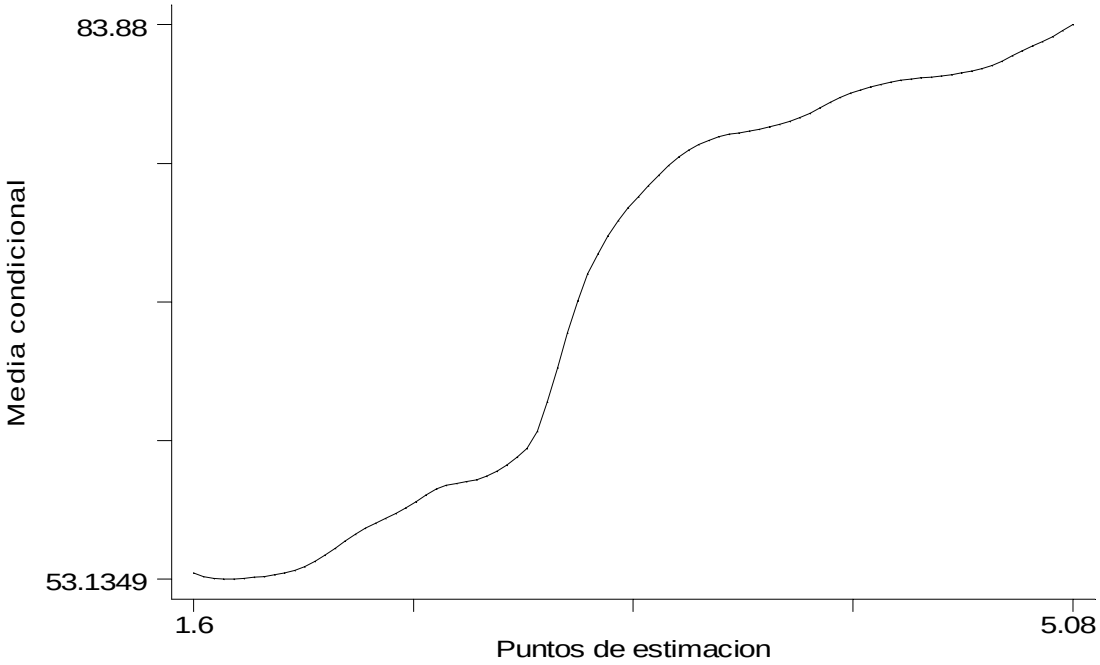

Figura 6.7 Aproximación por *PPPR* del estimador de regresión por kernel de Nadaraya-Watson para los datos del géiser "Old faithful". La amplitud de banda es 0.4, *M* = 10 y *k* = 4 (Cuártico)

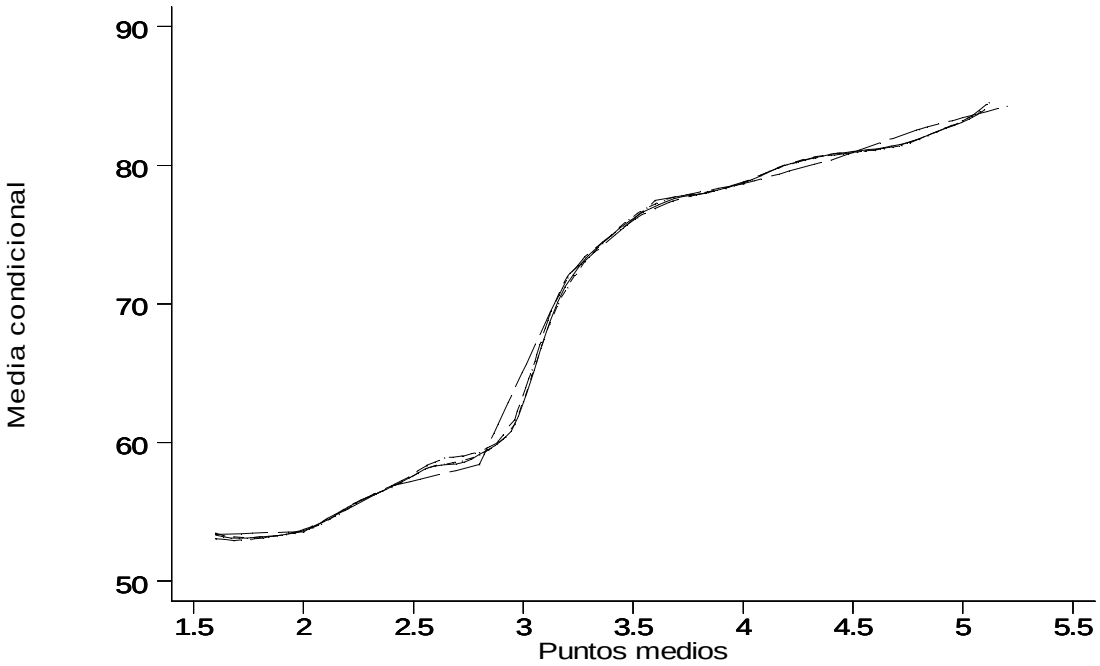

Figura 6.8 Estimador Nadaraya-Watson utilizando el algoritmo directo para regresión por kernel (línea continua) y aproximación por *PPPR* con *M* = 1 (línea con puntuación larga), *M* = 5 (línea con puntuación corta) y *M* = 10 (línea con puntuación muy corta). En todos los casos la amplitud de banda es de 0.4 con el kernel cuártico.

Se recomienda un valor de M = 10 como un buen compromiso en cuanto a velocidad de cálculo y precisión en la estimación (Härdle, 1991). A este respecto se incluye el Cuadro 6.1 con algunos reportes de tiempo utilizado para los cálculos utilizando los programas **kernreg.ado** y **warpreg.ado** en una computadora con procesador 486SX (25 MHz) con 6 Mb de RAM, sin coprocesador matemático y el programa regular de Stata (versión 3.1) corriendo bajo MS Windows 3.1.

Cuadro 6.1 Comparación del desempeño entre kernreg.ado y warpreg.ado

Archivo de datose	*n*	Programa y valores iniciales	Tiempo (en segundos)
Geyser.dta	272	kernreg, bw=0.4, kcode=4,npoints=100	388
"	"	", bw=0.65, ", "	410
"	"	", bw=0.8, ", "	420
"	"	", bw=1.75,","	475
Motcyc.dta	133	kernreg.ado, h=1, kcode=4, npoints=100	174
"	"	", bw=2.4, ", "	179
"	"	", bw=5, ", "	188
Geyser.dta	272	Warpreg, bw=0.4, M=10, kcode=4, (88 midpoints)	11
"	"	", bw=0.65, M=10 , kcode=4, (54 midpoints)	11
"	"	", bw=0.8, M=10, kcode=4, (45 midpoints)	11
"	"	", bw=1.75, M=10, kcode=4, (21 midpoints)	10
Motcyc.dta	133	warpreg.ado, h=1, M=10, kcode=4, (553 midpoints)	9
"	"	", bw=2.4, M=10, kcode=4, (231 midpoints)	7
"	"	", bw=5, M=10, kcode=4, (111 midpoints)	6

Nota: Los tiempos se tomaron desde el momento de oprimir la tecla ENTRADA una vez que se escribió el comando hasta la aparición en la pantalla del mensaje "more" en el resultado gráfico.

Los ahorros de tiempo utilizando los estimadores utilizando ***HDP-PPPR*** son impresionantes a pesar de que **warpreg.ado** escribe varios archivos temporales, llama un programa ejecutable externo el cual a su vez lee un archivo, hace los cálculos y escribe los resultados en otro archivo, para que posteriormente el paquete Stata lee estos resultados, dibuja la gráfica y llama de nuevo a los datos originales previamente grabados. En el programa **kernreg.ado** notamos una proporcionalidad directa entre h y el tiempo utilizado. Lo opuesto, una relación proporcional inversa (como era de esperarse) se observa entre h y el tiempo utilizado al utilizar **warpreg.ado**, especialmente con el conjunto de datos relativamente pequeño de la simulación de impacto de motocicletas.

6.5 Selección de la amplitud de banda

Para efectuar una estimación de densidad por regresión no paramétrica se requiere escoger un valor para h, el parámetro de suavización. Esta selección del tamaño de la vecindad para estimación es un problema cercanamente relacionado con la selección del grado para una regresión polinomial o la selección de variables en regresión múltiple. Como existe una "negociación" entre el sesgo y la varianza, uno debe lograr un buen "compromiso" que evite el obtener una curva que siga a cada uno

de los puntos (sobre-ajuste) o por el otro lado para obtener una curva que pierda características importantes de los datos (sobresuavizado) (Altman, 1992).

Como en el caso de la estimación de la densidad, se han propuesto varios métodos para escoger la amplitud de banda en la regresión por kernel. La forma más simple es probar varios valores y escoger aquél que resulte en una estimación de interés para el analista, esto es, una elección subjetiva. Este procedimiento es altamente flexible y se usa como introducción a los estudiantes inexpertos. Sin embargo, es conveniente tener métodos objetivos para producir estimaciones automáticas o para lograr consistencia en los resultados de varios investigadores (Altman, 1992). A este respecto, hay varios métodos "plug-in" además de otros enfoques entre los que destaca la validación cruzada (estimación repetitiva dejando una observación fuera). A continuación se presenta un programa para el paquete Stata que utiliza la estimación por ***PPPR*** para estimar una amplitud de banda "automática" por medio de la combinación de funciones penalizantes y validación cruzada que utiliza algoritmos modificados de Härdle (1991).

En la estimación de densidad, el enfoque utilizado es tener una aproximación de la amplitud de banda que minimiza al error cuadrado integrado medio asintótico (**ECIMA**) por medio del cálculo del error de secuencias de amplitudes. En la regresión, sin embargo, existen otros estimadores de la discrepancia entre $\hat{m}_h$ y la curva desconocida descrita por *m*:

- Error Cuadrado Promedio ***ECP*(*h*)**
- Error Cuadrado Integrado ***ECI*(*h*)**
- Error Cuadrado Condicionado ***ECC*(*h*)**
- Error Cuadrado Integrado Medio ***ECIM*(*h*)**

Cualquiera de estas medidas de distancia puede aplicarse a la estimación de la amplitud de banda óptima puesto que conducen asintóticamente a la misma amplitud de banda (Härdle y Marron, 1986). Entonces, por conveniencia, se sugiere utilizar la distancia más fácil de calcular, o sea el Error Cuadrado Promedio (Härdle, 1990, 1991). No obstante, debido a ciertas dificultades con las funciones minimizantes resultantes, la técnica generalmente empleada es una combinación de los términos de corrección de sesgo y la estimación "dejando una observación fuera" $\hat{m}_{hi}(X_i)$. Este enfoque hace uso de una de los selectores por función penalizante listados en el Cuadro 6.2 cuyo propósito es penalizar amplitudes de banda demasiado pequeñas. En esta técnica, el error de predicción (derivado del ***ECP***) se ajusta por los términos de corrección que resultan en una estimación de $G(h)$, el error ajustado de predicción. Estas funciones dan pesos diferentes al sesgo y varianza de $\hat{m}_h$. La función *S* de Shibata reduce el sesgo mientras que la *T* de Rice reduce la varianza. Sin importar estas diferencias, cualquiera de estos selectores conduce substancialmente a la misma amplitud de banda "óptima". Como puede verse en el Cuadro 6.2, la validación cruzada puede entenderse en términos de funciones penalizantes. Aunque se sabe que la velocidad de convergencia hacia este óptimo es lenta, no es posible derivar alguno mejor, y el valor estimado seleccionado a menudo trabaja bien aún para tamaños de muestra moderados (Härdle, *et al.*, 1988; Härdle, 1990; Altman, 1992).

Cuadro 6.2 Funciones penalizantes (Adaptadas de Härdle, 1991)

Nombre de la función penalizante	Expresion general
1) Selector de modelo de Shibata, (Shibata, 1981)	$\Xi S(u) = 1 + 2u$
2) Validación cruzada generalizada (Craven y Whaba, 1979; Li, 1985)	$\Xi GCV(u) = 1/(1 - u)^2$
3) Criterio de información de Akaike (Akaike, 1970)	$\Xi AIC(u) = \exp(2u)$
4) Error finito de predicción (Akaike, 1974)	$\Xi FPE(u) = (1 + u)/(1 - u)$
5) T de Rice (Rice, 1984)	$\Xi T(u) = 1/(1 - 2u)$

Un algoritmo para calcular directamente $G(h)$ es fácil de implementar, pero con la gran desventaja de que los pasos para una h específica son cuadráticos en el número de observaciones (Härdle, 1991). Por lo tanto, es deseable un método más eficiente, el cual afortunadamente existe y se basa en la aproximación por el método ***HDP-PPPR***. En este ámbito, la optimización del ***ECM*** en relación a la cantidad de suavización es equivalente a optimizar M. Para lograr esto, es necesario suponer una amplitud dada para δ la cual describe el nivel de precisión de la escala de medición utilizada en los datos.

Definiendo $W_{Mi}[z]$ y $\hat{m}_M[z]$ como los valores correspondientes de $W_{Mi}[x]$ y $\hat{m}_M[x]$ en el intervalo z, y suponiendo la función de peso $w(\bullet)$ es constante dentro de los intervalos se obtiene:

$$Y_{\bullet z}^2 = \sum_{i=1}^{n} Y_i^2 \, I(X_i \in B_z) \tag{9}$$

Finalmente, la expresión general para la función penalizante selectora para la regresión por ***PPPR*** es (Härdle, 1991):

$$G(M) = n^{-1} \sum_{z \in Z} \{Y_{\bullet z}^2 - 2\,\hat{m}_M[z] Y_{\bullet z} + \hat{m}_M^2[z] n_z\} \Xi(n^{-1} W_M[z]) w[z] \tag{10}$$

Como en los programas de validación cruzada para la selección de amplitud de banda óptima en la estimación de densidad (Salgado-Ugarte, *et al.*, 1995b), el algoritmo contiene los pasos típicos del método ***PPPR*** sólo que con ligeras modificaciones:

- agrupamiento de los datos en intervalos
- creación de pesos (elección de una función ponderal)
- ponderación de las frecuencias en los intervalos

La principal diferencia es que después de agrupar a los datos, se efectúa un paso repetitivo (loop) a través de un intervalo especificado de valores para el parámetro M (δ se mantiene constante), se incluye la creación de los pesos y una ponderación modificada para aplicarse sólo a los intervalos no vacíos y sus vecinos también no vacíos. Este procedimiento es lineal en el número de observaciones en contraste con el algoritmo directo que depende del cuadrado de n (Härdle, 1991).

En esta obra se presenta una modificación de los algoritmos y programas de Härdle (1991) para estimar $G(M)$. El programa para el paquete estadístico Stata **gwarpreg.ado** llama a un programa ejecutable escrito en Turbo Pascal (**gwarpren.exe**) contiene los pasos arriba citados. Este programa se presenta en el Capítulo 7.

El programa lleva a cabo los cálculos y los resultados son leídos y graficados en Stata. La primera gráfica mostrada es el valor del puntaje de $G(M)$ dependiendo de M, seguida de otra gráfica del puntaje pero ahora dependiendo del valor de la amplitud de banda. Estas dos gráficas auxilian en encontrar y seleccionar los intervalos con valores mínimos de puntaje por un procedimiento de ensayo y error. Si el intervalo analizado (definido con valores de M y por tanto de amplitud de banda) contiene un mínimo en los valores de $G(M)$ se podrá detectar en las gráficas. Además de los desplegados gráficos, el programa muestra en pantalla un listado de los cinco valores más bajos para el puntaje y sus correspondientes valores de M y amplitud de banda. De esta forma, el programa **gwarpreg.ado** proporciona toda la información necesaria para investigar y encontrar a la

amplitud de banda óptima. De otra forma, al variar el valor de delta y especificando un intervalo diferente para *M* (otros valores inicial y final) podemos localizar el mínimo de la función al observar los gráficos y los cuadros resultantes. Por ejemplo, para los datos de impacto de motocicletas, podemos especificar un valor de delta de 0.2, la función selectora de Shibata, el kernel Cuártico, un valor inicial de *M* de 5 y final de 20 así como un valor de frontera de 0.l escribimos:

```
. gwarpreg accel time, d(0.2) s(1) k(4) me(20)
```

Como resultado se obtienen dos gráficas y el cuadro siguiente:

```
Error Ajustado de Predicción G(M) para regresión PPPR Nadaraya-Watson
-------------------------------------------------------------------------
Valor de M = 8        Puntaje = 534.06671       Banda =   1.6000
Valor de M = 9        Puntaje = 537.53998       Banda =   1.8000
Valor de M = 7        Puntaje = 539.06793       Banda =   1.4000
Valor de M = 10       Puntaje = 542.87994       Banda =   2.0000
Valor de M = 11       Puntaje = 548.75659       Banda =   2.2000
```

Para este ejemplo, de las gráficas y cuadro se sugiere un valor óptimo de 1.6 para la amplitud de banda. Con el fin de corroborar este resultado, se probaron cada una de las funciones de selección y hasta donde fue posible, los mismos valores iniciales. Los resultados se resumen en el Cuadro 6.2

Cuadro 6.2 Resultados obtenidos con gwarpreg.ado usando varias funciones selectoras.

delta	Función Selectora	Tipo de Kernel	Valor inicial de M	Valor final de M	Valor de Frontera	Amplitud de banda óptima
0.2	*S*	Quartic	5	20	0.1	1.6
0.3	*VCG*	Quartic	5	20	0.1	2.4
0.2	*CIA*	Quartic	5	20	0.1	1.8
0.3	*EFP*	Quartic	5	20	0.1	2.1
0.3	*T*	Quartic	5	20	0.1	2.7

Los correspondientes resultados gráficos obtenidos con el programa **gwarpreg.ado** se incluyen en la Figura 6.8 (cada gráfica individual fue agregada con auxilio del Editor de Gráficas de Stata llamado “Stage”). Hay que señalar que debido a la necesidad de utilizar valores diferentes para delta, se optó por utilizar las gráficas del puntaje como función de los valores de *M*. Las curvas muestran tendencias similares con un bien definido valor mínimo, excepto en el caso de la *T* de Rice, la cual sugiere la existencia de un mínimo local.

Siguiendo los razonamientos de Härdle (1991) para los datos del géiser “Old Faithful”, es posible obtener una curva con dos valores mínimos indicando amplitudes de banda de 0.65 y 1.75 (Figura 6.10).

El valor de 0.22 para los datos de simulación de crecimiento fueron obtenidos utilizando este programa con un valor para **delta** de 0.02, diferentes funciones selectoras, kernel cuártico y valor final para *M* de 40. Algunas veces las funciones penalizantes selectoras no produjeron un mínimo claramente definido (***S***, ***CIA***); por contraste, las otras presentaron curvas con valores mínimos bien definidos que variaron de 0.2 (***VCG*** y ***EFP***) a 0.22 (***T*** de Rice). Por tanto, se escogió aquella con el mínimo mas claro (***T*** de Rice).

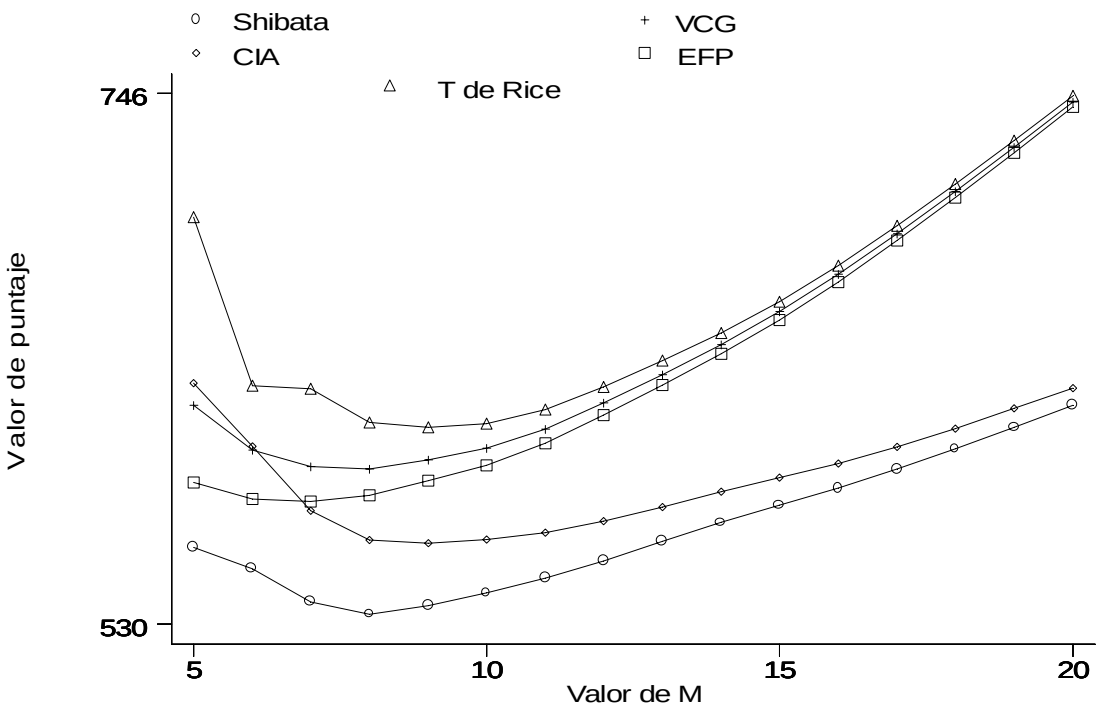

Figura 6.9 Puntaje de funciones penalizantes para los datos de simulación de impactos de motocicletas

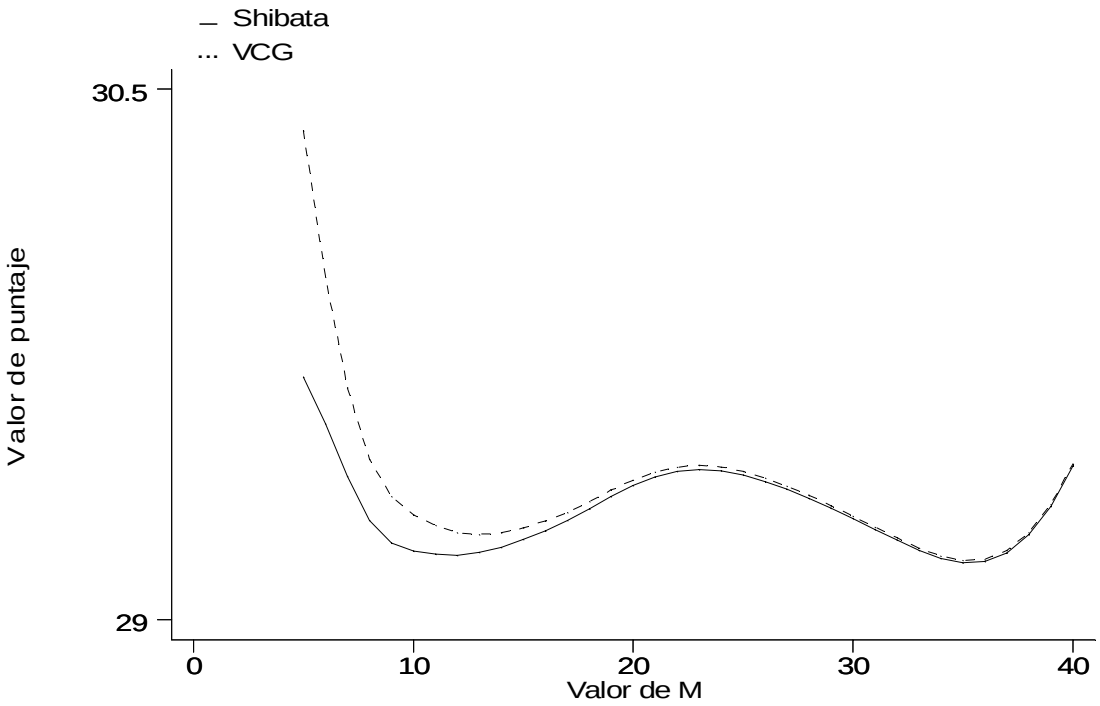

Figura 6.10 Puntaje de funciones penalizantes (línea continua = Shibata y línea punteada = VCG) para los datos del géiser "Old faithful" con $\delta = 0.05$.

En los tres ejemplos, la amplitud de banda automática parece ser una buena elección que permite obtener una curva suavizada con las características principales de la función original. Este valor puede utilizarse como referencia para considerar alternativas más ruidosas (con valores de amplitud de banda menores) o más suaves (amplitudes de banda mayores).

6.6 Otros métodos: estimación de los k-vecinos más cercanos (*k-NN*)

En lugar de considerar un intervalo fijo en la vecindad de *x* (es decir, una amplitud fija de banda) para calcular un promedio local de observaciones Y_i como sucede en la regresión por kernel, es posible considerar un tamaño de muestra fijo con vecindades variables en *X*. Este enfoque es utilizado en la estimación de regresión de los k vecinos más cercanos ó *k-NN* por sus siglas en inglés (***k Nearest Neighbor***), la cual puede definirse como:

$$\hat{m}_k(x) = n^{-1} \sum_{i=1}^{n} W_{k\,i}(x) Y_i \qquad (9)$$

De acuerdo a Härdle (1991) en esta expresión, la secuencia de pesos se define a través de la totalidad del conjunto de índices:

$$\{W_{ki}(x)\}_{i=1}^{n} = J_x = \{i \mid X_i \text{ es una de las k observaciones más cercanas a } x\} \qquad (10)$$

y la secuencia de los *k-NN* pesos definida para este conjunto de observaciones vecinas es:

$$W_{ki}(x) = \begin{cases} n/k & \text{if } i \in J_x \\ 0 & \text{otherwise} \end{cases} \qquad (11)$$

En este procedimiento el parámetro de suavización es *k*, el número de puntos vecinos y como en la regresión por kernel, existe un compromiso entre el sesgo y la varianza. Por lo tanto, incrementando *k* se obtienen estimaciones más suaves aunque con un sesgo mayor. Existen dos casos extremos:

1. cuando $k = n$, la estimación es igual al promedio de la variable de respuesta
2. cuando $k = 1$, la estimación es una función interpolante que se ajusta a cada uno de los valores de respuesta.

Resulta conveniente definir estimaciones *k-NN* simétricas, esto es, calculando el promedio de observaciones *Y* en número *k/2* a la izquierda y *k/2* a la derecha. Si *X* está ordenada, entonces la estimación *k-NN* puede calcularse recursivamente utilizando la siguiente fórmula (Härdle, 1990; 1991):

$$\hat{m}_k(X_{i+1}) = m_k(X_i) + k^{-1}(Y_{i+[k/2]+1} - Y_{i-[k/2]}) \qquad (12)$$

para $i = [k/2] + 1, \ldots, n - [k/2]$; donde $[u]$ indica el mayor entero menor o igual a *u*. La ventaja de este algoritmo es que es lineal en el número de observaciones (para detalles y mayor información al respecto consultar a Härdle, 1990).

Además, este rápido procedimiento de cálculo es importante porque permite actualizar ajustes polinomiales locales. Por esto, es posible calcular un suavizador bivariado con ponderación local (LOWESS por sus siglas en inglés: LOcally WEighted Scatterplot Smoother) con un costo de cómputo muy modesto (Härdle, 1990; Cook & Weisberg, 1994).

El programa para el paquete estadístico Stata que calcula al estimador de regresión *k-NN* se presenta en el Capítulo 7 de la presente obra.

A continuación se muestran los resultados obtenidos con dos diferentes valores para *k* aplicados a cada uno de los tres conjuntos de datos considerados anteriormente: los datos del géiser "Old faithful", los datos de la simulación de impactos de motocicletas y los datos de simulación de crecimiento (Figuras 6.11 a 6.16). Estas estimaciones parecen poco suaves comparadas con las producidas por los métodos para estimar la regresión por kernel, pero proporcionan en términos generales el mismo mensaje. Para los datos del géiser, indican la existencia de dos grupos con una pendiente ligeramente positiva.

En el caso de los datos de simulación de impactos de motocicletas, las estimaciones *k-NN* son en algunos aspectos mejores (en el sentido de ajuste a los puntos) que las estimaciones por regresión por kernel, especialmente al inicio de la serie y en las curvas indicadas por los puntos.

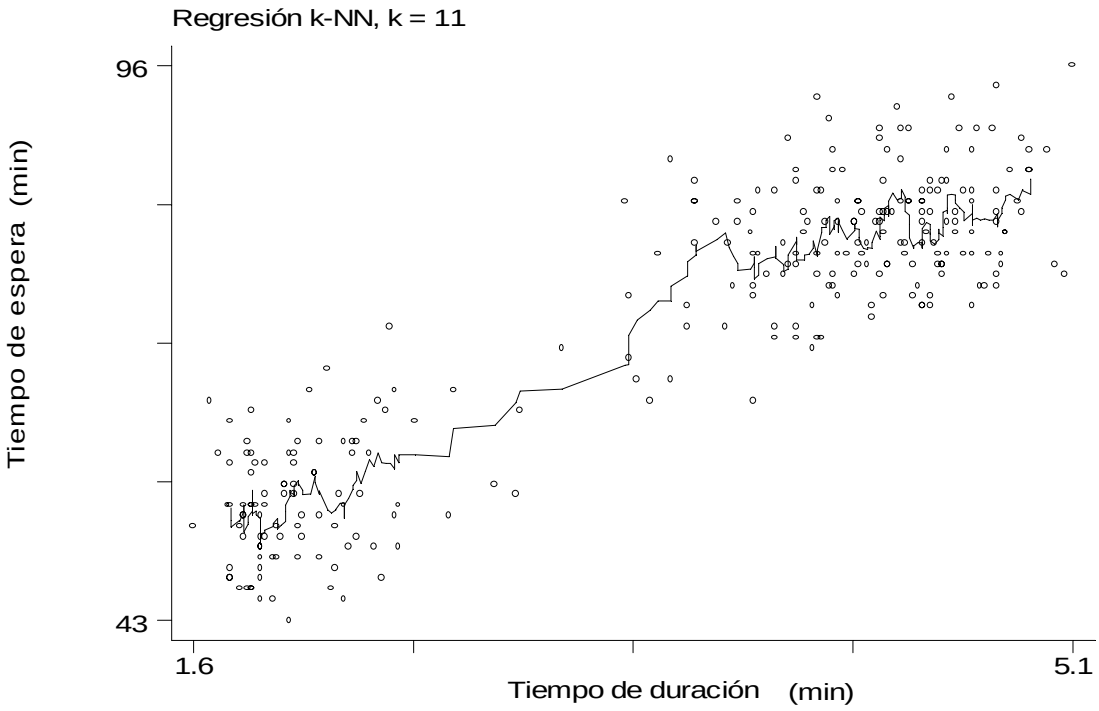

Figura 6.11 Diagrama bivariado de dispersión para los datos del géiser "Old faithful" y estimación de regresión *k-NN* con *k* = 11.

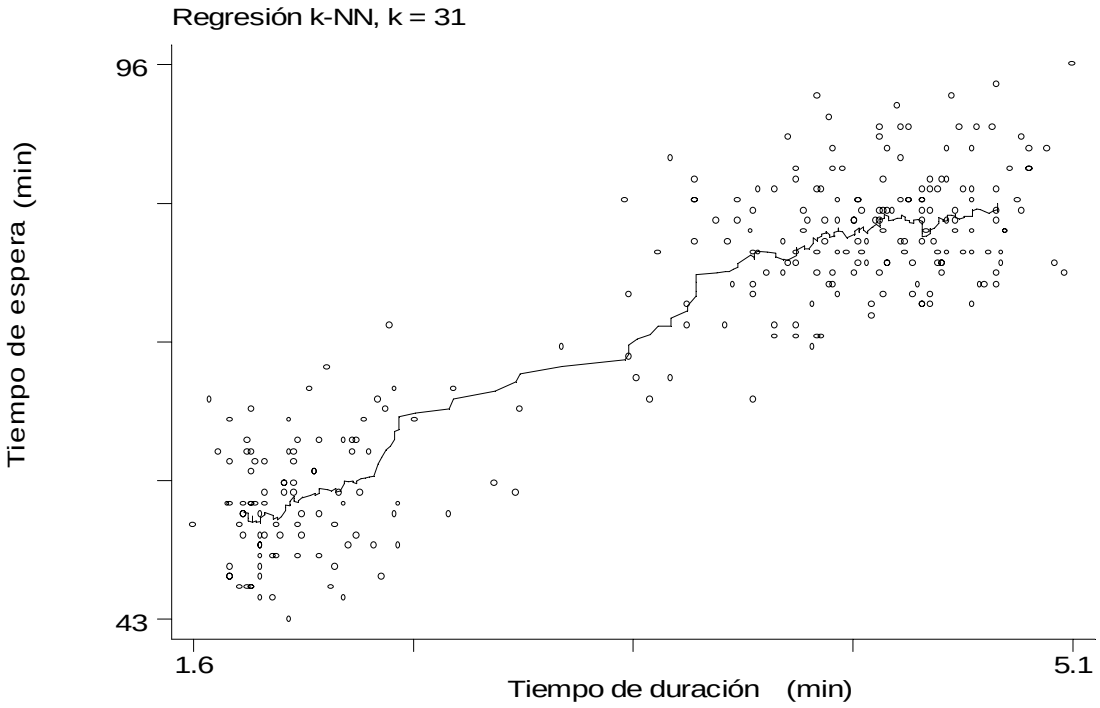

Figura 6.12 Diagrama bivariado de dispersión para los datos del géiser "Old faithful" y estimación de regresión *k-NN* con *k* = 31.

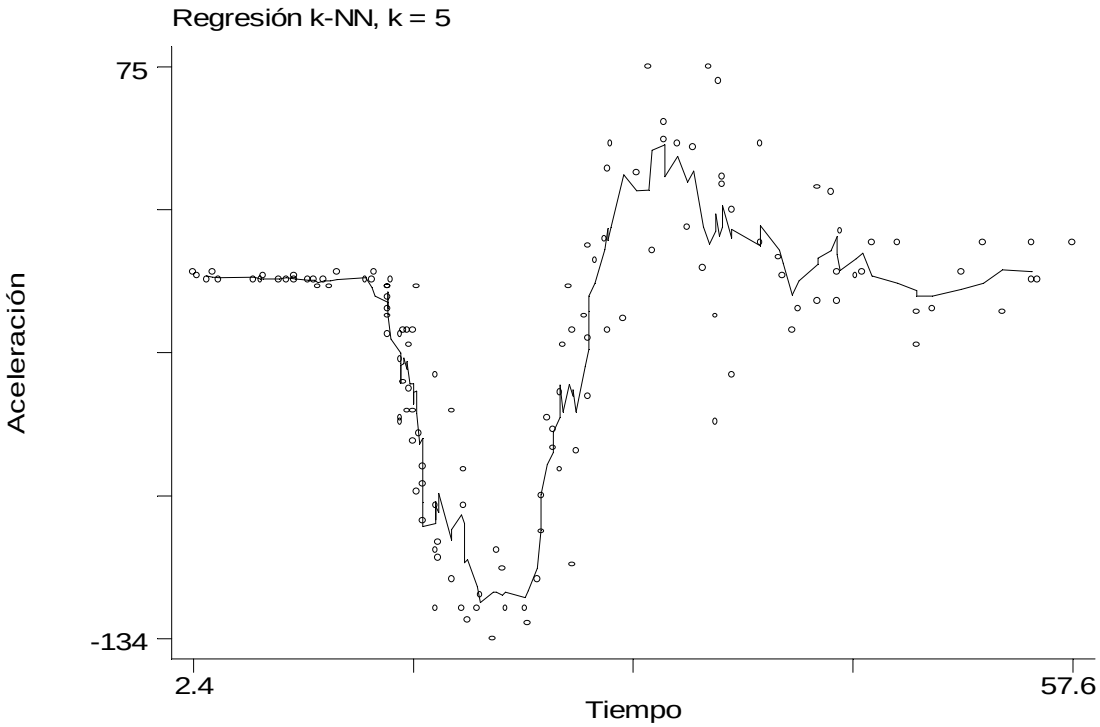

Figura 6.13 Diagrama bivariado de dispersión para los datos de impactos de motocicletas y estimación de regresión *k-NN* con *k* = 5.

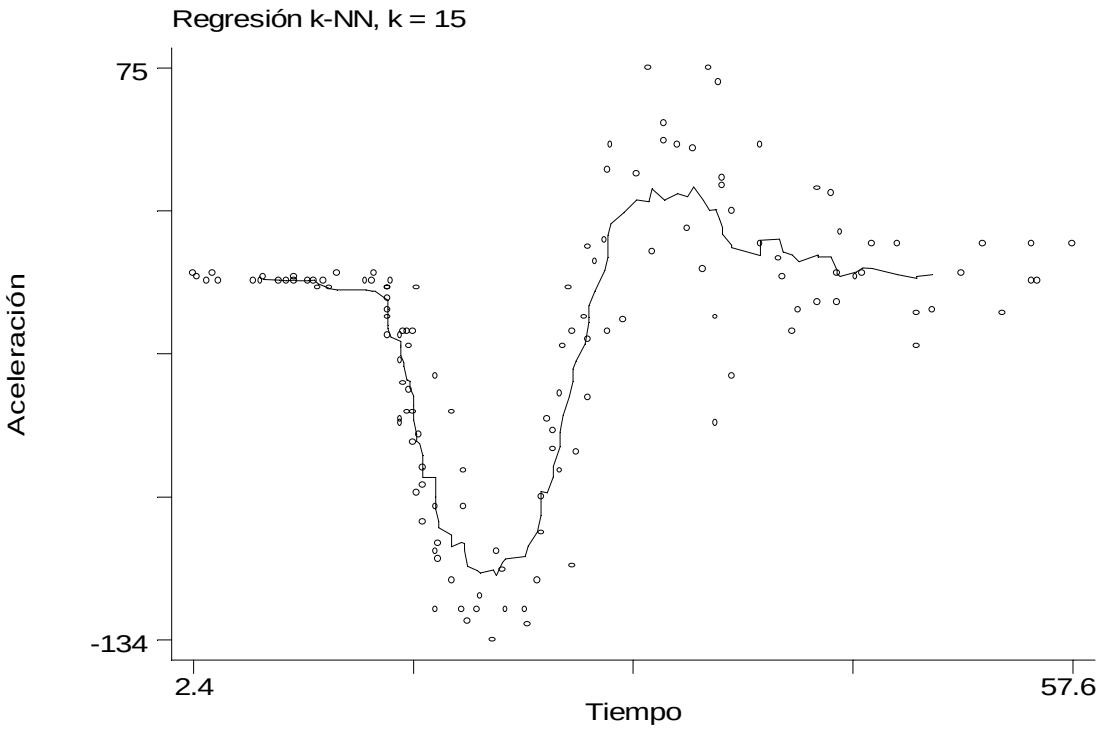

Figura 6.13 Diagrama bivariado de dispersión para los datos de impactos de motocicletas y estimación de regresión *k-NN* con *k* = 15.

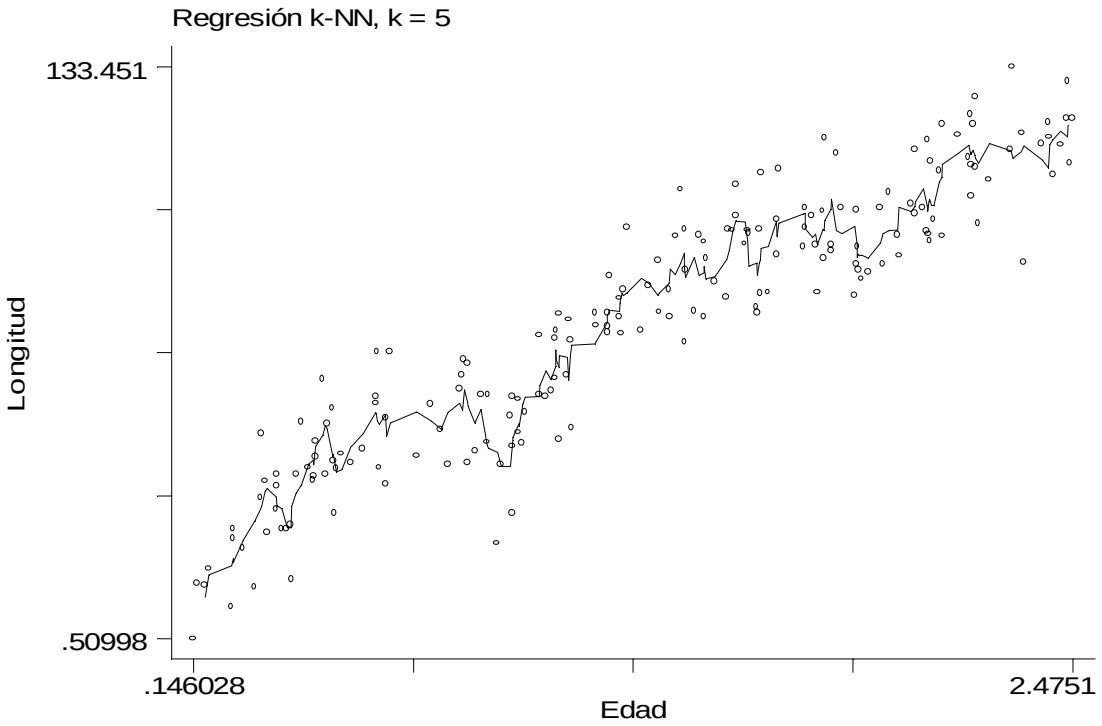

Figura 6.14 Diagrama bivariado de dispersión para los datos de simulación de crecimiento y estimación de regresión *k-NN* con *k* = 5.

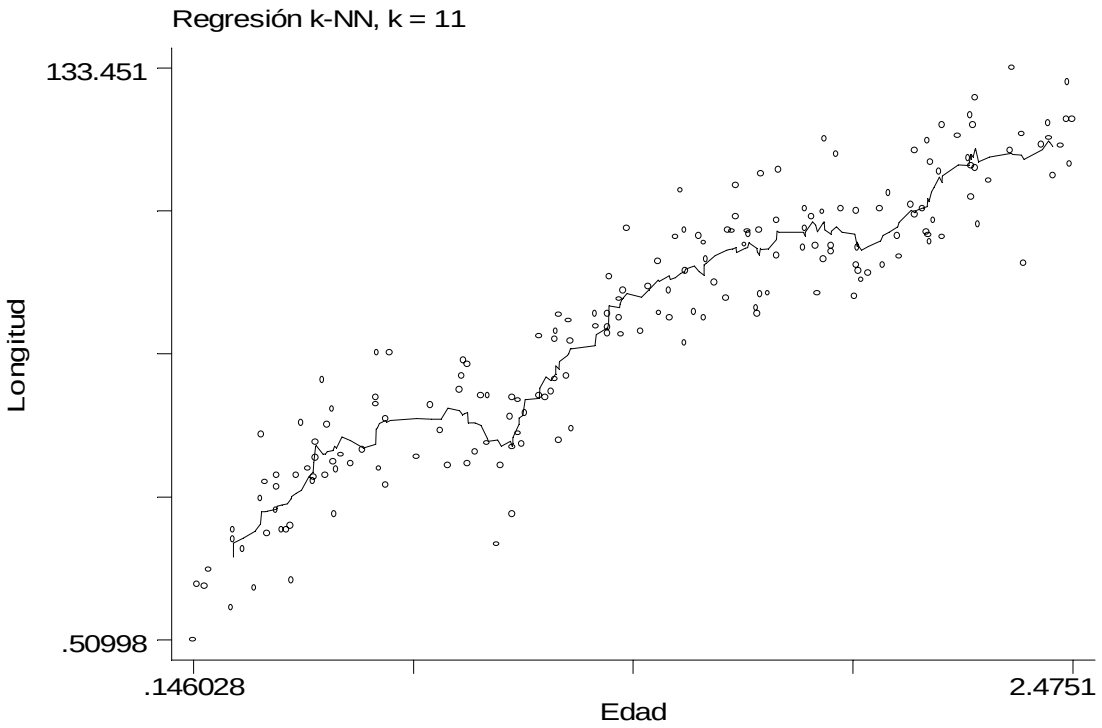

Figura 6.15 Diagrama bivariado de dispersión para los datos de simulación de crecimiento y estimación de regresión *k-NN* con *k* = 11.

Respecto a los datos de simulación de crecimiento se aprecia el carácter estacional de la función original con una correspondencia cercana incluso en el número de ciclos (tres). Sin embargo, un inconveniente es que la curva suavizada que sigue a los puntos simulados pueden llegar a describir lo que sería un "encogimiento" del cuerpo que no es realista para los peces o los bivalvos, organismos que poseen un esqueleto rígido. La función original no toma valores más allá de un incremento de cero en el tamaño. No obstante, la recuperación de la información contenida en los puntos simulados es notable.

Los estimadores no paramétricos de la regresión, a causa de su flexibilidad, son una herramienta muy importante para explorar y descubrir lo que a veces puede ser una relación bivariada muy compleja. Una vez que la función media es estimada es posible calcular su derivada así como características de la curva (máximos, mínimos o raíces). Además, la regresión no paramétrica puede utilizarse en combinación con métodos paramétricos de regresión para la construcción y validación de modelos (Altman, 1992). Estos suavizadores pueden emplearse para la estimación de modelos aditivos generalizados (Hastie y Tibshirani, 1990) en problemas multivariados de regresión (Scott, 1992; Cook y Weisberg, 1994).

6.7 Estimadores no paramétricos adicionales

Existen otros estimadores no paramétricos de regresión. Por ejemplo, en el paquete estadístico Stata podemos citar:

- **Estimadores locales de nivel**. Los procedimientos más simples de regresión no paramétrica son versiones locales de estimadores de nivel. En Stata podemos utilizar medianas (estimador de localización) para conectar bandas de puntos de datos (las bandas representan una vecindad o intervalo en X) al graficar; además, es posible utilizar el comando **ksm.ado** sin opciones (aparte de la amplitud de banda) para obtener un suavizador basado en promedios móviles o utilizar este programa con la opción de peso (además de la amplitud de banda) para obtener un promedio móvil con ponderación tricúbica.

- **Suavizadores "spline"**. Un "spline" es un suavizador lineal muy relacionado con la regresión por kernel. Por lo tanto, al utilizar el comando **graph** con la opción "c" para conectar a los puntos y escogiendo un número adecuado de bandas (el parámetro de suavización) es posible obtener en forma gráfica a estos estimadores no paramétricos de la regresión.

- **Suavizadores de regresión local**. Con el comando **ksm.ado** es posible calcular estimadores de regresión local ponderada y no ponderada (LOWESS) (consultar a Salgado-Ugarte y Shimizu, 1995 para profundizar acerca de algunas mejoras a este comando).

- **Suavizadores no lineales resistentes**. El comando **smooth** de Stata permite explorar los valores de respuesta por medio de combinaciones diferentes de suavizadores resistentes basados en la mediana.

Capítulo 7. Programas computarizados para estimación no paramétrica

7.1 Introducción

Este capítulo presenta muy brevemente los programas desarrollados para ejecutara los cálculos de todos los procedimientos no paramétricos descritos en los capítulos previos. Esta presentación incluye el nombre de los programas, su sintaxis de acuerdo al lenguaje de programación del paquete estadístico Stata, y en algunos casos, algunas notas sobre su desempeño. El listado completo de todos los programas se incluye en el apéndice 1.

7.2 Programas para estimación de densidad por kernel

7.2.1 Trazas de densidad

El programa para calcular trazas de densidad utilizando una función ponderal uniforme (Chambers, *et al.*, 1983) es **boxdetra.ado**. Su sintaxis es:

```
Boxdetra nomvar amplitud tradevar
```

Donde `nomvar` es la variable a procesar, `amplitud` es el valor numérico para el intervalo *h* y `tradevar` es el nombre de la variable nueva que contendrá las trazas de densidad estimadas. Los resultados pueden ser graficados y los puntos se pueden conectar mediante interpolación lineal o bien con lineas a saltos. Es posible combinar el resultado con diagramas univariados de dispersión o de caja en los márgenes.

Por ejemplo, para reproducir la Figura 2.1 podemos escribir en Stata las siguientes órdenes:

```
. use trocolen
. boxdetra length 40 den40
. graph den40 length, s(.) c(l) l1(Densidad) b2("Longitud estándar
(mm)") two o ne box ylab(0,.002,.004,.006) xlab
```

El primer comando llama al archivo de datos trocolen.dta, el segundo realiza la traza de densidad con amplitud de intervalo igual a 40 y guarda los resultados en una variable nueva llamada “den40”. El tercer comando es el encargado de dibujar la gráfica que viene a ser un gráfico X-Y en donde la variable den40 es Y y length es X. Después de la coma vienen varias opciones: s(.) indica que el símbolo de graficación para los puntos sea un punto pequeño; c(l) especifica conectar a los puntos mediante líneas rectas. Los títulos de los ejes se especifican con l1(Densidad) y b2(“Longitud estándar (mm)”) en donde las comillas se agregan para poder escribir paréntesis dentro del subtítulo. Las opciones two, one y box se agregan para especificar un gráfico X-Y y para agregar los diagramas univariados de dispersión y de caja respectivamente. Finalmente, con las opciones ylab(0,.002,.004,.006) y xlab se especifican números redondeados convenientes para los ejes vertical y horizontal. Para una explicación más detallada se pueden consultar los manuales y la ayuda del programa Stata (Stata Corporation, 1997, 1999).

Un inconveniente de este programa es que lleva a cabo *n* veces el comando `summary` para obtener las estimaciones de densidad y de esta forma con una muestra grande puede tardar algún tiempo para terminar.

El programa **boxdetr2.ado** toma en cuenta la misma función ponderal pero considera 50 puntos uniformemente espaciados desde el valor mínimo al máximo de los datos. La sintaxis de este comando es:

```
boxdetr2 nomvar amplitud tradevar punmedvar
```

Donde `nomvar`, `amplitud` y `tradevar` se definen como en el caso anterior y `punmedvar` es la variable que contiene los 50 puntos a lo largo del recorrido de las datos y que utilizados para la estimación.

Estos sencillos estimadores de la densidad eliminan la discontinuidad local de los intervalos del histograma pero no por completo. Una razón para su variabilidad es la forma rectangular de la función ponderal uniforme. Para una suavización adicional de la estimación de densidad podemos considerar una función diferente que varíe gradualmente a lo largo del intervalo *h*, por ejemplo una función coseno que se define:

$$W(u)=\begin{cases}1+\cos 2\pi u, & \text{si } |u|<1/2;\\ 0, & \text{en cualquier otro caso.}\end{cases}$$

Esta función coseno es empleada por el programa **cosdetra.ado**, cuya sintaxis:

```
cosdetra nomvar amplitud tradevar punmedvar
```

Como en el caso de **boxdetr2.ado**, este programa utiliza 50 puntos uniformemente espaciados en el recorrido de los datos.

7.2.2 Estimadores de densidad por kernel

El programa **kernsim.ado** calcula un estimador de densidad simple como se describe en el Capítulo 2 considerando (para ahorrar tiempo y número de cálculos, como ha sido sugerido por Chambers *et al.*, 1983 y por Fox, 1990) un número finito de puntos uniformemente espaciados (cincuenta). Puesto que este estimador simple integra a la unidad entre $x_{(1)} - h$ y $x_{(n)} + h$, estos puntos se localizan (de manera convencional) de $x_{(1)} - h + (d/2)$ a $x_{(n)} + h + (d/2)$, donde d es el intervalo definido por *Recorrido/50*, y el *recorrido* va de $x_{(1)} - h$ hasta $x_{(n)} + h$. Esta convención es adoptada en todos los programas subsecuentes para el paquete estadístico Stata (excepto **adgakern.ado**). La conveniencia de esta definición es que estos puntos (excepto el último) pueden considerarse como puntos medios de intervalos de manera análoga a los histogramas. Lo anterior con fines de comparación y uso de procedimientos derivados tales como la determinación y caracterización de componentes Gaussianos en la distribución de frecuencias (densidades).

La sintaxis del programa **kernsim.ado** es:

```
kernsim nomvar semiamp densivar punmedvar
```

donde `nomvar` es la variable a analizar, `semiamp` es la amplitud de intervalo (llamada por Fox "window half width" en inglés), `densivar` es la variable que contendrá las densidades estimadas y `punmedvar` contendrá los 50 puntos explicados arriba. Para reproducir la Figura 2.6 (Capítulo 2) en Stata basta teclear el siguiente comando:

```
. kernsim l 20 den20 med20
. graph den20 med20, c(l) s(o) ylab(0,.002,.004,.006) l1(Densidad)
b2("Longitud estándar (mm)") xlab
```

Esta figura es muy parecida a la obtenida por **boxdetr2.ado** utilizando un intervalo de 40, esto es, una distribución con al menos cuatro modas. Esto era de esperarse debido a que la elección de la amplitud de intervalo se hizo para obtener versiones equivalentes en suavidad. Como se ha señalado anteriormente, el valor de *h* controla el grado de suavidad con resultados más suaves para valores altos de intervalo.

En esta sección se comentan algunas implementaciones de funciones ponderales (kerneles) de Epanechnikov y Gausiana. Estos programas pueden modificarse de manera sencilla para calcular otros estimadores de densidad por kernel.

El programa **kernepa.ado** estima la densidad por medio de una función kernel de Epanechnikov. Su sintaxis es:

```
kernepa nomvar semiamp densivar punmedvar
```

donde los argumentos se definen como para **kernsim.ado**. La Figura 2.7 puede reproducirse por medio de los siguientes comandos:

```
. kernepa length 20 dene20 mede20
. graph dene20 mede20, c(s) s(.) ylab(0,.002,.004,.006) l1(Densidad) b2("Longit
> ud estándar (mm)") xlab
```

El programa **kerngaus.ado** usa la función kernel Gaussiana para la estimación de densidad, con una sintaxis semejante:

```
kerngaus nomvar semiamp densivar punmedvar
```

Para reproducir la Figura 2.9, en Stata podemos teclear:

```
. kerngaus length 15 deng15 medg15
. graph deng15 medg15, c(s) s(.) ylab(0,.002,.004,.006) l1(Densidad) b2("Longit
> ud estándar (mm)") xlab
```

El programa **adgakern.ado** calcula una estimación de densidad por kernel Gaussiano con amplitud variable utilizando el algoritmo descrito en la Sección 2.4 del Capítulo 2 (sin iteración). Este comando tiene la sintaxis siguiente:

```
adgakern nomvar semiamp densivar
```

Con este programa podemos reproducir la Figura 2.11 (estimación de densidad por kernel de amplitud variable utilizando la amplitud óptima de banda = 20.18) con las órdenes siguientes:

```
. adgakern length 20.18 denavop
. graph denavop length, two one box c(s) s(.) ylab xlab l1(Densidad)
  b2("Longitud estándar (mm)")
```

Para ahorrar cálculos y tiempo, el programa **adgaker2.ado** sólo considera 50 puntos uniformemente espaciados después de los pasos 1) y 2) del algoritmo en la Sección 2.4. Su sintaxis es:

```
adgaker2 nomvar semiamp densivar punmedvar
```

```
. adgaker2 length 15 de2av15 medav15
. graph de2av15 medav15, c(s) s(.) ylab xlab l1(Densidad)
b2("Longitud estándar (mm)")
```

7.2.3 Comentarios respecto al desempeño de los programas

A pesar de su simplicidad, estos estimadores de densidad por kernel no se propusieron sino hasta 1956 en el artículo de Rosenblatt. Una razón para este retraso es quizás la tediosa tarea de su cálculo y por tanto, sólo cuando se hizo disponible algún poder de cómputo fue posible de llevar a cabo. Debido a esto, los estimadores no paramétricos por kernel son ejemplos idóneos para implementación computarizada.

Sin embargo, estos procedimientos no están incluidos en ninguno de los paquetes computarizados estadísticos de amplio uso. El lenguaje de programación del paquete estadístico Stata proporciona un ambiente adecuado para programar tales procedimientos y para graficar los resultados, aunque, como ha sido señalado por Fox (1990), su ejecución tiende a ser lenta. El Cuadro 7.1 contiene el tiempo empleado por cada uno de los programas anteriormente descritos utilizando una computadora PC con procesador 486SX, 25 MHz y 6 Mb de RAM trabajando con los datos de la trucha coralina (n = 316):

Cuadro 7.1. Tiempo aproximado de ejecución of los diferentes programas implementados

Nombre del programa	Tiempo utilizado (minutos)
Boxdetra.ado	11
Boxdetr2.ado	3
Cosdetra.ado	3
Kernsim.ado	3
Kernepa.ado	4
Kerngaus.ado	3.5
Adgakern.ado	50
Adgaker2.ado	25

En estos programas se consideraron algunas de las sugerencias para reducir los cálculos y acelerar la ejecución. En todos los programas que utilizan un kernel Gaussiano (**kerngaus.ado**, **adgakern.ado** y **adgaker2.ado**) los cálculos se realizan considerando solo $|z| < 2.5$. Además, la mayoría de los programas presentados utilizan 50 puntos uniformemente espaciados a lo largo del recorrido de los datos. Si $n < 50$, entonces es necesario utilizar el comando de Stata para establecer el número de observaciones a 50 (“set observations 50”) antes de la estimación de densidad con cualquiera de los programas que emplean este número de puntos.

Utilizando estos programas como base, pueden emplearse otras funciones kernel diferentes. El número de puntos puede modificarse también si se requiere. Silverman (1986) y Fox (1990) proporcionan procedimientos adicionales y alternativos así como detalles de los mismos.

Por otra parte, estos programas pueden aplicarse a los datos presentados por Chamber, *et al.* (1983) y aquellos utilizados por Fox (1990) quien los adaptó (corrigió) de Leinhardt y Wasserman (1979), obteniendo esencialmente los mismos resultados.

Como una verificación adicional, los programas correspondientes en Turbo Pascal fueron escritos y aplicados a estos lotes de datos, obteniendo los mismos resultados (excepto por errores despreciables atribuidos al redondeo) que los obtenidos con los programas para Stata.

7.3 Programas para estimadores por HDP, PPPR y de densidad por kernel (revisión)

7.3.1 Histogramas desplazados promedio (HDP) y promedio ponderado de puntos redondeados (PPPR)

El cálculo de los histogramas desplazados promedio (***HDP***) y de su generalización conocida como el promedio ponderado de puntos redondeados (***PPPR***) se ha implementado como una combinación de archivos ejecutables de Turbo Pascal con programas del paquete estadístico Stata (archivos ado). Los archivos del programa fuente y su ejecutable son **warpings.pas** y **warpings.com** respectivamente. La versión en Turbo Pascal consiste de 8 procedimientos (subrutinas) además del programa principal. El procedimiento **DataInput** pide se especifique un archivo ASCII con los datos tal como los que se crean con el comando "**outfile**" de Stata. Este archivo debe contener sólo una columna de valores (un vector de datos) sin valores faltantes. Este procedimiento pide además que el usuario especifique la amplitud de ventana (*h*), el número de histogramas a desplazar y promediar (*M*) y posteriormente, la subrutina **SelectKerne**l permite especificar el tipo de función ponderal (kernel) de acuerdo con el siguiente código:

1 = Uniforme
2 = Triangular (original de los HDP)
3. = Epanechnikov
4 = Cuártico (biponderado)
5 = Triponderado
6 = Gaussiano

Una vez que esta información se ha proporcionado, **SortData** arregla en orden ascendente a las observaciones, **InitialCalc** determina el origen de la trama de acuerdo a *h* y *M*, **BinmeshCalc** cuenta las observaciones que caen en cada intervalo; **CreateWeight** calcula la función ponderal de acuerdo con la opción seleccionada por el usuario; **WeightBins** calcula el producto de la frecuencia con la función ponderal para cada intervalo no vacío. Finalmente, **ResulFile** crea un archivo ASCII con dos vectores: las estimaciones de densidad y sus correspondientes puntos medios. Este archivo se nombra automáticamente se nombra "**resfile**".

Dentro de Stata se ha implementado la estimación por **PPPR** utilizando el ejecutable warpings.com. La invocación del programa ejecutable así como otras operaciones gráficas y numéricas se incluyen en los siguientes tres archivos ado:

warpstep.ado
warpoly.ado
warping.ado

Los primeros dos programas permiten explorar gráficamente a los datos utilizando las combinaciones deseadas de amplitud de ventana (*h*) y número de histogramas a promediar (*M*): **warpstep.ado** permite calcular estimación de densidad por PPPR y presentarla gráficamente en forma de histograma (versión a pasos); por otra parte **warpoly.ado** calcula y dibuja los resultados utilizando interpolación lineal para conectar los puntos estimados de manera semejante a como el polígono de frecuencia (versión poligonal). El resultado único de estos programas es un desplegado gráfico de la estimación de densidad. Por ejemplo, para obtener el HDP de los cinco histogramas descritos en la Sección 3.2 (Capítulo 3) correspondiente a los datos de precipitación de nieve, se teclea el nombre de alguno de los tres programas arriba listados y proporcionando la información solicitada en la forma que sigue:

```
. warpstep

TYPE THE PATH, NAME AND EXTENSION OF TEXT DATA FILE
bufsnow.raw

THE NUMBER OF VALUES READ IS :        63

!!!!!WARNING!!!!!
IF THIS IS NOT CORRECT, PLEAS INTERRUPT AND
USE STATA COMMAND outfile TO GENERATE AN ASCII
FILE WITH THE DESIRED DATA VECTOR

GIVE THE VALUE OF THE BANDWIDTH ‘h’
10

GIVE THE NUMBER OF HISTOGRAMS TO SHIFT AND AVERAGE
5

SPECIFY THE WEIGHT FUNCTION

1 = Uniform;  2 = Triangle (ASH); 3 = Epanechnikov
4 = Quartic;  5 = Triweight;      6 =  Gaussian
2
(64 observations read)
```

El resultado se presenta en la última gráfica de la Figura 3.1; esta es el ***HDP*** de los datos de precipitación de nieve utilizando una amplitud de banda de 10, función ponderal triangular y promediando 5 histogramas desplazados (combinación que pruduce 64 pares de densidad-puntos medios). Esta estimación no depende más del origen y sugiere una estructura trimodal para la densidad desconocida.

Para obtener los resultados numéricos utilizamos **warping.ado** el cual tiene la sintaxis siguiente:

```
. warping denvar midvar
```

donde denvar y midvar son las nuevas variables que incorporan a las estimaciones de densidad y los correspondientes puntos medios utilizados para su cálculo. Como en **warpstep** y **warpoly** es necesario proporcionar el nombre del archivo **ASCII** con los datos, la amplitud de ventana (*h*), el número de intervalos desplazados (*M*) y el tipo de función ponderal. Como ejemplo, para reproducir el ***HDP*** de la Figura 3.1 se tecletaría:

`. warping densidad punmed` (se introducen después la información requerida: bufsnow.raw, 10, 5 y 2)

```
. gen inter=punmed[2] – punmed[1]
. gen cbajo= punmed-inter/2
. graph densidad cbajo, s(.) c(J) border
```

(en la Figura 3.1 se realizó una edición adicional con el Editor de Gráficas Stage)

Todos los programas ado utilizan el mismo programa ejecutable Turbo Pascal (**warpings.com**).

7.3.2 Estimadores de densidad por kernel (revisión)

En esta sección se han incluido archivos ado (macros) para calcular a los estimadores de densidad por kernel presentados anteriormente con algunos ajustes y algunos otros nuevos. Estos programas son:

kernsim.ado
kerntria.ado
kernepa.ado
kernquar.ado
kerntriw.ado
kerncos.ado
kerngaus.ado

Estos programas calculan respectivamente la densidad por medio de kernel uniforme, triangular, Epanechnikov, cuártico (biponderado), triponderado, Gaussiano y coseno respectivamente (dentro de Stata es posible obtener ventanas de ayuda al teclear “help kernel”). Cada programa considera una malla de 50 puntos uniformemente distribuidos de $x_{(1)} - h -$ (recorrido*0.1) a $x_{(n)} + h +$ (recorrido*0.1). De acuerdo con Härdle (1991), siguiendo la sugestión de Gasser *et al.* (1985), es posible sólo comparar estimaciones de densidad correspondientes a kerneles con el mismo soporte. Estos autores sugirieron utilizar el intervalo [-1,1] como estándar. Tomando en cuenta estas consideraciones, las ecuaciones para los archivos ado arriba listados se ajustaron de acuerdo al siguiente cuadro:

Cuadro 7.2 Ecuaciones generales de los kerneles implementados	
KernelUniforme	K(z)
Uniforme	½ I(|z| ≤ 1)
Triangular (HDP)	(1 - |z|) I(|z| ≤ 1)
Epanechnikov	¾ $(1 - z^2)$ I(|z| ≤ 1)
Cuártico (biponderado)	$(15/16)(1 - z^2)^2$ I(|z| ≤ 1)
Triponderado	$(35/32))(1 - z^2)^3$ I(|z| ≤ 1)
Coseno	$(\pi/4)\cos((\pi/2)z)$ I(|z| ≤ 1)
Gaussiano	$(1/\sqrt{2\pi}) \exp((-1/2)z^2)$

La sintaxis general de los archivos ado es de la forma en que se describe en la sección 7.2.2:

```
kernel nomvar semiamp denvar punmedvar
```

en donde "kernel" se refiere a alguno de los ado files arriba citados. Por ejemplo, para calcular un kernel cuártico (biponderado) con amplitud de ventana de 10 podemos teclear:

```
. use bufsnow
. kernquar snow 10 den10 pmed10
(salida omitida)

. graph den10 pmed10, xlab ylab c(s) s(.) border
```

7.3.3 Algunos comentarios adicionales

Es suficiente utilizar una $M \geq 5$ para obtener una estimación adecuada de cualquiera de los estimadores de densidad implementados por medio del procedimiento HDP-PPPR. Como convención particular, en esta obra se recomienda utilizar $M = 15$ en todas las estimaciones. Dependiendo del número de observaciones y del intervalo que tienen, esto puede resultar en varios cientos de puntos. Debido a la definición del arreglo (array) en el lenguaje Turbo Pascal, el número máximo válido de sub-intervalos es 2000 (con los datos de precipitación de nieve de la ciudad de Búffalo, este número se alcanza al utilizar un valor para *M* mayor de 150). Este máximo es en verdad bastante grande y es absolutamente innecesario probar valores de *M* mayores de 15, excepto en el caso de la investigación de multimodalidad donde se pueden utilizar valores hasta de 40.

Esta colección de programas ejecutables y ado tiene varias limitaciones: debido a que **warpstep** y **warpoly** utilizan el comando **"infile"**, es necesario no tener datos en la memoria o el programa se detiene con un error si existe alguno. Se prefirió este enfoque en lugar de la inclusión de borrar la memoria con el comando **"drop _all"** en los archivos ado. El archivo de datos debe estar en formato **ASCII** con una sola columna para el vector de datos. No es posible utilizar los selectores **"if"** ó **"in"** con estos programas (incluidos los **EDK's**). No obstante, trabajando con los datos de interés, el usuario, en un paso previo, puede seleccionar cualquier subconjunto de observaciones, descartando los valores faltantes y grabándolos con el comando **"outfile"**. Después de usar warping es necesario salvar los resultados en un archivo y después eliminarlos de la memoria antes de estimar nuevamente la densidad por cualquiera de estos programas. La inclusión de la función de peso Gaussiana en **warpings.com** requirió el realizar un ajuste adicional en la amplitud de ventana.

7.4 Programas para selección de amplitud de intervalo/banda en la estimación univariada de densidad

7.4.1 Reglas prácticas para estimadores univariados de densidad

La programación de las reglas presentadas en el Capítulo 4 como macros del paquete Stata (archivo ado) es directa. En los programas que se presentan a continuación se utilizan varios comandos de Stata tales como **summarize, detail** en combinación con el uso de las funciones **_result() y min()**. El programa para calcular estas reglas se ha llamado **bandw.ado** y tiene la siguiente sintaxis:

```
bandw nomvar [if exp] [in range]
```

El resultado de este comando es un cuadro con (se incluye número de ecuación del Capítulo 4):

- Reglas para número de intervalos de histogramas (1) y (5)
- Regla para el número sobresuavizado de intervalos para polígonos de frecuencia (9)
- Reglas para amplitud de intervalo de histogramas (2), (3), (4), (6) y (7)
- Reglas para amplitud de intervalo de polígonos de frecuencia (8) y (10)
- Reglas para amplitud de banda de estimadores de densidad por kernel (15), (14) y (16)

En este apartado se incluye sólo su uso para obtener los resultados que ya se han discutido en el Capítulo 4 en relación con los datos de precipitación de nieve y longitud de bagres.

```
. use bufsnow
. bandw snow

_____________________________________________________________
Some practical number of bins and binwidth-bandwidth rules
for univariate density estimation using histograms,
frequency polygons (FP) and kernel estimators
=============================================================

Sturges' number of bins =                          6.9773
Oversmoothed number of bins <=                     5.0133
-------------------------------------------------------------
FP oversmoothed number of bins <=                  5.4091
=============================================================

Scott's Gaussian binwidth =                       20.8641
Freedman-Diaconis robust binwidth =               17.4413
Terrell-Scott's oversmoothed binwidth >=          20.2262
Oversmoothed Homoscedastic binwidth >=            22.2292
Oversmoothed robust binwidth >=                   22.6999
-------------------------------------------------------------
FP Gaussian binwidth =                            22.2680
FP oversmoothed binwidth >=                       24.1323
=============================================================

Silverman's Gaussian kernel bandwidth =            9.3215
Haerdle's 'better' Gaussian kernel bandwidth =    10.9787
Scott's Gaussian kernel oversmoothed bandwidth =  11.8487
_____________________________________________________________
```

```
. use catfish
. bandw bodlen

_____________________________________________________________
Some practical number of bins and binwidth-bandwidth rules
for univariate density estimation using histograms,
frequency polygons (FP) and kernel estimators
=============================================================

Sturges' number of bins =                          12.2521
Oversmoothed number of bins <=                     16.9595
-------------------------------------------------------------
FP oversmoothed number of bins <=                  11.2383
=============================================================

Scott's Gaussian binwidth =                        15.0376
Freedman-Diaconis robust binwidth =                16.0466
Terrell-Scott's oversmoothed binwidth >=           13.2079
Oversmoothed Homoscedastic binwidth >=             16.0215
Oversmoothed robust binwidth >=                    20.8847
-------------------------------------------------------------
FP Gaussian binwidth =                             26.1322
FP oversmoothed binwidth >=                        28.3200
=============================================================

Silverman's Gaussian kernel bandwidth =            10.9391
Haerdle's 'better' Gaussian kernel bandwidth =     12.8838
Scott's Gaussian kernel oversmoothed bandwidth =   13.9048
_____________________________________________________________
```

7.4.2 Validación cruzada por mínimos cuadrados

Siguiendo a Scott y Terrell (1987) así como a Härdle (1991) se modificaron los programas de este último autor para escribir programas con el procedimiento PPPR para estimación de validación cruzada por mínimos cuadrados (VCMC) y validación cruzada sesgada (VCS) aplicadas a la estimación de densidad por kernel. Como los autores citados discuten, el enfoque de la PPPR es un método eficiente para cálculo que nos permite localizar la amplitud de banda "óptima" y llevar a cabo simulaciones que consideran un gran número de observaciones y repeticiones si se desea. Para lograr lo anterior se fija el valor de la amplitud pequeña δ (producto del desplazamiento del origen de la trama para cálculo). Puesto que $h = M \times \delta$, para encontrar el parámetro de suavización "óptimo" tan sólo basta localizar el valor "óptimo" para M. De manera semejante a las estimaciones de densidad presentadas anteriormente, estos programas utilizan archivos ejecutables de rutinas en lenguaje Turbo Pascal que realizan los cálculos (**l2cvwarf.exe** y **bcvwarpf.exe**) y que escriben los resultados en archivos de texto (formato ASCII). Estos resultados son llamados por archivos ado los cuales se encargan de realizar el procesamiento subsecuente.

La sintaxis del programa para realizar VCMC es la siguiente:

```
l2cvwarp varname [if exp] [in exp], delta(#) kercode(#)
     [mstart(#) mend(#) gen(cvval mval hval)
     nograph graph_options]
```

donde `delta` especifica la amplitud de banda pequeña resultante del desplazamiento de histogramas por promediar. Este valor es interpretado como una medida de la precisión de las observaciones, las cuales son redondeadas hasta el nivel que indica; `kercode` permite escoger la función ponderal de acuerdo a los siguientes códigos:

1. Uniforme
2. Triangular
3. Epanechnikov
4. Cuártica (Biponderada)
5. Triponderada
6. Gaussiana

Estos valores se requieren para correr el programa (si no se especifican entonces se detiene y despliega un mensaje de error). Las opciones `mstart` y `mend` permiten determinar el intervalo de valores para *M* considerados para la búsqueda del mínimo. El valor pre-establecido para `mstart` es 1; si `mend` no se especifica, entonces se fija a un valor igual aproximadamente a un tercio del recorrido de las observaciones (lo que produce una estimación inicial adecuada en la mayoría de los casos). Con la opción `gen` es posible definir tres variables que contendrán respectivamente, el valor de puntaje de la VC, valores de *M* y *h*. Finalmente, es posible suprimir el desplegado gráfico o manipular sus propiedades.

El programa produce como resultados una gráfica del puntaje de VC contra *M* la cual permite localizar el intervalo con el mínimo si lo sugiere (recordar que $h = M \times \delta$,). Además de la gráfica, se despliega un cuadro con la lista de los cinco valores de puntaje más bajos y sus correspondientes valores de *M* y *h*. De esta forma, es posible obtener la amplitud de banda "óptima" (aquella para el valor de puntaje VC mínimo). Si la gráfica no presenta de manera clara un mínimo, pero lo sugiere al mostrar una línea con tendencia al decremento monótono, entonces se pueden escoger otros valores para `mstart` y `mend` de forma que un mínimo quede abarcado.

Con este comando, es posible estimar el parámetro de suavización óptimo para los estimadores de densidad por kernel. Especificando el kernel triangular se puede obtener la amplitud de banda óptima para el HDP e el correspondiente estimador de densidad por kernel triangular (incrementando *M* en la estimación por PPPR). Las otras funciones ponderales se aplican a sus estimadores de densidad respectivos.

Por ejemplo, utilizando los datos de precipitación de nieve:

```
. l2cvwarp snow, delta(1) kercode(6) mstart(3) mend(30) xlog xlab xline(11.85)
```

(La gráfica correspondiente a la Figura 4.5 aparece en la pantalla, así como el Cuadro 4.5)

Para estimar el parámetro de suavización por medio de la Validación Cruzada Sesgada (VCS) se utiliza el programa bcvwarp.ado, el cual tiene la siguiente sintaxis:

```
bcvwarp varname [if exp] [in exp], delta(#) kercode(#)
      [mstart(#) mend(#) gen(bcvval mval hval)
      nograph graph_options
```

donde todas las opciones son interpretadas como en el caso de l2cvwarp.ado excepto que sólo se han implementado dos tipos de funciones kernel las cuales se han codificado como sigue:

1 Cuártico (Biponderado)
2 Triponderado.

De manera similar a la VCMC, el resultado predefinido del programa es la gráfica del valor de validación cruzada sesgada contra el intervalo de *M* utilizado para los cálculos; esta gráfica ayuda a localizar el mínimo de la curva resultante. Puede examinarse también el cuadro con los cinco valores más bajos del puntaje de VCS con sus correspondientes valores para *M* y *h*.

La aplicación a los datos de cantidad de nieve caída se presentan a continuación:

```
. bcvwarp snow, d(1) k(2) ms(15) me(70) xlog xlab xline(35.29)
```

(Los resultados son la Figura 4.6 y el Cuadro 4.6)

7.4.3 Programas mejorados para estimación univariada de densidad

En esta sección se presenta una versión revisada e integrada de los programas (introducidos como una serie de comandos separados en secciones anteriores) para la estimación de densidad por kernel a traves de métodos discretizados/interpolados y por el HDP-PPPR. Las estimaciones de densidad del Capítulo 4 fueron realizadas con estos comandos. Estas versiones tienen una sintaxis mas propia del paquete estadístico Stata e incluyen varias opciones de cálculo y graficación. La única característica faltante es la posibilidad de utilizar pesos en los cálculos.

Como se discutió en capítulos previos, por razones de cálculo y gráficas se requiere estimar las densidades por kernel sólo en un número discreto de puntos. De acuerdo al análisis de Jones (1989) los programas aquí presentados que utilizan una serie de puntos para cálculo así como las versiones para HDP-PPPR pertenecen a los estimadores que este autor clasifica como discretizados (parecidos a histogramas) o interpolados (semejantes a polígono de frecuencia).

Para esta clase de estimación de la densidad se introduce el programa revisado kerneld.ado cuya sintaxis es:

```
kerneld nomvar [if exp] [in range],
     bwidth(#) kercode(#) npoint(#) [gen(densivar gridvar)
     nograph graph_options]
```

donde bwidth(#) especifica la amplitud de banda (h), kercode(#) permite escoger la función ponderal (kernel) de acuerdo al siguiente código:

1. Uniforme
2. Triangular
3. Epanechnikov
4. Cuártico (biponderado)
5. Triponderado
6. Gaussiano
7. Coseno

npoint(#) se refiere al número de puntos uniformemente espaciados (trama) a lo largo del intervalo de nomvar a ser considerados para el cálculo; debe tenerse cuidado de emplear un número suficiente de puntos; menos de 50 pueden resultar en una estimación muy burda. La amplitud de banda, función ponderal y el número de puntos para cálculo no son opcionales; si no se especifican el programa se interrumpirá con un mensaje de error. Con la opción gen(densivar gridvar) es posible crear dos variables conteniendo los valores de densidad estimados y sus correspondientes

puntos de cálculo. Una vez que la información requerida se ha dado, el programa dibuja las densidades conectadas por lineas a lo largo de los puntos para cálculo (estimación por interpolación lineal ó parecida a PF); esta gráfica puede modificarse con las opciones propias de la combinación de comandos graph, twoway de Stata. Es posible utilizar la opción (J) para obtener un estimador semejante a histograma, pero esto afecta en forma adversa a la representación gráfica (Jones, 1989). La salida gráfica puede suprimirse utilizando la opción nograph.

Ejemplos:

```
. kerneld snow, bwidth(8) kercode(6) npoint(100)
. kerneld bodlen, b(8) k(4) np(100) gen(denopt gridopt) nograph
```

La versión nueva para la estimación de densidad por HDP-PPPR se incluye en el programa warpden.ado, el cual tiene la sintaxis siguiente:

```
warpden nomvar [if exp] [in range],
     bwidth(#) mval(#) kercode(#) [step nosort
     gen(densivar gridvar) nograph graph_options]
```

donde bwidth(#) es el parámetro de suavización h (amplitud de intervalo para histogramas, polígonos de frecuencia e histogramas desplazados promediados o su versión interpolante lineal, amplitud de banda para estimadores de densidad por kernel), mval(#) es el número de histogramas desplazados promediados considerados para el cálculo; kercode(#) permite especificar la función ponderal de acuerdo al mismo esquema de codificación utilizado para kerneld.ado (excepto que el kernel coseno no está incluido). De manera semejante al comando kerneld, estos tres parámetros no son opcionales, sino que deben especificarse para correr el programa. Con la opción step es posible escoger la versión semejante a histograma (HDP); la estimación pre-definida es la interpolación lineal (semejante a PF). Por medio de la opción nosort se puede evitar un ordenamiento de los datos si estos se han ordenado previamente. Se sugiere para propósitos de exploración cuando se requieren estimaciones repetidas con diferentes valores del parámetro de suavización, ordenar a la variable de interés en primer lugar y después incluir la opción nosort. Esto puede ahorrar tiempo especialmente cuando se tienen muchas observaciones (miles de datos). El resto de las opciones tienen un significado similar a aquellas para kerneld.ado, con la excepción de que técnicamente, los puntos utilizados para estimar densidades pueden interpretarse como los puntos medios de los intervalos menores resultantes del desplazamiento de histogramas. Este comando así como sus predecesores, utiliza un archivo ejecutable en Turbo Pascal para efectuar los cálculos, por medio del comando shell de Stata, pero ahora la transferencia de los parámetros y los resultados es automática sin ninguna entrada interactiva por parte del usuario.

Ejemplos:

```
. warpden snow, b(14.5) m(10) k(6)
. warpden bodlen, b(8) m(10) k(4) step
```

El último programa incluido aquí es warpdens.ado, el cual tiene la misma sintaxis que warpden, pero la “s” en el nombre indica que es un programa totalmente escrito en lenguaje de Stata. Esto significa que los cálculos se realizan dentro del paquete estadístico sin necesidad de ningún programa ejecutable. Por esto, este comando puede utilizarse en todas las plataformas en donde se ejecute Stata (en lo particular se han probado bajo MSDOS, Windows 3..11, Windows 95, Windows 98 y MacIntosh). El usuario es advertido de que esta versión puede no ser tan rápida como warpden, especialmente con un alto valor para M (se sugiere utilizar un valor de 10).

7.5 Prueba de multimodalidad de Silverman (bootstrap suavizado)

Este es un programa sencillo que se utiliza en combinación con el comando boot de Stata para crear muestras repetitivas (bootstrap). El archivo bootsam.ado lleva a cabo los cálculos requeridos para suavizar las muestras. Su sintaxis es:

```
boot bootsam, arguments(nomvar critbw) iterate(#)
```

donde la primera opción permite incluir argumentos para el comando boot; nomvar es el nombre de la variable que contiene las observaciones originales para muestreo; critbw es el valor de la amplitud de banda crítica para un número especificado de modas y con iterate es posible controlar el número de muestras.

Ejemplo:

```
. set seed 12345
. boot bootsam, arg(blfemin 24.46) iterate(100)
```

En primer lugar se establece la semilla para el generador de números aleatorios. Se requiere de un valor diferente para cada simulación para tener una serie distinta de números aleatorios. Los resultados de este comando son 100 muestras bootstrap extraídas de la variable original *blfemin*, usando una banda crítica de 24.46 de acuerdo a las expresiones de Silverman (1981b). Las muestras contienen a la variable *ysm* la cual contiene los valores suavizados de las muestras, *blfemin* es decir la variable original (repetida), *_rep* que es una variable indicadora del número de la muestra y *_obs*, el número de observaciones por muestra. Excepto *ysm* y *_rep* se recomienda borrar de la memoria a las demás variables (por el comando drop).

Cada muestra es examinada entonces para contar el número de modas de la estimación de densidad por kernel utilizando la amplitud crítica de banda. Esto podría realizarse mediante el uso del comando mejorado warpdenm.ado que tiene la siguiente sintaxis:

```
warpdenm nomvar [if exp] [in range], bwidth(#) kercode(#)
     mval(#) [ step numodes modes npoints gen(denvar midvar)
     nograph graph_options ]
```

Sólo se explican aquí las opciones nuevas:

`numodes` reporta el número de modas (maximos) en la estimación de densidad.
`modes` produce una lista de las modas estimadas (localizadas en los puntos utilizados para el cálculo de la densidad).
`npoints` da el número de puntos utilizados para la estimación.

La opción numodes en combinación con el sistema de búsqueda binaria hace posible encontrar iterativamente las amplitudes críticas de banda. Para lograr esta tarea recomendamos utilizar mval(30) u otro número de histogramas desplazados que produzcan un número conveniente de puntos, más para amplitudes menores y menos para amplitudes mayores. Es posible utilizar la opción npoints para ver el número de puntos utilizados en la estimación. Debe considerarse también

la precisión de la escala original para evitar usar un número excesivo de decimales en las amplitudes de banda.

Es posible examinar cada muestra para contar el número de modas en la estimación de densidad por kernel gaussiano con la banda crítica. De esta forma, para la primera muestra se tecletaría:

```
. warpdenm ysm if _rep==1, b(24.46) m(30) k(6) numo mo
```

para obtener el número de modas (de manera más precisa que con la gráfica) e inclusive, sus estimaciones.

Se ha automatizado el último paso del algoritmo de Silverman en el comando silvtest.ado. Este programa calcula el valor p para un número especificado de modas en la estimación de densidad por kernel gaussiano para cada una de las muestras repetitivas (bootstrap), contando las modas resultantes y calculando del total de repeticiones la fracción de estimadas con más modas que el número probado. Este comando tiene la siguiente sintaxis:

```
silvtest smvar repndx, critbw(#) mval(#) nuri(#)
     nurf(#) cnmodes(#)[nograph graph_options]
```

En este comando *smvar* es la variable suavizada bootstrap y *repndx* es el índice de repetición. Las opciones son:

`critbw` es la amplitud crítica de banda para el número de modas a probar.
`mval` es el número de histogramas desplazados promediados utilizados para calcular las estimaciones de densidad requeridas.
`nurf` permite a uno especificar el número final de las repeticiones requeridas.
`cnmodes` se refiere al número crítico de modas, esto es, el número de modas a probar.
`nuri` permite especificar el número inicial de replicación (en caso de la necesidad de dividir el cálculo en varias partes debido a memoria limitada). El valor pre-establecido es uno.
`nograph` suprime el desplegado de la gráfica.
`graph_options` son cualquiera de las opciones permitidas con `graph, twoway`.

Si el usuario no proporciona las opciones `critbw, mval, nurfin` y `cnmodes`, el programa se detiene y despliega un mensaje de error en la pantalla.

El programa permite el examen de la estimación de densidad por kernel para cada una de las muestras bootstrap (la Figura 7.1 muestra la estimación para la primera muestra), cuenta el número de modas y calcula el *valor-p* para el número de modas de prueba. Como ejemplo, aquí se presenta la estimación del *valor-p* para una moda para los datos del bagre estuarino *Cathorops melanopus* descritos en el Capítulo 5. Primeramente se pulsa el comando y las opciones para obtener una salida con el número de repetición y su correspondiente número de modas. Después de contar las modas de la última muestra, el programa muestra el *valor-p* incluyendo los números utilizados para el cálculo. De esa forma es posible extraer otra muestra repetitiva (bootstrap) y acumular el número de estimaciones de densidad con un número de modas mayor que el probado para calcular el *valor-p* en relación al gran total de repeticiones.

```
. silvtest ysm _rep, cr(25.26) m(30) nurf(50) cnm(1)
bs sample     1             Number of modes = 1
bs sample     2             Number of modes = 1
```

```
bs sample     3              Number of modes = 1
(output omitted)
bs sample     49             Number of modes = 1
bs sample     50             Number of modes = 1

Critical number of modes =      1

Pvalue =           0 / 50 =    0.0000
```

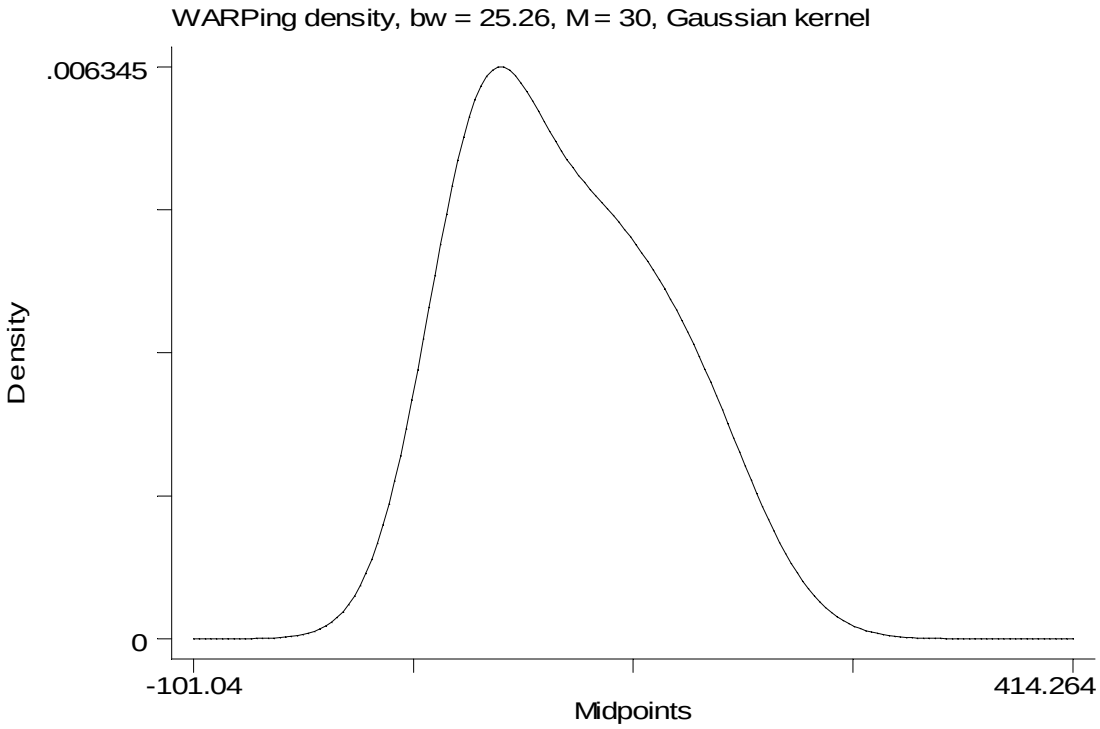

Figura 7.1 Estimación de densidad por kernel gaussiano utilizando la amplitud crítica de banda para una moda (25.26) para la primera muestra bootstrap, datos de longitud patrón de bagres n = 1116.

7.6 Programas para regresión no paramétrica

En esta sección se presentan los archivos ado de Stata algunas veces combinados con programas ejecutables en Turbo Pascal para calcular los estimadores de regresión no paramétrica descritos en el Capítulo 6. Los comandos implementados son:

- regresión por kernel utilizando un algoritmo de interpolación; kernreg.ado
- regresión por kernel utilizando el método HDP-PPPR; warpreg.ado
- estimador de regresión por k-vecinos más cercanos: knnreg.ado

Además, se presenta un programa para la selección de amplitud de banda en regresión por kernel: gwarpreg.ado así como un programa de simulación de crecimiento.

7.6.1 Programa para simulación de crecimiento de acuerdo a la función de crecimiento de von Bertalanffy ajustada para cese estacional de crecimiento.

Los comandos de Stata que ha continuación se listan se utilizaron para simular una serie de datos que siguen variaciones estacionales de crecimiento para probar los programas de regresión no paramétrica:

```
. set obs 200
. set seed 12345
. generate age=2.5*uniform()
. gen error=invnorm(uniform())*sqrt(100)
. sort age
. gen cumngt=age*0.27
. gen tprime=age-cumngt
. gen capq=2*_pi/(1-0.27)
. generate smaq=0.76*(tprime-0.1)+(0.76/capq)*(sin(capq*(tprime-
0.16))- sin(capq *(.1-.16)))
. generate ltfun=156*(1-exp(-smaq))
. generate ltsim=ltfun+error
. drop if ltsim<0 | ltfun<0
```

Estos datos simulados están graficados en la Figura 6.3 (Capítulo 6). Las observaciones de la variable de respuesta se han generado por la función de crecimiento de von Bertalanffy ajustada para cese estacional de crecimiento propuesta por Pauly *et al.* (1992).

$$L_t = L_\infty [1 - \exp(-q)]$$

En la cual

$$q = K(t' - t_0) + K/Q[\sin Q(t' - t_s) - \sin Q(t_0 - t_s)]$$

donde $Q = 2\pi/(1 - NGT)$. La variable *t'* se obtiene de la substracción de la edad real (*t*) el tiempo total de falta de crecimiento ocurrido hasta la edad *t*. Podemos identificar en esta relación bivariada que la variable de respuesta es la longitud a la edad ($Y_i = L_{ti}$) y la variable independiente es la edad ajustada ($X_i = t'_i$).

Los valores de la variable independiente están uniformemente distribuidos en el intervalo [0,2.5] y los errores agregados (ε) son Gaussianos con $\mu = 0$ y varianza $\sigma^2 = 100$. El gráfico de dispersión de los valores simulados (función + error) y la función original se muestran en la citada Figura 6.3 del Capítulo 6.

La versión automatizada de esta serie de comandos de Stata para simulación de crecimiento se incluye en el archivo-ado growsim.ado con una sintaxis simple de:

```
growsim nuobs seed agemax var ngt kval t0 ts linf ltvar ltsimvar
agevar
```

donde nuobs es el número de observaciones a generarse, seed es un número para inicializar el generador de números pseudoaleatorios, agemax es la edad máxima en el intervalo simulado, var es

la varianza del error; posteriormente siguen las constantes del modelo arriba descrito (ngt = NGT, kval = K, t0 = t_0, ts = t_s y linf = L_∞) y finalmente tres variables nuevas que contengan los resultados de la función (ltvar), la función más el error (ltsimvar) y las edades simuladas (agevar). Usando este programa podemos reproducir los datos de simulación de crecimiento del Capítulo 6 como sigue:

```
. growsim 200 12345 2.5 100 0.27 0.76 0.10 0.16 156 ltfun ltsim
age
```

Los valores de los parámetros son aquellos ajustados por Pauly, *et al.* (1992) correspondientes a los datos de crecimiento del salmón del Atlántico (*Salmo salar*) en un arroyo escocés (Egglishaw, 1970).

7.6.2 Estimador de Nadaraya Watson (regresión por kernel).

En capítulos previos (2-5) y articulos (Salgado-Ugarte *et al.* 1993) se han introducido algunos programas para calcular densidades univariadas usando funciones kernel a través de estimadores discretizados o interpolados (Jones, 1989). Es relativamente sencillo modificar estos programas para estimar el numerador y denominador de la ecuación 6.7. Esta implementación está incluida en el archivo kernreg.ado. La sintaxis de este comando es:

```
kernreg yvar xvar [if exp] [in range], bwidth(#) kercode(#)
     npoint(#) [gen(mhvar gridvar) nograph graph_options
```

donde bwidth(#) especifica *h*, el parámetro de suavización; kercode(#) permite escoger la función ponderal (kernel) de acuerdo al siguiente código:

1. Uniforme
2. Triangular
3. Epanechnikov
4. Cuártico (Biponderado)
5. Triponderado
6. Gaussiano
7. Coseno

npoint(#) se refiere al número de puntos igualmente espaciados (trama) de *xvar* a ser considerados para la estimación. La amplitud de banda, función ponderal y el número de puntos para la trama no son opcionales, deben especificarse para poder correr el programa (de otra forma, se detendrá y mostrará un mensaje de error). gen(*mhvar gridvar*) permite crear dos variables nuevas para contener la estimación de las medias condicionales (estimación de regresión no paramétrica) y sus correspondientes puntos en la trama de *x*. De manera pre-establecida el programa produce la gráfica de los valores de respuesta estimados (conectados por líneas) contra los correspondientes valores de *x*. De forma adicional, es posible usar las opciones gráficas. Finalmente, nograph suprime el desplegado de la gráfica.

Las curvas incluidas en la Figuras 6.1-6.3 (Capítulo 6) fueron obtenidas utilizando este comando, se guardaron en disco y se editaron con el editor gráfico Stage. Cada estimación se puede reproducir pulsando:

Datos del géiser “Old Faithful”

```
. kernreg wait dura, bwidth(0.8) kercode(4) npoint(100) gen(mp8
gridp8)
```

Datos de impacto de motocicletas:

```
. kernreg accel time, bwidth(2.4) kercode(4) npoint(100) gen(m2p4
grid2p4)
```

Datos de simulación de crecimiento (usando las abreviaciones de las opciones):

```
. kernreg ltsim age, b(0.22) k(4) np(100) gen(mp22 gridp22)
```

7.6.3 Aproximación PPPR del estimador Nadaraya-Watson.

De manera semejante al capítulo que introdujo a los HDP y a la estimación de densidad por PPPR, se presentan en esta sección varios programas que utilizan este eficiente procedimiento para la estimación no paramétrica de la regresión basados en los algoritmos y programas descritos en Härdle (1991) y Scott (1992). Además, se utilizaron como guía los programas en lenguaje "C" (revisiones de las versiones incluidas en Härdle, 1991) obtenidas de la librería electrónica Statlib por transferencia de archivos (ftp). El archivo ado warpreg.ado llama un programa ejecutable en Turbo-Pascal, warpregf.exe el cual es una versión modificada del programa presentado anteriormente warpings.exe llamado en los archivos ado para la estimación de densidad por PPPR. Este programa ejecutable TP para la regresión no paramétrica, lee de dos archivos ASCII generados por el comando outfile de Stata a los pares x-y, la amplitud de banda, el número de intervalos pequeños (M) y la función ponderal; posteriormente lleva a cabo los cálculos necesarios y produce otro archivo ASCII con los resultados. Las funciones ponderales (kernel) siguen el acostumbrado código:

1. Uniforme
2. Triangular
3. Epanechnikov
4. Cuártico (Biponderado)
5. Triponderado
6. Gaussiano

En esta obra se ha incluido al kernel Gaussiano (no considerado en los programas originales) haciendo un ajuste a la amplitud de banda. La sintaxis para el comando warpreg.ado es:

```
warpreg yvar xvar [if exp] [in range], bwidth(#) mval(#)
   kercode(#) [sort gen(mhvar midvar) nograph graph_options]
```

donde bwidth(#) especifica h, el parámetro de suavización; mval(#) representa el número de histogramas desplazados promediados utilizados para la estimación de las densidades requeridas; kercode(#) especifica el tipo de kernel (función ponderal); sort se utiliza para indicar que los datos han sido ordenados por x y y; gen(mhvar midvar) permite generar dos nuevas variables conteniendo la estimación de regresión y los puntos medios usados para el cálculo. Finalmente, es posible utilizar las opciones gráficas (de la combinación graph, twoway de Stata) o suprimir la salida gráfica (usando nograph). El resultado pre-establecido de este comando es una gráfica con la

correspondiente curva de regresión junto con los puntos medios considerados en adición a algunas etiquetas informativas.

Por ejemplo, usando los datos del géyser "Old faithful", para obtener la figura 6.7 se teclea:

```
. warpreg wait dura, bwidth(0.4) mval(10) kercode(4)
```

7.6.4 Selección de banda por funciones penalizantes.

Como en los programas para selección de amplitud de banda óptima por validación cruzada para estimación de densidad (Salgado-Ugarte, *et al*. 1995b), el algoritmo contiene los pasos típicos del método PPPR ligeramente modificados:

- Agrupación de los datos
- Creación de pesos
- Ponderación de intervalos

La principal diferencia es que después del agrupamiento de los datos se lleva a cabo un ciclo en el intervalo especificado para el parámetro M (δ se mantiene constante) y se incluye la creación de pesos y una ponderación modificada de los intervalos no vacíos y sus vecinos también no-vacíos. Este procedimiento es lineal en el número de observaciones en contraste con el algoritmo directo que depende en el cuadrado de *n* (Härdle, 1991).

En esta obra se ha modificado el algoritmo y programas de Härdle (1991) para estimar G(M). El archivo ado gwarpreg.ado llama un programa ejecutable de Turbo Pascal (gwarpren.exe) el cual contiene los pasos arriba mencionados. La sintaxis del comando gwarpreg es:

```
gwarpreg yvar xvar [if exp] [in range], delta(#) selec(#)
    kercode(#) [mstart(#) mend(#) bound(#) gen(score mvalue
    bandwidth)]
```

donde `delta` permite especificar el grado deseado de precisión para la estimación, `selec` se refiere a la función penalizante de acuerdo a los números del Cuadro 6.2 (Capítulo 6) y `kercode` es el código de kernel (1 = Uniforme, 2 = Triangular, 3 = Epanechnikov, 4 = Cuártico (biponderado), 5 = Triponderado y 6 = Gaussiano). Estas opciones (`delta`, `selec` y `kercode`) deben escribirse para que el programa funcione; si no se introducen, el programa se detiene y muestra un mensaje de error. Además, el programa permite definir el intervalo de valores de *M* con el fin de buscar el valor de puntaje mínimo y la proporción de los datos a considerar para la estimación. De manera preestablecida, el programa considera `mstart` = 5, mend = int(0.3*max(*x*)-min(*x*))/delta) y un valor de frontera (`bound`) de 0.1, lo que significa que aproximadamente el 90% de los datos se consideran para la estimación. Este parámetro de frontera se incluye para reducir la influencia de los puntos extremos (en las fronteras del intervalo donde ocurren) en la estimación (un valor alto recortará correspondientemente el número de observaciones a considerar). Es posible especificar otro valor para la frontera si se desea, pero generalmente no ocurre variación significativa en la amplitud de banda óptima resultante (Härdle, 1990). Finalmente, es posible generar tres variables nuevas con los valores para: *G(M)* (llamada `score`), *M* (`mvalue`) y las amplitudes de banda correspondientes. Estas variables con los resultados pueden usarse para generar gráficos personalizados que documenten la existencia del mínimo.

Después de proporcionar toda esta información, el programa realiza los cálculos y los resultados son importados y graficados por el paquete Stata. La primera gráfica es el valor del puntaje (score) G(M) en relación a M, seguida por otra gráfica también del puntaje, pero ahora en función del valor de la amplitud de banda. Estas dos gráficas son auxiliares en la búsqueda y selección de los intervalos que contienen valores de puntaje mínimos por medio de un procedimiento de ensayo y error. Si el intervalo analizado contiene un mínimo se mostrará en los gráficos. Despues de las gráficas, un listado de los cinco valores más bajos para los puntajes y sus correspondientes valores de M y amplitud de banda se muestran en la pantalla. De esta forma gwarpreg.ado proporciona toda la información necesaria para investigar la amplitud de banda óptima. De otra forma, al variar el valor de delta y al especificar valores iniciales y finales diferentes para M, podemos localizar el mínimo de la función al observar las gráficas y cuadros resultantes. Por ejemplo, utilizando los datos de impacto de motocicletas (Capítulo 6), una `delta` de 0.2, la función selectora de Shibata (1), kernel cuártico (4), `mstart` = 5, `mend` = 20 y un valor de frontera de 0.1 se teclearía:

```
. gwarpreg accel time, d(0.2) s(1) k(4) me(20)
```
(se omiten gráficas)

```
------------------------------------------------------------------------
Adjusted Prediction error G(M) for WARPing Nadaraya-Watson regression
------------------------------------------------------------------------
M-value = 8        Score value = 534.06671       Bandwidth =   1.6000
M-value = 9        Score value = 537.53998       Bandwidth =   1.8000
M-value = 7        Score value = 539.06793       Bandwidth =   1.4000
M-value = 10       Score value = 542.87994       Bandwidth =   2.0000
M-value = 11       Score value = 548.75659       Bandwidth =   2.2000
```

Para este ejemplo se sugiere un valor de amplitud de banda igual a 1.6 después de observar las gráficas y el cuadro de valores mínimos. Con el fin de comprobar los resultados se han usado cada una de las funciones de selección incluidas y hasta donde fuera posible, los mismos valores de entrada. Un resumen de los resultados se incluye en el Cuadro 6.3 del Capítulo 6.

Siguiendo a Härdle (1991), para los datos del geyser se obtiene una curva con dos mínimos locales que indican amplitudes de banda de 0.65 y 1.75 al dar el siguiente comando:

```
. gwarpreg wait dura, d(0.05) sel(2) k(4) me(40)
```

Para los datos de simulación de crecimiento, el comando empleado utilizando a *T* se incluye a continuación:

```
. gwarpreg ltsim age, d(0.02) sel(5) k(4) me(40)
```

7.6.5 Estimada por *k*-Vecinos más cercanos (*k*-*NN*).

En esta sección se presenta un sencillo programa del estimador *k-NN* que toma en cuenta la sugerencia de ordenar *xvar* y *yvar* (XploRe Systems, 1993). El comando `knnreg.ado` tiene la siguiente sintaxis:

knnreg *yvar xvar* [if *exp*] [in *range*], knum(#) [gen(mkvar) nograph *graph_options*]

donde knum especifica el número de vecinos a usarse para la estimación. Este comando, en forma preestablecida, dibuja el gráfico de dispersión con la regresión estimada sobrepuesta. De manera opcional, es posible generar una nueva variable con los resultados de la estimación y utilizar las opciones gráficas o suprimir las gráficas. La nueva variable con los resultados contendrá los valores faltantes correspondientes a los extremos de los valores de respuesta ordenados (esto debido a que su cálculo se basa en las diferencias de valores desplazados sin ajuste en los extremos). Es posible obtener los residuales y evaluar el ajuste como de costumbre. En los programas previos, esto puede lograrse sólo por medio de procedimientos especiales (Altman, 1992; Gasser, *et al.* 1986).

7.6.6 Algunas notas finales.

- Los programas **kernreg.ado** y **knnreg.ado** pueden usarse por todos los usuarios de Stata. Por otra parte, **warpreg.ado** y **gwarpreg.ado** pueden instalarse sólo para aquellos con Stata para DOS. Para usuarios de UNIX sería necesario compilar nuevamente con un Pascal de UNIX o traducir los programas a C para luego compilar y realizar un archivo ejecutable.

- Utilizando los algoritmos y programas de Härdle (1991) como una base, todos los procedimientos fueron adaptados a Turbo Pascal de Borland (Ver. 7.0) como aplicaciones individuales. Luego, versiones simplificadas de las rutinas necesarias (procedimientos en Turbo Pascal) se compilaron y se guardaron como archivos ejecutables que se integraron a los comandos de Stata por medio del comando shell. Al momento, estos procedimientos TP pueden procesar 1000 observaciones y anchos de intervalo pequeños.

- Cuando hubo duda, se usaron como guía los listados de las funciones del programa estadístico S y el código fuente en C disponible en la librería electrónica Statlib.

- Los resultados de estos programas se verificaron por comparación con las gráficas y cuadros del libro del Dr. Härdle (1991), con los resultados de las funciones de S en la computadora del Centro de Cómputo de la Universidad de Tokio y con la lista de resultados obtenidos por las funciones de S proporcionadas amablemente por el Dr. Brian Ripley. Se llevaron a cabo algunas comparaciones adicionales con los resultados obtenidos con XploRe (version 3.1) (XploRe Systems, 1993).

- El conjunto completo de funciones de S y código fuente en C de Härdle (1991) utilizadas para implementar los procedimientos basados en el PPPR y los archivos ado se obtuvieron a través de e-mails de Statlib. Brian Ripley amablemente proporcionó la lista de resultados de estas funciones para su comparación y validación. Con la ayuda del Dr. J. Hilbe, el Dr. Hárdle proporcionó una copia de su programa XploRe, el cual fue usado para llevar a cabo comprobaciones adicionales.

7.7 Programas en versiones para Windows o actualizados

Para concluir este capítulo incluiré algunas versiones especiales de los programas explicados en secciones anteriores y algunos otros nuevos resultantes de su actualización.

7.7.1 Versiones para Windows (3.1 y posteriores).

Si bien, con la llegada de Windows 95, la separación entre el MS-DOS y Windows desapareció, en la época de Windows 3.1 era necesario modificar un poco los programas que corrían en la versión para MS-DOS de Stata.

El primer programa es warpdenw.ado, versión para Windows de warpden.ado. Su sintaxis y opciones son semejantes con la única diferencia en el nombre del comando (que termina con una "w" para indicar que trabaja en Windows.

```
warpdenw nomvar [if exp] [in range],
     bwidth(#) mval(#) kercode(#) [step nosort
     gen(densivar gridvar) nograph graph_options]
```

Ejemplos:

```
. warpdenw snow, b(14.5) m(10) k(6)
. warpdenw bodlen, b(8) m(10) k(4) step
```

Los programas para validación cruzada también requirieron su modificación para trabajar en Windows 3.1. En el caso de la VCMC se escribió **l2cvwarw.ado**:

```
l2cvwarw varname [if exp] [in exp], delta(#) kercode(#)
     [mstart(#) mend(#) gen(cvval mval hval)
     nograph graph_options]
```

Tanto la sintaxis como las opciones son las mismas que para l2cvwarp.ado. Una diferencia es que esta versión sólo dibuja la gráfica del puntaje de VC contra el valor de *M* y no la amplitud de banda directamente.

Ejemplo:

```
. l2cvwarw snow, delta(1) kercode(6) mstart(3) mend(30) xlog xlab xline(11.85)
```

Para la VCS se escribió **bcvwarpw.ado** el cual tiene la siguiente sintaxis:

```
bcvwarpw varname [if exp] [in exp], delta(#) kercode(#)
     [mstart(#) mend(#) gen(bcvval mval hval)
     nograph graph_options
```

La única diferencia es en cuanto al nombre del comando y a que también dibuja sólo la gráfica de puntaje de VC contra el valor de *M*.

Ejemplo:

```
. bcvwarpw snow, d(1) k(2) ms(15) me(70) xlog xlab xline(35.29)
```

Para la regresión no paramétrica, se requirió escribir **warpregw.ado**, con la siguiente sintaxis:

```
warpregw yvar xvar [if exp] [in range], bwidth(#) mval(#)
   kercode(#) [sort gen(mhvar midvar) nograph graph_options]
```

Ejemplo:

```
. warpregw wait dura, bwidth(0.4) mval(10) kercode(4)
```

El equivalente para calcular el puntaje de VC y elegir la mejor amplitud de banda, es **gwarprew.ado**, con la siguiente sintaxis:

```
gwarprew yvar xvar [if exp] [in range], delta(#) selec(#)
    kercode(#) [mstart(#) mend(#) bound(#) gen(score mvalue
    bandwidth)]
```

De nuevo, el comando es semejante en sintaxis y opciones respecto a gwarpreg.ado, pero sólo dibuja la gráfica del puntaje de VC contra el valor de *M*.

Ejemplos:

```
. gwarprew accel time, d(0.2) s(1) k(4) me(20)

. gwarprew wait dura, d(0.05) sel(2) k(4) me(40)

. gwarprew ltsim age, d(0.02) sel(5) k(4) me(40)
```

7.7.2 Versiones actualizadas

Las siguientes son versiones actualizadas de programas que originalmente fueron escritos de manera muy sencilla. Entre estos se incluyen aquellos para trazas de densidad y los escritos para estimadores de densidad por kernel de amplitud de banda variable.

Para calcular traza de densidad con la función ponderal cuadrada (boxcar) se escribió boxdent.ado, el cual tiene la siguiente sintaxis:

```
boxdent varname [if exp] [in range], hval(#) [gen(denvar) nograph
    graph_options]
```

Opciones:

`hval` es la constante que especifica la amplitud de la ventana alrededor de cada punto de los datos.

`gen(denvar)` permite generar una nueva variable conteniendo los valores de traza de densidad.

`nograph` suprime el desplegado gráfico.

`graph_options` se refiere a cualquiera de las opciones permitidas con el comando `graph, twoway`.

A semejanza con **boxdetra.ado** este procedimiento lleva a cabo resúmenes condicionales para cada una de las observaciones en el conjunto de datos. Por tanto, el tiempo que requiere para su cálculo es proporcional al número de los datos y se requiere paciencia.

Ejemplo:

```
. boxdent ozone, h(75) gen(dtrace)
```

Calcula y muestra la gráfica con la traza de densidad generando además la variable "dtrace" con los resultados.

Es posible dibujar la gráfica con los resultados generados escribiendo lo siguiente:

```
. graph dtrace ozone, xlab ylab c(l) s(p)
```

La versión actualizada para calcular trazas de densidad con un número de puntos uniformemente distribuidos es **dentrace.ado**, con la siguiente sintaxis:

```
dentrace varname [if exp] [in range], hval(#) fcode(#) [npoints(#)
    gen(denvar midvar) nograph graph_options]
```

Opciones

`hval(#)` permite establecer el ancho de la ventana (banda)

`fcode(#)` permite indicar el código para la función de peso: 1 cuadrada; 2 coseno

`npoints(#)` se usa para especificar el número de puntos uniformemente espaciados para la estimación

`gen` se utiliza para generar dos variables nuevas: "denvar" con los valores de densidad y "midval" la variable con los puntos usados para el cálculo.

`nograph` y `graph_options` como en los casos anteriores.

`hval` y `fcode` no son opcionales. Si el usuario no las proporciona, entonces el programa para y muestra un mensaje de error en la pantalla.

Aún cuando dentrace considera de manera pre-establecida sólo 50 puntos uniformemente espaciados, el tiempo requerido para su cálculo es proporcional al número de observaciones. Puede requerirse paciencia.

Ejemplo

```
. dentrace ozone, h(75) f(1) gen(dtrace midpt)
```

Después de algún tiempo (dependiendo del número de observaciones) los resultados son mostrados como una gráfica de densidad y se generan dos variables “dtrace” y “midpt” con los resultados.

Es posible dibujar la gráfica con los resultados generados tecleando:

```
. graph dtrace midpt, xlab ylab c(l) s(p)
```

La versión actualizada de **adgakern.ado** representa una mejora considerable sobre el programa original. En primer lugar, el nombre del programa (**varwiker.ado**) refleja la denominación recomendada en la actualidad por los especialistas (es decir, EDK de amplitud de banda variable). Como su predecesor, **varwiker.ado** estima la densidad por medio del kernel Gaussiano a través de dos pasos pero grafica directamente los resultados. Este programa actualizado tiene la siguiente sintaxis:

```
varwiker varname [if exp] [in range], bwidth(#) [gen(denvar)
    nograph graph_options]
```

Opciones:

`bwidth(#)` permite especificar la amplitud de banda (al final representa el valor medio geométrico de las amplitudes utilizadas en el cálculo).

`gen` se utiliza para generar una variable nueva “denvar” con los valores de densidad estimados.

`nograph` y `graph_options` como en casos anteriores.

bwidth no es opcional. Si el usuario no la proporciona, entonces el programa para y muestra un mensaje de error en la pantalla.

Este programa primero estima la densidad usando un kernel Gaussiano con amplitud de banda fija y usa esta estimación para determinar pesos locales inversamente proporcionales a la densidad preliminar. Estos pesos son utilizados posteriormente para ajustar la amplitud de banda de tal forma que se haga angosta a densidades altas (reteniendo detalles) y amplias a bajas densidades (eliminando ruido).

Debido a que este programa requiere del cálculo de pesos locales para cada observación individual con base en una estimación preliminar de la densidad, el tiempo requerido para su cómputo es proporcional al número de datos. Se requiere paciencia.

Ejemplo:

```
. varwiker infmorat,  b(20) gen(vdensity)
```

Después de algún tiempo (que depende del número de observaciones) los resultados se grafican y se genera una variable nueva “vdensity” con los valores de densidad estimados.

Los valores de densidad pueden ser usados para repetir y modificar la gráfica tecleando:

```
. graph vdensity infmorat, c(s) s(.)
```

En la pantalla aparecerán las estimaciones de densidad para cada uno de los valores originales de la variable de interés. Es posible combinar esta gráfica con las opciones `oneway` y `box` del comando gráfico de Stata.

La versión actualizada de **adgaker2.ado** es el programa **varwike2.ado**. A semejanza con su predecesor, después de calcular la densidad y los pesos, para la estimación final de la densidad con amplitudes variables de banda se emplea un número discreto de puntos uniformemente espaciados (50 de manera pre-establecida). Sin embargo y como en el caso de **varwiker.ado**, este nuevo programa grafica directamente los resultados. Adicionalmente, se han agregado opciones muy útiles al trabajar con distribuciones multimodales. La sintaxis de **varwike2.ado** es:

```
varwike2 varname [if exp] [in range], bwidth(#) [npoints(#)
    numodes modes gen(denvar gridvar) nograph graph_options]
```

Opciones

bwidth(#) permite especificar (finalmente como un valor medio geométrico) la amplitud de banda. Este valor no es opcional. Si el usuario no lo proporciona, el programa se detiene y muestra un mensaje de error en la pantalla.

`npoint(#)` posibilita especificar el número de puntos uniformemente espaciados en el recorrido de la variable de interés que son usados para la estimación de densidad. El valor pre-establecido es de 50 puntos.

`numodes` se utiliza para mostrar el número de modas en la estimación de densidad.

`modes` permite hacer una lista de los valores estimados para cada una de las modas en la densidad. La opción `numodes` debe ser incluida para que funcione.

`gen` permite generar la variable “denvar” con los valores de densidad estimados en los puntos dados por “gridvar”.

`nograph` y `graph_options` como en los casos anteriores.

De nueva cuenta, debido a la necesidad de calcular pesos locales para cada una de las observaciones y a pesar de utilizar en el paso final un número discreto de puntos uniformemente espaciados, el tiempo requerido para su cálculo puede ser considerable si el valor de n es alto. Puede requerirse de la paciencia del usuario.

Ejemplo:

```
. varwike2 infmorat, b(20) gen(vdensity midpt)
```

Después de algún tiempo (que depende del número de observaciones y del número de puntos para la estimación) los resultados son graficados y se generan las variables vdensity y midpt con los resultados (densidad y punto de cálculo respectivamente).

```
. varwike2 infmorat, b(20) numo mo nog
```

El número de modas y sus valores estimados aparecen como un cuadro en la ventana de Resultados de Stata (sin gráfica).

Estos dos últimos programas (**varwiker.ado** y **varwike2.ado**) se presentan en Salgado-Ugarte, (2000).

7.8 Programa EDK 2000

A continuación se presenta un breve tutorial para el programa EDK 2000 tal como se presenta en el archivo readme.txt que lo acompaña:

El programa EDK 2000 es un ejecutable escrito en Visual Basic Ver. 5.0 (Microsoft Corporation, 1997) que integra los procedimientos más conocidos para el cálculo de los estimadores de densidad por kernel, incluidas las “trazas de densidad” presentadas en Chambers, *et al*. (1983) y el estimador de densidad por kernel gaussiano de amplitud variable (Fox, 1990; Salgado-Ugarte, *et al*. 1993). Esta versión todavía es muy simple pero con la ventaja de funcionar en el ambiente Windows.

Programa EDK2000 Versión 1.0 Enero, 2000

Conjunto de programas para la estimación no paramétrica de la densidad por medio de diferentes funciones ponderales (kerneles) para datos univariados.

Requerimientos:
Sistema 486 o superior (Pentium recomendado)
S.O. Windows 95 o 98 (Recomendado)

Guía breve de uso:

A la fecha se proporciona la siguiente información. Versiones posteriores contarán con ayuda dentro del programa.

El conjunto de programas en el diskette de distribución debe copiarse a una carpeta propia en el disco duro.

Para su ejecución puede utilizarse el Explorador de Windows abriendo la carpeta donde se encuentra el programa y oprimiendo dos veces (doble click) el botón del ratón en el icono del archivo ejecutable "edk2m1.exe".

Se abre una ventana de presentación que se cierra al oprimir el botón del ratón en el centro de dicha ventana. Enseguida aparece la pantalla principal del programa con los siguientes menúes:

Archivo
- Abrir
- Guardar resultados
- Cerrar

Estimar
- Traza de densidad
  - Directa
  - Discreta
- EDK
  - Directa
  - Discreta

EDKVariable(Gauss)
Directa
Discreta
Graficar
Puntos
Líneas
Ayuda
Acerca de

El menú Archivo-Abrir permite la lectura de un archivo de texto con los datos numéricos. Estos deben estar escritos en una solo columna, sin encabezado. El nombre del archivo debe escribirse completo, incluyendo su extensión.
Al abrir el archivo, en la parte izquierda de la pantalla se actualizan los datos sobre "Archivo leído" y "Número de observaciones".

Con los datos en la memoria es posible abrir el Menú Estimar. La opción "Traza de densidad" permite calcular los estimadores presentados por Chambers, *et al*. 1983.

La opción directa utiliza los valores de los datos para la estimación de la densidad y al elegirla se abre un cuadro de diálogo en el que se necesita especificar la función ponderal (cuadrada o coseno) y la amplitud de ventana.
Se cuenta con un botón para el cálculo de ventanas de acuerdo a las expresiones para valores óptimos de Silverman (1986) y Härdle (1991), así como a la sobresuavizada de Scott (1992). Detalles sobre estas expresiones se encuentran en Salgado-Ugarte, *et al*. (1993; 1995a y 1995b). Una vez especificado el ancho de la ventana (sea el recomendado por la formula o algún otro valor especificado en la ventana de texto por el usuario) se oprime el botón "Calcular densidad". Si no se especifica valor para el ancho de ventana, aparece un mensaje de error y se regresa al cuadro de diálogo.

Al finalizar el cálculo aparece el mensaje "Cálculo Terminado". Se debe entonces oprimir la cruz del ángulo superior derecho del cuadro de diálogo para cerrar el mismo. Con los resultados en la memoria se puede pasar al menú Graficar y si se desea a la opción Archivo-Guardar Resultados (ver abajo para estas opciones de menú).

La opción discreta permite hacer el cálculo considerando un numero de puntos uniformemente espaciados (50 como valor pre-establecido). Asimismo se requiere especificar la función ponderar (cuadrada o coseno). Se tiene el botón para calcular ancho de ventana y la densidad. Se cierra la ventana y se puede ir al menú Graficar o a la opción Archivo-Guardar Resultados (ver abajo para estas opciones de menú).

La opción "EDK" permite calcular los Estimadores de Densidad por Kernel presentados por Härdle (1991) y Scott (1992). Para detalles ver los artículos de Salgado-Ugarte (1993; 1995a; 1995b). Esta opción puede hacerse de manera directa y discreta. En este último caso el número de puntos pre-establecido es de 50. El usuario puede utilizar un número mayor si lo requiere. Ambas opciones abren un cuadro de diálogo en el que se debe especificar una función ponderal de las siguientes:

Cuadrada, Triangular, Epanechnikov, Biponderada, Triponderada, Gaussiana y Coseno.

Se cuenta con el botón para "Calcular Bandas" (dos óptimas y una sobresuavizada) las cuales automáticamente se ajustan al tipo de función ponderal (kernel) especificado de acuerdo a la idea de

los "kerneles equivalentes" de Scott (1992) para producir estimaciones con el mismo grado de suavización.

Una vez especificado en ancho de la banda (ventana) se oprime el botón "Calcular Densidad" y al aparecer el mensaje "Calculo terminado" se debe cerrar el cuadro. Si no se especifica valor para el ancho de banda y se oprime el botón de "Calcular Densidad" aparece un mensaje de error y se regresa al cuadro de diálogo.

Con los resultados en la memoria se puede pasar al menú Graficar y si se desea a la opción Archivo-Guardar Resultados (ver abajo para estas opciones de menú).

La opción EDKVariable(Gauss) permite calcular el Estimador de Densidad por Kernel Gaussiano con banda de amplitud variable (ver Fox, 1990 y Salgado-Ugarte, *et al*. 1993). Este estimador resulta muy conveniente ya que proporciona detalles de la distribución donde los datos abundan (empleando bandas angostas) y disminuye el ruido donde los datos son escasos (con bandas mas anchas). Para lograr esto el algoritmo realiza una estimación preliminar de la densidad, la cual es usada como referencia para en un segundo paso, calcular la densidad tomando a los valores iniciales como factores de ajuste a la banda.

A semejanza de los procedimientos anteriores están incluidas dos opciones: Directa y Discreta. En la directa se utilizan todos los datos para el cálculo y en la discreta se usa un número de puntos uniformemente espaciados en la segunda parte del algoritmo de cálculo (lo que resulta mas eficiente).

Debido a la cantidad de cálculos requeridos, esta opción puede llevar un tiempo considerable si el número de los datos es elevado (miles de observaciones) por lo que se pide paciencia al usuario.

La opción "WARP" permite el cálculo de "histogramas desplazados promediados" y a los Estimadores de densidad por kernel. Es un procedimiento muy eficiente para calcular a estos estimadores ya que reduce considerablemente el número de operaciones. Al elegirla aparece un cuadro de diálogo para especificar una de las siguientes funciones ponderales (kerneles):

Cuadrada, Triangular, Epanechnikov, Biponderada, Triponderada y Gaussiana.

Se cuenta con los botones para calcular las bandas y la densidad. El número pre-establecido de histogramas a promediar es de 5, valor que conduce a un histograma desplazado promediado. Utilizando un valor mayor (10 o más) el resultado es equivalente a un Estimador de Densidad por Kernel. Se recomienda utilizar un valor de 10 para no aumentar el número de operaciones y disminuir la eficiencia del método.

La opción "Graficar" permite el desplegado gráfico de los resultados en el cuadro de imágenes de la pantalla principal del programa. Estas rutinas producen un gráfico con el eje horizontal representando a la variable de interés y el vertical a la densidad (frecuencia). La opción "Puntos" como su nombre lo indica, presenta los resultados como puntos aislados, y la de "Lineas" conecta los valores de densidad con líneas rectas y además, presenta a cada uno de los datos como puntos en el eje horizontal. Las rutinas gráficas incluidas son muy simples (presentan valores numéricos máximos y mínimos para cada eje y un encabezado con el estimador utilizado, el número de puntos usado para la estimación en el caso discreto, el kernel y la amplitud de banda) pero son de utilidad ya que permiten observar de una manera muy clara la forma de la distribución de los datos. Si se desean gráficos de mejor presentación, los resultados deberán guardarse (con la opción Archivo-Guardar Resultados) e importarse a otro programa capaz de realizar gráficos para presentaciones.

La opción Archivo-Guardar Resultados, permite guardar los resultados en un archivo de texto que puede leerse en otros programas para hacer gráficas con presentación profesional.

Los métodos directos producen un archivo de texto con dos columnas:

dato, densidad

Los métodos discretos tienen como resultado dos columnas de la siguiente manera:

punto medio, densidad

Al escoger esta opción se abre un cuadro de diálogo en el que se escribe el nombre del archivo con los resultados. Se recomienda incluir la extensión "txt" (ejemplo: resulta.txt) para facilitar su importación como archivo de texto en otros paquetes computarizados.

El menú Ayuda contiene una opción única (Acerca de) que abre un cuadro de diálogo con información del programa EDK2000. Este cuadro tiene dos botones: Aceptar, que cierra el cuadro de información e
Información.. que proporciona información del sistema de cómputo que se está utilizando.

Para reporte de fallas e información adicional, contactar a

Dr. Isaías H. Salgado Ugarte
isalgado@servidor.unam.mx

Bibliografía

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

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

Chambers, J.M., W.S. Cleveland, B. Kleiner y P.A. Tukey 1983. *Graphical methods for data analysis*. Belmont, CA: Wadsworth.

Comparini, A. y E. Gori, 1986. Estimating modes and antimodes of multimodal densities. *Metron (Italian Statistical Review)*, **44**: 307-332.

Cook, R. D., y S. Weisberg, 1982. *Residuals and Influence in Regression*. Chapman and Hall, Londres.

Cook, R. D., y S. Weisberg, 1994. *An Introduction to Regression Graphics*. John Wiley & Sons. Nueva York.

Cox, D.R. 1966. Notes on the analysis of mixed frequency distributions. *The British Journal of Mathematical and Statistical Psychology*, **19**: 39-47.

Doane, D.P. 1976. Aesthetic frequency classifications. *The American Statistician*, **30**: 181-183.

Efron, B. 1982. *The jackknife, the bootstrap, and other resampling plans*. Philadelphia: Society for Industrial and Applied Mathematics.

Egglishaw, H.J., 1970. Production of salmon and trout in a stream in Scotland. *Journal of Fish Biology*, **1**: 117-136.

Emerson J.D. y D.C. Hoaglin 1983. Stem-and-leaf displays. In: *Understanding robust and exploratory data analysis*, ed. Hoaglin, D.C., F. Mosteller y J.W. Tukey, 7-30. Nueva York, John Wiley & sons.

Epanechnikov, V.A. 1969. Nonparametric estimation of a multidimensional probability density. *Theor. Probab. Appl*., 14: 153-158.

Fisher, R.A., 1932. *Statistical methods for research workers*. Cuarta edición. Oliver and Boyd, Edinburgh.

Fisher, R.A., 1958. *Statistical methods for research workers*. XIII edición. Oliver and Boyd, Edinburgh.

Fox, J. 1990. Describing univariate distributions. In: *Modern Methods of Data Analysis*, eds. J. Fox y J.S. Long, 58-125. Newbury Park, CA: Sage publications.

Freedman, D. y P. Diaconis 1981a. On the histogram as a density estimator: L theory, *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, **57**: 453-476.

Freedman, D. y P. Diaconis 1981b. On the maximum deviation between the histogram and the underlying density. *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, **58**: 139-167.

Frigge, M., D.C. Hoaglin, y B. Iglewicz, 1989. Some implementations of the boxplot. *The American Statistician*, **43**: 50-54.

Gasser, T., H.G. Müller y V. Mammitzsch. 1985. Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society, Series B*, **47**: 238-252.

Gasser, T., L. Sroka, y C. Jennen-Steinmetz, 1986. Residual variance and residual pattern in nonlinear regression. *Biometrika*, **73**: 625-633.

Goeden, G.B. 1978. A monograph of the coral trout, *Plectropomus leopardus* (Lacépède). *Res. Bull. Fish. Serv. Queensl.*, 1: 42 p.

Good, I.J. y R.A. Gaskins 1980. Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *Journal of the American Statistical Association*, **75**: 42-73.

Hall, P., S.J. Sheather, M.C. Jones, y J.S. Marron, 1991. On optimal data-based bandwidth selection in kernel density estimation. *Biometrika* **27**: 228-254.

Hamilton, L.C., 1992. Quartiles, outliers, and normality: some Monte Carlo results. *Stata Technical Bulletin* 6. Reprinted in *Stata Technical Bulletin Reprints*, vol 3, pp. 92-95.

Härdle, W., 1990. *Applied Nonparametric Regression* (Econometric Society Monograph Series No. 19), Cambridge University Press, Nueva York.

Härdle, W. 1991. *Smoothing Techniques. With Implementations in S*. Springer-Verlag. Nueva York

Härdle, W., y J.S. Marron, 1986. Random approximations to an error criterion of nonparametric statistics. *Journal of Multivariate Analysis*, **20**: 91-113.

Härdle, W. y D.W. Scott. 1988. Smoothing in low and high dimensions by weighted averaging using rounded points. *Technical report* 88-16, Rice University.

Härdle, W., P. Hall, y J.S. Marron, 1988. How far are automatically chosen regression smoothing parameters from their optimum. *Journal of the American Statistical Association*, **83**: 86-95.

Hartigan, J.A. y P.M. Hartigan, 1985. The Dip test of unimodality. *The annals of Statistics*, **13**: 70-84.

Hartigan, P.M., 1985. Computation of the DIP statistic to test for unimodality (Algorithm AS217). *Applied Statistics*, **34**: 320-325.

Hastie, T. y Tibshirani, R. 1990. *Generalized Additive Models*. Chapman y Hall, Londres.

Hilbe, J. 1993. Generalized linear models. *Stata Technical Bulletin* 11: 20-28.

Hilbe, J., 1994a. Negative binomial regression. *Stata Technical Bulletin* 18: 2-5.

Hilbe, J., 1994b. Comment on Royston's revision of glm. *Stata Technical Bulletin* 18: 11-13.

Hoaglin, D.C. 1983. Letter values: A set of selected order statistics. In: *Understanding robust and exploratory data analysis*, ed. Hoaglin, D.C., F. Mosteller y J.W. Tukey, 33-57. Nueva York, John Wiley & sons.

Izenman , A.J., y C. Sommer, 1988. Philatelic mixtures and multimodal densities. *Journal of the American Statistical Association*, **83**(404): 941-953.

Jones, M.C., 1989. Discretized and interpolated kernel density estimates. *Journal of the American Statistical Association*, **84**(407): 733-741.

Leinhardt, S. y S.S. Wasserman 1979. Exploratory data analysis: An introduction to selected methods. In: *Sociological methodology* 1979 ed. K.F. Schuessler. San Francisco, Jossey-Bass.

Marron, J.S. 1986. Will the art of smoothing ever become a science? *Contemporary Mathematics*. **59**: 169-178.

Marron, J.S. y D. Nolan, 1988. Canonical kernels for density estimation. *Statistics and Probability letters* 7: 195-199.

Microsoft Corporation, 1997. *Visual Basic Versión 5.0 Manual del programador*. Microsoft Corporation. E.U.A.

Müller, D.W. y G. Sawitzki, 1991. Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association*, **86**(415): 738-746.

Nadaraya, E.A. 1964. On estimating regression. *Theory of Probability and its Applications*, **10**: 186-190.

Parzen, E. 1979. Nonparametrical statistical data modeling. *Journal of the American Statistical Association*, **74**: 105-131.

Pauly, D. 1988. Fisheries research and the demersal fisheries of South-east Asia. In: *Fish Population Dynamics: the Implications for Management*, ed. J.A. Gulland, 329-348., John Wiley and Sons, Chichester

Pauly, D., M. Soriano-Bartz, J. Moreau, y A. Jarre-Teichmann, 1992. A new model accounting for seasonal cessation of growth in fishes. *Australian Journal of Marine and Freshwater Resources*, **43**: 1151-1156.

Roeder, K., 1990. Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, **85**(411):617-624.

Rosenblatt, M. 1956. Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist*. **27**: 832-837.

Royston, P., 1994a. Generalized linear models: revision of glm. *Stata Technical Bulletin* **18**: 6-11.

Royston, P., 1994b. Generalized linear models: revision of glm. Rejoinder. *Stata Technical Bulletin* **19**: 17.

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

Salgado-Ugarte, I.H. 1985. Algunos aspectos biológicos del bagre *Arius melanopus* Gunther (Osteichthyes: Arridae) en el Sistema Lagunar de Tampamachoco, Ver. B.S. Thesis, Carrera de Biología, E.N.E.P. Zaragoza, Universidad Nacional Autónoma de México.

Salgado-Ugarte, I.H. 1992. *El análisis exploratorio de datos biológicos. Fundamentos y aplicaciones.* ENEP Zaragoza UNAM y Marc Ediciones. Mexico.

Salgado-Ugarte, I.H., 1995. Nonparametric methods for fisheries data analysis and their application in conjunction with other statistical techniques to study biological data of the Japanese sea bass *Lateolabrax japonicus* in Tokyo Bay. Tesis de doctorado en Biociencia Acuática desarrollada en el Departamento de Pesquerías, Facultad de Agricultura de la Universidad de Tokio, Tokio, Japón, 389 p.

Salgado-Ugarte, I.H. y J. Curts-García. 1992. Resistant smoothing using Stata. *Stata Technical Bulletin* **7**: 8-11.

Salgado-Ugarte, I.H. y J. Curts-García. 1993. Twice reroughing procedure for resistant nonlinear smoothing. *Stata Technical Bulletin* **11**: 14-16.

Salgado-Ugarte, I.H. y M. Shimizu. 1995. Robust scatterplot smoothing: enhancements to Stata's ksm. *Stata Technical Bulletin* **25**: 23 - 26.

Salgado-Ugarte, I.H., M. Shimizu, y T. Taniuchi. 1993. Exploring the shape of univariate data using kernel density estimators. *Stata Technical Bulletin* **16**: 8-19.

Salgado-Ugarte, I.H., M. Shimizu, y T. Taniuchi, 1995a. ASH, WARPing, and kernel density estimation for univariate data. *Stata Technical Bulletin* **26**: 2-10.

Salgado-Ugarte, I.H., M. Shimizu, y T. Taniuchi, 1995b. Practical rules for bandwidth selection in univariate density estimation. *Stata Technical Bulletin*, **27**: 5-19.

Salgado-Ugarte, I.H., M. Shimizu y T. Taniuchi, 1997. Nonparametric assessment of multimodality for univariate data. *Stata Technical Bulletin* **38**: 27-35.

Salgado-Ugarte, I.H. 2000. Exploring the use of variable bandwidth kernel density estimators. *Stata Technical Bulletin* **XX**: xx-xx.

Scott, D.W. 1976. Nonparametric probability density estimation by optimization theoretic techniques. Ph.D. thesis, Department of Mathematical Sciences, Rice University.

Scott, D.W. 1979. On optimal and data-based histograms. *Biometrika*, **66**: 605-610.

Scott, D.W. 1985a. Frequency polygons: Theory and application. *Journal of the American Statistical Association*, 80(390): 348-354.

Scott, D.W. 1985b. Averaged shifted histograms: effective nonparametric density estimators in several dimensions. *Annals of Statistics*, 13: 1024-1040.

Scott, D.W. 1992. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, Nueva York.

Scott, D.W. y G.R. Terrell, 1987. Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, **82**(400): 1131-1146.

Sheather, S.J., y M.C. Jones, 1991. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, B*, **53**(3): 683-690.

Silverman, B.W. 1978. Choosing the window width when estimating a density. *Biometrika*, **65**: 1-11.

Silverman, B.W. 1981a. Density estimation for univariate an bivariate data. In *Interpreting Multivariate Data*, ed. V. Barnett, 37-53, John Wiley and Sons, Chichester.

Silverman, B.W. 1981b. Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society, B*, **43**: 97-99.

Silverman, B.W. 1983. Some properties of a test for multimodality based on kernel density estimates. In *Probability, Statistics and Analysis*, ed. J.F.C. Kingman and G.E.H. Reuter: 248-259, Cambridge University Press, Cambridge.

Silverman, B.W., 1985. Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *Journal of the Royal Statistical Society, Series B*, **47**: 1-52.

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

Sparre, P., E. Ursin y S.C. Venema, 1989. *Introduction to tropical fish stock assessment. Part 1. Manual*. FAO Fisheries Technical Paper. 306.1. Rome, FAO: 57-123.

StataCorp. 1995. *Stata Statistical Software: Release 4.0*. College Station, TX. Stata Corporation.

StataCorp. 1997. *Stata Statistical Software: Release 5.0*. College Station, TX. Stata Corporation.

StataCorp. 1999. *Stata Statistical Software: Release 6.0*. College Station, TX. Stata Corporation.

Sturges, H.A. 1926. The choice of a class interval. *Journal of the American Statistical Association*, **21**: 65-66.

Tarter, M.E. y R.A. Kronmal 1976. An introduction to the implementation and theory of nonparametric density estimation. *The American Statistician*, **30**: 105-112.

Taylor, C.C., 1989. Bootstrap choice of the smoothing parameter in kernel density estimation. *Biometrika*, **76**: 705-712.

Terrell, G.R., 1990. The maximal smoothing principle in density estimation. *Journal of the American Statistical Association*, **85**(410): 470-477.

Terrell, G.R. and D.W. Scott, 1985. Oversmoothed nonparametric density estimates. *Journal of the American Statistical Association*, **80**(389): 209-214.

Tukey, J.W. 1977. *Exploratory data analysis*. Reading, MA. Addison-Wesley.

Watson, G.S. 1964. Smooth regression analysis. *Sankhyä*, Series A, **26**: 359-372.

Wegman, E.J. 1972a. Non parametric probability density estimation: I. A summary of available methods. *Technometrics*, **14**: 533-546.

Wegman, E.J. 1972b. Nonparametric probability density estimation: II. A comparison of density estimation methods. *Journal of Statistical Computation and Simulation*, **1**: 225-245.

Weisberg, S. 1980. *Applied Linear Regression*. John Wiley & Sons, New York.

Wilk M.B. y R. Gnanadesikan 1968. Probability plotting methods for the analysis of data. *Biometrika* **55**(1): 1-17.

XploRe Systems, 1993. *XploRe — A computing Environment for eXploratory Regression and Data Analysis*. *Version 3.1*. Institut für Statistik und Ökonometrie. Berlin.

La presente obra contiene los siguientes temas:

Métodos Univariados:

- Histogramas y polígonos de frecuencia
- Trazas de densidad
- Estimadores de densidad por kernel (EDKs)
- EDKs de amplitud de banda variable
- Reglas prácticas para elección de amplitud de intarvalo/banda
- Elección de banda por validación cruzada
- Prueba bootstrap de multimodalidad

Métodos Bivariados:

- Regresión por kernel
- Regresión por vecinos cercanos
- Elección de banda en regresión no paramétrica

Programas computarizados:

- Programas para estimación no paramétrica
- Paquete EDK2000

Este libro puede utilizarse como texto complementario en cursos (licenciatura y posgrado) de Estadística, Bioestadística, o bien como texto principal sobre los principales métodos de suavización no paramétrica, en ciencias biológicas, de la salud, de la conducta e inclusive en las ciencias socio-económicas ó fisico-matemáticas. Los programas computarizados incluidos permiten el uso práctico de todos los métodos presentados.

ISBN 970-32-0292-6