Métodos no paramétricos avanzados

Daniela Rodriguez
Departamento de Matemática y Instituto de Cálculo, FCEyN,
Universidad de Buenos Aires y CONICET.

1 Introducción

La inferencia estadística comúnmente se focaliza sobre funciones de distribución que son puramente paramétricas o puramente no paramétricas. En los modelos paramétricos se comienza haciendo supuestos rígidos sobre la estructura de los datos para luego estimar de la manera más eficiente posible los parámetros que definen su estructura. Un modelo paramétrico razonable produce inferencias precisas mientras que un modelo erróneo posiblemente conducirá a conclusiones equivocadas.

Sin embargo, en la mayoría de las aplicaciones, los modelos paramétricos constituyen una aproximación al modelo subyacente, y la búsqueda de un modelo adecuado suele no ser sencilla. Es aquí donde, las técnicas de estimación no paramétricas surgen como una alternativas más fexibles a los modelos paramétricos.

Como punto en común, los métodos no paramétricos explotan la idea de suavizado local, que solamente utiliza las propiedades de continuidad o diferenciabilidad local de la función a estimar. El éxito del suavizado local depende de la presencia de una cantidad suficiente de observaciones alrededor de cada punto de interés, para que éstas puedan proveer la información adecuada para la estimación. Así mismo, los procedimientos de estimación no paramétricos pueden ayudar en el inicio de la investigación a descubrir la estructura probabilística que gobierna los datos de modo que los supuestos del análisis paramétrico estén bien fundamentados.

La idea básica en estimación no paramétrica es usar los datos para realizar la inferencia haciendo la menor cantidad de supuestos que sea posible. En el contexto de este curso nos referiremos a inferencia no paramtrica como un conjunto de técnicas que tratan de mantener el número de supuestos tan bajo como sea posible. Nos focalizaremos en dos problemas: Estimación de la densidad Estimación de la regresión.

2 Estimación de la densidad

Una característica básica que describe el comportamiento de una variable aleatoria X es su función de densidad. El conocimiento de la función de densidad nos ayuda en muchos aspectos. Por ejemplo, si tenemos un conjunto de observaciones generadas a partir de la densidad f y queremos conocer cuantas observaciones caen en un conjunto podemos calcular a partir de la función de densidad f la probabilidad de que la variable aleatoria X pertenezca a ese determinado conjunto como una integral sobre dicho conjunto, es decir

$$P(X \in A) = \int_A f(x)dx.$$

Si este valor es alto para un cierto conjunto A comparado con la probabilidad sobre otro conjunto B, de manera informal se podría decir que dado un conjunto de observaciones, hay una alta probabilidad de encontrar una observación en la región A y baja en la region B, es decir, la función de densidad nos dirá donde las observaciones ocurren más frecuentemente.

En la mayoría de los estudios prácticos no se conoce la función de densidad de X directamente. Y en su lugar sólo contamos con un conjunto de observaciones $X_1, \ldots X_n$ que suponemos independientes, idénticamente distribuídas y con función de densidad f desconocida. Nuestro objetivo es estudiar como estimar la función de densidad basándonos en la muestra aleatoria $X_1, \ldots X_n$.

El histograma es el estimador de la densidad más antiguo y popular. Para calcularlo se necesita un origen y un ancho para poder especificar los intervalos. $I_j = (x_0+jh, x_0+(j+1)h]$ donde $(j = ..., -1, 0, 1, ...)$ en cada intervalo el histograma cuenta el número de observaciones que caen en el. Luego se dibuja el histograma de manera que el area bajo de cada barra sea proporcional al numero de observaciones que caen en el intervalo. Una descripción más formal del histograma puede encontrarse en el ejercicio 1. En el libro de Härdle podemos esncontrar diversas propiedades estadisticas tales como el calculo del sesgo y la varianza y el estudio de su convergencia. Sin embargo el histograma tiene algunas desventajas tales como: Es constante sobre intervalos. Los resultados dependen del origen. Elección de h. Lenta velocidad de convergencia. Las discontinuidades en el estimador se deben al procedimiento y no a la distribución subyacente.

Los métodos de estimación no paramétricos han surgido con el objetivo de dar una respuesta a este problema y han sido ampliamente estudiados. En este Capítulo estudiaremos propuetas para la estimación de la función de densidad y estudiaremos sus propiedades e implementación.

2.1 Estimación por Núcleos

Sea $X_1, \ldots, X_n$ una muestra aleatoria con función de densidad $f(x)$. Como mencionamos anteriormente, el problema consiste en estimar $f(x)$ a partir de las observaciones. En primer lugar, intentaremos dar una idea intuitiva de la estimación de la función de densidad por núcleos.

Si X es una variable aleatoria con densidad f continua en x,

$$f(x) \quad = \quad F'(x) = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h}$$

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

Por otro lado un estimador natural de $P(x-h<X<x+h)$ es simplemente considerar la proporción de la muestra que cae en el intervalo $(x-h, x+h)$. Entonces dado un h suficientemente pequeño podemos deducir el siguiente estimador de $f(x)$,

$$\widetilde{f}(x) = \frac{1}{2h}\frac{\#\{X_i : X_i \in (x-h, x+h)\}}{n}.$$

Esencialemte, este estimador cuenta la cantidad de observaciones que "caen"en un entorno de radio h alrededor de x. Asímismo si consideramos F_n la distribución empírica podemos escribir a $\widetilde{f}(x)$ como

$$\widetilde{f}(x) = \frac{F_n(x+h) - F_n(x-h)}{2h}.$$

Observemos que este estimador es diferente del histograma, pues el histograma parte de una grilla o partición fija de la recta y para estimar $f(x)$ se calcula la proporción de obervaciones del intervalo que contiene al punto x dividido la longitud del intervalo. De esta manera, la densidad de dos puntos x y x' que se encuentran en el mismo intarvalo se estimatan por el mismo valor. Sin embargo, el estimador anterior calcula la proporción de obervaciones de un entorno del punto x, es decir de un intervalo centrado en x, por lo tanto por mas que x' se encuentre en el entorno de x al estimar $f(x)$, la estimación de $f(x')$ puede variar pues varía el entorno de x'.

Otra forma de expresar el estimador $\widetilde{f}(x)$ es de la siguiente manera,

$$\widetilde{f}(x) = \frac{1}{2h}\frac{\#\{X_i : X_i \in (x-h, x+h)\}}{n} = \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{2}I(|x-X_i|<h).$$

luego si definimos la función w como $w(x) = \frac{1}{2}I_{(|x|<1)}$, tenemos que $\widetilde{f}(x)$ es equivalente a

$$\widetilde{f}(x) = \sum_{i=1}^{n}\frac{1}{nh}w\Big(\frac{x-X_i}{h}\Big). \tag{1}$$

Notemos que $w \geq 0$, $\int w(s)ds = 1$, además, para cada $1 \leq i \leq n$ tenemos que $w\Big(\frac{x-X_i}{h}\Big) = \frac{1}{2}$ si y solo si $X_i \in (x-h, x+h)$, es decir la función w le otorga un peso uniforme a cada observación X_i en el entorno $(x-h, x+h)$ y 0 a cada observación fuera del entorno. A la funciíon w se la denomina núcleo uniforme o de Parzen.

Sin embargo, uno podría estar interesado en darle mayor peso a las observaciónes más cercanas a x. Esto se lograría facilmente reemplazando la función de peso o núcleo w por una función K no negativa que verifique la condición $\int K(x)dx = 1$. Ademas, si consideramos una función de pesos K con mayor suavidad obtendríamos un estimador más suave. En general los pesos utilizados decrecen de manera suave, dándole así menor pesos a las observaciones más alejadas del punto x. Algunas opciones posibles de núcleos, podrían ser

Funciones de Núcleo o Kernel	
Kernel	$K(u)$
Uniforme	$\frac{1}{2} \mathrm{I}(\lvert u \rvert \leq 1)$
Triangular	$(1 - \lvert u \rvert)\, \mathrm{I}(\lvert u \rvert \leq 1)$
Epanechnikov	$\frac{3}{4}(1 - u^2)\, \mathrm{I}(\lvert u \rvert \leq 1)$
Quartic (Biweight)	$\frac{15}{16}(1 - u^2)^2\, \mathrm{I}(\lvert u \rvert \leq 1)$
Triweight	$\frac{35}{32}(1 - u^2)^3\, \mathrm{I}(\lvert u \rvert \leq 1)$
Gaussiano	$\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}u^2)$
Coseno	$\frac{\pi}{4} \cos(\frac{\pi}{2}u)\, \mathrm{I}(\lvert u \rvert \leq 1)$

Tabla 1: Diferentes funciónes núcleos.

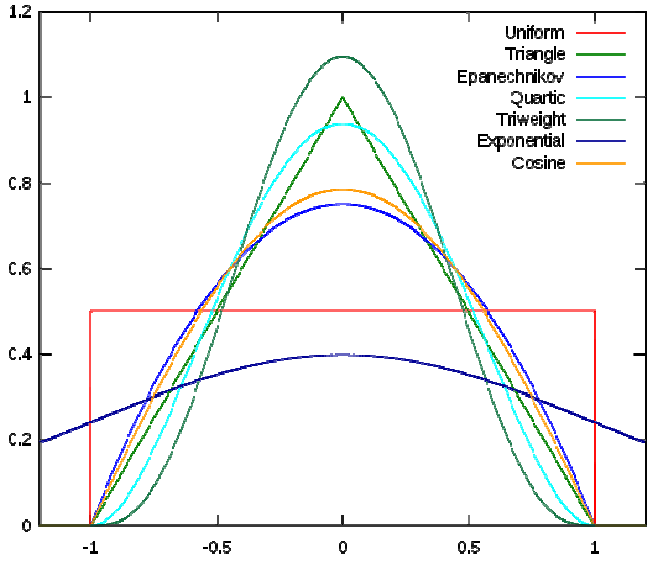

Figura 1: Funciónes núcleos.

De esta manera obtenemos el estimador que constituye uno de los estimadores no

paramétricos mas estudiados, que fue definido por Rosenblatt (1959)

$$\widetilde{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big) \tag{2}$$

donde K es una función núcleo, $h = h_n$ es llamado el *paramétro de suavizado* o *ancho de ventana* y satisface $h_n \to 0$ si $n \to \infty$.

Estos estimadores se construyen en cada punto del eje real de acuerdo con los valores muestrales más cercanos al mismo, es decir se considera un entorno alrededor de cada punto donde se desea estimar la densidad y basados en las observaciones que se encuentran en ese entorno se construye el estimador, dándole mayor peso a aquellas observaciones más cercanas y menor peso aquellas más alejadas, dentro del entorno. Para establecer los pesos se suele utilizar diversas funciones de ponderación llamadas núcleos. Los entornos están dados a partir de un parámetro de suavizado o ventana, para hacernos una idea de los mismos podemos imaginarnos una bola centrada en el punto a estimar cuyo radio corresponde justamente al ancho de banda o ventana.

El paramétro de suavizado suele ser un punto crucial en el proceso de estimación, ya que como su nombre lo indica se encuentra altamente relacionado con el nivel de suavización que se introduce en la estimación. En la Figura 2 obervamos la influencia de la elección de la ventana para un conjunto de datos y en la Figura 3 podemos apreciar la influencia del núcleo en la estimación.

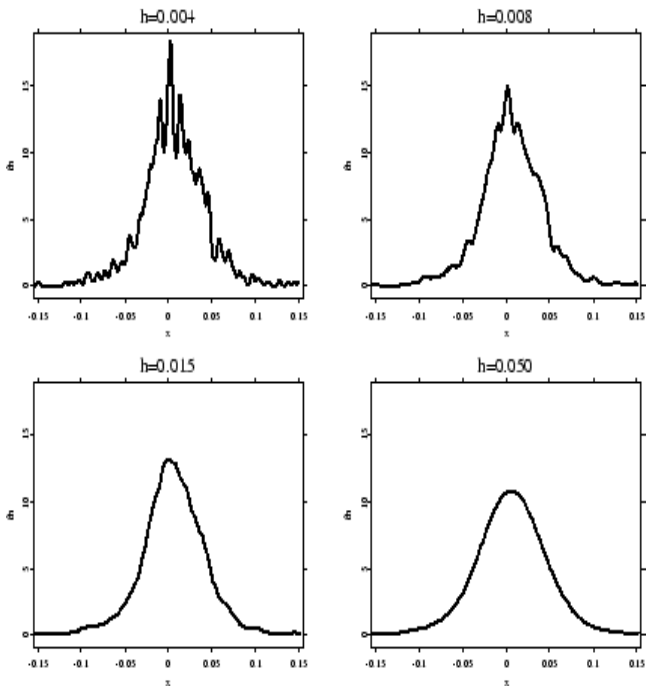

Figura 2: Estimador de densidad para diferentes anchos de banda. Datos correspondientes a rentabilidad de acciones.

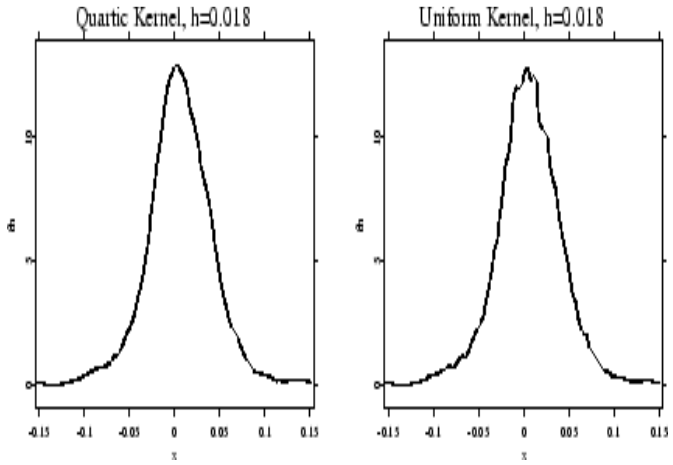

Figura 3: Estimador de densidad para diferentes núcleos. Datos correspondientes a rentabilidad de acciones.

Las propiedades del estimador de la densidad dependen de la elección del núcleo y del ancho de la ventana. La combinación de la función de ponderación, el ancho de la ventana y el tamańo de muestra hacen a la bondad de la estimación resultante. Ventanas demasiado pequeñas derivarán en estimadores muy variables ya que en cada punto los entornos carecerán de suficientes observaciones en las cuales basar la estimación. Por otra parte, un ventana demasiado grande producirá estimadores muy suaves, que no lograrán captar la estructura local de la densidad dando lugar a estimadores sesgados.

Notemos que si $\int_{+\infty}^{-\infty} K(x)dx = 1$ y $K \geq 0$, entonces el estimador $\widetilde{f}$ es también una función de densidad. $\int_{+\infty}^{-\infty} \widetilde{f}(x)dx = 1$. Pues,

$$\int_{+\infty}^{-\infty} \widetilde{f}(x)dx = \int_{+\infty}^{-\infty} \sum_{i=1}^{n} \frac{1}{nh} K\Big(\frac{x - X_i}{h}\Big)dx =$$

$$\frac{1}{n}\sum_{i=1}^{n} \int_{+\infty}^{-\infty} \frac{1}{h} K\Big(\frac{x - X_i}{h}\Big)dx = \frac{1}{n}\sum_{i=1}^{n} \int_{+\infty}^{-\infty} K(s)ds = 1$$

Por otro lado las condiciones de suavidad que le exijamos al núcleo también las heredará la función de densidad. Es decir, si el núcleo es una función continua también lo será el estimador de densidad asociada a el.

Obervemos que fijados la ventana y el núcleo, el estimador de densidad es único para el conjunto de datos dado. Y no depende de el "origen "como el histograma. La elección del núcleo suele ser una función positiva para garantizar que el estimador sea efectivamente una

densidad, sin embargo en algunas circunstancias pueden considerarse núcleos con algunos valores negativos que no siempre implicará que el estimador resultante tome también valores negativos.

2.2 Propiedades

Una de las primeras propiedades que se estudia de un estimador es el análisis del sesgo y la varianza.

Proposición : Bajo los siguientes supuestos

i) f es 2-veces derivable tal que $\int f''(s)ds < \infty$.

ii) $\int K = 1$, $\int K(s)sds = 0$ y $\int K(s)s^2ds < \infty$.

Tenemos que $E\left[\widetilde{f}(x)\right] = f(x) + \frac{h^2}{2}f''(x)\mu_2(K) + o(h^2)$ si $h \to 0$ para cada x. Donde $\mu_2(K) = \int s^2K(s)ds$

demostración:

$$\begin{aligned} E\left[\widetilde{f}(x)\right] &= E\Big(\frac{1}{hn}\sum_{i=1}^{n}K\Big(\frac{x-X_i}{h}\Big)\Big) = \frac{1}{n}\sum_{i=1}^{n}E\Big(\frac{1}{h}K\Big(\frac{x-X_i}{h}\Big)\Big) \\ &= E\Big(\frac{1}{h}K\Big(\frac{x-X_i}{h}\Big)\Big) = \frac{1}{h}\int_{+\infty}^{-\infty}K\Big(\frac{x-u}{h}\Big)f(u)du \\ &= \int_{+\infty}^{-\infty}K(y)f(x-hy)dy \end{aligned}$$

Haciendo el desarrollo de Taylor de f de orden 2 centrado en x tenemos

$$\begin{aligned} E\left[\widetilde{f}(x)\right] &= \int_{+\infty}^{-\infty}K(y)f(x-hy)dy \\ &= \int_{+\infty}^{-\infty}K(y)[f(x)+f'(x)hy+\frac{f''(x)}{2}y^2h^2+o(h^2)]dy \\ &= f(x)\int_{+\infty}^{-\infty}K(y) + f'(x)h\int_{+\infty}^{-\infty}K(y)ydy + h^2\frac{f''(x)}{2}\int_{+\infty}^{-\infty}K(y)y^2dy + o(h^2) \end{aligned}$$

y por las hipótesis sobre el núcleo concluimos la demostración.

Por lo tanto el sesgo del estimador es

$$Sesgo\left(\widetilde{f}(x)\right) = h^2\frac{f''(x)}{2}\int_{+\infty}^{-\infty}K(y)y^2dy + o(h^2)$$

Este resultado muestra que si la ventana es mayor, el sesgo aumentará y para obtener menor sesgo habría entonces que considerar ventanas más pequeñas. Por otro lado el sesgo depente de $f''(x)$ es decir depende de la curvatura de la función. Por ejmeplo, el sesgo será negativo si la derivada segunda es negativa o equivalentemente si la función tiene un máximo local.

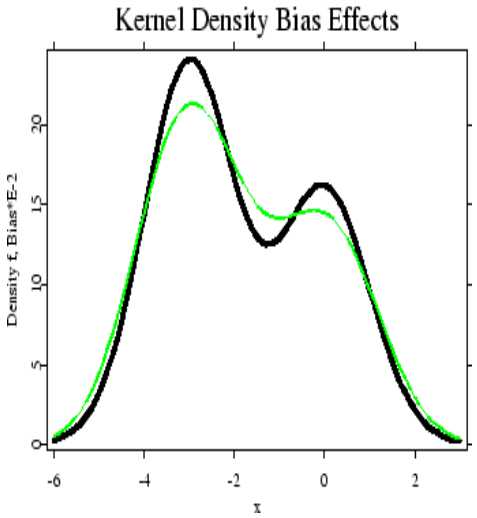

Figura 4: Estimador de densidad (en verde) y verdadera densidad (en negro).

Bajo las mismmas hipótesis introducidas anteriormente, probaremos que la varianza del estimador es

$$Var(\widetilde{f}(x)) = \frac{1}{nh}\|K\|^2 f(x) + o(\frac{1}{nh}).$$

Sea $K_h(x) = \frac{1}{h}K(x/h)$, como las X_i son i.i.d

$$\begin{aligned} Var(\widetilde{f}(x)) &= n^{-2}Var(\sum_{i=1}^{n} K_h(x - X_i)) = n^{-2}\sum_{i=1}^{n} Var(K_h(x - X_i)) \\ &= n^{-1}Var(K_h(x - X_1)) = n^{-1}[E(K_h^2(x - X_1)) - E^2(K_h(x - X_1))] \\ &= n^{-1}[E(K_h^2(x - X_1)) - (f(x) + o(h^2))^2] \end{aligned}$$

Usando los mismos argumentos que antes, es decir, cambio de variable y un desarrollo de Taylos tenemos que

$$E(K_h^2(x - X_1)) = h^{-1}\int K^2(s)f(x - hs)ds = h^{-1}\|K\|_2^2 f(x) + o(h).$$

Este resultado nos dice que si elegimos nh grandes podremos dar un estimador con varianza mas pequeña y análogamente si $\|K\|_2^2 = \int K^2$ es pequeña, es decir el núcleo es mas bien chato.

De esta forma hemos calculado el error cuadrático medio del estimador (ECM) para cada x,

$$ECM(\widehat{f}(x)) = h^4\frac{(f''(x))^2}{4}\mu_2^2(K) + o(h^4) + \frac{1}{nh}\|K\|^2 f(x) + o(\frac{1}{nh})$$

Como conclusión nuevamente tenemos un compromiso entre sesgo y varianza. Pues h pequeños derivarán en estimadores con menor sesgo mientras que al aumenter el ancho de banda lograremos disminuir la varianza. En la siguiente figura podemos apreciar este efecto.

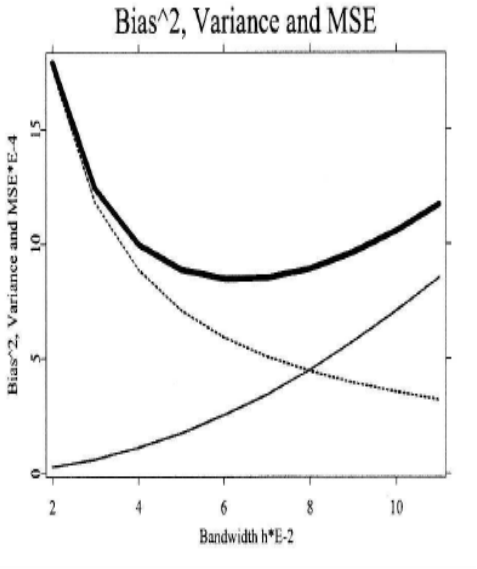

Figura 5: Sesgo al cuadrado (linea sólida); varianza (linea punteada) y error cuadrático medio (linea solida gruesa).

Un corolario que se desprende de lo analizado anteriormente es la consistencia debil del estimador. Hemos probado que si $h \to 0$ y $nh \to \infty$ tenemos que $\widetilde{f}(x) \xrightarrow{p} f(x)$ para cada x. Se pueden obtener resultados maś fuertes de consistencia, como la consistencia uniforme en x, pero escapan los objetivos de estas notas.

2.3 Selección del Núcleo y la ventana.

La Figura 5 de la sección anterior nos muestra el compromiso entre el sesgo y la varianza reflejando así la importancia de la selección apropiada del ancho de banda. Una elección natural de la ventana sería considerar aquella que minimice el ECM. Recordemos que

$$ECM(\widehat{f}(x)) \quad = \quad h^4 \frac{(f''(x))^2}{4} \mu_2^2(K) + \frac{1}{nh} \|K\|^2 f(x) + o(\frac{1}{nh}) + o(h^4)$$

luego si $h \to 0$ y $nh \to \infty$, podemos despreciar los terminos de menor orden y buscaremos el valor de h que hace mínimo

$$ECM(\widehat{f}(x)) \approx h^4 \frac{(f''(x))^2}{4} \mu_2^2(K) + \frac{1}{nh} \|K\|^2 f(x)$$

simplemente derivando respecto de h e igualando a 0 obtenemos que

$$h_{opt}(x) = \left(\frac{\|K\|^2 f(x)}{(f''(x))^2 \mu_2^2(K) n} \right)^{1/5} = \left(\frac{\|K\|^2 f(x)}{(f''(x))^2 \mu_2^2(K)} \right)^{1/5} n^{-1/5}$$

De esta manera, hemos encontrado la ventana óptima que depende de cantidades desconocidas como $f(x)$ y $f''(x)$ y de constantes que son funciones del núcleo. Ademas de no poder calcularla en la práctica la ventana obtenida es local, es decir, depende del punto x donde se esta estimando. Alguno de estos inconvenientes pueden ser solucionados considerando el error cuadrático medio integrado (MISE o ECMI). Más precisamente en lugar de considerar $ECM(\widehat{f}(x))$ estudiaremos $\int ECM(\widehat{f}(x))dx$

$$ECMI(\widehat{f}(x)) \quad = \quad h^4 \frac{\int (f''(x))^2 dx}{4} \mu_2^2(K) + \frac{1}{nh} \|K\|^2 + o(\frac{1}{nh}) + o(h^4)$$

o simplemente despreciando los terminos pequeños, el error cuadrático medio integrado asintótico ($AMISE$ o $ECMIA$)

$$ECMIA(\widehat{f}(x)) \quad = \quad h^4 \frac{\int (f''(x))^2 dx}{4} \mu_2^2(K) + \frac{1}{nh} \|K\|^2.$$

Por lo tanto análogamente a lo desarrollado anteriormente obtenemos la siquiente ventana optima

$$h_{opt} = \left(\frac{\|K\|^2}{\|f\|^2 \mu_2^2(K)} \right)^{1/5} n^{-1/5}$$

que ya no depende del punto x donde se este estimando y no depende del valor de $f(x)$ pero que aun depende de $\|f''\|$ que es desconocido. Antes de dar una posible solución a este problema notemos que si calculamos el valor del $ECMIA$ en el valor de h_{opt} obtenido, tenemos

$$ECMIA(\widehat{f}(x))(h_{opt}) = \frac{5}{4} \left(\|f''\| \mu_2(K) \right)^{2/5} \|K\|^{8/5} n^{-4/5}$$

Obviamente al aumentar el tamaño de muestra obtendremos un $ECMIA$ mas pequeño. Pero lo interesante a remarcar es que si hubieramos realizado el mismo análisis del $ECMIA$ en el caso del histograma la velocidad óptima hibiese sido de $n^{2/3}$ en lugar de $n^{-4/5}$, dejando así otro argumento a favor de la superioridad del estimador de densidad basado en núcleos.

Para seleccionar el parámetro de suavizado existen varias alternativas, aquí presentaremos dos métodos Convalidación cruzada y un método Plug–in.

El método plug–in es una alternativa usual de estimación que consiste en reemplazar parámetros desconocidos de una expresión por estimadores. Por lo tanto, a fin de obtener un estimador para la ventana optima bastará con dar un estimador de $\|f''\|$ ya que las constantes que dependen del núcleo pueden ser calculadas una ves fijado este.

Silverman dio una propueta bajo el supuesto de normalidad de f. En este caso si f es normal se puede calcular $\|f''\|^2 = \sigma^{-5} \frac{3}{8\sqrt{\pi}}$ y luego estimar σ.

De esta forma en el caso en que el núcleo también sea gaussiano tendriamos que $\widehat{h}_{opt} = 1.06\widehat{\sigma}n^{-1/5}$. Claramente uno puede objetar la regla de Silverman, pues asumir normalidad en el contexto no paramétrico es bastante desacertado siendo que f es deconocido. Pero en la práctica se ha visto que para densidades unimodales y cerca de la simetría, la ventana estimada provee resultados razonables. Otra alternativa que se denomina método Plug-in refinado es considerar un estimador no paramétrico de la derivada segunda. Este podría calcularse derivando dos veces el estimador de la densidad con una ventana preliminar, que podría ser la presentada anteriormente.

Los métodos de convalidación cruzada no realizan ningún supuesto sobre la familia a la que pertenece f. La idea es considerar una medida entre f y su estimador $\widetilde{f}$, en este caso tomaremos el error cuadrático integrado ECI

$$\begin{aligned} ECI(h) &= \int (\widetilde{f}(x) - f(x))^2 dx \\ &= \int \widetilde{f}^2(x)dx - 2\int \widetilde{f}(x)f(x)dx + \int f^2(x)dx \\ &= \int \widetilde{f}^2(x)dx - 2E(\widetilde{f}(x)) + \int f^2(x)dx \end{aligned}$$

Notemos que $\int f^2(x)dx$ no depende de h; $\int \widetilde{f}^2(x)dx$ es calculable con los datos y $E(\widetilde{f}(x))$ puede ser estimada por $E(\widetilde{\widetilde{f}(x)}) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{f}_{-i}(X_i)$ donde $\widetilde{f}_{-i}(X_i)$ corresponde al estimador de la densidad calculado sin la observación i evaluado en la X_i, es decir,

$$\widetilde{f}_{-i}(x) = \frac{1}{h(n-1)} \sum_{j=1;i\neq j}^{n} K((x - X_j)/h).$$

De esta forma, podemos estimar la ventana óptima como $\widehat{h}_{cv} = argmin_h CV(h)$ donde

$$CV(h) = \int \widetilde{f}^2(x)dx - \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1;i\neq j}^{n} K((X_i - X_j)/h).$$

Para facilitar el calculo se puede probar que

$$\int \widetilde{f}^2(x)dx = \frac{1}{n^2 h} \sum_i \sum_j K * K((X_i - X_j)/h)$$

donde $K * K(u) = \int K(u - v)K(v)dv$.

Podríamos decir que no hay un método que sea mejor a los demás. Lo mejor a la hora de poner en práctica el cálculo del estimador será usar distintos métodos y comparar las estimaciones obtenidas.

En cuanto a la selección del núcleo en general se considerá un núcleos simétrico y unimodal siendo fáciles de interpretar.

En cuanto a la elección del núcleo, recordemos que el error cuadrático medio integrado asintótico calculado en la ventana óptima es

$$ECMIA(\widehat{f}(x))(h_{opt}) = \frac{5}{4} \left(\|f''\|\mu_2(K)\right)^{2/5} \|K\|^{8/5} n^{-4/5}$$

luego, si $T(K) = (\mu_2(K))^2 \|K\|^8)^{1/5}$ podemos buscar el núcleo K que minimiza $T(K)$. Epanechnikov mostró que entre todos los núcleos de soporte compacto no negativos

$$K(u) = \frac{3}{4}\left(\frac{1}{15^{1/5}}\right)\left(1 - \left(\frac{u}{15^{1/5}}\right)^2\right) I(|u| \leq 15^{1/5}).$$

Como vemos en la siguiente tabla, si bien el núcleo de Epanechnikov es óptimo, los cocientes son muy cercanos a 1. El que más difiere es el Uniforme que da un incremento del 6%, por lo que en la práctica la elección del núcleo no es tan importante para la eficiencia del estimador.

Nucleo	T(K)	$T(K)/T(K_{epan})$
Uniform	0.3701	1.0602
Triangle	0.3531	1.0114
Epanechnikov	0.3491	1.0000
Quartic	0.3507	1.0049
Triweight	0.3699	1.0595
Gaussian	0.3633	1.0408
Cosine	0.3494	1.0004

Tabla 2: Comparación de los núcleos.

2.4 Extensión al caso multivariado

En ciertas situaciónes uno puede estar interesado en estimar la densidad en un contexto multivariado mas que en una dimensión. Por lo tanto resulta interesante extender la propuesta anterior cuando trabajamos con mayor dimensión. Consideremos el caso de una densidad sobre R^d. Obervemos una muestra de tamaño n de vectores aleatorios $\mathbf{X}_i$ donde

$$\mathbf{X}_i = \begin{pmatrix} X_{i1} \\ \vdots \\ X_{id} \end{pmatrix} \qquad i = 1, \ldots, n.$$

El objetivo será estimar la densidad $f(\mathbf{x}) = f(x_1, \ldots, x_d)$. La extensión natural de la propuesta realizada anteriormente sería considerar

$$\begin{aligned} \widehat{f}(\mathbf{x}) &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^d} K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^d} K\left(\frac{x_1 - X_{i1}}{h}, \ldots, \frac{x_d - X_{id}}{h}\right) \end{aligned}$$

donde K es un núcleo multivariado es decir $K : R^d \to R$

En este caso se ha elegido utilizar la misma ventana h en todas las componentes, pero no es necesario. Podríamos tomar una ventana distinta en cada componente. Si tomamos $h = (h_1, \ldots, h_d)'$ tendríamos

$$\widehat{f}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_1 \ldots h_d} K\left(\frac{x_1 - X_{i1}}{h_1}, \ldots, \frac{x_d - X_{id}}{h_d}\right).$$

El núcleo multivariado puede der elegido como un núcleo multiplicativo es decir $K(\mathbf{u}) = K_1(u_1)\dots K_d(u_d)$ donde K_j $1 \leq j \leq d$ es un núcleo univariado. Luego

$$\widehat{f}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_1 \dots h_d} K_1\left(\frac{x_1 - X_{i1}}{h_1}\right), \dots, K_d\left(\frac{x_d - X_{id}}{h_d}\right).$$

Otra alternativa es usar un verdadero núcleo multivariado, por ejemplo el núcleo multivariado de Epanechnikov

$$K(\mathbf{u}) \propto (1 - \mathbf{u}'\mathbf{u})I(\mathbf{u}'\mathbf{u} \leq 1).$$

Los núcleos multivariados también se pueden obtener a partir de los núcleos univariados de la siguiente forma.

$$K(\mathbf{u}) \propto K(\|\mathbf{u}\|).$$

Un enfoque más general propone considerar una matriz H (no singular)

$$\widehat{f}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{det(H)} K\left(H^{-1}(X - \mathbf{X}_i)\right).$$

es decir el caso de todas las ventanas iguales correspondería a $H = hI_d$. Para este estimador se pueden calcular al igual que en el caso de $d = 1$ el sesgo y la varianza obteniendo una expresión para el $ECMIA$ que permite calcular la ventana óptima. En este caso $h_{opt} \sim n^{-1/(4+d)}$ y $ECMIA(h_{opt}) \sim n^{-4/(d+4)}$. Como vemos la tasa de convergencia disminuye enormemente si la comparamos con la del caso $d = 1$. Esto es lo que se conoce como la maldición de la dimensión. Por esta razón estos estimadores se utilizan para dimensiones muy bajas, $d = 2$ o 3.

$n^{-4/(4+d)}$	$d=1$	$d=2$	$d=3$	$d=5$	$d=10$
$n = 100$	0.025	0.046	0.072	0.129	0.268
$n = 1000$	0.004	0.010	0.019	0.046	0.139
$n = 100'000$	$1.0\,10^{-4}$	$4.6\,10^{-4}$	$13.9\,10^{-4}$	0.006	0.037

Tabla 3: Comparación de la dimensión y el error.

Los criterios de selección de la ventana introducidos anteriormente pueden ser extendidos en el caso multivariado.

2.5 Vecinos Más Cercanos

Si aplicamos este estimador a datos procedentes de distribuciones con colas pesadas, con una ventana suficientemente pequeña para estimar bien la parte central de la distribución no lograremos estimar correctamente las colas de la distribución. Mientras que con un valor de ventana grande para la correcta estimación de las colas no podremos ver los detalles que ocurren en la parte principal de la distribución. Para superar estos defectos, se propuso un estimador conceptualmente similar al estudiado por Rosenblatt pero cuyos entornos no son fijos sino que se adaptan al punto en el cual se está estimando. Estos estimadores se conocen con el nombre de estimadores por vecinos más cercanos con núcleos.

Como mencionamos anteriormente el problema de escoger el valor de ancho de banda es no trivial. Pues un h demasiado pequeño tiene como efecto que la varianza del estimador aumente demasiado ya que son pocas las observaciones considerados es cada punto. Mientras que un valor demasiado alto da resultados con un alto sesgo debido a que se promedian demasiadas observaciones que no logran captar la tendencia o forma de la curva a estimar. A este compromiso en la elección del valor de h se le denomina compromiso sesgo-varianza.

Una manera de dar una solución a este problema es considerar entornos variables. Es decir, en lugar de fijar un ancho de ventana y a partir de los valores muestrales que caen en él estimar la función de densidad, la idea sería construir en cada punto donde deseamos estimar entornos que contengan una cantidad fija de observaciones. Más precisamente, sea $d(x, y) = |x - y|$ la distancia entre dos puntos x, y. Consideremos para cada valor de x las distancias $d(x, X_i)$ para $1 \leq i \leq n$ y llamemos $d_i(x)$ a las distancias ordenadas, es decir $d_i(x) = (d(x, X_i))^{(i)}$, el estadístico de orden i de las distancias al punto x.

Definimos el estimador de densidad por el método del *k-ésimo vecinos más cercanos* como

$$\widehat{f}(x) = \frac{k}{2nd_k(x)}. \tag{3}$$

Con el fin de comprender un poco mejor esta definición, recordemos que, por lo visto en (1), para una muestra de tamańo n, uno esperaría aproximadamente $2hnf(x)$ observaciones dentro del intervalo $[x - h, x + h]$ para cada $h > 0$. Por otro lado exactamente k observaciones caerán dentro del intervalo $[x - d_k(x), x + d_k(x)]$, entonces es razonable esperar que k sea aproximadamente como $2d_k(x)nf(x)$. Y de aquí obtenemos el estimador de k vecinos más cercanos propuesto en (3).

Mientras que (1) está basado en un número de observaciones que yacen en un intervalo de longitud fija centrado en el punto de interés, el estimador de *k-ésimo vecinos más cercanos* es inversamente proporcional al tamaño del intervalo que contiene un número k de observaciones dado. Es posible generalizar el estimador de *k-ésimo vecinos más cercanos* combinando (1) con (3) obteniendo así el siguente estimador

$$\widehat{f}(x) = \frac{1}{nd_k(x)} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{d_k(x)}\Big) \tag{4}$$

donde K es una función núcleo con las mismas propiedades que las definidas anteriormente, $k = k_n$ es una sucesión tal que $k_n \to \infty$ si $n \to \infty$ y $d_k(x)$ es la distancia entre x y el k-ésimo vecino más cercano.

2.6 Intervalos y Bandas de confianza

Para obtener intervalos de confianza es necesario calcular la distribución del estimador. Hasta el momento se desconoce la distribución exacta pero si es posible obtener el comportamiento asintótico del estimador. Bajo ciertas condiciones de regularidad

1. $h_n \to 0$
2. $nh_n \to \infty$
3. x tiene densidad f continua en x y dos veces diferenciable

4. $K : I\!R \to I\!R$ es acotado, $\int K = 1$ y $\int u^2 K(u) > 0$ y con soporte compacto.

se puede probar si $h_n = cn^{-1/5}$

$$\sqrt{nh}(\widetilde{f}(x) - f(x)) \xrightarrow{\mathcal{D}} N(\frac{c^{5/2}}{2} f''(x) m_2(K), f(x)\|K\|^2).$$

Luego resulta el siguiente intervalo de confianza de nivel aproximado $1 - \alpha$

$$\left[\widetilde{f}(x) - \frac{h^2}{2} f''(x) m_2(K) - z_{\alpha/2}\sqrt{\frac{f(x)\|K\|^2}{nh}}, \widetilde{f}(x) - \frac{h^2}{2} f''(x) m_2(K) + z_{\alpha/2}\sqrt{\frac{f(x)\|K\|^2}{nh}}\right]$$

si h es pequeña se puede despreciar el término que involucra a la derivada segunda y utilizar el siguiente intervalo

$$\left[\widetilde{f}(x) - z_{\alpha/2}\sqrt{\frac{\widetilde{f}(x)\|K\|^2}{nh}}, \widetilde{f}(x) + z_{\alpha/2}\sqrt{\frac{\widetilde{f}(x)\|K\|^2}{nh}}\right]$$

de lo contrario podemos estimar la derivada segunda, derivando un es de núcleos usando una ventana g.

Es importante notar que este intervalo es sólo para $f(x)$ y no para toda la densidad. Para deducir bandas de confianza para toda la función es necesario emplear otras técnicas. Bickel y Rosenblatt (1973) probaron el siguiente resultado: sea f una función de densidad definida sobre el $(0, 1)$, $h_n = n^{-\delta} \in (1/5, 1/2)$, entonces para todo $x \in (0, 1)$.

$$\lim_{n\to\infty} P\left(\widetilde{f}(x) - \sqrt{\frac{\widetilde{f}(x)\|K\|^2}{nh}}\left\{\frac{z}{2\delta\log n} + d_n\right\}^{-1/2} \le f(x) \le \widetilde{f}(x) + \sqrt{\frac{\widetilde{f}(x)\|K\|^2}{nh}}\left\{\frac{z}{2\delta\log n} + d_n\right\}^{-1/2}\right) = \exp\{-2\exp\{-z\}\}$$

donde $d_n = (2\delta\log n)^{1/2} + (2\delta\log n)^{-1/2}\log\left\{\frac{\|K'\|_2}{2\pi\|K\|_2}\right\}$. Entonces para hallar una banda de confianza de nivel α bastará encontrar el valor de z que satisface $\exp(-2\exp(-z)) = 1 - \alpha$. Por ejemplo si $\alpha = 0.05$ luego $z \approx 3.663$

El siguiente ejemplo corresponde a datos de la ganancia promedio en horas de 534 trabajadores elegidos al azar en Estados Unidos en durante mayo de 1985.

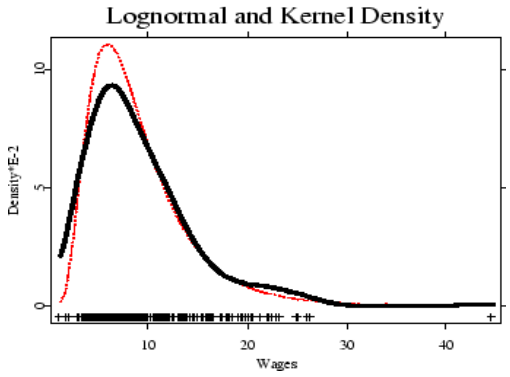

Figura 6: En rojo estimador paramétrico de la lognormal en negro estimador no paramétrico (núcleo cuadrático, $h = 5$).

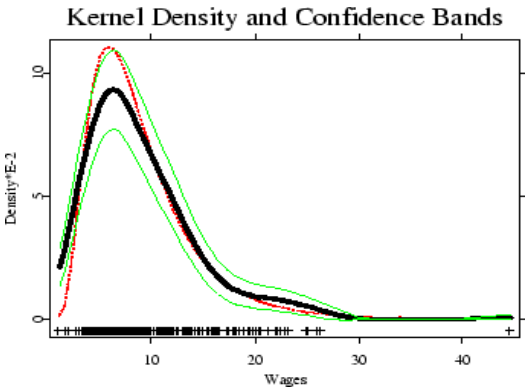

Figura 7: En rojo estimador paramétrico de la lognormal en negro estimador no paramétrico (núcleo cuadrático, $h = 5$) y en verde los intervalos de confianza.

Podemos ver que el estimador paramétrico entorno a la moda se encuentra fuera de la banda de confianza por lo tanto rechazaríamos la hippótesis de que la verdadera distribución es la densidad lognormal. Sin embargo, la estimación paramétrica parece capturar bastante bien la forma de la distribución. Los test o intervalos no paramétricos suelen tener pérdida de eficiencia pero es posible encontrar test noparametricos que tengan mejor velocidad de convergencia.

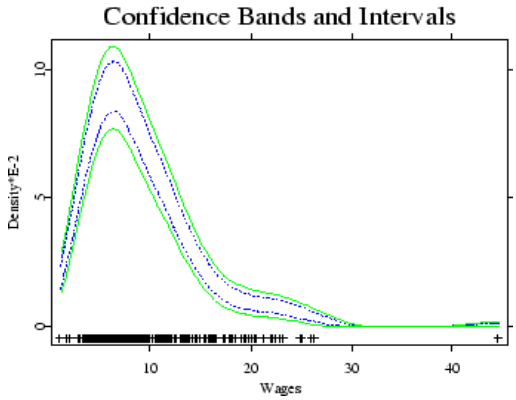

Figura 8: Bandas e intervalos de confianza.

2.7 Paquetes y librerías de R

Paquete base: "density(x) ". Las opciones principales son

- **bw**

  bw.nrd0 implementa la ventana para un ncleo Gaussiano.

  bw.nrd es una variante introducida por de Scott (1992) usando un factor de corrección factor 1.06.

  bw.ucv y bw.bcv implementan la unbiased y la biased cross-validation (minimiza la AMISE en lugar del ISE, usando un plug-in de la derivada)

- **kernel** gaussian, epanechnikov, rectangular, triangular, biweight, cosine, optcosine

  Las instrucciones para dibujar serían

```
plot(density(precip, n = 1000))
rug(precip)  #esta instruccin dibuja la muestra en el eje x.
lines(density(precip, bw="nrd"), col = 2)
lines(density(precip, bw="ucv"), col = 3)
```

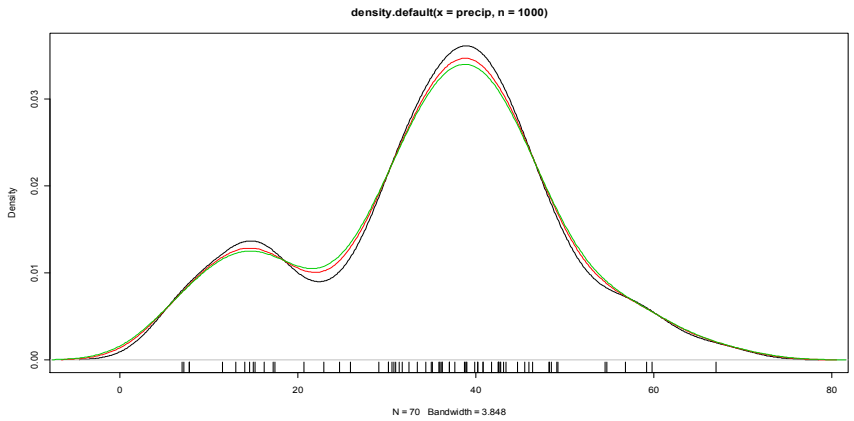

Figura 9:

Paquete "KernSmooth "

- bkde : estimador de la densidad con núcleos normal, epanech, box, biweight, triweight.
- bkde2: estimador de la densidad con núcleo normal en $2D$.
- bkfe: estima la derivada drv con núcleo normal de una densidad.
- dpih: elección de h para un histograma.
- dpik: : elección de h por plug-in para estimador de núcleo.

Paquete "sm"

- sm.density: ajuste de una densidad en dimensión 1, 2 o 3.
  sm.density(x, h, model = "none", weights = NA, group=NA, ...)
  Usan núcleo normal y si el parámetro h no está elige una ventana óptima para normal.

```
y <- rnorm(50)
sm.density(y)

y <- cbind(rnorm(50), rnorm(50))
sm.density(y, display = "image")

 y <- cbind(rnorm(50), rnorm(50), rnorm(50))
 sm.density(y)
```

2.8 Ejercicios

1. Consideremos el histograma como estimador de la densidad. Mas precisamente tomemos una partición de la recta $I\!R = \bigcup_{j=-\infty}^{\infty} B_j$ donde $B_j = [(j-1)h, jh)$. Si deseamos estimar f en el punto x, sea j_0 tal que $x \in B_{j_0}$ luego $\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} I(X_i \in B_{j_0})$.

Probar que $Sesgo(\widehat{f}(x)) = ((j_0 - \frac{1}{2})h - x)f'(j_0 - \frac{1}{2})h-) + o(h)$.

Al igual que con el estimador de núcleos se puede calcular el error cuadrático medio integrado asintótico del histograma

$$ECMIA(\widehat{f}) = (nh)^{-1} + \frac{h^2}{12}\|f'\|^2$$

A partir del ECMIA calcule la ventana óptima de histograma.

2. (a) Generar 100 datos que provengan de una densidad $f(x)$ mezcla de normales $0.4N(-1,1) + 0.6N(2,1)$, es decir

$$f(x) = 0.4\phi_1(x) + 0.6\phi_2(x),$$

donde ϕ_1 es la densidad de una normal con media -1 y varianza 1 y ϕ_2 es la densidad de una normal con media 2 y varianza 1.

   (b) Graficar en R dicha densidad.

   (c) Calcular la ventana óptima del histograma para el caso de la densidad del punto a)

   (d) Realizar histogramas con ventanas $h = 0.2$, 1, 2, 10 y con la ventana obtenida en c). Comparar. Una opción sencilla para realizar histogramas indicando el h y el x_0 es usando la instrucción **truehist**de la library(MASS).

3. Simular 500 datos provenientes de una distribución $N(\mu, \sigma^2)$.

   (a) Calcular la ventana óptima h_0 para el AMISE para los datos generados. Realice el histograma correspondiente.

   (b) Adapte la ventana de a) para el caso en que debe estimar los parámetros de la distribución. Realice el histograma correspondiente y compare.

   (c) Repita la generación de los 500 datos normales varias veces, realice los histogramas de a) y b) y vea cómo se comportan los resultados a lo largo de las repeticiones.

4. Implementar en R el estimador de densidad utilizando vecinos más cercanos.

5. Implementar en R el método de selección de ventana de validación cruzada.

6. Los datos que se hallan en el archivo buffalo.txt, corresponden a la mediciones de cantidad de nieve caída (en pulgadas) en Buffalo en los inviernos de 1910/1911 a 1972/1973. Estudiar el ajuste de la función de densidad basado en los distintos métodos introducidos. Compare los resultados con la estimación paramétrica correspondiente a alguna familia que considere apropiada..

7. Simular 500 datos provenientes de una distribución $N(2,1)$.

   (a) Estimar la densidad normal a partir de los datos.

   (b) Graficar en un mismo plot la verdadera densidad, la densidad estimada por núcleos y la normal estimada. Comparar.

   (c) Para los dos estimadores calcular y graficar el error relativo como

$$ER(\widehat{f}(x)) = \frac{\widehat{f}(x)}{f(x)} - 1$$

Comparar los gráficos obtenidos.

3 Regresión no paramétrica.

3.1 Regresión No Paramétrica: Modelos No Paramétricos.

Las curva de regresión describe la relación entre dos variables, una variable explicativa X y una variable respuesta Y. Una vez observado X, el valor medio de Y está dada por la función de regresión y en muchas situaciones es de gran interés tener algo de conocimiento sobre esta relación.

Dada una muestra (X_i, Y_i) $i = 1, \ldots, n$ el objetivo es estimar la esperanza condicional es decir $m(X_i) = E(Y_i|X_i)$ sin realizar ningun supuesto sobre la función m como lo puede ser la linealidad, la monotonía, una relación cuadrática, etc. Esta relación también es comunmente modelada como

$$Y_i = m(X_i) + \varepsilon_i \qquad i = 1, \ldots, n$$

donde ε_i son variables aleatorias independientes con media 0 que denota la variación de Y_i alreededor del $m(X_i)$.

Antes de continuar recordemos como calcular la esperanza condicional en el caso de densidad conjunta. Sean X e Y dos variables aleatorias con densidad conjunta $f(x, y)$. La esperanza condicional de Y dado $\mathbf{X} = x$ puede calcularse como

$$\begin{aligned} E(Y|X = x) &= \int y f(y|x) dy = \int y \frac{f(x,y)}{f_X(x)} \\ &= = \frac{r(x)}{f_X(x)} = m(x) \end{aligned}$$

Un ejemplo sencillo si consideramos $f(x, y) = x + y$ si $0 < x < 1$ e $0 < y < 1$, es fácil calcular la densidad marginal $f_X(x) = x + \frac{1}{2}$, si $0 < x < 1$. Luego

$$E(Y|X = x) = \int y \frac{x+y}{x+\frac{1}{2}} dy = \frac{\frac{1}{2}x + \frac{1}{2}}{x + \frac{1}{2}} = m(x).$$

Como se puede ver en este ejemplo la estructura de dependencia dada por la esperanza condicional no es lineal. El objetivo de esta sección es proveer mecanismos de estimación para la función m con el menor número de supuestos.

3.2 Estimación por Núcleos.

En primer lugar estudiaremos el estimador propuesto por Nadaraya–Watson (1964). Notemos que por lo visto anteriormente si (X, Y) tiene densidad conjunta

$$m(x) = \int y \frac{f(x,y)}{f_X(x)}.$$

Por lo tanto como $f(x, y)$ y $f_X(x)$ son desconocidas una idea sencilla sería hacer un plug–in, es decir reemplazar estas funciones de densidad por estimadores que pueden ser por ejemplo los estudiados en el capítulo anterior. De esta manera podemos considerar

$$\widehat{f}_{h,g}(x, y) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right) \frac{1}{g} K\left(\frac{y - Y_i}{g}\right)$$

y

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

Calculemos entonces $\int y\widehat{f}_{h,g}(x,y)dy$,

$$\begin{aligned}
\int y\widehat{f}_{h,g}(x,y)dy &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - X_i}{h}\right)\int \frac{y}{g}K\left(\frac{y - Y_i}{g}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - X_i}{h}\right)\int (sg + Y_i)K(s) \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - X_i}{h}\right)Y_i
\end{aligned}$$

Entonces un estimador de m queda definido como

$$\widehat{m}_h(x) = \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-X_i}{h}\right)Y_i}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-X_i}{h}\right)} = \frac{\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)Y_i}{\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)}.$$

Veamos como funciona el estimador en un ejemplo. El siguiente gráfico muestra un conjunto de datos simulados.

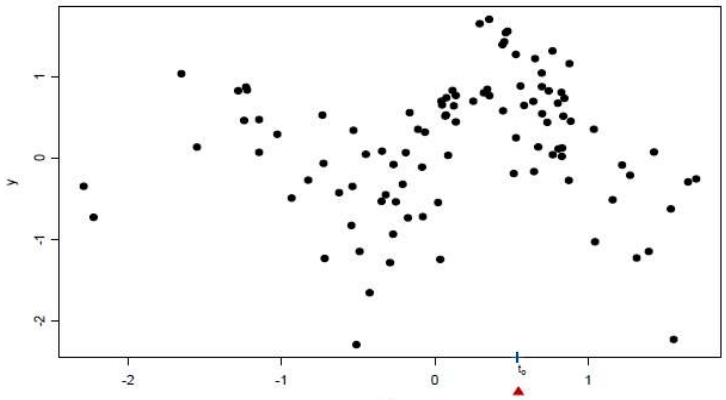

Figura 10: Ejemplo de datos simulados.

En el punto t_0 estimaremos $m(t_0) = E(y|t = t_0)$, el estimador de Nadaraya–Watson podemos interpretarlo como un promedio local, es decir

$$\widehat{m}_h(x) = \sum_{i=1}^{n} W_{ni}(t_0)Y_i$$

donde $W_{ni}(t_0) = \frac{K\left(\frac{t_0 - X_i}{h}\right)}{\sum_{j=1}^{n}K\left(\frac{t_0 - X_j}{h}\right)}$ y $\sum_{i=1}^{n} W_{ni}(t_0) = 1$

Mas precisamente, el estimador actúa promediando localmente las observaciónes Y_i con pesos que dependen de la cercanía de las variables X_i al punto t_0 donde queremos estimar.

Si consideremos el siguiente núcleo uniforme $K(u) = \frac{1}{2} I_{[-1,1]}(u)$ y una ventana $h = 0.3$ luego miraremos el entorno $(t_0 - 0.3, t_0 + 0.3)$ y promediaremos las obervaciones Y_i tales que su respectiva X_i pertenece a dicho entorno.

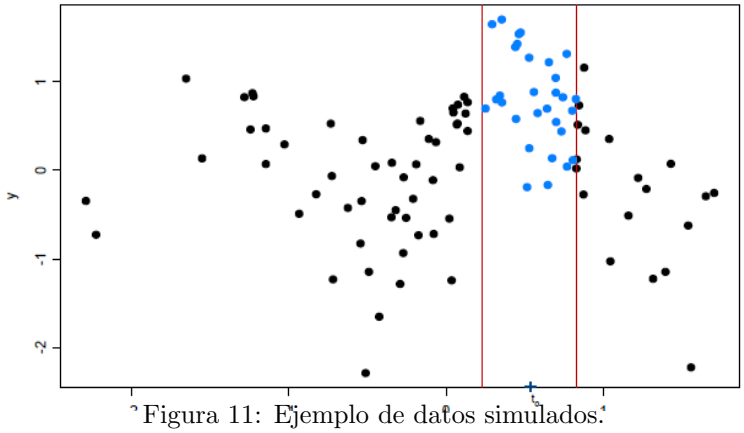

Figura 11: Ejemplo de datos simulados.

Al igual que en el caso de los estimadores de densidad, el papel de la ventana es muy importante en el proceso de estimación y como podemos ver en la siguiente figura determina el grado de suavidad de la función estimada.

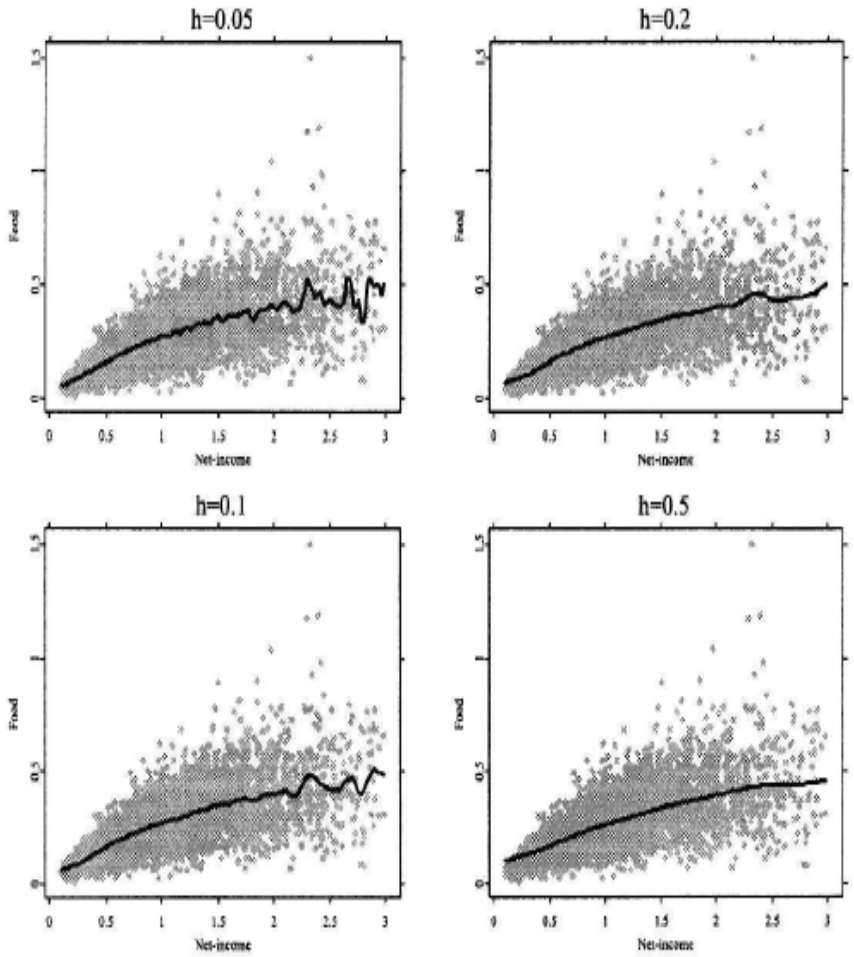

Figura 12: Promedio de ingresos vs promedio de gasto en alimentos en Inglaterra en 1973 $n = 7125$.

En general se puede ver que con ventanas muy pequeñas el estimador tiende a interporlar los datos en los puntos de la muestra, mientras que ventanas muy grandes tienden a estimadores constantes alrededeor de $\bar{Y}$.

Otra forma de pensar el estimador de Nadaraya–Watson es la siguiente. Si consideramos la siguiente función

$$M(\theta) = \sum_{i=1}^{n} W_{ni}(x)(Y_i - \theta)^2$$

y buscamos para cada x, θ que minimiza M es fácil ver que

$$argmin_\theta M(\theta) = \sum_{i=1}^{n} W_{ni}(x)(Y_i - \widehat{m}(x))^2.$$

Es decir, el estimado no paramétrico de regresión es un estimador de mínimos cuadrados ponderados, donde los pesos son calculados de manera local en el punto x donde queremos estimar.

Bajo ciertas hipótesis se puede calcular una expresión para el error cuadrático medio

$$ECM(\widehat{m}(x)) = \frac{1}{nh}\frac{\sigma^2(x)}{f(x)}\|K\|^2 + \frac{h^4}{4}\left[m''(x) + 2\frac{m'(x)f'(x)}{f(x)}\right]^2 \mu_2^2(K) + o((nh)^{-1}) + o(h^4)$$

donde $\mu_2^2(K) = \int u^2 K(u)du$ y $\sigma^2(x) = \text{var}(Y|X)$. Por lo tanto ver que si $h \to 0$ y $nh \to \infty$ $\widehat{m}(x) \xrightarrow{p} m(x)$ cuando $n \to \infty$. Además, al igual que en el caso de estimación de densidad podemos calcular la ventana óptima minimizando el $ECMA$ en función de h. En este caso $h_{opt} \approx n^{-1/5}$ y reemplazando la ventana obtenida en la expresión del $ECMA$ tenemos que $ECMA(h_{opt}) = O(n^{-4/5})$. Como era de esperar el estimador noparamétrico tiene velocidad de convergencia más lenta que los estimadores de regresión lineal y tiene el mismo orden que el estimador no paramétrico de la densidad.

3.3 Vecinos Más Cercanos.

Los estimadores de núcleos definidos anteriormente pueden ser vistos como un promedio ponderado de la variable de respuesta en un intervalo fijo determinado por h alrededor de x.

El estimador de k vecinos más cercanos también puede ser visto como un promedio ponderado de la respuesta pero en un entorno de ancho variable: los valores que intervienen ahora en el promedio corresponden a las k observaciones cuyos valores de X son los k más cercanos al punto de interés x.

$$\widehat{m}_k(x) = \frac{1}{n}\sum_{i=1}^{n} W_{ki}(x)Y_i$$

donde $W_{ki}(x) = \frac{n}{k}$ si X_i es una de las k observaciones más cercanas a x y 0 en caso contrario. El paramétro k esta relaciónado con la suavidad de la estimación aumentar k llevará a un estimador más suave. Cuando x se encuentra en una región rala, los puntos que caen en el intervalo pueden estar lejos de x y dar como resultados estimadores con alto sesgo.

El estimador anterior, puede pensarse como un estimador de núcleos con núcleo uniforme si llamamos $d_k(x)$ a la mayor distancia entre x y su k-ésimo vecino más cercano podemos escribir al estimador de la siguiente manera

$$\widehat{m}_k(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x-X_i}{d_k(x)}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{d_k(x)}\right)}.$$

y de esta manera podría generalizarse usando otro núcleo y no sólo el uniforme, esto da lugar a los estimadores denominados k vecino con núcleos.

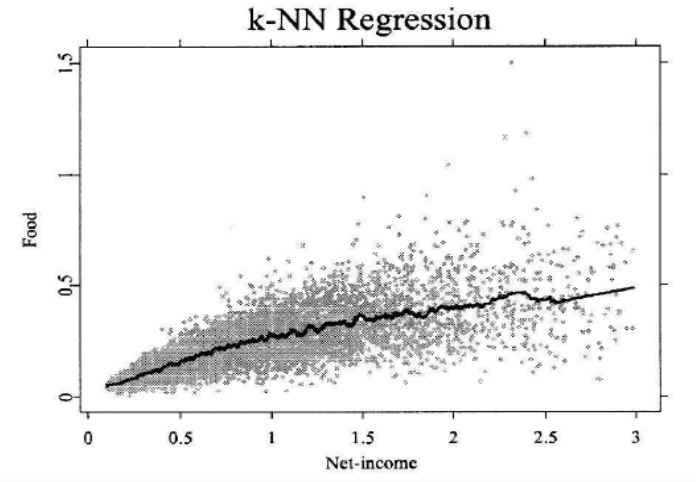

Figura 13: Promedio de ingresos vs promedio de gasto en alimentos en Inglaterra en 1973 $n = 7125$ usando k vecinos $k = 101$.

Se pueden obtener expresiones para el sesgo y la varianza de los estimadores. El siguiente cuadro compara ambas expresiones.

	núcleo	k- NN
sesgo	$h^2 \frac{(m''f+2m'f')(x)}{2f(x)} \mu_2(K)$	$\left(\frac{k}{n}\right)^2 \frac{(m''f+2m'f')(x)}{8f^3(x)} \mu_2(K)$
varianza	$\frac{\sigma^2(x)}{nhf(x)} \|\|K\|\|_2^2$	$\frac{2\sigma^2(x)}{k} \|\|K\|\|_2^2$

Tabla 4: Comparación de ordenes de estimadores.

Notemos que si $\frac{h^2}{2f(x)} = \left(\frac{k}{n}\right)^2 \frac{1}{8f^3(x)}$, es decir si $k = 2nhf(x)$ los sesgos coinciden, aunque esto depende de la distribución marginal de x que es desconocida, ademas con esta misma restricción también coincidirían las varianzas y por lo tanto los estimadores serían equivalentes. Lo importante a resaltar en este punto es que el número de vecinos debe tener el mismo orden que nh .

3.4 Polinomios Locales.

El método de polinomios locales consiste como su nombre lo indica en aproximar a la funciones de regresión localmente en cada punto x por un polinomio cuyo grado es determinado

por el usuario. Si consideramos el desarrollo de Taylor de la función de regresión

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

Esto sugiere una regresión polinomial local de la siguiente manera,

$$\min_{\beta} \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1(X_i - x) - \ldots - \beta_p(X_i - x)]^2 K_h(x - X_i)$$

donde β es el vector de coeficiente $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$, es importante notar que $\beta = \beta(x)$ es decir depende fuertemente del punto donde se esta estimando. Recordemos que el estimador de núcleos podía verse también como un problema de mínimos cuadrados, en este caso sería equivalente a ajustar polinomios locales con $p = 0$.

Comparando entonces el desarrollo de Taylor con la definición de la regresión polinomial, el estimador de la función de regresión sería entonces $\widehat{m}_{p,h} = \widehat{\boldsymbol{\beta}}_0(x)$ y los demás coeficientes sirven para estimar las primeras p derivadas de la función de regresión de la siguiente manera, la derivada $r-$ésima

$$\widehat{m}^{(r)}_{p,h}(x) = \nu! \widehat{\boldsymbol{\beta}}_r(x).$$

En la práctica el problema de como calcular los coeficientes puede resolverse como mínimos cuadrados pesados obteniendo una expresión explícita para los coeficientes de $\widehat{\boldsymbol{\beta}}(x)$.

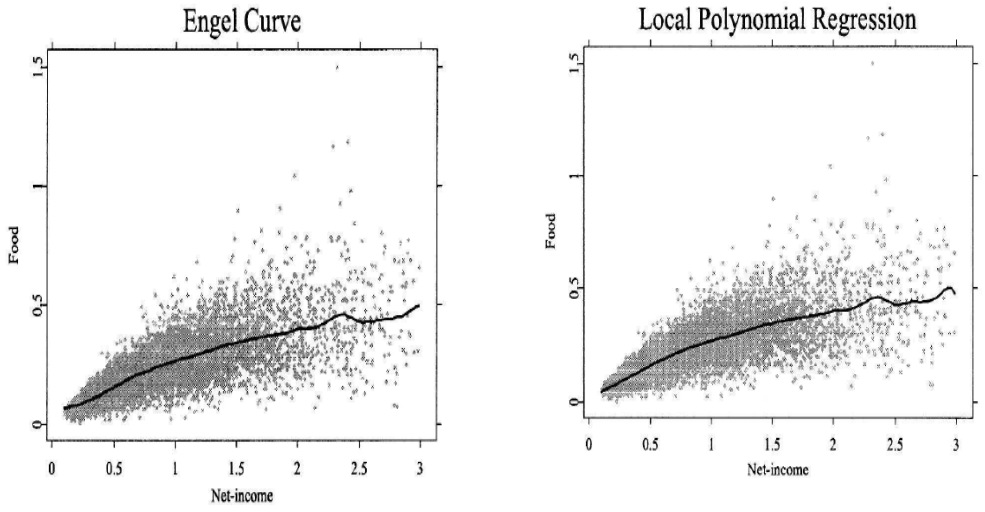

Figura 14: Polinomios locales $p = 0$ y $p = 1$.

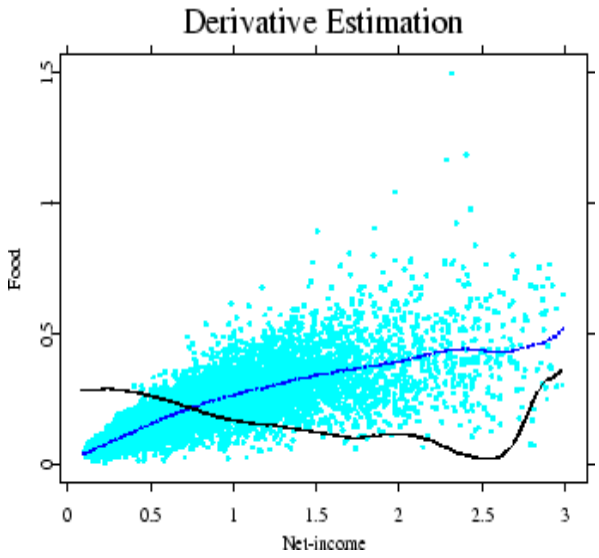

Figura 15: Estimador de regresión por polinomios locales $p = 2$ y de la primer derivada.

3.5 Método de Splines.

La motivación del método de splines es considerar la suma del cuadrado de los residuos, es decir ajustar la función m que minimice

$$\sum_{i=1}^{n}\{Y_i - m(X_i)\}^2.$$

Pero este enfoque no es demasiado bueno ya que la solución nos dara una función que interporla los datos, es decir $m(X_i) = Y_i$, provocando en muchas circunstancias estimadores con mucha varianza. La idea entonces es introducir un termino de penalización, que castiga las funciones que oscilan demasiado. Esto es posible sumando un termino de restricción

$$\|m''\|^2 = \int m''^2(x)dx.$$

Más precisamente, se busca la función m que minimice

$$S_\lambda(m) = \sum_{i=1}^{n}\{Y_i - m(X_i)\}^2 + \lambda\|m''\|^2$$

el paramétro λ juega un papel parecido al de la ventana, valores pequeños de λ producirán estimadores cercanos a la interpolación mientras que valores grandes derivarán en estimadores cercanos a una función lineal.

Se puede ver que si consideramos las funciones dos veces diferenciables en el intervalo $[X_{(1)}, X_{(n)}]$ la única solución está dada por el spline cúbico $\widehat{m}_\lambda$ que consiste en polinomios cúbicos

$$p_i(x) = \alpha_i + \beta_i x + \gamma_i x^2 + \delta_i x^3$$

con $i = 1, \ldots, n-1$, entre puntos adyacentes $X_{(i)}$ y $X_{(i+1)}$ Es decir, entre dos observaciones consecutivas se ajusta un polinomio de grado 3 y en cada uno de los nodos o observaciones los polinomios se "pegan "bien es decir coinciden hasta sus derivadas segundas.

Todas las restricciones impuestas a la minimización resulta en un sistema de ecuaciones lineales que se puede resolver con una cantidad de cálculos de orden n.

Puede verse tambien que el spline es lineal en Y, es decir que se puede escribir como $\sum_{i=1}^{n} w_{\lambda,i}(x)Y_i$.

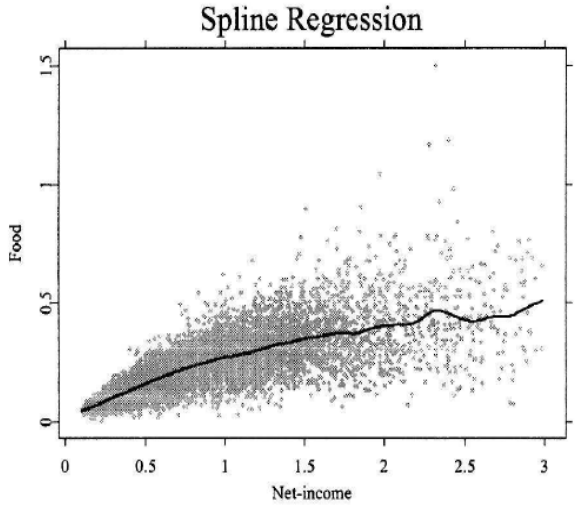

Figura 16: Estimador de splines $\lambda = 0.005$.

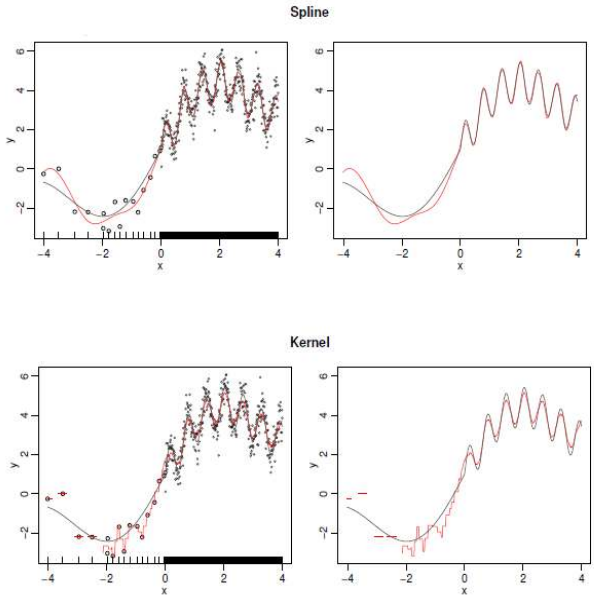

Figura 17: Comparación de splines y núcleos.

3.6 Selección del Parámetro de Suavizado: Validación Cruzada y Métodos Plug-in

Como vimos todos los los estimadores presentados dependen de un parámetro de suavizado, ya sea la ventana en núcleos, la cantidad de vecinos en vecinos más cercanos, el grado del polinomio y la ventana en polinomios locales y el parámetro λ en el método de splines. En esta sección solo nos concentraremos en la selección del parámetro de suavizado de un estimador de regresión por núcleo pero en cada caso existen propuestas para la selección de los distintos parámetros.

Si usásemos como en el caso de la estimación de la densidad la minimización del $MISE$ o $AMISE$ tendríamos el problema de que más constantes desconocidas deberían ser estimadas. Y ademas los resultados llevarian asintóticamente al mismo suavizado que otra medida que introduciremos a continuación, el error cuadrático promedio (ASE).

$$ASE(\widehat{m}_h) = \frac{1}{n}\sum_{i=1}^{n}\{\widehat{m}_h(X_i) - m(X_i)\}^2 w(X_i)$$

donde w es una función que le otroga menos pesos a las observaciones i que estan apartadas. La dificultad aquí es que el ASE contiene a $m(x)$ que es desconocida. Una forma sencilla de resolver este inconveniente es reemplazar $m(X_i)$ por Y_i

$$p(h) = \frac{1}{n}\sum_{i=1}^{n}\{\widehat{m}_h(X_i) - Y_i\}^2 w(X_i)$$

pero como $\widehat{m}_h(X_i)$ es calculada con la misma observación i la ventana que minimice tenderá a ser muy pequeña interprolando las función en las obervaciones. Entonces es conveniente

estimar la función sin la obervación i cuando evaluamos en la obervación i. De esta manera definimos el criterio de validación cruzada de la siguiente manera. Buscamos h que minimice $CV(h)$ donde

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\{\widehat{m}_{h,-i}(X_i) - Y_i\}^2 w(X_i)$$

donde $\widehat{m}_{h,-i}(X_i) = \frac{\sum_{i\neq j} K_h(X_i - X_j)Y_j}{\sum_{i\neq j} K_h(X_i - X_j)}$.

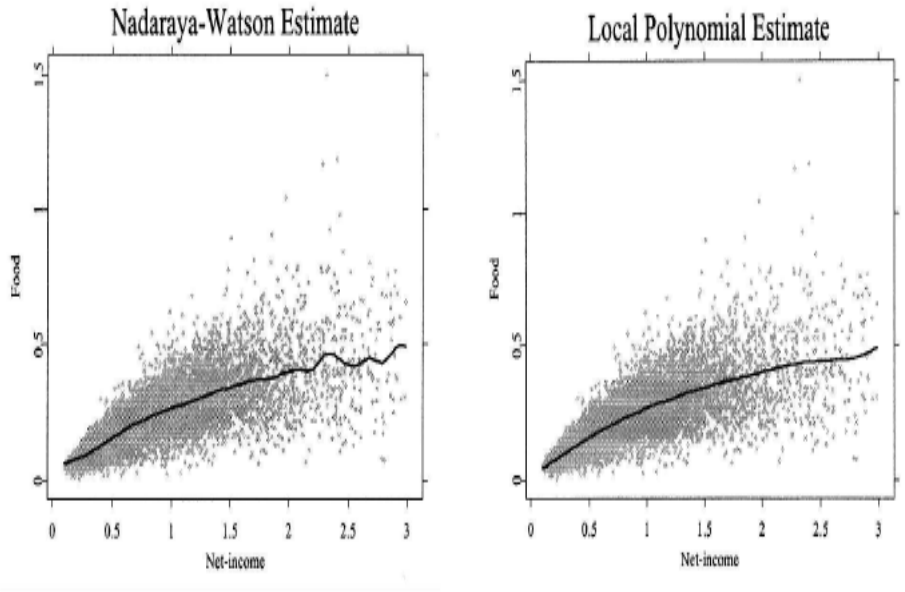

Figura 18: Estimador de Nadaraya-Watson Núcleos Cuadrático $h_{cv} = 0.15$. Estimador lineal local $(p = 1)$ $h_{cv} = 0.56$.

3.7 Inferencia con Regresión No Paramétrica.

Para obtener intervalos de confianza es necesario tener el comportamiento asintótico de los estimadores. Bajo ciertas hip'ótesis, se puede probar que el estimador de Nadaraya Watson es asintóticamente normal, mas precisamente si $h = cn^{-1/5}$ entonces

$$\sqrt{n^{4/5}}(\widehat{m}_h(x) - m(x)) \xrightarrow{\mathcal{D}} N(b_x, \nu_x^2)$$

donde

$$b_x = c^2\mu_2(K)\left(\frac{m''(x)}{2} + \frac{m'(x)f_X'(x)}{f_X(x)}\right)$$

y

$$\nu_x^2 = \frac{\sigma^2(x)\|K\|_2^2}{cf_X(x)}.$$

A partir de este resultado podemos hallar un intervalo asintótico de nivel $1-\alpha$

$$\left[\widehat{m}_h(x) - z_{\alpha/2}\sqrt{\frac{\widehat{\sigma}^2(x)\|K\|_2^2}{nh\widehat{f}_X(x)}}, \widehat{m}_h(x) + z_{\alpha/2}\sqrt{\frac{\widehat{\sigma}^2(x)\|K\|_2^2}{nh\widehat{f}_X(x)}}\right]$$

donde $\frac{1}{n}\sum_{i=1}^n W_{hi}(x)(Y_i - \widehat{m}_h(X_i))^2$ y $W_{hi}(x)$ son los pesos de Nadaraya–Watson.

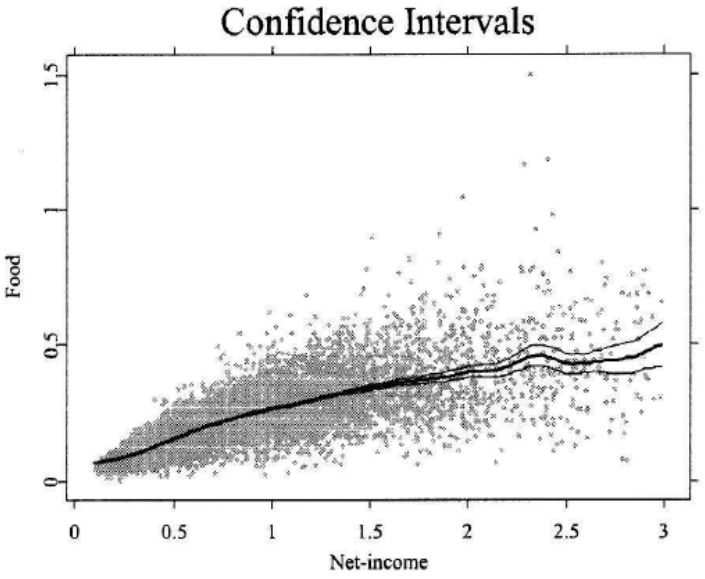

Figura 19: Intervalos de confianza de nivel 0.95, estimador de Nadaraya-Watson Núcleos Cuadrático $h = 0.2$.

Como en el caso de estimación de densidades podemos obtener bandas de confianza uniformes para $m(x)$ bajo condiciones más restrictivas. Este resultado puede encontrarse en Bickel y Rosenblatt (1973).

A la hora de realizar test de hipótesis en la estimación no paramétrica existen bastantes diferencias respecto de la estimación paramétrica. Ya que al no haber paramétros de interés, no podemos testear la significación de los mismos. En este caso las preguntas típicas en este contexto son del tipo, tiene efecto x sobre y? la función m es significativamente diferente de un modelo paramétrico? es por ejemplo lineal?

Un aspecto que es interesante es que la equivalencia que existe entre intervalos y tests en el contexto paramétrico no es muy útil aquí . Los órdenes de convergencia óptimos son distintos en estimación y test no paramétricos, por lo tanto por ejemplo el parámetro de suavizado debe determinarse por separado y la construcción de bandas de confianza alrededor de un estimador no paramétrico para decidir si es significativamente distinto de una función lineal, es muy conservativo y por lo tanto ineficiente para testear.

Existen diferentes test que se proponen en estes constexto, por ejemplo los estudiados por Härdle y Mammen (1993) que consideran el siguiente estadístico,

$$T = n\sqrt{h} \int \{\widehat{m}_h(x) - m_{\widehat{\theta}}(x)\}^2 w(x) dx$$

con el fin de testar $H_0 : m(x) = m_\theta(x)$. Sin embrago, el mayor problema práctico que enfrenta este test es la lentitud de su convergencia.

Para evitar esto, existen algunas alternativas que estudiaremos en el siguiente Capítulo tales como las técnicas de bootstrap, que permiten aproximar los valores críticos correspondientes a una distribución basada en una muestra finita.

3.8 Caso Multivariado.

El estimador de Nadaraya–Watson, puede extenderse al caso en que observamos un vector $\mathbf{X}$ de covariables d-dimensional. En este caso queremos estimar

$$E(Y|\mathbf{X}) = E(Y|X_1, \ldots, X_d) = m(\mathbf{X})$$

Al igual que en caso univariado $m(\mathbf{X}) = \frac{\int y f(y,x) dy}{f_X(x)}$, por lo tanto en este caso debemos hacer plug-in pero con estimadores de densidad multivariados

$$\widehat{f}(y, x) = \frac{1}{n} \sum_{i=1}^{n} K_h(Y_i - y) K_{\mathbf{H}}(\mathbf{X}_i - x)$$

y el estimador no paramétrico de m quedará

$$\widehat{m}_{\mathbf{H}}(x) = \frac{\sum_{i=1}^n K_{\mathbf{H}}(\mathbf{X}_i - x) Y_i}{\sum_{i=1}^n K_{\mathbf{H}}(\mathbf{X}_i - x)}.$$

Este estimador corresponde nuevamente al caso del estimador constante local, pero puede generalizarse fácilmente al caso de un polinomio local. Por ejemplo en el caso de estimador lineal local tendríamos que minimizar en función de β_0 y β_1

$$\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{X}_i - x)(Y_i - \beta_0 - {\beta_1}^T(\mathbf{X}_i - x))^2.$$

El siguiente ejemplo muestra un ejemplo simulado, son 500 puntos uniformes en el $[0, 1] \times [0, 1]$ con $m(x) = sin(2\pi x_1) + x_2$ y $\varepsilon_i \sim N(0, 1/4)$.

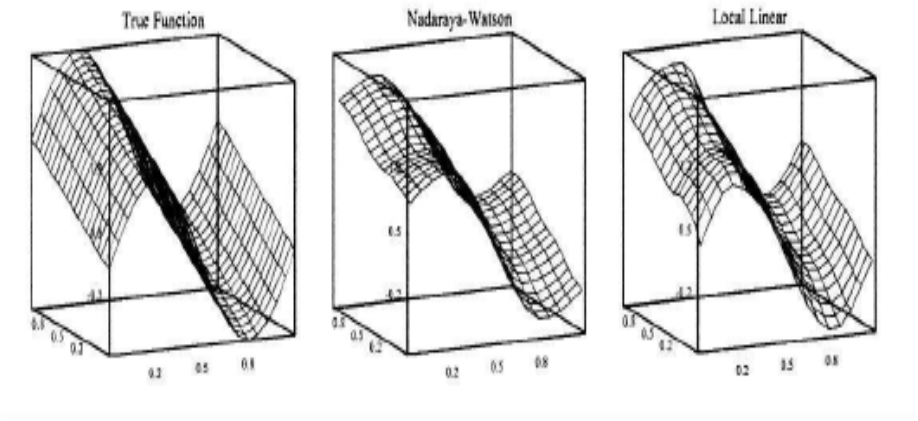

Figura 20: Comportamiento de los estimadors en el caso multivariado $h_1 = h_2 = 0.3$.

Para el caso en que $H = hI$ tenemos una expresión para el $ECMA$ y obtener a partir de el la ventana óptima que como en el caso multivariado de estimación de la densidad la ventana ptima es de la forma $h_{opt} \sim n^{-1/(4+d)}$ y la velocidad del $ECMA$ es $n^{-d/(4+d)}$, es a medida que la dimensión aumenta la velocidad de convergencia se reduce rápidamente.

3.9 Paquetes y librerías de R

Paquete "KernSmooth "

- locpoly: estimador por polinomios locales de la densidad o regresión

Paquete "sm"

- sm.regression: estimador de regresión con 1 o 2 covariables. El default es un ajuste local lineal. El parámetro model="no effect"o "linear"permite realizar tests y realizar bandas de referencia.

  ```
  sm.regression(x,y,col="red",model="linear")
  sm.regression(x,y,col="red",model="no effect")
  ```

  Con display="se"construye banda de variabilidad (son bandas en las que no se tiene en cuenta el sesgo y en realidad son intervalos de confianza para $E(\widehat{m}(x))$ en lugar de $m(x)$. Tiene nivel individual no global.

- sm.sigma: estima el desvío standard de los residuos de una regresión no paramétrica con una covariable

- ksmooth: calcula el estimador de Nadaraya-Watson

```
ksmooth(x, y, kernel = c("box", "normal"), bandwidth =

 0.5,range.x = range(x), n.points = max(100, length(x)), x.points)

plot(speed, dist) lines(ksmooth(speed, dist, "normal", bandwidth=2), col=2)
 lines(ksmooth(speed, dist, "normal", bandwidth=5), col=3)
```

- loess: ajusta un polinomio local

```
loess(formula, data, weights, subset, na.action, model = FALSE, span =

0.75, enp.target, degree = 2, parametric = FALSE, drop.square = FALSE,

 normalize = TRUE, family = c("gaussian", "symmetric"), method = c("loess",
 "model.frame"), control = loess.control(...), ...)

Para span<1 indica la proporcin de puntos que entran en el entorno

cars.lo <- loess(dist ~ speed, cars, span=.5)
```

- loess.smooth: grafica la curva calculada por loess

```
loess.smooth(x, y, span = 2/3, degree = 1, family = c("symmetric",

"gaussian"), evaluation = 50, ...)
```

3.10 Ejercicios

1. El archivo SLID del paquete car de *R* contiene datos de salarios de la provincia de Ontario, Canadá, de 1994 correspondiente a la Encuesta Canadiense de Trabajo e Ingreso Dinámico. Son en total 7425 observaciones con las siguientes variables (notar que hay un número muy grande missings):

wages	Ingreso por hora
education	Años de escolaridad
age	Edad en años
sex	Female, Male
language	Inglés Francés u otro.

   Usando gráficos de scatter plot y estimadores de regresión no paramétricos estudiar la relación de wages con education y también la de wages con age. Si esta relación parece no lineal, cuál podría ser una transformación adecuada? Tener en cuenta la distribución de wages para este último punto. Qué pasa si se estudia cada sexo por separado? Se observa lo mismo?

2. El archivo contracep.txt contiene los datos de anticoncepción en 50 países en desarrollo (Robey, Shea, Rutstein y Morris, 1992). Las variables son:

region	Africa, Asia, Latin.Amer, Near.East
tfr	Tasa total de fertilidad (niños por mujeres)
contraceptors	Porcentaje de mujeres que usan anticonceptivos entre las mujeres casadas en edad fértil

Sugiero leer los datos como read.csv(file="c:\\ ... ").

A fin de explorar la relación entre estas variables:

(a) Construir un scatter plot de tfr (respuesta) vs. contraceptors (covariable).

(b) Ajustar una recta de mínimos cuadrados y graficar.

(c) Sobreimponer un ajuste usando loess.smooth (tener en cuenta que si el parmétro span es < 1 se indica la proporción de puntos que interviene en el entorno y que en family si se indica "gaussian" el ajuste es por mínimos cuadrados).

(d) Cómo caracterizaría la relación entre las dos variables?

3. Generar 100 obsevaciones de acuerdo al modelo

$$\begin{aligned} Y_i &= (\sin(2\pi X_i^3))^3 + \varepsilon_i \\ X_i &\sim U(0,1) \\ \varepsilon_i &\sim N(0, 0.1) \end{aligned}$$

(a) Graficar el scatter plot de x vs. y. Superponer el estimador de Nadaraya-Watson usando la rutina ksmooth con los valores de default. Cuál son esos valores? Qué n ucleo se usó?

(b) Idem a) usando la ventana 0.1.

(c) Repetir a) y b) usando el núcleo normal. Comparar con los resultados anteriores.

(d) Graficar el scatter plot de x vs. y. Superponer el estimador calculado por la rutina loess.smooth con los valores de default. Superponer usando span=0.25. Comparar. Qué valor usó de span con el default?

(e) Repetir d) usando la opción family="gaussian", qué efecto tiene esto? En d) qué family se usó?

(f) Superponer en todos los gráficos la verdadera curva. Comparar.

4 Métodos basados en remuestreo.

Los procedimientos estadísticos necesitan en muchas ocaciones conocer determinadas características de la distribución de los estadísticos o los estimadores utilizados. Por ejemplo, a la ahora de obtener un test de hipótesis o intervalos de confianza se necesitan los percentiles de la distribución del estadístico empleado. Cuando llevamos a cabo un problemas de estimación es importante tener alguna medida de la exactitud o precisión como por ejemplo el error cuadrático medio del estimador obtenido. El enfoque clásico procura calcular la distribución del estadistico bajo el modelo determinado, sin embargo, la mayoría de las veces es muy difícil o imposible obtener fórmulas exactas y explícitas de la distribución o de las medidas de exactitud.

Los métodos de remuestreo buscan reemplazan las técnicas clásico evaluando los estadísticos en remuestras obtenidas a partir de los datos originales, y obteniendo mediante estos valores estimadores de las medidas de exactitud o de la distribución del estadístico. Los métodos de remuestreo más populares en la literatura estadística son el jackknife de Quenouville (1949) y Tukey (1958), y el bootstrap de Efron (1979).

Un pilar fundamental del método bootstrap lo constituye, el principio plug-in que puede interpretarse como la sustitución de la distribución subyacente F por un estimador $\widehat{F}$ de ésta. Usualmente, para estimar la distribución F, se utiliza la función de distribución empírica $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{Xi \leq x\}}$, donde $I_{\{Xi \leq x\}} = 1$ si $X_i \leq x$ y 0 en otro caso (ver en Efron (1979) y Efron y Tibshirani (1993)). Otra alternativa utilizada cuando tenemos densidad f es utilizar una $\widehat{F}_h$ asociada a un estimador de tipo kernel de la densidad (ver, por ejemplo, Cuevas y Romo (1997)).

4.1 Motivación del principio bootstrap.

Como mencionamos anteriormente el bootstrap de Efron (1979) tiene como pieza fundamental la utilización del plug-in que constituye uno de los métodos más simples a fin de obtener un estimador de un parámetro poblacional $\theta = T(F)$, donde T es un funcional definido en una clase de funciones de distribución y F es la distribución que genera los datos. Un estimador plug-in es simplemente considerar $\widehat{\theta} = T(\widehat{F})$, donde $\widehat{F}$ es un estimador de F. El ejemplo más sencillo lo constituye la media, es decir si deseamos estimar $\mu = E(X)$ donde $X \sim F$ como $\mu = E(X) = \int x \; dF$ luego si consideramos la distribución empírica de F bastará con computar el estimador $\widehat{\mu} = \int x \; d\widehat{F} = \bar{X}$.

En general el método bootstrap puede ser visto de la siguiente manera. Consideremos $\mathbf{X} = (X_1, \ldots, X_1)$ un conjunto de datos generados de acuerdo a una distribución F, y sea $T(\mathbf{X})$ un estadístico cuya distribución deseamos conocer o estimar, que llamaremos $\mathcal{L}(T, F)$.

Para fijar ideas supongamos que $X_i \sim N(\mu, 1)$ y consideremos el estadistico $T(\mathbf{X}) = \sqrt{n}(\bar{X} - \mu)$, en este caso conocemos exactamente su distribución es $N(0, 1)$ es decir $\mathcal{L}(T, F) = N(0, 1)$ así mismo si consideramos $\bar{X}$ un estimador de μ y deseamos calcular su varianza $var(\bar{X}) = var(X_1)/n$ es decir prodríamos calcular precentiles y realizar test o intervalos de confianza o dar alguna medida de precisión o exactitud del estimador. Claramente en este ejemplo es fundamental conocer la distribución de las variables para determinar matemáticamente la distribución del estadístico, aunque por la sencillez del ejemplo y el TCL del límite también podríamos concluir un resultado similar.

Sin embargo en muchas situaciones no solo que no es posible realizar ciertas suposiciones

sino que también se dificulta el calculo teórico.

El método bootstrap propone estimar $\mathcal{L}(T, F)$ a través del método plug-in con $\mathcal{L}^*(T^*, \widehat{F})$, es decir con la distribución del estadístico $T^* = T(\mathbf{X}^*)$ donde $\mathbf{X}^* = (X_1^*, \ldots, X_n^*)$ generados a partir de $\widehat{F}$.

Como hacemos esto en la práctica?

Consideremos $X_1, \ldots, X_n$ un muestra aleatoria con distribución F. En primer lugar creamos una nueva muestra de tamanõ n de la muestra original que realizaremos a partir de un muestreo con reemplazo $\mathbf{X}^* = (X_1^*, \ldots, X_n^*)$. Esta muestra se denomina muestra Bootstrap. Es importante notar que muestrar con reemplazo es equivalente a obtener una muestra con distribución $\widehat{F}$.

En un segundo paso, calculamos el estadístico o estimador de interés $T(\mathbf{X}^*)$ por lo tanto este estadístico tendrá distribución $\mathcal{L}^*(T^*, \widehat{F})$. Si bien no conoceamos la distribución $\mathcal{L}(T, F)$ pues no conocemos F o es difícil de calcular es posible que si podamos calcular la distribución bootstrap $\mathcal{L}^*(T*, \widehat{F})$ pues $\widehat{F}$ si es conocida. En este caso hemos calculado entonces una aproximación de $\mathcal{L}(T, F)$ que es lo que estabamos buscando.

Un ejemplo sencillo de esto pero que no desarrellaremos sus cuentas es la mediana, si $T(\mathbf{X}) = X_{(n+1/2)}$ si n es impar entonces la distribución de la mediana dependerá de F pero la distribución de $T(\mathbf{X}^*) = X^*_{(n+1/2)}$ es posible calcularla completamente. Pues $T(\mathbf{X}^*) = X^*_{(n+1/2)}$ puede tomar solo los valores de la muestra original es decir es discreta en $X_1, \ldots, X_n$.

Pero que sucede si la distribución bootstrap tampoco es posible calcularla? Es aquí donde aparece el verdadero provecho del método bootstrap. Pues la distribución bootstrap es siempre aproximable a partir de Monte Carlo. El punto clave es el hecho que podemos generar tantas muestras bootstrap como querramos. Luego si generamos B nuevas muestras de tamanõ n, para cada una de las B muestras calculamos el estadístico o estimador de interés, obteniendo B estadísticos $T(\mathbf{X}^*)_i$ para $1 \leq i \leq B$. De esta manera obtenemos una "muestra "de estadísticos y a partir de ellos podemos calcular la distribución empírica basada en $T(\mathbf{X}^*)_i$ para $1 \leq i \leq B$

4.2 Estimación del sesgo y precisión de un estimador.

4.3 Bootstrap paramétrico

4.4 Bootstrap no paramétrico

4.4.1 Regresión

En esta sección estudiaremos como realizar un test para determinar si la función de regresión pertenece a un determinado modelo paramétrico. Más precisamente, la hipótesis nula corresponde a $H_0 : m(x) = m_\theta(x)$ con $\theta \in \Theta$. Por ejemplo $m_\theta(x) = \theta_1 x + \theta_0$.

Utilizaremos lo que se conoce como wild bootstrap (Wu, 1986). Esta implementación al caso de un test no paramétrico se debe a Härdle y Mammen (1993). La idea aquí es resamplear sobre los residuos $\widehat{\varepsilon_i} = Y_i - \widehat{m}(X_i)$ para $i = 1, \ldots, n$ obtenidos bajo H_0.

Cada residuo bootstrap ε_i^* se muestrea de una distribución que coincide con la de $\widehat{\varepsilon_i}$

hasta los tres primeros momentos:

$$E(\varepsilon_i^*) = 0 \qquad E(\varepsilon_i^{*2}) = \widehat{\varepsilon_i}^2 \qquad E(\varepsilon_i^{*3}) = \widehat{\varepsilon_i}^3$$

Los pasos a seguir son los siguientes:

Paso 1. Estimar la función de regresión m_θ bajo H_0, y construir los residuos $\widehat{\varepsilon_i} = Y_i - \widehat{m}(X_i)$

Paso 2. Para cada X_i muestrea un residuo bootstrap ε_i^* de manera tal que

$$E(\varepsilon_i^*) = 0 \qquad E(\varepsilon_i^{*2}) = \widehat{\varepsilon_i}^2 \qquad E(\varepsilon_i^{*3}) = \widehat{\varepsilon_i}^3$$

Paso 3. Generar muestras bootstrap (Y_i^*, X_i) mediante $Y_i^* = \widehat{m}(X_i) + \varepsilon_i^*$

Paso 4. Calcular el estadístico T^* de la misma forma que el original

$$T = n\sqrt{h} \int \{\widehat{m}_h(x) - m_{\widehat{\theta}}(x)\}^2 w(x) dx$$

pero utilizando las muestras bootstrap.

Paso 5. Repetir los pasos 2 a 4 nboot (del orden de cientos o miles) y determinar los percentiles de los estadsticos T^* obtenidos.

Una manera sencilla de obtener residuos bootstrap que cumplan con el Paso 2, es considerando una distribución discreta a dos valores conocida como golden cut method:

$$a = \frac{1 - \sqrt{5}}{2} \widehat{\varepsilon_i}$$

y

$$b = \frac{1 + \sqrt{5}}{2} \widehat{\varepsilon_i}$$

que ocurren con probabilidad $q = frac5 + \sqrt{5}10$ y $1 - q$, respectivamente.

5 Conclusiones

1. Ventajas
   (a) Hipótesis libre de distribuciones
   (b) Mínima cantidad de supuestos
   (c) Son más potentes cuando las hipótesis parametricas no se cumplen.
   (d) Son fáciles de entender y calcular
2. Desventanjas
   (a) Menos potentes frente a sus alternativas parametricas
   (b) Las hipótesis nulas son mas complejas

(c) Se necesita mayor número de observaciones respecto a los test nóparametricos para lograr una misma potencia.

3. Cuando usar

   (a) cuando los supuestos no parámetricos son desconocidos.

   (b) si las hipótesis no envuelven un parámetro poblacional

References

[1] Azzalini, A. y Bowman, A. (1999). Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations. Oxford Statistical Science Series

[2] Härdle, W. (1991) Applied Nonparametric Regression. Econometric Society Monographs

[3] Härdle, W., Mller , M., Sperlich, S. y Werwatz, A. (2004) Nonparametric and Semiparametric Models

[4] W. Härdle. (1999) Applied Nonparametric Regression. Econometric Society Monographs.

[5] Horowitz, Joel (2009) Semiparametric and Nonparametric Methods in Econometrics Springer Series in Statistics

[6] Ruper, D., Wand, M, Carrol. R. (2003) Semiparametric Regression. Cambridge.

[7] Tsiatis, Anastasios. (2010) Semiparametric Theory and Missing Data. Springer Series in Statistics)

[8] Wasserman, L. (2006) All of Nonparametric Statistics. Springer Texts in Statistics.