UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales
Departamento de Matemática

Tesis de Licenciatura

Estimadores de Naradaya Watson aplicados a datos funcionales

Romina Cornistein

Directora: Dra. Mariela Sued

Marzo de 2013

Agradecimientos

Esta tesis es el resultado final de mis años como estudiante de Matemática. Este camino lo recorrí con personas hermosas, indispensables, luminosas.

En primer lugar le quiero agradecer a Mariela, por ser parte fundamental de este trabajo, por ayudarme a crecer en cada momento que necesité de su guía y por tener las puertas de su oficina siempre abiertas.

A mis compañeros, por el hermoso grupo que hemos construido. A Lu, Pau, Flor S., Luquitas, Pablo, Flor V. y Lucho. A Marian y sus entrañables Criaturas. A Manu, por las tardes en que parametrizamos al mundo. A Marie F., Vero M., por ser excelentes compañeras de cursada. A Iru y Lu S. Al seminario de Birra de los Viernes, Juli G., Anita, Flor Z., Pau K, por hacer tan divertido este último tramo que por momentos resultó interminable.

Al Grupo de Mareas del Servicio del Hidrografía Naval por su apoyo incondicional, en especial a Enrique y Mónica F.

A mi familia, por apoyarme siempre, y sobrevivir a mis malos humores.

A los Frenkel, por hacerme parte de su familia.

A *Estelita,* por el camino recorrido.

A Maguita, Vero Singer, Denu y Tami Colodenco por llenar mis días de ternura y optimismo.

A Brian, por sus lecciones de coraje.

Índice general

Introducción

El análisis estadístico frecuentemente se concentra en modelos parámetricos. En este enfoque, la estimación se reduce a determinar eficientemente el valor de finitos parámetros mediante una muestra, para utilizarlos posteriormente para construir intervalos de confianza, test de hipótesis, etc. Este paradigma supone que las variables estudiadas siguen la estructura fijada por el modelo, suposición que puede no ser cierta si el modelo propuesto es indadecuado. Mas aun, puede ocurrir que las familias parametricas no proporcionen un buen ajuste.

Para sortear esta dificultad, en las últimas décadas ha cobrado auge la estadística no paramétrica, que relaja los supuestos a familias funcionales más generales de dimensión infinita, pudiendo ser implementados a una gran cantidad de problemas. El costo de esta mayor flexibilidad, es la menor eficiencia que poseen respecto de los estimadores paramétricos, ya que se cuenta con la misma información para estimar objetos de dimensión finita en el primer caso e infinita en el segundo.

Los primeros trabajos se concentraron en la estimación de la función de densidad para una distribución continua y fueron desarrollados por Rosenblatt [1], Parzen [2] y Whittle [3]. Para cada punto, se construye un entorno a partir del cual se le da mayor peso a las observaciones más cercanas al mismo. En este proceso, entra en juego la función de peso, que llamaremos núcleo y la longitud del entorno, que llamaremos ventana o parámetro de suavizado. Posteriormente, Naradaya [4] y Watson [5] abordan el problema de regresión. En este contexto, se quiere expresar el vínculo existente entre la variable de respuesta y las covariables explicativas cuando existe poco conocimiento a priori de su forma. Los estimadores no paramétricos proporcionan formas suaves de su relación y explotan la idea de promedios ponderados locales alrededor del punto sobre el cual se estima. En ambos problemas, la bondad de la estimación está relacionada con la cantidad de observaciones disponibles alrededor del punto en el cual se contruye el entorno: ventanas demasiado pequeñas filtrarán muchos datos ocasionando estimadores muy variantes, mientras que entornos demasiado grandes producirán estimadores casi constantes y muy suaves.

Por otra parte, en muchos campos de la ciencia aplicada se deben enfrentar problemas de inferencia estadística que incluyen datos aleatorios presentados en forma de curvas. Más precisamente, se cuenta con una alta cantidad de observaciones discretizadas en una grilla. Una primera idea es modelar cada elemento de la grilla como una variable unidimensional y realizar el análisis clásico multivariado. En este enfoque, se ignora la correlación de las muestras y se debe enfrentar los problemas de alta dimensionalidad de las covariables, debido a la baja relación entre éstas y la información disponible. Por este motivo, surge el tratamiento funcional de esta clase de datos, conocido como Análisis de Datos Funcionales (FDA), popularizado por Ramsay & Silverman [13]. Ferraty & Vieu [14] extienden los estimadores de regresión por núcleos al caso de covariables funcionales y obtienen resultados de consistencia casi completa.

Esta tesis de divide en tres capítulos. En el capítulo 1, se introducen los estimadores de densidad y de regresión por núcleos, obteniéndose resultados de consistencia en probabilidad y se exploran las técnicas de obtensión del parámetro de suavizado más utilizadas. En el capítulo 2, se extienden los estimadores de regresión para el caso de covariables funcionales y se prueban resultados de

consistencia casi completa bajo hipótesis de continuidad de la función de regresión, cotas de los momentos condicionales de la respuesta, la regularidad del núcleo y la distribución de los procesos. Asimismo, se extienden estos resultados para el caso de covariables finito dimensionales. En el capítulo 3, se implementan los estimadores de regresión con covariables funcionales al problema de datos faltantes, realizando un estudio de simulación bajo el mismo escenario que Ferraty *et al.* [16].

Capítulo 1

Estimación no paramétrica

Supongamos que X es una variable aleatoria continua y estamos interesados en estimar estimar f, su función de densidad. Cuando no se cuenta con supuestos fuertes sobre su naturaleza, utilizar un método paramétrico para estimarla puede llevar a resultados erróneos. Es por ello, que los estimadores no paramétricos surgieron como una alternativa más flexible y entre los mas estudiados, se encuentran los estimadores de densidad por núcleos introducidos por Rosenblatt [1] y Parzen [2]. A continuación, intentaremos dar una idea intuitiva de su construcción.

1.1. Estimadores de la densidad por núcleos

Sea $X_1, \cdots, X_n$ una muestra aleatoria con densidad $f(x)$. Nuestro objetivo será estimar $f(x)$ a partir de la misma. Sabemos que para todo $h > 0$, $f(x)$ satisface

$$P\left(x \in (x-h, x+h)\right) = \int_{x-h}^{x+h} f(t)\, dt.$$

Si aproximamos a la probabilidad de un conjunto con la proporción de ocurrencia, podemos reemplazar el lado izquierdo de esta igualdad por $\frac{\sharp\{X_i \in (x-h, x+h)\}}{n}$, mientras que el lado derecho por $f(x)2h$, si h es suficientemente pequeño y f es continua en x. Despejando podemos aproximar $f(x)$ por la siguente expresión:

$$f(x) \approx \frac{\sharp\{X_i \in (x-h, x+h)\}}{2nh}. \tag{1.1}$$

De (1.1) podemos deducir un estimador reemplazando las aproximaciones por igualdades

$$\hat{f}(x) = \frac{\sharp\{X_i \in (x-h, x+h)\}}{2nh}.$$

Observemos que $\hat{f}(x)$ se puede escribir como

$$\frac{1}{nh} \sum_{i=1}^{n} \frac{1}{2} \mathbf{I}_{[-1,1]} \left(\frac{x - X_i}{h} \right),$$

donde $\mathbf{I}_A(t)$ es la función indicadora del conjunto A, valiendo 1 cuando t pertenece a A y 0 caso contrario.

Si llamamos $w(x) = \frac{1}{2}\mathbf{I}_{[-1,1]}(x)$, tenemos entonces que $\hat{f}(x) = \sum_{i=1}^{n} \frac{1}{nh} w\left(\frac{x - X_i}{h}\right)$. Notemos que $w(x)$ es una función de densidad pues $w \geq 0$, $\int w(x)dx = 1$, de lo cual deducimos que $\hat{f}(x)$ también lo es. En efecto, $\hat{f}(x) \geq 0$ y

$$\begin{aligned}
\int \hat{f}(x)dx &= \int \frac{1}{2nh} \sum_{i=1}^{n} w\left(\frac{x - X_1}{h}\right) dx \\
&= \int \frac{1}{2nh} \sum_{i=1}^{n} \mathbf{I}\left(x \in (X_i - h, X_i + h)\right) dx \\
&= \frac{1}{2nh} \sum_{i=1}^{n} \int_{X_i - h}^{X_i + h} dx \\
&= \frac{1}{2nh} \sum_{i=1}^{n} 2h \\
&= 1.
\end{aligned}$$

Observemos que $w(\cdot)$ otorga un peso uniforme igual a $\frac{1}{2}$ a aquellas observaciones X_i ubicadas en un entorno h de x e igual a cero en el resto. Si quisiésemos dar mayor peso a observaciones más cercanas podríamos reemplazar a $w(x)$ por otra función $K(x)$ que también verifique:

- $K \geq 0$
- $\int K(x)dx = 1$.

y otro posible estimador $\hat{f}_K$ sería:

$$\hat{f}_K(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right). \tag{1.2}$$

A la función K se la denomina *función de peso* o *núcleo* y a h se lo llama *ventana* o *parámetro de suavizado*. A esta familia $\hat{f}_K$ de estimadores se los define como *Estimadores por núcleos* o *Estimadores de Naradaya-Watson*. Observemos también que $\hat{f}_K$ hereda el grado de suavidad de K. Por otra parte, el comportamiento del estimador dependerá de los valores que tome h. Suponiendo que K tiene soporte compacto

- Si $h \to 0$, el $K\left(\frac{x - X_i}{h}\right)$ valdrá 0 si $x \neq X_i$, y $\frac{K(0)}{nh}$ si $x = X_i$ para algún i. Entonces, el estimador dará 0 si $x \neq X_i$ y un valor muy grande si $x = X_i$.
- Si $h \to +\infty$, $K\left(\frac{x - X_i}{h}\right) \to K(0)$ para $1 \leq i \leq n$, por lo cual $\hat{f}_K(x) \to 0$.

1.1.1. Propiedades estadísticas

Nuestro objetivo será buscar condiciones mediante las cuales se garantize la convergencia en probalidad de los estimadores por núcleos a medida que agrandamos el tamaño de la muestra. En lo que sique, supondremos que K satisface las siguientes condiciones

H1. $K \geq 0$ y $\int K(u)\, du = 1$.

H2. $\int uK(u)\, du = 0$ y $\int uK^2(u)\, du = 0$.

H3. $\int |s^i||K^j(s)|ds < +\infty$, para $i = 1, 2, \frac{5}{2}$ y $j = 1, 2$.

Bajo ciertas hipótesis de regularidad sobre $f(x)$, obtenemos el siguiente resultado:

Proposición 1.1.1. *Sea f una función de densidad $C^2(\mathbb{R})$ con $f''(\cdot)$ absolutamente continua tal que $||f'''||_{L^2} < \infty$ y K un núcleo univariado que satisaface las hipótesis **H1**, **H2** y **H3**.*

Si $n \to \infty$, $h \to 0$ y $nh \to \infty$, entonces

$$\hat{f}_K(x) \xrightarrow{P} f(x).$$

Demostración. La idea será probar que bajo las hipótesis de la Proposición, el error cuadrático medio $ECM(\hat{f}_K(x)) = \mathbb{E}\left((\hat{f}_K(x) - f(x))^2\right)$ tiende a 0. Dado que $ECM(\hat{f}_K(x)) = Sesgo(\hat{f}_K(x))^2 + \mathbb{V}\text{ar}(\hat{f}_k(x))$, analizaremos cada término por separado.

$Sesgo(\hat{f}_K(x)) = \mathbb{E}(\hat{f}_K(x)) - f(x)$:

$$\begin{aligned}
\mathbb{E}(\hat{f}_K(x)) &= \mathbb{E}\left(\frac{1}{nh}\sum_i^n \left(\frac{x - X_i}{h}\right)\right) \\
&= \frac{n}{nh}\,\mathbb{E}\left(K\left(\frac{x - X}{h}\right)\right) \\
&= \frac{1}{h}\int K\left(\frac{x-u}{h}\right) f(u)du.
\end{aligned}$$

Mediante el cambio de variable $y = \frac{x-u}{h}$ y en vista de que $\int K(u)\,du = 1$ (**H1**) obtenemos que

$$\mathbb{E}(\hat{f}_K(x)) - f(x) = \int K(y)\,(f(x - hy) - f(x))\,dy.$$

Haciando el desarrollo de Taylor de f alrededor de x y expresando el resto en la forma integral obtenemos la siguiente cota,

$$f(x - hy) - f(x) = -f'(x)hy + \frac{f''(x)}{2}y^2h^2 + \int_{x-yh}^{x} \frac{f'''(t)}{2!}(x - t)^2\,dt.$$

Con lo cual,

$$\begin{aligned}
|\,\mathbb{E}(\hat{f}_K(x)) - f(x)| &= \left|\int K(y)\left[-f'(x)hy + \frac{f''(x)}{2}h^2y^2 + \int_{x-yh}^{x}\frac{f'''(t)}{2!}(x-t)^2\,dt\right]dy\right| \\
&= \left|-f'(x)h\int K(y)y\,dy + h^2\frac{f''(x)}{2}\int k(y)y^2dy + \int\int_{x-yh}^{x}\frac{f'''(t)}{2!}(x-t)^2\,dt\,K(y)\,dy\right| \\
&\leq \left|h^2\frac{f''(x)}{2}\right|\int K(y)y^2\,dy + \left|\int\frac{1}{2!}\left(||f'''||_{L^2}\right)\left(\int_{x-yh}^{x}|x-t|^4\,dt\right)^{\frac{1}{2}}K(y)\,dy\right| \\
&\leq \left|h^2\frac{f''(x)}{2}\right|\int K(y)y^2\,dy + \frac{||f'''||_{L^2}}{2!\sqrt{5}}h^{\frac{5}{2}}\int|y|^{\frac{5}{2}}K(y)\,dy \\
&\leq \left|h^2\frac{f''(x)}{2}\right|\int K(y)y^2\,dy + o(h^2)
\end{aligned}$$

Por lo tanto,

$$|\,Sesgo(\hat{f}_K(x))| \leq h^2 \left|\frac{f''(x)}{2}\right| \int K(y)y^2\,dy + o(h^2).$$

Analizemos $\mathbb{V}\mathrm{ar}(\hat{f}_K(x))$:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(\hat{f}_K)(x) &= \mathbb{V}\mathrm{ar}\left(\frac{1}{nh}\sum_{i=1}^{n} K(x-X_i)\right) && (1.3)\\
&= \frac{1}{n^2h^2} n\,\mathbb{V}\mathrm{ar}\left(K\left(\frac{x-X}{h}\right)\right) && (1.4)\\
&\leq \frac{1}{nh^2}\,\mathbb{E}\left(K^2\left(\frac{x-X}{h}\right)\right). && (1.5)
\end{aligned}$$

Estudiemos $\mathbb{E}\left(\frac{1}{h^2}K^2\left(\frac{x-X}{h}\right)\right)$:

Haciendo el mismo cambio de variable

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{h^2}K^2\left(\frac{x-X}{h}\right)\right) &= \frac{1}{h}\int K^2(y) f(x-hy)\,dy\\
&\leq \frac{1}{h}\left|\int K^2(y)\left(f(x) - f'(x)hy + \frac{f''(x)}{2}h^2y^2\right)dy\right| + Ch^{\frac{5}{2}}\int K^2(y)|y|^{\frac{5}{2}}\,dy\\
&\leq \frac{1}{h}f(x)\int K^2(y)\,dy + \frac{1}{2h}|f''(x)|h^2\int K^2(y)y^2\,dy + \frac{1}{h}o(h^2)\int K^2(y)|y|^{\frac{5}{2}}\,dy.
\end{aligned}$$

Podemos acotar $\mathbb{V}\mathrm{ar}(\hat{f}_K(x))$ de la siguiente forma

$$\mathbb{V}\mathrm{ar}(\hat{f}_K(x)) \leq \frac{1}{nh}f(x)\int K^2(y)\,dy + o\left(\frac{1}{nh}\right).$$

Finalmente,

$$ECM(\hat{f}_K(x)) \leq h^4\frac{f''^2(x)}{4}\left(\int K(y)y^2dy\right)^2 + o(h^4) + \frac{f(x)}{nh}\int K^2(y)dy + o\left(\frac{1}{nh}\right).$$

Lo cual implica que

$$\hat{f}_K(x) \xrightarrow{P} f(x),$$

en vista de que $n \to \infty$, $h \to 0$ y $nh \to \infty$. □

Observación 1.1.2. Es fácil ver que bajo las mismas condiciones sobre el núcleo, n y h, el estimador resulta consistente pidiendo solo condiciones locales de continuidad. En efecto,

$$\begin{aligned}
\mathbb{E}(\hat{f}_K(x)) &= \int \frac{1}{h}K\left(\frac{x-u}{h}\right)f(u)du && (1.6)\\
&= K_h * f(x),\ \text{donde } K_h(x) = \tfrac{1}{h}K(\tfrac{x}{h}). && (1.7)
\end{aligned}$$

Por resultados de teoría de la medida [6], sabemos que siendo K un es núcleo integrable, acotado tal que $\frac{K(x)}{x} \to 0$ si $|x| \to \infty$ y f continua en x e integrable se verifica

$$K_h * f(x) \xrightarrow{h\to 0} f(x),$$

lo que prueba que el estimador es asintóticamente insesgado. Por otra parte, de la ecuación (1.5) y observando que $\frac{K^2}{\int K^2}$ es también un núcleo, $K_h^2 * f(x) \xrightarrow{h\to 0} f(x) \int K^2(y)\, dy$ y podemos deducir que

$$\mathbb{V}\mathrm{ar}(\hat{f}_K)(x) \leq \frac{1}{nh} K_h^2 * f(x) \to 0. \tag{1.8}$$

Esta propiedad permite relajar las hipótesis e implementar el estimador para variables más generales.

De la estructura del $ECM(\hat{f}_K(x))$, podemos ver que si consideramos h muy grandes, el sesgo aumenta mientras que la varianza disminuye. Por otra parte, si el producto nh es muy pequeño, la varianza no tenderá a cero. Es decir, fijado n, la elección de h resulta fundamental en la implementación de $\hat{f}_K(x)$ para equilibrar el sesgo y la varianza. En el siguiente ejemplo, podremos apreciar la influencia de este factor en el resultado de la estimación.

Ejemplo 1. Se estimó la densidad de una muestra aleatoria de tamaño 100 con distribución normal estandar para distintas elecciones núcleos y ventanas. En las Figuras 1.1 y 1.2 se pueden apreciar los resultados obtenidos. Los núcleos utilizados fueron los siguientes:

- **Gaussiano:**
$$K_G(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$
- **Uniforme:**
$$K_U(x) = \begin{cases} \frac{1}{2} & \text{si } |x| \leq 1 \\ 0 & \text{si } |x| > 1 \end{cases}$$
- **Triangular:**
$$K_T(x) = \begin{cases} 1 - |x| & \text{si } |x| \leq 1 \\ 0 & \text{si } |x| > 1 \end{cases}$$
- **Cuadrático:**
$$K_C(x) = \begin{cases} \frac{3}{4}(1 - x^2) & \text{si } |x| \leq 1 \\ 0 & \text{si } |x| > 1 \end{cases}$$

Como se puede apreciar en la figura 1.2 es necesario contar con un método de selección de h. A continuación presentaremos los métodos más utilizados.

1.2. Selección de la ventana

1.2.1. Método *plug-in*

Consideremos el *Error Cuadrático Medio Asintótico* de $\hat{f}_K(x)$ ($ECMA$), que resulta de despreciar los términos pequeños en (1.1.1):

$$ECMA(\hat{f}_K(x)) = \frac{h^4 f''^2(x)}{4} \left(\int y^2 K(y) dy \right)^2 + \frac{f(x)}{nh} \|K\|_{L_2}^2$$

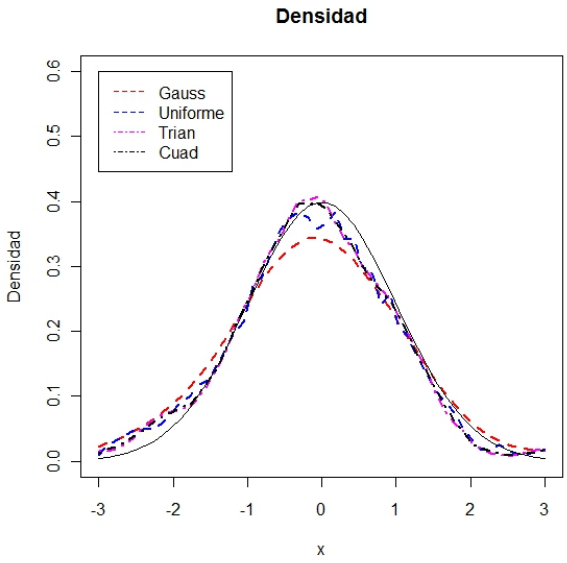

Figura 1.1: *Estimadores de densidad para distintos núcleos y* $h = 0.4$. *Podemos observar que los resultados obtenidos son semejantes, tal como lo indica la bibliografía [8]. Las diferencia principal es la suavidad, heredada del núcleo.*

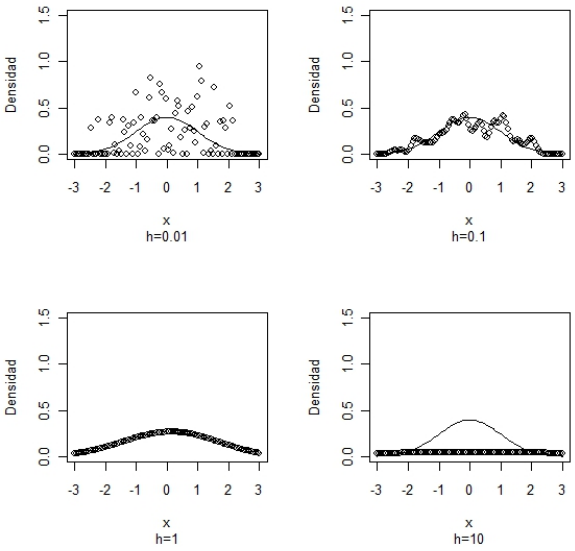

Figura 1.2: *Estimadores de densidad para distintas ventanas implementado con el núcleo Gaussiano. Los resultados obtenidos para distintas ventanas pueden diferir significativamente: para valores pequeños de* h *el estimador es muy variante y poco suave, mientras que para valores grandes de* h*, el estimador es casi constante. Por este motivo,* h *también es llamado "Parámetro de uavizado".*

Un criterio para elegir h consistirá en minimizar la expresión anterior para cada x fijo. Nos bastará derivar la expresión respecto de h e igualar a 0, obteniendo así la siguiente elección óptima

$$h_{opt} = \left(\frac{\|K\|_{L_2}^2 f(x)}{f''(x) \int y^2 K(y) dy} \right)^{\frac{1}{5}} n^{-\frac{1}{5}}.$$

Observemos que la ventana óptima depende del punto x, y además de valores desconocidos como $f(x)$ y $f''(x)$. Una forma de solucionar la dependencia local consiste en buscar h que minimize el *Error Cuadrático Medio Asintótico Integrado* (*ECMAI*),

$$ECMAI(\hat{f}_K) = \int ECMA\, \hat{f}_K(x)\, dx = \int \frac{h^4}{4} f''^2(x)\, dx \left(\int y^2 K(y)\, dy \right)^2 + \frac{1}{nh} \|K\|_{L_2}^2.$$

Derivando nuevamente respecto de h , deducimos $\tilde{h}_{opt}$:

$$\tilde{h}_{opt} = \left(\frac{\|K\|_{L_2}^2}{\|f''\|_{L_2}^2 \int K(y) y^2 dy} \right)^{\frac{1}{5}} n^{-\frac{1}{5}}. \tag{1.9}$$

Para calcular $\tilde{h}_{opt}$, necesitamos $\|f''\|_{L_2}^2$. Para estimarlo, podemos implementar una técnica *plug-in*, que consiste en reemplazarla por un estimador no paramétrico de $\|\hat{f}''\|^2$, utilizando una ventana prefijada. Cuando podemos suponer regularidad sobre f, podemos aplicar *La Regla Referencia Normal* desarrollada por Silverman [7] que consiste en suponer que f sigue una distribución normal $N(\mu, \sigma^2)$. Si utilizamos el núcleo Gaussiano, reemplazaremos en la ecuación (1.9) por los siguientes valores:

$$\int y^2 K_G(y)\, dy = 1,$$

$$\int K_G^2(y)\, dy = \frac{1}{2\sqrt{\pi}},$$

$$||f''||_{L_2}^2 = \sigma^{-5} \frac{3}{8\sqrt{\pi}}.$$

Reemplazando esta información en la ecuación 1.9, obtenemos

$$\tilde{h}_{opt} = \left(\frac{4\hat{\sigma}}{3n} \right)^{\frac{1}{5}} \approx \frac{1.06}{n^{\frac{1}{5}}} \hat{\sigma}$$

donde $\hat{\sigma}$ es un estimador consistente de σ. Generalmente se propone

$$\hat{\sigma} = \text{mín} \left\{ \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}, \frac{\hat{Q}}{1.34} \right\},$$

donde $\hat{Q}$ representa un estimador del rango intercuantil. Recordemos que el rango intercuantil es el precentil 75 menos el percentil 25. La razón por la cual se divide por 1.34 radica en el hecho que $\frac{Q}{1.34}$ resulta un estimador consistente de σ el caso de que la variable siga una distribución $N(\mu, \sigma^2)$.

1.2.2. Elección de ventana mediante convalidación cruzada

Los métodos de convalidación consisten en minimizar una medida de distancia entre f y $\hat{f}_K$, por ejemplo la norma en $L^2(\mathbb{R})$,

$$||\hat{f}_K(x) - f(x)||_{L_2} = \sqrt{\int_{-\infty}^{+\infty} (\hat{f}_K(x) - f(x))^2 \, dx}, \tag{1.10}$$

con esta distancia, definiremos

$$h_{opt} = \min_h ||(\hat{f}_K(x)) - f(x)||^2_{L_2}.$$

Debido a que $f(x)$ es desconocida, h_{opt} no puede ser estimado por la muestra. A fines prácticos, nos convendrá expresar la distancia de la siguiente forma

$$\begin{aligned} ||(\hat{f}_K(x)) - f(x)||^2_{L_2} &= \int (\hat{f}^2_K(x) - f(x))^2 \, dx \\ &= \int \hat{f}^2_K(x) \, dx - 2 \int \hat{f}_K(x) f(x) \, dx + f^2(x) \, dx. \end{aligned}$$

Observemos que el último término no depende de h, con lo cual minimizar $||(\hat{f}(x)) - f(x)||^2_{L_2}$ será equivalente a minimizar $\int \hat{f}^2_K(x) \, dx - \int \hat{f}_K(x) f(x) \, dx$. Veamos cómo estimar cada término en función de la muestra

$$\int \hat{f}^2_K(x) \, dx = \int \left(\frac{1}{nh} \sum_i^n K\left(\frac{x - X_i}{h} \right) \right)^2 dx \tag{1.11}$$

$$= \frac{1}{n^2 h^2} \sum_{i,j}^n \int K\left(\frac{x - X_i}{h} \right) K\left(\frac{x - X_j}{h} \right) dx, \tag{1.12}$$

$$= \frac{1}{n^2 h} \sum_{i,j}^n \int K(y) K\left(y - \left(\frac{X_j - X_i}{h} \right) \right) dy \tag{1.13}$$

$$= \frac{1}{n^2 h} \sum_{i,j}^n K * K\left(\frac{X_i - X_j}{h} \right). \tag{1.14}$$

El término $\int \hat{f}_K(x) f(x) \, dx$ podemos interpretarlo como la esperanza que tendría la variable aleatoria extra $\hat{f}_K(\tilde{X})$, donde estamos pensado que $X_1, \ldots, X_n$ se encuentran fijas y $\tilde{X}$ tiene como función de densidad a f. Como no contamos con muestras adicionales, estimaremos la esperanza por $\widehat{\mathbb{E}}\left(\hat{f}_K(X) \right) = \frac{1}{n} \sum_i^n \hat{f}_{K,-i}(X_i)$, donde $\hat{f}_{K,-i}(X_i)$ corresponde al estimador *leave-one-out* de $f(X_i)$ y que consiste en calcular el estimador de densidad sin la i-ésima observación, es decir

$$\hat{f}_{K,-i}(X_i) = \frac{1}{h(n-1)} \sum_{i \neq j}^n K\left(\frac{X_i - X_j}{h} \right). \tag{1.15}$$

Uniendo (1.14) y (1.15) queda definido $\hat{h}_{CV}$,

$$\hat{h}_{CV} = \underset{h}{argmin} \, CV(h) = \underset{h}{argmin} \, \frac{1}{n^2 h} \sum_{i,j}^n K * K\left(\frac{X_i - X_j}{h} \right) - \frac{2}{hn(n-1)} \sum_{i \neq j}^n K\left(\frac{X_i - X_j}{h} \right). \tag{1.16}$$

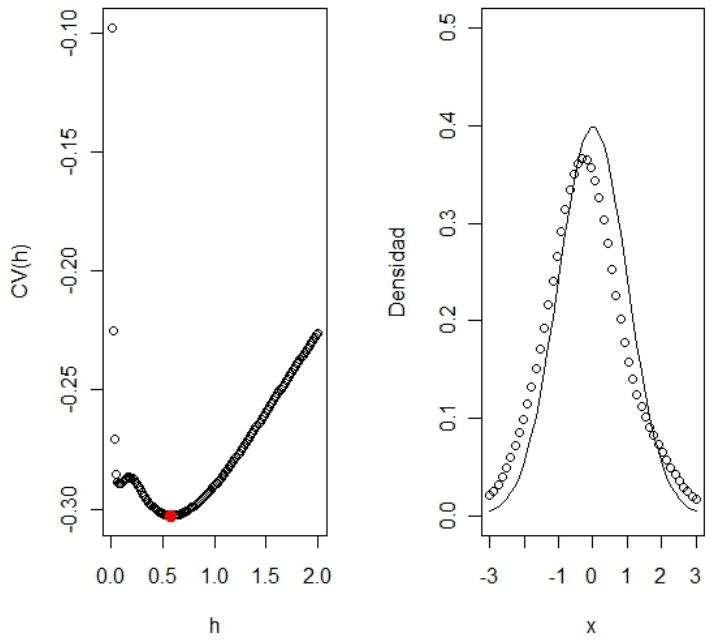

Figura 1.3: Función de pérdida $CV(h)$ y función de densidad calculada para h_{min}.

Para evaluar cúan buena es la aproxmación de $\hat{h}_{CV}$, consideraremos que es *asintóticamente óptimo* si

$$\frac{||\hat{f}_{K,h=\hat{h}_{CV}}(x) - f(x)||^2_{L_2}}{||\hat{f}_{K,h=h_{opt}}(x) - f(x)||^2_{L_2}} \xrightarrow{c.t.p.} 1.$$

Para funciones con densidad acotada, El Teorema de Stone [9] prueba que $\hat{h}_{CV}$ es asintóticamente óptimo.

Se calculó la función de pérdida $CV(h)$ dada por 1.16 para la muestra del Ejemplo 1. Siendo K el núcleo gaussiano, $K * K(x) = \frac{1}{\sqrt{2 \cdot 2\pi}} e^{-\frac{1}{2 \cdot 2} x^2}$. A partir del gráfico, se halló el mínimo, y se gráfico la densidad para el h_{min} hallado. En la Figura 1.3 se pueden observar los resultados obtenidos.

1.3. Extensión al caso multivariado

Consideremos una densidad $f(\vec{x})$ definida sobre $\mathbb{R}^d$. Supongamos que contamos con una muestra aleatoria $\vec{X_i} = (X_{i1}, \cdots, X_{id})$, $1 \leq i \leq d$. Para un cubo pequeño centrado en $\vec{x}$ de lado h $(C_h(\vec{x}))$,

podemos aproximar la probabilidad de dicho cubo por

$$\begin{aligned} P\left(C_h(\vec{x})\right) &\approx \frac{\sum_{i=1}^{n} \mathbf{I}_{C_h(\vec{x})}(\vec{x}_i)}{n} \\ f(\vec{x})h^d &\approx \frac{\sum_{i=1}^{n} \mathbf{I}_{C_h(\vec{x})}(\vec{x}_i)}{n} \\ f(\vec{x}) &\approx \frac{\sum_{i=1}^{n} \mathbf{I}_{C_h(\vec{x})}(\vec{x}_i)}{nh^d} \end{aligned}$$

Análogamente al caso unidemensional, el estimador $\hat{f}_K(\vec{x})$ de Naradaya-Watson consistirá en reemplazar $\mathbf{I}_{C_h(\vec{X})}$ por un núcleo multivariado $K : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$, quedando definido de la siguiente manera,

$$\hat{f}_K(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^d} K\left(\frac{\vec{x} - \vec{X}_i}{h}\right). \tag{1.17}$$

Para funciones $C^2(\mathbb{R}^d)$, con derivadas parciales terceras integrables en cada coordenada Härdle [10] obtiene condiciones suficientes de consistencia pidiendo que $h \to 0$ y $nh^d \to +\infty$. Mas aun, muestra que

$$|\, Sesgo(\hat{f}_K(\vec{x}))| \leq C_1(\vec{x})h^2 + o(h^2),$$

mientras que

$$\mathbb{V}\mathrm{ar}(\hat{f}_K(\vec{x})) \leq C_2(\vec{x})\frac{1}{nh^d}.$$

El cambio de la relación entre h y n en la varianza puede interpretarse de la siguiente forma: fijado n, para obtener una buena estimación de $\hat{f}_K(\vec{x})$, h deberá incrementarse con la dimensión de forma tal que haya suficientes observaciones en el cubo de lado h alrededor de X. En consecuencia, el lado del cubo se incrementará en cada dirección de manera que se obtengan suficientes datos para promediar. Consideremos, por ejemplo, la distribución uniforme en el cubo unitario de $\mathbb{R}^{10}$. Un cubo de lado $0,5$ cuyo volumen es $(\frac{1}{2})^{10}$ tendrá una probabilidad aproximada de $0,001$. Entonces, se necesitará un promedio 1000 observaciones para que al menos una de ellas pertenezca a dicho cubo. Esta relación entre la cantidad de datos necesarios y la dimensión resulta de caracter exponencial y es llamada *"La Maldición de la Dimensión"*. Por este fenómeno, estos estimadores son usados para dimensiones bajas.

Se pueden construír núcleos multivariados, a partir de núcleos univariados proponiendo $K(\vec{x}) = K(\|\vec{x}\|)$, donde $\|\ .\ \|$ es una norma en $\mathbb{R}^d$. Otra alternativa es considerar núcleos mutiplicativos $K_{mult}(\vec{x}) = K(x_1)\dots K(x_d)$ y utilizar en cada coordenada distintas ventanas, obteniendo el siguiente estimador alternativo para $\hat{f}_K(\vec{x})$

$$\hat{f}_K(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_d \dots h_d} K\left(\frac{x_1 - x_{i_1}}{h_1}\right)\dots K\left(\frac{x_d - x_{i_d}}{h_d}\right).$$

1.4. Regresión no paramétrica

Cuando hablamos de regresión entre dos variables aleatorias X e Y, intentamos describir el comportamiento *promedio* de Y, que llamaremos *variable de respuesta*, en presencia de X que

llamaremos *covariable*. Más precisamente, estamos pensando que contamos con una muestra de v.a.i.i.d. (X_i, Y_i), $1 \leq i \leq n$, para las cuales se verifica el siguiente modelo

$$Y = r(X) + \epsilon \tag{1.18}$$

donde $r : \mathbb{R} \to \mathbb{R}$ es la función de regresión desconocida y ϵ es el error con $\mathbb{E}(\epsilon|X) = 0$ y $\mathbb{V}\text{ar}(\epsilon|X) = \sigma^2$. De esta última igualdad se sigue que

$$r(X) = \mathbb{E}(Y|X).$$

Supondremos X e Y variables aleatorias continuas con distribución conjunta f_{XY} tal que $\mathbb{E}(|Y|) < \infty$. Recordemos cómo hallar $\mathbb{E}(Y|X = x)$, si conocemos $f_{XY}(x, y)$ y $f_X(x)$. Para $f_X(x) > 0$,

$$\begin{aligned} \mathbb{E}(Y|X = x) &= \int y f_{X|Y}(y|x)\, dy & (1.19)\\ &= \int y \frac{f_{XY}(x, y)}{f_X(x)}\, dy & (1.20)\\ &= \frac{m(x)}{f_X(x),} & (1.21) \end{aligned}$$

siendo $m(x)$

$$m(x) = \int y f_{XY}(x, y)\, dy. \tag{1.22}$$

De 1.20 podemos deducir un estimador de $r(x)$ practicando una técnica *plug-in*, es decir, reemplazando $f_{XY}(x, y)$ y $f_X(x)$ por sus respectivos estimadores de Naradaya-Watson

$$\hat{r}_K(x) = \frac{\int y \hat{f}_{K,XY}(x, y)\, dy}{\hat{f}_{K,X}(x)}$$

donde propondremos

$$\begin{aligned} \hat{f}_{K,XY}(x, y) &= \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) K\left(\frac{y - Y_i}{h}\right)\\ \hat{f}_{K,X}(x) &= \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right). \end{aligned}$$

Aquí consideramos núcleos multiplicativos para estimar la densidad conjunta f_{XY} y tomaremos en cada coordanada la misma ventana h.

Miremos $\hat{m}_K(x) = \int y\hat{f}_{XY}(x,y)\,dy$:

$$\begin{aligned}\int y\hat{f}_{K,XY}(x,y)\,dy &= \sum_{i=1}^{n}\int \frac{y}{nh^2}K\left(\frac{x-X_i}{h}\right)K\left(\frac{y-Y_i}{h}\right)\,dy \\ &= \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)\int \frac{y}{h}K\left(\frac{y-Y_i}{h}\right)\,dy \\ &= \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)\int (sh+Y_i)K(s)\,ds, \text{ bajo } H1 \text{ y } H2 \\ &= \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)Y_i.\end{aligned}$$

Con lo cual, queda definido el *Estimador de Regresión de Nadaraya-Watson,*

$$\hat{r}_K(x) = \frac{\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)Y_i}{\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)}. \tag{1.23}$$

En caso de que el denominador sea 0, el numerador también lo será y el estimador en esos puntos no se encontrará definido.

Observación 1.4.1. Notemos que si llamamos $W_i(x) = \frac{K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)}$ se cumplen: i) $W_i(x) \geq 0$ y ii) $\sum_i^n W_i(x) = 1$, con lo cual $\hat{r}_K(x) = \sum_{i=1}^{n} W_i(x)Y_i$ resulta un promedio de las observaciones Y_i ponderado locamente por el peso $W_i(x)$. Más aun, es fácil demostrar derivando la función $\sum_{i=1}^{n} W_i(x)(Y_i - a)^2$ respecto de a que

$$\hat{r}_K(x) = \underset{a\in\mathbb{R}}{argmin}\sum_{i=1}^{n} W_i(x)(Y_i - a)^2.$$

Análogamente a los estimadores de densidad, $\hat{r}_K(x)$ depende de la elección adecuada de h. Resulta interesante analizar el comportamiento del estimador en los casos de que h tome valores extremos. Suponiendo nuevamente que K tiene soporte compacto,

- Si $h \to 0$, el $K\left(\frac{x-X_i}{h}\right)$ valdrá 0 salvo que $x = X_i$. En el primer caso $\hat{r}_K(x)$ no se encontrará definido y en el segundo $\hat{r}_K(X_i) = Y_i$. Es decir, el estimador interpolará los datos.
- Si $h \to +\infty$, para todo x, $K\left(\frac{x-X_i}{h}\right) \to K(0)$, $W_i(x) \to \frac{1}{n}$ y $\hat{r}_K(x) \to \bar{Y}$.

Una posible extensión del estimador de regresión por núcleos para $x \in \mathbb{R}^d$ estará dada por

$$\hat{r}_K(\vec{x}) = \frac{\sum_{i=1}^{n}K\left(\frac{\|\vec{x}-\vec{X}_i\|}{h}\right)Y_i}{\sum_{i=1}^{n}K\left(\frac{\|\vec{x}-\vec{X}_i\|}{h}\right)}. \tag{1.24}$$

Härdle [10] demuestra la consistencia de esta propuesta.

1.4.1. Propiedades estadísticas

Cuando deseamos deducir las propiedades estadísticas de $\hat{r}_K$, nos encontramos frente a la dificultad de que tanto el numerador como el denominador que lo constituyen son variables aleatorias. En lugar de acotar el *ECM*, estudiaremos el comportamiento asintótico de cada uno de los factores que conforman el estimador, con el objetivo de probar la convergencia en probabilidad.

Lema 1.4.2. *Sean X e Y variables aleatorias continuas y K un núcleo univariado que cumplen las siguientes hipótesis:*

- *K satisface las hipótesis **H1**,**H2** y **H3**.*
- *$f_X(x) > 0$, $f(\cdot) \in C^2(\mathbb{R})$, con $f''(.)$ absolutamente continua y $||f'''||_{L^2} < \infty$.*
- *$m(\cdot) \in C^2(\mathbb{R})$, con $m''(\cdot)$ absolutamente continua y $||m'''||_{L^2} < \infty$.*
- *$s^2(\cdot) = \mathbb{E}(Y^2|X = \cdot) \in C^2(\mathbb{R})$, con $(s^2)''(\cdot)$ absolutamente continua y $||(s^2)'''||_{L^2} < \infty$.*

Entonces,

$$ECM(\hat{m}_K(x)) \leq h^4 \frac{(m(x)'')^2}{4} \left(\int K(y) y^2 \, dy \right)^2 + o(h^4) + \frac{s^2(x) f_X(x)}{nh} \int K^2(y) \, dy + o\left(\frac{1}{nh}\right).$$

Demostración.

Analizemos $|\, Sesgo(\hat{m}_K(x))| = \left| \mathbb{E}(\hat{m}_K(x)) - m(x) \right|$

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h} \right) Y_i \right) &= \frac{1}{h} \mathbb{E}\left(K\left(\frac{x - X}{h} \right) Y \right) \\
&= \frac{1}{h} \mathbb{E}\left(\mathbb{E}(Y|X) K\left(\frac{x - X}{h} \right) \right) \\
&= \frac{1}{h} \mathbb{E}\left(r(X) K\left(\frac{x - X}{h} \right) \right) \\
&= \frac{1}{h} \int r(u) K\left(\frac{x - u}{h} \right) f_X(u) \, du.
\end{aligned}$$

Si hacemos el desarrollo de Taylor de la función $r(u) f_X(u) = m(u)$ alredededor de x como en la proposición 1.1.1 obtenemos que

$$|\, \mathbb{E}(\hat{m}_K(x)) - m(x)| = |\, Sesgo(\hat{m}_K(x))| \leq \frac{h^2}{2} |(m(x)'')| \int y^2 K(y) \, dy + o(h^2). \tag{1.25}$$

Veamos $\mathbb{V}\text{ar}\left(\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) Y_i\right)$:

$$\begin{aligned}
\mathbb{V}\text{ar}\left(\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) Y_i\right) &= \frac{1}{nh^2}\,\mathbb{V}\text{ar}\left(K\left(\frac{x-X}{h}\right) Y\right) \\
&\leq \frac{1}{nh^2}\,\mathbb{E}\left(\left(K\left(\frac{x-X}{h}\right) Y\right)^2\right) \\
&\leq \frac{1}{nh^2}\,\mathbb{E}\left(\mathbb{E}\left(K\left(\frac{x-X}{h}\right)^2 Y^2 \middle| X\right)\right) \\
&= \frac{1}{nh^2}\,\mathbb{E}\left(\mathbb{E}(Y^2|X) K\left(\frac{x-X}{h}\right)^2\right) \\
&= \frac{1}{nh^2}\int K\left(\frac{x-u}{h}\right)^2 s^2(u) f_X(u)\,du \\
&= \frac{1}{nh} f_X(x) s^2(x) \int K^2(s)\,ds + o\left(\frac{1}{nh}\right).
\end{aligned}$$

Por lo tanto, deducimos que

$$\mathbb{V}\text{ar}\left(\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) Y_i\right) \leq \frac{1}{nh} f_X(x) s^2(x) \int K^2(s)\,ds + o\left(\frac{1}{nh}\right). \tag{1.26}$$

Finalmente, $ECM(\hat{m}(x))$ nos queda acotado por:

$$ECM(\hat{m}_K(x)) \leq \frac{h^4}{4}\left(m(x'')\right)^2 \left(\int y^2 K(y)\,dy\right)^2 + o(h^4) + \frac{1}{nh} f_X(x) s^2(x) \int K^2(s)\,ds + o\left(\frac{1}{nh}\right).$$

□

Proposición 1.4.3. *Bajo las mismas hipótesis que el Lema 1.4.2, $\hat{r}_K(x)$ resulta un estimador débilmente consistente de $r(x)$.*

Demostración. De (1.25) y (1.26) concluimos que el numerador converge en probabilidad a $m(x)$ si $n \to \infty$, $h \to 0$ y $nh \to \infty$. Observemos que en la Proposición 1.1.1 probamos que el denominador converge en probabilidad $f_X(x)$ bajo las mismas hipótesis. Finalmente, mediante las propiedades de la convergencia en probabilidad con respecto al cociente, siendo $f_X(x) > 0$, podemos asegurar que $\hat{r}_K(x) = \frac{\hat{m}_K(x)}{\hat{f}_{K,X}(x)}$ converge en probabilidad a $r(x)$. □

1.4.2. Método de selección de ventana

Como señalamos anteriormente para la densidad, la elección del parámetro de suavizado resulta fundamental para obtener una buena estimación. En el caso de regresión, también nos interesará minimizar alguna distancia entre $r(x)$ y $\hat{r}_K(x)$, por ejemplo:

- **Error Cuadrático Promediado**

$$ECP(h) = \frac{1}{n}\sum_{i=1}^{n} \left(r(X_i) - \hat{r}_K(X_i)\right)^2 w(X_i),$$

donde $w(.)$ es una función de no negativa cuya finalidad es darle menor influencia a las observaciones próximas a la frontera del soporte de X, en el cual el estimador cuenta con menor observaciones y en consecuencia pierde calidad. Observemos que $ECP(h)$ es una variable aletoria que depende de $X_1, \cdots, X_n$.

- **Error Cuadrático Integrado**

$$ECI(h) = \int \left(r(x) - \hat{r}_K(x)\right)^2 w(x) f_X(x)\, dx.$$

$ECI(h)$ es también una variable aleatoria, que posee como peso adicional de $w(x)$ la función de densidad $f_X(x)$.

- **Error Cuadrático Promediado Esperado**

$$ECPE(h) = \mathbb{E}(ECP(h)|X_1, \cdots, X_n).$$

Observemos que $\hat{r}_K(x)$ es una variable aleatoria que depende de $Y_1, \cdots, Y_n$ y $X_1, \cdots, X_n$, por lo cual, $ECPE(h)$ es una variable aleatoria que consiste en considerar la esperanza condicional de $ECP(h)$ respecto de $X_1, \cdots, X_n$.

- **Error Cuadrático Integrado Esperado**

$$ECIE(h) = \mathbb{E}(ECI(h)).$$

$ECIE(h)$ no es una variable aleatoria, pues se computa el $ECI(h)$ sobre todos las posibles muestras de X e Y.

Härdle y Marron [11] encuentran condiciones bajo las cuales demuestran la equivalencia entre las distancias propuestas anteriormente. Formalmente, prueban que

$$\sup_{h \in H_n} \left| \frac{d(h) - ECIE(h)}{ECIE(h)} \right| \xrightarrow{ctp.} 0,$$

donde $H_n = [n^{\delta-1}, n^{-\delta}]$, $0 < \delta < \frac{1}{2}$ y $d(h) = ECP(h), ECI(h), ECPE(h)$.

Este resultado nos permite elegir la distancia sobre la cual optimizar h. La más sencilla de estimar resulta $ECP(h)$ y sobre ella basaremos nuestro criterio de selección. En lo que sigue, tomaremos $w(X_i) = 1$.

Método de convalización cruzada

El cómputo de $ECP(h)$ resulta imposible en términos prácticos, debido a que $r(X_i)$ es desconocida. Una aproximación de $ECP(h)$ consistirá en reemplazar a $r(X_i)$ por Y_i, quedandando definida la siguiente distancia

$$R(h) = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{r}_K(X_i)\right)^2. \tag{1.27}$$

La aproximación de $R(h)$ por $ECP(h)$ puede verse cuantificarse por la siguiente relación

$$\begin{aligned} R(h) &= \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{r}_K(X_i) \pm r(X_i)\right)^2 && (1.28) \\ &= \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 + ECP(h) - \frac{2}{n} \sum_{i=1}^{n} \epsilon_i (\hat{r}_K(X_i) - r(X_i)), && (1.29) \\ & && (1.30) \end{aligned}$$

donde $\epsilon_i = Y_i - r(X_i)$.

El primer término de $R(h)$ no depende de h y una vez obtenida la muestra es un número fijo. Por el contrario, el último si es dependiente de h. Como hipótesis adicional, supondremos que los errores ϵ_i, donde $1 \leq i \leq n$, resultan condicionalmente independientes respecto de $X_1, \cdots, X_n$.

Analizemos la esperanza condicional de $-2\epsilon_i(\hat{r}_K(X_i) - r(X_i))$:

$$\begin{aligned}
\mathbb{E}\left(-2\epsilon_i(\hat{r}_K(X_i) - r(X_i))\middle| X_1, \cdots, X_n\right) &= -2\,\mathbb{E}\left(\epsilon_i(\hat{r}_K(X_i) - r(X_i))\middle| X_1, \cdots, X_n\right) \\
&= -2\,\mathbb{E}\left(\epsilon_i\left(\sum_{j=1}^{n} W_j(X_i)Y_j - r(X_i)\right)\middle| X_1, \cdots, X_n\right) \\
&= -2\sum_{i=1}^{n}\mathbb{E}\left(\epsilon_i\left(\sum_{j=1}^{n} W_j(X_i)(r(X_j)+\epsilon_j) - r(X_i)\right)\middle| X_1, \cdots, X_n\right) \\
&= -2\,\mathbb{E}\left(\epsilon_i|X_1, \cdots, X_n\right)\left(\sum_{j=1}^{n} W_j(X_i)(r(X_j) - r(X_i))\right) - \\
&\quad -2\sum_{j=1}^{n} W_j(X_i)\,\mathbb{E}(\epsilon_i\epsilon_j|X_1, \cdots, X_n),
\end{aligned}$$

donde $W_j(X_i) = \frac{K\left(\frac{X_i - X_j}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{X_i - X_j}{h}\right)}$.

Siendo ϵ_i y ϵ_j condicionalmente independientes dado $X_1, \ldots, X_n$ obtenemos que $\mathbb{E}\left(\epsilon_i\epsilon_j\middle| X_1, \cdots, X_n\right) = 0$ para $i \neq j$,

$$\begin{aligned}
\mathbb{E}\left(-2\epsilon_i(\hat{r}_K(X_i) - r(X_i))\middle| X_1, \cdots, X_n\right) &= -2\sum_{j=1}^{n} W_j(X_i)\,\mathbb{E}\left(\epsilon_i\epsilon_j|X_1, \cdots, X_n\right) \\
&= -2W_i(X_i)\sigma^2
\end{aligned}$$

De aquí concluimos que este término tiene esperanza con signo negativo. Este término tendría esperanza nula si no utilizaríamos el par $(X_i.Y_i)$ simultáneamente para estimar $R(h)$ y $r(X_i)$. Intuitivamente, podemos pensar que $R(h)$ tenderá a elegir h pequeños que interpolen a los datos. Una forma de compensarlo es reemplazar a $\hat{r}_K(X_i)$ por su estimador *leave-one-out*,

$$\hat{r}_{K,-i}(X_i) = \frac{\sum_{j\neq i}^{n} K\left(\frac{X_i - X_j}{h}\right) Y_j}{\sum_{j\neq i}^{n} K\left(\frac{X_i - X_j}{h}\right)},$$

y considerar la función de Pérdida de Convalidación Cruzada, $(CV(h))$ definida por

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{r}_{K,-i}(X_i)\right)^2.$$

Fijadas $X_1, \cdots, X_n$, $CV(h)$ resulta una función que difiere de $ECP(h)$ por $\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2$ independiente de h y un término adicional que consiste en un promedio de una variable que tiene esperanza

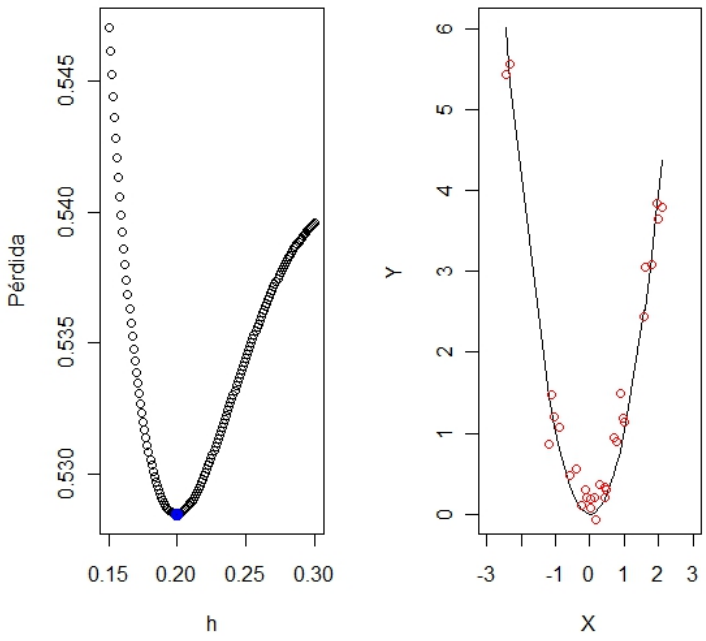

Figura 1.4: *Función CV(h).*

nula. El procedimiento para selccionar h consistirá en elegir aquel que minimize $CV(h)$, quedando así definido

$$\hat{h}_{CV} = \underset{h}{argmin} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{r}_{K,-i}(X_i)\right)^2 .$$

Si llamamos h_{min} aquel que minimiza la pérdida $ECP(h)$, Härde [12] prueba que

$$\frac{ECP(\hat{h}_{CV})}{ECP(h_{min})} \xrightarrow{P} 1.$$

En consecuencia, $\hat{h}_{CV}$ heredará las mismas propiedades asintóticas de h_{min}. En particular el resultado de equivalencia entre $ECP(h)$, $ECI(h)$, $ECPE(h)$ y $ECPE(h)$.

Ejemplo 2. Se generó una muestra de tamaño 30 bajo el modelo de regresión

$$Y_i = X_i^2 + \epsilon_i,$$

donde $X_i \sim U[-3, 3]$ y $\epsilon_i \sim N(0, 0.5)$ para $1 \leq i \leq 30$. Luego se implementó el estimador de regresión *leave-one-out* de Naradaya-Watson a cada X_i eligiendo el núcleo Gaussiano, y se eligió h mediante el criterio de Convalidación Cruzada. Los resultados obtenidos pueden apreciarse en la Figura 1.4.2.

Capítulo 2

Regresión funcional

En muchas áreas de investigación los datos recolectados provienen de funciones. Imaginemos por ejemplo una curva $X(t)$, de la cual observamos $\{X(t_j)\}_{j=1\dots J.}$. Antiguamente, cada $X(t_j)$ era tratada como una variable aleatoria unidimensional. Los avances computacionales hacen posible hacer la grilla cada vez más fina, de modo que podríamos considerar estas muestras como observaciones de la familia continua $\mathcal{X} = \{X(t), t \in I \subseteq \mathbb{R})\}$. Los métodos estadísticos tradicionales no son adecuados cuando ignoramos el carácter funcional de las mismas, primero porque la relación del tamaño de la muestra y el número de covariables es muy baja: recordemos que en el contexto no paramétrico los estimadores de Naradaya-Watson eran usados para bajas dimensiones debido a *"La Maldición de la Dimensión"*. Segundo, debido a la existencia de fuertes correlaciones entre las variables explicativas provocando por ejemplo, el mal condicionamiento del problema de regresión lineal multivariado. Es por ello que en los ultimos años se desarrollaron técnicas de modelización que tuvieron en cuenta la estructura funcional de esta clase de datos.

Motivados por el tratamiento funcional de los datos, será preciso introducir el concepto de *variable aleatoria funcional.* Intentatemos construir su definición recurriendo a los conceptos conocidos al caso multidimensional. Sea $(\Omega, \mathcal{A}, \mathcal{P})$ un espacio de probabilidad, $||.||$ una norma definida en $\mathbb{R}^n$ y $\mathcal{B}(\mathbb{R}^n)$ la σ-*álgebra* de Borel inducida por $||.||$. Recordemos que un vector aleatorio $\Omega \xrightarrow{\vec{X}} \mathbb{R}^n$ es una función $\mathcal{A}$-medible, es decir, una aplicación que para todo boreliano $\mathcal{B} \in \mathcal{B}(\mathbb{R}^n)$, $\vec{X}^{-1}(\mathcal{B})$ resulta un conjunto de $\mathcal{A}$. En el caso funcional, reemplazaremos $\mathbb{R}^n$ por un espacio de métrico (E, d) inmerso en un espacio de Banach $(E, ||.||)$ de dimensión infinita.

Definición 2.0.4. Sea $(\Omega, \mathcal{A}, \mathcal{P})$ un espacio de probabilidad, (E, d) un espacio métrico de dimensión infinita y $\mathcal{B}(E)$ la σ-álgebra de Borel inducida por d . Diremos que $\mathcal{X} : \Omega \to E$ es una *variable aleatoria funcional (vf.)* si para todo $\mathcal{B} \in \mathcal{B}(E)$, $\mathcal{X}^{-1}(\mathcal{B}) \in \mathcal{A}$. Llamaremos *dato funcional (df.)* a una observación χ de $\mathcal{X}$ y *muestra aleatoria funcional (maf.)* a un conjunto $\chi_1 \dots \chi_n$ de observaciones de $\mathcal{X}$ independientes entre sí.

En este trabajo supondremos que para toda variable aleatoria real y continua $Y : \Omega \to \mathbb{R}$ con primer momento finito se encontrará garantizada la existencia de $\mathbb{E}(Y|\mathcal{X} = \chi)$. Nuestro objetivo será extender los estimadores de regresión no paramétricos introducidos en el capítulo anterior a espacios funcionales. Estudiaremos el modelo de regresion no paramétrico con covariable funcional que cumple las siguientes hipótesis

$$Y = r(\mathcal{X}) + \epsilon \tag{2.1}$$
$$\mathbb{E}(\epsilon|\mathcal{X}) = 0. \tag{2.2}$$

Si bien la frontera entre estimación paramétrica y no parámetrica en el caso multivariado resulta

evidente, deberemos describir con mayor precisión dichos conceptos en estos espacios. Utilizaremos la defincón introducida por Ferraty [14].

Definición 2.0.5. Sea $\mathcal{X}$ una vaf. definida sobre E y $r : E \to \mathbb{R}^d$ una función que depende de la distribución de $\mathcal{X}$. Diremos que un *modelo de estimación funcional es paramétrico* cuando $r \in \mathcal{C}$, donde $\mathcal{C}$ es un espacio indexado por un conjunto finito de elementos de E. Por otra parte, diremos que un *modelo de estimación funcional es no paramétrico* cuando $\mathcal{C}$ no sea indexable por un conjunto finito de elementos de E.

Ejemplo 3. Sean Y *va.*, $H = L^2[0,1]$, dotado del producto interno $< g, f >= \int_0^1 g(x)f(x)\,dx$. Llamemos $r(\chi) = \mathbb{E}(Y|\mathcal{X} = \chi)$. El modelo de regresión lineal funcional estará asociado a restringir $r(\chi)$ al conjunto

$$\mathcal{C} = \left\{ r(\chi) = T(\chi), \text{ donde } T : L^2[0,1] \to \mathbb{R} \text{ es un operador lineal y continuo} \right\}.$$

Aplicando el Teorema de Representación de Riesz, T admite un único representante $\phi \in L^2[0,1]$ que satisface

$$r(\chi) =< \phi, \chi >= \int_0^1 \phi(x)\chi(x)\,dx.$$

Por lo tanto $\mathcal{C}$ es indexable a único parámetro ϕ de $L^2[0,1]$.

Ejemplo 4. Sean Y y H como en el ejemplo 3, un modelo de regresión general consistirá en restringir a $r(\chi)$ al conjunto

$$\mathcal{C} = \left\{ r(\chi) = T(\chi), \text{ donde } T : L^2[0,1] \to \mathbb{R} \text{ es un operador continuo} \right\}.$$

Observemos que $\mathcal{C}$ no es indexable a un conjunto finito de elementos de E y por lo tanto este modelo de regresión es no paramétrico.

2.1. Regresión funcional no paramétrica

Recordemos el estimador de regresión Naradaya-Watson para covariables $\vec{x} \in \mathbb{R}^d$ estaba dado por la siguiente expresión

$$\hat{r}_K(\vec{x}) = \frac{\sum_{i=1}^n K\left(\frac{||\vec{x}-\vec{X}_i||}{h}\right) Y_i}{\sum_{i=1}^n K\left(\frac{||\vec{x}-\vec{X}_i||}{h}\right)}. \tag{2.3}$$

La extensión natural en el caso funcional para $\chi \in (E, d)$ será entonces

$$\hat{r}_K(\chi) = \frac{\sum_{i=1}^n K\left(\frac{d(\chi,\chi_i)}{h}\right) Y_i}{\sum_{i=1}^n K\left(\frac{d(\chi,\chi_i)}{h}\right)}. \tag{2.4}$$

Análogamente al caso multidimensional, el estimador se encontrará definido siempre que $\sum_{i=1}^n K\left(\frac{d(\chi,\chi_i)}{h}\right) \neq 0$. Observemos que el modelo subyacente a 2.4 es no parámetrico, y asumiremos que $r(\chi)$ pertenece a la familia de operadores continuos.

Debido a que evaluaremos a K en argumentos positivos, consideraremos núcleos con soporte en los reales no negativos. Usaremos en consecuencia núcleos asimétricos, en particular los derivados del ejemplo 1:

- **Gaussiano asimétrico:**

$$K_{GA}(x) = \begin{cases} \frac{\sqrt{2}}{\sqrt{\pi}} e^{-\frac{1}{2}x^2} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

- **Uniforme asimétrico:**

$$K_{UA}(x) = \begin{cases} 1 & \text{si } 0 \leq x \leq 1 \\ 0 & \text{caso contrario} \end{cases}$$

- **Triangular asimétrico:**

$$K_{TA}(x) = \begin{cases} 2(1-x) & \text{si } 0 \leq x \leq 1 \\ 0 & \text{caso contrario} \end{cases}$$

- **Cuadrático asimétrico:**

$$K_{CA}(x) = \begin{cases} \frac{3}{2}(1-x^2) & \text{si } 0 \leq x \leq 1 \\ 0 & \text{caso contrario} \end{cases}$$

Observación 2.1.1. En el capítulo anterior mostramos que la elección de la ventana era importante para obtener una buena estimación, mientras que en el ejemplo 1 notamos que los resultados eran similares para distintas elecciones de núcleos. Los mismo ocurrirá- sin tener en consideración los costos computacionales- con las posibles normas que midan la distancia entre elementos de $\mathbb{R}$, debido a que las mismas son equivalentes. No obstante, esto no ocurre en espacios de dimensión infinita, elementos que son cercanos respecto de una norma $\| \cdot \|_1$ no necesariamente lo serán respecto de otra $\| \cdot \|_2$.

2.2. Consistencia

Nuestro objetivo será describir las condiciones que garantizen la convergencia del estimador de regresión dado por la ecuación 2.4 a medida que se incremente el tamaño de la muestra. La principal dificultad radica en adaptar en la medida de lo posible la demostración de consistencia para el caso finito dimensional sin tener como herramienta la suposición de una función de densidad que nos permita controlar el sesgo y la varianza. En la demostración para el caso real, se probó la consistencia en forma separada de los elementos del cociente dado por 1.2, y se forzó la aparición en cada término del factor nh,

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

En la demostración de esta proposición, probamos que el numerador tiende $m(x)f_X(x)$, mientras que el denominador a $f_X(x)$. Siendo que en el contexto funcional no contamos con densidades reemplazaremos nh por $n\,\mathbb{E}\left(K\left(\frac{x-X}{h}\right)\right)$.

Notación 2.2.1. Sea $\chi \in (E, d)$ y $h > 0$, llamaremos

$$\varphi_\chi(h) = P(\mathcal{X} \in B(\chi, h)),$$

donde $B(\chi, h)$ es la bola de radio h y centro χ inducida por d.

En lo que sigue, restrigiremos los núcleos a las clases definidas a continuación.

Definición 2.2.2. Sea $K : \mathbb{R} \to \mathbb{R}_{\geq 0}$ un núcleo univariado.

1. Diremos que $K : \mathbb{R} \to \mathbb{R}_{\geq 0}$ es un *núcleo de tipo I* si $\int K = 1$ y existen $0 < C_1 < C_2 < \infty$ tales que
$$C_1 \mathbf{I}_{[0,1]} \leq K \leq C_2 \mathbf{I}_{[0,1]}.$$
2. Diremos que $K : \mathbb{R} \to \mathbb{R}_{\geq 0}$ es un *núcleo de tipo II* si $\int K = 1$, está soportado en $[0,1]$, es acotado, continuo en $x = 1$, derivable en $(0,1)$ y para $-\infty < C_1 < C_2 < 0$, K' se encuentra acotada por
$$C_1 \leq K' \leq C_2.$$

Observemos que la primer clase incluye al núcleo Uniforme asimétrico, mientras que la segunda a los núcleos asimétricos Triangular y Cuadrático.

Lema 2.2.3. *Sea K un núcleo univariado de tipo I. Entonces existen $0 \leq C \leq C' < \infty$ tales que*
$$C\varphi_{\mathcal{X}}(h) \leq \mathbb{E}\left(K\left(\frac{d(\chi, \mathcal{X})}{h}\right)\right) \leq C'\varphi_{\mathcal{X}}(h)$$

Demostración. Como K es un núcleo univariado de tipo I y $\mathbf{I}_{[0,1]}(\frac{d(\chi,\mathcal{X})}{h}) = \mathbf{I}_{B(\chi,h)}(\mathcal{X})$, se satisface que
$$C_1 \mathbf{I}_{B(\chi,h)}(\mathcal{X}) \leq K\left(\frac{d(\chi, \mathcal{X})}{h}\right) \leq C_2 \mathbf{I}_{B(\chi,h)}(\mathcal{X}),$$
tomando esperanza en cada término, y señalando que $\mathbb{E}\left(\mathbf{I}_{B(\chi,h)}\right) = \varphi_{\chi}(h)$ obtenemos que
$$C_1\varphi_{\mathcal{X}}(h) \leq \mathbb{E}\left(K\left(\frac{d(\chi, \mathcal{X})}{h}\right)\right) \leq C_2\varphi_{\mathcal{X}}(h).$$
La igualdad buscada se obtiene tomando $C_1 = C$ y $C_2 = C'$. □

Lema 2.2.4. *Sea K un núcleo univariado de tipo II y supongamos que φ_{χ} satisface la siguiente propiedad*
$$\exists C_3 > 0,\ \exists \epsilon_0 > 0 \text{ tales que para todo } \epsilon < \epsilon_0, \int_0^{\epsilon} \varphi_{\chi}(u)\, du > C_3 \epsilon \varphi_{\chi}(\epsilon).$$
Entonces para h suficientemente pequeño, existen $0 \leq C \leq C' < \infty$ tales que
$$C\varphi_{\chi}(h) \leq \mathbb{E}\left(K\left(\frac{d(\chi, \mathcal{X})}{h}\right)\right) \leq C'\varphi_{\chi}(h).$$

Demostración.

Sea $Y(\omega) = \frac{d(\chi,\mathcal{X})}{h}(\omega)$ variable aleatoria no negativa y $dP_Y(y)$ la medida en $\mathbb{R}$ inducida por Y. Nos resultará útil señalar que
$$\int_0^1 dP_Y(y) = P\left(Y \in (0,1)\right) = P\left(\frac{d(\chi, \mathcal{X})}{h} \leq 1\right) = \varphi_{\chi}(h).$$
Por otro lado,
$$\mathbb{E}\left(K\left(\frac{d(\chi, \mathcal{X})}{h}\right)\right) = \int_{\Omega} K\left(\frac{d(\chi, \mathcal{X})}{h}\right)(\omega)\, dP(\omega) = \int_{\mathbb{R}} K(y)\, dP_Y(y),$$

la última integral vale $\int_0^1 K(y)\,dP_Y(y)$, pues el soporte de K está contenido en $[0,1]$. Además, como K es derivable en ese intervalo, para todo $y \in (0,1)$ vale que $K(y) = K(0) + \int_0^y K'(u)\,du$, lo cual implica

$$\begin{aligned}
\mathbb{E}\left(K\left(\frac{d(\chi,\mathcal{X})}{h}\right)\right) &= \int_0^1 K(0)\,dP_Y(y) + \int_0^1 \left(\int_0^y K'(u)\,du\right) dP_Y(y) && (2.5)\\
&= K(0)\varphi_\chi(h) + \int_0^1 \int_0^1 K'(u) I_{[u,1]}(y)\,du\,dP_Y(y) && (2.6)\\
&= K(0)\varphi_\chi(h) + \int_0^1 K'(u)\left(\int_u^1 dP_Y(y)\right) du && (2.7)\\
&= K(0)\varphi_\chi(h) + \int_0^1 K'(u) P\left(u \le Y \le 1\right)\,du && (2.8)\\
&= K(0)\varphi_\chi(h) + \int_0^1 K'(u)\left(\varphi_\chi(h) - \varphi_\chi(hu)\right)\,du && (2.9)\\
&= K(0)\varphi_\chi(h) + \varphi_\chi(h)\int_0^1 K'(u)\,du - \int_0^1 K'(u)\varphi_\chi(hu)\,du && (2.10)\\
&= K(0)\varphi_\chi(h) + \varphi_\chi(h)\left(K(1) - K(0)\right) - \int_0^1 K'(u)\varphi_\chi(hu)\,du && (2.11)\\
&= -\int_0^1 K'(u)\varphi_\chi(hu)\,du && (2.12)\\
&= -\int_0^h \frac{1}{h} K'\left(\frac{t}{h}\right)\varphi_\chi(t)\,dt && (2.13)
\end{aligned}$$

Observemos que la igualdad (2.7) es verdadera por el Teorema de Fubini, y (2.12) debido a que $K(1) = 0$ al ser un núcleo de tipo II.
Entonces, para $h < \epsilon_0$ acotamos (2.13) usando la hipótesis sobre $\varphi_\chi(t)$, y la cota para K'

$$-\int_0^h \frac{1}{h} K'\left(\frac{t}{h}\right)\varphi_\chi(t)\,dt > -C_1C_3\varphi_\chi(h).$$

Obteniendo finalmente la cota inferior

$$\mathbb{E}\left(K\left(\frac{d(\chi,\mathcal{X})}{h}\right)\right) \ge -C_3C_1\varphi_\chi(h). \tag{2.14}$$

Consideremos $C' = sup_{t\in[0,1]} K(t)$, entonces $\mathbb{E}\left(K\left(\frac{d(\chi,\mathcal{X})}{h}\right)\right) \le C'\varphi_\chi(h)$, pues K es un núcleo

acotado soportado en $[0,1]$. Si $C = -C_3C_1$ hemos demostrado que

$$C\varphi_\chi(h) \leq \mathbb{E}\left(K\left(\frac{d(\chi,\mathcal{X})}{h}\right)\right) \leq C'\varphi_\chi(h).$$

□

Observación 2.2.5. Observemos que para $\vec{x} \in \mathbb{R}^d$ el resultado es válido pidiendo que $f(\vec{x})$ sea positiva y continua en $\vec{x}$. Eso garantiza que $f(\cdot)$ sea positiva en una bola de radio ϵ_0 centrada en $\vec{x}$.

En efecto, si $f(\vec{x}) > 0$ por continuidad, en el intervalo $[0,\epsilon_0]$ alcanzará un mínimo $m > 0$ y un máximo $M > 0$. Por lo tanto, para todo $0 < t \leq \epsilon_0$, valdrá la siguiente desigualdad

$$m\frac{\pi^{\frac{d}{2}}}{\gamma(\frac{d}{2}+1)}t^d < \varphi_{\vec{x}}(t) = \int_{B_d(\vec{x},t)} f(\vec{s})\, ds_1 \ldots ds_d < M\frac{\pi^{\frac{d}{2}}}{\gamma(\frac{d}{2}+1)}t^d,$$

siendo $\frac{\pi^{\frac{d}{2}}}{\gamma(\frac{d}{2}+1)}t^d = |B_d(\vec{x},t)|$.

Sea $\epsilon \leq \epsilon_0$

$$\begin{aligned}
\int_0^\epsilon \varphi_{\vec{x}}(t)\, dt &> \int_0^\epsilon m\frac{\pi^{\frac{d}{2}}}{\gamma(\frac{d}{2}+1)}t^d\, dt && (2.15)\\
&= m\frac{\pi^{\frac{d}{2}}}{\gamma(\frac{d}{2}+1)}\frac{\epsilon^{d+1}}{(d+1)} && (2.16)\\
&> \frac{mM}{(d+1)}\epsilon\varphi_{\vec{x}}(\epsilon), \text{ aquí aplicamos la cota superior para } t=\epsilon. && (2.17)
\end{aligned}$$

2.2.1. Convergencia en probabilidad

En esta sección demostraremos la convergencia en probabilidad del estimador 2.4. Así como en el caso finito dimensional pedimos que $h \to 0$, $n \to \infty$ y el cociente $\frac{1}{nh} \to \infty$, ahora pediremos lo mismo para n y h y la relación entre n y $\varphi_\chi(h)$. Al modelo de regresión dado en el ejemplo **??**, agregaremos hipótesis de continuidad sobre $\mathbb{E}(Y^2|\chi)$.

Teorema 2.2.6. *Sean Y y $r(\chi)$ que cumplen el modelo de regresión continuo dado por la ecuación (2.1). Supongamos que*

C1. *Sea $\chi \in \mathcal{X}$, tal que $\varphi_\chi(h) > 0$ para todo $h > 0$.*

C2. *K es un núcleo de tipo I (o II) y se verifican el Lema 2.2.3 (o 2.2.4).*

C3. *$\lim h = 0$, $\lim \varphi_\chi(h) = 0$ y $\lim \frac{1}{n\varphi_\chi(h)} = 0$ cuando $n \to \infty$.*

C4. *$r(\cdot)$ y $\mathbb{E}\left(Y^2|\mathcal{X} = \cdot\right) = \sigma_2(\cdot)$ son funciones continuas y acotadas en un entorno de χ.*

Entonces,

$$\hat{r}_K(\chi) \xrightarrow{P} r(\chi).$$

Demostración. Llamemos

$$\Delta_i = \frac{K\left(\frac{d(\chi,\mathcal{X}_i)}{h}\right)}{\mathbb{E}\left(K\left(\frac{d(\chi,\mathcal{X})}{h}\right)\right)}.$$

Observemos que si se cumple $C1$ y el lema 2.2.3 (o 2.2.4), $\mathbb{E}\left(K\left(\frac{d(\chi,\mathcal{X})}{h}\right)\right) > 0$ y además, por construcción $\mathbb{E}(\Delta_i) = 1$

Para la demostración será conveniente normalizar $\hat{r}(\chi)$ por el factor $n\,\mathbb{E}\left(K\left(\frac{d(\chi,\mathcal{X})}{h}\right)\right)$.

$$\hat{r}_K(\chi) = \frac{\hat{r}_{K,2}(\chi)}{\hat{r}_{K,1}(\chi)}, \tag{2.18}$$

siendo

$$\hat{r}_{K,1}(\chi) = \frac{1}{n}\sum_{i=1}^{n}\Delta_i,$$

$$\hat{r}_{K,2}(\chi) = \frac{1}{n}\sum_{i=1}^{n}\Delta_i Y_i.$$

Nos resultará útil escribir la distancia entre $\hat{r}_K(\chi)$ y $r(\chi)$ de la siguiente forma:

$$\hat{r}_K(\chi) - r(\chi) = \frac{1}{\hat{r}_{K,1}(\chi)}\left\{(\mathbb{E}(\hat{r}_{K,2}(\chi)) - r(\chi)) + (\hat{r}_{K,2}(\chi) - \mathbb{E}(\hat{r}_{K,2}(\chi))) - r(\chi)(\hat{r}_{K,1}(\chi) - 1)\right\}. \tag{2.19}$$

Bastará con demostrar que cada término del lado derecho tiende en probabilidad a 0. Nos bastará con probar los siguientes resultados:

(i) $\lim_{n\to\infty}\mathbb{E}\left(\hat{r}_{K,2}(\chi)\right) = r(\chi)$.

(ii) $\hat{r}_{K,2}(\chi) - \mathbb{E}\left(\hat{r}_{K,2}(\chi)\right) \xrightarrow{P} 0$.

(iii) $\hat{r}_{K,1}(\chi) \xrightarrow{P} 1$.

Veamos **(i)**:

$$\begin{aligned}
|r(\chi) - \mathbb{E}(\hat{r}_{K,2}(\chi))| &= |r(\chi) - \mathbb{E}\left(Y_1\Delta_1\right)| \\
&= |r(\chi) - \mathbb{E}\left(\mathbb{E}\left(Y_1\Delta_1|\mathcal{X}_1\right)\right)| \\
&= |r(\chi) - \mathbb{E}\left(r(\mathcal{X}_1)\Delta_1\right)| \\
&= |\,\mathbb{E}\left((r(\chi) - r(\mathcal{X}_1))\,\Delta_1\right)| \\
&\leq \mathbb{E}\left(|r(\chi) - r(\mathcal{X}_1)|\Delta_1\right) \text{ observemos que } \Delta_1 \text{ está soportado en } B(\chi,h). \\
&\leq \mathbb{E}\left(\sup_{\chi'\in B(\chi,h)}|r(\chi) - r(\chi')|\Delta_1\right) \\
&= \sup_{\chi'\in B(\chi,h)}|r(\chi) - r(\chi')|\,\mathbb{E}(\Delta_1), \text{ recordemos que } \mathbb{E}(\Delta_1) = 1, \\
&= \sup_{\chi'\in B(\chi,h)}|r(\chi) - r(\chi')| \\
&\leq \epsilon, \text{ para } h \text{ suficientemente pequeño, pues } r(\cdot) \text{ es una función continua en } \chi.
\end{aligned}$$

Por lo tanto, hemos probado que $\lim_{n\to\infty} \mathbb{E}\left(\hat{r}_2(\chi)\right) = r(\chi)$.

Veamos **(ii)**:

Bastará probar que $\mathbb{V}\text{ar}(\hat{r}_{K,2}(\chi)) \to 0$.

$$\begin{aligned}
\mathbb{V}\text{ar}(\hat{r}_{K,2}(\chi)) &= \mathbb{V}\text{ar}\left(\frac{1}{n}\sum_{i=1}^{n}\Delta_i Y_i\right) && (2.20)\\
&= \frac{1}{n}\,\mathbb{V}\text{ar}\left(\Delta_1 Y_1\right) && (2.21)\\
&\leq \frac{1}{n}\,\mathbb{E}\left(\left(\Delta_1 Y_1\right)^2\right) && (2.22)\\
&= \frac{1}{n}\,\mathbb{E}\left(\mathbb{E}\left(\Delta_1^2 Y_1^2 | \mathcal{X}_1\right)\right) && (2.23)\\
&= \frac{1}{n}\,\mathbb{E}\left(\mathbb{E}\left(Y_1^2|\mathcal{X}_1\right)\Delta_1^2\right) && (2.24)\\
&= \frac{1}{n}\,\mathbb{E}\left(\sigma_2(\mathcal{X}_1)\Delta_1^2\right) && (2.25)\\
&\leq \frac{1}{n}\,\mathbb{E}\left(\sup_{\chi'\in B(\chi,h)}|\sigma_2(\chi)|\Delta_1^2\right) && (2.26)\\
&\leq \frac{1}{n}\sup_{\chi'\in B(\chi,h)}|\sigma_2(\chi')|\,\mathbb{E}\left(\Delta_1^2\right). && (2.27)
\end{aligned}$$

En vista de la continuidad de $\sigma_2(.)$, podemos afirmar que $\sup_{\chi'\in B(\chi,h)}|\sigma_2(\chi')| \leq |\sigma_2(\chi)| + 1 = C_\chi < \infty$, para h suficientemente pequeño.

Analizemos $\mathbb{E}(\Delta_1^2)$:

Sea $C = \sup_{x\in[0,1]} K(x)$, entonces $K^2 \leq C^2 I_{[0,1](x)}$. Tomando esperanzas en ambos miembros, se verificará

$$\mathbb{E}\left(K^2\left(\frac{d(\chi,\mathcal{X})}{h}\right)\right) \leq C^2\varphi_\chi(h).$$

Como K es un núcleo de tipo I (o II) en vista del Lema 2.2.3 (o 2.2.4) existirán $\tilde{C}_1, \tilde{C}_2$ tales que

$$\tilde{C}_1^2\varphi_\chi(h)^2 \leq \left(\mathbb{E}\left(K\left(\frac{d(\chi,\mathcal{X})}{h}\right)\right)\right)^2 \leq \tilde{C}_2^2\varphi_\chi(h)^2.$$

Combinando ambas desigualdades, obtenemos que

$$\mathbb{E}(\Delta_1^2) \leq \frac{C^2}{\tilde{C}_1^2\varphi_\chi(h)}.$$

Por lo tanto, podemos acotar 2.27 de la siguiente forma:

$$\mathbb{V}\text{ar}(\hat{r}_2(\chi)) \leq \frac{1}{n}\sup_{\chi'\in B(\chi,h)}|\sigma_2(\chi')|\,\mathbb{E}\left(\Delta_1^2\right) \leq \frac{C^2 C_\chi}{\tilde{C}_1^2}\frac{1}{n\varphi_\chi(h)} \qquad (2.28)$$

En vista de $C3$, para n grande y h pequeño, 2.28 prueba que $\mathbb{V}\text{ar}(\hat{r}_{K,2}(\chi)) \to 0$ y $\hat{r}_{K,2}(\chi) - \mathbb{E}\left(\hat{r}_{K,2}(\chi)\right) \xrightarrow{P} 0$.

Veamos **(iii)**:

Este resultado se deriva de aplicar **(ii)** a la variable constante $Y_i = 1$.

En vista de los resultados (**i**), (**ii**) y (**iii**) hemos probado la convergencia en probabilidad de $\hat{r}_K(\chi)$. □

2.2.2. Convergencia casi completa

El objetivo de esta sección es mostrar la *convergencia casi completa* (*a.co.*) del estimador de regresión.

Definición 2.2.7. Diremos que X_n converge a X en forma *casi completa* si

$$\forall \epsilon > 0,\ \sum_{n\in\mathbb{N}} P\left(|X_n - X| > \epsilon\right) < \infty,$$

y que

$$X_n - X = O_{a.co.}(u_n),\ \text{para } u_n \to 0$$

si y solo si,

$$\exists \epsilon_0 > 0,\ \sum_{n\in\mathbb{N}} P\left(|X_n - X| > \epsilon_0 u_n\right) < \infty.$$

La convergencia *casi completa* es una forma de convergencia más fuerte que la convergencia casi segura. Este resultado y las propiedades que preserva se encuentran detallados en el Apéndice A.

Teorema 2.2.8. *Sean Y y $r(\chi)$ que cumplen el modelo de regresión continuo dado por la ecuación (2.1). Supongamos que*

H1. *Sea $\chi \in \mathcal{X}$, tal que $\varphi_\chi(h) > 0$ para todo $h > 0$.*

H2. *K es un núcleo de tipo I (o II) y se verifica el Lema 2.2.3 (o 2.2.4).*

H3. *$\lim h = 0$, $\lim \varphi_\chi(h) = 0$ y $\lim \frac{\log n}{n\varphi_\chi(h)} = 0$ cuando $n \to \infty$.*

H4. *Existen $h > 0$ y $C > 0$ tales que para todo $m \geq 2$ se verifica*

$$\sup_{\chi' \in B_\chi(h)} \mathbb{E}(|Y|^m | \mathcal{X} = \chi) \leq C^m.$$

Entonces,

$$\lim_{n\to\infty} \hat{r}_K(\chi) = r(\chi)\ \ a.co.$$

Demostración. Para probar este resultado, consideraremos la descomposicion dada en 2.19y demostraremos los siguientes resultados:

(i) $\lim_{n\to\infty} \mathbb{E}\left(\hat{r}_{K,2}(\chi)\right) = r(\chi)$.

(ii) $\hat{r}_{K,1}(\chi) - 1 = O_{a.co.}\left(\sqrt{\frac{\log(n)}{n\varphi_\chi(h)}}\right)$.

(iii) $\hat{r}_{K,2}(\chi) - \mathbb{E}\left(\hat{r}_{K,2}(\chi)\right) = O_{a.co.}\left(\sqrt{\frac{\log(n)}{n\varphi_\chi(h)}}\right)$.

Veamos **(i)**:

Este resultado lo hemos demostrado en el Teorema 2.2.6.

Veamos **(ii)**:

Debemos ver que existe $\epsilon_0 > 0$ que verifique

$$\sum_{n\in\mathbb{N}} P\left(|\hat{r}_{K,1}(\chi) - 1| > \epsilon_0\sqrt{\frac{\log(n)}{n\varphi_\chi(h)}}\right) < \infty, \text{ es decir,}$$

$$\sum_{n\in\mathbb{N}} P\left(\frac{1}{n}\left|\sum_{i=1}^{n}\Delta_i - 1\right| > \epsilon_0\sqrt{\frac{\log(n)}{n\varphi_\chi(h)}}\right) < \infty.$$

Para demostrar la última desigualdad usaremos el siguiente Lema:

Lema 2.2.9. *Sean Z_i va. independientes con media 0 tal que para $m \geq 2$ se cumple que $\mathbb{E}(|Z_i|^m) \leq C_m H^{2(m-1)}$, donde $C_m \leq \frac{m!}{2}C^2$, $H > 0$ y $C > 0$. Entonces, para todo $\epsilon > 0$,*

$$P\left(\left|\sum_{i=1}^{n} Z_i\right| > n\epsilon\right) \leq 2\exp\left(\frac{-\epsilon^2 n}{2H^2(C^2+\epsilon)}\right).$$

El Lema es un Corolario inmediato de la Desigualdad de Bernstein y se encuentra demostrado en el Apéndice. Tomando $Z_i = \Delta_i - 1$, acotaremos adecuadamente $\mathbb{E}(|Z_i^m|)$ y veremos que existen $C > 0$ y $H > 0$ de forma tal que se satisfagan las condiciones del Lema.

Recordemos que si K es un núcleo de tipo I (o II), como procedimos anteriormente podemos acotar superiormente K. En consecuencia existirá $C_2 > 0$ tal que

$$K^m\left(\frac{d(\chi,\mathcal{X}_1)}{h}\right) \leq C_2^m I_{[0,1]}\left(\frac{d(\chi,\mathcal{X}_1)}{h}\right).$$

Tomando esperanza a ambos lados obtenemos que

$$\mathbb{E}\left(K^m\left(\frac{d(\chi,\mathcal{X}_1)}{h}\right)\right) \leq C_2^m\varphi_\chi(h).$$

Por otro lado, por el Lema 2.2.3 (o 2.2.4), exististirá C_1 tal que

$$(C_1\varphi_\chi(h))^m \leq \left(\mathbb{E}\left(K\left(\frac{d(\chi,\mathcal{X}_1)}{h}\right)\right)\right)^m$$

Combinando ambas desigualdades, obtenemos que para todo $m \geq 2$

$$\mathbb{E}(\Delta_1^m) \leq \left(\frac{C_1}{C_2}\right)^m \varphi_\chi(h)^{-m+1}. \tag{2.29}$$

Miremos $\mathbb{E}(|Z_i|^m) = \mathbb{E}(|\Delta_1 - 1|^m)$:

$$\begin{aligned}\mathbb{E}(|\Delta_1 - 1|^m) &= \mathbb{E}\left(\left|\sum_{j=0}^{m}\binom{m}{j}\Delta_1^j(-1)^{m-j}\right|\right)\\ &\leq \sum_{j=0}^{m}\binom{m}{j}\mathbb{E}\left(\Delta_1^j\right)\\ &\leq \sum_{j=0}^{m}\binom{m}{j}\left(\frac{C_2}{C_1}\right)^j\varphi_\chi(h)^{-j+1}, \text{ la desigualdad es válida por (2.29)}\end{aligned}$$

En vista de que $\varphi_\chi(h) \to 0$, para h pequeño y $j < m$, $\varphi_\chi(h)^{-j+1} \leq \varphi_\chi(h)^{-m+1}$. Entonces

$$\mathbb{E}(|\Delta_1 - 1|^m) \leq \left(\sum_{j=0}^{m}\binom{m}{j}\left(\frac{C_2}{C_1}\right)^j\right)\varphi_\chi(h)^{-m+1} \tag{2.30}$$

$$= \left(\frac{C_2}{C_1}+1\right)^m\varphi_\chi(h)^{-m+1} \tag{2.31}$$

Estamos en condiciones de aplicar el Lema 2.2.9, considerando $C_m = \left(\frac{C_2}{C_1}+1\right)^m$ y $H^2 = \varphi_\chi(h)^{-1}$. Notemos que se cumplen las hipótesis del Lema, pues para m grande $\left(\frac{C_2}{C_1}+1\right)^m \leq \frac{m!}{2}$ y podemos tomar C suficientemente grande de manera que tal que la desigualdad tambien valga para los primeros m términos. Entonces,

$$P\left(\frac{1}{n}\left|\sum_{i=1}^{n}\Delta_i - 1\right| > \epsilon\right) \leq 2\exp\left(\frac{-\epsilon^2 n}{2\varphi_\chi(h)^{-1}(C^2+\epsilon)}\right).$$

Como el resultado anterior vale para todo $\epsilon > 0$, consideremos $\epsilon_n = \epsilon_0\sqrt{\frac{\log n}{n\varphi_\chi(h)}}$,

$$P\left(\frac{1}{n}\left|\sum_{i=1}^{n}\Delta_i - 1\right| > \epsilon_0\sqrt{\frac{\log n}{n\varphi_\chi(h)}}\right) \leq 2\exp\left(\frac{-\left(\epsilon_0\sqrt{\frac{\log n}{n\varphi_\chi(h)}}\right)^2 n}{2\varphi_\chi(h)^{-1}\left(C^2+\epsilon_0\sqrt{\frac{\log n}{n\varphi_\chi(h)}}\right)}\right) \tag{2.32}$$

$$= 2\left(\frac{1}{n}\right)^{\frac{\epsilon_0^2}{2\left(C^2+\epsilon_0\sqrt{\frac{\log n}{n\varphi_\chi(h)}}\right)}} \tag{2.33}$$

Debido a que por $H3$, $\sqrt{\frac{\log n}{n\varphi_\chi(h)}} \to 0$, para $n \geq n_0$ se cumplirá $\sqrt{\frac{\log n}{n\varphi_\chi(h)}} < 1$. Por lo tanto, podemos elegir ϵ_0 de forma tal que

$$\frac{\epsilon_0^2}{2\left(C^2+\sqrt{\frac{\log n}{n\varphi_\chi(h)}}\right)} \geq \frac{\epsilon_0^2}{2(C^2+1)} = \alpha > 1.$$

Finalmente, obtenemos que para $n \geq n_0$

$$P\left(\frac{1}{n}\sum_{i=1}^{n}|\Delta_i - 1| > \epsilon_0\sqrt{\frac{\log n}{n\varphi_\chi(h)}}\right) \leq 2\left(\frac{1}{n}\right)^{\frac{\epsilon_0^2}{2\left(C^2+\epsilon_0\sqrt{\frac{\log n}{n\varphi_\chi(h)}}\right)}} \leq 2\left(\frac{1}{n}\right)^{\frac{\epsilon_0^2}{2(C^2+1)}} \leq \frac{2}{n^\alpha}$$

Por lo tanto,

$$\sum_{n\in\mathbb{N}} P\left(\frac{1}{n}\left|\sum_{i=1}^{n}\Delta_i - 1\right| > \epsilon_0\sqrt{\frac{\log n}{n\varphi_\chi(h)}}\right) \leq \sum_{n<n_0} P\left(\frac{1}{n}\left|\sum_{i=1}^{n}\Delta_i - 1\right| > \epsilon_0\sqrt{\frac{\log n}{n\varphi_\chi(h)}}\right) + \sum_{n\geq n_0}\frac{2}{n^\alpha} < \infty$$

Con lo cual, hemos probado que $\hat{r}_{K,1}(\chi) - 1 = O_{a.co.}\left(\sqrt{\frac{\log n}{n\varphi_\chi(h)}}\right)$.

Veamos **(iii)**:

Ahora buscamos la convergencia casi completa de $\hat{r}_{K,2}(\chi)$. Veremos que podremos aplicar nuevamente el Lema 2.2.9, tomando ahora $Z_i = Y_i\Delta_i - \mathbb{E}(Y_1\Delta_1)$ y nuevamente $H^2 = \varphi_\chi(h)^{-1}$.

Debemos acotar $\mathbb{E}\left(|Y_1\Delta_1 - \mathbb{E}(Y_1\Delta_1)|^m\right)$

$$|\left(Y_1\Delta_1 - \mathbb{E}(Y_1\Delta_1)\right)^m| = \left|\sum_{j=0}^{m}\binom{m}{j}\left(Y_1\Delta_1\right)^j\left(\mathbb{E}(Y_1\Delta_1)\right)^{m-j}\right| \tag{2.34}$$

$$\mathbb{E}\left(|Y_1\Delta_1 - \mathbb{E}(Y_1\Delta_1)|^m\right) \leq \sum_{j=0}^{m}\binom{m}{j}\mathbb{E}(|Y_1\Delta_1^j|)|\,\mathbb{E}(Y_1\Delta_1)|^{m-j} \tag{2.35}$$

Sabemos de la ecuación 2.29 que para todo $j \geq 2$, $\mathbb{E}(\Delta_1^j) \leq \left(\frac{C_2}{C_1}\right)^j\varphi_\chi(h)^{-j+1}$. Por otro lado, dado el soporte compacto de K

$$\mathbb{E}(|Y_1^j\Delta_1^j|) \leq \sup_{\chi'\in B(\chi,h)}\sigma_j(\chi')\,\mathbb{E}(\Delta_1^j).$$

Las desigualdades anteriores nos permiten acotar (2.35) de la siguiente forma

$$\mathbb{E}\left(|Y_1\Delta_1 - \mathbb{E}(Y_1\Delta_1)|^m\right) \leq \sum_{j=0}^{m}\binom{m}{j}\sup_{\chi'\in B(\chi,h)}\sigma_j(\chi')\left(\frac{C_2}{C_1}\right)^j\varphi_\chi(h)^{-j+1}|\,\mathbb{E}(Y_1\Delta_1)|^{m-j}.$$

Miremos $|\,\mathbb{E}(Y_1\Delta_1)|^{m-j}$:

$$|\,\mathbb{E}(Y_1\Delta_1)|^{m-j} \leq \sup_{\chi'\in B(\chi,h)}|r(\chi')\,\mathbb{E}(\Delta_1)|^{m-j} \tag{2.36}$$

$$\leq \sup_{\chi'\in B(\chi,h)}|r(\chi')|^{m-j},\ \text{pues } \mathbb{E}(\Delta_1) = 1. \tag{2.37}$$

Volviendo a la desigualdad (2.35),

$$\begin{aligned}
\mathbb{E}\left(|Y_1\Delta_1 - \mathbb{E}(\Delta_1Y_1)|^m\right) &\leq \sum_{j=0}^{m}\binom{m}{j}\sup_{\chi'\in B(\chi,h)}\sigma_j(\chi')\left(\frac{C_2}{C_1}\right)^j\varphi_\chi(h)^{-j+1}\sup_{\chi'\in B(\chi,h)}|r(\chi')|^{m-j} \\
&\leq \left\{\sum_{j=0}^{m}\binom{m}{j}\sup_{\chi'\in B(\chi,h)}\sigma_j(\chi')\left(\frac{C_2}{C_1}\right)^j\sup_{\chi'\in B(\chi,h)}|r(\chi')|^{m-j}\right\}\varphi_\chi(h)^{-m+1} \\
&\leq \sup_{\chi'\in B(\chi,h),0\leq j\leq m}\sigma_j(\chi')\left(\frac{C_2}{C_1} + \sup_{\chi'\in B(\chi,h)}|r(\chi')|\right)^m\varphi_\chi(h)^{-m+1}
\end{aligned}$$

Para todo $\chi' \in B(\chi, h)$ se cumple

$$\begin{aligned}\sigma_j(\chi') &= \mathbb{E}(|Y^j|I_{Y\leq 1}|\mathcal{X}=\chi') + \mathbb{E}(|Y^j|I_{Y>1}|\mathcal{X}=\chi') \\ &\leq 1 + \mathbb{E}(|Y^m|I_{Y>1}|\mathcal{X}=\chi'), \text{ para } j \leq m \\ &\leq 1 + C^m, \text{ esto es válido en vista de } H\mathcal{3}, \\ &\leq (1+C)^m\end{aligned}$$

Podemos acotar $\sup_{\chi'\in B(\chi,h),0\leq j\leq m}\sigma_j(\chi')$ obteniendo

$$\sup_{\chi'\in B(\chi,h),0\leq j\leq m}|\sigma_j(\chi')| \leq (1+C)^m.$$

En vista de $H4$, podemos acotar $C_m = \sup_{\chi'\in B(\chi,h),0\leq j\leq m}\sigma_j(\chi')\left(\frac{C_2}{C_1} + \sup_{\chi'\in B(\chi,h)}|r(\chi')|\right)^m$ de la siguiente manera,

$$C_m \leq (1+C)^m\left(\frac{C_2}{C_1} + \sup_{\chi'\in B(\chi,h)}|r(\chi')|\right)^m \leq \frac{m!}{2}C'$$

donde C' es elegida de forma tal que los para todo $m \in \mathbb{N}$ se cumpla

$$(1+C)^m\left(\frac{C_2}{C_1} + \sup_{\chi'\in B(\chi,h)}|r(\chi')|\right)^m \leq C'\frac{m!}{2}.$$

Estamos en condiciones de aplicar el Lema 2.2.9, obteniendo que para $n \geq n_0$ existe $\alpha > 1$ tal que

$$P\left(\frac{1}{n}\left|\sum_{i=1}^n Y_i\Delta_i - \mathbb{E}(Y_1\Delta_1)\right| \geq \epsilon_0\sqrt{\frac{\log n}{n\varphi_\chi(h)}}\right) \leq 2\exp\left(\frac{-\left(\epsilon_0\sqrt{\frac{\log n}{n\varphi_\chi(h)}}\right)^2 n}{2\varphi_\chi(h)^{-1}\left(C'^2 + \epsilon_0\sqrt{\frac{\log n}{n\varphi_\chi(h)}}\right)}\right) \leq \frac{2}{n^\alpha}$$

Finalmente, la serie quedará acotada de la siguiente forma

$$\begin{aligned}\sum_{n=1\in\mathbb{N}} P\left(\frac{1}{n}\left|\sum_{i=1}^n Y_i\Delta_i - \mathbb{E}(Y_1\Delta_1)\right| \geq \epsilon_0\sqrt{\frac{\log n}{n\varphi_\chi(h)}}\right) &\leq \sum_{n<n_0} P\left(\frac{1}{n}\left|\sum_{i=1}^n Y_i\Delta_i - \mathbb{E}(Y_1\Delta_1)\right| \geq \epsilon_0\sqrt{\frac{\log n}{n\varphi_\chi(h)}}\right) \\ &\quad + \sum_{n\geq n_0}\frac{2}{n^\alpha} \\ &< \infty.\end{aligned}$$

Hemos probado entonces **(iii)**.

En vista de **(i)**, **(ii)** y **(iii)**, hemos probado la convergencia casi completa de $\hat{r}_K(\chi)$. □

Observación 2.2.10. Este resultado también puede ser aplicado para el caso de covariables finito dimensionales desarrollado en el Capítulo 1. Mas aun, con esta propuesta hemos probado la convergencia casi completa del estimador de regresión- que incluye a la convergencia casi segura y en probabilidad- para núcleos de tipo I o II y pidiendo que los momentos condicionales $\mathbb{E}(|Y|^m|\vec{X})$ se encuentren localmente controlados por una cota exponencial.

Por otra parte, podemos aplicar este resultado a la estimación de la densidad $f(\vec{x})$. Teniendo en cuenta que

$$\begin{aligned}
\hat{f}_K(\vec{x}) &= \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{||\vec{x}-\vec{X}_i||}{h}\right) && (2.38)\\
&= \mathbb{E}\left(\frac{1}{h^d}K\left(\frac{||\vec{x}-\vec{X}_i||}{h}\right)\right)\frac{1}{n}\sum_{i=1}^{n}\frac{K\left(\frac{||\vec{x}-\vec{X}_i||}{h}\right)}{\mathbb{E}\left(K\left(\frac{||\vec{x}-\vec{X}_i||}{h}\right)\right)} && (2.39)\\
&= \mathbb{E}\left(\frac{1}{h^d}K\left(\frac{||\vec{x}-\vec{X}_i||}{h}\right)\right)\frac{1}{n}\sum_{i=1}^{n}\Delta_i && (2.40)\\
&= \mathbb{E}\left(\frac{1}{h^d}K\left(\frac{||\vec{x}-\vec{X}||}{h}\right)\right)\hat{r}_{K,1}(\vec{x}). && (2.41)
\end{aligned}$$

Bajo las hipótesis del Teorema,

- $\mathbb{E}\left(\frac{1}{h^d}K\left(\frac{||\vec{x}-\vec{X}||}{h}\right)\right) \to f(\vec{x})$.
- $\lim \hat{r}_{K,1} \xrightarrow{a.co.} 1$

Entonces, $\hat{f}_K(\vec{x}) \xrightarrow{a.co.} f(\vec{x})$.

Capítulo 3

Aplicaciones

3.1. Modelo MAR

Al realizar un análisis estadístico, puede ocurrir que algunos datos no sean observados. La presencia de valores perdidos no puede ser ignorado en el análisis, debido a que puede tener repercusiones graves que van desde la pérdida de potencia del estudio hasta la aparición de sesgos inaceptables. En algunos casos, la razón por la cual faltan puede estar relacionada con la respuesta y en otros no. Sea Y_i la variable respuesta de interés, consideremos la variable dicotómica A_i definida de la siguiente forma

$$A_i = \begin{cases} 1 & \text{si la variable } Y_i \text{ es observada} \\ 0 & \text{si la variable } Y_i \text{ no es observada.} \end{cases}$$

Decimos que los datos *faltan completamente al azar (Missing Completely At Random MCAR)*, cuando la ausencia de la observacion Y_i está determinada por un mecanismo ajeno a la variable de estudio. Un ejemplo estaría dado en un análisis del ingreso medio de la población, si las personas que no nos proporcionan su salario tirasen una moneda y en función del resultado decidiesen proporcionar información o no. En este contexto, la estimación del ingreso medio no se vería afectada, pues el grupo que proporciona información resulta representativo. Sin embargo, si la ausencia de información es mayor en la población con altos ingresos, la falta de información no sería azarosa y la estimación del salario medio se vería sesgado negativamente por este factor.

Bajo el modelo MCAR las variables Y y A son independientes, por lo cual

$$Y_i \sim Y_i | A_i = 1.$$

En consecuencia

$$\mu = \mathbb{E}(Y_i) = \mathbb{E}(Y_i | A_i = 1), \tag{3.1}$$

Esta relación sugiere que un estimador consistente para la esperanza de Y resultará de promediar las variables Y_i observadas, es decir,

$$\hat{\mu}_{Obs} = \frac{\sum_{i=1}^{n} A_i Y_i}{\sum_{i=1}^{n} A_i}. \tag{3.2}$$

Por la Ley de los grandes Números, $\hat{\mu}_{Obs} \xrightarrow{P} \mathbb{E}(Y_i | A_i = 1) = \mathbb{E}(Y_i)$. Observemos que si no se cumple la relación dada por (3.1), este estimador podría no ser consistente, lo cual explica por qué estimar

el ingreso medio con la información observada cuando no estamos en el modelo MCAR no sería una buena decisión.

A menudo los datos no están perdidos completamente al azar, pero la ausencia puede ser clasificada como *desaparecida al azar (Missing At Randon MAR)*. La perdida de datos es MAR cuando el mecanismo que produce la ausencia de una observación es independiente de la variable de estudio, conocidas ciertas variables explicativas que llamaremos $\mathcal{X}$ y son observadas siempre. En principio, $\mathcal{X}$ puede pertencer a cualquier espacio abstracto, finito o infinito dimensional. Un ejemplo de MAR estaría dado si la pérdida de valores en la variable ingreso dependiese de la edad del sujeto, pero dentro de cada clase la razón de ausencia de información no se encontrase relacionada con el mismo. Formalmente, decimos que los datos se pierden al azar cuando

$$(A, Y)|\mathcal{X} \text{ son independientes.}$$

En consecuencia se cumplirá la siguiente relación

$$Y|\mathcal{X} \sim Y|(\mathcal{X}, A = 1).$$

Si llamamos $r(\mathcal{X}_i) = \mathbb{E}(Y_i|\mathcal{X}_i)$, podemos deducir que $r(\mathcal{X}_i)$ coincide con $\mathbb{E}(Y_i|\mathcal{X}_i, A_i = 1)$, dando origen a la siguiente igualdad

$$\mu = \mathbb{E}(Y) = \mathbb{E}(r(\mathcal{X})) = \mathbb{E}(\mathbb{E}(Y|(\mathcal{X}, A = 1))). \tag{3.3}$$

De (3.3) podemos motivar un estimador de μ, promediando los valores predichos de cada observación mediante un estimador de función de regresión formada por aquellos pares $(\mathcal{X}_i, Y_i, A_i = 1)$. En nuestro caso, $\hat{r}_K(\mathcal{X})$ será el estimador de *leave-one-out* de Naradaya-Watson y $\mathcal{X}$ pertenecerá a un espacio funcional (E, d), dando origen a

$$\hat{\mu}_{Reg} = \frac{1}{n}\sum_{i=1}^{n} \hat{r}_{K,-i}(\mathcal{X}_i) \tag{3.4}$$

Recordemos que $\hat{r}_{K,-i}(\mathcal{X}_i)$ fue introducido en el capítulo 1 y consiste en predecir Y_i excluyendo de la función de regresión el par $(\mathcal{X}_i, Y_i)$. El motivo por el cual se quita dicho par en el estimador es evitar que el mismo tienda a interpolar a los datos. Entonces, $\hat{r}_{-i}(\mathcal{X}_i)$ quedará determinado de la siguiente forma

$$\hat{r}_{K,-i}(\mathcal{X}_i) = \frac{\sum\limits_{j\neq i}^{n} A_j Y_j K\left(\frac{d(\mathcal{X}_i,\mathcal{X}_j)}{h}\right)}{\sum\limits_{j\neq i}^{n} A_j K\left(\frac{d(\mathcal{X}_i,\mathcal{X}_j)}{h}\right)},$$

donde K es un núcleo univariado y $h > 0$ es el parámetro de suavizado.

Otra alternativa propuesta por Cheng [15] para $\mathcal{X} \in \mathbb{R}^d$, es reemplazar sólo a las respuestas faltantes por sus valores predichos y hacer un promedio mixto entre los valores predichos y las respuestas observadas.

Es fácil ver que bajo MAR, $\mathbb{E}(AY + (1 - A)r(\mathcal{X})) = \mu$:

$$\begin{aligned}
\mathbb{E}\left(AY + (1 - A)r(\mathcal{X})\right) &= \mathbb{E}\left(\mathbb{E}\left(AY + (1 - A)r(\mathcal{X})|\mathcal{X}\right)\right) \\
&= \mathbb{E}\left(\mathbb{E}(A|\mathcal{X})\,\mathbb{E}(Y|\mathcal{X}) + \mathbb{E}(1 - A|\mathcal{X})r(\mathcal{X})\right) \\
&= \mathbb{E}\left(\mathbb{E}(A|\mathcal{X})r(\mathcal{X}) + \mathbb{E}(1 - A|\mathcal{X})r(\mathcal{X})\right) \\
&= \mathbb{E}(r(\mathcal{X})) \\
&= \mu
\end{aligned}$$

Esto motiva proponer el siguiente estimador

$$\hat{\mu}_{Mix} = \frac{1}{n}\sum_{i=1}^{n}\left(A_i Y_i + (1-A_i)\hat{r}_{K,-i}(\mathcal{X}_i)\right). \tag{3.5}$$

La familia de estimadores de Hotvitz-Thompson se basan en la siguiente relación que se verifica bajo el modelo MAR

$$\mathbb{E}(Y) = \mathbb{E}\left(\frac{AY}{\pi(\mathcal{X})}\right), \text{ siendo } \pi(\mathcal{X}) = P(A=1|\mathcal{X}). \tag{3.6}$$

En efecto,

$$\begin{aligned}
\mathbb{E}\left(\frac{AY}{\pi(\mathcal{X})}\right) &= \mathbb{E}\left(\mathbb{E}\left(\frac{AY}{\pi(\mathcal{X})}\middle|\mathcal{X}\right)\right) && (3.7)\\
&= \mathbb{E}\left(\frac{1}{\pi(\mathcal{X})}\,\mathbb{E}(AY|\mathcal{X})\right), \text{ bajo el modelo MAR,} && (3.8)\\
&= \mathbb{E}\left(\frac{1}{\pi(\mathcal{X})}\,\mathbb{E}(A|\mathcal{X})\,\mathbb{E}(Y|\mathcal{X})\right) && (3.9)\\
&= \mathbb{E}\left(\mathbb{E}(Y|\mathcal{X})\right) && (3.10)\\
&= \mathbb{E}(Y) && (3.11)
\end{aligned}$$

La ecuación (3.10) es válida, debido a que siendo A una variable binaria, $\mathbb{E}(A|\mathcal{X}) = P(A=1|\mathcal{X})$. Como hipótesis adicional, supondremos que $\pi(\mathcal{X}) > 0$ para todo $\mathcal{X} \in E$.

Entonces de (3.6) podemos deducir otro estimador para μ, promediando los valores $\frac{AY}{\pi(\mathcal{X})}$ observados. Siendo $\pi(\mathcal{X})$ desconocida será aproximado por el estimador *leave-one-out* de Naradaya-Watson, $\hat{\pi}_{K,-i}(\mathcal{X}_i)$

$$\hat{\pi}_{-i}(\mathcal{X}_i) = \frac{\sum_{j\neq i}^{n} A_j K\left(\frac{d(\mathcal{X}_i,\mathcal{X}_j)}{h}\right)}{\sum_{j\neq i}^{n} K\left(\frac{d(\mathcal{X}_i,\mathcal{X}_j)}{h}\right)}.$$

dando origen al siguiente estimador

$$\hat{\mu}_{HT} = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i A_i}{\hat{\pi}_{-i}(\mathcal{X}_i)}, \tag{3.12}$$

Ferraty, Sued y Vieu [16] prueban que

$$\sqrt{n}(\hat{\mu}-\mu) = O_p(1),$$

donde $\hat{\mu} = \hat{\mu}_{Reg}, \hat{\mu}_{Mix}$ o $\hat{\mu}_{HT}$ y $O_p(1)$ significará que la sucesión está acotada en probabilidad. El resultado anterior implica la convergencia en probabilidad de cada estimador.

3.2. Estudio de simulación

En esta sección mostraremos los resultados de un estudio de simulación con el fin de evaluar los estimadores $\hat{\mu}_{Reg}$, $\hat{\mu}_{Mix}$ y $\hat{\mu}_{HT}$ en el caso de covariables funcionales. Los datos fueron generados bajo el mismo escenario que Ferraty *et al.* [16]. También computaremos $\hat{\mu}_{Obs}$ para comparar los resultados obtenidos cuando no se tiene en cuenta las respuestas faltantes.

3.2.1. El modelo

- La covariable funcional $\mathcal{X}$ es una función definida en $[0, \pi]$, con la siguiente estructura

$$\mathcal{X}_i(t) = Z_i \cos(2t),$$

donde $i = 1, \cdots, 200$ y Z_i resultan v.a.i.i.d distribuídas de acuerdo a

$$Z_i \sim \begin{cases} F_1 \sim N(10, 36) & \text{con probabilidad } 0.3 \\ F_2 \sim N(0, 36) & \text{con probabilidad } 0.7, \end{cases}$$

y por cada curva son registrados 100 valores, $\{\mathcal{X}_i(t_j)\}_{t_1,\cdots,t_{100}}$ donde $t_1 = 0, \cdots, t_j = \frac{(j-1)}{99}\pi, \cdots, t_{100} = \pi$.

- Las respuestas son generadas mediante el modelo de regresión

$$Y_i = r(\mathcal{X}_i) + \epsilon_i,$$

donde $\epsilon_i \sim N(0, 0.05)$ independientes de $\mathcal{X}_i$. La función de regresión estará dada mediante el funcional

$$r(\mathcal{X}_i) = \frac{2}{\pi} \int_0^\pi \mathcal{X}_i^2(t)\, dt.$$

Es fácil ver que $r(\mathcal{X}_i) = Z_i^2$ quedando determinada $Y_i = Z_i^2 + \epsilon_i$. De esta forma, $\mathbb{E}(Y_i) = \mathbb{E}(Z_i^2)$. Para calcular $\mathbb{E}(Z_i^2)$ nos resultará conveniente condicionar a Z_i respecto de cada $F_{j=1,2}$,

$$\begin{aligned} \mathbb{E}(Z_i^2) &= \mathbb{V}\text{ar}(Z_i) + \mathbb{E}(Z_i)^2 && (3.13) \\ &= \mathbb{E}(\mathbb{V}\text{ar}(Z_i|Z_i \sim F_j)) + \mathbb{V}\text{ar}(\mathbb{E}(Z_i|Z_i \sim F_j)) + (\mathbb{E}(\mathbb{E}(Z_i|Z_i \sim F_j)))^2 && (3.14) \end{aligned}$$

Calculemos $\mathbb{E}(Z_i|F_j)$:

$$\mathbb{E}(Z_i|F_j) = \begin{cases} 10 & \text{si } Z_i \sim F_1, \text{ con probabilidad } 0.3 \\ 0 & \text{si } Z_i \sim F_2 \text{ con probabilidad } 0.7 \end{cases}$$

De esta forma, $\mathbb{E}(\mathbb{E}(Zi|F_j) = 0.3 * 10 = 3$ y $\mathbb{V}\text{ar}(\mathbb{E}(Z_i|F_j)) = 0.3 * 10^2 - (0.3 * 10)^2 = 21$. Por otra parte, $\mathbb{E}(\mathbb{V}\text{ar}(Z_i|F_j)) = 36$. Finalmente, $\mathbb{E}(Y_i)$ estará dada por

$$\begin{aligned} \mathbb{E}(Y_i) &= \mathbb{E}(Z_i^2) && (3.15) \\ &= 36 + 21 + 9 && (3.16) \\ &= 66 && (3.17) \end{aligned}$$

- El mecanismo de ausencia de respuestas dependerá de $\mathcal{X}$ mediante el esquema

$$P(A_i = 1|\mathcal{X}_i) = 1 - \frac{1}{1 + \exp(\alpha r(\mathcal{X}_i))},$$

donde $\alpha > 0$.
La variable A_i fue generada mediante el siguiente procedimiento: se generan aleatoriamente una variable U_i con distribución uniforme en el intervalo $[0, 1]$. Si $U_i < P(A_i = 1|\mathcal{X}_i)$, se designa $A_i = 1$, caso contrario $A_i = 0$. Posteriormente, de acuerdo al valor de A_i de decide descartar o no Y_i del conjunto de observaciones. Notemos que el procedimiento anterior asegura estar en condiciones $MAR.$, pues fijada $\mathcal{X}_i$, se obtienen Z_i^2, $Y_i = f(Z_i^2, \epsilon_i)$ y $A_i = g(Z_i^2, U_i)$. De la independencia entre ϵ_i y U_i se concluye la independecia condicional entre Y_i y A_i. El paramétro α controla el grado de dependencia entre $\mathcal{X}_i$ y A_i. En particular para $\alpha = 0$, resultan independientes, mientras que para valores altos tiende a aumentar la presencia de las respuestas. Por otra parte, observemos que este mecanismo censura principalmente a las curvas con menor amplitud provenientes de F_2.

3.2.2. Implementación

Las estimadores $\hat{\mu}_{Reg}$, $\hat{\mu}_{Mix}$ y $\hat{\mu}_{HT}$ fueron implementados en el entorno **R** mediante la función $funopare.cv$, desarrolada por Ferraty y Vieu. El código se encuentra disponible en *http://www.lsp.ups-tlse.fr/staph/npfda*. Los argumentos elegidos e implementados internamente por dicha función fueron los siguientes:

- El núcleo elegido fue el cuadrático asimétrico K_{CA}.
- El parámetro de suavizado fue obtenido minimizando la pérdida de convalidación cruzada a lo largo de una serie de bandas, tal como fue desarrollada en el capítulo 1.
- La distancia entre las funciones implementada fue la norma $L^2[0,\pi]$, aproximada por la distancia en esa norma de las primeras 11 proyecciones trigonométricas de la serie de Fourier ($p_{11}(.)$) desarrollada en el intervalo $[0,\pi]$.

  La representación de Fourier [17] para $\mathcal{X}(t) \in L^2[0,\pi]$ se encuentra dada por la siguiente expresión

$$\mathcal{X}(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n cos(2nt) + b_n sen(2nt),$$

  donde

$$a_n = \frac{2}{\pi} \int_0^{\pi} cos(2nt) f(t)\, dt,\ n = 0, 1, 2, \cdots,$$

$$b_n = \frac{2}{\pi} \int_0^{\pi} sen(2nt) f(t)\, dt,\ n = 1, 2, \cdots.$$

  La proyección de las primeras N funciones ($p_N(\mathcal{X})$) consiste en truncar la serie hasta $N \in \mathbb{N}$ fijo, es decir,

$$p_N(\mathcal{X}) = \frac{a_0}{2} + \sum_{n=1}^{N} a_n cos(2nt) + b_n sen(2nt).$$

  De modo tal que se $d(\mathcal{X}_i, \mathcal{X}_j)$ queda determinada de la siguiente forma

$$d(\mathcal{X}_i, \mathcal{X}_j) = \int_0^{\pi} (p_{11}(\mathcal{X}_i)(t) - p_{11}(\mathcal{X}_j(t)))^2\, dt.$$

3.2.3. Resultados y conclusiones

Resultados

Se implementaron los estimadores $\hat{\mu}_{Reg}$, $\hat{\mu}_{Mix}$, $\hat{\mu}_{HT}$ y $\hat{\mu}_{Obs}$ para una muestra de tamaño 200 y diferentes valores de α. Por cada experimento, se realizaron 100 replicaciones independientes de las cuales se obtuvieron un valor medio para cada estimador

$$\hat{\mu} = \frac{1}{200} \sum_{i=1}^{200} \hat{\mu}_i$$

y el Error Cuadrático Medio (MSE) dado por

$$MSE(\hat{\mu}) = \frac{1}{100} \sum_{i=1}^{100} (\hat{\mu}_i - 66)^2\,,$$

α	$\bar{A}$	$\hat{\mu}_{Reg}$	$\hat{\mu}_{Mix}$	$\hat{\mu}_{HT}$	$\hat{\mu}_{Obs}$
0	0.48	62.07 (66.32)	64.06 (49.96)	69.76 (214.05)	66.52 (86.25)
0.05	0.78	64.70 (44.93)	67.40 (47.39)	67.80 (48.24)	82.32 (333.06)
0.1	0.83	62.92 (47.95)	66.11 (39.12)	66.30 (39.39)	77.13 (176.12)
0.2	0.88	63.11 (51.62)	66.22 (44.03)	66.26 (44.07)	74.11 (116.29)
0.4	0.91	63.39 (55.61)	66.26 (51.79)	66.27 (52.00)	71.89 (94.32)
0.8	0.94	63.00 (56.54)	66.02 (47.22)	66.03(47.25)	69.92 (71.61)
1	0.94	62.99 (55.49)	66.40 (39.07)	66.40 (39.01)	69.89 (57.07)

Tabla 3.1: Valores medios (MSE) obtenidos para las 100 replicaciones.

donde $\hat{\mu}_i$ es la estimación obtenida en la replicación i-ésima y $\hat{\mu}$=$\hat{\mu}_{Reg}$, $\hat{\mu}_{Obs}$, $\hat{\mu}_{HT}$ o $\hat{\mu}_{Obs}$. También se computaron los valores medios de A. Los resultados son presentados en la Tabla 3.2.3. En la Figura 3.1 se presenta el Boxplot de los estimadores para $\alpha = 0.8$.

Conclusiones

- En casi todos los niveles, a excepción de $\alpha = 0$ en donde hay independencia entre el mecanismo de censura y las covariables, $\hat{\mu}_{Obs}$ que tiende a sobreestimar a μ. Esto se debe a que el mecanismo censura principalmente a las respuestas provenientes de F_2, y en consecuencia promedia amplitudes más grandes.

- $\hat{\mu}_{Reg}$ tiende a subestimar a μ. En la mayoría de los niveles, se obtienen los mejores resultados con $\hat{\mu}_{Mix}$ y $\hat{\mu}_{HT}$.

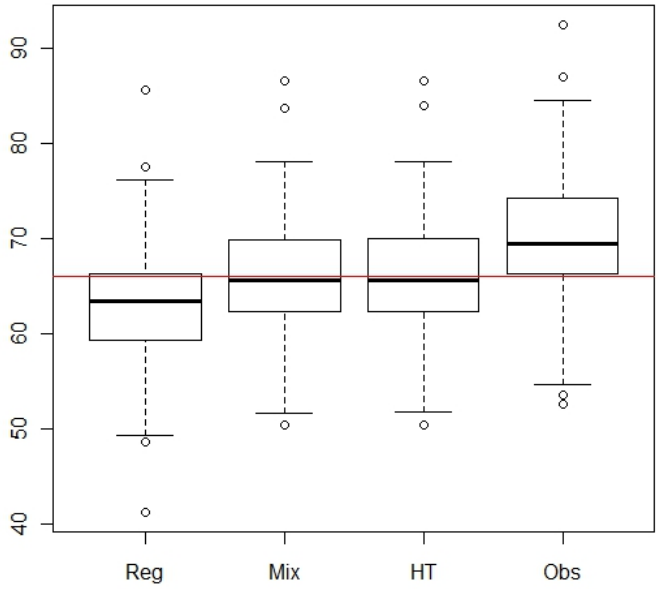

Figura 3.1: *Boxplots obtenidos para* $\alpha = 0.8$

Apéndice A

Convergencia casi completa

Definición A.0.1. Sea $(X_n)_{n\in\mathbb{N}}$ una sucesión de variables aleatorias reales. Diremos que X_n converge a X en forma *casi completa* si

$$\forall \epsilon > 0,\ \sum_{n\in\mathbb{N}} P\left(|X_n - X| > \epsilon\right) < \infty.$$

Notaremos esta clase de convergencia de la siguiente forma

$$\lim_{n\to\infty} X_n = X,\ a.co.$$

Observemos que si $X_n \to X, a.co.$ entonces de manera inmediata se deduce que $X_n \to X$ en probabilidad (P). Una condición necesaria para que la serie $\sum_{n\in\mathbb{N}} P\left(|X_n - X| > \epsilon\right)$ resulte sumable es que

$$\lim_{n\to\infty} P\left(|X_n - X| > \epsilon\right) = 0 \Leftrightarrow X_n \xrightarrow{P} X.$$

También incluye a la convergencia casi segura $(cs.)$. En efecto, si para todo $\epsilon > 0$,

$$\sum_{n\in\mathbb{N}} P\left(|X_n - X| > \epsilon\right) < \infty,$$

entonces por el Lema de Borel Cantelli, se verifica que

$$P\left(\overline{\lim_{n\to\infty}}\left\{|X_n - X| > \epsilon\right\}\right) = 0.$$

En consecuencia, el conjunto complementario dado por $\{\exists n,\ \forall m > n, |X_m - X| \leq \epsilon\}$ tendrá probabilidad 1.

Siendo ϵ cualquier número positivo, hemos probado que

$$P\left(\{\forall \epsilon, \exists n, \forall m > n, |X_m - X| \leq \epsilon\}\right) = 1 \Leftrightarrow X_n \xrightarrow{cs.} X.$$

Definición A.0.2. Sea $X_n \to X$ *a.co.* y $(u_n) \to 0$. Diremos que la *tasa de convergencia casi completa* de X_n a X es de orden $u_n \to 0$, si y solo si,

$$\exists \epsilon_0 > 0,\ \sum_{n\in\mathbb{N}} P\left(|X_n - X| > \epsilon_0 u_n\right) < \infty,$$

y lo escribiremos

$$X_n - X = O_{a.co.}(u_n).$$

La convergencia casi completa verifica las siguientes propiedades [14]:

- Si $X_n - X = O_{a.co.}(u_n)$, entonces:
  - *a)* $X_n - X = O_P(u_n)$
  - *b)* $X_n - X = O_{cs.}(u_n)$
- Si $\lim_{n\to\infty} X_n = L_X$ *a.co.* y $\lim_{n\to\infty} Y_n = L_Y$ *a.co.*, donde $L_X, L_Y \in \mathbb{R}$. Entonces,
  - *a)* $\lim_{n\to\infty} X_n + Y_n = L_X + L_Y$, *a.co.*
  - *b)* $\lim_{n\to\infty} X_n Y_n = L_X L_Y$, *a.co.*
  - *c)* $\lim_{n\to\infty} \frac{1}{Y_n} = \frac{1}{L_Y}$, *a.co.* siempre que $L_Y \neq 0$.
- Si $X_n - L_X = O_{a.co.}(u_n)$ y $Y_n - L_Y = O_{a.co.}(u_n)$,
  - *a)* $(X_n + Y_n) - (L_X + L_Y) = O_{a.co.}(u_n)$.
  - *b)* $X_n Y_n - L_X L_Y = O_{a.co.}(u_n)$.
  - *c)* $\frac{1}{Y_n} - \frac{1}{L_Y} = O_{a.co.}(u_n)$.

Notación A.0.3. Notaremos que la sucesión X_n está *acotada en probabilidad* de la siguiente forma

$$X_n = O_p(1).$$

Recordemos que la sucesión X_n está acotada en probabilidad si para todo $\epsilon > 0$, existe un $K > 0$ tal que para todo n se cumpla

$$P(|X_n| > K) < \epsilon.$$

Recordemos también la propiedad "0.Acotado" la cual permite afirmar que el producto de una sucesión acotada en probabilidad y una sucesión convergente a 0 en dicha medida, también resulta convergente a 0.

Apéndice B

Desigualdades exponenciales

Agradezco a la Dra. Pamela Llop por facilitar la siguiente demostración.

Proposición B.0.4 (**Desigualdad de Bernstein**). *Sea (Z_n) una sucesión de variables aleatorias independientes con media 0, b_i una sucesión de reales positivos y $a > 0$. Supongamos que*

$$\forall m \geq 2, \;\; \mathbb{E}(|Z_i^m|) \leq \frac{m!}{2} b_i^2 a^{m-2},$$

y sea $B_n^2 = \sum_{i=1}^{n} b_i^2$. Entonces,

$$\forall \epsilon > 0, \;\; P\left(\left|\sum_{i=1}^{n} Z_i\right| > \epsilon B_n\right) \leq 2\exp\left(-\frac{\epsilon^2}{2\left(1+\frac{\epsilon a}{B_n}\right)}\right)$$

Demostración. Nos bastará con probar que

$$P\left(\sum_{i=1}^{n} Z_i > \epsilon B_n\right) \leq \exp\left(-\frac{\epsilon^2}{2\left(1+\frac{\epsilon H}{B_n}\right)}\right).$$

Sean $S_n = \sum_{i=1}^{n} Z_i$ y $\lambda > 0$. Se verificará

$$\begin{aligned}
P\left(S_n > \epsilon B_n\right) &= P\left(e^{\lambda S_n} > e^{\lambda \epsilon B_n}\right) && \text{(B.1)}\\
&\leq e^{-\lambda \epsilon B_n}\, \mathbb{E}\left(e^{\lambda S_n}\right), \text{ en vista de la Desigualdad de Markov.} && \text{(B.2)}\\
&= e^{-\lambda \epsilon B_n}\, \mathbb{E}\left(\prod_{i=1}^{n} e^{\lambda Z_i}\right) && \text{(B.3)}\\
&= e^{-\lambda \epsilon B_n} \prod_{i=1}^{n} \mathbb{E}\left(e^{\lambda Z_i}\right), \text{ por la independencia de } Z_i && \text{(B.4)}
\end{aligned}$$

Miremos $\mathbb{E}\left(e^{\lambda Z_i}\right)$:

$$\begin{aligned}
\mathbb{E}\left(e^{\lambda Z_i}\right) &= \mathbb{E}\left(\sum_{i=0}^{\infty}\frac{\lambda^k}{k!}Z_i^k\right)\\
&= \mathbb{E}\left(1+\lambda Z_i+\sum_{i=2}^{\infty}\frac{\lambda^k}{k!}Z_i^k\right)\\
&\leq 1+\mathbb{E}\left(\sum_{i=2}^{\infty}\frac{\lambda^k}{k!}|Z_i|^k\right),\ \text{ya que } \mathbb{E}(Z_i)=0,\\
&\leq 1+\sum_{i=2}^{\infty}\lambda^k k!\,\mathbb{E}(|Z_i|^k),\ \text{por el Teorema de Convergencia Monótona,}\\
&\leq 1+\sum_{i=2}^{\infty}\frac{\lambda^k}{2}b_i^2a^{k-2},\ \text{por hipótesis,}\\
&= 1+\frac{b_i\lambda^2}{2}\sum_{i=0}^{\infty}(\lambda a)^k,\ \text{siempre que } \lambda a<1\\
&= 1+\frac{b_i^2\lambda^2}{2}\left(\frac{1}{1-\lambda a}\right)\\
&\leq \exp\left(\frac{b_i^2\lambda^2}{2}\frac{1}{(1-\lambda a)}\right),\ \text{pues } 1+z\leq e^z \text{ si } z>0.
\end{aligned}$$

Volviendo a la desigualdad B.4

$$\begin{aligned}
P\left(S_n>\epsilon B_n\right) &\leq e^{-\lambda\epsilon B_n}\prod_{i=1}^{n}\mathbb{E}\left(e^{\lambda Z_i}\right)\\
&\leq e^{-\lambda\epsilon B_n}\prod_{i=1}^{n}e^{\frac{b_i^2\lambda^2}{2}\frac{1}{(1-\lambda a)}}\\
&= e^{-\lambda\epsilon B_n}e^{\sum_{i=1}^{n}\frac{b_i^2\lambda^2}{2}\frac{1}{(1-\lambda a)}}\\
&= e^{-\lambda\epsilon B_n}e^{\frac{\lambda^2}{2(1-\lambda a)}B_n^2}
\end{aligned}$$

Considerando $\lambda=\frac{\epsilon}{B_n+\epsilon a}$, se obtiene

- $\lambda a=\frac{\epsilon a}{B_n+\epsilon a}<1$,
- $\lambda\epsilon B_n=\frac{-\epsilon^2 B_n}{B_n+\epsilon a}$,
- $\frac{\lambda^2}{2(1-\lambda a)}B_n^2=\frac{\epsilon^2}{2\left(1+\frac{\epsilon a}{B_n}\right)}$
- $-\lambda\epsilon B_n+\frac{B_n^2\lambda^2}{2(1-\lambda a)}=\frac{-\epsilon^2}{2\left(1+\frac{\epsilon b}{B_n}\right)}$.

Finalmente, obtenemos

$$P\left(\sum_{i=1}^{n}Z_i>\epsilon B_n\right)\leq\exp\left(\frac{-\epsilon^2}{2\left(1+\frac{\epsilon a}{B_n}\right)}\right).$$

Análogamente,

$$P\left(-\sum_{i=1}^{n} Z_i < -\epsilon B_n\right) \leq \exp\left(\frac{-\epsilon^2}{2\left(1+\frac{\epsilon a}{B_n}\right)}\right).$$

Lo cual implica,

$$P\left(\left|\sum_{i=1}^{n} Z_i\right| > \epsilon B_n\right) \leq 2\exp\left(\frac{-\epsilon^2}{2\left(1+\frac{\epsilon a}{B_n}\right)}\right).$$

Queda entonces demostrada la Proposición.

□

Lema B.0.5. *Sean Z_i va. independientes con media 0 tal que para $m \geq 2$ se cumple que $\mathbb{E}(|Z_i|^m) \leq C_m H^{2(m-1)}$, donde $C_m \leq \frac{m!}{2}C^2$, $H > 0$ y $C > 0$. Entonces,*

$$\forall \epsilon > 0,\ P\left(\left|\sum_{i=1}^{n} Z_i\right| > n\epsilon\right) \leq 2\exp\left(\frac{-\epsilon^2 n}{2H^2(C^2+\epsilon)}\right).$$

Demostración. Para demostrar el Lema aplicaremos la Desigualdad de Bernstein considerando $a = H^2$, $b_i^2 = C^2 a$ y $B_n = C\sqrt{na}$. De modo que para todo $m \geq 2$ se cumple la siguiente desigualdad

$$\mathbb{E}(|Z_i|^m) \leq \frac{m!}{2}C^2 a a^{m-2}$$

Tomando $\epsilon_n = \epsilon\frac{\sqrt{n}}{\sqrt{a}C}$ obtenemos

$$\begin{aligned}
P\left(\left|\sum_{i=1}^{n} Z_i\right| > \epsilon_n B_n\right) &\leq 2\exp\left(\frac{-\epsilon_n^2}{2\left(1+\frac{\epsilon_n a}{B_n}\right)}\right) && \text{(B.5)}\\
P\left(\left|\sum_{i=1}^{n} Z_i\right| > \epsilon n\right) &\leq 2\exp\left(\frac{-\epsilon^2 n}{2aC^2\left(1+\frac{\epsilon\sqrt{n}a}{\sqrt{a}C\sqrt{n}aC}\right)}\right) && \text{(B.6)}\\
&\leq 2\exp\left(\frac{-\epsilon^2 n}{2a\left(C^2+\epsilon\right)}\right) && \text{(B.7)}\\
&\leq 2\exp\left(\frac{-\epsilon^2 n}{2H^2\left(C^2+\epsilon\right)}\right) && \text{(B.8)}
\end{aligned}$$

Queda entonces demostrado el Lema. □

Bibliografía

[1] Rosenblatt, M. (1955). Remarks on some nonparametric estimates of a density function. *Ann. Statist.,* 27, 832-837.

[2] Parzen, E. (1962). On the estimation of a probability density function and the mode. *Ann. Statist.,* 33, 1065-1076.

[3] Whittle, P.(1958). On the smoothing of probability density functions. *Journal of the Royal Statistical Society B.,* 20, 334-343.

[4] Nadaraya, E. (1965). On non nonparametric estimates for density functions and regression curves. *Theory prob. and appl.,* 10, 297-302.

[5] Watson, G. (1964). Smooth regression analysis. *Sankhya Series A,* 26, 359-372.

[6] Wheeden, R. and Zygmund, A. (1977). *Measure and integral. An introduction to real analysis.* CRC Press. Págs 1488-150.

[7] Jones, M. C., Marron, J. S. and Sheather, S. J.(1996). A brief survey of bandwidth selection for density estimation. *Journal of the Royal Statistical Society B.,* 91, 401-407.

[8] Wasserman, L. (2006). *All of Nonparametric Statistics.* Springer Series in Statistics, New York.

[9] Stone, C. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.,* 12, 1285-1297.

[10] Härdle W. (1991). *Smoothing techniques, with implementations in S.* Springer.

[11] Härdle W. and Marron S, (1985). Optimal bandwidth selection in nonparametric regression function estimation.*Ann. Statist.,* 13, 1465-1481.

[12] Härdle W. and Hall, P, (1988). How far are automatically chosen regression smoothing parameters from their optimum? (with discussion) *Journal of The American Statistical Association.,* 83, 86-101.

[13] Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis. 2nd Edition.* Springer Series in Statistics, New York.

[14] Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis. Theory and Practice.* Springer, New York.

[15] Cheng, P. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of The American Statistical Association.,* 89, 81-87.

[16] Ferraty F., Sued M. and Vieu P., (2012). Mean estimation with data missing at random for functional covariables. *Statistics: A Journal of Theoretical and Applied Statistics.* Taylor &Francis Group, Florida, 1-19.

[17] Lighthill, M. J. (1958) *Introduction to Fourier Analysis and Generalised Functions.* Cambridge University Press, Cambridge.

[18] Bernstein, S.N.(1946) *Probability Theory,* 4th ed. (in Russian). M.-L. Gostechizdat.