UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales

Departamento de Matemática

Tesis de Licenciatura

Estimación en modelos aditivos con respuestas faltantes

Alejandra Mercedes Martinez

Directora: Dra. Graciela L. Boente Boente

Lugar de trabajo: Instituto de Cálculo, FCEyN, UBA.

Fecha: 30 de Marzo de 2009

Estimación en modelos aditivos con respuestas faltantes

El modelo de regresión noparamétrico aditivo, supone que se tienen observaciones independientes $(\mathbf{x}_i, y_i)$, $1 \leq i \leq n$, $\mathbf{x}_i \in \mathbb{R}^d$ tales que $E\left(y_i|\mathbf{x}_i\right) = m(\mathbf{x}_i)$ con $m(\mathbf{x}) = \sum_{\alpha=1}^{d} g_\alpha(x_\alpha)$. Las funciones $g_\alpha : \mathbb{R} \to \mathbb{R}$ son las cantidades a estimar. Estimadores para este modelo han sido ampliamente estudiados en la literatura. En esta tesis, se introduce una clase de estimadores para las componentes de un modelo aditivo cuando las respuestas pueden ser faltantes, es decir, cuando se observan $(\mathbf{x}_i, y_i, \delta_i)$, $1 \leq i \leq n$ donde $\delta_i = 1$ si y_i es observada y $\delta_i = 0$ si y_i es faltante. El objetivo es estudiar estimadores de la función de regresión y cada una de sus componentes que se basen solamente en los datos existentes así como propuestas basadas en imputar las observaciones faltantes. Para ello, se supondrá que tenemos un mecanismo de pérdida de observaciones ignorable, es decir, que δ_i e y_i son condicionalmente independientes dado $\mathbf{x}_i$, o sea, $P\left(\delta_i = 1|y_i, \mathbf{x}_i\right) = P\left(\delta_i = 1|\mathbf{x}_i\right) = p\left(\mathbf{x}_i\right)$.

En esta tesis, se obtienen resultados resultados de consistencia fuerte para los estimadores propuestos y se realiza un estudio de simulación que permite comparar las distintas propuestas dadas para muestras moderadas.

Palabras Claves: Consistencia; Datos Faltantes; Estimadores de Núcleos; Modelos Aditivo; Suavizadores.

Agradecimientos

No hubiese podido llegar a esta instancia sin la ayuda, cariño y comprensión de mucha gente. Este agradecimiento es para todos ellos.

A mi mamá, por haber estado siempre conmigo, en las buenas y en las malas.

A mi hermano, por haberme dicho siempre "Viste! Te lo dije!"al enterarse del resultado de cada examen.

A mi papá, por haberme recordado cada vez lo cerca que estaba de llegar a la meta.

A mi tía Chiquita, por haberme aconsejado, acompañado y ayudado todos estos años de carrera.

A mis tíos y primos, que siempre se preocuparon por mi.

A mi abuela, porque sin sus rezos y sus santos no hubiese tenido la fuerza y entusiasmo necesarios para rendir cada final.

A mis amigos y seres queridos, porque siempre me apoyaron, estudiaron al lado mío y jamás dejaron que me sienta sola.

A mi directora y profesores, por responder siempre a mis dudas y ayudarme a progresar día a día.

Índice general

Capítulo 1

Introducción

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

Como hemos mencionado, la inferencia estadística comúnmente se focaliza sobre funciones de distribución que son puramente paramétricas o puramente noparamétricas. Un modelo paramétrico razonable produce inferencias precisas mientras que un modelo erróneo posiblemente conducirá a conclusiones equivocadas. Por otro lado, los modelos noparamétricos si bien están asociados con alta estabilidad tienen menor precisión. Recientemente, los modelos noparamétricos han ganado una importante atención en el estudio de fenómenos naturales con comportamiento de complejidad no lineal. Sean $(\mathbf{x}_i^{\intercal}, y_i)^{\intercal}$ observaciones independientes idénticamente distribuidas (i.i.d) tales que $y_i \in \mathbb{R}$, $\mathbf{x}_i \in \mathbb{R}^d$ y

$$y_i = m(\mathbf{x}_i) + \epsilon_i \qquad 1 \leq i \leq n. \tag{1.1}$$

donde los errores ϵ_i son independientes e independientes de $\mathbf{x}_i$.

El análisis del modelo (1.1) requiere de técnicas de suavizado multivariadas para la función m y por lo tanto, encuentra en aplicaciones la dificultad conocida como "maldición de la dimensión"que está asociada al hecho de que al aumentar la dimensión los entornos de un punto $\mathbf{x}$ se hacen cada vez más ralos. Es decir, se necesita un número exponencialmente mayor de datos para que dichos entornos contengan observaciones de la muestra. Por ejemplo, si tenemos observaciones $\mathbf{x}_i \sim \mathcal{U}([-1,1]^d)$, $1 \leq i \leq n$, independientes, o sea, $x_{ij} \sim \mathcal{U}([-1,1])$ son variables independientes e idénticamente distribuídas, el número esperado de observaciones que yacen en $[-0.1, 0.1]^d$ es igual a $n/10^d$. Este fenómeno se traduce en la tasa de convergencia de los estimadores que no es $\sqrt{n}$ como en el caso paramétrico sino que, por ejemplo, para el caso de estimadores basados en núcleos es de orden $(nh_n^d)^{\frac{1}{2}}$ siendo h_n la ventana o parámetro de suavizado utilizado en el cómputo del estimador. En los últimos años, para resolver este problema, diversos autores han tratado el

problema de reducción de la dimensión de las covariables en modelos de regresión noparamétrica. Hastie y Tibshirani (1990) introdujeron los modelos aditivos que generalizan los modelos lineales, resuelven el problema de "la maldición de la dimensión"y además son de fácil interpretación. Este nuevo planteo combina la flexibilidad de los modelos noparamétricos con la simple interpretación del modelo lineal estándar. En el mismo, se supone que $m(\mathbf{x}) = \sum_{j=1}^{d} g_j(x_j)$ donde las funciones $g_j : \mathbb{R} \to \mathbb{R}$ son las cantidades a estimar. Estimadores para este modelo han sido ampliamente estudiados en la literatura.

En el Capítulo 2, se recordarán los principales métodos de estimación en el modelo de regresión noparamétrica (1.1), mientras que en el Capítulo 3 se introducirá el modelo aditivo y se describirán procedimientos de estimación de sus componentes.

En muchas situaciones, sobre todo en estudios biomédicos, puede haber un conjunto de los puntos del diseño con respuestas faltantes. Un tema fundamental de interés es estudiar el impacto de las observaciones faltantes en el funcionamiento de los estimadores utilizados. El análisis de regresión lineal con datos faltantes fue desarrollado con Yates (1933) quién propuso sustituir las respuestas faltantes por predicciones basadas en mínimos cuadrados. Con esta idea de imputar valores faltantes por predicciones basadas en mínimos cuadrados, Cochran (1968) redujo el sesgo en estudios observacionales, y Afifi y Elashoff (1969) dieron resultados asintóticos sobre las propuestas basadas en el proceso de imputación. De aquí en más, muchos estudios se enfocaron en modelos de regresión lineal y modelos log–lineales con datos faltantes.

La inferencia básica en estos modelos considera una muestra aleatoria $(\mathbf{x}_i^{\mathrm{T}}, y_i, \delta_i)^{\mathrm{T}}$, $1 \leq i \leq n$ donde la covariable $\mathbf{x}_i$ es observada, mientras que la variable respuesta y_i no es completamente observada. Para indicar la presencia o ausencia de respuesta se introduce la variable indicadora δ_i que vale 1 si y_i se observa mientras que si y_i es faltante $\delta_i = 0$. Aún cuando en muchas situaciones, tanto las respuestas como la variables explicativas son faltantes, en esta tesis nos concentraremos en el caso en que los datos faltantes ocurren sólo en las respuestas. Esta situación ocurre en muchos estudios biológicos cuando las covariables pueden ser controladas. El simple patrón de datos faltantes descripto corresponde al esquema de muestra doble propuesto por Neyman (1938) donde primero se obtiene una muestra completa y luego, se miden algunas covariables porque es quizás menos costosa de obtener esta muestra de covariables que la de variables respuesta. Por otra parte, en muestras basadas en encuesta, estos datos faltantes usualmente ocurren en forma de no–respuestas. En esta tesis, supondremos que los datos son faltantes al azar (MAR) que debilita la condición de ser faltantes completamente al azar (MCAR). Sea $(Y, \mathbf{X}^{\mathrm{T}}, \delta)^{\mathrm{T}}$ un vector aleatorio con la misma distribución que $(\mathbf{x}_i^{\mathrm{T}}, y_i, \delta_i)^{\mathrm{T}}$. La suposición de MAR requiere la existencia de un mecanismo de aleatoridad, indicado por $p(\mathbf{X})$, tal que $P(\delta = 0|\mathbf{X}, Y) = P(\delta = 1|\mathbf{X}) = p(\mathbf{X})$, es decir, que el supuesto MAR asume que δ y la respuesta Y son condicionalmente independientes dadas las covariables $\mathbf{X}$. Por otro lado, MCAR es más restrictivo pues requiere que δ sea independiente tanto de $\mathbf{X}$ como de Y, es decir, $p(\mathbf{X})$ es idénticamente igual a una constante p entre 0 y 1. En la práctica, el supuesto MAR debe estar justificado por la naturaleza del experimento cuando resulte legítimo suponer que la ausencia de Y depende principalmente de $\mathbf{X}$.

En el Capítulo 4, se describen procedimientos de estimación cuando la respuesta es faltante. En particular, el caso de variables explicativas multidimensonales se presenta en la sección 4.3, donde se introducen, para el modelo aditivo, las distintas propuestas objeto de esta tesis. En el Capítulo 5 se prueban resultados de consistencia para todos los estimadores propuestos. Finalmente, los resultados de un estudio de simulación preliminar se dan en el Capítulo 6.

Capítulo 2

Regresión noparamétrica

2.1. Introducción

Una pregunta importante en muchos campos de la ciencia es la relación entre dos variables, digamos X e Y o entre un vector $\mathbf{X} \in \mathbb{R}^d$ y una variable Y. El análisis de regresión está relacionado a la pregunta de cómo la variable *dependiente* Y puede ser explicada por las variables *explicativas*, *regresoras* o *independientes* $\mathbf{X}$. Esto significa tratar de hallar una relación de la forma $Y = m(\mathbf{X})+\epsilon$, donde $m : \mathbb{R}^d \to \mathbb{R}$ es una función y ϵ es un error aleatorio correspondiente en muchos casos al error de medición. En muchos casos, la teoría del estudio que se está realizando no pone restricciones sobre m, es decir, la teoría no dice si m es lineal, cuadrática o creciente en X, en el caso unidimensional o condiones análogas en el caso multidimensional. Por lo tanto, depende del análisis empírico usar los datos correspondientes a las mediciones realizadas para descubrir más sobre m.

Contamos por lo tanto, con observaciones $(\mathbf{x}_i, y_i)$, $1 \leq i \leq n$ independientes e indénticamente distribuídas que satisfacen el modelo

$$y_i = m(\mathbf{x}_i) + \epsilon_i, \qquad 1 \leq i \leq n. \tag{2.1}$$

La ecuación (2.1) dice que la relación $Y = m(\mathbf{X})$ no se satisface exactamente sino que es afectada por una variable aleatoria ϵ, llamada error aleatorio. En la mayoría de las situaciones se supone que $E(\epsilon) = 0$ con lo cual, el modelo puede escribirse como

$$E(Y|\mathbf{X} = \mathbf{x}) = m(\mathbf{x}) \ , \tag{2.2}$$

la ecuación (2.2) dice que la relación entre la variable dependiente y la variable independiente se mantiene en promedio. El objetivo del análisis empírico es usar el conjunto de observaciones $(\mathbf{x}_i, y_i)$, $1 \leq i \leq n$ para estimar m.

Recordemos que en el enfoque paramétrico, usualmente se asume que m es lineal, o sea, $m(\mathbf{x}) = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$, y el problema de estimar m se reduce al problema de estimar α y $\boldsymbol{\beta} \in \mathbb{R}^d$. Sin embargo, esta aproximación no es siempre apropiada. El enfoque noparamétrico no supone restricciones a priori sobre m, salvo condiciones de suavidad o continuidad para obtener resultados de convergencia uniforme. Sin embargo, como veremos más adelante, hay un precio a pagar por esta flexibilidad.

Supongamos que $(\mathbf{X}, Y)$ es un vector aleatorio con función de densidad de probabilidad (pdf) conjunta $f(\mathbf{x}, y)$. El esquema muestral natural es considerar una muestra aleatoria de la distribución caracterizada por $f(\mathbf{x}, y)$. Esto es, considerar aleatoriamente observaciones de la forma $(\mathbf{x}_i, y_i)$, $1 \leq i \leq n$ independientes e idénticamente distribuídas con densidad $f(\mathbf{x}, y)$. Este esquema muestral será referido como un diseño *aleatorio*.

Sin embargo, hay aplicaciones (especialmente en las ciencias naturales) donde el investigador puede controlar los valores de las variables predictoras $\mathbf{X}$ que, por lo tanto, es fija, e Y es la única variable aleatoria. Este esquema es referido como un diseño *fijo*. En este caso, podemos pensar que la densidad $f_{\mathbf{X}}(\mathbf{x})$ es conocida (inducida por el investigador). Este conocimiento adicional simplificará la estimación de m.

En esta tesis, supondremos que los efectos son aleatorios. El caso de efectos fijos en regresión noparamétrica sin datos faltantes puede verse en Härdle (2004).

2.2. Regresión univariada basada en núcleos

2.2.1. Definición del estimador de regresión

La derivación del estimador de regresión en el caso de un diseño aleatorio cuando $d = 1$ es natural si se recuerda la definición de esperanza condicional

$$m(x) = E(Y|X = x) = \int y \frac{f(x, y)}{f_X(x)}\, dy = \frac{\int y f(x, y)\, dy}{f_X(x)}\ . \tag{2.3}$$

Tanto $f(x, y)$ como $f_X(x)$ en (2.3) son desconocidas y en consecuencia, podemos estimarlas utilizando los estimadores basados en núcleos definidos por Rosenblatt (1956) y Parzen (1962). El estimador de núcleos de $f_X(x)$ se define como

$$\widehat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i),$$

donde $K_h(x) = (1/h)K\left(u/h\right)$ con $K : \mathbb{R} \to \mathbb{R}$ una función par, tal que $\int K(u)\, du = 1$ y $K \geq 0$. Llamamos a K función núcleo.

Para estimar $f(x, y)$ utilizaremos el estimador de densidad basado en núcleos multiplicativo, es decir, obtenido como producto de núcleos

$$\widehat{f}_{n,(X,Y)}(x, y) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i) K_g(y - y_i).$$

Por lo tanto, podemos estimar el numerador $r(x)$ de (2.3) por la expresión

$$\widehat{r}_n(x) = \int y \widehat{f}_{n,(X,Y)}(x, y)\, dy \quad = \quad \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i) \int \frac{y}{g} K\left(\frac{y - y_i}{g}\right) dy$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n} K_h(x-x_i)\int (sg+y_i)K(s)\,ds \\
&= \frac{1}{n}\sum_{i=1}^{n} K_h(x-x_i)y_i
\end{aligned}
$$

y obtenemos de esta forma el estimador de Nadaraya (1964)–Watson (1964) de la función de regresión m

$$\widehat{m}_n(x) = \frac{\widehat{r}_n(x)}{\widehat{f}_n(x)} = \frac{\sum_{i=1}^{n} K_h(x-x_i)y_i}{\sum_{j=1}^{n} K_h(x-x_j)} \tag{2.4}$$

que es un estimador natural de la esperanza condicional ya que en el caso en que $K(u) = (1/2)\mathrm{I}_{[-1,1]}(u)$ obtenemos el promedio de las respuestas y_i correspondientes a valores de la covariable x_i cercanos a x, o sea, tales que $|x_i - x| \leq h$. Algunos puntos significativos son:

- Reescribiendo (2.4) como

$$\widehat{m}_n(x) = \sum_{i=1}^{n}\left(\frac{K_h(x-x_i)}{\sum_{j=1}^{n} K_h(x-x_j)}\right) y_i = \sum_{i=1}^{n} W_{ni}(x)y_i$$

  se revela que el estimador de Nadayara-Watson puede ser visto como un promedio pesado (local) de las variables respuesta y_i ya que $\sum_{i=1}^{n} W_{ni}(x) = 1$.

- Notemos que, tal como ocurre con los estimadores de densidad basados en núcleos, la ventana h determina el grado de suavidad de $\widehat{m}_n$. Si $h \to 0$, entonces $W_{ni}(x) \to 1$ si $x = x_i$ y no está definida en otro caso. Por lo tanto, en una observación x_i, $\widehat{m}_n$ converge a y_i, es decir, tenemos una interpolación de los datos. Por otro lado, si $h \to \infty$ luego, $W_{ni}(x) \to 1/n$ para todo x, y $\widehat{m}_n(x_i) \to \bar{y}$, es decir, el estimador es una función constante que asigna el promedio muestral a cada x. Lograr elegir h de manera de tener un buen compromiso entre sub y sobresuavidad es nuevamente un problema crucial.

- Si el denominador de $W_{ni}(x)$ es igual a 0, el numerador también es igual a cero y la estimación no está definida. Esto puede ocurrir en regiones donde los datos son escasos.

2.2.2. Propiedades estadísticas

En esta sección, nos interesa describir las condiciones bajo las cuales los estimadores de núcleos resultan consistentes. Limitaremos nuestra descripción a la consistencia débil de los estimadores y daremos una expresión para el error cuadrático medio asintótico.

Consideraremos el siguiente conjunto de hipótesis

A1 $\int |K(u)|\,du < \infty$,

A2 $\lim_{u\to\infty} uK(u) \to 0$,

A3 $E(Y^2) < \infty$,

A4 $h_n \to 0$, $nh_n \to \infty$.

El siguiente Teorema cuya demostración puede verse en Härdle (1990) establece la consistencia débil puntual de los estimadores $\widehat{m}_n(x)$ de la función de regresión. Para obtenerla basta mostrar que tanto el numerador como el denominador de $\widehat{m}_n(x)$ convergen a $m(x)f_X(x)$ y $f_X(x)$, respectivamente, de donde se obtiene fácilmente el resultado.

Teorema 2.2.1. *Supongamos que (x_i, y_i), $1 \leq i \leq n$ son independientes e idénticamente distribuídas, tales que $E(y_i|x_i) = m(x_i)$. Supongamos además que se cumplen **A1** a **A4**. Sea x un punto de continuidad de $m(x)$, $f_X(x)$ y de $\sigma^2(x) = Var(Y|X = x)$ tal que $f_X(x) > 0$, entonces el estimador definido en (2.4) cumple que $\widehat{m}_n(x) \xrightarrow{p} m(x)$.*

Como es bien sabido, el error cuadrático medio, que indicaremos MSE, tiene dos componentes, un término de varianza y uno de sesgo al cuadrado. Para el caso del estimador de regresión basado en núcleos, el desarrollo siguiente permite obtener esa descomposición (ver, Prakasa Rao, 1983). Como $\widehat{m}_n(x) = \widehat{r}_n(x)/\widehat{f}_n(x)$ tenemos

$$\begin{aligned}\widehat{m}_n(x) - m(x) &= \left\{\frac{\widehat{r}_n(x)}{\widehat{f}_n(x)} - m(x)\right\}\left[\frac{\widehat{f}_n(x)}{f_X(x)} + \left\{1 - \frac{\widehat{f}_n(x)}{f_X(x)}\right\}\right] \\ &= \frac{\widehat{r}_n(x) - m(x)\widehat{f}_n(x)}{f_X(x)} + \{\widehat{m}_n(x) - m(x)\}\frac{f_X(x) - \widehat{f}_n(x)}{f_X(x)}.\end{aligned}$$

Se puede ver que de los dos términos del lado derecho, el primero es el término lider en la distribución de $\widehat{m}_n(x) - m(x)$, mientras que el segundo término puede despreciarse. Por lo tanto, asintóticamente el MSE de $\widehat{m}_n(x)$ puede ser aproximado a partir de $E\left\{\dfrac{\widehat{r}_n(x) - m(x)\widehat{f}_n(x)}{f_X(x)}\right\}^2$.

Teorema 2.2.2. *Supongamos que (x_i, y_i), $1 \leq i \leq n$ son independientes e idénticamente distribuídas, tales que $E(y_i|x_i) = m(x_i)$. Supongamos además que se cumplen **A1** a **A4**. Sea x un punto de continuidad de $m(x)$, $m'(x)$, $m''(x)$, $f_X(x)$, $f'_X(x)$ y de $\sigma^2(x) = Var(Y|X = x)$ tal que $f_X(x) > 0$, entonces se tiene que*

$$\text{MSE}\{\widehat{m}_n(x)\} \approx \frac{1}{nh}\frac{\sigma^2(x)}{f_X(x)}\|K\|_2^2 + h^4\left\{\frac{m''(x)}{2} + \frac{m'(x)f'_X(x)}{f_X(x)}\right\}^2 \mu_2^2(K) = \text{AMSE}(n,h) \qquad (2.5)$$

donde $\mu_2(K) = \int u^2 K(u)du$ y $\|K\|_2^2 = \int K^2(u)du$.

El primer sumando de $\text{AMSE}(n,h)$ corresponde a la varianza mientras que el segundo al sesgo. Como vemos esta expresión muestra que es necesario un compromiso entre ambos ya que valores grandes de la ventana aumentan el sesgo mientras que disminuyen la varianza y recíprocamente. Sean

$$C_1 = \frac{\sigma^2(x)}{f_X(x)}\|K\|_2^2 \qquad \text{y} \qquad C_2 = \left\{\frac{m''(x)}{2} + \frac{m'(x)f'_X(x)}{f_X(x)}\right\}^2 \mu_2^2(K)$$

entonces, tenemos que $\text{AMSE}(n,h) = C_1/(nh) + h^4 C_2$. Minimizando esta expresión con respecto a h, obtenemos la ventana óptima, en el sentido de minimizar el error cuadrático asintótico es $h_{opt} = (C_1/4C_2)^{1/5} n^{-1/5}$, es decir, tiene orden $n^{-1/5}$. Por otra parte, $\text{AMSE}(n, h_{opt})$ tiene orden $n^{4/5}$, es decir, que la tasa de convergencia de estos estimadores es más lenta que la del estimador de mínimos cuadrados en el modelo de regresión lineal pero es la misma que la del estimador de la función de densidad.

2.3. Otros estimadores para el modelo de regresión univariada

2.3.1. Regresión polinomial local y estimación derivada

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

$$m(t) \approx m(x) + m'(x)(t-x) + \ldots + m^{(p)}(x)(t-x)^p \frac{1}{p!} \tag{2.6}$$

para todo punto t en un entorno del punto x. Esto sugiere una regresión *polinomial local*, que busca ajustar un polinomio de grado p en un entorno de x, incluyendo pesos basados en núcleos en el problema de minimización para penalizar observaciones correspondientes a valores de las covariables alejadas de x y dar mayor preponderancia a las respuestas y_i correspondientes a valores de x_i cercanos a x. Es decir, resolvemos

$$\widehat{\boldsymbol{\beta}}(x) = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \sum_{i=1}^{n} \{y_i - \sum_{k=0}^{p} \beta_k (x_i - x)^k\}^2 K_h(x - x_i), \tag{2.7}$$

donde $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^\intercal$ es el vector de coeficientes. El resultado es un estimador de mínimos cuadrados pesados con pesos $K_h(x - X_i)$. Usando la notación

$$\mathbf{P} = \begin{pmatrix} 1 & x_1 - x & (x_1-x)^2 & \ldots & (x_1-x)^p \\ 1 & x_2 - x & (x_2-x)^2 & \ldots & (x_2-x)^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x & (x_n-x)^2 & \ldots & (x_n-x)^p \end{pmatrix}, \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

$$\mathcal{W} = \begin{pmatrix} K_h(x-x_1) & 0 & \ldots & 0 \\ 0 & K_h(x-x_2) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & K_h(x-x_n) \end{pmatrix},$$

$\widehat{\boldsymbol{\beta}}(x)$ con la fórmula usual del estimador de mínimos cuadrados pesados

$$\widehat{\boldsymbol{\beta}}(x) = (\mathbf{P}^\intercal \mathcal{W} \mathbf{P})^{-1} \mathbf{P}^\intercal \mathcal{W} \mathbf{y}.$$

Es importante notar que, en contraste con la versión paramétrica de mínimos cuadrados, este estimador varía con x. Por lo tanto, esto es realmente una regresión local en el punto x. Denotemos las componentes de $\widehat{\boldsymbol{\beta}}(x)$ por $\widehat{\beta}_0(x), ..., \widehat{\beta}_p(x)$. El estimador polinomial local de la función regresora m se define como $\widehat{m}_{p,n}(x) = \widehat{\beta}_0(x)$ ya que en el desarrollo (2.6) corresponde al primer término. Toda la función $\widehat{m}_{p,n}$ se obtiene variando x en la regresión polinomial local anterior.

Observemos que para el caso en que $p = 0$, $\widehat{\boldsymbol{\beta}}(x)$ se reduce a $\widehat{\beta}_0(x)$, lo que significa que el estimador localmente constante no es más que el estimador Nadayara-Watson, es decir,

$$\widehat{m}_{0,n}(x) = \widehat{m}_n(x) = \frac{\sum_{i=1}^n K_h(x - x_i) y_i}{\sum_{i=1}^n K_h(x - x_i)}.$$

Para el caso en que $p = 1$, el estimador localmente lineal está dado por

$$\widehat{m}_{1,n}(x) = \mathbf{e_0}^{\intercal}(\mathbf{P}^{\intercal}\mathcal{W}\mathcal{X})^{-1}\mathbf{P}^{\intercal}\mathcal{W}\mathbf{y}$$

donde $\mathbf{e_0} = (\mathbf{1}, \mathbf{0}, \ldots, \mathbf{0})$. Este ajuste localmente lineal, a diferencia del ajuste por Nadayara-Watson, reacciona más sensiblemente en los extremos del ajuste y produce menos sesgo en la frontera del soporte. El error cuadrático medio asintótico del estimador de regresión localmente lineal está dado por

$$\text{AMSE}\{\widehat{m}_{1,n}(x)\} = \frac{1}{nh}\frac{\sigma^2(x)}{f_X(x)}\|K\|_2^2 + \frac{h^4}{4}\{m''(x)\}^2\mu_2^2(K).$$

El AMSE en este caso difiere del AMSE del estimador Nadayara-Watson solamente en el sesgo. Es fácil ver que el sesgo en el ajuste localmente lineal es independiente del diseño y desaparece si m es lineal. Entonces, el ajuste localmente lineal puede mejorar la estimación en regiones con observaciones dispersas. Se puede ver también que el sesgo del estimador localmente lineal tiene la misma forma que para el caso de un diseño fijo.

Para estimar funciones de regresión, los ordenes p usualmente utilizados son uno (localmente lineal) o tres (regresión localmente cúbica), en general ocurre que los órdenes impares son mejores que los órdenes pares. Tal como ocurre para otros métodos basados en núcleos, la amplitud h determina el grado de suavidad de $\widehat{m}_{p,n}$. Una ventaja adicional de la aproximación localmente polinómica es que provee una fácil manera de estimar las derivadas de la función m. La aproximación natural sería estimar m por $\widehat{m}$ y luego computar las derivadas de $\widehat{m}$. Sin embargo, un método alternativo y más eficiente se obtiene al comparar (2.6) con (2.7). A partir de esto, tenemos el estimador localmente polinomial de la derivada de orden ν de la función de regresión está dado por $\widehat{m}_{p,n}^{(\nu)}(x) = \nu!\widehat{\beta}_\nu(x)$. Usualmente, si se desea estimar la derivada de orden ν, el orden del polinomio es $p = \nu + 1$ o $p = \nu + 3$.

2.3.2. Estimador basado en vecinos más cercanos

El estimador basado en los k vecinos más cercanos ($k-$NN) puede verse como un promedio pesado de las variables respuesta en un entorno de x, con la importante diferencia de que la amplitud del entorno no es fija sino variable. Para ser más específicos, los valores de las respuestas usados para computar el promedio son aquellos que corresponden a los k valores observados de las covariables más cercanos al punto x, punto en el que queremos estimar $m(x)$. Formalmente, el estimador $k-$NN puede escribirse como

$$\widehat{m}_n(x) = \sum_{i=1}^{n} W_{ni}(x)y_i, \tag{2.8}$$

donde los pesos $\{W_{ni}(x)\}_{i=1}^n$ están definidos como

$$W_{ni}(x) = \begin{cases} 1/k & \text{si } i \in J_{k,x} \\ 0 & \text{caso contrario} \end{cases}$$

con el conjunto de índices $J_{k,x} = \{i : x_i \text{ es una de las } k \text{ observaciones más cercanas a } x\}$. Notemos que $k = k_n$ es el parámetro de suavidad de este estimador. Incrementar k hace más suave la estimación.

El estimador $k-$NN puede verse como un estimador basado en núcleos con núcleo uniforme $K(u) = (1/2)\mathrm{I}_{[-1,1]}(u)$ y amplitud variable $R = R(k)$, donde $R(k)$ es la máxima distancia entre x y sus k vecinos más cercanos. Es decir, en primer lugar ordenamos las distancias $D_i(x) = |x_i - x|$ de menor a mayor, sean $D^{(1)}(x) \leq D^{(2)}(x) \leq \ldots D^{(n)}(x)$ los estadísticos de orden, entonces $R(k) = D^{(k)}(x)$, $J_{k,x} = \{i : |x_i - x| \leq R(k)\}$ y

$$\widehat{m}_n(x) = \frac{\sum_{i=1}^n K_R(x - x_i)y_i}{\sum_{i=1}^n K_R(x - x_i)} .$$

Esta forma de escribir el estimador de vecinos más cercanos permite introducir una familia más general de estimadores que son los esimadores de vecinos más cercanos con núcleo. Sea K un núcleo, y sea $R = R(k)$ la distancia de x a su $k-$ésimo vecino más cercano, el estimador de vecinos más cercanos con núcleo se define como

$$\widehat{m}_n(x) = \frac{\sum_{i=1}^n K_R(x - x_i)y_i}{\sum_{i=1}^n K_R(x - x_i)} . \tag{2.9}$$

El siguiente resultado da condiciones para su consistencia débil, ya que obtiene aproximaciones para su sesgo y varianza.

Teorema 2.3.1. *Supongamos que* $\lim_{n\to\infty} k = \infty$, $\lim_{n\to\infty} k/n = 0$ *y se cumplen* ***A1*** *a* ***A3****. Entonces, si* $\widehat{m}_n(x)$ *indica el estimador definido en (2.9) se tiene que*

$$\begin{aligned} E\{\widehat{m}_n(x)\} - m(x) &\approx \frac{\mu_2(K)}{8f_X(x)^2}\left\{m''(x) + 2\frac{m'(x)f'_X(x)}{f_X(x)}\right\}\left(\frac{k}{n}\right)^2 \\ Var\{\widehat{m}_k(x)\} &\approx 2\|K\|_2^2\frac{\sigma^2(x)}{k} . \end{aligned}$$

Observemos que eligiendo $k = 2nhf_X(x)$ obtenemos un k-NN estimador que es aproximadamente idéntico a un estimador basado en núcleos con amplitud de intervalo h en los términos principales del MSE.

Existen otras familias de estimadores que son los basados en splines o en series ortogonales. Una descripción de los mismos y de sus principales propiedades puede verse en Prakasa Rao (1983).

2.4. Selección del parámetro de suavidad

Como hemos visto la ventana o el parámetro de suavizado h o el número de vecinos k tienen un papel fundamental en el momento de elegir el estimador ya que determinan la forma del estimador resultante. En esta sección, nos interesará describir métodos para elegir una "buena" ventana en el caso de estimación basada en núcleos. En primer lugar, la ventana deberá tener deseables propiedades teóricas. En segundo lugar, deberá ser aplicable en la práctica. Considerando la primera condición, se han propuesto diversas medidas de cuán cerca está la estimación a la verdadera función.

- El MSE mide el desvío al cuadrado del estimador $\widehat{m}_n$ de m en el punto x. Si estamos interesados en cuán bien estimamos a la función m debemos utilizar una medida global de cercanía del estimador a la verdadera función.

- El *error cuadrático integrado* (ISE) definido como

$$\text{ISE}(h) = \text{ISE}\{\widehat{m}_n\} = \int_{-\infty}^{\infty} \{\widehat{m}_n(x) - m(x)\}^2 w(x) f_X(x)\, dx$$

  es una medida de discrepancia global. La función $w(x)$ es una función que se introduce para dar distinto peso a diferentes regiones, por ejemplo, si queremos reducir el efecto de frontera o en regiones de datos ralos, para reducir la varianza en esa región. Sin embargo, $\text{ISE}(h)$ es una variable aleatoria dado que diferentes muestras producirán diferentes valores de $\widehat{m}_n(x)$, y por lo tanto diferentes valores de ISE, lo que complica su comparación.

- El *error cuadrático integrado medio* (MISE) se define como

$$\text{MISE}(h) = \text{MISE}\{\widehat{m}_n\} = E\{\text{ISE}(h)\} =$$
$$\int \dots \int \left[\int_{-\infty}^{\infty} \{\widehat{m}_n(x) - m(x)\}^2 w(x) f_X(x)\, dx \right] f(x_1, \dots, x_n, y_1, \dots, y_n)\, dx_1 \dots\, dx_n\, dy_1 \dots\, dy_n$$

  y tiene la ventaja de que no es una variable aleatoria.

- El *error cuadrático promediado* (ASE)

$$\text{ASE}(h) = \text{ASE}\{\widehat{m}_n\} = \frac{1}{n} \sum_{j=1}^{n} \{\widehat{m}_n(x_j) - m(x_j)\}^2 w(x_j)$$

  es una aproximación discreta de ISE, y tal como ocurre con ISE es una variable aleatoria y una medida de discrepancia global

- El *error cuadrático medio promediado* MASE

$$\text{MASE}(h)(x_{1,o}, \dots, x_{n,o}) = \text{MASE}\{\widehat{m}_n\} = E\{\text{ASE}(h) | x_1 = x_{1,o}, \dots, x_n = x_{n,o}\}$$

  es la esperanza condicional de ASE. Luego, variando los posibles valores de las variables aleatorias $x_1, \dots, x_n$, podemos ver a MASE como una variable aleatoria.

Una elección natural de medida de discrepancia para derivar una regla para la elección de h sería MISE o su versión asintótica AMISE ya que esta medida es la utilizada para seleccionar la ventana óptima en el caso de estimación de densidades o de regresión. Sin embargo, como hemos visto la ventana óptima depende de las derivadas de la función de regresión y de densidad que son desconocidas.

Discutiremos dos aproximaciones para su aplicabilidad: convalidación cruzada y términos de penalidad. Por simplicidad, nos restringiremos a la selección de la amplitud del intervalo para el estimador de Nadayara-Watson. Para este estimador se puede ver que ASE, ISE y MISE conducen asintóticamente al mismo nivel de suavidad (ver Härdle, Müller, Sperlich y Werwatz, 2004). Por lo tanto, podemos usar el criterio que es más fácil de calcular y manipular la versión discreta $\text{ASE}(h)$.

2.4.1. El error cuadrático promediado

Queremos encontrar la amplitud de intervalo h que minimice $\text{ASE}(h)$. Observemos que

$$\text{ASE}(h) = \frac{1}{n}\sum_{i=1}^{n} m^2(x_i)w(x_i) + \frac{1}{n}\sum_{i=1}^{n} \widehat{m}_n^2(x_i)w(x_i) - 2\frac{1}{n}\sum_{i=1}^{n} m(x_i)\widehat{m}_n(x_i)w(x_i).$$

Por lo tanto, su esperanza condicional, MASE, está dada por

$$\begin{aligned} &\text{MASE}(h)(x_{1,o},\ldots,x_{n,o}) = E\{\text{ASE}(h)|x_1 = x_{1,o},\ldots,x_n = x_{n,o}\} \\ = \;& \frac{1}{n}\sum_{i=1}^{n}\left[Var\{\widehat{m}_n(x_i)|x_1 = x_{1,o},\ldots,x_n = x_{n,o}\} + \text{sesgo}^2\{\widehat{m}_n(x_i)|x_1 = x_{1,o},\ldots,x_n = x_{n,o}\}\right] w(x_i). \end{aligned}$$

Ambos $\text{MASE}(h)$ y $\text{ASE}(h)$ dependen de m, la función que queremos estimar. Luego, si, por ejemplo, lo que queremos es graficarlos será necesario reeplazarla por una aproximación basada en la muestra. Una manera sencilla e ingenua de hacer ésto es reemplazar $m(x_i)$ por las respuestas y_i, es decir, considerar

$$p(h) = \frac{1}{n}\sum_{i=1}^{n}\{y_i - \widehat{m}_n(x_i)\}^2 w(x_i),$$

llamado *estimador resustituído* y que es esencialmente una suma pesada de residuos cuadrados (RSS). Sin embargo, hay una problema con esta aproximación ya que y_i es usado en $\widehat{m}_n(x_i)$ para autopredecirse con lo cual $p(h)$ puede hacerse arbitrariamente pequeño cuando $h \to 0$. Se puede ver también que $p(h)$ es sesgado como estimador de $\text{ASE}(h)$. En las dos secciones que siguen se describen dos maneras de resolver este problema.

2.4.2. Convalidación cruzada

El método de convalidación cruzada resuelve el problema descripto para $p(h)$, en que y_i es usado en $\widehat{m}_n(x_i)$ para autopredecirse, ya que utiliza el estimador *leave-one-out*

$$\widehat{m}_{n,-i}(x_i) = \frac{\sum_{j\neq i} K_h(x_i - x_j)y_j}{\sum_{j\neq i} K_h(x_i - x_j)}\ .$$

Esto es, en la estimación de $\widehat{m}_n$ en x_i la $i-$ésima observación es excluída. Esto resulta en la función de convalidación cruzada

$$\text{CV}(h) = \frac{1}{n}\sum_{i=1}^{n}\{y_i - \widehat{m}_{n,-i}(x_i)\}^2 w(x_i).$$

Se puede mostrar que minimizar $\text{CV}(h)$ es (en promedio) equivalente a minimizar $\text{ASE}(h)$. Podemos concluir entonces que tomando como regla para la elección de la amplitud de intervalo aquella que consiste en "elegir h que minimice $\text{CV}(h)$", esta regla resulta ser teóricamente deseable y aplicable en la práctica.

2.4.3. Funciones penalizadoras

Recordemos que $E\{p(h)|x_1,\ldots,x_n\} \neq E\{\text{ASE}(h)|x_1,\ldots,x_n\}$. Sin embargo, esta discrepancia entre ambas no es del todo importante siempre y cuando la ventana h que minimiza $E\{p(h)|x_1 = x_{1,o},\ldots,x_n = x_{n,o}\}$ sea la misma que la que minimiza $E\{\text{ASE}(h)|x_1 = x_{1,o},\ldots,x_n = x_{n,o}\}$. Desafortunadamente, ésto no ocurre. La aproximación de la función penalizadora corrige el sesgo multiplicando $p(h)$ por un factor de corrección. La "versión corregida" de $p(h)$ puede escribirse como

$$G(h) = \frac{1}{n}\sum_{i=1}^{n}\{y_i - \widehat{m}_n(x_i)\}^2\Psi\left(W_{ni}(x_i)\right)w(x_i)$$

con función de corrección Ψ. Si consideramos a Ψ con expansión de Taylor de primer orden dada por $\Psi(u) = 1 + 2u + O(u^2)$ y al reemplazarla en $G(h)$ ignoramos los términos de orden superior, se puede ver que asintóticamente $G(h)$ es igual a $\text{ASE}(h)$. Varias funciones penalizadoras han sido introducidas, cada una de ellas da distinto peso relativo a la varianza y sesgo de $\widehat{m}_n(x)$. Una descripción de las mismas puede verse en Härdle, Müller, Sperlich y Werwatz (2004).

Si denotamos $\widehat{h}$ a la amplitud de intervalo que minimiza $G(h)$ y $\widehat{h}_0$ a la que minimiza el $\text{ASE}(h)$ entonces, para $n \to \infty$

$$\frac{\text{ASE}(\widehat{h})}{\text{ASE}(\widehat{h}_0)} \xrightarrow{p} 1 \text{ y } \frac{\widehat{h}}{\widehat{h}_0} \xrightarrow{p} 1\ .$$

Por lo tanto, elegir la amplitud de intervalo minimizando $G(h)$ es otra "buena" regla de elección de amplitud para la estimación de regresión basada en núcleos. Resultados análogos pueden obtenerse para regresión local polinomial.

2.5. Regresión multivariada basada en núcleos

En esta sección, estaremos interesados en especificar cómo la variable respuesta Y depende del vector de variables $\mathbf{X}$, o sea, nuestro objetivo será estimar la esperanza condicional

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

donde $\mathbf{X} = (X_1,\ldots,X_d)^{\text{T}}$. Como en la sección 2.2, supongamos que el vector $(\mathbf{X}, Y)$ es un vector aleatorio con función de función de densidad conjunta $f(\mathbf{x}, y)$, entonces

$$E(Y|\mathbf{X}) = \int y f(y|\mathbf{x})\,dy = \frac{\int y f(\mathbf{x},y)\,dy}{f_{\mathbf{X}}(\mathbf{x})} = \frac{r(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}.$$

Reemplazaremos la densidad multivariada $f(\mathbf{x}, y)$ por un estimador de densidad basado en núcleos definido por

$$\widehat{f}_{n,\mathbf{H}}(y,\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_h(y_i - y)\mathcal{K}_{\mathbf{H}}(\mathbf{x}_i - \mathbf{x})$$

donde $K : \mathbb{R} \to \mathbb{R}$ es una función núcleo, $\mathbf{H}$ es una matriz no singular y $\mathcal{K} : \mathbb{R}^d \to \mathbb{R}$ es tal que $\mathcal{K} \geq 0$ y $\int \mathcal{K}(\mathbf{u})\,d\mathbf{u} = 1$ y $\mathcal{K}_{\mathbf{H}}(\mathbf{u}) = (\det(\mathbf{H}))^{-1}\mathcal{K}(\mathbf{H}^{-1}\mathbf{u})$. Llamaremos a $\mathcal{K}$ función núcleo multivariada.

Reemplazaremos también $f_{\mathbf{X}}(\mathbf{x})$ por el estimador de densidad basado en núcleos

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

Obtenemos así el estimador generalizado de Nadaraya-Watson definido por

$$\widehat{m}_{\mathbf{H}}(\mathbf{x}) = \frac{\sum_{i=1}^{n} \mathcal{K}_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i) y_i}{\sum_{i=1}^{n} \mathcal{K}_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i)}. \tag{2.10}$$

Por lo tanto, el estimador de regresión multivariado basado en núcleos es nuevamente una suma pesada de respuestas observadas y_i.

2.5.1. Propiedades estadísticas

En Härdle, Müller, Sperlich y Werwatz (2004) se muestra que, si $\int \mathbf{u}\mathcal{K}(\mathbf{u})\, d\mathbf{u} = 0$, $\int \mathbf{u}\mathbf{u}^{\intercal}\mathcal{K}(\mathbf{u})\, d\mathbf{u} = \mu_2(\mathcal{K})\mathbf{I}_d$, bajo condiciones de regularidad, en el interior del soporte de $f_{\mathbf{X}}$

$$\begin{aligned}
\text{Sesgo}\{\widehat{m}_{\mathbf{H}}|\mathbf{x}_1, \ldots, \mathbf{x}_n\} &\approx \mu_2(\mathcal{K})\frac{\nabla m(\mathbf{x})^{\intercal}\mathbf{H}\mathbf{H}^{\intercal}\nabla f(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} + \frac{1}{2}\mu_2(\mathcal{K})\text{traza}\left(\mathbf{H}^{\intercal}\mathcal{H}_m(\mathbf{x})\mathbf{H}\right) \\
\text{Var}\{\widehat{m}_{\mathbf{H}}|\mathbf{x}_1, \ldots, \mathbf{x}_n\} &\approx \frac{1}{n\det\mathbf{H}}\|\mathcal{K}\|_2^2\frac{\sigma(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}
\end{aligned}$$

donde $\nabla m(\mathbf{x})$ y $\mathcal{H}_m(\mathbf{x})$ indican el gradiente y hessiano de m, respectivamente.

Un caso particular de $\mathcal{K}_{\mathbf{H}}(u)$ corresponde a la situación en que la matrix de suavizado $\mathbf{H} = h_n\mathbf{I}$. En ese caso, (2.10) puede escribirse como

$$\widehat{m}_n(\mathbf{x}) = \frac{\displaystyle\sum_{i=1}^{n} \mathcal{K}\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n}\right) y_i}{\displaystyle\sum_{j=1}^{n} \mathcal{K}\left(\frac{\mathbf{x} - \mathbf{x}_j}{h_n}\right)}. \tag{2.11}$$

Los siguientes resultados pueden verse en Devroye (1978) y establecen la consistencia fuerte uniforme sobre compactos del estimador definido en (2.11). Estos resultados serán utilizados posteriormente en el Capítulo 5 para obtener la consistencia de los estimadores para el modelo aditivo cuando las respuestas son faltantes al azar.

Sea $(\mathbf{x}_i, y_i)_{i=1}^n$ una sucesión de vectores aleatorios independientes e idénticamente distribuídos en $\mathbb{R}^d \times \mathbb{R}$. Sea $(\mathbf{X}, Y)$ un vector aleatorio con la misma distribución que $(\mathbf{x}_i, y_i)$ y $m(\mathbf{x}) = E[Y|\mathbf{X} = \mathbf{x}]$. Denotemos por μ a la medida de probabilidad de $\mathbf{X}$.

Definición 2.5.1. Diremos que $(y_i)_{i=1}^n$ es una sucesión *uniformemente acotada* si $|Y - m(\mathbf{X})| \leq c$ casi seguramente para algún $c < \infty$.

Definición 2.5.2. Diremos que $(y_i)_{i=1}^n$ es una sucesión *uniformemente Gaussiana generalizada* si para algún $\sigma \geq 0$ y $c \geq 0$ se cumple

$$\sup_{\mu} E[e^{\lambda(Y - m(\mathbf{X}))}|\mathbf{X}] \leq e^{\frac{\sigma^2\lambda^2}{2(1-|\lambda|c)}}, \text{ para todo } \quad |\lambda| \leq \frac{1}{c}\,.$$

Observación 2.5.1. Si las observaciones son uniformemente acotadas es fácil ver que son uniformemente Gaussianas generalizadas. También lo son si son normales, es decir, si $Y|\mathbf{X} = \mathbf{x} \sim N(m(\mathbf{x}), \sigma^2(\mathbf{x}))$ y $\sup_{\mathbf{x}\in\mathbb{R}^d} \sigma^2(\mathbf{x}) < \infty$).

Sea $\|.\|$ una norma en $\mathbb{R}^d$. Consideraremos las siguientes hipótesis que pueden dividirse en tres grupos, hipótesis para el núcleo $\mathcal{K}$, hipótesis sobre la sucesión h_n y una hipótesis sobre la medida de probabilidad μ.

Hipótesis sobre el núcleo

K1. $\mathcal{K} : \mathbb{R}^d \to \mathbb{R}$ es no negativa, acotada y $\int \mathcal{K}(\mathbf{u})\, d\mathbf{u} < \infty$.

K2. $\mathcal{K}(\mathbf{x}) = K(\|\mathbf{x}\|)$ para alguna función no decreciente K tal que

i) $u^d K(u) \to 0$ cuando $u \to \infty$,

ii) $K(u^*) > 0$ para algún $u^* > 0$.

Hipótesis sobre el parámetro de suavizado

H1. $h_n \to 0$ y $nh_n^d / \log n \to \infty$.

Hipótesis sobre la medida marginal de X

X1. Existen $a, b > 0$ tal que $\inf_{\mathbf{x}\in A} \mu(\mathcal{S}(\mathbf{x}, r)) \geq ar^d$, para todo $r \in [0, b]$, donde $\mathcal{S}(\mathbf{x}, r)$ es la esfera cerrada con centro $\mathbf{x}$ y radio r.

Observación 2.5.2. Supongamos que $\mathbf{X}$ tiene densidad $f_{\mathbf{X}}$ continua y estrictamente positiva en su soporte $\text{sop}(f_X)$. Luego, si A es compacto y $A \subset \text{sop}(f_X)$, se cumple **X1**.

Luego, tenemos el siguiente teorema

Teorema 2.5.3. *Sea $(\mathbf{x}_i, y_i)_{i=1}^n$ una sucesión de vectores aleatorios independientes e idénticamente distribuídos y tal que $(y_i)_{i=1}^n$ es una sucesión uniformemente Gaussiana generalizada. Sea $\widehat{m}_n(\mathbf{x})$ el estimador de Nadaraya-Watson definido en (2.11). Sea A un conjunto compacto y supongamos que se cumplen **K1**, **K2**, **H1** y **X1** y que además m es acotada y continua en el soporte de μ, entonces*

$$\sup_{\mathbf{x}\in A} |\widehat{m}_n(\mathbf{x}) - m(\mathbf{x})| \xrightarrow{c.s.} 0 \, .$$

Capítulo 3

Modelos aditivos y efectos marginales

3.1. Introducción

Sea Y la variable dependiente y $\mathbf{X}$ un vector d-dimensional de variables explicativas. Consideremos la estimación de la función de regresión $m(\mathbf{X}) = E(Y|\mathbf{X})$. Hemos visto que cuando $d = 1$ y la función de regresión tiene segunda derivada continua, la tasa del error cuadrático medio asintótico era $n^{4/5}$. Para el caso multidimensional, Stone (1985) mostró que la tasa de convergencia óptima para estimar m es $n^{\kappa/(2\kappa+d)}$ con κ el índice de suavidad de m. Por lo tanto, altos valores de d llevan a una tasa de convergencia más lenta. Este hecho está asociado al fenómeno de entornos cada vez más ralos al aumentar la dimensión que hemos descripto en el Capítulo 1. Para resolver este problema, se introdujeron los modelos aditivos. Una función de regresión m satisface un modelo aditivo si tiene la forma

$$m(\mathbf{X}) = c + \sum_{\alpha=1}^{d} g_\alpha(X_\alpha) \tag{3.1}$$

donde g_α son funciones noparamétricas unidimensionales que operan en cada elemento del vector de variables explicativas. Stone (1985) mostró también que bajo un modelo aditivo la tasa óptima para estimar m es la tasa uno–dimensional $n^{\kappa/(2\kappa+1)}$. Hablamos entonces de una reducción de la dimensión. Los modelos aditivos combinan la flexibilidad de los modelos noparamétricos con la simple interpretación del modelo lineal estándar.

Buja, Hastie y Tibshirani (1989) y Hastie y Tibshirani (1990) propusieron un procedimiento iterativo denominado *backfitting* para estimar las componentes aditivas. Más recientemente, Tjøstheim y Auestad (1994) y Linton y Nielsen (1995) introdujeron un método no iterativo para estimar los efectos marginales que definimos como la esperanza de Y respecto del error aleatorio ϵ y de todas las covariables excepto X_α que se deja fija. El efecto marginal dice cómo varía Y en promedio al variar X_α. Notemos que los efectos marginales coinciden con las componentes aditivas g_α, excepto por la constante c, si la verdadera función de regresión m es efectivamente aditiva. La idea de este método consiste en primero estimar el funcional multidimensional m y luego usar el procedimiento denominado *integración marginal* que describiremos en la sección 3.2 para obtener los efectos marginales.

Antes de comenzar con la descripción del procedimiento de integración marginal, recordemos

el modelo que utilizaremos de aquí en más. Para mayor generalidad supondremos un modelo de regresión heteroscedástico, es decir, supondremos que $Y = m(\mathbf{X}) + \sigma(\mathbf{X})\epsilon$ donde $E(\epsilon|\mathbf{X}) = 0$ y $Var(\epsilon|\mathbf{X}) = 1$. En general, llamaremos $u = \sigma(\mathbf{X})\epsilon$ a la componente del error. Consideraremos el problema de estimar, basados en la muestra aleatoria $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, la esperanza condicional de Y dado $\mathbf{X}$, o sea, la función $m(\mathbf{x})$ que supondremos satisface el modelo aditivo (3.1). Las funciones componentes $g_\alpha(.)$ de (3.1) explican el impacto específico de la componente particular X_α en la respuesta Y. Para que el modelo sea identificable, necesitamos agregar alguna restricción. Consideraremos que

$$E_{X_\alpha}\{g_\alpha(X_\alpha)\} = 0 \qquad \forall\alpha\ . \tag{3.2}$$

Luego, $E(Y) = c$. La constante c puede ser fácilmente estimada con un orden de convergencia $\sqrt{n}$ mediante el estimador $\widehat{c} = \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$, por lo que en cuanto al comportamiento asintótico de los estimadores de g_α no hay pérdida de generalidad al suponer que $c = 0$.

3.2. Estimador de integración marginal

Este procedimiento busca estimar los efectos marginales de las variables regresoras X_α. Indiquemos por $\mathbf{X}_{\underline{\alpha}}$ al vector de todas las variables explicatorias menos la X_α, es decir, $\mathbf{X}_{\underline{\alpha}} = (X_1, \ldots, X_{\alpha-1}, X_{(\alpha+1)}, \ldots, X_d)^{\intercal}$ y por $f_{\underline{\alpha}}$ su densidad conjunta que suponemos existe. El efecto marginal de una variable explicatoria dice cómo cambia Y en promedio si esa variable varía. En otras palabras, el efecto marginal representa la esperanza "condicional" que indicaremos $E_{\epsilon, X_{\underline{\alpha}}}(Y|X_\alpha)$, donde la esperanza no es sólo respecto del error de distribución sino también de todas las variables regresoras excepto X_α que permanece fija. Observemos que hasta este momento habíamos suprimido el ϵ en las esperanzas condicionales, éste es el único caso donde necesitamos mencionar explícitamente en qué distribución estamos calculando la esperanza.

Como ya lo hemos mencionado, cuando se cumple el modelo aditivo (3.1), los efectos marginales corresponden a las funciones componentes aditivas $g_\alpha + c$. El estimador aquí está basado en la idea de integración, proveniente de la siguiente observación. Notemos por f_α la densidad marginal de X_α. Por (3.2) tenemos que $E_{X_\alpha} g_\alpha(X_\alpha) = \int g_\alpha(t) f_\alpha(t)\, dt = 0$, para todo $1 \le \alpha \le d$. Si ahora $m(\mathbf{X})$ cumple el modelo aditivo (3.1), tenemos que

$$\int m(\mathbf{x}) f_{\underline{\alpha}}(\mathbf{x}_{\underline{\alpha}}) \prod_{k\neq\alpha} dX_k \;=\; E_{\mathbf{X}_{\underline{\alpha}}}\left\{ c + g_\alpha(X_\alpha) + \sum_{k\neq\alpha} g_k(X_k) \right\} = c + g_\alpha(X_\alpha). \tag{3.3}$$

Notemos que la expresión anterior coincide con $E_{\epsilon, \mathbf{X}_{\underline{\alpha}}}(Y|X_\alpha)$. Para ilustrar el procedimiento de integración marginal daremos un ejemplo.

Ejemplo 3.2.1. *Supongamos que tenemos un proceso de generación de datos que cumplen el modelo*

$$Y = 4 + X_1^2 + 2\sin X_2 + u,$$

donde $X_1 \sim U[-2, 2]$ *y* $X_2 \sim U[-3, 3]$ *y* u *es una variable aleatoria tal que* $E(u) = 0$ *pero eventualmente heteroscedástica. La función de regresión es igual a* $m(x_1, x_2) = E(Y|\mathbf{X} = \mathbf{x}) = 4 + x_1^2 + 2\sin x_2$. *En consecuencia, tenemos esperanzas marginales*

$$E_{X_2}\{m(X_1, X_2)\} = \int_{-3}^{3} \frac{1}{6}\left\{4 + X_1^2 + 2.\sin u\right\} du = 4 + X_1^2$$

$$E_{X_1}\{m(X_1, X_2)\} = \int_{-2}^{2} \frac{1}{4}\left\{4 + u^2 + 2.\sin X_2\right\}\, du = \frac{16}{3} + 2\sin X_2\ .$$

Las componentes aditivas deben estar normalizadas de modo tal que $E_{X_\alpha} g_\alpha(X_\alpha) = 0$. *Por lo tanto, obtenemos las siguientes funciones componentes* $g_1(x_1) = x_1^2 - \frac{4}{3}$ *y* $g_2(x_2) = 2\sin x_2$ *y la constante* c *es igual a* $c = \frac{16}{3}$.

Para estimar simultáneamente las funciones y sus derivadas se combina el procedimiento de integración marginal con aproximaciones localmente polinomiales. Una aproximación basada en polinomios de orden 0, o sea, ajustando localmente una constante, es posible si sólo interesa estimar la función de regresión y no las derivadas de cada componente, esta aproximación es la que presentaremos en la sección 3.2.1. Una propuesta que mejora a la anterior en presencia de un número alto de covariables se presenta en la sección 3.2.2. Discutiremos brevemente en la sección 3.2.3 cómo estimar los términos de interacción cuando el modelo aditivo no se cumple exactamente.

3.2.1. Estimación de los efectos marginales

A fin de estimar el efecto marginal $g_\alpha(x)$, la ecuación (3.3) sugiere el siguiente procedimiento. En primer lugar, estimar la función m con un suavizador multidimensional $\widetilde{m}$ y luego integrar sobre todas las variables salvo X_α respecto de un estimador de la densidad $f_{\underline{\alpha}}$. En la estimación, el procedimiento de integración puede ser reemplazado por un promedio tomado en todas las componentes que no son la de interés, es decir, sobre $\mathbf{X}_{\underline{\alpha}}$. El estimador resulta ser entonces

$$\widehat{\{g_\alpha(x) + c\}} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{m}(x, \mathbf{x}_{i\alpha}),$$

donde por simplicidad de notación, indicamos por $(x, \mathbf{x}_{i\alpha})$ al vector que tiene la componente α igual a x y todas las demás iguales a x_{ij}, $j \neq \alpha$. Notemos que para obtener los efectos marginales, sólo integramos $\widetilde{m}$ sobre todas las otras direcciones $\underline{\alpha}$. En caso de aditividad, estos efectos marginales son las funciones componentes aditivas g_α más la constante c que, como ya hemos mencionado, puede ser estimada consistentemente con tasa $\sqrt{n}$ usando $\widehat{c} = \bar{y}$. Por lo tanto, un posible estimador para g_α es

$$\widehat{g}_\alpha(x) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{m}(x, x_{i\alpha}) - \bar{y}.$$

El estimador obtenido centrando las marginales, es decir, si definimos el estimador de g_α como

$$\widehat{g}_\alpha(x) = \widehat{\{g_\alpha(x) + c\}} - \frac{1}{n}\sum_{i=1}^{n}\widehat{\{g_\alpha(x_{i\alpha}) + c\}} = \frac{1}{n}\sum_{i=1}^{n}\widetilde{m}(x, x_{i\alpha}) - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n}\sum_{\ell=1}^{n}\widetilde{m}(x_{i\alpha}, \mathbf{x}_{\ell\underline{\alpha}}) \qquad (3.4)$$

tiene el mismo comportamiento asintótico que el obtenido centrando usando $\bar{y}$.

Nos queda por discutir cómo obtener un estimador razonable $\widetilde{m}(x_\alpha, \mathbf{x}_{\underline{\alpha}})$. En principio, éste podría ser cualquier estimador noparamétrico multivariado de los descriptos en el Capítulo 2. En particular, si usamos el estimador de núcleos

$$\widetilde{m}_n(x_\alpha, \mathbf{x}_{\underline{\alpha}}) = \frac{\sum_{i=1}^{n} K_{h_n}(x_{i\alpha} - x_\alpha)\mathcal{K}_{\mathbf{H}}(\mathbf{x}_{i\underline{\alpha}} - \mathbf{x}_{\underline{\alpha}})y_i}{\sum_{j=1}^{n} K_{h_n}(x_{j\alpha} - x_\alpha)\mathcal{K}_{\mathbf{H}}(\mathbf{x}_{j\underline{\alpha}} - \mathbf{x}_{\underline{\alpha}})}\ ,$$

obtenemos

$$\widehat{g}_\alpha(x_\alpha) = \frac{1}{n}\sum_{\ell=1}^{n} \frac{\sum_{i=1}^n K_{h_n}(x_{i\alpha} - x_\alpha)\mathcal{K}_{\mathbf{H}}(\mathbf{x}_{i\underline{\alpha}} - \mathbf{x}_{\underline{\alpha}})y_i}{\sum_{j=1}^n K_{h_n}(x_{j\alpha} - x_\alpha)\mathcal{K}_{\mathbf{H}}(\mathbf{x}_{j\underline{\alpha}} - \mathbf{x}_{\underline{\alpha}})} - \bar{y}.$$

Estimadores basado en polinomios locales pueden verse en Härdle, Müller, Sperlich y Werwatz (2004).

3.2.2. Estimación de tasa óptima basada en integración marginal en presencia de muchas covariables

Para el caso bivariado, Linton y Nielsen (1995) mostraron que integrar el estimador de Nadaraya-Watson produce estimadores de las componentes marginales que son asintóticamente normales con tasa de convergencia óptima. Algunos desarrollos heurísticos, basadas en la consistencia del estimador piloto sugieren que el estimador no convergería con tasa de convergencia óptima en presencia de más de cuatro covariables. Para resolver este problema, Hengartner y Sperlich (2005) propusieron un estimador inicial $\widetilde{m}$ de la función de regresión internamente normalizado tal que la integración marginal aplicada a este estimador da como resultado estimadores de las componentes marginales con tasa óptima. La demostración dada por dichos autores revela que el estimador piloto sobresuavizará las variables a ser integradas y que el estimador resultante es en sí mismo un suavizador de regresión de menor dimensión. Describiremos brevemente su propuesta.

Consideremos la muestra de observaciones i.i.d. $(\mathbf{x}_i, y_i) \in \mathbb{R}^{d+1}$, $1 \leq i \leq n$ y subdividamos las componentes de las covariables $\mathbf{x}_i = (\mathbf{x}_{i,1}^\intercal, \mathbf{x}_{i,2}^\intercal)^\intercal \in \mathbb{R}^d$, donde $\mathbf{x}_{i,j} \in \mathbb{R}^{d_j}$, $j = 1, 2$, $d_1 + d_2 = d$. Supongamos además que las observaciones tienen densidad conjunta $f(\mathbf{x}, y) = f(y|\mathbf{x})f(\mathbf{x})$. Definamos la medida producto Q en $\mathbb{R}^d$ como $Q_2(\mathbf{x}_2) = Q(\mathbb{R}^{d_1}, \mathbf{x}_2)d\mathbf{x}_1$ y sea $qd\mathbf{x} = dQ$, $q_2 d\mathbf{x}_2 = dQ_2$. Sea $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2)^\intercal$. Luego, el impacto marginal de $\mathbf{X}_1$ en Y es

$$\eta_1(\mathbf{u}_1) = E_Q[m(\mathbf{X}_1, \mathbf{X}_2)|\mathbf{X}_1 = \mathbf{u}_1] - \mu = \int_{\mathbb{R}^{d_2}} m(\mathbf{u}_1, \mathbf{u}_2)q_2(\mathbf{u}_2)d\mathbf{u}_2 - \mu \tag{3.5}$$

con μ una constante arbitraria. Por lo tanto, η_1 es la proyección $L_2(Q)$ de m en el espacio de funciones de $\mathbf{u}_1$.

Supongamos que la esperanza condicional $m(\mathbf{u})$ es aditiva en $\mathbf{u}_1$ y $\mathbf{u}_2$, es decir, $m(\mathbf{u}) = \mu + g_1(\mathbf{u}_1) + g_2(\mathbf{u}_2)$. Para asegurar la identificabilidad del modelo, agregamos como antes las condiciones

$$E_Q[g_1(\mathbf{X}_1)] = 0 = E_Q[g_2(\mathbf{X}_2)]. \tag{3.6}$$

Obviamente, en este caso, $\eta_1 = g_1$.

Como hemos mencionado, la idea central del método de integración es imitar (3.5) y estimar la función de impacto marginal $\eta_1(\mathbf{u}_1)$ integrando un estimador piloto $m_n(\mathbf{u})$ de la función de regresión multivariada $m(\mathbf{u})$ con respecto a la medida de probabilidad $Q_2(\mathbf{u}_2)$ en $\mathbb{R}^{d_2}$.

Para ilustrar esto, consideremos la aplicación del método de integración al suavizador de regresión multivariado Nadaraya-Watson

$$\widetilde{m}_n(\mathbf{u}) = \frac{\sum_{j=1}^n W_{nj}(\mathbf{u})y_j}{\sum_{j=1}^n W_{nj}(\mathbf{u})},$$

con pesos

$$W_{nj}(\mathbf{u}) = \prod_{\ell=1}^{d} \frac{1}{h_\ell} K\left(\frac{u_\ell - x_{j,\ell}}{h_\ell}\right),$$

siendo K un núcleo fijo y $h_1, \ldots, h_d > 0$ los parámetros de suavizado. Para que este estimador sea consistente se requiere que $\lim_{n\to\infty} n \prod_{\ell=1}^{d} h_\ell = \infty$. Para funciones de regresión m diferenciables de orden 2, Hengartner y Sperlich (2005) mostraron que el análisis asintótico para el estimador de η_1 dado por

$$\widehat{\eta}_1(\mathbf{u}_1) = \int \widetilde{m}_n(\mathbf{u}_1, \mathbf{u}_2) q_2(\mathbf{u}_2)\, d\mathbf{u}_2 - \int \widetilde{m}_n(\mathbf{x}) q(\mathbf{x})\, d\mathbf{x}$$

revela que tiene varianza y sesgo cuadrado que son de orden $(n \prod_{j=1}^{d_1} h_j)^{-1}$ y $\max_{1\le\ell\le d} h_\ell^4$ respectivamente. Esto es porque la integración actúa como un promedio reduciendo la varianza del estimador piloto pero dejando el orden del sesgo cuadrado sin cambios. Un balance óptimo para el sesgo cuadrado y varianza es posible solamente cuando $d < 4 + d_1$. Esto muestra que el estimador de integración marginal no tiene tasa óptima en presencia de una cantidad arbitraria de covariables si utilizamos como estimador inicial el estimador de Nadaraya–Watson. En otras palabras, el método de integración sufre de la *maldición de la dimensión*.

Linton (1997) probó que la varianza del estimador de integración marginal crece al crecer la correlación entre las covariables con lo cual, resulta ineficiente, al menos cuando el modelo es puramente aditivo.

Para simplificar la notación, supongamos que $\mu = 0$ y $d_1 = 1$, o sea, $\mathbf{u} = (u_1, \mathbf{u}_2)$ con $\mathbf{u}_2 = (u_2, \ldots, u_d)$, $\mathbf{x}_{i,1} = x_{i,1}$, $\mathbf{x}_{i,2} = (x_{i,2}, \ldots, x_{i,d})^{\mathrm{T}}$, y $d_2 = d - 1$. Hengartner y Sperlich (2005) demostraron que bajo ciertas condiciones de regularidad la integración del estimador piloto Nadaraya-Watson centrado tiene distribución asintótica normal, es decir, $\sqrt{nh_1}(\widehat{\eta}_1(u_1) - \eta_1(u_1)) \xrightarrow{D} N(b_1(u_1), \tau^2(u_1))$, donde, tal como ocurre con el estimador de regresión de Nadaraya-Watson, el sesgo depende explícitamente de la densidad de las covariables. Esto es indeseable. Por otra parte, como el cómputo del estimador requiere de una integración numérica multivariada, el número de operaciones requeridas es exponencial en la cantidad de covariables. En resumen, este estimador no se libera completamente de la *maldición de la dimensión*.

La idea básica para evitar la maldición de la dimensión, y al mismo tiempo reducir el costo computacional, gira entorno de integrar un estimador piloto adaptado al problema, seguido por una apropiada centralización de la componente aditiva estimada. Supongamos por el momento que conocemos la densidad $f(\mathbf{x})$ de las covariables, y consideremos la estimación de la componente $g_1(u_1)$ de una función de regresión separable $m(u_1, \mathbf{u}_2) = g_1(u_1) + g_2(\mathbf{u}_2)$ que es s veces continuamente diferenciable en u_1.

Sean K_ℓ, $1 \le \ell \le d$ núcleos suaves de orden s, es decir, tales que $\int u^k K_\ell(u)\, du = 0$ para $k = 0, 1, \ldots, s-1$ y $\int u^s K_\ell(u)\, du \neq 0$. Dadas las ventanas $h_1, \ldots, h_d > 0$, llamaremos *estimador de regresión multivariado internamente normalizado* a

$$\widetilde{m}_n(u_1, \ldots, u_d) = \frac{1}{n} \sum_{j=1}^{n} \left(\prod_{\ell=1}^{d} \frac{1}{h_\ell} K_\ell\left(\frac{u_\ell - x_{j,\ell}}{h_\ell}\right) \right) \frac{y_j}{f(\mathbf{x}_j)}.$$

Este estimador piloto es más sencillo de integrar que el estimador de Nadaraya-Watson, y produce un estimador de $\varphi_1 = \eta_1 + \mu$ que puede escribirse como

$$\begin{aligned}
\widetilde{\varphi}_1(u_1) &= \int_{\mathbb{R}^{d-1}} \widetilde{m}_n(u_1, \mathbf{u}_2) q_2(\mathbf{u}_2)\, d\mathbf{u}_2 \\
&= \frac{1}{n}\sum_{j=1}^{n} \frac{1}{h_1} K_1\left(\frac{u_1 - x_{j,1}}{h_1}\right) \frac{1}{f(x_{j,1})} \times \left\{ y_j \int_{\mathbb{R}^{d-1}} \left(\prod_{\ell=2}^{d} \frac{1}{h_\ell} K_\ell\left(\frac{u_\ell - x_{j,\ell}}{h_\ell}\right)\right) \frac{q_2(\mathbf{u}_2)}{f(\mathbf{x}_{j,2}|x_{j,1})}\, d\mathbf{u}_2 \right\} \\
&= \frac{1}{n}\sum_{j=1}^{n} \frac{1}{h_1} K_1\left(\frac{u_1 - x_{j,1}}{h_1}\right) \frac{\zeta_{n,j}}{f(x_{j,1})},
\end{aligned}$$

donde $\zeta_{n,j}$ indica la expresión entre llaves. Hengartner y Sperlich (2005) mostraron que este estimador alcanza la tasa óptima univariada y que más aún las ventanas $h_2, \ldots, h_d$ no necesitan siquiera converger a 0, es decir, podemos sobresuavizar en las demás coordenadas.

En la mayoría de las situaciones la densidad de las covariables no es conocida y por lo tanto, será necesario considerar un estimador de la densidad conjunta. Consideraremos el estimador de la densidad basado en núcleos. De esta forma el estimador inicial queda definido por

$$\widetilde{m}_n(\mathbf{u}) = \frac{1}{n}\sum_{j=1}^{n} \mathcal{K}_{\mathbf{h}}(\mathbf{u} - \mathbf{x}_j) \frac{y_j}{\widehat{f}_n(\mathbf{x}_j)}$$

con $\widehat{f}_n(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^{n} \mathcal{K}_{\mathbf{h}}(\mathbf{x} - \mathbf{x}_j)$ el estimador de la densidad f basada en núcleos y $\mathcal{K}$ es el núcleo producto de los d núcleos univariados K_ℓ con ventanas $\mathbf{h} = (h_1, \ldots, h_d)^{\mathrm{T}}$. Podemos entonces definir el estimador de φ_1 como $\widetilde{\varphi}_1(u_1) = \int_{\mathbb{R}^{d-1}} \widetilde{m}_n(u_1, \mathbf{u}_2) q_2(\mathbf{u}_2)\, d\mathbf{u}_2$. Por lo tanto, estimamos la función de impacto marginal $\eta_1(u_1) = E_Q[m(X_1, \mathbf{X}_2|X_1 = u_1)] - E_Q[m(\mathbf{X})]$ con su contraparte empírica

$$\widehat{\eta}_1(u_1) = \widetilde{\varphi}_1(u_1) - \int \widetilde{\varphi}_1(v_1) q_1(v_1)\, dv_1 = \int_{\mathbb{R}^{d-1}} \widetilde{m}_n(u_1, \mathbf{u}_2) q_2(\mathbf{u}_2)\, d\mathbf{u}_2 - \int_{\mathbb{R}} \widetilde{m}_n(v_1, \mathbf{u}_2) q_2(\mathbf{u}_2) q_1(v_1)\, d\mathbf{u}_2 dv_1.$$

Análogamente a lo obtenido en el caso en que la densidad se suponía conocida, intercambiando los órdenes de integración y la suma en el primer término de la derecha, tenemos que

$$\widetilde{\varphi}_1(u_1) = \int_{\mathbb{R}^{d-1}} \widetilde{m}_n(u_1, \mathbf{u}_2) q_2(\mathbf{u}_2)\, d\mathbf{u}_2 = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{h_1} K\left(\frac{u_1 - x_{j,1}}{h_1}\right) \frac{\zeta_{n,j}}{\widehat{f}_n(x_{j,1})},$$

donde

$$\zeta_{n,j} = \frac{y_j}{\widehat{f}_n(\mathbf{x}_{j,2}|x_{j,1})} \int_{\mathbb{R}^{d-1}} \left(\prod_{\ell=2}^{d} \frac{1}{h_\ell} K\left(\frac{u_\ell - x_{j,\ell}}{h_\ell}\right)\right) q_2(\mathbf{u}_2)\, d\mathbf{u}_2$$

Este es un suavizador de regresión internamente normalizado del arreglo triangular de pares aleatorios i.i.d. $(X_{j,1}, \zeta_{n,j})$. Bajo un modelo aditivo de la forma $E[Y|\mathbf{X} = \mathbf{u}] = g_1(u_1) + g_2(\mathbf{u}_2)$, Hengartner y Sperlich (2005) estudiaron el comportamiento asintótico del estimador y mostraron que el estimador de integración centrado satisface, bajo ciertas condiciones de regularidad, $\sqrt{nh_1}(\widehat{\eta}_1(u_1) - \eta_1(u_1)) \xrightarrow{D} N(b(u_1), \tau^2(u_1))$. La varianza asintótica τ^2 es la misma que la del estimador basado en el estimador de Nadaraya–Watson, mientras que el sesgo asintótico difiere ya que se elimina el sesgo debido al sobresuavizado del estimador piloto en las variables a integrar. Bajo el modelo aditivo $\widehat{\eta}_1(u_1)$ provee un estimador asintóticamente normal de la componente aditiva g_1.

Por otra parte, como las componentes aditivas estimadas son asintóticamente independientes, este resultado puede ser utilizado para demostrar teoremas sobre la distribución asintótica para la reconstrucción aditiva de la función de regresión m, por ejemplo, cuando $d_j = 1$ para todo j y definimos

$$\widehat{m}_n(\mathbf{x}) = \sum_{\ell=1}^{d} \widehat{\eta}_n(x_\ell) + \int_{\mathbb{R}^d} \widetilde{m}_n(\mathbf{z}) q(\mathbf{z})\, d\mathbf{z}.$$

3.2.3. Términos de interacción

Como ya lo habíamos remarcado, la integración marginal estima los efectos marginales. Estos son idénticos a las componentes aditivas si el modelo es verdaderamente aditivo. ¿Pero qué ocurre si el modelo subyacente no es puramente aditivo? ¿Cómo se comportan los estimadores cuando hay cierta interacción entre las variables explicatorias, por ejemplo dada por un término adicional $g_{\alpha j}(X_\alpha, X_j)$?

Una obvia debilidad del modelo aditivo es que estas interacciones son olvidadas por completo. Por esta razón extenderemos el modelo de regresión con interacciones de a pares, resultando en

$$m(\mathbf{x}) = c + \sum_{\alpha=1}^{d} g_\alpha(x_\alpha) + \sum_{1 \le \alpha < j \le d} g_{\alpha j}(x_\alpha, x_j). \tag{3.7}$$

Aquí usamos $1 \le \alpha < j \le d$ para asegurarnos que incluímos cada interacción de un par sólo una vez. En otras palabras, estamos suponiendo que $g_{\alpha j} = g_{j\alpha}$. Principalmente, podemos considerar también términos de interacción de orden mayor a dos, pero esto haría la visualización e interpretación casi imposibles. Más aún, la ventaja de evitar la maldición de la dimensión se perdería paso a paso. Luego, nos restringiremos solamente al caso de interacciones bivariadas.

Para el estimador de integración marginal, términos de interacción bivariados han sido estudiados por Sperlich, Tjøstheim y Yang (2002). Ellos proveen propiedades asintóticas y adicionalmente introducen tests para chequear la significancia de las interacciones.

Para la estimación de (3.7) con integración marginal debemos extender nuestra condición de identificabilidad (3.2) de modo a incluir condiciones que involucren los términos de interacción

$$\int g_{\alpha j}(x_\alpha, x_j) f_\alpha(x_\alpha)\, dx_\alpha = \int g_{\alpha j}(x_\alpha, x_j) f_j(x_j)\, dx_j = 0, \tag{3.8}$$

con f_α, f_j las densidades marginales de X_α y X_j respectivamente.

Como antes, las ecuaciones (3.2) y (3.8) no deben ser consideradas como restricciones sino como condiciones de identificabilidad. Siempre es posible trasladar las funciones g_α y g_j sin cambiar las formas funcionales o la función de regresión total. Es decir, en este modelo, las componentes se identifican salvo una constante aditiva.

De acuerdo a la definición de $\mathbf{X}_{\underline{\alpha}}$, notemos por $\mathbf{X}_{\underline{\alpha j}}$ a la variable aleatoria $(d-2)$–dimensional que se obtiene al remover X_α y X_j de $\mathbf{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$. Denotaremos las densidades marginales de X_α, $\mathbf{X}_{\underline{\alpha j}}$ y $\mathbf{X}$ por $f_\alpha(x_\alpha)$, $f_{\underline{\alpha j}}(\mathbf{x}_{\underline{\alpha j}})$ y $f(\mathbf{x})$ respectivamente.

Si consideramos nuevamente el procedimiento de integración marginal, obtenemos

$$\theta_\alpha(x_\alpha) = \int m(x_\alpha, \mathbf{x}_{\underline{\alpha}}) f_{\underline{\alpha}}(\mathbf{x}_{\underline{\alpha}})\, dx_{\underline{\alpha}},\ 1 \leq \alpha \leq d \tag{3.9}$$

$$\theta_{\alpha j}(x_\alpha, x_j) = \int m(x_\alpha, x_j, \mathbf{x}_{\underline{\alpha j}}) f_{\underline{\alpha j}}(\mathbf{x}_{\underline{\alpha j}})\, d\mathbf{x}_{\underline{\alpha j}}, \tag{3.10}$$

$$c_{\alpha j} = \int g_{\alpha j}(x_\alpha, x_j) f_{\alpha j}(x_\alpha, x_j)\, dx_\alpha dx_j \tag{3.11}$$

para todo par $1 \leq \alpha < j \leq d$. Se puede ver que

$$\theta_{\alpha j}(x_\alpha, x_j) - \theta_\alpha(x_\alpha) - \theta_j(x_j) + \int m(\mathbf{x}) f(\mathbf{x})\, d\mathbf{x} = g_{\alpha j}(x_\alpha, x_j) + c_{\alpha j}\ ,$$

con lo cual, centrando esta función de manera apropiada nos daría la función de interacción de interés.

Usando el mismo procedimiento de estimación descripto anteriormente, es decir, reemplazando las esperanzas por promedios y la función m por un pre-estimador apropiado, obtenemos estimaciones para g_α y para los términos de interacción $g_{\alpha j}$. Una descripción puede verse en Härdle, Müller, Sperlich y Werwatz (2004).

Un punto importante a tener en cuenta en integración marginal es que la estimación de las funciones componentes unidimensionales no es afectada por la inclusión de términos de interacción. Esto significa que, si estimamos bajo el modelo (3.1) o el modelo (3.7) no cambian los resultados de la estimación de las funciones marginales g_α.

Capítulo 4

Estimación noparamétrica de la función de regresión con datos faltantes

4.1. Introducción

El análisis de regresión de datos faltantes fue desarrollado a partir de Yates (1933) quien propuso sustituir los datos faltantes por predicciones basadas en mínimos cuadrados. Con esta idea de imputar valores faltantes en el modelo de regresión lineal, Cochran (1968) redujo el sesgo en estudios observacionales, y Afifi y Elashoff (1969) establecieron resultados sobre sus propiedades asintóticas. De aquí en más, muchos estudios se enfocaron en modelos de regresión lineal con errores normales y modelos log-lineales con datos faltantes.

En un modelo de regresión, la inferencia básica comienza considerando la muestra aleatoria $(\mathbf{x}_i, y_i, \delta_i)$, $1 \leq i \leq n$. Aquí las covariables $\mathbf{x}_i$ son observadas, mientras que la variable respuesta y_i no es completamente observada. Para indicar la presencia o ausencia de respuesta introdujimos la variable indicadora δ_i que toma el valor 1 si y_i se observa mientras que si y_i está faltante $\delta_i = 0$, para $1 \leq i \leq n$. El simple patrón de datos faltantes considerado típicamente se asocia con el esquema de muestra doble propuesto por Neyman (1938) para componer la falta de información de la respuesta Y considerando más (y posiblemente menos costosas) observaciones en las covariables $\mathbf{X}$. En la práctica, estos datos faltantes usualmente ocurren en forma de no–respuestas en muestras basadas en encuestas.

Un tema de fundamental interés es estudiar el impacto de las observaciones faltantes en el funcionamiento de los estimadores basados en núcleos. El efecto de tal falta es precisamente cuantificado con el error cuadrático medio asintótico (AMSE) del suavizador local utilizado. Un estimador imputado que ajusta el efecto de la falta sustituyendo las observaciones faltantes con la respectiva estimación basada en núcleos y que se describe en este Capítulo puede mejorar los resultados obtenidos. El AMSE muestra claramente cómo la función de núcleos y el valor de la ventana usada para construir los sustitutos afectan el funcionamiento del estimador imputado.

En este capítulo, supondremos que los datos son faltantes faltantes al azar (MAR). Con una

aproximación puramente noparamétrica para discutir los datos faltantes, la suposición de MAR requiere la existencia de un mecanismo de aleatoridad, denotado por $p(\mathbf{X})$, tal que

$$P(\delta = 0|\mathbf{X}, Y) = P(\delta = 1|\mathbf{X}) = p(\mathbf{X}) \tag{4.1}$$

vale casi seguramente. Por otro lado, MCAR demandaría que δ sea independiente tanto de $\mathbf{X}$ como de Y, es decir, $p(\mathbf{X})$ es idénticamente igual a una constante p entre 0 y 1. En la práctica, (4.1) debe estar justificada por la naturaleza del experimento cuando resulte legítimo suponer que la ausencia de Y depende principalmente de $\mathbf{X}$. El modelo de regresión noparamétrica que consideraremos para los datos incompletos $(\mathbf{x}_i, y_i, \delta_i)$ estará dado por

$$y_i = m(\mathbf{x}_i) + u_i \tag{4.2}$$

para $1 \leq i \leq n$. Aquí m es la función de regresión, $\mathbf{x}_i$ son los puntos del diseño, u_i son los errores aleatorios tales que $E(u_i) = 0$. En este esquema permitimos un modelo heteroscedástico, en el cual la varianza de u_i puede suponerse dependientes de $\mathbf{x}_i$, $\mathrm{Var}(u_i) = \sigma^2(\mathbf{x}_i)$, es decir, $u_i = \sigma(\mathbf{x}_i)\epsilon_i$ con $E(\epsilon_i) = 0$ y $\mathrm{Var}(\epsilon_i) = 1$.

En contraste de la discusión sobre la estimación de ciertos parámetros globales como la media de Y, o la función de distribución de Y, el objetivo de este Capítulo es estudiar el comportamiento del estimador noparamétrico de regresión. Para datos incompletos, es de importancia considerar un estimador de regresión apropiado y examinar el impacto de las observaciones faltantes en su funcionamiento.

En este Capítulo, presentamos primero los resultados dados por Chu y Chen (1993) y por González–Manteiga y Pérez–Gonzalez (2004) que estudiaron el estimador lineal local (LLS) en el caso de covariables unidimensionales pues, como hemos mencionado, , en el caso de datos completos, tiene la ventaja de tener menor sesgo asintótico en regiones acotadas del soporte de la densidad de X y mejor comportamiento en regiones cercanas a la frontera. Luego, en la sección 4.3 introduciremos los estimadores que son la propuesta de esta tesis para el caso de modelos aditivos.

Para construir el estimador simplificado en el caso de datos incompletos, un esquema sencillo usualmente usado es el bien conocido método de cancelación de a pares (*pairwise deletion method*). Dicho método elimina los puntos del diseño $\mathbf{x}_i$ en $\{(\mathbf{x}_i, y_i, \delta_i)\}_{i=1}^n$ con $\delta_i = 0$ y trata el resto de los datos como un conjunto de datos completos, de esta forma se obtiene un estimador llamado *estimador simplificado*. Por otro lado, una alternativa natural que sigue la idea de Yates (1933), sugiere usar el estimador simplificado en $\mathbf{x}_i$ para imputar la observación faltante y_i, es decir, cuando $\delta_i = 0$. Esto provee un estimador llamado *imputado*. Es de importancia notar que el estimador imputado es un estimador basado en núcleos en dos etapas que posiblemente use dos núcleos y dos amplitudes diferentes.

Para el caso de covariables unidimensionales, Chu y Chen (1993) y González–Manteiga y Pérez–Gonzalez (2004) compararon ambos estimadores basándonos en sus mínimos errores cuadráticos medios asintóticos (AMSE). Si las dos amplitudes usadas en las dos etapas de la construcción del estimador imputado (basado en suavizadores localmente lineales) no son del mismo orden de magnitud, el estimador imputado será inferior al simplificado en términos de AMSE si la ventana del estimador imputado tiene mayor orden que la del simplificado. Mientras que, eligiendo apropiadamente dos núcleos diferentes con las mismas amplitudes para el estimador imputado, puede verse numéricamente que es, en general, preferible usar el estimador imputado. Es interesante que esencialmente el estimador imputado no produce mayor ventaja cuando usamos el mismo núcleo y

la misma amplitud en ambas etapas de su construcción. En este caso, sin embargo, si la amplitud usada en el paso de imputación del estimador imputado es más pequeña que la usada en la construcción de la segunda etapa, el estimador imputado puede ser mejor. En la sección 4.2 resumimos estos resultados.

4.2. Estimadores locales lineales para el caso de covariables univariadas y respuestas faltantes

Chu y Chen (1993) propusieron el siguiente procedimiento para estimar la función de regresión cuando hay respuestas faltantes. Para ello, utilizaron los estimadores localmente lineales, indicados por LLS, que describimos en la sección 2.3 del Capítulo 2. Presentaremos brevemente su enfoque.

4.2.1. Propuesta de Chu y Chen (1993)

El LLS se puede implementar fácilmente usando los datos existentes. Sea $K : \mathbb{R} \to \mathbb{R}$ una función núcleo. Nuestro primer candidato a estimador de $m(x)$ denotado por SLLS es el que se obtiene al eliminar todos los pares incompletos y aplicando a la muestra de datos completos el suavizador local lineal, es decir, minimizando en a_0 y a_1 la función

$$\sum_{i=1}^{n} [y_i - a_0 - a_1(x - x_i)]^2 K((x - x_i)/h)\delta_i. \tag{4.3}$$

Como siempre h es la ventana cuya magnitud es la necesaria para lograr el grado de suavidad local requerido y el valor de h deberá verificar $h \to 0$ y $nh \to \infty$ cuando $n \to \infty$.

Sean $\widehat{a}_0(x)$ y $\widehat{a}_1(x)$ los valores que minimizan el problema (4.3). Es fácil ver que $\widehat{a}_0(x)$ puede expresarse como

$$\widehat{a}_0(x) = \frac{\sum_{i=1}^{n} \alpha_i y_i}{\sum_{i=1}^{n} \alpha_i}$$

donde $\alpha_i = [S_2 - (x - x_i)S_1]K((x - x_i)/h)\delta_i$, con

$$S_\ell = \sum_{i=1}^{n} (x - x_i)^\ell K((x - x_i)/h)\delta_i$$

para todo $\ell \geq 0$. Para evitar que el denominador de $\widehat{a}_0(x)$ sea 0, el estimador simplificado, que indicaremos por SLLS, de $m(x)$ que consideraremos será el dado por

$$\widehat{m}_{\mathrm{S}}(x) = \frac{\sum_{i=1}^{n} \alpha_i y_i}{\sum_{i=1}^{n} \alpha_i + n^{-2}}.$$

El subíndice S en $\widehat{m}_{\mathrm{S}}(x)$ se debe justamente al hecho de estar basado en las observaciones existentes. Se puede ver que el comportamiendo asintótico de $\widehat{m}_{\mathrm{S}}(x)$ no está afectado por la cantidad n^{-2} en el denominador. Sin embargo, para pequeños valores de n, el valor n^{-2} podría ser más grande que $\sum_{i=1}^{n} \alpha_i$ y por lo tanto, tener una notoria influencia sobre $\widehat{m}_{\mathrm{S}}(x)$.

El SLLS $\widehat{m}_{\text{S}}(x)$ sigue el concepto de la eliminación de a pares, lo que claramente facilita su cálculo en la práctica. También resulta natural considerar un estimador en dos etapas, el estimador imputado localmente lineal que indicaremos por ILLS. En una primer etapa, se puede usar un núcleo K y una amplitud h para la construcción de un predictor $\widehat{m}_{\text{S}}(x_i)$ para cada respuesta y_i faltante. Luego, en la segunda etapa consideremos un núcleo L y una amplitud g con los datos obtenidos completando la muestra mediante los predictores definidos. Específicamente, minimizamos la cantidad

$$\sum_{i=1}^{n} [y_i^* - b_0 - b_1(x - x_i)]^2 L((x - x_i)/g) \tag{4.4}$$

donde $y_i^* = \delta_i y_i + (1-\delta_i)\widehat{m}_{\text{S}}(x_i)$. Nuevamente es fácil encontrar una expresión para los valores $\widehat{b}_0(x)$ y $\widehat{b}_1(x)$ que minimizan el problema (4.4) y que corresponde a la dada en la sección 2.3 del Capítulo 2. El estimador ILLS de $m(x)$ se define como

$$\widehat{m}_{\text{I}}(x) = \frac{\sum_{i=1}^{n} \beta_i y_i^*}{\sum_{i=1}^{n} \beta_i + n^{-2}}$$

donde $\beta_i = [T_2 - (x - x_i)T_1]L((x - x_i)/g)$, con

$$T_\ell = \sum_{i=1}^{n} (x - x_i)^\ell L((x - x_i)/g),$$

para $\ell \geq 0$. Aquí el subíndice I en $\widehat{m}_{\text{I}}(x)$ se refiere al hecho de que estamos imputando las observaciones faltantes usando predictores basados en una estimación inicial.

Los dos estimadores $\widehat{m}_{\text{S}}(x)$ y $\widehat{m}_{\text{I}}(x)$ son consistentes si la hipótesis de MAR o MCAR es correcta. Si $p(X) = 1$, ambos estimadores son simplemente el LLS. Por otro lado, si $p(X) = 0$, las estimaciones locales son apenas significativas. Luego, el comportamiento asintótico de los dos estimadores que serán estudiados a continuación supone la función $p(X)$ toma valores siempre positivos. Más aún, será necesario suponer que en el soporte de X el ínfimo de la función $p(X)$ es positivo.

4.2.2. Propiedades asintóticas

El objetivo de esta sección es comparar $\widehat{m}_{\text{S}}(x)$ y $\widehat{m}_{\text{I}}(x)$ respecto de su error cuadrático medio asintótico, AMSE. Para esto, además de la hipótesis MAR en (4.1), necesitaremos algunas notaciones y condiciones extras. Sea $f_X(x)$ y $m''(x)$ la densidad de X y la segunda derivada de m respectivamente, en cada punto de interés x. Las condiciones bajo las cuales se obtiene expansiones para el AMSE son las siguientes.

C1 $f_X(x) > 0$, $p(x) > A > 0$ y las funciones f_X, p, σ^2, y m'' son Lipschitz continuas en un entorno de x.

C2 *a)* el núcleo K tiene soporte en $[-1, 1]$, $K \geq 0$, $\int K(u)du = 1$. Más aún, K es una función par y Lipschitz continua.

b) el núcleo L tiene soporte en $[-1, 1]$, $L \geq 0$, $\int L(u)du = 1$. Más aún, L es una función par y Lipschitz continua.

C3 *a*) la amplitud h_n satisface que $\lim_{n\to\infty} h_n = 0$ con $\lim_{n\to\infty} nh_n = \infty$.

b) la amplitud g_n satisface que $\lim_{n\to\infty} g_n = 0$ con $\lim_{n\to\infty} ng_n = \infty$.

Estas condiciones son condiciones usuales, sin embargo, es importante notar que la elección de L con su amplitud g_n constituirán el punto crucial del siguiente teorema que da la varianza y sesgo asintóticos de $\widehat{m}_{\text{S}}(x)$ y $\widehat{m}_{\text{I}}(x)$. En lo que sigue la notación $\psi_n \triangleright \phi_n$ indica que $\lim_{n\to\infty} \psi_n/\phi_n = 0$.

Teorema 4.2.1. *Bajo los supuestos **C1** a **C3**, la varianza y el sesgo de $\widehat{m}_{\text{S}}(x)$ y de $\widehat{m}_{\text{I}}(x)$ pueden expresarse asintóticamente a través de*

$$\begin{aligned}
Var[\widehat{m}_{\text{S}}(x)] &= n^{-1}h_n^{-1}\sigma^2(x)f(x)^{-1}p(x)^{-1}\int K^2(z)\,dz + o(n^{-1}h_n^{-1} + h_n^4) && (4.5)\\
Sesgo[\widehat{m}_{\text{S}}(x)] &= \frac{1}{2}m''(x)h_n^2\int z^2K(z)\,dz + o(h_n^2), && (4.6)\\
Var[\widehat{m}_{\text{I}}(x)] &= n^{-1}g_n^{-1}\sigma^2(x)f(x)^{-1}v_i(x) + o(n^{-1}g_n^{-1} + g_n^4 + h_n^4) && (4.7)\\
Sesgo[\widehat{m}_{\text{I}}(x)] &= \frac{1}{2}m''(x)\left[g_n^2\int z^2L(z)\,dz + h_n^2q(x)\int z^2K(z)\,dz\right] + o(h_n^2 + g_n^2) && (4.8)
\end{aligned}$$

respectivamente, donde $q(x) = 1 - p(x)$. Por otra parte, $v_i(x)$, para $i = 1, 2, 3$, indica las funciones

$$\begin{aligned}
v_1(x) &= p(x)^{-1}\int L^2(z)\,dz + (g_n/h_n).\left(q(x)^2p(x)^{-1}\int K^2(z)\,dz + 2q(x)K(0)\right)\\
v_2(x) &= p(x)^{-1}\int (p(x)L(z) + q(x)\alpha^{-1}A(z))^2\,dz\\
v_3(x) &= p(x)^{-1}\int L^2(z)\,dz
\end{aligned}$$

donde la función A está definida por $A(z) = \int L(u)K((z-u)/\alpha)\,du$, en los casos en que $h_n \triangleright g_n$, $h_n = \alpha g_n$, para alguna constante $\alpha > 0$, y $g_n \triangleright h_n$, respectivamente.

En el caso en que la hipótesis de MCAR valga, los resultados del Teorema 4.2.1 siguen valiendo con $p(x)$ y $q(x)$ reemplazados por una constante $p \in (0,1)$ y $q = 1 - p$, respectivamente.

Observación 4.2.2. (*El efecto de observaciones faltantes en el* SLLS). Por (4.5), las observaciones faltantes tiene el efecto de incrementar la varianza asintótica del $\widehat{m}_{\text{S}}(x)$ en un factor de escala $p(x)^{-1}$. Acorde con este resultado, cuantas más obervaciones faltantes ocurrieron en un entorno del punto x, más grande es la varianza asintótica de $\widehat{m}_{\text{S}}(x)$. Por otro lado, por (4.6), las observaciones faltantes no tienen efecto en el sesgo asintótico de $\widehat{m}_{\text{S}}(x)$.

Observación 4.2.3. (*Estimación de $p(x)$*). El valor de $p(x)$ puede estimarse utilizando un LLS sobre las variables (x_i, δ_i). Como en (4.4), reemplazando y_i^* por δ_i, obtenemos que el estimador $\widehat{p}(x)$ de $p(x)$ puede expresarse como

$$\widehat{p}(x) = \frac{\sum_{i=1}^n \beta_i\delta_i}{\sum_{i=1}^n \beta_i + n^{-2}}\ .$$

Bajo las hipótesis anteriores, si p tiene segunda derivada continua p'', Fan (1993) mostró que la varianza y el sesgo de $\widehat{p}(x)$ pueden expresarse asintóticamente como

$$\begin{aligned}\mathrm{Var}[\widehat{p}(x)] &= n^{-1}h^{-1}f(x)^{-1}p(x)q(x)\int K^2 + o(n^{-1}h^{-1}+h^4),\\ \mathrm{Sesgo}[\widehat{p}(x)] &= \frac{1}{2}h^2p^{(2)}(x)\int z^2K + o(h^2),\end{aligned}$$

respectivamente.

Observación 4.2.4. (*Comparación entre los comportamientos del* SLLS *y el* ILLS). Ahora compararemos el comportamiento del $\widehat{m}_{\mathrm{S}}(x)$ y el $\widehat{m}_{\mathrm{I}}(x)$ en el AMSE cuando $K = L$. Por (4.5)-(4.8), si $g_n \triangleright h_n$, el AMSE de $\widehat{m}_{\mathrm{S}}(x)$ es el mismo que el de $\widehat{m}_{\mathrm{I}}(x)$. En este caso, $\widehat{m}_{\mathrm{S}}(x)$ y $\widehat{m}_{\mathrm{I}}(x)$ tienen el mismo comportamiento en la estimación de $m(x)$. Por otro lado, si $h_n \triangleright g_n$, el comportamiento de $\widehat{m}_{\mathrm{S}}(x)$ es mejor que el de $\widehat{m}_{\mathrm{I}}(x)$, en el sentido de menor valor de AMSE y eficiencia computacional.

Sin embargo, si $h_n = \alpha g_n$, para $\alpha > 0$, el comportamiento de $\widehat{m}_{\mathrm{I}}(x)$ puede ser mejor que el de $\widehat{m}_{\mathrm{S}}(x)$, en el sentido de menor valor de AMSE. Los estimadores que propondremos se basan en promedios locales, una aproximación basada en ajustes lineales locales también sería posible pero implica un mayor costo computacional teniendo en cuenta que las covariables son de dimensión $d > 1$.

4.3. Estimadores en modelos de regresión noparamétrica aditivos con respuestas faltantes

En esta sección supondremos que las observaciones $(\mathbf{x}_i, y_i, \delta_i)$ son i.i.d. y cumplen (4.1) y satisfacen el modelo (4.2) donde m cumple el modelo aditivo (3.1). Porpondremos dos familias de estimadores,

- los basados en la muestra completa, o sea, eliminando todos los pares incompletos y que llamaremos simplificados, por analogía a la sección anterior
- los obtenidos imputando las observaciones faltantes mediante el estimador simplificado.

Para el estimador imputado, el procedimiento de imputación puede realizarse usando predictores basados en el modelo aditivo o predictores basados en un estimador de la función de regresión, como el de Nadaraya–Watson, que para dimensiones pequeñas no presentará problemas computacionales asociados a la *maldición de la dimensión*. Por otra parte, como describimos en la sección 3.2.1, los estimadores del modelo aditivo podrán basarse en el estimador de Nadaraya–Watson o en el estimador internamente normalizado propuesto por Hengartner y Sperlich (2005).

Consideraremos funciones núcleos multivariadas $\mathcal{K}$ y $\mathcal{L}$ tales que $\mathcal{K} : \mathbb{R}^d \to \mathbb{R}$, $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$, $\mathcal{K} \geq 0$, $\mathcal{L} \geq 0$, $\int \mathcal{K}(\mathbf{u})\,d\mathbf{u} = 1$, $\int \mathbf{u}\mathcal{K}(\mathbf{u})\,d\mathbf{u} = 0$, $\int \mathbf{u}\mathbf{u}^{\mathrm{T}}\mathcal{K}(\mathbf{u})\,d\mathbf{u} = \mu_2(\mathcal{K})\mathbf{I}_d$, $\int \mathcal{L}(\mathbf{u})\,d\mathbf{u} = 1$, $\int \mathbf{u}\mathcal{L}(\mathbf{u})\,d\mathbf{u} = 0$, $\int \mathbf{u}\mathbf{u}^{\mathrm{T}}\mathcal{L}(\mathbf{u})\,d\mathbf{u} = \mu_2(\mathcal{L})\mathbf{I}_d$. Por otra parte, $\mathbf{H}$ y $\mathbf{\Gamma}$ indicarán matrices en $\mathbb{R}^{d\times d}$ no singulares y $\mathcal{K}_{\mathbf{H}}(\mathbf{u}) = (\det(\mathbf{H}))^{-1}\mathcal{K}(\mathbf{H}^{-1}\mathbf{u})$.

4.3.1. Estimador simplificado para el modelo aditivo

De acuerdo a lo descripto en la sección 3.2.1 y utilizando el conjunto de datos completo $\{(\mathbf{x}_i, y_i)\}_{\delta_i=1}$ podemos introducir dos estimadores preliminares de $\widetilde{m}$ que indicaremos $\widetilde{m}_{\mathrm{S}}^{(1)}$ y $\widetilde{m}_{\mathrm{S}}^{(2)}$ y que están definidos como

$$\widetilde{m}_{\mathrm{S}}^{(1)}(\mathbf{x}) = \frac{\sum_{i=1}^{n} \mathcal{K}_{\mathbf{H}}(\mathbf{x}-\mathbf{x}_i)\delta_i y_i}{\sum_{j=1}^{n} \mathcal{K}_{\mathbf{H}}(\mathbf{x}-\mathbf{x}_j)\delta_j} \tag{4.9}$$

$$\widetilde{m}_{\mathrm{S}}^{(2)}(\mathbf{x}) = \sum_{i=1}^{n} \frac{\mathcal{K}_{\mathbf{H}}(\mathbf{x}-\mathbf{x}_i)\delta_i y_i}{\sum_{j=1}^{n} \mathcal{K}_{\mathbf{H}}(\mathbf{x}_i-\mathbf{x}_j)\delta_j} \,. \tag{4.10}$$

Sea $\widehat{c}$ un estimador de $c = E(Y)$, por ejemplo, podemos considerar los siguientes estimadores de c

$$\widehat{c}^{(1)} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{m}_{\mathrm{S}}^{(1)}(\mathbf{x}_i) \qquad\qquad \widehat{c}^{(2)} = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i y_i}{p(\mathbf{x}_i)} \,,$$

donde el último supone que se conoce la probabilidad de falta de respuesta p y ajusta el promedio al hecho de tener respuestas faltantes.

En base a los estimadores definidos en (4.9) y (4.10) se obtienen, respectivamente, dos estimadores de las funciones marginales, el primero basado en el estimador de Nadaraya–Watson y el segundo internamente corregido

$$\widehat{g}_{\alpha,\mathrm{S}}^{(1)}(x_\alpha) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{m}_{\mathrm{S}}^{(1)}(x_\alpha, \mathbf{x}_{\underline{\alpha}i}) - \widehat{c} \tag{4.11}$$

$$\widehat{g}_{\alpha,\mathrm{S}}^{(2)}(x_\alpha) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{m}_{\mathrm{S}}^{(2)}(x_\alpha, \mathbf{x}_{\underline{\alpha}i}) - \widehat{c}. \tag{4.12}$$

De esta manera, el estimador de la función de regresión que hace uso del modelo aditivo queda definido, respectivamente, por

$$\widehat{m}_{\mathrm{S}}^{(1)}(\mathbf{x}) = \sum_{\alpha=1}^{d} \widehat{g}_{\alpha,\mathrm{S}}^{(1)}(x_\alpha) + \widehat{c} \tag{4.13}$$

$$\widehat{m}_{\mathrm{S}}^{(2)}(\mathbf{x}) = \sum_{\alpha=1}^{d} \widehat{g}_{\alpha,\mathrm{S}}^{(2)}(x_\alpha) + \widehat{c} \,. \tag{4.14}$$

4.3.2. Estimador imputado para el modelo aditivo

Como en Chu y Chen (1993), podemos intentar mejorar la estimación de la función de regresión imputando los datos faltantes. En base a los estimadores descriptos, surgen naturalmente tres

maneras de realizar la imputación que indicaremos $y_{i,\mathrm{I}}^{(0)}$, $y_{i,\mathrm{I}}^{(1)}$ y $y_{i,\mathrm{I}}^{(2)}$ y que se definen como

$$\begin{aligned} y_{i,\mathrm{I}}^{(0)} &= \widetilde{m}_{\mathrm{S}}^{(1)}(\mathbf{x}_i), \\ y_{i,\mathrm{I}}^{(1)} &= \widehat{m}_{\mathrm{S}}^{(1)}(\mathbf{x}_i) \\ y_{i,\mathrm{I}}^{(2)} &= \widehat{m}_{\mathrm{S}}^{(2)}(\mathbf{x}_i). \end{aligned}$$

Observemos que la primera de ellas no hace uso del modelo aditivo y por lo tanto, será adecuada cuando se producen alejamientos de dicho modelo. Tiene el inconveniente de sufrir la *maldición de la dimensión*. Los otros predictores de las observaciones se basan en el supuesto de aditividad y es de esperar que tengan mejor comportamiento si el modelo se cumple. De esta forma, con la muestra imputada $(\mathbf{x}_i, \widehat{y}_i^{(\ell)})$ donde

$$\widehat{y}_i^{(\ell)} = \begin{cases} y_i & \text{si } \delta_i = 1 \\ y_{i,\mathrm{I}}^{(\ell)} & \text{si } \delta_i = 0 \end{cases}$$

definimos nuevos estimadores iniciales para la función de regresión. El estimador de Nadaraya–Watson y el internamente corregido que quedan definidos, para $\ell = 0, 1, 2$, por

$$\widetilde{m}_{\mathrm{I}}^{(1,\ell)}(\mathbf{x}) = \frac{\displaystyle\sum_{i=1}^{n} \mathcal{L}_{\mathbf{\Gamma}}(\mathbf{x} - \mathbf{x}_i)\widehat{y}_i^{(\ell)}}{\displaystyle\sum_{j=1}^{n} \mathcal{L}_{\mathbf{\Gamma}}(\mathbf{x} - \mathbf{x}_j)} \tag{4.15}$$

$$\widetilde{m}_{\mathrm{I}}^{(2,\ell)}(\mathbf{x}) = \sum_{i=1}^{n} \frac{\mathcal{L}_{\mathbf{\Gamma}}(\mathbf{x} - \mathbf{x}_i)\widehat{y}_i^{(\ell)}}{\displaystyle\sum_{j=1}^{n} \mathcal{L}_{\mathbf{\Gamma}}(\mathbf{x}_i - \mathbf{x}_j)} \tag{4.16}$$

dando origen a seis posibles estimadores. Mencionemos que como en el caso unidimensional, en esta segunda etapa otro núcleo y otra ventana pueden ser usados.

Con estos nuevos estimadores de m podemos definir los siguientes estimadores de las marginales, para $\ell = 0, 1, 2$,

$$\widehat{g}_{\alpha,\mathrm{I}}^{(1,\ell)}(x_\alpha) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{m}_{\mathrm{I}}^{(1,\ell)}(x_\alpha, \mathbf{x}_{\underline{\alpha}i}) - \widehat{c} \tag{4.17}$$

$$\widehat{g}_{\alpha,\mathrm{I}}^{(2,\ell)}(x_\alpha) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{m}_{\mathrm{I}}^{(2,\ell)}(x_\alpha, \mathbf{x}_{\underline{\alpha}i}) - \widehat{c}\,. \tag{4.18}$$

Finalmente el estimador imputado de la función de regresión, a partir de estos estimadores de las marginales, queda definido por

$$\begin{aligned} \widehat{m}_{\mathrm{I}}^{(1,\ell)}(\mathbf{x}) &= \sum_{\alpha=1}^{d} \widehat{g}_{\alpha,\mathrm{I}}^{(1,\ell)}(x_\alpha) + \widehat{c} \\ \widehat{m}_{\mathrm{I}}^{(2,\ell)}(\mathbf{x}) &= \sum_{\alpha=1}^{d} \widehat{g}_{\alpha,\mathrm{I}}^{(2,\ell)}(x_\alpha) + \widehat{c}\,. \end{aligned}$$

Capítulo 5

Consistencia de los estimadores propuestos para modelos de regresión noparamétrica aditivos con respuestas faltantes

5.1. Introducción

En este Capítulo probaremos la consistencia fuerte sobre compactos de los estimadores simplificado e imputado definidos en la sección 4.3 del Capítulo 4. En la sección 5.2, se enuncian las hipótesis necesarias y la notación utilizada en las demostraciones. En la sección 5.3 se establece la consistencia uniforme de los estimadores simplificados $\widetilde{m}_{\mathrm{S}}^{(1)}(\mathbf{x})$ y $\widetilde{m}_{\mathrm{S}}^{(2)}(\mathbf{x})$ definidos en (4.9) y (4.10). En la sección 5.4, se obtiene la consistencia fuerte de los estimadores de la función de regresión basados en pesos de Nadaraya–Watson e internamente corregidos cuando se imputan las observaciones faltantes mediante un estimador fuertemente y uniformemente consistente. Por simplicidad en el enunciado de los resultados, en esta sección como en la anterior los estimadores están expresados utilizando el mismo núcleo $\mathcal{K}$. En particular, se deduce la consistencia fuerte de los estimadores $\widetilde{m}_{\mathrm{I}}^{(1,\ell)}(\mathbf{x})$ y $\widetilde{m}_{\mathrm{I}}^{(2,\ell)}(\mathbf{x})$, definidos en (4.15) y (4.16) si el núcleo $\mathcal{L}$ cumple las hipótesis necesarias y si la matriz $\mathbf{r} = \gamma_n \mathrm{I}_d$ donde I_d la matriz identidad de $\mathbb{R}^{d\times d}$ con el parámetro de suavizado γ_n satisfaciendo las condiciones requeridas a la ventana. En la sección 5.5, se obtiene la consistencia fuerte de los dos estimadores de la media de Y definidos en la sección 4.3. De esta forma, para obtener la consistencia de las componentes aditivas se podrá suponer, sin pérdida de generalidad, que $c = 0$. Finalmente, en la sección 5.6 obtendremos la consistencia de las componentes aditivas $g_{\alpha,\mathrm{S}}^{(1)}(x_\alpha)$ y $g_{\alpha,\mathrm{S}}^{(2)}(x_\alpha)$ definidas en (4.11) y (4.12), respectivamente. Los resultados de la sección 5.6 se prueban en un marco general y sólo requieren consistencia uniforme del estimador de la regresión utilizado en el proceso de integración marginal. Por lo tanto, incluyen no sólo a los estimadores simplificados antes mencionados, sino que también incluyen a los estimadores imputados $\widehat{g}_{\alpha,\mathrm{I}}^{(1,\ell)}(x_\alpha)$ y $\widehat{g}_{\alpha,\mathrm{I}}^{(2,\ell)}(x_\alpha)$ definidos en (4.17) y (4.18), respectivamente.

5.2. Hipótesis y notación

Sean $(\mathbf{x}_i, y_i, \delta_i)$, $1 \leq i \leq n$, i.i.d con la misma distribución que $(\mathbf{X}, Y, \delta)$ donde $\mathbf{X} = (X_1, \ldots, X_d)^{\intercal} \in \mathbb{R}^d$, $\delta_i = 1$ si y_i es observada y $\delta_i = 0$ si y_i es faltante. A lo largo de este Capítulo, consideremos las siguientes hipótesis

D1. $Y = m(\mathbf{X}) + \sigma(\mathbf{X})\epsilon$ con ϵ tal que $E(\epsilon) = 0$ y $Var(\epsilon) = 1$.

D2. $\mathbf{X}$ tiene densidad f_X de soporte compacto, Lipschitz continua y acotada fuera del cero y del infinito en su soporte $\mathcal{C} = \text{sop}(f_X)$.

D3. $P(\delta = 1|\mathbf{X}, Y) = E[\delta|\mathbf{X}, Y] = E[\delta|\mathbf{X}] = p(\mathbf{X})$, con $p : \mathbb{R}^d \to \mathbb{R}$ continua en $\mathcal{C}$ y tal que $i(p) = \inf_{\mathbf{x}\in\mathcal{C}} p(\mathbf{x}) > 0$.

D4. $m : \mathbb{R}^d \to \mathbb{R}$ y $\sigma : \mathbb{R}^d \to \mathbb{R}^+$ son continuas en $\mathcal{C}$.

D5. Los errores ϵ son independientes de $(\mathbf{X}, \delta)$. Más aún, la sucesión $(\epsilon_i)_{i=1}^n$ es uniformemente Gaussiana generalizada.

D6. La sucesión $(\epsilon_i^2)_{i=1}^n$ es uniformemente Gaussiana generalizada.

De ahora en más indicaremos por $u_j = \sigma(\mathbf{x}_j)\epsilon_j$ y por $u = \sigma(\mathbf{X})\epsilon$, de modo que el modelo dado por **D1** se escribe $Y = m(\mathbf{X}) + u$.

Por simplicidad, supondremos que la matriz $\mathbf{H}$ utilizada en el proceso de suavizado es de la forma $\mathbf{H} = h_n \mathrm{I}_d$ con I_d la matriz identidad de $\mathbb{R}^{d\times d}$. De esta manera, llamaremos

$$\mathcal{K}_{h_n}(\mathbf{x}) = \mathcal{K}_{\mathbf{H}}(\mathbf{x}) = \frac{1}{h_n^d}\mathcal{K}\left(\frac{\mathbf{x}}{h_n}\right).$$

Por completitud en los enunciados, recordaremos a continuación las hipótesis **K1**, **K2** y **H1** dadas en la Sección 2.5.1, que son hipótesis sobre el núcleo y el parámetro de suavizado.

K1. $\mathcal{K} : \mathbb{R}^d \to \mathbb{R}$ es no negativa, acotada y $\int \mathcal{K}(\mathbf{u})\, d\mathbf{u} < \infty$.

K2. $\mathcal{K}(\mathbf{x}) = K(\|\mathbf{x}\|)$ para alguna función no decreciente $K : \mathbb{R}^+ \to \mathbb{R}^+$ tal que

i) $u^d K(u) \to 0$ cuando $u \to \infty$,

ii) $K(u^*) > 0$ para algún $u^* > 0$.

H1. $h_n \to 0$ y $nh_n^d / \log n \to \infty$.

Las siguientes hipótesis se utilizarán para mostrar la consistencia de los estimadores de los efectos marginales bajo el modelo aditivo (3.1). Bajo **D2**, la función de densidad de la componente X_α, que indicaremos f_α tiene soporte compacto que denotaremos $\mathcal{C}_\alpha = \text{sop} f_\alpha$.

A1. $m(\mathbf{x}) = \sum_{\alpha=1}^{d} g_\alpha(x_\alpha) + c$.

A2. $Eg_\alpha(X_\alpha) = 0$ para todo $1 \leq \alpha \leq d$.

A3. g_α funciones continuas en $\mathcal{C}_\alpha$ para todo $1 \leq \alpha \leq d$.

Dada una función $g : \mathbb{R}^d \to \mathbb{R}$ indicaremos por $i(g) = \inf_{\mathbf{x}\in\mathcal{C}} g(\mathbf{x})$ y por $\|g\|_{0,\infty} = \sup_{\mathbf{x}\in\mathcal{C}} |g(\mathbf{x})|$. Por otra parte, si $g : \mathbb{R} \to \mathbb{R}$ indicaremos por $i_\alpha(g) = \inf_{x\in\mathcal{C}_\alpha} g(x)$ y por $\|g\|_{\alpha,\infty} = \sup_{x\in\mathcal{C}_\alpha} |g(x)|$.

Por último, llamaremos $\widehat{m}_Z(\mathbf{x})$ al estimador de Nadaraya–Watson de la función de regresión, $E(Z|\mathbf{X})$, basado en las observaciones $(\mathbf{x}_i, z_i)$ calculado con el núcleo $\mathcal{K}$ y ventana h_n, o sea,

$$\widehat{m}_Z(\mathbf{x}) = \frac{\sum_{i=1}^n \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right) z_i}{\sum_{i=1}^n \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)}\,.$$

5.3. Convergencia uniforme casi segura del estimador simplificado

Comenzaremos probando la consistencia del estimador simplificado $\widetilde{m}_{\mathrm{S}}^{(1)}$ definido en (4.9).

Lema 5.3.1. *Bajo **D1** a **D5**, **K1**, **K2** y **H1**, tenemos que*

a) $\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}_{\delta Y}(\mathbf{x}) - p(\mathbf{x})m(\mathbf{x})| \xrightarrow{c.s.} 0$

b) $\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}_{\delta}(\mathbf{x}) - p(\mathbf{x})| \xrightarrow{c.s.} 0.$

DEMOSTRACIÓN. Comenzaremos probando a). Observemos que, como $\delta Y = \delta m(\mathbf{X}) + \delta u$,

$$\begin{aligned}\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}_{\delta Y} - p(\mathbf{x})m(\mathbf{x})| &= \sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}_{\delta m}(\mathbf{x}) + \widehat{m}_{\delta u}(\mathbf{x}) - p(\mathbf{x})m(\mathbf{x})| \\ &\leq \sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}_{\delta m}(\mathbf{x}) - p(\mathbf{x})m(\mathbf{x})| + \sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}_{\delta u}(\mathbf{x})|.\end{aligned}$$

Por lo tanto, bastará probar que

$$\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}_{\delta m}(\mathbf{x}) - p(\mathbf{x})m(\mathbf{x})| \xrightarrow{c.s.} 0 \tag{5.1}$$

$$\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}_{\delta u}(\mathbf{x})| \xrightarrow{c.s.} 0 \tag{5.2}$$

Por **D4**, m es acotada en $\mathcal{C}$, luego, la sucesión de variables $(\delta_i m(\mathbf{x}_i))_{i=1}^n$ es una sucesión de variables aleatorias independientes, idénticamente distribuídas y uniformemente acotadas tales que $E[\delta m(\mathbf{X})|\mathbf{X}] = m(\mathbf{X})E[\delta|\mathbf{X}] = m(\mathbf{X})p(\mathbf{X})$. De donde, por la Observación 2.5.1, resulta que $(\delta_i m(\mathbf{x}_i))_{i=1}^n$ es una sucesión uniformemente Gaussiana generalizada. Usando el Teorema 2.5.3 se deduce (5.1).

Veamos ahora que la sucesión de variables aleatorias independientes e idénticamente distribuídas $(\delta_i u_i)_{i=1}^n$ también resulta ser una sucesión uniformemente Gaussiana generalizada. Usando que los errores ϵ son independientes de δ y de $\mathbf{X}$ y que $E[\epsilon] = 0$ resulta que $E[\delta u|\mathbf{X}] = p(\mathbf{X})\sigma(\mathbf{X})E[\epsilon] = 0$. Sea $\lambda \in \mathbb{R}$, tenemos que

$$\begin{aligned}E\left[e^{\lambda\delta u}|\mathbf{X}=\mathbf{x}\right] &= E\left[e^{\lambda\delta\sigma(\mathbf{x})\epsilon}|\mathbf{X}=\mathbf{x}\right] = E\left[E\left[e^{\lambda\delta\sigma(\mathbf{x})\epsilon}|\mathbf{X}=\mathbf{x},Y\right]|\mathbf{X}=\mathbf{x}\right] = \\ &= E\left[e^0(1-p(\mathbf{x})) + p(\mathbf{x})\,e^{\lambda\sigma(\mathbf{x})\epsilon}|\mathbf{X}=\mathbf{x}\right] = 1 - p(\mathbf{x}) + p(\mathbf{x})E\left[e^{\lambda\sigma(\mathbf{x})\epsilon}\right]\end{aligned}$$

Como $(\epsilon_i)_{i=1}^n$ es una sucesión de variables aleatorias independientes e idénticamente distribuídas y uniformemente Gaussiana generalizada, existen $\tau \geq 0$ y $c \geq 0$ tales que si $\phi < \frac{1}{c}$, entonces

$$E\left(e^{\phi\epsilon}\right) \leq e^{\frac{\tau^2\phi^2}{2(1-|\phi|c)}}$$

Por **D4** σ es acotada en $\mathcal{C}$, luego si $d = c\|\sigma\|_{0,\infty}^2$ y $\tilde{\tau} = \tau\|\sigma\|_{0,\infty}$ resulta que para todo $\lambda \leq 1/d$ como $\phi = \lambda\sigma(\mathbf{x}) \leq \frac{1}{c}$ entonces

$$\begin{aligned}\sup_{\mathbf{x}\in\mathcal{C}} E\left[e^{\lambda\sigma(\mathbf{x})\epsilon}\right] &\leq \sup_{\mathbf{x}\in\mathcal{C}} e^{\frac{\tau^2\lambda^2\sigma^2(\mathbf{x})}{(1-|\lambda|\sigma(\mathbf{x})c)}} \\ &\leq \sup_{\mathbf{x}\in\mathcal{C}} e^{\frac{\tau^2\lambda^2\|\sigma\|_{0,\infty}^2}{(1-|\lambda|c\|\sigma\|_{0,\infty}^2)}} = e^{\frac{\tilde{\tau}^2\lambda^2}{(1-|\lambda|d)}} .\end{aligned}$$

De esta manera, si $\lambda \leq 1/d$, $1 \leq e^{\tilde{\tau}^2\lambda^2/(1-|\lambda|d)}$ y entonces, resulta que

$$\begin{aligned}\sup_{\mathbf{x}\in\mathcal{C}} E\left[e^{\lambda\delta u}|\mathbf{X}=\mathbf{x}\right] &\leq (1-p(\mathbf{x})) + p(\mathbf{x})e^{\frac{\tilde{\tau}^2\lambda^2}{(1-|\lambda|d)}} \\ &\leq (1-p(\mathbf{x}))e^{\frac{\tilde{\tau}^2\lambda^2}{(1-|\lambda|d)}} + p(\mathbf{x})e^{\frac{\tilde{\tau}^2\lambda^2}{(1-|\lambda|d)}} = e^{\frac{\tilde{\tau}^2\lambda^2}{(1-|\lambda|d)}} ,\end{aligned}$$

con lo cual $(\delta_j u_j)_{j=1}^n$ es una sucesión uniformemente Gaussiana generalizada. Como además es una sucesión de variables aleatorias independientes e idénticamente distribuídas, del Teorema 2.5.3 obtenemos (5.2).

La parte b) se deduce de (5.1) tomando $m \equiv 1$ o utilizando el Teorema 2.5.3 y el hecho que la sucesión de variables aleatorias independientes e idénticamente distribuídas $(\delta_i)_{i=1}^n$ es una sucesión uniformemente acotada y por lo tanto una sucesión uniformemente Gaussiana generalizada. □

Teorema 5.3.1. *Bajo **D1** a **D5**, **K1**, **K2** y **H1**, tenemos que* $\sup_{\mathbf{x}\in\mathcal{C}} |\widetilde{m}_{\mathrm{S}}^{(1)}(\mathbf{x}) - m(\mathbf{x})| \xrightarrow{c.s.} 0$.

DEMOSTRACIÓN. Se deduce inmediatamente del Lema 5.3.1. Efectivamente, sea $\mathcal{N}$ el conjunto de probabilidad 0 tal que $\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}_{\delta Y}(\mathbf{x}) - p(\mathbf{x})m(\mathbf{x})| \not\to 0$ o $\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}_{\delta}(\mathbf{x}) - p(\mathbf{x})| \not\to 0$, luego por **D3**, como $i(p) > 0$, si $\omega \notin \mathcal{N}$ se tiene que para $n \geq n_0$

$$\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}_{\delta}(\mathbf{x}) - p(\mathbf{x})| \leq \frac{i(p)}{2}$$

de donde ,

$$|\widehat{m}_{\delta}(\mathbf{x})| \geq |p(\mathbf{x})| - \sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}_{\delta}(\mathbf{x}) - p(\mathbf{x})| \geq i(p) - \frac{i(p)}{2} = \frac{i(p)}{2}.$$

Observemos que $\widetilde{m}_{\mathrm{S}}^{(1)}$ puede escribirse como

$$\widetilde{m}_{\mathrm{S}}^{(1)}(\mathbf{x}) = \frac{\sum_{i=1}^n \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\delta_i y_i}{\sum_{i=1}^n \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\delta_i} = \frac{\widehat{m}_{\delta Y}(\mathbf{x})}{\widehat{m}_{\delta}(\mathbf{x})} .$$

Por lo tanto, como por **C4** m es acotada en $\mathcal{C}$, si $\omega \notin \mathcal{N}$ y $n \geq n_0$, se obtiene

$$\begin{aligned}
\sup_{\mathbf{x}\in\mathcal{C}}\left|\widetilde{m}_{\mathrm{S}}^{(1)} - m(\mathbf{x})\right| &= \sup_{\mathbf{x}\in\mathcal{C}}\left|\frac{\widehat{m}_{\delta Y}(\mathbf{x}) - m(\mathbf{x})\widehat{m}_{\delta}(\mathbf{x})}{\widehat{m}_{\delta}(\mathbf{x})}\right| \leq \frac{\sup_{\mathbf{x}\in\mathcal{C}}|\widehat{m}_{\delta Y}(\mathbf{x}) - m(\mathbf{x})\widehat{m}_{\delta}(\mathbf{x})|}{\inf_{\mathbf{x}\in\mathcal{C}}|\widehat{m}_{\delta}(\mathbf{x})|} \\
&\leq \frac{\sup_{\mathbf{x}\in\mathcal{C}}|\widehat{m}_{\delta Y}(\mathbf{x}) - m(\mathbf{x})p(\mathbf{x}) + m(\mathbf{x})(p(\mathbf{x}) - \widehat{m}_{\delta}(\mathbf{x}))|}{\inf_{\mathbf{x}\in\mathcal{C}}|\widehat{m}_{\delta}(\mathbf{x})|} \\
&\leq \frac{\sup_{\mathbf{x}\in\mathcal{C}}|\widehat{m}_{\delta Y}(\mathbf{x}) - m(\mathbf{x})p(\mathbf{x})| + \sup_{\mathbf{x}\in\mathcal{C}}|m(\mathbf{x})|\sup_{\mathbf{x}\in\mathcal{C}}|\widehat{m}_{\delta}(\mathbf{x}) - p(\mathbf{x})|}{\inf_{\mathbf{x}\in\mathcal{C}}|\widehat{m}_{\delta}(\mathbf{x})|} \\
&\leq \frac{2}{i(p)}\left\{\sup_{\mathbf{x}\in\mathcal{C}}|\widehat{m}_{\delta Y}(\mathbf{x}) - m(\mathbf{x})p(\mathbf{x})| + \|m\|_{0,\infty}\sup_{\mathbf{x}\in\mathcal{C}}|\widehat{m}_{\delta}(\mathbf{x}) - p(\mathbf{x})|\right\},
\end{aligned}$$

de donde se deduce que $\sup_{\mathbf{x}\in\mathcal{C}}|\widetilde{m}_{\mathrm{S}}^{(1)}(\mathbf{x}) - m(\mathbf{x})| \xrightarrow{c.s.} 0$. □

Procederemos ahora a probar la convergencia uniforme casi segura del estimador $\widetilde{m}_{\mathrm{S}}^{(2)}$ definido en (4.10). Para ello probaremos un Lema cuya demostración es inmediata que resultará de utilidad.

Lema 5.3.2. *Sea $\mathcal{A}$ un compacto, $b(\mathbf{x})$ y $f(\mathbf{x})$ dos funciones continuas en $\mathcal{A}$, tales que $\inf_{\mathbf{x}\in\mathcal{A}} f(\mathbf{x}) > 0$. Sean $\widehat{f}(\mathbf{x}) = \widehat{f}_n(\mathbf{x})$ y $\widehat{a}(\mathbf{x}) = \widehat{a}_n$ tales que*

$$\begin{aligned}
\sup_{\mathbf{x}\in\mathcal{A}}|\widehat{f}(\mathbf{x}) - f(\mathbf{x})| &\xrightarrow{c.s.} 0 \\
\sup_{\mathbf{x}\in\mathcal{A}}\left|\frac{\widehat{a}(\mathbf{x})}{\widehat{f}(\mathbf{x})} - b(\mathbf{x})\right| &\xrightarrow{c.s.} 0
\end{aligned}$$

entonces $\sup_{\mathbf{x}\in\mathcal{C}}|\widehat{a}(\mathbf{x}) - b(\mathbf{x})f(\mathbf{x})| \xrightarrow{c.s.} 0$.

DEMOSTRACIÓN. Observemos que

$$\begin{aligned}
\sup_{\mathbf{x}\in\mathcal{A}}|\widehat{a}(\mathbf{x}) - b(\mathbf{x})f(\mathbf{x})| &= \sup_{\mathbf{x}\in\mathcal{A}}|\widehat{a}(\mathbf{x}) - b(\mathbf{x})\widehat{f}(\mathbf{x}) + b(\mathbf{x})\widehat{f}(\mathbf{x}) - b(\mathbf{x})f(\mathbf{x})| \\
&\leq \sup_{\mathbf{x}\in\mathcal{A}}|\widehat{a}(\mathbf{x}) - b(\mathbf{x})\widehat{f}(\mathbf{x})| + \sup_{\mathbf{x}\in\mathcal{A}}|b(\mathbf{x})|\sup_{\mathbf{x}\in\mathcal{C}}|\widehat{f}(\mathbf{x}) - f(\mathbf{x})| \\
&\leq \sup_{\mathbf{x}\in\mathcal{A}}\left|\frac{\widehat{a}(\mathbf{x})}{\widehat{f}(\mathbf{x})} - b(\mathbf{x})\right|\sup_{\mathbf{x}\in\mathcal{A}}|\widehat{f}(\mathbf{x})| + \sup_{\mathbf{x}\in\mathcal{A}}|b(\mathbf{x})|\sup_{\mathbf{x}\in\mathcal{A}}|\widehat{f}(\mathbf{x}) - f(\mathbf{x})| \\
&\leq \sup_{\mathbf{x}\in\mathcal{A}}\left|\frac{\widehat{a}(\mathbf{x})}{\widehat{f}(\mathbf{x})} - b(\mathbf{x})\right|\left[\sup_{\mathbf{x}\in\mathcal{A}}|f(\mathbf{x})| + \sup_{\mathbf{x}\in\mathcal{A}}|\widehat{f}(\mathbf{x}) - f(\mathbf{x})|\right] \\
&\quad + \sup_{\mathbf{x}\in\mathcal{A}}|b(\mathbf{x})|\sup_{\mathbf{x}\in\mathcal{A}}|\widehat{f}(\mathbf{x}) - f(\mathbf{x})|
\end{aligned}$$

Como $b(\mathbf{x})$ y $f(\mathbf{x})$ son funciones continuas en $\mathcal{A}$ compacto, $\sup_{\mathbf{x}\in\mathcal{A}}|b(\mathbf{x})| < \infty$ y $\sup_{\mathbf{x}\in\mathcal{A}}|f(\mathbf{x})| < \infty$ de donde se concluye el resultado. □

Teorema 5.3.2. *Bajo **D1** a **D6**, **K1**, **K2** y **H1**, tenemos que $\sup_{\mathbf{x}\in\mathcal{C}}|\widetilde{m}_{\mathrm{S}}^{(2)}(\mathbf{x}) - m(\mathbf{x})| \xrightarrow{c.s.} 0$.*

DEMOSTRACIÓN. Dado $\mathbf{x} \in \mathcal{C}$, denotemos por $\pi(\mathbf{x}) = p(\mathbf{x}) f_X(\mathbf{x})$ y definamos

$$\begin{aligned}
\widehat{\pi}(\mathbf{x}) &= \frac{1}{nh_n^d} \sum_{i=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right) \delta_i \\
\widehat{f}(\mathbf{x}) &= \frac{1}{nh_n^d} \sum_{j=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right) \\
\widehat{r}(\mathbf{x}) &= \frac{1}{\widehat{\pi}(\mathbf{x})} - \frac{1}{\pi(\mathbf{x})} \, .
\end{aligned}$$

Como $y_j = m(\mathbf{x}_j) + u_j$, tenemos que

$$\begin{aligned}
\widetilde{m}_{\mathrm{S}}^{(2)}(\mathbf{x}) &= \sum_{j=1}^{n} \frac{\mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right) \delta_j m(\mathbf{x}_j)}{\sum_{i=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}_i-\mathbf{x}_j}{h_n}\right) \delta_i} + \sum_{j=1}^{n} \frac{\mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right) \delta_j u_j}{\sum_{i=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}_i-\mathbf{x}_j}{h_n}\right) \delta_i} \\
&= \frac{1}{nh_n^d} \sum_{j=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right) \frac{\delta_j m(\mathbf{x}_j)}{\widehat{\pi}(\mathbf{x}_j)} + \frac{1}{nh_n^d} \sum_{j=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right) \frac{\delta_j u_j}{\widehat{\pi}(\mathbf{x}_j)} \\
&= \frac{1}{nh_n^d} \sum_{j=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right) \delta_j m(\mathbf{x}_j) \widehat{r}(\mathbf{x}_j) + \frac{1}{nh_n^d} \sum_{j=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right) \delta_j u_j \widehat{r}(\mathbf{x}_j) \\
&\quad + \frac{1}{nh_n^d} \sum_{j=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right) \frac{\delta_j u_j}{\pi(\mathbf{x}_j)} + \frac{1}{nh_n^d} \sum_{j=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right) \frac{\delta_j m(\mathbf{x}_j)}{\pi(\mathbf{x}_j)} \\
&= B_1(\mathbf{x}) + B_2(\mathbf{x}) + B_3(\mathbf{x}) + B_4(\mathbf{x})
\end{aligned}$$

Bastará mostrar que

a) $\sup_{\mathbf{x}\in\mathcal{C}} |B_1(\mathbf{x})| \xrightarrow{c.s.} 0$

b) $\sup_{\mathbf{x}\in\mathcal{C}} |B_2(\mathbf{x})| \xrightarrow{c.s.} 0$

c) $\sup_{\mathbf{x}\in\mathcal{C}} |B_3(\mathbf{x})| \xrightarrow{c.s.} 0$

d) $\sup_{\mathbf{x}\in\mathcal{C}} |B_4(\mathbf{x}) - m(\mathbf{x})| \xrightarrow{c.s.} 0$

Para obtener a) y b) empezaremos mostrando que

$$\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{r}(\mathbf{x})| \xrightarrow{c.s.} 0 \tag{5.3}$$

Tenemos que

$$\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{r}(\mathbf{x})| = \sup_{\mathbf{x}\in\mathcal{C}} \left| \frac{1}{\widehat{\pi}(\mathbf{x})} - \frac{1}{\pi(\mathbf{x})} \right| = \sup_{\mathbf{x}\in\mathcal{C}} \left| \frac{\pi(\mathbf{x}) - \widehat{\pi}(\mathbf{x})}{\widehat{\pi}(\mathbf{x})\pi(\mathbf{x})} \right| \leq \frac{\sup_{\mathbf{x}\in\mathcal{C}} |\pi(\mathbf{x}) - \widehat{\pi}(\mathbf{x})|}{i(\widehat{\pi})\; i(\pi)}.$$

Observemos que $\widehat{m}_\delta(\mathbf{x}) = \widehat{\pi}(\mathbf{x})/\widehat{f}(\mathbf{x})$ y por el Lema 5.3.1, $\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}_\delta(\mathbf{x}) - p(\mathbf{x})| \xrightarrow{c.s.} 0$. Por otra parte, bajo **D2**, **K1**, **K2** y **H1** tenemos que (ver Prakasa Rao, 1983)

$$\sup_{\mathbf{x}\in\mathcal{C}} \left| \widehat{f}(\mathbf{x}) - f_X(\mathbf{x}) \right| \xrightarrow{c.s.} 0 \, , \tag{5.4}$$

con lo cual por el Lema 5.3.2, como f y p son funciones continuas en $\mathcal{C}$, se obtiene que

$$\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{\pi}(\mathbf{x}) - \pi(\mathbf{x})| \xrightarrow{c.s.} 0. \tag{5.5}$$

Observemos que $i(p)$ $i(f_X) \leq i(\pi)$, de donde por **D2** y **D3**, $i(\pi) > 0$. Sea $\mathcal{N}_1$ el conjunto de probabilidad 0 tal que $\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{\pi}(\mathbf{x}) - \pi(\mathbf{x})| \not\to 0$. Luego, si $\omega \notin \mathcal{N}_1$ para $n \geq n_1$

$$\inf_{\mathbf{x}\in\mathcal{C}} |\widehat{\pi}(\mathbf{x})| \geq i(\pi) - \sup_{\mathbf{x}\in\mathcal{C}} |\widehat{\pi}(\mathbf{x}) - \pi(\mathbf{x})| \geq \frac{i(\pi)}{2} \ ,$$

de donde,

$$\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{r}(\mathbf{x})| \leq \frac{2}{i(\pi)} \sup_{\mathbf{x}\in\mathcal{C}} |\widehat{\pi}(\mathbf{x}) - \pi(\mathbf{x})|,$$

y por lo tanto, de (5.5), se deduce (5.3).

a) Veamos que $\sup_{\mathbf{x}\in\mathcal{C}} |B_1(\mathbf{x})| \xrightarrow{c.s.} 0$. Observemos que

$$\begin{aligned} |B_1(\mathbf{x})| &\leq \frac{1}{nh_n^d} \sum_{j=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right) |\delta_j m(\mathbf{x}_j)\widehat{r}(\mathbf{x}_j)| \\ &\leq \|m\|_{0,\infty} \sup_{\mathbf{x}\in\mathcal{C}} |\widehat{r}(\mathbf{x})| \frac{1}{nh_n^d} \sum_{j=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right) = \|m\|_{0,\infty} \sup_{\mathbf{x}\in\mathcal{C}} |\widehat{r}(\mathbf{x})| \ \widehat{f}(\mathbf{x}) \ . \end{aligned}$$

Luego, tenemos que

$$\begin{aligned} \sup_{\mathbf{x}\in\mathcal{C}} |B_1(\mathbf{x})| &\leq \|m\|_{0,\infty} \sup_{\mathbf{x}\in\mathcal{C}} |\widehat{r}(\mathbf{x})| \sup_{\mathbf{x}\in\mathcal{C}} \widehat{f}(\mathbf{x}) \\ &\leq \|m\|_{0,\infty} \sup_{\mathbf{x}\in\mathcal{C}} |\widehat{r}(\mathbf{x})| \left(\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{f}(\mathbf{x}) - f_X(\mathbf{x})| + \|f_X\|_{0,\infty} \right) \ , \end{aligned}$$

de donde usando (5.3) y (5.4) se obtiene a).

b) Veamos que $\sup_{\mathbf{x}\in\mathcal{C}} |B_2(\mathbf{x})| \xrightarrow{c.s.} 0$. Usando la desigualdad de Cauchy-Schwartz, obtenemos

$$\begin{aligned} |B_2(\mathbf{x})| &= \frac{1}{nh_n^d} \left| \sum_{j=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right) \delta_j u_j \widehat{r}(\mathbf{x}_j) \right| \\ &\leq \left[\frac{1}{nh_n^d} \sum_{i=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right) \delta_j^2 u_j^2 \right]^{\frac{1}{2}} \left[\frac{1}{nh_n^d} \sum_{j=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right) \widehat{r}(\mathbf{x}_j)^2 \right]^{\frac{1}{2}} \\ &\leq \left[\frac{1}{nh_n^d} \sum_{i=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right) \delta_j^2 u_j^2 \right]^{\frac{1}{2}} \widehat{f}(x)^{\frac{1}{2}} \sup_{\mathbf{x}\in\mathcal{C}} |\widehat{r}(\mathbf{x})| \end{aligned}$$

de donde

$$\sup_{\mathbf{x}\in\mathcal{C}} |B_2(\mathbf{x})| \leq \sup_{\mathbf{x}\in\mathcal{C}} \left[\frac{1}{nh_n^d} \sum_{i=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right) \delta_j^2 u_j^2 \right]^{\frac{1}{2}} \left[\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{f}(x) - f_X(\mathbf{x})| + \|f_X\|_{0,\infty} \right]^{\frac{1}{2}} \sup_{\mathbf{x}\in\mathcal{C}} |\widehat{r}(\mathbf{x})|$$

Como $(u_j^2)_{j=1}^n$ es una sucesión de variables aleatorias independientes, idénticamente distribuídas y uniformemente Gaussiana generalizada, la sucesión $(\delta_j u_j^2)_{j=1}^n$ también lo es. La demostración de esta propiedad es análoga a la dada en el Lema 5.3.1.Por otra parte, **D1** y **D5** implican que $E[\delta u^2|\mathbf{X}=\mathbf{x}] = p(\mathbf{x})\sigma^2(\mathbf{x})E[\epsilon^2] = p(\mathbf{x})\sigma^2(\mathbf{x})$. Luego,, usando que σ y p son funciones continuas en $\mathcal{C}$, por el Teorema 2.5.3, se obtiene

$$\sup_{\mathbf{x}\in\mathcal{C}}\left|\frac{\sum_{i=1}^n \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right)\delta_j u_j^2}{\sum_{j=1}^n \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right)} - p(\mathbf{x})\sigma^2(\mathbf{x})\right| \xrightarrow{c.s.} 0\ ,$$

de donde,

$$\sup_{\mathbf{x}\in\mathcal{C}}\left|\frac{1}{nh_n^d}\sum_{i=1}^n \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right)\delta_j u_j^2 - p(\mathbf{x})\sigma^2(\mathbf{x})f_X(\mathbf{x})\right| \xrightarrow{c.s.} 0. \tag{5.6}$$

Por lo tanto,

$$\begin{aligned}\sup_{\mathbf{x}\in\mathcal{C}}\left|\frac{1}{nh_n^d}\sum_{i=1}^n \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right)\delta_j u_j^2\right| \quad &\leq \quad \sup_{\mathbf{x}\in\mathcal{C}}\left|\frac{1}{nh_n^d}\sum_{i=1}^n \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right)\delta_j u_j^2 - p(\mathbf{x})\sigma^2(\mathbf{x})f_X(\mathbf{x})\right| \\ &\quad + \sup_{\mathbf{x}\in\mathcal{C}}\left|p(\mathbf{x})\sigma^2(\mathbf{x})f(\mathbf{x})\right|.\end{aligned}$$

Pero $\sup_{\mathbf{x}\in\mathcal{C}}\left|p(\mathbf{x})\sigma^2(\mathbf{x})f(\mathbf{x})\right| \leq \|\sigma^2\|_{0,\infty}\|f\|_{0,\infty} < \infty$ por **D2** y **D4**. Luego, tenemos que

$$\begin{aligned}\sup_{\mathbf{x}\in\mathcal{C}}|B_2(\mathbf{x})| \quad &\leq \quad \left\{\sup_{\mathbf{x}\in\mathcal{C}}\left|\frac{1}{nh_n^d}\sum_{i=1}^n \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right)\delta_j u_j^2 - p(\mathbf{x})\sigma^2(\mathbf{x})f_X(\mathbf{x})\right| + \|\sigma^2\|_{0,\infty}\|f\|_{0,\infty}\right\}^{\frac{1}{2}} \\ &\quad \times \left[\sup_{\mathbf{x}\in\mathcal{C}}|\widehat{f}(x) - f_X(\mathbf{x})| + \|f_X\|_{0,\infty}\right]^{\frac{1}{2}} \sup_{\mathbf{x}\in\mathcal{C}}|\widehat{r}(\mathbf{x})|\end{aligned}$$

de donde, por (5.3), (5.4) y (5.6), se obtiene b).

c) Queremos probar que

$$\sup_{\mathbf{x}\in\mathcal{C}}|B_3(\mathbf{x})| = \sup_{\mathbf{x}\in\mathcal{C}}\left|\frac{1}{nh_n^d}\sum_{j=1}^n \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right)\frac{\delta_j u_j}{\pi(\mathbf{x}_j)}\right| \xrightarrow{c.s.} 0$$

Esto se obtiene a partir del hecho que $\left(\frac{\delta_j u_j}{\pi(\mathbf{x}_j)}\right)_{j=1}^n$ es una sucesión de variables aleatorias independientes, idénticamente distribuídas y uniformemente Gaussiana generalizada. La demostración de esta propiedad es análoga a dada en el Lema 5.3.1, tomando $d = c\|\sigma\|_{0,\infty}^2\ i(\pi)^{-1} < \infty$. Como además

$$E\left(\frac{\delta u}{\pi(\mathbf{X})}|\mathbf{X}=\mathbf{x}\right) = \frac{p(\mathbf{x})\sigma(\mathbf{x})}{\pi(\mathbf{x})}E(\epsilon) = 0\ ,$$

por el Teorema 2.5.3 obtenemos que

$$\sup_{\mathbf{x}\in\mathcal{C}}\left|\frac{\sum_{j=1}^n \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right)\frac{\delta_j u_j}{\pi(\mathbf{x}_j)}}{\sum_{i=1}^n \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)}\right| \xrightarrow{c.s.} 0.$$

De donde usando (5.4), se deduce que $\sup_{\mathbf{x}\in\mathcal{C}}|B_3(\mathbf{x})| \xrightarrow{c.s.} 0$, lo que concluye la demostración de c).

d) Veamos ahora que $\sup_{\mathbf{x}\in\mathcal{C}}|B_4(\mathbf{x})-m(\mathbf{x})| \xrightarrow{c.s.} 0$. Como $\left(\frac{\delta_j m(\mathbf{x}_j)}{\pi(\mathbf{x}_j)}\right)_{j=1}^{n}$ es una sucesión de variables aleatorias independientes, idénticamente distribuídas y uniformemente acotada, por la Observación 2.5.1, es una sucesión uniformemente Gaussiana generalizada, y además

$$E\left[\frac{\delta m(\mathbf{X})}{\pi(\mathbf{X})}|\mathbf{X}=\mathbf{x}\right] = \frac{p(\mathbf{x})m(\mathbf{x})}{\pi(\mathbf{x})} = \frac{m(\mathbf{x})}{f(\mathbf{x})},$$

luego, por el Teorema 2.5.3, resulta que

$$\sup_{\mathbf{x}\in\mathcal{C}}\left|\frac{1}{\widehat{f}(\mathbf{x})}\frac{1}{nh_n^d}\sum_{j=1}^{n}\mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right)\frac{\delta_j m(\mathbf{x}_j)}{\pi(\mathbf{x}_j)} - \frac{m(\mathbf{x})}{f(\mathbf{x})}\right| \xrightarrow{c.s.} 0.$$

Con lo cual, usando nuevamente (5.4) obtenemos que

$$\sup_{\mathbf{x}\in\mathcal{C}}\left|\frac{1}{nh_n^d}\sum_{j=1}^{n}\mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right)\frac{\delta_j m(\mathbf{x}_j)}{\pi(\mathbf{x}_j)} - m(\mathbf{x})\right| \xrightarrow{c.s.} 0$$

lo que concluye la demostración de d) y por lo tanto, la del Teorema. □

5.4. Consistencia del estimador de regresión imputado

Teorema 5.4.1. *Supongamos que se cumplen **D1** a **D5**, **K1**, **K2**, **H1**. Sea $\widetilde{m}$ un estimador de la función de regresión que converge uniformemente casi seguramente a m en $\mathcal{C}$, es decir, tal que $\sup_{\mathbf{x}\in\mathcal{C}}|\widetilde{m}(\mathbf{x})-m(\mathbf{x})| \xrightarrow{c.s.} 0$. Sea $\widehat{y}_i = \delta_i y_i + (1-\delta_i)\widetilde{m}(\mathbf{x}_i)$, y definamos*

$$\widehat{m}(\mathbf{x}) = \frac{\sum_{i=1}^{n}\mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\widehat{y}_i}{\sum_{j=1}^{n}\mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right)}.$$

Entonces, $\sup_{\mathbf{x}\in\mathcal{C}}|\widehat{m}(\mathbf{x})-m(\mathbf{x})| \xrightarrow{c.s.} 0$

DEMOSTRACIÓN. Observemos que

$$\begin{aligned}\widehat{m}(\mathbf{x})-m(\mathbf{x}) &= \frac{\sum_{i=1}^{n}\mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\delta_i y_i}{\sum_{j=1}^{n}\mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right)} + \frac{\sum_{i=1}^{n}\mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)(1-\delta_i)\widetilde{m}(\mathbf{x}_i)}{\sum_{j=1}^{n}\mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right)} - m(\mathbf{x}) \\ &= B_1(\mathbf{x}) + B_2(\mathbf{x}) + B_3(\mathbf{x})\end{aligned}$$

donde

$$\begin{aligned}
B_1(\mathbf{x}) &= \frac{\sum_{i=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right) \delta_i y_i}{\sum_{j=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right)} - p(\mathbf{x})m(\mathbf{x}) = \widehat{m}_{\delta Y}(\mathbf{x}) - p(\mathbf{x})m(\mathbf{x}) \\
B_2(\mathbf{x}) &= \frac{\sum_{i=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right) (1-\delta_i) m(\mathbf{x}_i)}{\sum_{j=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right)} - (1-p(\mathbf{x}))m(\mathbf{x}) = \widehat{m}_{(1-\delta)m}(\mathbf{x}) - (1-p(\mathbf{x}))m(\mathbf{x}) \\
B_3(\mathbf{x}) &= \frac{\sum_{i=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right) (1-\delta_i) \left[\widetilde{m}(\mathbf{x}_i) - m(\mathbf{x}_i)\right]}{\sum_{j=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right)} .
\end{aligned}$$

Por el Lema 5.3.1, $\sup_{\mathbf{x}\in\mathcal{C}} |B_1(\mathbf{x})| \xrightarrow{c.s.} 0$. Por otra parte, definiendo $\Delta = 1-\delta$, obtenemos como en (5.1) que $\sup_{\mathbf{x}\in\mathcal{C}} |B_2(\mathbf{x})| \xrightarrow{c.s.} 0$. Finalmente, tenemos que

$$\begin{aligned}
|B_3(\mathbf{x})| &= \left| \frac{\sum_{i=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right) (1-\delta_i)[\widetilde{m}(\mathbf{x}_i) - m(\mathbf{x}_i)]}{\sum_{j=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right)} \right| \\
&\leq \frac{1}{\widehat{f}(\mathbf{x})} \frac{1}{nh_n^d} \sum_{i=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right) (1-\delta_i) \sup_{\mathbf{x}\in\mathcal{C}} |\widetilde{m}(\mathbf{x}) - m(\mathbf{x})| \leq \\
&\leq \sup_{\mathbf{x}\in\mathcal{C}} |\widetilde{m}(\mathbf{x}) - m(\mathbf{x})| ,
\end{aligned}$$

de donde, se obtiene que $\sup_{\mathbf{x}\in\mathcal{C}} |B_3(\mathbf{x})| \xrightarrow{c.s.} 0$, lo que concluye la demostración. □

Teorema 5.4.2. *Supongamos que se cumplen **D1** a **D6**, **K1**, **K2**, **H1**. Sea $\widetilde{m}$ un estimador de la función de regresión que converge uniformemente casi seguramente a m en $\mathcal{C}$, es decir, tal que $\sup_{\mathbf{x}\in\mathcal{C}} |\widetilde{m}(\mathbf{x}) - m(\mathbf{x})| \xrightarrow{c.s.} 0$. Sea $\widehat{y}_i = \delta_i y_i + (1-\delta_i)\widetilde{m}(\mathbf{x}_i)$, y definamos*

$$\widehat{m}(\mathbf{x}) = \sum_{i=1}^{n} \frac{\mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right) \widehat{y}_i}{\sum_{j=1}^{n} \mathcal{K}\left(\frac{\mathbf{x}_i-\mathbf{x}_j}{h_n}\right)}.$$

Entonces, $\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}(\mathbf{x}) - m(\mathbf{x})| \xrightarrow{c.s.} 0$.

DEMOSTRACIÓN. Usando que $y_i = m(\mathbf{x}_i) + u_i$ obtenemos que

$$\begin{aligned}
\widehat{y}_i &= \delta_i \left(m(\mathbf{x}_i) + u_i\right) + (1-\delta_i)\widetilde{m}(\mathbf{x}_i) \\
&= \delta_i u_i + m(\mathbf{x}_i) + (1-\delta_i) \left(\widetilde{m}(\mathbf{x}_i) - m(\mathbf{x}_i)\right) .
\end{aligned}$$

Por lo tanto, podemos reescribir $\widehat{m}$ como

$$\begin{aligned}
\widehat{m}(\mathbf{x}) &= \frac{1}{nh_n^d}\sum_{i=1}^n \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\frac{m(\mathbf{x}_i)}{\widehat{f}(\mathbf{x}_i)} + \frac{1}{nh_n^d}\sum_{i=1}^n \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\frac{\delta_i u_i}{\widehat{f}(\mathbf{x}_i)} \\
&\quad + \frac{1}{nh_n^d}\sum_{i=1}^n \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)(1-\delta_i)\frac{(\widetilde{m}(\mathbf{x}_i)-m(\mathbf{x}_i))}{\widehat{f}(\mathbf{x}_i)} \\
&= \frac{1}{nh_n^d}\sum_{i=1}^n \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\frac{m(\mathbf{x}_i)}{f_X(\mathbf{x}_i)} + \frac{1}{nh_n^d}\sum_{i=1}^n \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\frac{\delta_i u_i}{f_X(\mathbf{x}_i)} \\
&\quad + \frac{1}{nh_n^d}\sum_{i=1}^n \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\delta_i u_i\left[\frac{1}{\widehat{f}(\mathbf{x}_i)} - \frac{1}{f_X(\mathbf{x}_i)}\right] \\
&\quad + \frac{1}{nh_n^d}\sum_{i=1}^n \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\frac{(1-\delta_i)\,(\widetilde{m}(\mathbf{x}_i)-m(\mathbf{x}_i))}{f_X(\mathbf{x}_i)} \\
&\quad + \frac{1}{nh_n^d}\sum_{i=1}^n \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)m(\mathbf{x}_i)\left[\frac{1}{\widehat{f}(\mathbf{x}_i)} - \frac{1}{f_X(\mathbf{x}_i)}\right] \\
&\quad + \frac{1}{nh_n^d}\sum_{i=1}^n \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)(1-\delta_i)\,(\widetilde{m}(\mathbf{x}_i)-m(\mathbf{x}_i))\left[\frac{1}{\widehat{f}(\mathbf{x}_i)} - \frac{1}{f_X(\mathbf{x}_i)}\right] \\
&= \sum_{i=1}^6 B_i(\mathbf{x})\ .
\end{aligned}$$

Queremos ver que

a) $\sup_{\mathbf{x}\in\mathcal{C}}|B_1(\mathbf{x})-m(\mathbf{x})| \xrightarrow{c.s.} 0$

b) $\sup_{\mathbf{x}\in\mathcal{C}}|B_i(\mathbf{x})| \xrightarrow{c.s.} 0$ para $i \geq 2$.

a) **D2** y **D4** implican que $\left(\frac{m(\mathbf{x}_i)}{f_X(\mathbf{x}_i)}\right)_{i=1}^n$ es una sucesión de variables aleatorias independientes, idénticamente distribuídas y uniformemente acotadas, es uniformemente Gaussiana generalizada. Luego, por el Teorema 2.5.3 y como

$$E\left[\frac{m(\mathbf{X})}{f_X(\mathbf{X})}|\mathbf{X}=\mathbf{x}\right] = \frac{m(\mathbf{x})}{f_X(\mathbf{x})}\ ,$$

resulta que

$$\sup_{\mathbf{x}\in\mathcal{C}}\left|\frac{1}{\widehat{f}(\mathbf{x})}\frac{1}{nh_n^d}\sum_{i=1}^n \mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\frac{m(\mathbf{x}_i)}{f_X(\mathbf{x}_i)} - \frac{m(\mathbf{x})}{f_X(\mathbf{x})}\right| \xrightarrow{c.s.} 0\ ,$$

de donde usando (5.4) obtenemos a).

b) Para probar que $\sup_{\mathbf{x}\in\mathcal{C}}|B_2(\mathbf{x})| \xrightarrow{c.s.} 0$, observemos que $z_i = u_i/f_X(\mathbf{x}_i) = m^\star(\mathbf{x}_i) + \sigma^\star(\mathbf{x}_i)\epsilon_i$ donde $m^\star \equiv 0$ y $\sigma^\star(\mathbf{x}) = \sigma(\mathbf{x})/f_X(\mathbf{x})$ cumplen **D4**, luego por el Lema 5.3.1

$$\sup_{\mathbf{x}\in\mathcal{C}}|\widehat{m}_{\delta Z}(\mathbf{x})| \xrightarrow{c.s.} 0\ ,$$

con lo cual como $B_2(\mathbf{x}) = \widehat{m}_{\delta Z}(\mathbf{x})\,\widehat{f}(\mathbf{x})$ y usando (5.4) se obtiene el resultado.

Para probar que $\sup_{\mathbf{x}\in\mathcal{C}} |B_i(\mathbf{x})| \xrightarrow{c.s.} 0$, con $i > 4$, usamos las siguientes acotaciones

$$\begin{aligned}
|B_4(\mathbf{x})| &\leq \frac{1}{nh_n^d}\sum_{i=1}^{n}\mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\frac{\sup_{\mathbf{x}\in\mathcal{C}}|\widetilde{m}(\mathbf{x}_i)-m(\mathbf{x}_i)|}{i(f_X)} \leq \sup_{\mathbf{x}\in\mathcal{C}}|\widehat{f}(\mathbf{x})|\frac{\sup_{\mathbf{x}\in\mathcal{C}}|\widetilde{m}(\mathbf{x}_i)-m(\mathbf{x}_i)|}{i(f_X)}\\
|B_5(\mathbf{x})| &\leq \|m\|_{0,\infty}\sup_{\mathbf{x}\in\mathcal{C}}\left|\frac{1}{\widehat{f}(\mathbf{x}_i)}-\frac{1}{f_X(\mathbf{x}_i)}\right|\sup_{\mathbf{x}\in\mathcal{C}}|\widehat{f}(\mathbf{x})|\\
|B_6(\mathbf{x})| &\leq \sup_{\mathbf{x}\in\mathcal{C}}\left|\frac{1}{\widehat{f}(\mathbf{x}_i)}-\frac{1}{f_X(\mathbf{x}_i)}\right|\sup_{\mathbf{x}\in\mathcal{C}}|\widetilde{m}(\mathbf{x}_i)-m(\mathbf{x}_i)|\sup_{\mathbf{x}\in\mathcal{C}}|\widehat{f}(\mathbf{x})|
\end{aligned}$$

que permiten deducir el resultado de (5.4), **D2** y del hecho que $\sup_{\mathbf{x}\in\mathcal{C}}|\widetilde{m}(\mathbf{x}_i)-m(\mathbf{x}_i)| \xrightarrow{c.s.} 0$.

Falta probar que $\sup_{\mathbf{x}\in\mathcal{C}}|B_3(\mathbf{x})| \xrightarrow{c.s.} 0$. La demostración sigue argumentos análogos a los utilizados para probar b) en la demostración del Teorema 5.3.2 y por esta razón se omite, concluyendo la demostración. □

5.5. Consistencia de los estimadores $\widehat{c}_1$ y $\widehat{c}_2$

Teorema 5.5.1. *Sea $\widetilde{m}$ un estimador de la función de regresión que converge uniformemente casi seguramente a m en $\mathcal{C}$, es decir, tal que $\sup_{\mathbf{x}\in\mathcal{C}}|\widetilde{m}(\mathbf{x})-m(\mathbf{x})| \xrightarrow{c.s.} 0$. Supongamos que se cumplen **D1**, **A1** y **A2**. Si*

$$\widehat{c} = \frac{1}{n}\sum_{i=1}^{n}\widetilde{m}(\mathbf{x}_i)$$

entonces $\widehat{c} \xrightarrow{c.s.} c$.

De aquí deducimos el siguiente Corolario.

Corolario 5.5.2. Supongamos que se verifican **A1**, **A2**, **A3**, **D1** a **D5**, **K1**, **K2** y **H1**, entonces $\widehat{c}_1 = \frac{1}{n}\sum_{i=1}^{n}\widetilde{m}_{\mathrm{s}}^{(1)}(\mathbf{x}_i) \xrightarrow{c.s.} c$.

DEMOSTRACIÓN DEL TEOREMA 5.5.1.Bajo **A1** $m(\mathbf{x}) = c + \sum_{\alpha=1}^{d} g_\alpha(x_\alpha)$. Luego,

$$\begin{aligned}
|\widehat{c}-c| &= \left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{m}(\mathbf{x}_i)-c\right| = \left|\frac{1}{n}\sum_{i=1}^{n}\left\{\widetilde{m}(\mathbf{x}_i)-\left(c+\sum_{\alpha=1}^{d}g_\alpha(x_{\alpha i})\right)\right\}+\frac{1}{n}\sum_{i=1}^{n}\sum_{\alpha=1}^{d}g_\alpha(x_{\alpha i})\right|\\
&= \left|\frac{1}{n}\sum_{i=1}^{n}\{\widetilde{m}(\mathbf{x}_i)-m(\mathbf{x}_i)\}+\sum_{\alpha=1}^{d}\left\{\frac{1}{n}\sum_{i=1}^{n}g_\alpha(x_{\alpha i})\right\}\right|
\end{aligned}$$

$$\begin{aligned} &\leq \frac{1}{n}\sum_{i=1}^{n}|\widetilde{m}(\mathbf{x}_i)-m(\mathbf{x}_i)|+\sum_{\alpha=1}^{d}\left|\frac{1}{n}\sum_{i=1}^{n}g_\alpha(x_{\alpha i})\right| \\ &\leq \sup_{\mathbf{x}\in\mathcal{C}}|\widetilde{m}(\mathbf{x})-m(\mathbf{x})|+\sum_{\alpha=1}^{d}\left|\frac{1}{n}\sum_{i=1}^{n}g_\alpha(x_{\alpha i})\right| \end{aligned}$$

Como por hipótesis $\sup_{\mathbf{x}\in\mathcal{C}}|\widetilde{m}(\mathbf{x})-m(\mathbf{x})| \xrightarrow{c.s.} 0$, bastará probar que

$$\sum_{\alpha=1}^{d}\left|\frac{1}{n}\sum_{i=1}^{n}g_\alpha(x_{\alpha i})\right| \xrightarrow{c.s.} 0 \tag{5.7}$$

Como existe $E[|g_\alpha(X_\alpha)|]<\infty$ y por **A2**, $Eg_\alpha(X_\alpha)=0$, por la Ley Fuerte de los Grandes Números se cumple que para todo $1\leq\alpha\leq d$,

$$\frac{1}{n}\sum_{i=1}^{n}g_\alpha(x_{\alpha i}) \xrightarrow{c.s.} 0.$$

de donde se obtiene (5.7), lo que concluye la demostración del Teorema. □

Teorema 5.5.3. *Si se cumplen **D1** a **D4**, **A1** y **A2** se tiene que*

$$\widehat{c}_2=\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i y_i}{p(\mathbf{x}_i)} \xrightarrow{c.s.} c$$

DEMOSTRACIÓN. Sea $z_i=\delta_i y_i/p(\mathbf{x}_i)$, la sucesión $\{z_i\}_{i=1}^n$ es una sucesión de variables aleatorias independientes e idénticamente distribuídas tales que

$$|z_1|\leq\frac{|y_1|}{i(p)}\leq\frac{|m(\mathbf{x}_1)|+\sigma(\mathbf{x}_1)|\epsilon_1|}{i(p)}\,,$$

luego, como m y σ son continuas en $\mathcal{C}$ por **D1**, $E(|z_i|)<\infty$. Por otra parte, **D3** y **A2** implican que

$$E\left[\frac{\delta Y}{p(\mathbf{X})}|\mathbf{X}=\mathbf{x}\right]=\frac{p(\mathbf{x})}{p(\mathbf{x})}E[Y|\mathbf{X}=\mathbf{x}]=c\,.$$

Luego, por la Ley Fuerte de los Grandes Números, se obtiene el resultado. □

5.6. Consistencia de componentes aditivas

Recordemos que por simplicidad, la notación $m(x_\alpha,\mathbf{x}_{\underline{\alpha}i})$ indica al valor de la función m calculado en el vector $\mathbf{x}$ cuya coordenada α es x_α y las demás coordenadas coinciden con las de $\mathbf{x}_i$.

Teorema 5.6.1. *Supongamos que se cumplen **D2**, **A1** a **A3**. Sea $\widehat{c}$ un estimador consistente de c y $\widetilde{m}(\mathbf{x})$ un estimador de la función de regresión tal que $\sup_{\mathbf{x}\in\mathcal{C}}|\widetilde{m}(\mathbf{x})-m(\mathbf{x})| \xrightarrow{c.s.} 0$ y definamos*

$$\widehat{g}_\alpha(x_\alpha)=\frac{1}{n}\sum_{i=1}^{n}\widetilde{m}(x_\alpha,\mathbf{x}_{\underline{\alpha}i})-\widehat{c},$$

entonces

a) $\sup_{x\in\mathcal{C}_\alpha} |\widehat{g}_\alpha(x_\alpha) - g_\alpha(x_\alpha)| \xrightarrow{c.s.} 0$

b) $\widehat{m}(\mathbf{x}) = \sum_{\alpha=1}^{d} \widehat{g}_\alpha(x_\alpha) + \widehat{c}$ *converge uniformemente casi seguramente a* $m(\mathbf{x})$ *en* $\mathcal{C}$*, o sea,* $\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}(\mathbf{x}) - m(\mathbf{x})| \xrightarrow{c.s.} 0$.

DEMOSTRACIÓN. Veamos primero a). Sea $1 \le \alpha \le d$. Tenemos que

$$\begin{aligned}\sup_{x_\alpha\in C_\alpha} |\widehat{g}_\alpha(x_\alpha) - g_\alpha(x_\alpha)| \;\le\; & \sup_{x_\alpha\in C_\alpha} \left| \frac{1}{n}\sum_{i=1}^{n} \widetilde{m}(x_\alpha, \mathbf{x}_{\underline{\alpha}i}) - m(x_\alpha, \mathbf{x}_{\underline{\alpha}i}) \right| + |\widehat{c} - c| \\ & + \sup_{x_\alpha\in C_\alpha} \left| \frac{1}{n}\sum_{i=1}^{n} m(x_\alpha, \mathbf{x}_{\underline{\alpha}i}) - c - g_\alpha(x_\alpha) \right| = B_1 + B_2 + B_3.\end{aligned}$$

La convergencia unifome de $\widetilde{m}$ y el hecho que $B_1 \le \sup_{\mathbf{x}\in\mathcal{C}} |\widetilde{m}(\mathbf{x}) - m(\mathbf{x})|$, muestran que $B_1 \xrightarrow{c.s.} 0$. Por otra parte, $B_2 \xrightarrow{c.s.} 0$ pues $\widehat{c}$ es un estimador consistente de c. Luego, para ver a) bastará probar que $B_3 \xrightarrow{c.s.} 0$. Como m satisface el modelo aditivo dado en **A1**

$$\begin{aligned} B_3 &= \sup_{x_\alpha\in C_\alpha} \left| \frac{1}{n}\sum_{i=1}^{n} \left\{ \sum_{\tau=1,\tau\neq\alpha}^{d} g_\tau(x_{\tau i}) + g_\alpha(x_\alpha) - g_\alpha(x_\alpha) \right\} \right| = \sup_{x_\alpha\in C_\alpha} \left| \frac{1}{n}\sum_{i=1}^{n} \sum_{\tau=1,\tau\neq\alpha}^{d} g_\tau(x_{\tau i}) \right| \\ &= \left| \sum_{\tau=1,\tau\neq\alpha}^{d} \frac{1}{n}\sum_{i=1}^{n} g_\tau(x_{\tau i}) \right| . \end{aligned}$$

Como $E|g_\tau(X_\tau)| < \infty$ y se cumple **A2**, la Ley Fuerte de los Grandes Números implica que, para $1 \le \tau \le d$,

$$\frac{1}{n}\sum_{i=1}^{n} g_\tau(x_{\tau i}) \xrightarrow{c.s.} 0 \,,$$

de donde, se obtiene que

$$\sum_{\tau=1,\tau\neq\alpha}^{d} \frac{1}{n}\sum_{i=1}^{n} g_\tau(X_{\tau i}) \xrightarrow{c.s.} 0 \,,$$

concluyendo la demostración de a).

b) Tenemos que

$$\begin{aligned} \sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}(\mathbf{x}) - m(\mathbf{x})| &= \sup_{\mathbf{x}\in\mathcal{C}} \left| \sum_{\alpha=1}^{d} \widehat{g}_\alpha(x_\alpha) - g_\alpha(x_\alpha) + \widehat{c} - c \right| \\ &\le |\widehat{c} - c| + \sum_{\alpha=1}^{d} \sup_{x_\alpha\in C_\alpha} |\widehat{g}_\alpha(x_\alpha) - g_\alpha(x_\alpha)| \end{aligned}$$

El resultado se deduce ahora de a) y de la consistencia de $\widehat{c}$. □

Los Teoremas 5.3.1, 5.3.2 y 5.6.1 permiten obtener la consistencia de las propuestas dadas en la sección 4.3.1.

Corolario 5.6.2. *Supongamos que se cumplen **D1** a **D5** y **A1** a **A3**. Sea $\mathcal{K}$ un núcleo multivariado que cumple **K1** y **K2** y sea $\mathbf{H} = h_n I_d$ donde h_n cumplen **H1**. Entonces, los estimadores $\widehat{g}^{(1)}_{\alpha,\mathrm{S}}$, $\widehat{g}^{(2)}_{\alpha,\mathrm{S}}$, $\widehat{m}^{(1)}_{\mathrm{S}}$ y $\widehat{m}^{(2)}_{\mathrm{S}}$ definidos en la sección 4.3.1 convergen uniformemente casi seguramente a g_α y m, respectivamente, para $1 \leq \alpha \leq d$, es decir, para $\ell = 1, 2$ y $1 \leq \alpha \leq d$, $sup_{x \in \mathcal{C}_\alpha} |\widehat{g}^{(\ell)}_{\alpha,\mathrm{S}}(x) - g_\alpha(x)| \xrightarrow{c.s.} 0$ y $sup_{\mathbf{x} \in \mathcal{C}} |\widehat{m}^{(\ell)}_{\mathrm{S}}(\mathbf{x}) - m(\mathbf{x})| \xrightarrow{c.s.} 0$.*

Por otra parte, del Teorema 5.4.1 y del Corolario 5.6.2 se obtienen los siguientes resultados que muestra la consistencia de los estimadores propuestos en la sección 4.3.2 cuando en el segundo paso se utiliza el estimador de Nadaraya–Watson.

Corolario 5.6.3. *Supongamos que se cumplen **D1** a **D5**, **A1** a **A3**. Sean $\mathcal{K}$ y $\mathcal{L}$ núcleos multivariados que cumplen **K1** y **K2** y sean $\mathbf{H} = h_n I_d$ y $\mathbf{\Gamma} = \gamma_n I_d$ donde h_n y γ_n cumplen **H1**. Entonces, los estimadores $\widetilde{m}^{(1,\ell)}_{\mathrm{I}}$, para $1 \leq \alpha \leq d$ y $\ell = 0, 1, 2$, definidos en la sección 4.3.2, convergen uniformemente casi seguramente a m en $\mathcal{C}$, es decir, $sup_{\mathbf{x} \in \mathcal{C}} |\widetilde{m}^{(1,\ell)}_{\mathrm{I}}(\mathbf{x}) - m(\mathbf{x})| \xrightarrow{c.s.} 0$.*

Corolario 5.6.4. *Supongamos que se cumplen **D1** a **D6**, **A1** a **A3**. Sean $\mathcal{K}$ y $\mathcal{L}$ núcleos multivariados que cumplen **K1** y **K2** y sean $\mathbf{H} = h_n I_d$ y $\mathbf{\Gamma} = \gamma_n I_d$ donde h_n y γ_n cumplen **H1**. Entonces, los estimadores $\widehat{g}^{(1,\ell)}_{\alpha,\mathrm{I}}$ y $\widehat{m}^{(1,\ell)}_{\mathrm{I}}$ para $1 \leq \alpha \leq d$, $\ell = 0, 1, 2$, definidos en la sección 4.3.2, convergen uniformemente casi seguramente a g_α y m respectivamente, es decir, $sup_{x \in \mathcal{C}_\alpha} |\widehat{g}^{(1,\ell)}_{\alpha,\mathrm{I}}(x) - g_\alpha(x)| \xrightarrow{c.s.} 0$ y $sup_{\mathbf{x} \in \mathcal{C}} |\widehat{m}^{(1,\ell)}_{\mathrm{I}}(\mathbf{x}) - m(\mathbf{x})| \xrightarrow{c.s.} 0$.*

Por otra parte, del Teorema 5.4.2 y del Corolario 5.6.2 se obtienen los siguientes resultados que muestran la consistencia de los estimadores propuestos en la sección 4.3.2 cuando en el segundo paso se utiliza el estimador corregido internamente.

Corolario 5.6.5. *Supongamos que se cumplen **D1** a **D5**, **A1** a **A3**. Sean $\mathcal{K}$ y $\mathcal{L}$ núcleos multivariados que cumplen **K1** y **K2** y sean $\mathbf{H} = h_n I_d$ y $\mathbf{\Gamma} = \gamma_n I_d$ donde h_n y γ_n cumplen **H1**. Entonces, los estimadores $\widetilde{m}^{(1,\ell)}_{\mathrm{I}}$, para $1 \leq \alpha \leq d$ y $\ell = 0, 1, 2$, definidos en la sección 4.3.2, convergen uniformemente casi seguramente a m en $\mathcal{C}$, es decir, $sup_{\mathbf{x} \in \mathcal{C}} |\widetilde{m}^{(2,\ell)}_{\mathrm{I}}(\mathbf{x}) - m(\mathbf{x})| \xrightarrow{c.s.} 0$.*

Corolario 5.6.6. *Supongamos que se cumplen **D1** a **D6**, **A1** a **A3**. Sean $\mathcal{K}$ y $\mathcal{L}$ núcleos multivariados que cumplen **K1** y **K2** y sean $\mathbf{H} = h_n I_d$ y $\mathbf{\Gamma} = \gamma_n I_d$ donde h_n y γ_n cumplen **H1**. Entonces, los estimadores $\widehat{g}^{(2,\ell)}_{\alpha,\mathrm{I}}$ y $\widehat{m}^{(2,\ell)}_{\mathrm{I}}$ para $1 \leq \alpha \leq d$, $\ell = 0, 1, 2$, definidos en la sección 4.3.2, convergen uniformemente casi seguramente a g_α y m respectivamente, es decir, $sup_{x \in \mathcal{C}_\alpha} |\widehat{g}^{(2,\ell)}_{\alpha,\mathrm{I}}(x) - g_\alpha(x)| \xrightarrow{c.s.} 0$ y $sup_{\mathbf{x} \in \mathcal{C}} |\widehat{m}^{(2,\ell)}_{\mathrm{I}}(\mathbf{x}) - m(\mathbf{x})| \xrightarrow{c.s.} 0$.*

Capítulo 6

Estudio de Monte Carlo

6.1. Condiciones de la simulación

En este Capítulo, se describen los resultados de un estudio de simulación cuyo objetivo es comparar el comportamiento de los estimadores $\widetilde{m}_{\text{S}}^{(1)}$, $\widehat{m}_{\text{S}}^{(1)}$, $\widehat{m}_{\text{S}}^{(2)}$, $\widehat{m}_{\text{I}}^{(1,0)}$, $\widehat{m}_{\text{I}}^{(1,1)}$, $\widehat{m}_{\text{I}}^{(2,2)}$, $\widehat{g}_{\alpha,\text{S}}^{(1)}$, $\widehat{g}_{\alpha,\text{S}}^{(2)}$, $\widehat{g}_{\alpha,\text{I}}^{(1,0)}$, $\widehat{g}_{\alpha,\text{I}}^{(1,1)}$ y $\widehat{g}_{\alpha,\text{I}}^{(2,2)}$, $1 \leq \alpha \leq d$, definidos en las secciones 4.3.1 y 4.3.2 a fin de determinar si el hecho de imputar mejora la estimación. Se realizaron 500 replicaciones en las que se generaron muestras aleatorias independientes $\{(\mathbf{x}_i^{\text{T}}, y_i, \delta_i)\}_{i=1}^n$ de tamaño $n = 500$. Para ello, primero generamos observaciones $(\mathbf{x}_i^{\text{T}}, z_i)$ tales que

$$z_i = m(\mathbf{x}_i) + u_i, \qquad 1 \leq i \leq n$$

donde $\mathbf{x}_i = (x_{i1}, x_{i2}) \sim U([0,1] \times [0,1])$, $u \sim N(0, \sigma^2)$ con $\sigma = 0.5$, $m : \mathbb{R}^2 \to \mathbb{R}$ es una función aditiva de la forma

$$m(x_1, x_2) = 4 + 24\left(x_1 - \frac{1}{2}\right)^2 + 2\pi \sin(\pi x_2)\,. \tag{6.1}$$

Los modelos para generar las respuestas faltantes pueden describirse como siguen. Definimos $y_i = z_i$ si $\delta_i = 1$ y faltante en otro caso, donde $\{\delta_i\}_{i=1}^n$ se generaron según un mecanismo de pérdida de observaciones ignorable, MAR, es decir, $P(\delta_i = 1 | y_i, \mathbf{x}_i) = P(\delta_i = 1 | \mathbf{x}_i) = p(\mathbf{x}_i)$ con p alguna de las siguientes funciones

- $p_1(\mathbf{x}) \equiv 1$ que corresponde a la situación de muestras completas.
- $p_2(\mathbf{x}) \equiv 0.8$ que corresponde a un mecanismo de pérdida de datos completamente al azar.
- $p_3(\mathbf{x}) = 0.4 + 0.5(\cos(2x_1x_2 + 0.4))^2$

La probabilidad $p_1(\mathbf{x})$ permite comparar los estimadores propuestos, simplificados e imputados, con el estimador de regresión que podría ser evaluado si contáramos con el conjunto total de datos. Observemos que este estimador, al que nos referiremos como *estimador de datos completos*, no puede calcularse en la práctica. El objetivo es detectar qué estimador, simplificado o imputado, da errores cuadráticos más cercanos a los obtenidos si no hubiese respuestas faltantes.

Además, x_{i1}, x_{i2}, δ_i y u_i se generaron de forma a ser independientes entre sí.

Para identificar las funciones componentes marginales debemos tener en cuenta que, por **A2**, sus esperanzas son 0. Luego, para el modelo (6.1), tenemos que $c = 10$ y las componentes aditivas son

$$g_1(x_1) = 24\left(x_1 - \frac{1}{2}\right)^2 - 2 \qquad g_2(x_2) = 2\pi \sin(\pi x_2) - 4\ .$$

Respecto del procedimiento de suavizado, hemos utilizado el núcleo de Epanechnikov producto, tanto en los pasos previos (estimador simplificado) como posteriores a la imputación de los datos faltantes. Recordemos que esta función núcleo es de la forma $\mathcal{K}(\mathbf{x}) = K(x_1)K(x_2)$ donde $K(u) = \frac{3}{4}(1-u^2)\mathrm{I}_{[-1,1]}(u)$.

El comportamiento de un estimador $\widehat{m}$ de m se midió utilizando una aproximación del error cuadrático integrado definido en la sección 2.4 y calculado para cada una de las replicaciones como

$$\text{ISE}(\widehat{m}) = \frac{1}{\ell^2}\sum_{s=1}^{\ell}\sum_{j=1}^{\ell}\left(m\left(\mathbf{u}_{js}\right) - \widehat{m}\left(\mathbf{u}_{js}\right)\right)^2\ ,$$

donde $\mathbf{u}_{js} = (j/\ell, s/\ell)$, $1 \le j, s \le \ell$, $\ell = 50$. Una aproximación del MISE se obtuvo promediando sobre las 500 replicaciones el ISE.

Por otra parte, para evitar la influencia en el ISE del efecto de frontera que se observa en el estimador de Nadaraya–Watson, se definió una medida ponderada que corresponde a introducir una función de pesos $\mathcal{W}(\mathbf{x}) = W(x_1)W(x_2)$ con $W(t) = I_{(\tau,1-\tau)}(t)$ en el ISE, donde τ indica la ventana utilizada en el cálculo del estimador. Definimos entonces

$$\text{WISE}(\widehat{m}) = \frac{1}{\ell^2}\sum_{s=1}^{\ell}\sum_{j=1}^{\ell}\left(m\left(\mathbf{u}_{js}\right) - \widehat{m}\left(\mathbf{u}_{js}\right)\right)^2 \mathcal{W}(\mathbf{u}_{js})\ ,$$

donde $\mathbf{u}_{js} = (j/\ell, s/\ell)$, $1 \le j, s \le \ell$, $\ell = 50$. El valor WMISE indica el promedio del WISE sobre las replicaciones.

Medidas similares se utilizaron para comparar los estimadores de las componentes aditivas g_α.

Para comparar los estimadores, se utilizaron, en este estudio preliminar, distintas ventanas ya que un proceso de búsqueda utilizando convalidación cruzada era costoso computacionalmente. Las ventanas que hemos utilizado son

- para el estimador simplificado, definido en (4.9) o (4.10), h =0.15, 0.2, 0.25 y 0.3,
- para el estimador imputado, definido en (4.15) o (4.16), γ =0.1, 0.15, 0.2, 0.25 y 0.3. Esta selección de ventanas se hizo sobre la base de los resultados obtenidos por Chu y Chen (1993) y por González–Manteiga y Pérez–Gonzalez (2004) que muestran que, en el caso de covariables unidimensionales, el estimador imputado debe utilizar una ventana menor o igual que la simplificado utilizado para imputar las respuestas faltantes.

6.2. Elección del estimador de la media de Y

Recordemos que por la condición **A2**, $Eg_\alpha(X_\alpha) = 0$ y por lo tanto, $E(Y) = c = 10$ bajo el modelo aditivo (6.1). En la sección 4.3, hemos introducido dos estimadores de c definidos como

$$\begin{aligned} \widehat{c}^{(1)} &= \frac{1}{n}\sum_{i=1}^{n} \widetilde{m}_{\mathrm{S}}^{(1)}(\mathbf{x}_i) \\ \widehat{c}^{(2)} &= \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i y_i}{p(\mathbf{x}_i)} \,. \end{aligned}$$

Para decidir qué estimador de la constante c elegir, hemos realizado un estudio de simulación preliminar basado 1000 replicaciones y muestras de tamaño 500 para comparar estos dos estimadores así como los estimadores de las componentes aditivas, $\widehat{g}_{\alpha,\mathrm{S}}^{(\ell)}$, $\ell = 1, 2$, $\alpha = 1, 2$, asociados a cada uno de ellos. Para comparar los estimadores de c se utilizaron medias, desvíos y errores cuaráticos medios calculados sobre las replicaciones mientras que para los estimadores de g_α se calculó el MISE del estimador resultante. La ventana utilizada para el estimador $\widehat{c}_1$ fue $h_c = 0.2$, que corresponde a una amplitud cercana a la óptima elegida por *convalidación cruzada*.

A modo de ejemplo, mostraremos los resultados utilizados cuando el mecanismo de pérdida de observaciones está dado por $p_3(\mathbf{x})$. El Cuadro 6.1(a) presenta las medidas resumen de los estimadores de c mientras que el MISE de los estimadores de g_α asociados a cada estimador de c se pueden ver en el Cuadro 6.1(b). Notemos que el desvío de $\widehat{c}_2$ es 3 veces mayor que el de $\widehat{c}_1$ y esto se debe a que el mecanismo de pérdida elegido, induce una pérdida promedio de observaciones cercana al 40 %. Vale la pena mencionar, que cuando no hay pérdida de observaciones (o sea, cuando $p \equiv 1$), el estimador de $\widehat{c}_1$ es igual a $\bar{y}$. En este caso, error cuadrático medio de $\widehat{c}_1$ es igual a 0.0143062, o sea, la pérdida de observaciones sólo produce un aumento del 27 % en el error cuadrático del estimador. Por otra parte, $\widehat{c}_2$ también da valores de MISE mayores que $\widehat{c}_1$ tanto para los estimadores de las componentes aditivas basados en el estimador de Nadaraya–Watson como para el internamente corregido. Esto nos indica que resulta conveniente trabajar con el estimador $\widehat{c}_1$. Por este motivo, todas las estimaciones, antes y después de imputar, utilizarán este estimador de c.

6.3. Resultados

6.3.1. Comportamiento de los estimadores simplificados

El Cuadro 6.2 presenta los MISE para los estimadores simplificados bajo los distintos esquemas de pérdida de datos para las distintas ventanas consideradas en el proceso de suavizado. Por otra parte, en el Cuadro 6.3 se dan los WMISE de los estimadores de las funciones marginales cuando utilizamos el estimador simplificado de Nadaraya–Watson o internamente corregido.

Observemos que bajo el esquema de pérdida de datos p_1, no hay datos faltantes. Por lo tanto, los predictores $\widehat{y}_i^{(\ell)}$, $\ell = 0, 1, 2$ definidos en la sección 4.3 coinciden con y_i, con lo cual, en este caso, los estimadores $\widehat{m}_{\mathrm{S}}^{(1)}$, $\widehat{m}_{\mathrm{I}}^{(1,0)}$, $\widehat{m}_{\mathrm{I}}^{(1,1)}$ coinciden si $\gamma = h$, así como también los estimadores $\widehat{m}_{\mathrm{S}}^{(2)}$ y $\widehat{m}_{\mathrm{I}}^{(2,2)}$. Por esta razón, si $p(\mathbf{x}) \equiv 1$ sólo podemos comparar los estimadores $\widetilde{m}_{\mathrm{S}}$, $\widehat{m}_{\mathrm{S}}^{(1)}$ y $\widehat{m}_{\mathrm{S}}^{(2)}$ entre sí. Cuando no hay respuestas faltantes, el estimador $\widetilde{m}_{\mathrm{S}}$ es el mejor estimador en término de los MISE para

cualquier valor de h, entre los tres considerados. Los MISE de $\widehat{m}_{\mathrm{S}}^{(1)}$ son ligeramente superiores a los del estimador de Nadaraya–Watson $\widetilde{m}_{\mathrm{S}}$, mientras que los de $\widehat{m}_{\mathrm{S}}^{(2)}$, que está basado en el estimador internamente corregido y en el modelo aditivo, son por lo menos cuatro veces mayores que los de $\widetilde{m}_{\mathrm{S}}$.

Respecto a los estimadores de las componentes marginales, cuando no hay respuestas faltantes, los estimadores $\widehat{g}_{1,\mathrm{S}}^{(1)}$ y $\widehat{g}_{2,\mathrm{S}}^{(1)}$ tienen menor MISE que $\widehat{g}_{1,\mathrm{S}}^{(2)}$ y $\widehat{g}_{2,\mathrm{S}}^{(2)}$, respectivamente. Sin embargo, los estimadores $\widehat{g}_{2,\mathrm{S}}^{(1)}$ y $\widehat{g}_{2,\mathrm{S}}^{(2)}$ tienen un MISE similar cuando las ventanas utilizadas son 0.20 y 0.25. Observemos que $\widehat{g}_{1,\mathrm{S}}^{(2)}$ tiene valores de MISE casi 8 veces mayores que los de $\widehat{g}_{1,\mathrm{S}}^{(1)}$ lo cual podría explicar que las diferencias observadas en los MISE de $\widehat{m}_{\mathrm{S}}^{(1)}$ y $\widehat{m}_{\mathrm{S}}^{(2)}$ pueden atribuirse a la estimación de esta componente. El Cuadro 6.3 permite apreciar que, para las ventanas más pequeñas, los estimadores $\widehat{g}_{1,\mathrm{S}}^{(1)}$ y $\widehat{g}_{2,\mathrm{S}}^{(1)}$ tienen un mejor comportamiento fuera de la frontera. Los errores altos observados en el Cuadro 6.2 pueden atribuirse entonces a la mala estimación en puntos cercanos a la frontera. A medida que incrementamos el tamaño de la ventana, el estimador $\widehat{g}_{1,\mathrm{S}}^{(2)}$ logra reducir el MISE más de lo que lo hace $\widehat{g}_{1,\mathrm{S}}^{(1)}$, pero no ocurre lo mismo para su equivalente estimador de la componente g_2.

Los comentarios dados para $p \equiv 1$ son válidos para los otros mecanismos de pérdida de respuestas. Por otra parte, para todos los esquemas de pérdida de datos, los menores valores de MISE de $\widetilde{m}_{\mathrm{S}}$ y $\widehat{m}_{\mathrm{S}}^{(1)}$ se obtienen con h =0.15, mientras que para el estimador $\widehat{m}_{\mathrm{S}}^{(2)}$ el menor MISE se alcanza cuando la ventana utilizada es h =0.2. Esto parece ser un reflejo de lo que está ocurriendo con los estimadores de las funciones marginales. Los estimadores de g_1 tienen un MISE menor cuando la ventana es lo más chica posible, y lo mismo ocurre con el estimador $\widehat{g}_{2,\mathrm{S}}^{(1)}$. Sin embargo, el estimador $\widehat{g}_{2,\mathrm{S}}^{(2)}$ tiene notoriamente menor MISE cuando h =0.2.

Si comparamos los estimadores mediante el MISE, recomendaríamos utilizar, entre los tres estimadores simplificados, el estimador $\widetilde{m}_{\mathrm{S}}$ pues resulta ser más preciso que los otros dos bajo los esquemas de pérdida considerados en este estudio. Esto resulta llamativo ya que este estimador no supone que el modelo es aditivo, mientras que los otros dos sí. Sin embargo, ya hemos mencionado que los errores cuadráticos del estimador $\widehat{m}_{\mathrm{S}}^{(1)}$ no difieren en gran medida de los de $\widetilde{m}_{\mathrm{S}}$. Por otra parte, la ventaja asintótica de $\widehat{m}_{\mathrm{S}}^{(2)}$ sobre $\widehat{m}_{\mathrm{S}}^{(1)}$ se observa si $d > 4$ lo que permitiría explicar lo que se observa en nuestro estudio de simulación.

Por otra parte, si utilizamos el WMISE como criterio de bondad del estimador, para las dos ventanas más pequeñas, los estimadores $\widehat{g}_{1,\mathrm{S}}^{(1)}$ y $\widehat{g}_{2,\mathrm{S}}^{(1)}$ se comportan mejor que los estimadores $\widehat{g}_{1,\mathrm{S}}^{(2)}$ y $\widehat{g}_{2,\mathrm{S}}^{(2)}$. Sin embargo, para las dos ventanas más grandes, el estimador $\widehat{g}_{1,\mathrm{S}}^{(2)}$ resulta ser un mejor estimador de g_1 que $\widehat{g}_{1,\mathrm{S}}^{(1)}$, pero su equivalente $\widehat{g}_{2,\mathrm{S}}^{(2)}$ se convierte en un peor estimador de g_2. Debemos notar que los resultados del WMISE para ventanas grandes como h =0.30 se basan en las observaciones que yacen en el intervalo $[0.30, 0.70]$, o sea, en promedio en el cálculo del error sobre el 40 % de los datos.

6.3.2. Comportamiento de los estimadores imputados

Los Cuadros 6.4, 6.5 y 6.7 presentan los MISE de los estimadores imputados de m, g_1 y g_2, respectivamente, bajo los esquemas de pérdida de datos p_2 y p_3 y para los distintos pares de ventanas considerados en el proceso de suavizado. Por otra parte, en el Cuadro 6.6 y 6.8 se dan los WMISE de los estimadores imputados de las funciones marginales g_1 y g_2, respectivamente.

Para ambos mecanismos de pérdida de respuestas p_2 y p_3, los menores MISE de los estimadores $\widehat{m}_{\text{I}}^{(1,0)}$ y $\widehat{m}_{\text{I}}^{(1,1)}$, que se basan en imputar las observaciones mediante los estimadores de Nadaraya–Watson $\widetilde{m}_{\text{S}}$ y $\widehat{m}_{\text{S}}^{(1)}$ respectivamente, ocurren cuando $(h,\gamma) = (0.15, 0.1)$. Respecto al estimador $\widehat{m}_{\text{I}}^{(2,2)}$, el menor MISE puede observarse cuando $(h,\gamma) = (0.3, 0.15)$ bajo p_2 y cuando $(h,\gamma) = (0.25, 0.15)$ bajo p_3. Sin embargo, el valor de MISE en este último caso es más de cuatro veces superior a los de $\widehat{m}_{\text{I}}^{(1,0)}$ y $\widehat{m}_{\text{I}}^{(1,1)}$ y el mínimo valor del MISE de $\widehat{m}_{\text{I}}^{(2,2)}$ no parece diferir tanto del valor de MISE cuando $(h,\gamma) = (0.2, 0.15)$. Vale la pena mencionar que los MISE de los estimadores $\widehat{m}_{\text{I}}^{(1,0)}$ y $\widehat{m}_{\text{I}}^{(1,1)}$ son similares (del mismo orden), sin importar el tamaño de ventana que se haya utilizado para el cálculo del estimador simplificado. Sin embargo, para estos estimadores, cuanto más chicas son las ventanas usadas tanto para imputar las observaciones faltantes (h, usada para el cálculo del estimador simplificado) como para evaluar el estimador imputado (γ), mejor es la estimación. No ocurre lo mismo con el estimador $\widehat{m}_{\text{I}}^{(2,2)}$. Para éste, los errores parecieran comportarse más irregularmente. El mínimo se alcanza como dijimos en $(h,\gamma) = (0.3, 0.15)$ bajo p_2 y en $(h,\gamma) = (0.25, 0.15)$ bajo p_3, y los errores más altos se alcanzan en los extremos, o sea, asociados a utilizar ventanas grandes ya sea para predecir las observaciones mediante el estimador simplificado como para calcular el estimador imputado, o bien, ventanas chicas para ambos casos.

Respecto de la estimación de la componente aditiva g_1, tanto bajo p_2 como bajo p_3, los estimadores $\widehat{g}_{1,\text{I}}^{(1,0)}$ y $\widehat{g}_{1,\text{I}}^{(1,1)}$ tienen menor MISE cuando usamos los pares de ventanas más pequeñas para imputar las observaciones y para calcular el estimador imputado, pero en ninguno de los casos difieren mucho entre sí. El menor valor del MISE se obtiene cuando $(h,\gamma) = (0.15, 0.10)$. Sin embargo, para el estimador $\widehat{g}_{1,\text{I}}^{(2,2)}$ convendría usar las ventanas $(h,\gamma) = (0.3, 0.15)$ y los valores de MISE obtenidos son siempre sustancialmente mayores que los de los otros dos estimadores. El Cuadro 6.6 nos indica que a medida que aumenta el tamaño de la ventana γ utilizada para el cálculo del estimador imputado, disminuyen los valores del WMISE del estimador $\widehat{g}_{1,\text{I}}^{(2,2)}$ y se vuelven más parecidos a los otros dos estimadores de g_1 hasta lograr, bajo el mecanismo de pérdida p_2, un WMISE aún mejor para $\gamma = h$. Por otra parte, en este cuadro se aprecia que las diferencias entre los WMISE de los estimadores $\widehat{g}_{1,\text{I}}^{(1,0)}$ y $\widehat{g}_{1,\text{I}}^{(1,1)}$ resultan bastante mayores. Ambos estimadores alcanzan el mínimo cuando se usa una ventana para el imputado igual a $\gamma = 0.10$, pero el resultado de WMISE obtenido para $\widehat{g}_{1,\text{I}}^{(1,0)}$ es la mitad del de $\widehat{g}_{1,\text{I}}^{(1,1)}$, en ambos esquemas de pérdida de respuestas.

El Cuadro 6.7 muestra los MISE de los estimadores imputados de g_2. Tal como ocurría para g_1, los MISE de los estimadores $\widehat{g}_{2,\text{I}}^{(1,0)}$ y $\widehat{g}_{2,\text{I}}^{(1,1)}$ son más chicos cuando utilizamos las ventanas más pequeñas, es decir, $(h,\gamma) = (0.15, 0.10)$. Sin embargo, para el estimador $\widehat{g}_{2,\text{I}}^{(2,2)}$, el menor MISE ocurre cuando $(h,\gamma) = (0.20, 0.20)$ bajo p_2 y p_3 . En el Cuadro 6.8 se puede apreciar que, como para g_1, el estimador $\widehat{g}_{2,\text{I}}^{(1,0)}$ es siempre bastante mejor que $\widehat{g}_{2,\text{I}}^{(1,1)}$, mientras que el estimador $\widehat{g}_{2,\text{I}}^{(2,2)}$ tiene un valor de WMISE de por lo menos el doble que el de $\widehat{g}_{2,\text{I}}^{(1,1)}$, siendo en algunos casos mucho mayor. Como vemos, el efecto de frontera es determinante en el estimador de Nadaraya–Watson ya las mayores

diferencias entre $\widehat{g}_{2,\mathrm{I}}^{(1,0)}$ y $\widehat{g}_{2,\mathrm{I}}^{(1,1)}$ se visualizan dentro del intervalo $[\gamma, 1-\gamma]$. El efecto de frontera hace que el MISE de ambos estimadores sea prácticamente el mismo para casi todos los pares de ventanas consideradas.

6.3.3. Comparación entre el estimador simplificado y el estimador imputado

En esta sección compararemos los resultados obtenidos para los MISE de los estimadores simplificados e imputados a fin de decidir si conviene imputar las respuestas faltantes.

Si comparamos los MISE de los estimadores $\widetilde{m}_{\mathrm{S}}^{(1)}$ y $\widehat{m}_{\mathrm{I}}^{(1,0)}$, dados en los Cuadros 6.2 y 6.4, podemos ver que el hecho de imputar mejora la estimación en todos los casos salvo cuando $h-\gamma = 0$ o 0.5 o bajo p_2 cuando h=0.15. En la mayoría de los casos, cuando $h-\gamma = 0.5$, el MISE de los estimadores $\widetilde{m}_{\mathrm{S}}^{(1)}$ y $\widehat{m}_{\mathrm{I}}^{(1,0)}$ es bastante parecido. Lo mismo ocurre con $\widehat{m}_{\mathrm{S}}^{(1)}$ y $\widehat{m}_{\mathrm{I}}^{(1,1)}$. Los estimadores internamente normalizados se comportan de una manera diferente. Para valores chicos de h ($h = 0.15$ o 0.20), el estimador imputado es peor que el simplificado. Sin embargo, a medida que h es más grande, existen más valores de γ que logran disminuir el MISE. De la comparación anterior, podemos deducir que para los estimadores basados en Nadaraya–Watson imputar mejora la estimación en especial cuando la diferencia entre las ventanas h y γ es grande mientras que para los internamente normalizados, la ventaja del imputado se observa si uno elige ventanas h grandes para imputar las observaciones faltantes.

Respecto de los estimadores de las componentes marginales, g_1 y g_2, o sea, si comparamos $\widehat{g}_{1,\mathrm{S}}^{(1)}$ con $\widehat{g}_{1,\mathrm{I}}^{(1,1)}$ y $\widehat{g}_{2,\mathrm{S}}^{(1)}$ con $\widehat{g}_{2,\mathrm{I}}^{(1,1)}$, nuevamente observamos que logramos disminuir los errores al imputar los datos siempre y cuando la ventana usada para el imputado sea menor que la usada para el simplificado (ver Cuadros 6.2 y 6.5 y 6.2 y 6.7, respectivamente). Entre los estimadores $\widehat{g}_{1,\mathrm{S}}^{(2)}$ y $\widehat{g}_{1,\mathrm{I}}^{(2,2)}$ ocurre algo similar. Mientras la ventana h sea grande, es decir, tome valores 0.25 o 0.3, si la ventana γ es menor que h, imputar logra una mejoría en términos del MISE. Sin embargo, al comparar los MISE de los estimadores internamente normalizados de g_2, $\widehat{g}_{2,\mathrm{S}}^{(2)}$ y $\widehat{g}_{2,\mathrm{I}}^{(2,2)}$, podemos ver que la imputación no logra mejorar la estimación. Por otra parte, los estimadores $\widehat{g}_{1,\mathrm{I}}^{(1,0)}$ y $\widehat{g}_{2,\mathrm{I}}^{(1,0)}$ de de g_1 y g_2, respectivamente tienen valores de MISE menores que cualquier otro y parecen ser los que se deben recomendar a pesar de que para imputar las observaciones no hacen uso del modelo aditivo.

Vale la pena mencionar que los Cuadros 6.2 y 6.4 permiten observar que para los mecanismos de pérdida de observaciones p_2 y p_3, los menores valores de MISE de los estimadores $\widehat{m}_{\mathrm{I}}^{(1,0)}$, $\widehat{m}_{\mathrm{I}}^{(1,1)}$ y $\widehat{m}_{\mathrm{I}}^{(2,2)}$ son ligeramente superiores a los del caso en que no hay datos faltantes, o sea, cuando $p(\mathbf{x}) = p_1 \equiv 1$.

Finalmente podemos concluir que, en presencia de datos faltantes, la imputación logra mejorar la estimación cuando la ventana utilizada para imputar las observaciones, o sea, la del estimador simplificado, es suficientemente mayor que la del estimador imputado. Por otra parte, estimar la función de regresión en el primer paso, o sea para imputar las respuestas faltantes, con un estimador que no utilice integración marginal logra un estimador con menor MISE.

6.4. Cuadros

(a)

	$\widehat{c}_1$	$\widehat{c}_2$
Media	10.00002	9.99558
Desvío	0.13520	0.41963
ECM	0.01828	0.17611

(b)

	ℓ =1		ℓ =2	
	$\widehat{c}_1$	$\widehat{c}_2$	$\widehat{c}_1$	$\widehat{c}_2$
r =1	0.3046	2.1030	0.8131	2.8487
r =2	0.2891	0.3989	0.7800	1.1270

Cuadro 6.1: (a) Medidas resumen para los estimadores de la media de Y y (b) MISE de los estimadores $\widehat{g}^{(\ell}_{r,\mathrm{S}})$, $r, \ell = 1, 2$, en función del estimador de c utilizado bajo el mecanismo de pérdida p_3.

	$p_1(\mathbf{x}) \equiv 1$				$p_2(\mathbf{x}) \equiv 0.8$				$p_3(\mathbf{x}) = 0.4 + 0.5\cos^2(2x_1x_2 + 0.4)$			
$h =$	0.15	0.20	0.25	0.30	0.15	0.20	0.25	0.30	0.15	0.20	0.25	0.30
$\widetilde{m}^{(1)}_{\mathrm{S}}$	0.4341	0.6706	1.0172	1.4765	0.4476	0.6830	1.0290	1.4876	0.4799	0.7251	1.0880	1.5665
$\widehat{m}^{(1)}_{\mathrm{S}}$	0.5497	0.7646	1.0930	1.5357	0.5539	0.7694	1.0993	1.5431	0.5728	0.8025	1.1521	1.6185
$\widehat{m}^{(2)}_{\mathrm{S}}$	7.5561	2.4966	4.4023	7.7180	2.8241	2.5570	4.5312	7.9148	2.9052	2.6735	4.7758	8.3126
$\widehat{g}^{(1)}_{1,\mathrm{S}}$	0.1758	0.2851	0.4449	0.6540	0.1775	0.2864	0.4464	0.6558	0.1864	0.3030	0.4728	0.6930
$\widehat{g}^{(2)}_{1,\mathrm{S}}$	1.7465	2.0968	3.2514	4.8172	1.7652	2.1162	3.2828	4.8656	1.8161	2.1133	3.2826	4.8899
$\widehat{g}^{(1)}_{2,\mathrm{S}}$	0.1581	0.2663	0.4385	0.6755	0.1602	0.2689	0.4418	0.6790	0.1714	0.2879	0.4715	0.7206
$\widehat{g}^{(2)}_{2,\mathrm{S}}$	0.6087	0.2713	0.4566	0.9956	0.6388	0.2983	0.4950	1.0499	0.7650	0.3907	0.6045	1.1923

Cuadro 6.2: MISE de los estimadores simplificados de m, g_1 y g_2 bajo distintos esquemas de pérdida de datos y para distintas ventanas h.

	$p_1(\mathbf{x}) \equiv 1$				$p_2(\mathbf{x}) \equiv 0.8$				$p_3(\mathbf{x}) = 0.4 + 0.5\cos^2(2x_1x_2 + 0.4)$			
$h =$	0.15	0.20	0.25	0.30	0.15	0.20	0.25	0.30	0.15	0.20	0.25	0.30
$\widehat{g}^{(1)}_{1,\mathrm{S}}$	0.0611	0.0720	0.1019	0.1422	0.0635	0.0738	0.1034	0.1432	0.0669	0.0786	0.1093	0.1485
$\widehat{g}^{(2)}_{1,\mathrm{S}}$	0.3622	0.1329	0.0752	0.0470	0.3380	0.1340	0.0752	0.0469	0.3609	0.1373	0.0757	0.0473
$\widehat{g}^{(1)}_{2,\mathrm{S}}$	0.0576	0.0741	0.1149	0.1733	0.0596	0.0759	0.1163	0.1743	0.0635	0.0811	0.1230	0.1814
$\widehat{g}^{(2)}_{2,\mathrm{S}}$	0.2521	0.1262	0.2773	0.5256	0.2866	0.1499	0.3039	0.5547	0.3672	0.2339	0.4134	0.6802

Cuadro 6.3: WMISE de los estimadores simplificados de las funciones marginales bajo distintos esquemas de pérdida de datos y para distintas ventanas h.

		$p_2(\mathbf{x}) \equiv 0.8$					$p_3(\mathbf{x}) = 0.4 + 0.5\cos^2(2x_1x_2 + 0.4)$				
	$\gamma =$	0.10	0.15	0.20	0.25	0.30	0.10	0.15	0.20	0.25	0.30
$h = 0.30$	$\widehat{m}_{\mathrm{I}}^{(1,0)}$	0.5391	0.7049	0.9651	1.3263	1.7875	0.7320	0.9282	1.2163	1.5966	2.0660
	$\widehat{m}_{\mathrm{I}}^{(1,1)}$	0.5977	0.7601	1.0162	1.3728	1.8288	0.8553	1.0455	1.3260	1.6978	2.1578
	$\widehat{m}_{\mathrm{I}}^{(2,2)}$	4.2962	2.3859	3.8430	6.9949	11.2794	3.4810	3.4327	6.1222	10.1938	15.2371
$h = 0.25$	$\widehat{m}_{\mathrm{I}}^{(1,0)}$	0.5109	0.6649	0.9121	1.2632		0.6382	0.8164	1.0846	1.4505	
	$\widehat{m}_{\mathrm{I}}^{(1,1)}$	0.5840	0.7340	0.9761	1.3213		0.7903	0.9613	1.2202	1.5752	
	$\widehat{m}_{\mathrm{I}}^{(2,2)}$	5.3860	2.5261	3.3039	5.9619		4.4085	2.8614	4.4666	7.7448	
$h = 0.20$	$\widehat{m}_{\mathrm{I}}^{(1,0)}$	0.4869	0.6282	0.8633			0.5637	0.7215	0.9707		
	$\widehat{m}_{\mathrm{I}}^{(1,1)}$	0.5993	0.7115	0.9404			0.7475	0.8967	1.1345		
	$\widehat{m}_{\mathrm{I}}^{(2,2)}$	7.1745	2.9473	2.9293			6.6130	3.1146	3.4001		
$h = 0.15$	$\widehat{m}_{\mathrm{I}}^{(1,0)}$	0.4667	0.5959				0.5080	0.6462			
	$\widehat{m}_{\mathrm{I}}^{(1,1)}$	0.5699	0.6934				0.7240	0.8517			
	$\widehat{m}_{\mathrm{I}}^{(2,2)}$	9.3599	3.7323				10.6384	4.4380			

Cuadro 6.4: MISE de los estimadores imputados de m bajo distintos esquemas de pérdida de datos y para distintas ventanas h y γ.

		$p_2(\mathbf{x}) \equiv 0.8$					$p_3(\mathbf{x}) = 0.4 + 0.5\cos^2(2x_1x_2 + 0.4)$				
	$\gamma =$	0.10	0.15	0.20	0.25	0.30	0.10	0.15	0.20	0.25	0.30
$h = 0.30$	$\widehat{g}_{1,\mathrm{I}}^{(1,0)}$	0.1666	0.2515	0.3816	0.5558	0.7726	0.2615	0.3610	0.5035	0.6855	0.9047
	$\widehat{g}_{1,\mathrm{I}}^{(1,1)}$	0.1813	0.2648	0.3931	0.5652	0.7795	0.2895	0.3862	0.5252	0.7029	0.9175
	$\widehat{g}_{1,\mathrm{I}}^{(2,2)}$	2.1391	1.7668	2.7166	4.2193	6.0118	2.0124	2.2572	3.5491	5.2585	7.1955
$h = 0.25$	$\widehat{g}_{1,\mathrm{I}}^{(1,0)}$	0.1538	0.2331	0.3571	0.5268		0.2176	0.3085	0.4419	0.6176	
	$\widehat{g}_{1,\mathrm{I}}^{(1,1)}$	0.1728	0.2504	0.3722	0.5392		0.2542	0.3417	0.4708	0.6411	
	$\widehat{g}_{1,\mathrm{I}}^{(2,2)}$	2.4928	1.8076	2.5382	3.8980		2.2828	2.0608	3.0194	4.5130	
$h = 0.20$	$\widehat{g}_{1,\mathrm{I}}^{(1,0)}$	0.1427	0.2158	0.3342			0.1820	0.2632	0.3879		
	$\widehat{g}_{1,\mathrm{I}}^{(1,1)}$	0.1683	0.2371	0.3527			0.2282	0.3053	0.4246		
	$\widehat{g}_{1,\mathrm{I}}^{(2,2)}$	3.0376	1.9133	2.3818			2.9672	2.0758	2.6168		
$h = 0.15$	$\widehat{g}_{1,\mathrm{I}}^{(1,0)}$	0.1330	0.2001				0.1547	0.2264			
	$\widehat{g}_{1,\mathrm{I}}^{(1,1)}$	0.1607	0.2253				0.2115	0.2781			
	$\widehat{g}_{1,\mathrm{I}}^{(2,2)}$	3.7831	2.0997				4.2411	2.3689			

Cuadro 6.5: MISE de los estimadores imputados de $g_1$bajo distintos esquemas de pérdida de datos y para distintas ventanas h y γ.

		$p_2(\mathbf{x}) \equiv 0.8$					$p_3(\mathbf{x}) = 0.4 + 0.5\cos^2(2x_1x_2 + 0.4)$				
	$\gamma =$	0.10	0.15	0.20	0.25	0.30	0.10	0.15	0.20	0.25	0.30
$h = 0.30$	$\widehat{g}_{1,\mathrm{I}}^{(1,0)}$	0.0399	0.0493	0.0717	0.1133	0.1822	0.0586	0.0737	0.1026	0.1507	0.2246
	$\widehat{g}_{1,\mathrm{I}}^{(1,1)}$	0.0870	0.0894	0.1081	0.1436	0.1852	0.1260	0.1292	0.1524	0.1912	0.2316
	$\widehat{g}_{1,\mathrm{I}}^{(2,2)}$	0.7496	0.2314	0.1143	0.0781	0.0559	0.6390	0.2573	0.1720	0.1381	0.1075
$h = 0.25$	$\widehat{g}_{1,\mathrm{I}}^{(1,0)}$	0.0459	0.0551	0.0796	0.1279		0.0577	0.0720	0.1022	0.1557	
	$\widehat{g}_{1,\mathrm{I}}^{(1,1)}$	0.0873	0.0873	0.1022	0.1349		0.1183	0.1194	0.1370	0.1708	
	$\widehat{g}_{1,\mathrm{I}}^{(2,2)}$	0.9337	0.2950	0.1336	0.0831		0.8268	0.2932	0.1560	0.1093	
$h = 0.20$	$\widehat{g}_{1,\mathrm{I}}^{(1,0)}$	0.0521	0.0601	0.0854			0.0582	0.0701	0.0995		
	$\widehat{g}_{1,\mathrm{I}}^{(1,1)}$	0.0949	0.0862	0.0977			0.1173	0.1144	0.1261		
	$\widehat{g}_{1,\mathrm{I}}^{(2,2)}$	1.1951	0.3842	0.1671			1.2175	0.4136	0.1891		
h=0.15	$\widehat{g}_{1,\mathrm{I}}^{(1,0)}$	0.0595	0.0666				0.0625	0.0719			
	$\widehat{g}_{1,\mathrm{I}}^{(1,1)}$	0.0918	0.0860				0.1211	0.1132			
	$\widehat{g}_{1,\mathrm{I}}^{(2,2)}$	1.5302	0.4969				1.8117	0.6153			

Cuadro 6.6: WMISE de los estimadores imputados de g_1 bajo distintos esquemas de pérdida de datos y para distintas ventanas h y γ.

		$p_2(\mathbf{x}) \equiv 0.8$					$p_3(\mathbf{x}) = 0.4 + 0.5\cos^2(2x_1x_2 + 0.4)$				
	$\gamma =$	0.10	0.15	0.20	0.25	0.30	0.10	0.15	0.20	0.25	0.30
h=0.30	$\widehat{g}_{2,\mathrm{I}}^{(1,0)}$	0.1591	0.2398	0.3737	0.5650	0.8132	0.2608	0.3586	0.5088	0.7118	0.9657
	$\widehat{g}_{2,\mathrm{I}}^{(1,1)}$	0.1699	0.2493	0.3814	0.5706	0.8164	0.2856	0.3805	0.5270	0.7259	0.9752
	$\widehat{g}_{2,\mathrm{I}}^{(2,2)}$	1.2967	0.4773	0.5231	1.0029	1.7879	1.2026	0.7487	1.0625	1.7629	2.7428
h=0.25	$\widehat{g}_{2,\mathrm{I}}^{(1,0)}$	0.1443	0.2186	0.3455	0.5310		0.2106	0.2988	0.4381	0.6328	
	$\widehat{g}_{2,\mathrm{I}}^{(1,1)}$	0.1592	0.2318	0.3563	0.5392		0.2433	0.3280	0.4627	0.6521	
	$\widehat{g}_{2,\mathrm{I}}^{(2,2)}$	1.5433	0.5185	0.4104	0.7701		1.3890	0.6188	0.6849	1.1898	
h=0.20	$\widehat{g}_{2,\mathrm{I}}^{(1,0)}$	0.1319	0.1994	0.3195			0.1713	0.2487	0.3773		
	$\widehat{g}_{2,\mathrm{I}}^{(1,1)}$	0.1576	0.2166	0.3338			0.2131	0.2863	0.4094		
	$\widehat{g}_{2,\mathrm{I}}^{(2,2)}$	1.9786	0.6393	0.3472			1.9065	0.7111	0.4688		
h=0.15	$\widehat{g}_{2,\mathrm{I}}^{(1,0)}$	0.1216	0.1827				0.1425	0.2093			
	$\widehat{g}_{2,\mathrm{I}}^{(1,1)}$	0.1456	0.2041				0.1938	0.2556			
	$\widehat{g}_{2,\mathrm{I}}^{(2,2)}$	2.4982	0.8607				2.8980	1.0862			

Cuadro 6.7: MISE de los estimadores de g_2 bajo distintos esquemas de pérdida de datos y para distintas ventanas h y γ.

		$p_2(\mathbf{x}) \equiv 0.8$					$p_3(\mathbf{x}) = 0.4 + 0.5\cos^2(2x_1x_2 + 0.4)$				
	$\gamma =$	0.10	0.15	0.20	0.25	0.30	0.10	0.15	0.20	0.25	0.30
h=0.30	$\widehat{g}_{2,\mathrm{I}}^{(1,0)}$	0.0418	0.0545	0.0833	0.1362	0.2213	0.0653	0.0849	0.1217	0.1823	0.2733
	$\widehat{g}_{2,\mathrm{I}}^{(1,1)}$	0.0825	0.0869	0.1129	0.1612	0.2226	0.1267	0.1310	0.1625	0.2152	0.2770
	$\widehat{g}_{2,\mathrm{I}}^{(2,2)}$	0.7590	0.2615	0.3481	0.6208	0.9205	0.7758	0.5522	0.7993	1.1568	1.4726
h=0.25	$\widehat{g}_{2,\mathrm{I}}^{(1,0)}$	0.0451	0.0572	0.0873	0.1452		0.0595	0.0775	0.1143	0.1783	
	$\widehat{g}_{2,\mathrm{I}}^{(1,1)}$	0.0813	0.0832	0.1055	0.1504		0.1140	0.1168	0.1422	0.1896	
	$\widehat{g}_{2,\mathrm{I}}^{(2,2)}$	0.8975	0.2491	0.2494	0.4749		0.8280	0.3762	0.4955	0.7913	
h=0.20	$\widehat{g}_{2,\mathrm{I}}^{(1,0)}$	0.0494	0.0595	0.0892			0.0567	0.0710	0.1052		
	$\widehat{g}_{2,\mathrm{I}}^{(1,1)}$	0.0855	0.0810	0.0995			0.1098	0.1085	0.1276		
	$\widehat{g}_{2,\mathrm{I}}^{(2,2)}$	1.1259	0.2845	0.1807			1.1152	0.3448	0.2920		
h=0.15	$\widehat{g}_{2,\mathrm{I}}^{(1,0)}$	0.0554	0.0633				0.0586	0.0688			
	$\widehat{g}_{2,\mathrm{I}}^{(1,1)}$	0.0844	0.0800				0.1117	0.1051			
	$\widehat{g}_{2,\mathrm{I}}^{(2,2)}$	1.4693	0.3736				1.7003	0.4768			

Cuadro 6.8: WMISE de los estimadores de g_2 bajo distintos esquemas de pérdida de datos y para distintas ventanas h y γ.

Capítulo 7

Conclusiones

En esta tesis se han propuesto estimadores para el modelo aditivo con respuestas faltantes y se ha demostrado la consistencia de los mismos. Mediante un estudio de simulación se han comparado algunos de los estimadores propuestos y se ha concluído que

- la imputación de los datos mejora la estimación siempre y cuando la diferencia entre la ventana del primer y la del segundo paso sea grande
- algunos de los estimadores propuestos se comportan bien fuera de la frontera, mientras que otros se comportan de manera más homogénea en todo el soporte de las covariables y que
- imputar los datos faltantes sin la suposición de que el modelo es aditivo y calcular finalmente, basado en la muestra imputada, el estimador que utilice integración marginal es una buena propuesta para estimar la función de regresión bajo un modelo aditivo, dado que reduce los errores cuadráticos de los demás estimadores y lleva a valores del MISE ligeramente superiores a los del estimador que hubiésemos calculado si no hubiéramos observado datos faltantes.

El estudio del comportamiento asintótico de los estimadores propuestos será objeto de estudio posterior y nos permitirá validar teóricamente lo observado en el estudio de Monte Carlo.

Bibliografía

[1] Afifi, A. y Elashoff, R. (1969). Missing observations in multivariate statistics III: large sample analysis of simple linear regression. *J. Amer. Statist. Assoc.*, **64**, 337-358.

[2] Boente, G. y Fraiman, R. (1991). Strong Uniform Convergence Rates for Some Robust Equivariant Nonparametric Regression Estimates for Mixing Processes. *International Satistical Review*, **59**, 355-372.

[3] Boente, G., González-Manteiga, W. y Pérez-González, A. (2009). Robust nonparametric estimation with missing data. *Journal of Statistical Planning and Inference*, **139**, 571-592.

[4] Bosq, D. (1998). *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction*. Springer-Verlag.

[5] Buja, A., Hastie, T. y Tibshirani, R. (1989). Linear smoothers and additive models (with discussion). *Annals of Statistics*, **17**. 453-555.

[6] Chu, C. y Chen, P. (1993). Nonparametric regression estimation with missing data. *Journal of Statistical Planning and Inference*, **48**, 85-99.

[7] Chu, C. y Chen, P. (1996). Kernel estimation of distribution functions and quantiles with missing data. *Statistica Sinica*, **6**, 63-78.

[8] Cochran, W. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, **24**, 295-313.

[9] Devroye, L. (1978). The uniform convergence of the Nadaraya-Watson regression function estimate. *The Canadian Journal of Statistics*, **6**, 179-191.

[10] González–Manteiga, W. y Pérez–González, A. (2004). Nonparametric mean estimation with missing data. *Comm. Statist. Theory Methods*, **33**, 277-303.

[11] Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.*, **21**, 196-216.

[12] Härdle, W., Müller, M., Sperlich, S. y Werwatz, A. (2004). *Nonparametric and Semiparametric Models*. Springer.

[13] Hastie, T.J. y Tibshirani, R.J. (1990). *Generalized Additive Models*. Monographs on Statistics and Applied Probability No. 43. Chapman and Hall, London.

[14] Hengartner, N. y Sperlich, S. (2005). Rate optimal estimation with the integration method in the presence of many covariates. *Journal of Multivariate Analysis*, **95**, 246-272.

[15] Linton, O. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, **1**, 93-100.

[16] Linton, O. (1997). Efficient estimation of additive nonparametric regression models. *Biometrika*, **84**, 469-473.

[17] Linton, O. y Nielsen, J. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, **82**, 93-101.

[18] Nadaraya, E. (1964). On estimating regression. *Theory probability and Applications*, **10**, 186-190.

[19] Parzen, E. (1962). On the estimation of a probability density function and the mode. *Annals of Mathematical Statistics*, **33**, 1065-1076.

[20] Prakasa Rao, B. (1983) *Nonparametric Functional Estimation*. Academic Press.

[21] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, **27**, 832-837.

[22] Sperlich, S. Tjøstheim, D. y Yang, L. (2002) Nonparametric estimation and testing of interaction in additive models. *Econometric Theory*, **18**, 197-251.

[23] Stone, C.J. (1985). Additive regression and other nonparametric models. *Annals of Statistics*, **13**, 689-705.

[24] Tjøstheim, D. y Auestad, B. (1994) Nonparametric identification of nonlinear time series: Selecting significant lags. *Journal of the American Statistical Association*, **89**, 1410-1430.

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

[26] Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Emporium J. Exp. Agriculture*, **1**, 129-142.