Universidad de Granada

Facultad de Ciencias

Departamento de Estadística e Investigación Operativa

TRABAJO FIN DE MASTER:

ESTIMACION NO PARAMETRICA Y SEMIPARAMETRICA EN POBLACIONES FINITAS

Flor Alba Ruiz Arias

Tutor: Pr. Dr. D. Ismael Sánchez Borrero

GRANADA
Septiembre, 2012

Trabajo Presentado por Flor Alba Ruiz Arias

Dirigido por Ismael Ramón Sánchez Borrego

Máster Oficial en Estadística Aplicada

DEPARTAMENTO DE ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

UNIVERSIDAD DE GRANADA

Índice general

Capítulo 1

Métodos de regresión no paramétrica y semiparamétrica

1.1. Introducción

Los métodos de regresión no paramétrica y semiparamétrica son herramientas estadísticas útiles que reciben la atención de numerosos investigadores de diversas áreas de la Estadística. Estos métodos permiten realizar análisis de datos, predicciones y hacer inferencia sin tener que especificar completamente un modelo paramétrico para los datos. En el contexto de las encuestas por muestreo, su empleo es mucho menos conocido. En este trabajo nos centramos en los métodos no paramétricos y semiparamétricos en dos áreas importantes de la estadística: la estimación de densidades y la estimación de la función de regresión.

En la próxima sección, se describe brevemente su aplicación en el contexto de datos independientes e idénticamente distribuidos (*iid*). Las siguientes secciones parten del supuesto que las observaciones provienen de encuestas complejas.

En términos generales los métodos no paramétricos son aquellos que no asumen una forma paramétrica de las características principales de interés en los datos (aunque podría haber supuestos paramétricos sobre algunas de las "características", por ejemplo, la varianza en el caso de la regresión). Por el contrario los métodos semiparamétricos usan una combinación de especificación paramétrica y no paramétrica para las características principales de interés.

1.2. Métodos tipo núcleo

Los métodos tipo núcleo son empleados tanto para estimación de densidades como para estimación de la función de regresión. Se describe a continuación el estimador tipo núcleo de la función de densidad en el caso univariante. Sean $X_1, ..., X_n$ y suponemos los x_i son *i.i.d.* con función de densidad desconocida $f_x(\cdot)$. Suponemos además que esta función es suave en el sentido de continuidad y diferenciabilidad. En la monografía de Wand y Jones (1995) se incluye una completa revisión de estos métodos.

Dado x, un estimador de la densidad tipo núcleo $\widehat{f}_x(x;h)$ se define

$$\widehat{f}_x(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right), \tag{1.1}$$

donde $K(\cdot)$ denota la función núcleo y el parámetro h es el *ancho de banda* o *parámetro de suavizado*. La función núcleo $K(\cdot)$ es habitualmente una densidad de probabilidad simétrica. Esta función determina qué pesos asigna a cada observación basándose en las distancias de las observaciones de la muestra para construir el estimador.

El ancho de banda h determina la suavidad del estimador $\widehat{f}_x(x;h)$: valores pequeños de h dan lugar a estimaciones más "apuntadas" y valores grandes implican estimaciones más suavizadas. El ancho de la banda determina el equilibrio entre el sesgo y la varianza del estimador de densidad tipo núcleo $\widehat{f}_x(x;h)$. Con un valor grande de h se obtiene mayor sesgo y menor varianza que con un valor h pequeño.

Consideramos el problema de estimación de la función de regresión. Podemos distinguir dos tipos de modelos de regresión atendiendo a los supuestos que se establecen sobre la función de regresión m;

1. Modelo de regresión paramétrica: asume que la función de regresión tiene forma predeterminada.

2. Modelo de regresión no paramétrica: sólo supone hipótesis de suavidad (en el sentido de continuidad y diferenciabilidad) sobre la función de regresión m. Tampoco se asume ninguna forma predefinida como el anterior para la función de regresión.

La regresión no paramétrica tiene cuatro propósitos principales para estimar una curva de regresión:

1. Proporciona un método versátil para estudiar la relación general entre dos variables

2. Permite dar predicciones de las observaciones, aunque éstas no tengan referencia a ningún modelo paramétrico fijo

3. Proporciona una herramienta para encontrar falsas observaciones para estudiar la influencia de los puntos aislados

4. Contribuye a la creación de un método flexible de imputacion de los valores faltantes

En regresión no paramétrica es relevante el estimador de la función de regresión llamado estimador lineal local tipo núcleo, introducido entre otros por Fan and Gijbels (1996), que se destaca por sus buenas propiedades teóricas y prácticas. Los métodos no paramétricos son más apropiados cuando no se tiene conocimiento previo de la relación entre las variables objeto de estudio puesto que sólo parten de supuestos de suavidad sobre la función de regresión. Estos métodos son computacionalmente costosos debido al gran número de operaciones que involucran y son sólo aplicables en la práctica con la ayuda de un programa informático.

Ahora bien, supongamos que tenemos un conjunto de datos con observaciones $(X_1, Y_1), \cdots, (X_n, Y_n)$ y estamos interesados en estimar la función de regresión $m(\cdot)$ en el modelo

$$Y_i = m(x_i) + \varepsilon_i. \tag{1.2}$$

Por simplicidad, suponemos que ε_i son *i.i.d.* con media 0 y varianza σ^2. El método de estimación tipo núcleo más habitual es la *regresión polinómica local*, y en particular el estimador lineal local tipo núcleo. (ver la monografía de Fan and Gijbels, 1996). Sea q el grado del estimador de regresión polinómica local. Dado x, el estimador $\widehat{m}(x)$ está definido por $\widehat{\beta}_0$, donde los $\widehat{\beta}_0, \ldots, \widehat{\beta}_q$ se obtienen resolviendo el siguiente problema de mínimos cuadrados ponderados:

$$\min_{\beta} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) \Big(Y_i - \beta_0 - \beta_1\big(x_i - x\big) - \cdots - \beta_q\big(x_i - x\big)^q\Big)^2.$$

Este estimador puede ser escrito en notación matricial como

$$\widehat{m}(x) = e_1^T \left(\mathbf{X}_x^T \mathbf{W}_x(h) \mathbf{X}_x\right)^{-1} \mathbf{X}_x^T \mathbf{W}_x(h) \mathbf{Y}, \tag{1.3}$$

con $e_1 = (1, 0, \ldots, 0)^T$, $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, la matriz $\mathbf{W}_x$ está dada por

$$\mathbf{W}_x = \operatorname{diag}\{K((x_1 - x)/h), \ldots, K((x_n - x)/h)\}$$

y la matriz de diseño $\mathbf{X}_x$ se define como:

$$\mathbf{X}_x = \begin{bmatrix} 1 & x_1 - x & \cdots & (x_1 - x)^q \\ \vdots & \vdots & & \vdots \\ 1 & x_n - x & \cdots & (x_n - x)^q \end{bmatrix}. \tag{1.4}$$

El equilibrio entre el sesgo y la varianza del estimador $\widehat{m}(x)$ depende del ancho de banda h. Es evidente que el estimador polinómico local puede escribirse como una combinación lineal de los Y_i, $\widehat{m}(x) = \sum w_i(x) Y_i$. Esto será de utilidad al aplicar este método en el contexto de las encuestas por muestreo.

En Wand y Jones (1995) y en Fan y Gijbels (1996) puede encontrarse más información sobre las propiedades teóricas y áreas de aplicación de la regresión polinómica local.

1.3. Métodos splines y otros métodos

Los métodos splines suelen emplearse para la estimación de la función de regresión. Consideremos de nuevo un conjunto de datos con observaciones $(X_1, Y_1), \ldots, (X_n, Y_n)$, que siguen el modelo (1.2) con errores *i.i.d.*. La función de regresión $m(\cdot)$ se supone suave en el sentido de continuidad y diferenciabilidad. Suponemos además que se aproxima bien mediante una *función spline*. Las funciones spline polinómicas están definidas por:

$$m(x; \boldsymbol{\beta}) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{j=1}^{J} \beta_{p+j} (x - \kappa_j)_+^p, \tag{1.5}$$

donde $p \geq 1$ es el grado del spline, $\kappa_1, \ldots, \kappa_j$ son los *nodos* y la función $(\cdot)_+^p$ está dada por

$$(x - \kappa)_+^p = \begin{cases} (x - \kappa)^p & \text{si } x > \kappa \\ 0 & \text{en otro caso.} \end{cases}$$

Los modelos spline lineal ($p = 1$) y cúbico ($p = 3$) son elecciones habituales en la práctica. Otras formulaciones de la función spline $m(x; \boldsymbol{\beta})$ son posibles, en las que el conjunto de *funciones base*
$\left\{1, x, \ldots, x^p, (x - \kappa_1)_+^p, \ldots, (x - \kappa_J)_+^p\right\}$ es reemplazado por un conjunto diferente. Por ejemplo, los B-splines (Boor, 2001) son un conjunto de funciones base muy empleadoas con mejores propiedades teóricas que la del propio spline polinómico. La mayoría de estas formulaciones, incluyendo B-splines, pueden escribirse de forma equivalente al spline polinómico antes introducido, por lo que restringimos nuestra atención a (1.5).

Aunque existen diferentes métodos de regresión por splines, nos centramos en la regresión spline penalizada, debido a su facilidad de uso y a su buena aplicabilidad en las encuestas por muestreo. Una excelente visión general de este método y sus aplicaciones en el amplio contexto de la regresión, se proporciona en Ruppert et al. (2003). Es evidente que $m\left(x;\boldsymbol{\beta}\right)$ es esencialmente una función paramétrica y que las desviaciones del polinomio de orden p-ésimo pueden solamente ocurrir en los nodos, de forma que, la flexibilidad del spline como representación de una función desconocida está determinada por el número y la localización de los nodos. Para garantizar que $m\left(x;\boldsymbol{\beta}\right)$ es lo suficientemente flexible, el spline penalizado establece un número grande de nodos J, esto es, al menos $J = n/4$ y la localización de ellos en ciertos cuantiles adecuados de los x_i.

El ajuste del modelo spline a las observaciones se realiza mediante la minimización por mínimos cuadrados, pero añadiendo una penalización para asegurar la existencia de una solución y reducir el potencial aumento en la varianza debido a la gran cantidad de parámetros que se necesitan estimar. El estimador de $m(\cdot)$ es $m(x;\widehat{\boldsymbol{\beta}})$ empleando (1.5), donde $\widehat{\boldsymbol{\beta}}$ es el estimador que hace mínimo

$$\sum_{i=1}^{n}\left(Y_i-\beta_0-\beta_1x_i-\cdots-\beta_px_i^p-\sum_{j=1}^{J}\beta_{p+j}\left(x_i-\kappa_j\right)_+^p\right)^2+\lambda\sum_{j=1}^{J}\beta_{p+j}^2, \qquad (1.6)$$

donde λ es el parámetro de suavizado. Este parámetro juega un papel análogo al ancho de banda h en los métodos de regresión tipo núcleo, puesto que determina el equilibrio entre el sesgo y la varianza del estimador $m(\cdot;\widehat{\boldsymbol{\beta}})$. Valores grandes de λ, produce un sesgo mayor y menor varianza que valores pequeños de λ. Dado que sólo la parte no polinómica de los coeficientes spline está penalizada en (1.6), λ determina la cantidad de desviación desde una función polinómica de grado p-ésimo.
Debido a la naturaleza flexible de la función spline (1.5) y a la presencia del parámetro λ, la regresión spline penalizada se considera habitualmente un método no paramétrico. Sin embargo, comparte muchas características con la regresión paramétrica debido a que el número de parámetros es fijo (en $J+p+1$) y el estimador se deduce como solución de un problema de mínimos cuadrados.

Existen otros métodos de regresión spline tales como los de regresión spline (no penalizada) y los de regresión spline suavizada. En el primero, se determina una función spline con un número pequeño de nodos y la función se ajusta sin penalización de modo que se necesita una cuidadosa atención en la ubicación de los nodos a fin de evitar un elevado sesgo. En la regresión spline suavizada, el enfoque de formulación es diferente al anterior, el estimador es esencialmente equivalente a un

spline polinómico como (1.5) pero con un nodo en cada observación x_i y un término de penalización en la derivada de la función.

Descomposiciones ortogonales, en particular la descomposición de ondas (Vidakovi'c, 1999), es un método de regresión no paramétrico con buenas propiedades estadísticas que es aplicable en situaciones donde la función media no es necesariamente suave. Redes neuronales (Ripley, 1996) son una clase de métodos conceptualmente relacionados con la regresión spline penalizada, en la que los parámetros se encuentran a través de la regresión no lineal. Finalmente, los métodos basados en la clasificación, tales como árboles de regresión y clasificación (Breiman et al., 1984) y splines de regresión adaptativa multivariante (MARS) (Friedman, 1991) pueden utilizarse como métodos de regresión no paramétrica.

1.4. Ajuste de modelos más complejos

Es posible extender los métodos presentados anteriormente al caso multivariante, pero la llamada "maldición de dimensionalidad" hace que sea impracticable a partir de una dimensión igual o superior a tres. Esto implica que la flexibilidad del modelo tiene que disminuir a medida que la dimensión en el espacio de variables aumenta para obtener ajustes satisfactorios. Esto podría hacerse incrementando la cantidad de suavizado (mediante el uso de un ancho de banda o del parámetro de suavizado) o usando un número reducido de nodos (en el caso de los splines), pero planteamientos de mayor utilidad son los de reemplazar el modelo completo no paramétrico por un modelo restringido con más especificaciones. Se discuten dos casos especiales importantes de estos modelos: modelos aditivos y modelos semiparamétricos.

Sea $\mathbf{X}_i = (X_{1i}, \ldots, X_{iD})^T$ un vector de D covariables. En los modelos aditivos, el modelo (1.2) se reemplaza por

$$Y_i = m_1 (X_{1i}) + \cdots + m_D X_{Di} + \varepsilon_i, \tag{1.7}$$

donde las funciones $m_d(\cdot)$ son habitualmente univariantes y bajo supuestos de suavidad, no restringidas a pertenecer a una familia paramétrica específica. Este modelo fue dado a conocer por Hastie y Tibshirani (1990), quienes propusieron métodos de estimación basados en el algoritmo backfitting. Este enfoque, que se lleva a cabo en S-Plus y R, se basa en la aplicación iterativa de métodos de una dimensión tales como la regresión polinómica local o la regresión spline para los residuos a partir de los ajustes con respecto a las otras variables. Mientras que otros métodos de ajuste del modelo (1.7) han sido propuestos desde entonces, el algoritmo backfitting sigue siendo popular hoy en día.

En un modelo semiparamétrico, el término no paramétrico se combina con un término paramétrico. Sea $\mathbf{Z}_i = (Z_{1i}, \ldots, Z_{Pi})^T$ que representa las covariables adicionales que forman parte de dicho término paramétrico. Un modelo habitual semiparámetrico viene dado por

$$Y_i = m(X_i) + \mathbf{Z}_i^T \boldsymbol{\beta} + \varepsilon_i, \tag{1.8}$$

donde el término no paramétrico $m(X_i)$, puede ser multivariante y por tanto modelado como un modelo aditivo. *Backfitting* puede ser también aplicado para ajustar el modelo (1.8), pero existen en literatura otros métodos especialmente diseñados para modelos semiparamétricos.

Capítulo 2

Métodos no paramétricos en poblaciones finitas

2.1. Introducción

Consideremos el empleo de métodos no paramétricos para realizar inferencia sobre una población $U = \{1, \ldots, i, \ldots, N\}$ y se consideran las variables objeto de estudio y_i, z_i, etc. Suponemos que para cada $i \in U$, se observa un vector auxiliar $\mathbf{x}_i$. Sea $t_x = \sum_{i \in U} \mathbf{x}_i$. Una muestra probabilística $s \subset U$ de tamaño fijo se extrae de acuerdo a un diseño de muestral $p(\cdot)$, donde $p(s)$ es la probabilidad de que la muestra sea seleccionada. Sea $\pi_i = \mathrm{P}[i \in s] = \sum_{s: i \in s} p(s) > 0$ y $\pi_{ij} = \mathrm{P}[i, j \in s]$ para todo $i, j \in U$.

Pretendemos obtener estimaciones puntuales y un intervalo de confianza asociado a un parámetro poblacional, como el total $t_y = \sum_{i \in U} y_i$, o la media $\overline{y}_U = N^{-1} t_y$. Una proporción es un caso particular de la media poblacional, con y_i igual a una función indicadora. En particular, la función de distribución de una poblacional finita, denotada por $F_y(z) = N^{-1} \sum_{i \in U} I_{\{y_i \leq z\}}$ con $I_{\{A\}} = 1$ si A es cierto, y 0 en otro caso, es una proporción para cada z fijo. Otros parámetros poblacionales finitos de interés incluyen las razones $\sum_{i \in U} y_i / t_y = \sum_{i \in U} z_i$ y los vectores de coeficientes de regresión,

$$\mathbf{B} = \left(\sum_{i \in U} \mathbf{x}_i^T \mathbf{x}_i \right)^{-1} \sum_{i \in U} \mathbf{x}_i y_i.$$

Cada uno de estos ejemplos se construye a partir de los totales de una población finita, de forma que estimar el total poblacional de una variable y, constituye un problema recurrente en este contexto.

El estimador de Horvitz-Thompson de t_y,

$$\widehat{t}_y = \sum_{i \in s} \frac{y_i}{\pi_i}, \tag{2.1}$$

es un estimador insesgado del total poblacional t_y con varianza bajo el diseño, dada por

$$\mathrm{Var}_p\left(\widehat{t}_y\right) = \sum_{i,j \in U} \left(\pi_{ij} - \pi_i \pi_j\right) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}. \tag{2.2}$$

No obstante, si se dispone de variables auxiliares, pueden obtenerse estimadores más eficientes que $\widehat{t}_y$.

Sea el modelo de superpoblación, ξ, dado por

$$y_i = \mu\left(\mathbf{x}_i\right) + \varepsilon_i, \tag{2.3}$$

donde ε_i es una sucesión de variables aleatorias independientes con media cero y varianza $v\left(\mathbf{x}_i\right)$. Es habitual considerar modelos de superpoblación paramétricos, y en particular, de tipo lineal, esto es $\mu\left(\mathbf{x}_i\right) = \mathbf{x}_i^T \boldsymbol{\beta}$. La desventaja potencial de los estimadores obtenidos por este modelo de superpoblación es la ineficiencia cuando el modelo está mal especificado. Si el modelo de regresión no se ajusta bien a los datos no hay ninguna mejora con respecto al estimador de Horvitz-Thompson y esto puede conducir a una pérdida de eficiencia. Para evitar esto, se reemplaza la especificación paramétrica por una no paramétrica, en la que $\mu\left(\cdot\right)$ es una función suave en el sentido de continuidad y diferenciabilidad.

Una vez el modelo es ajustado con los datos de la muestra, hay al menos dos formas de incorporar estas predicciones dentro de la estimación del total de la población finita. La primera es mediante un enfoque *basado en el modelo*, en el que los valores del modelo ajustado $\tilde{\mu}\left(\mathbf{x}_i\right)$ se emplean para predecir únicamente los valores no muestreados de y:

$$\widehat{t}_{MB} = \sum_{i \in U \setminus s} \tilde{\mu}\left(\mathbf{x}_i\right) + \sum_{i \in s} y_i. \tag{2.4}$$

En general, los estimadores basados en el modelo de este tipo, son insesgados y eficientes cuando $\mu\left(\mathbf{x}_i\right)$ y $v\left(\mathbf{x}_i\right)$ están correctamente especificados, pero sesgados e incluso inconsistentes, si el modelo es incorrecto. Inspirados por la aplicabilidad general de los modelos no paramétricos, Kuo (1988), Dorfman (1992) y Chambers et al. (1993) desarrollaron estimadores basados en el modelo empleando regresión no paramétrica.

La segunda manera de incorporar las predicciones al modelo es el *modelo-asistido*, que evita los posibles problemas de mala especificación del modelo a través de un ajuste del sesgo en el diseño. La estimación en el modelo-asistido está basada en las predicciones $\widehat{\mu}_i$ para todos los elementos de la población y se incluye una corrección del sesgo del diseño en esa predicción. El estimador modelo-asistido resultante es de la forma

$$\widehat{t}_{MA} = \sum_{i\in U} \widehat{\mu}_i + \sum_{i\in s} \frac{y_i - \widehat{\mu}_i}{\pi_i}. \tag{2.5}$$

Los métodos no paramétricos, como el basado en el modelo y el modelo-asistido pueden ser empleados para mejorar la precisión de los estimadores de la función de distribución. (véase el trabajo de Johnson et al., 2008)

2.2. Regresión tipo núcleo en poblaciones finitas

Se describe a continuación la construcción de un estimador modelo-asistido. Sea K una función núcleo continua y h es el parámetro ancho de banda. Presentamos el estimador polinómico local tipo núcleo de grado q basado en los datos poblacionales. Sea $\mathbf{y}_U = [y_i]_{i\epsilon U_N}$ el vector de dimensión N de y_i en la población finita. Definimos la matriz $N \times (q+1)$

$$\mathbf{X}_{Ui} = \begin{pmatrix} 1 & x_1 - x_i & \dots & (x_1 - x_i)^q \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_N - x_i & \dots & (x_N - x_i)^q \end{pmatrix} = [1 \quad x_j - x_i \quad \dots \quad (x_j - x_i)^q]_{j\epsilon U},$$

y sea la matriz,

$$\mathbf{W}_{Ui} = diag\left\{ \frac{1}{h} K\left(\frac{x_j - x_i}{h} \right) \right\}_{j\epsilon U}.$$

Sea e_r el vector con un 1 en la posición $r-$ésima y ceros en el resto. El estimador polinómico local de la función de regresión en x_i basado en la población, viene dado por

$$\mu_i = \mathbf{e}_1'(\mathbf{X}_{Ui}'\mathbf{W}_{Ui}\mathbf{X}_{Ui})^{-1}\mathbf{X}_{Ui}'\mathbf{W}_{Ui}\mathbf{y}_U = \mathbf{w}_{Ui}'\mathbf{y}_U, \tag{2.6}$$

que estará bien definido si $\mathbf{X}_{Ui}'\mathbf{W}_{Ui}\mathbf{X}_{Ui}$ es invertible.

Si los μ_i son conocidos, entonces un estimador insesgado de t_y basado en el diseño está dado por el estimador de diferencia generalizado

$$t_y^* = \sum_{i\epsilon s} \frac{y_i - \mu_i}{\pi_i} + \sum_{i\epsilon U} \mu_i, \tag{2.7}$$

y su varianza sería

$$Var_p(t_y^*) = \sum_{i,j\epsilon U}(\pi_{ij} - \pi_i\pi_j)\frac{y_i - \mu_i}{\pi_i}\frac{y_j - \mu_j}{\pi_j}. \tag{2.8}$$

El estimador poblacional μ_i es el estimador tradicional de regresión polinómico local para la función desconocida $\mu(\cdot)$. Sin embargo, no puede ser calculado, porque sólo conocemos los y_i de $s \subset U$. Sea $\mathbf{y}_s = [y_i]_{i\epsilon s}$ el vector de dimensión n de y_i obtenido de la muestra. Definimos la matriz de dimensión $n \times (q+1)$

$$\mathbf{X}_{si} = [1 \quad x_j - x_i \quad \ldots \quad (x_j - x_i)^q]_{j\epsilon s},$$

y sea

$$\mathbf{W}_{si} = diag\left\{\frac{1}{\pi_i h}K\left(\frac{x_j - x_i}{h}\right)\right\}_{j\epsilon s}.$$

Un estimador muestral de μ_i basado en el diseño viene dado por

$$\hat{\mu}_i^o = \mathbf{e}_1'(\mathbf{X}_{si}'\mathbf{W}_{si}\mathbf{X}_{si})^{-1}\mathbf{X}_{si}'\mathbf{W}_{si}\mathbf{y}_s = \mathbf{w}_{si}'^o\mathbf{y}_s, \tag{2.9}$$

que está definido si $\mathbf{X}_{si}'\mathbf{W}_{si}\mathbf{X}_{si}$ es invertible. Si sustituimos los $\hat{\mu}_i^o$ en (2.7) obtenemos el estimador de regresión polinómico local para el total poblacional

$$\tilde{t}_y^o = \sum_{i\epsilon s}\frac{y_i - \hat{\mu}_i^o}{\pi_i} + \sum_{i\epsilon U}\hat{\mu}_i^o. \tag{2.10}$$

El estimador muestral (dado en (2.9)) difiere significativamente del tradicional estimador de regresión polinómico local. La presencia de las probabilidades de inclusión en los $\mathbf{w}_{si}^o$ hacen que el estimador $\hat{\mu}_i$ basado en la muestra sea un estimador consistente bajo el diseño, que se basa en un parámetro ancho de banda h fijo, no necesariamente óptimo. En principio, el estimador (2.9) puede no estar definido para cierto $i\epsilon U$: si para alguna muestra s, hay menos de $q+1$ observaciones en el dominio de definición del núcleo para un x_i, entonces la matriz $\mathbf{X}_{si}^{'}\mathbf{W}_{si}\mathbf{X}_{si}$ sería singular. Esto no es un problema en la práctica, ya que se puede seleccionar un ancho de banda lo suficientemente grande para que $\mathbf{X}_{si}^{'}\mathbf{W}_{si}\mathbf{X}_{si}$ sea invertible para todos los x_i. Pero hay ocasiones en las que esta situación no puede excluirse teóricamente siempre y cuando se considera un ancho de banda fijo para una determinada población. A continuación vamos a considerar un estimador muestral que existe para cualquier muestra $s\epsilon U$. El estimador muestral para μ_i viene dado por

$$\hat{\mu}_i = \mathbf{e}_1'\left(\mathbf{X}_{si}'\mathbf{W}_{si}\mathbf{X}_{si} + diag\left\{\frac{\delta}{N^2}\right\}_{j=1}^{q+1}\right)^{-1}\mathbf{X}_{si}'\mathbf{W}_{si}\mathbf{y}_s = \mathbf{w}_{si}'\mathbf{y}_s \tag{2.11}$$

para $\delta > 0$. El término δN^{-2} en el denominador garantiza que el estimador esté bien definido para todo $s \subset U$. Otro posible ajuste consistiría en reemplazar la elección habitual del núcleo (como puede ser un núcleo gaussiano). En la práctica esta elección sólo aumenta la complejidad computacional del estimador lineal. Finalmente el estimador de regresión polinómico local para (2.11) está dado por

$$\tilde{t}_y = \sum_{i \epsilon s} \frac{y_i - \hat{\mu}_i}{\pi_i} + \sum_{i \epsilon U} \hat{\mu}_i. \tag{2.12}$$

Breidt and Opsomer (2000) estudiaron el diseño y las propiedades teóricas del modelo del estimador polinómico local, demostrando que este estimador es consistente bajo el diseño y asintóticamente insesgado bajo un conjunto de condiciones de regularidad. Asintóticamente, el error cuadrático medio bajo el diseño de $\widehat{t}_{MA}$ es equivalente a la varianza del estimador de diferencia generalizado,

$$\mathrm{MSE}_p\left(\widehat{t}_{MA}\right) = \mathrm{E}_p\left(\widehat{t}_{MA} - t_y\right)^2 \approx \sum_{i,j \in U} \left(y_i - \mu_i\right)\left(y_j - \mu_j\right) \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}. \tag{2.13}$$

Un estimador de diseño asintóticamente insesgado y consistente para $\mathrm{MSE}_p\left(\widehat{t}_{MA}\right)$ está dado por

$$\widehat{V}\left(\widehat{t}_{MA}\right) = \sum_{i,j \in s} \left(y_i - \widehat{\mu}_i\right)\left(y_j - \widehat{\mu}_j\right) \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{1}{\pi_{ij}}. \tag{2.14}$$

Dado que cada $\widehat{\mu}_i$ es una combinación lineal de los y_i de la muestra, el estimador modelo-asistido (2.5), puede también ser escrito en la misma forma, esto es, $\widehat{t}_{MA} = \sum_s \omega_{is} y_i$.

En el estudio de simulación de Breidt y Opsomer (2000), el estimador lineal local tipo núcleo es competitivo con el estimador clásico de regresión, cuando la función de regresión subyacente es lineal, pero tiene un rendimiento superior a dicho estimador, cuando la función de regresión subyacente no lo es. El estimador presenta además un rendimiento superior a otros estimadores paramétricos y no paramétricos, tanto para el modelo-asistido como para los basados en el modelo. El estimador de Breidt y Opsomer (2000) alcanza una eficiencia muy superior al estimador de Horvitz-Thompson, y al estimador de regresión cúbica y al de post-estratificación.

Aunque el rendimiento de un estimador no paramétrico depende de la selección del parámetro ancho de la banda, los resultados de los estudios de simulación fueron bastante insensibles a esta elección.

2.3. Regresión por splines en poblaciones finitas

Estudiamos a continuación la estimación no paramétrica de la función de regresión por splines penalizados en poblaciones finitas. Para el modelo de superpoblación (2.3), suponemos que la función $\mu(\cdot)$ puede ser escrita como (1.5). Se define

$$\mathbf{x}_i^T = \left(1, x_i, \ldots, x_i^q, (x_i - \kappa_1)_+^q, \ldots, (x_i - \kappa_j)_+^q\right),$$

$\mathbf{X_s} = \left[\mathbf{x_i^T}\right]_{\mathbf{i}\in\mathbf{s}}$, $\mathbf{X_U} = \left[\mathbf{x_i^T}\right]_{\mathbf{i}\in\mathbf{U}}$, y $\prod_s = \text{diag}\{\pi_i\}_{i\in s}$. Además, se define la matriz diagonal $\mathbf{A}_\lambda = \text{diag}\{0, \ldots, 0, \lambda, \ldots, \lambda\}$, con $q+1$ ceros sobre la diagonal principal y J constantes de penalización λ.

El estimador spline muestral de $\mu(x_i)$, solución del problema de mínimos cuadrados penalizados (1.6), está dado por

$$\tilde{\mu}(x_i) = \mathbf{x}_i^T \left(\mathbf{X}_s^T \mathbf{X}_s + \mathbf{A}_\lambda\right)^{-1} \mathbf{X}_s^T \mathbf{y}_s. \tag{2.15}$$

Empleando $x_i = \pi_i$ en (2.12) e insertándolo en (2.4), Zheng y Little (2003) han propuesto un estimador basado en el modelo en el contexto del muestreo en poblaciones finitas que emplea los splines penalizados. En Zheng y Little (2004) se extiende dicho estimador al muestreo en dos etapas.

Para introducir un estimador modelo-asistido de regresión en poblaciones finitas, se define

$$\mu_i = \mathbf{x}_i^T \left(\mathbf{X}_U^T \mathbf{X}_U + \mathbf{A}_\lambda\right)^{-1} \mathbf{X}_U^T \mathbf{y}_U,$$

y se estima por su versión muestral, mediante

$$\widehat{\mu}_i = \mathbf{x}_i^T \left(\mathbf{X}_s^T \Pi_s^{-1} \mathbf{X}_s + \mathbf{A}_\lambda\right)^{-1} \mathbf{X}_s^T \Pi_s^{-1} \mathbf{y}_s.$$

Se inserta esta estimación en (2.5) y se obtiene el estimador de regresión spline penalizado modelo-asistido propuesto por Breidt et al. (2005).

Este estimador tiene propiedades teóricas similares a las del estimador linal local tipo núcleo, como la consistencia y la insesgadez asintótica bajo el diseño (bajo condiciones suaves). En estudios de simulación incluidos en Breidt et al. (2005), se muestra que este estimador es en general, muy similar al estimador de regresión polinómica local. Sin embargo, la regresión spline penalizada presenta algunas ventajas sobre los métodos tipo núcleo que los convierte en un método atractivo en el contexto de las poblaciones finitas. Estos métodos permiten la incorporación de múltiples covariables, combinaciones de variables categóricas, así como términos paramétricos y no paramétricos, como se muestra en Aerts et al. (2002). Otra ventaja importante es que estos estimadores pueden calcularse con relativa sencillez, incluso considerando conjuntos grandes de datos o conjuntos de datos en regiones con escasez de

éstos. Por último, estos métodos resultan más fáciles de implementar en el contexto general del muestreo en poblaciones finitas.

Otra clase de estimadores no paramétricos modelo-asistido, basados en splines, ha sido estudiada por Goga (2004,2005). Goga utiliza la regresión por splines no penalizada donde el dominio de la variable auxiliar se divide en un número de nodos, una función base B-spline se asocia a cada nodo, y el número de éstos tiende a infinito, de forma que los B-splines son densos sobre todo el dominio. Goga (2005) muestra que el estimador de regresión por spline es asintóticamente insesgado bajo el diseño y consistente. Propone una aproximación de la varianza basada en el diseño y muestra que la varianza esperada es asintóticamente equivalente al límite inferior de Godambe-Joshi. El estudio de simulación muestra que el estimador de regresión por splines tiene buenas propiedades. Goga (2004) aplica esta metodología para la construcción de estimadores modelo-asistidos en el caso de muestreo en dos etapas, con toda la información auxiliar disponible en cada una de estas etapas.

2.4. Otros métodos de suavizamiento en poblaciones finitas

Breidt et al. (2007) extienden el estimador polinómico local al modelo semiparamétrico (1.8) y muestran que el estimador modelo-asistido semiparamétrico es consistente bajo el diseño y asintóticamente normal.

Montanari y Ranalli (2005) introducen las redes neuronales como una técnica de suavizamiento multivariante para métodos de estimación modelo-asistido. Un problema con ambos métodos es que éstos no conducen a estimadores calibrados para totales poblacionales de variables auxiliares. Además, el estimador resultante no puede escribirse como suma ponderada de los y_i. Tanto Montanari como Ranalli (2005) y Opsomer et al. (2008) aplican métodos de calibración del modelo, al originalmente propuesto por Wu y Sitter (2001) como una manera de obtener expresiones calibradas para sus estimadores.

Sea $\widehat{\mu}_i$, que denota el ajuste obtenido por cualquier ajuste por red neuronal. El método de calibración emplea la misma expresión del estimador modelo-asistido $\widehat{t}_{MA}$ en (14), pero reemplaza los $\widehat{\mu}_i$ por $\widehat{\mu}_i^*=\widehat{\mu}_i\widehat{\beta}$, con $\widehat{\beta}$ el coeficiente estimado de regresión de los y_i sobre $\widehat{\mu}_i$, empleando mínimos cuadrados ponderados bajo el diseño. El estimador resultante está calibrado para $\sum_U \widehat{\mu}_i$.

2.5. Selección del parámetro de suavizado

Los métodos de regresión no paramétrica requieren la especificación de uno o varios parámetros de suavizado, como el ancho de banda en la regresión tipo núcleo

o el parámetro de penalización en la regresión por splines. Este parámetro tiene una gran importancia para el rendimiento del estimador no paramétrico.

Es deseable contar con un método de selección del parámetro de suavizado que proporcione parámetros óptimos en el sentido de alcanzar un equilibrio entre la varianza y el sesgo del estimador. Esto es, encontrar un valor del parámetro que haga posible un ajuste razonable a los datos sin que aumente excesivamente la variabilidad del estimador.

Opsomer y Miller (2005) proponen un método de selección del ancho de banda para el estimador lineal local tipo núcleo que estima el ancho de banda h haciendo mínimo el error cuadrático medio del diseño (2.10). Si se hace mínimo el estimador tradicional de este error, la $\widehat{V}\left(\widehat{t}_{MA}\right)$ en (2.11), tiende a seleccionar anchos de banda que son demasiado pequeños, por lo que proponen un estimador basado en el método clásico de selección del ancho de banda por validación cruzada. El estimador es el valor de h que hace mínima la función

$$\mathrm{CV}\left(h\right) = \sum_{i,j \in s} \left(y_i - \widehat{\mu}_i^{(-)}\right)\left(y_j - \widehat{\mu}_j^{(-)}\right) \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{1}{\pi_{ij}},$$

sobre una red de anchos de banda. $\widehat{\mu}_i^{(-)}$ es el estimador (2.9) donde se omite la observación i-ésima de la muestra.
El cálculo de $\mathrm{CV}\left(h\right)$ puede simplificarse dado que $\widehat{\mu}_i^{(-)}$ puede ser escrita como una función de $\widehat{\mu}_i$, por lo que sólo es necesario ajustar el estimador lineal local tipo núcleo una vez para cada valor de h. En Opsomer y Miller (2005) se muestran las buenas propiedades teóricas de este método, así como su buena aplicabilidad en diferentes situaciones y problemas.

Capítulo 3

Métodos no paramétricos en estimación de áreas pequeñas

Consideramos la estimación en áreas pequeñas, como aplicación final de los métodos no paramétricos en el contexto de las encuestas por muestreo. Cowling, et al. (1996) presentan dos aplicaciones, en las que se emplea la regresión no paramétrica en el contexto de la Estadística espacial. En dos trabajos recientes, los métodos no paramétricos han sido considerados en el contexto donde los métodos clásicos de estimación de áreas pequeñas eran aplicados tradicionalmente. Mukhopadhay y Maiti (2004) proponen una extensión del modelo de área-nivel en la que la función lineal, se reemplaza por una función no paramétrica estimada mediante la regresión tipo núcleo. Opsomer et al. (2008) consideran el uso de regresión spline penalizado en este contexto.

Supongamos que la población contiene T áreas pequeñas de interés. El modelo de nivel-área no paramétrico estudiado por Mukhopadhay y Maiti (2004) está dado por

$$y_t = m\left(x_t\right) + u_t + \varepsilon_t, \tag{3.1}$$

donde u_t y ε_t están independientemente distribuidos bajo una distribución $\mathcal{N}\left(0, \sigma_u^2\right)$ y $\mathcal{N}\left(0, D_t\right)$ con D_t conocida, respectivamente. Si $m\left(\cdot\right)$ es una función lineal, este modelo se conoce como el modelo Fay-Herriot (Fay and Harriot, 1979). El objetivo de los métodos de estimación en áreas pequeñas es predecir $\tilde{y}_t = m\left(x_t\right) + u_t$. El procedimiento de predicción se inicia mediante la estimación $m\left(\cdot\right)$ a través del estimador de regresión polinómica local constante $(q = 0)$ (1.3), de modo que la matriz $\mathbf{X}_x$ en (1.4) es sustituida por un vector de unos (como se hizo en la sección 1.4). La varianza en áreas pequeña σ_u^2 se estima por $\widehat{\sigma}_u^2 = \sum_{t=1}^{T} \left\{\left(y_t - \widehat{m}\left(x_t\right)\right)^2 - D_t\right\} / T$, y

la estimación de $\tilde{y}_t$ se define por

$$\widehat{y}_t = \widehat{\gamma}_t y_t + (1 - \widehat{\gamma}_t)\, \widehat{m}\,(x_t)\,, \tag{3.2}$$

con $\widehat{\gamma}_t = \widehat{\sigma}_u^2 / (\widehat{\sigma}_u^2 + D_t)$. Mukhopadhay y Maiti (2004) obtienen una aproximación asintótica del error cuadrático medio de la predicción $\widehat{y}_t$, $\mathrm{E}\,(\widehat{y}_t - \tilde{y}_t)^2$, y un estimador *plug-in* para dicha cantidad.

La regresión spline penalizada (ver Wand, 2003), proporciona un enfoque práctico apropiado para la aplicación de modelos no paramétricos en la estimación de áreas pequeñas. Opsomer et al. (2008) extienden el enfoque del modelo lineal mixto elemento-nivel en la estimación en áreas pequeñas descrito en Battese et al. (1988) al contexto en el que la función de regresión puede ser estimada con métodos no paramétricos o semiparamétricos.

El modelo está dado por

$$y_i = m\,(x_i) + \mathbf{d}_i^T \mathbf{u} + \varepsilon_i, \tag{3.3}$$

donde $\mathbf{d_i} = (d_{1i}, \ldots, d_{Ti})^T$ es un vector de indicadores con $d_{ti} = 1$ si el elemento i está en el área pequeña t y cero en caso contrario, $\mathbf{u} = (u_1, \ldots, u_T)^T$ es un vector de efectos mutuamente independientes en áreas pequeñas con media 0 y varianza σ_u^2, y ε_i es el error aleatorio con media 0 y varianza σ_ε^2, independiente de $\mathbf{u}$.

La función $m\,(\cdot)$ se expresa como una función spline en (1.5). Tomando en consideración el trabajo de Wand (2003), reescribimos esto como $m\,(x_i) = \mathbf{x}_i^T \boldsymbol{\beta} + z_i^T \gamma$, con $\mathbf{x}_i = (1, x_i, \ldots, x_i^p)^T$,
$z_i = \left((x_i - \kappa_1)_+^p, \cdots, (x_i - \kappa_J)_+^p\right)^T$, $\boldsymbol{\beta}$ un vector de parámetros desconocidos, y $\boldsymbol{\gamma}$ un vector de variables aleatorias independientes con media 0 y varianza σ_γ^2. El modelo completo puede escribirse

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + z_i^T \boldsymbol{\gamma} + \mathbf{d}_i^T \mathbf{u} + \varepsilon_i. \tag{3.4}$$

El término $z_i^T \boldsymbol{\gamma}$ es una desviación aleatoria de la tendencia lineal poblacional y $\mathbf{d}_i^T \mathbf{u}$ es el efecto aleatorio para un área pequeña i. El objetivo de la estimación en áreas pequeñas es ahora la predicción de $\tilde{y}_t = \overline{\mathbf{x}}_t^T \boldsymbol{\beta} + \overline{z}_t^T \boldsymbol{\gamma} + u_t$, donde suponemos que $\overline{\mathbf{x}}_t^T$ y $\overline{z}_t^T$ son conocidas.
Opsomer et al. (2008) proponen la estimación de máxima verosimilitud empírica restringida (MVER) para estimar los parámetros $\boldsymbol{\beta}$, σ_γ^2, σ_u^2, σ_ε^2, y predecir $\tilde{y}_t$ mediante

$$\widehat{y}_t = \overline{\mathbf{x}}_t^T \widehat{\boldsymbol{\beta}} + \overline{z}_t^T \widehat{\boldsymbol{\gamma}} + \widehat{u}_t, \tag{3.5}$$

con

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\mathbf{Y} \\
\widehat{\boldsymbol{\gamma}} &= \widehat{\sigma}^2_\gamma\mathbf{Z}^T\widehat{\mathbf{V}}^{-1}\left(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right) \\
\widehat{\mathbf{u}} &= \widehat{\sigma}^2_\gamma\mathbf{D}^T\widehat{\mathbf{V}}^{-1}\left(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right),
\end{aligned}$$

donde $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ y $\mathbf{Y}$, $\mathbf{Z}$ y $\mathbf{D}$ se definen análogamente. Además, $\widehat{\mathbf{V}}$ es la matriz de varianzas-covarianzas estimada de $\mathbf{Y}$ obtenida insertando las estimaciones MVER de los parámetros de la varianza en la matriz de varianzas-covarianzas de $\mathbf{Y}$.

La aproximación asintótica del error cuadrático medio de la predicción de $\widehat{y}_t$ se demuestra en Opsomer et al. (2008), que generaliza directamente a la obtenida en ausencia del efecto aleatorio del spline.

Bibliografía

[1] Aerts, M., Claeskens, G., Wand, M. (2002). Some theory for penalized spline generalized additive models. *Journal of Statistical Planning and Inference 103, 455–470*

[2] Battese, G.E., Harter, R.M., Fuller, W.A. (1988). An error components model for prediction of county crop area using survey and satellite data. *Journal of the American Statistical Association 83, 28–36.*

[3] C. de Boor (2001), A Practical Guide to Splines (Revised Edition), Springer-Verlag, New York. 33

[4] Breidt, F.J., Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling.

[5] Breidt, F.J., Opsomer, J.D. (2009). Nonparametric and semiparametric estimation in complex surveys, in Handbook of Statistics - Sample Surveys: Inference and Analysis, Vol. 29B, D. Pfeffermann and C.R. Rao (editors), The Netherlands: North-Holland, 103-120.

[5] Breiman, L., Friedman, J.H., Olshen, R., Stone, C.J. (1984). *Classification and Regression Trees.*Wadsworth, Belmont, CA.

[6] Chambers, R.L., Dorfman, A.H., Wehrly, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association* ***88*** *(421), 268–277.*

[7] Cowling, A., Chambers, R., Lindsay, R., Parameswaran, B. (1996). Applications of spatial smoothing to survey data. *Survey Methodology* ***22****, 175–183.*

[8] Dorfman, A.H. (1992). Non-parametric regression for estimating totals in finite populations. *ASA Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA, pp. 622–625.

[9] Fan, J. & Gijbels, I. (1996). Local Polynomial Modelling and Its Applications, Chapman and Hall, London. .

[10] Friedman, J.H. (1991). Multivariate adaptive regression splines (with discussion). *The Annals of Statistics* **19**, 1–141.

[11] Goga, C. (2004). Estimation de l'évolution d'un total en présence d'information auxiliaire: une approche par splines de régression. *Comptes Rendus de l'Académie des Sciences Paris Ser.* I ***339***, 441–444.

[12.] Goga, C. (2005). Réduction de la variance dans les sondages en présence d'information auxiliaire: une approche non paramétrique par splines de régression. *The Canadian Journal of Statistics* ***33***, 163–180.

[13] Hastie, T.J., Tibshirani, R.J. (1990). *Generalized Additive Models.* Chapman and Hall,Washington, DC.

[14] Kuo, L. (1988). Classical and prediction approaches to estimating distribution functions from survey data. *ASA Proceedings of the Section on Survey Research Methods*,American Statistical Association, Alexandria, VA, 280–285.

[15] Montanari, G.E., Ranalli, M.G. (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association* ***100***(472), 1429–1442.

[16] Mukhopadhay, P., Maiti, T. (2004). Two-stage nonparametric approach for small area estimation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA, 4058–4065.

[17] Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G., Breidt, F.J. (2008). Nonparametric small area estimation using penalised spline regression. *Journal of the Royal Statistical Society, Series* B, **70**, 265–286.

[18] Opsomer, J.D., Miller, C.P. (2005). Selecting the amount of smoothing in nonparametric regression estimation for complex surveys. *Journal of Nonparametric Statistics* **17**, 593–611.

[19] Pfeffermann, D., Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya, Series B* 61(Pt. 1): 166–186.

[20] Ruppert, D., Wand, M.P., Carroll, R.J. (2003). *Semiparametric Regression.* Cambridge University Press, Cambridge, UK.

[21] Vidakovi'c, B. (1999). *Statistical Modeling by Wavelets*. JohnWiley & Sons, Inc, New York.

[22] Wand, M.P. (2003). Smoothing and mixed models. *Computational Statistics* 18, 223–249.

[23] Wu, C., Sitter, R.R. (2001a). A model-calibration approach to using complete auxiliary information from survey data. J*ournal of the American Statistical Association* **96**(453), 185–193.

[24] Zheng, H., Little, R.J.A. (2003). Penalized spline model-based estimation of finite population total from probability-proportional-to-size samples. *Journal of Official Statistics* **19**, 99–117.