ESTIMADORES ROBUSTOS EN MODELOS DE REGRESIÓN SEMI-FUNCIONALES PARCIALMENTE LINEALES

Alejandra Vahnovan[†] y Graciela Boente[‡]

[†] *Universidad Nacional de La Plata, ANPCyT, Argentina, avahnovan@mate.unlp.edu.ar*
[‡] *Universidad de Buenos Aires, CONICET, Argentina, gboente@dm.uba.ar*

Resumen: En este trabajo proponemos estimadores robustos para modelos de regresión parcialmente lineales semi-funcionales. Presentamos además, resultados de convergencia de estos estimadores y evaluamos el comportamiento de las propuestas robustas y clásicas mediante un estudio de simulación.

Palabras clave: *robustez, métodos semiparamétricos, datos funcionales*
2000 AMS Subject Classification: 62G08 - 62G35

1. INTRODUCCIÓN

La mayoría de los procedimientos estadísticos clásicos están basados en modelos con hipótesis rígidas, tales como errores normales, observaciones equidistribuidas, etc. Bajo estas hipótesis se deducen procedimientos óptimos. Por ejemplo, para el caso de regresión el procedimiento óptimo es el de mínimos cuadrados; para modelos paramétricos en general, los procedimientos óptimos clásicos son los estimadores de máxima verosimilitud. Sin embargo, estos métodos son muy sensibles al incumplimiento de las hipótesis que los generaron, tales como la presencia en la muestra de observaciones atípicas. Los procedimientos estadísticos robustos tienen como objetivo permitir inferencias válidas cuando el modelo no se cumple exactamente y al mismo tiempo ser altamente eficientes bajo el modelo. Por otra parte, los modelos clásicamente usados son paramétricos y la suposición es que la muestra de observaciones proviene de una familia paramétrica conocida. 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), además 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 además de procedimientos estadísticos robustos, métodos noparamétricos para analizar los datos finito-dimensionales. En muchas situaciones uno se enfrenta con datos que, o bien son, o bien parecen provenir de un proceso suave. En tales situaciones, discretizar el proceso para estudiar las componentes de variación mediante las técnicas usuales de componentes principales, regresión paramétrica o noparamétrica no parece ser lo más indicado. Por otra parte, las técnicas de Análisis Multivariado no darán información sobre las derivadas o las integrales de las funciones que se suponen yacen bajo los datos. Es por eso, que se han desarrollado en los últimos años diversas técnicas de análisis para datos funcionales entre las que podemos mencionar el análisis de componentes principales funcional y la regresión paramétrica funcional, entre otros. Como es bien sabido, en muchos casos la relación entre la variable respuesta y la explicativa no se conoce exactamente y en ese caso, los modelos noparamétricos ofrecen una opción más flexible tanto en el caso finito-dimensional como en el caso en que las variables explicativas provienen de un proceso suave. En [1] consideraron el problema de predecir una variable aleatoria real a partir de variables explicativas funcionales mediante una aproximación noparamétrica basada en núcleos. Para una descripción de distintos procedimientos para datos funcionales ver [2].

El objetivo de los modelos semiparamétricos parcialmente lineales es permitir a algunas de las variables explicativas actuar de manera libre (noparamétrica), mientras que otras están controladas por medio de una relación paramétrica (lineal). Hay numerosos trabajos en los que se considera el caso en el que las variables explicativas toman valores reales, pero recientemente se ha incrementado el interés en muchos campos en los que las observaciones estadísticas son curvas, por lo que los datos funcionales se han convertido en el objeto de atención de muchas investigaciones. Aquí, se combinará la flexibilidad de un modelado parcialmente lineal junto con la reciente metodología para el tratamiento noparamétrico de datos funcionales.

2. EL MODELO

El Modelo Parcialmente Lineal Semi-Funcional (PLSF) puede definirse como

$$Y = Z'\beta + g(X) + \epsilon \tag{1}$$

donde Y es una variable aleatoria (v.a.), $Z \in \mathbb{R}^p$, X es una v.a. funcional, es decir toma valores en un espacio infinito-dimensional $\mathcal{H}$, β es un parámetro desconocido y $g : \mathcal{H} \to \mathbb{R}$ es un operador suave que no se supone lineal. En este trabajo, consideramos a $\mathcal{H}$ un espacio semimétrico y denotamos d a la semimétrica asociada. Suponemos tener n observaciones (Y_i, Z_i, X_i), $i = 1, \ldots, n$, independientes e idénticamente distribuídas (i.i.d.), que satisface (1). El problema estadístico consiste en estimar el operador g y el parámetro multivariado β . Usualmente, se supone que $E(\epsilon_i|X_i) = 0$, $E(\epsilon_i^2|X_i = X) = \sigma^2(X)$.

Denotemos por $\phi_0(x) = E(Y|X = x)$ y $\phi(x) = (\phi_1(x), ..., \phi_p(x))'$ donde $\phi_j(x) = E(Z_j|X = x)$. Entonces, tenemos que $g(x) = \phi_0(x) - \beta'\phi(x)$ y luego $Y - \phi_0(x) = \beta'(Z - \phi(x)) + \epsilon$. Esto sugiere que si proponemos a priori estimadores de $\phi_0(x)$ y $\phi(x)$, que denotaremos $\widehat{\phi}_0(x)$ y $\widehat{\phi}(x)$ respectivamente, podemos obtener un estimador para β y, finalmente, un estimador para g.

3. ESTIMADORES CLÁSICOS

El enfoque noparamétrico clásico estima las esperanzas condicionales con

$$\widehat{\phi}_0(x) = \sum_{i=1}^{n} w_{n,h}(x, X_i)Y_i \qquad \widehat{\phi}_j(x) = \sum_{i=1}^{n} w_{n,h}(x, X_i)Z_{ij},$$

donde $w_{n,h}(x, X_i) = K(d(x, X_i)/h) \Big/ \sum_{j=1}^{n} K(d(x, X_j)/h)$, K es una función núcleo y h una sucesión de números reales estrictamente positiva. Luego,

$$\begin{cases} \widehat{\beta}_h = (\widetilde{Z}_h'\widetilde{Z}_h)^{-1}\widetilde{Z}_h'\widetilde{Y}_h \\ \widehat{g}_h(x) = \sum_{i=1}^{n} w_{n,h}(x, X_i)(Y_i - Z_i'\widehat{\beta}_h), \end{cases}$$

donde $\widetilde{Y}_h = (I_n - W_h)Y$, $\widetilde{Z}_h = (I_n - W_h)Z$, I_n la matriz identidad de $n \times n$ y W_h la matriz de pesos $(W_h)_{i,j} = w_{n,h}(X_i, X_j)$.

Los pesos utilizados en los estimadores son la versión funcional de los pesos de Nadaraya-Watson. El estimador de $g(x)$ está basado en un promedio de las variables respuesta por lo que es muy sensible a observaciones atípicas, particularmente a aquellas que se encuentran en el entorno del punto X. Por otra parte, para la estimación de β se considera un estimador de mínimos cuadrados sobre los residuos con lo que dicho estimador también será sensible a la presencia de datos anómalos.

4. ESTIMADORES ROBUSTOS

Los métodos estadísticos robustos tienen como objetivo permitir inferencias válidas cuando el modelo no se cumple exactamente. Teniendo en cuenta la sensibilidad de los estimadores de (β, g) definidos en [3], introduciremos estimadores robustos para el modelo PLSF extendiendo una propuesta en tres pasos dada para el caso finito-dimensional en [4] al caso funcional utilizando el M-estimador local funcional definido en [5].

4.1. PROCEDIMIENTO EN TRES PASOS

Paso 1: Estimamos $\phi_0(x)$ y $\phi_j(x)$ con un suavizado robusto, como las medianas locales o M-estimadores locales. Denotamos con $\widehat{\phi}_0(x)$ y $\widehat{\phi}_j(x)$ a los estimadores obtenidos y $\widehat{\phi}(x) = (\widehat{\phi}_1(x), \ldots, \widehat{\phi}_p(x))'$.
Paso 2: El estimador de β, $\widehat{\beta}$, se define utilizando cualquier estimador de regresión robusto sobre los residuos $Y_i - \widehat{\phi}_0(X_i)$ y $Z_i - \widehat{\phi}(X_i)$.

Paso 3: El estimador de la función de regresión g se define como $\widehat{g}(x) = \widehat{\phi}_0(x) - \widehat{\beta}'\widehat{\phi}(x)$.

En el Paso 1, calculamos las medianas locales $\widehat{\phi}_{0,med}(x)$ y $\widehat{\phi}_{j,med}(x)$ como la mediana de las funciones de distribución condicionales empíricas $\widehat{F}_0(y|X=x)$ y $\widehat{F}_j(z|X=x)$, que están definidas como

$$\widehat{F}_0(y|X=x) = \sum_{i=1}^{n} w_{n,h}(x, X_i) I_{(-\infty,y]}(y_i),$$

$$\widehat{F}_j(z|X=x) = \sum_{i=1}^{n} w_{n,h}(x, X_i) I_{(-\infty,z]}(z_{ij}),\ 1 \leq j \leq p.$$

Por otro lado, los M-estimadores locales de la regresión de Y versus X y de cada componente de Z versus X, $\widehat{\phi}_{0,M}(x)$ y $\widehat{\phi}_{j,M}(x)$ respectivamente, se basan en los definidos en [5]. Es decir, están definidos implícitamente como la solución única de

$$\sum_{i=1}^{n} w_{n,h}(x, X_i)\psi\left(\frac{Y_i - \widehat{\phi}_{0,M}(x)}{\widehat{s}_0(x)}\right) = 0 \quad \text{y} \quad \sum_{i=1}^{n} w_{n,h}(x, X_i)\psi\left(\frac{Z_{ij} - \widehat{\phi}_{j,M}(x)}{\widehat{s}_j(x)}\right) = 0,$$

donde en cada caso ψ es una función a valores reales que satisface algunas condiciones de regularidad y $\widehat{s}_j$ son estimadores locales de escala robustos. Posibles elecciones para la función ψ son la función de Huber o la bicuadrada, mientras que las escalas $\widehat{s}_0(x)$ y $\widehat{s}_j(x)$ pueden tomarse como la MAD local correspondiente a $\widehat{F}_0(y|X=x)$ y $\widehat{F}_j(z|X=x)$ respectivamente.

4.2. CONSISTENCIA

Bajo ciertas condiciones de regularidad y algunas restricciones sobre el tamaño de $S_{\mathcal{H}}$ que involucran la $\epsilon-$entropía de Kolmogorov, entre otros, se tiene que, para cualquier conjunto compacto $S_{\mathcal{H}} \subset \mathcal{H}$

- $\sup_{x \in S_{\mathcal{H}}} |\widehat{\phi}_j(x) - \phi_j(x)| \xrightarrow{as} 0$
- $\widehat{\beta} \xrightarrow{as} \beta$
- $\widehat{g}(x) = \widehat{\phi}_0(x) - \widehat{\beta}'\widehat{\phi}(x)$ es uniformemente consistente sobre compactos.

5. ESTUDIO DE MONTE CARLO

Realizamos un estudio de simulación cuando el parámetro de regresión tiene dimensión 2 y comparamos el comportamiento del estimador de mínimos cuadrados con los estimadores obtenidos suavizando en el Paso 1 con la mediana local y con un M-estimador local con función de score bicuadrada. Luego de suavizar la variable respuesta y las covariables de la regresión Z se calcularon, entre otros, estimadores de β de tipo M con función bicuadrada y de Huber.

Realizamos 1000 replicaciones generando muestras independientes de tamaño $n = 100$ siguiendo el modelo

$$Y_i = Z_{i1}\beta_1 + Z_{i2}\beta_2 + g(X_i) + \epsilon_i, \qquad 1 \leq i \leq n$$

con

- Z_{ij} y ϵ_i i.i.d. con distribución normal
- $X_i(z) = a_i(z - 0{,}5)^2 + b_i\,(z \in [0,1])$ con a_i y b_i i.i.d. con distribución uniforme
- $\beta = (-1, 3)'$
- $g(X_i) = \exp(-8f(X_i)) - \exp(-12f(X_i))$ donde $f(X_i) = sign(X_i'(1) - X_i'(0))\sqrt{3\int_0^1 (X_i'(z))^2 dz}$.

La suavidad de las curvas X_i nos permite considerar la semimétrica basada en la norma L_2

$$d(X, X^*) = \left(\int_0^1 (X'(z) - X^{*\prime}(z))^2 dz \right)^{1/2}.$$

Se compararon los resultados obtenidos con los estimadores para conjuntos de datos normales con los obtenidos bajo diferentes contaminaciones del modelo, obteniendo el desempeño esperado para los estimadores robustos propuestos.

REFERENCIAS

[1] FERRATY, F., MAS, A. Y VIEU, PH., *Nonparametric regression on functional data: inference and practical aspects*, Australian & New Zealand Journal of Statistics, Volume 49, issue 3 (2007), pp. 267-286..

[2] FERRATY, F. Y VIEU, PH. , *Nonparametric Functional Data Analysis: Theory and Practice. Springer Series in Statistcs*, Springer, New York, (2006).

[3] ANEIROS. G. Y VIEU, PH., *Semi-functional partial linear model*,Statist.& Prob. Lett., 11 (2006), pp. 1102-1110.

[4] BIANCO, A. Y BOENTE, G., *Robust estimators in semiparametric partly linear regression models*, J. Statist. Plann. Inference, 122 (2004), pp. 229-252.

[5] AZZEDINE, N., LAKSACI, A. Y OUD-SAID, E., *On the robust nonparametric regression estimation for functional regressors*, Statistics & Probability Letters, Volume 78, Issue 18, (2008), pp. 3216-3221.