Marí, Gonzalo
Cuesta, Cristina
Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística

INTERVALOS BOOTSTRAP PARA REGRESIONES SPLINE PENALIZADAS BAJO EL ENFOQUE DE MODELOS MIXTOS

Resumen

Durante la última década, las técnicas estadísticas asociadas a los suavizados y regresiones semi-paramétricas han sido objeto de grandes avances teóricos y de gran divulgación en la práctica profesional. En particular, las regresiones spline penalizadas (P-spline) consideradas bajo el enfoque de los modelos mixtos constituyen una herramienta muy simple pero efectiva y gozan de la ventaja del sustento teórico de los modelos mixtos al momento de hacer inferencias. Sin embargo, los intervalos de confianza que surgen de él a menudo son puestos en discusión y como alternativa se puede recurrir a la construcción de intervalos de incertidumbre generados por procedimientos bootstrap. En este trabajo se presenta la construcción de diferentes tipos de intervalos bootstrap teniendo en cuenta un modelo de efectos fijos y uno de efectos aleatorios. En particular se presentan los intervalos bootstrap de tipo paramétrico, empírico y *Wild* tanto para el caso de p-spline bajo un modelo de efectos fijos como uno de efectos aleatorios. Se ejemplifica la construcción de estos intervalos con un conjunto de datos sobre tasa de mortalidad materna y tasa de nacimientos en adolescentes por país. Estas variables son estudiadas en el contexto de una investigación sobre causas de muertes maternas que actualmente está llevando a cabo la Organización Mundial de la Salud.

1. Metodología

Sea el modelo

$$y/x = \mu(x) + \varepsilon$$

donde ε representa el error aleatorio y $\mu(x)$ es una función suave pero que no esta preespecificada. A $\mu(x)$ la podemos descomponer en $x\beta + zu$, donde x es una matriz que representa un polinomio (que en este trabajo se considera lineal) y z una matriz de alta dimensión compuesto por funciones de bases tales como funciones truncadas, B-splines, funciones radiales, etc. En particular, la expresión de las funciones truncadas es $(x - N_k)_+$ de $x_+ = x$ para $x > 0$ y cero en otro caso, y N_k son los nodos de la función. De modo que la dimensión de z queda determinado por la cantidad de nodos empleados para el ajuste.

En definitiva el modelo propuesto es:

$$f(x_i) = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} \beta_{ik}(x_i - N_k)_+ + \varepsilon_i \quad (1)$$

donde $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$. La formulación de este modelo puede reescribirse como

$$y = x\beta + zu + \varepsilon = C\theta + \varepsilon \quad (2)$$

siendo $C = (x||z)$ y $\theta = (\beta'||u')'$.

Si se impone una penalidad a los coeficientes de u, se puede estimar θ minimizando el criterio

$$\|y - C\theta\|^2 - \lambda\theta' D\theta \quad (3)$$

En el caso de bases truncadas, es conveniente elegir *D* como la matriz identidad.El coeficiente lambda actúa como un parámetro de suavizado. La estimación de θa partir de (3) es entonces

$$\hat{\theta} = (C'C + \lambda D)^{-1} C'y \quad (4)$$

Asumiendo que u es aleatorio con $u \sim N(0, \sigma_u^2)$, las estimaciones que surgen de θ coinciden con las obtenidas en (4) bajo el mejor estimador predictor lineal insesgado (BLUEP) con $\hat{\lambda}^2 = \frac{\sigma_\varepsilon^2}{\sigma_u^2}$, donde $\hat{\lambda}$ se obtiene estimando las componentes de variancia por máxima verosimilitud o máxima verosimilitud restringida.

Si bien se pueden construir intervalos de confianza para $C\hat{\theta}$a partir del modelo (2) desde un enfoque clásico, el objetivo en este trabajo es construir intervalos bootstrap.

2. Intervalos de confianza para modelos de efectos fijos

2.1. Intervalos de confianza clásico

Bajo los supuestos planteados para el modelo (2) se puede definir un intervalo de confianza de la siguiente forma

$$C\hat{\theta} \pm \mathrm{z}_{1-\frac{\alpha}{2}} \widehat{\mathrm{st.dev}}(C\hat{\theta})$$

donde

$$\mathrm{st.dev}\left(C\hat{\theta}\right) = \sigma_\varepsilon \sqrt{\mathrm{C_x}(\mathrm{C'C} + \lambda \mathrm{D})^{-1}\mathrm{C'C}(\mathrm{C'C} + \lambda \mathrm{D})^{-1}\mathrm{C_x'}}$$

donde $\mathrm{D} = \mathrm{diag}(0_2, \mathrm{I_K})$

2.2. Intervalos de confianza bootstrap para el modelo de efectos fijos

El método Bootstrap es un método de replicación desarrollado por Efron (1979). Consiste en la reutilización de la muestra original, a partir de la cual se obtienen estimaciones de o de los parámetros de interés aplicando el mismo estimador a cada muestra bootstrap. A partir de estas estimaciones se pueden obtener estimaciones de variancia e intervalos de confianza.

Bajo el modelo (2) se presentan tres tipos de estimaciones bootstrap: paramétrico, empírico y Wild. A continuación se describen los pasos a seguir para obtener cada uno de estos intervalos.

2.2.1. Bootstrap paramétrico

a) Obtener una estimación de σ_ε^2
b) Calcular $y^* = C\hat{\theta} + \varepsilon^*$, donde ε^*se genera de una distribución N*(0,*$\hat{\sigma}_\varepsilon^2 I$*)*
c) Se obtiene $\hat{y}^*$ a partir del modelo (2)
d) Se repiten los pasos b) y c) B veces
e) Para cada valor de *x* obtener el percentil de 2,5% y del 97,5% de la distribución de $\hat{y}^*$

2.2.2. Bootstrap empírico

a) Estimar los residuales a partir del modelo (2)
b) Obtener $\varepsilon^* = \{\varepsilon_i^*\}_{\boldsymbol{i=1,\dots,n}}$, una muestra aleatoria con reemplazo (de tamaño n) de los residuales obtenidos en a)
c) Calcular $y^* = C\hat{\theta} + \varepsilon^*$, donde los ε^*son los obtenidos en el punto b)
d) Postular el modelo (2) con los pares (*x,* y^*) y obtener (*x,* $\hat{y}^*$)
e) Repetir los pasos b) hasta d) B veces
f) Para cada valor de *x* obtener el percentil de 2,5% y del 97,5% de la distribución de $\hat{y}^*$

2.2.3. Bootstrap Wild

a) Estimar los residuales a partir del modelo ….
b) Obtener $\varepsilon^* = \{\varepsilon_i^*\}_{i=1,\dots,n}$de una distribución de 2 puntos con masa $a_i = \hat{\varepsilon}_i(1-\sqrt{5})/2$ y $b_i = \hat{\varepsilon}_i(1+\sqrt{5})/2$ y probabilidad muestral $P(\varepsilon_i^* = a_i) = (5+\sqrt{5})/10$
c) Calcular $y^* = C\hat{\theta} + \varepsilon^*$, donde los ε^*son los obtenidos en el punto b)
d) Postular el modelo (2) con los pares (*x,* y^*) y obtener (*x,* $\hat{y}^*$)
e) Repetir los pasos b) hasta d) B veces
f) Para cada valor de *x* obtener el percentil de 2,5% y del 97,5% de la distribución de $\hat{y}^*$

3. Aplicación

Durante el año 2000, la Cumbre de la ONU se reunió para debatir sobre los objetivos de desarrollo del Milenio y presentó un plan de acción mundial para alcanzar ocho objetivos relacionados fundamentalmente con la lucha contra la pobreza, nuevos compromisos para la salud de las mujeres y los niños y otras iniciativas contra la pobreza, el hambre y la enfermedad. Se puso como fecha límite, de re-evaluación de cumplimiento de los objetivos, el año 2015.

Entre las principales acciones a tomar se priorizó la disminución de las muertes maternas. Es decir, se enfatizo la necesitad de disminuir los casos relacionados con muertes de mujeres durante el embarazo, parto y puerperio.

Para ello, desde la Organización Mundial de la Salud se arbitraron muchos esfuerzos dedicados a investigar cuales eran las causas de muertes más frecuentes en cada país con el objetivo de informarlas y tomar acciones gubernamentales tendientes a disminuirlas.

A fin de estimar las causas de muertes por país se construyeron modelos estadísticos donde algunas de las variables utilizadas como predictoras fueron:

1) variables macro económicas del país (producto bruto interno per cápita, porcentaje del gasto del gobierno dedicado a salud, etc.),

2) variables relacionadas con la situación de salud neonatal (tales como cobertura de cuidados pre-natales , etc.),

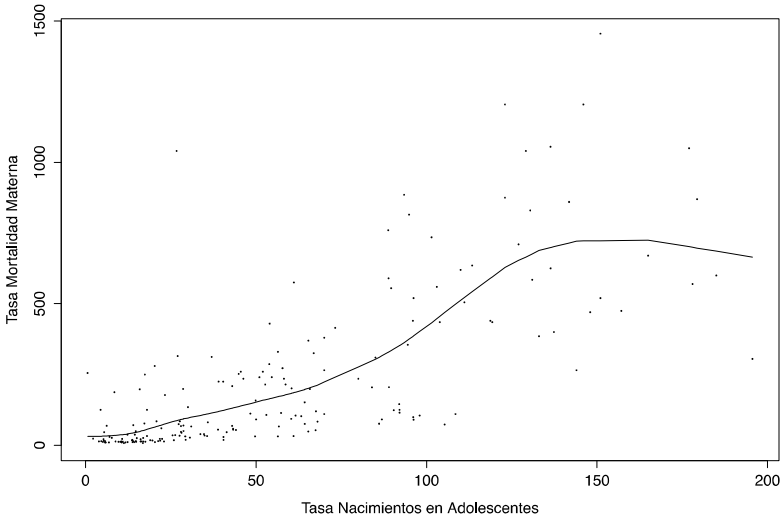
3) variables relacionadas con las características de la población (tales como tasa de fecundidad general, tasa de nacimientos en adolescentes, etc).

Asimismo, se categorizaron los países de acuerdo a Nivel de desarrollo (desarrollados, en vías de desarrollo) y de acuerdo a nivel económico (alto, medio-alto, medio-bajo y bajo).

Como primer paso se estudiaron las relaciones entre estas variables indicadoras. A fin de ilustrar la metodología desarrollada en este trabajo se muestra como ejemplo la relación entre la tasa de mortalidad materna y la tasa de nacimientos en adolescentes.

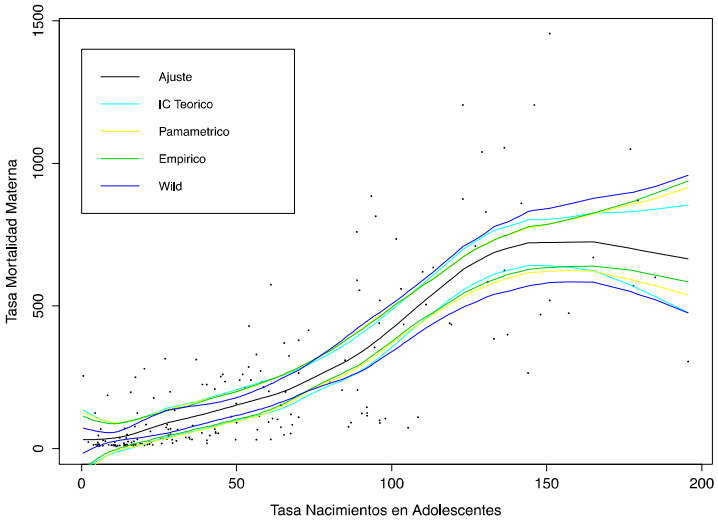
En la Figura 1 se muestra el diagrama de dispersión de los datos junto con el ajuste correspondiente a una regresión p-spline lineal penalizada con 36 nodos y λ igual a 3873,95

Figura 1. Regresión p-spline penalizada para las variables Tasa de Mortalidad Materna y Tasa de Nacimientos en Adolescentes de países



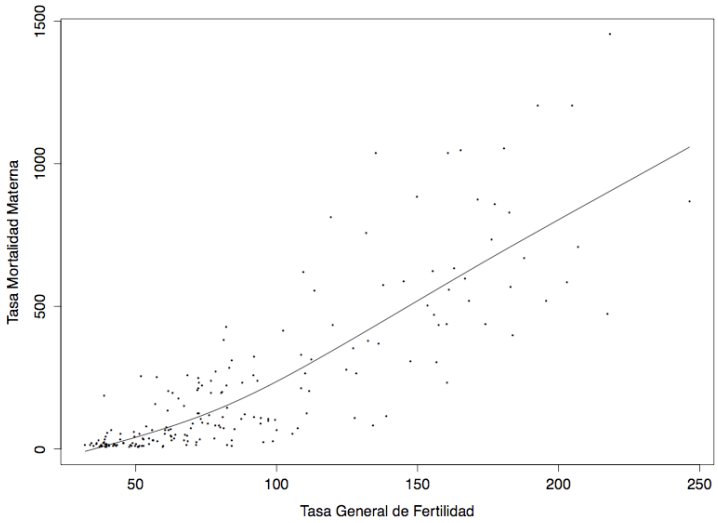
En la siguiente figura se muestran los intervalos de confianza construidos bajo el método clásico y bajos las tres opciones de intervalos Bootstrap

Figura 2. Intervalos de Confianza Clásico y Boostrap para la regresión p-spline penalizada para las variables Tasa de Mortalidad Materna y Tasa de Nacimientos en Adolescentes de países



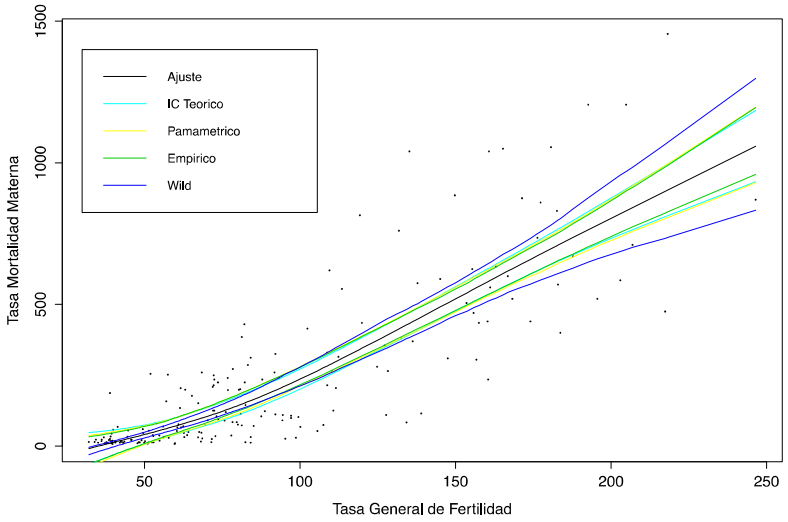
En la Figura 3 se muestra el diagrama de dispersión de los datos junto con el ajuste correspondiente a una regresión p-spline lineal penalizada con 37 nodos y λ igual a 75900.91 entre las variables Tasa Gerenal de Fertilidad y Tasa de Mortalidad Materna.

Figura 3. Regresión p-spline penalizada para las variables Tasa de Mortalidad Materna y Tasa General de Fertilidad



En la siguiente figura se muestran los intervalos de confianza construidos bajo el método clásico y bajos las tres opciones de intervalos Bootstrap

Figura 4. Intervalos de Confianza Clásico y Boostrap para la regresión p-spline penalizada para las variables Tasa de Mortalidad Materna y Tasa General de Fertilidad



4. Conclusión

En esta aplicación se ha observado que los intervalos de confianza clásicos no difieren demasiado de los intervalos bootstrap empírico y paramétrico. Los intervalos bootstrap de tipo Wild, en cambio, sugieren intervalos de mayor amplitud. A partir de este trabajo se sugiere un estudio mas detallado del comportamiento de estos intervalos bootstrap en situaciones especiales tales como tamaños de muestra pequeños, cambios de curvatura más pronunciada o diferente variabilidad en los datos. Asimismo se propone extender el estudio a situaciones donde los términos asociados a los nodos se consideran aleatorios y por tanto deberá considerarse su variabilidad para la construcción de los intervalos bootstrap asociados.

REFERENCIAS BIBLIOGRÁFICAS

Kauermann, G., Claeskens, G. & Opsomer, J.D. (2009). Bootstrapping for Penalized Spline-Regression, *Journal of Computational and GraphicalStatistics*, 18, 126-146.

Härdle, W., Marron, J.S. (1991). Bootstrap Simulataneous Error Bars for Nonparametric Regression. *Annals of statistics*, 19 (2), 778-796.

Efron, B., Tibshirani, R. (1993). *An Introduction to the Bootstrap*, New York: Chapman& Hall.

Efron, B. (1979). Bootstrap Methods: another look at the jackknife*. Annals of Statistics,* 7 , 1-26.