**Cristina Cuesta**
**Gonzalo Marí**
**Nicolás Zino**
*Instituto de Investigaciones Teóricas y Aplicadas en Estadística (IITAE)*

# INTERVALOS DE CONFIANZA BOOTSTRAP BAJO INCUMPLIMIENTO DE SUPUESTOS EN REGRESIONES SPLINES PENALIZADAS

## 1. Introducción a los modelos P-splines

Consideremos el modelo

$$\mathbf{y}/\mathbf{x} = \boldsymbol{\mu}(\mathbf{x}) + \boldsymbol{\varepsilon} \tag{1}$$

donde $\boldsymbol{\varepsilon}$ representa a los errores aleatorios y $\boldsymbol{\mu}(\mathbf{x})$ se asume una función suave, no especificada. Resulta conveniente descomponer a $\boldsymbol{\mu}(\mathbf{x})$ en una matriz de baja dimensión $\mathbf{X}$ que puede contener una forma polinómica y en una componente de alta dimensión $\mathbf{Z}$, compuesta por bases truncadas (aunque admite otras), cuyas expresiones son de la forma $(x - N_k)_+$, donde $(x)_+ = x$ para $x > 0$ y $0$ en otro caso, y siendo $N_k$ los nodos de la función. Esto nos conduce a la siguiente formulación del modelo:

$$\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \mathbf{z}\mathbf{u} + \boldsymbol{\varepsilon} = \mathbf{C}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \tag{2}$$

donde $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I})$ y siendo $\mathbf{C} = (\mathbf{x}||\mathbf{z})$ *y* $\boldsymbol{\theta} = (\boldsymbol{\beta}'||\mathbf{u}')'$.

La estimación de $\boldsymbol{\theta}$ a través de un ajuste paramétrico simple puede provocar inconvenientes de cálculo debido a la alta dimensionalidad de $\mathbf{C}$. Entonces, $\boldsymbol{\theta}$ puede ser estimado, imponiendo una penalidad a los coeficientes de $\mathbf{u}$, lo que conduce a minimizar el criterio:

$$(\mathbf{y} - \mathbf{C}\boldsymbol{\theta})'(\mathbf{y} - \mathbf{C}\boldsymbol{\theta}) - \lambda \mathbf{u}'\mathbf{A}\mathbf{u}$$

donde $\mathbf{A}$ es una matriz de penalidad adecuadamente elegida y $\lambda$ es el parámetro de suavizado. Para el caso de bases truncadas, resulta conveniente elegir a $\mathbf{A}$ como la matriz identidad.

La estimación de $\theta$ resulta entonces:

$$\widehat{\boldsymbol{\theta}} = (\mathbf{C}'\mathbf{C} + \lambda\mathbf{D})^{-1}\mathbf{C}'\mathbf{y}$$

con $\mathbf{D} = diag(\mathbf{O}, \mathbf{A}) = diag(\mathbf{0_2}, \mathbf{I_k})$

## 2. Supuestos del modelo

El cumplimiento de los supuestos en los modelos de regresión garantiza que las estimaciones obtenidas a través del método de mínimos cuadrados ordinarios sean los mejores estimadores lineales insesgados (BLUP). Cuando tales supuestos son violados, se generan problemas en los resultados alcanzados, haciendo que las estimaciones obtenidas no cumplan con algunas de las propiedades deseables.

A continuación se presentan los supuestos que deben cumplir los modelos de regresión con una única variable explicativa.

### 2.1. Variancia constante

Los errores del modelo de regresión deben tener variancia homogénea. Si la variancia del modelo crece o disminuye, ocurre la Heterocedastidad y su presencia repercute en la eficiencia de los estimadores, haciendo que las pruebas estadísticas carezcan de validez o que las inferencias sean erróneas.

### 2.2. No correlación

Se exige que los errores asociados a cada variable explicativa no estén relacionados en el tiempo. El problema que se presenta se llama Autocorrelación y afecta la validez estadística de los test que miden la significancia de dichas variables en el modelo.

### 2.3. Normalidad

Los residuos del modelo deben seguir una distribución normal. Es uno de los supuestos claves del modelo, dado que permite desarrollar las pruebas de hipótesis, basadas en los estadísticos t de Student y F de Snedecor.

## 3. Intervalos de confianza para modelos de efectos fijos

### 3.1. Intervalos de confianza clásico

Bajo los supuestos planteados para el modelo (2) se puede definir un intervalo de confianza para $\boldsymbol{\mu}(\mathbf{x})$ de la siguiente forma:

$$\mathbf{C}\widehat{\boldsymbol{\theta}} \pm z_{1-\frac{\alpha}{2}}\widehat{st.dev}\left(\mathbf{C}\widehat{\boldsymbol{\theta}}\right)$$

donde

$$\widehat{st.dev}\left(\mathbf{C}\widehat{\boldsymbol{\theta}}\right) = \sigma_{\varepsilon}\sqrt{\mathbf{C_x}\left(\mathbf{C}'\mathbf{C} + \lambda\mathbf{D}\right)^{-1}\mathbf{C}'\mathbf{C}\left(\mathbf{C}'\mathbf{C} + \lambda\mathbf{D}\right)^{-1}\mathbf{C}_{\mathbf{x}}'}$$

con $\mathbf{D} = diag(\mathbf{0_2}, \mathbf{I_k})$

### 3.2. Intervalos de confianza bootstrap

El método Bootstrap es un método de replicación desarrollado por Efron (1979). Consiste en la reutilización de la muestra original. De ésta se seleccionan un número determinado de veces muestras con reposición de igual tamaño (denominadas bootstrap), a partir de la cual se obtienen estimaciones de los parámetros de interés aplicando el mismo estimador a cada una de ellas. Luego se podrán 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.

#### 3.2.1. Bootstrap paramétrico

a. Obtener una estimación de $\sigma_{\varepsilon}^2$
b. Calcular $\mathbf{y}^* = \mathbf{C}\widehat{\boldsymbol{\theta}} + \boldsymbol{\varepsilon}^*$, donde $\boldsymbol{\varepsilon}^*$ es generado de una distribución $N(\mathbf{0}, \sigma_{\varepsilon}^2\mathbf{I})$
c. Se obtiene $\hat{\mathbf{y}}^*$ a partir del modelo (2)
d. Se repiten los pasos b. y c. B veces
e. Para cada valor de $x$ obtener los percentiles 2.5% y 97.5% de la distribución de $\hat{y}^*/x$

#### 3.2.2. Bootstrap empírico

a. Estimar los residuales a partir del modelo (2)
b. Obtener $\boldsymbol{\varepsilon}^* = \{\varepsilon_i^*\}_{i=1\cdots n}$, una muestra aleatoria con reemplazo de tamaño n de los residuales obtenidos en a.
c. Calcular $\mathbf{y}^* = \mathbf{C}\widehat{\boldsymbol{\theta}} + \boldsymbol{\varepsilon}^*$, donde los $\boldsymbol{\varepsilon}^*$ son los obtenidos en b.
d. Postular el modelo (2) con $(\mathbf{x}, \mathbf{y}^*)$ y obtener $(\mathbf{x}, \hat{\mathbf{y}}^*)$
e. Repetir los pasos b. hasta d. B veces
f. Para cada valor de $x$ obtener los percentiles 2.5% y 97.5% de la distribución de $\hat{y}^*/x$

### 3.2.3. Bootstrap Wild

a. Estimar los residuales a partir del modelo (2)
b. Obtener $\boldsymbol{\varepsilon}^* = \{\varepsilon_i^*\}_{i=1\cdots n}$, de una distribución de 2 puntos con masa $a_i = \hat{\varepsilon}_i\left(1-\sqrt{5}\right)/2$ y $b_i = \hat{\varepsilon}_i\left(1+\sqrt{5}\right)/2$ y probabilidad muestral $P(\hat{\varepsilon}_i = a_i) = \left(5+\sqrt{5}\right)/10$
c. Calcular $\mathbf{y}^* = \mathbf{C}\widehat{\boldsymbol{\theta}} + \boldsymbol{\varepsilon}^*$, donde los $\boldsymbol{\varepsilon}^*$ son los obtenidos en b.
d. Postular el modelo (2) con $(\mathbf{x}, \mathbf{y}^*)$ y obtener $(\mathbf{x}, \hat{\mathbf{y}}^*)$
e. Repetir los pasos b. hasta d. B veces
f. Para cada valor de $x$ obtener los percentiles 2.5% y 97.5% de la distribución de $\hat{y}^*/x$

## 4. Aplicación

Se realizó un estudio por simulación a partir del cual se obtuvieron 100 pares de valores $(x, y)$ utilizando la siguiente función de generación de datos:

$$y_i = 3x_i - 3x_i^2 + x_i^3 - 0.1x_i^4 + \varepsilon_i$$

con $x \in [0,3]$ y donde los $\varepsilon_i$ van a tener una distribución que dependerá de los escenarios de cumplimiento de los supuestos que se estén proponiendo.

### 4.1 Escenario 1: Errores Normales con variancia constante

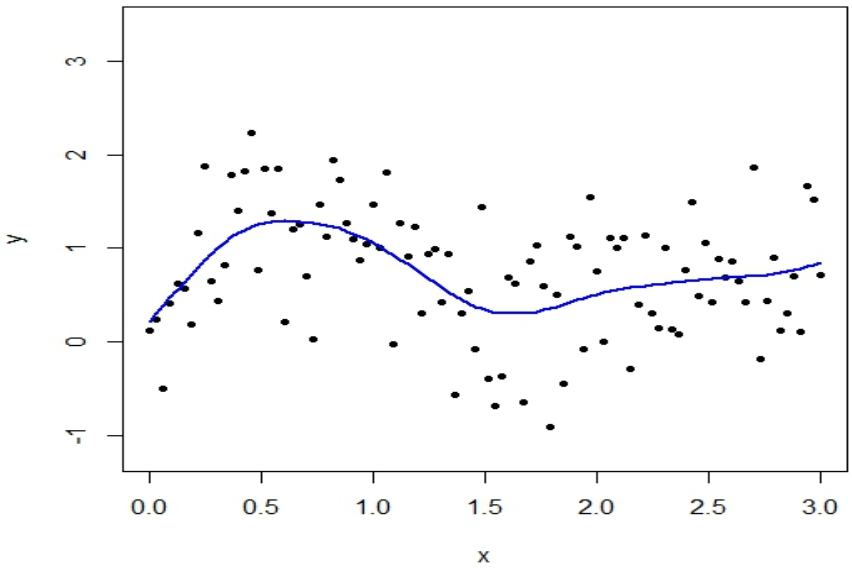
Se comenzó estudiando el comportamiento de los intervalos Bootstrap bajo cumplimiento de los supuestos de los modelos de regresión.

Para ello, los errores se generaron de la siguiente manera:

$$\varepsilon_i \widetilde{iid} N(0, \sigma_\varepsilon^2)$$

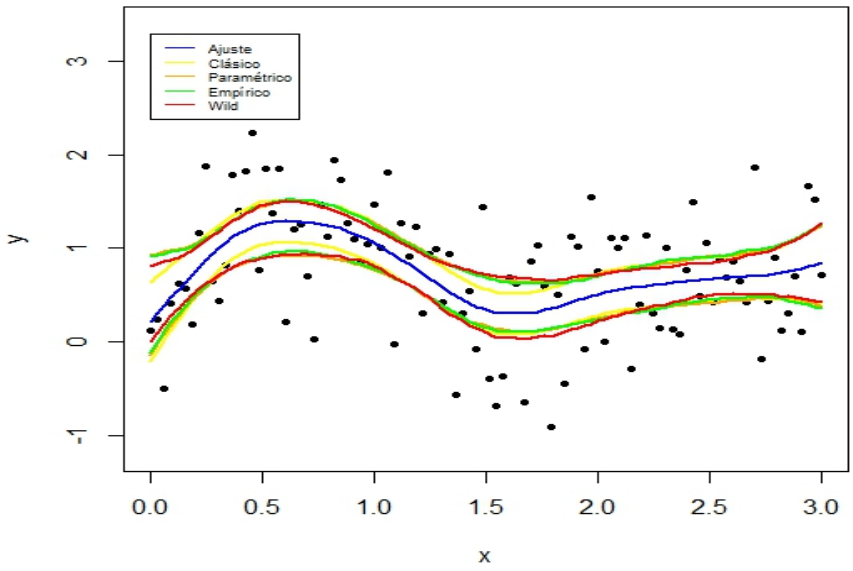
con $\sigma_\varepsilon^2 = 0.36$. La Figura 1 muestra el diagrama de dispersión de los datos simulados y el correspondiente ajuste bajo un modelo P-spline lineal de efectos fijos, con errores normales y variancia constante.

Figura 1. Regresión P-spline lineal bajo cumplimiento de los supuestos.




En la figura siguiente se presentan los intervalos de confianza clásicos y los diferentes tipos de intervalos Bootstrap.

Figura 2. Intervalos de Confianza Clásico y Bootstrap para la Regresión P-spline lineal bajo cumplimiento de los supuestos.



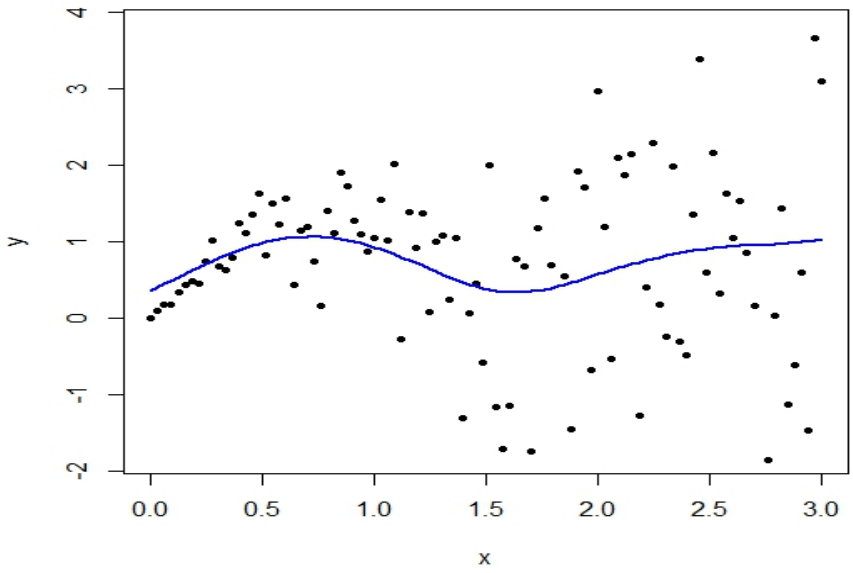

## 4.2 Escenario 2: Errores Normales con variancia no constante

En este caso, los intervalos Bootstrap se analizaron bajo condiciones de Heteroscedasticidad, la cual fue obtenida con los errores definidos como se muestra a continuación:

$$\varepsilon_{i\widetilde{iid}} N\left(0, \frac{\sigma_\varepsilon^2 x_i}{2}\right)$$

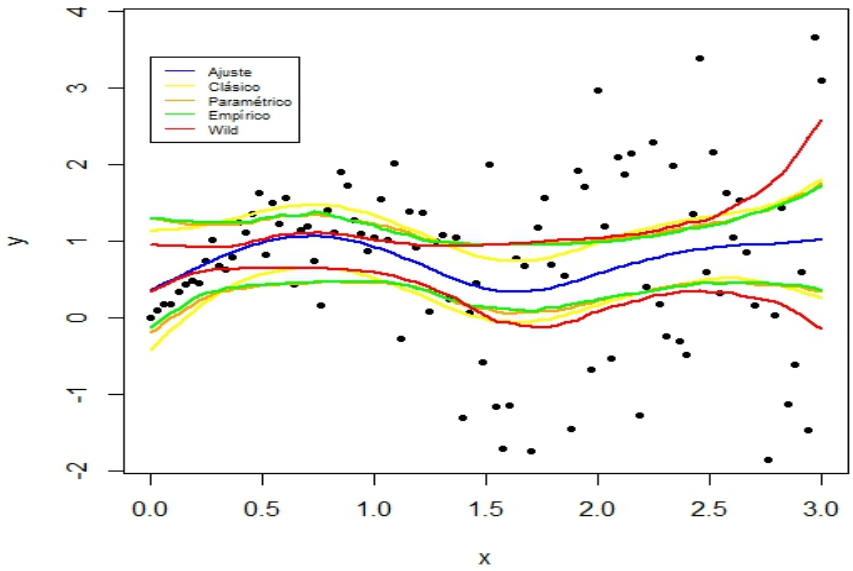
con $\sigma_\varepsilon^2 = 1.82$. La Figura 3 muestra el diagrama de dispersión de los datos simulados y el ajuste bajo un modelo P-spline lineal de efectos fijos, con errores normales y variancia no constante.

Figura 3. Regresión P-spline lineal bajo incumplimiento del supuesto de variancia constante.




En la figura siguiente se presentan los intervalos de confianza clásicos y los diferentes tipos de intervalos Bootstrap, bajo las condiciones descriptas.

Figura 4. Intervalos de Confianza Clásico y Bootstrap para la Regresión P-spline lineal, bajo incumplimiento del supuesto de variancia constante.



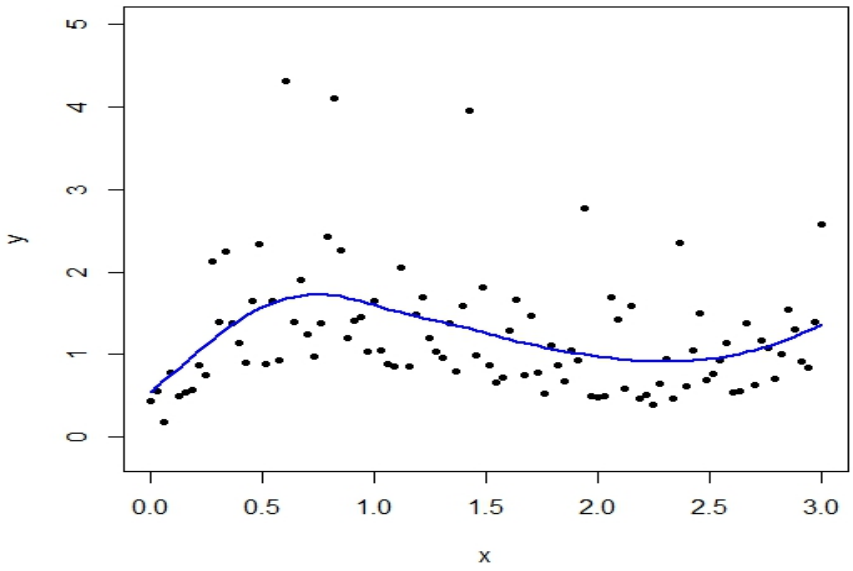

### 4.3 Escenario 3: Errores No Normales

En este escenario, los intervalos Bootstrap se estudiaron bajo la situación de errores provenientes de una distribución Gamma:

$$\varepsilon_{i\widetilde{iid}} Gamma(\alpha, \beta)$$

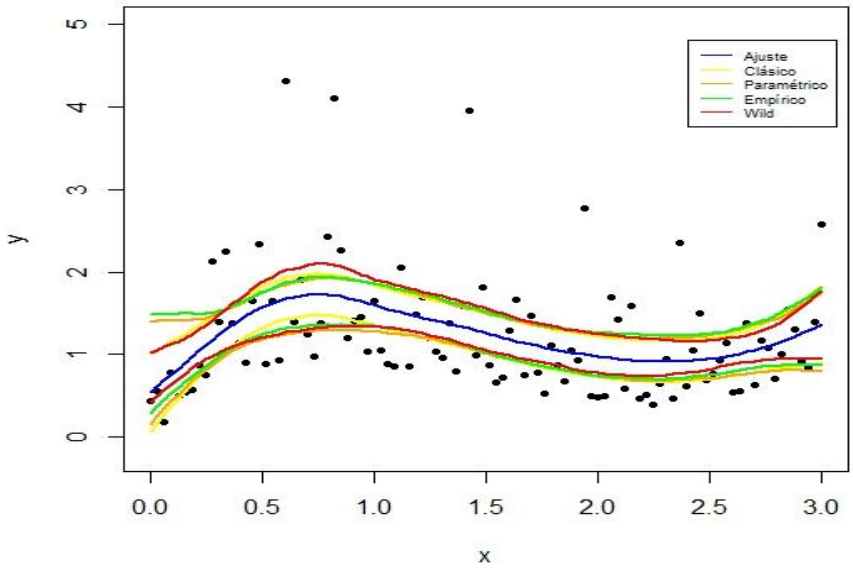
con $\alpha = 1$ y $\beta = 2/3$. La Figura 5 muestra el diagrama de dispersión de los datos simulados y el correspondiente ajuste bajo un modelo P-spline lineal de efectos fijos, con errores generados a partir de una distribución Gamma.

Figura 5. Regresión P-spline lineal con errores no normales.




En la figura siguiente se presentan los intervalos de confianza clásicos y los diferentes tipos de intervalos Bootstrap, bajo las condiciones descriptas.

Figura 4. Intervalos de Confianza Clásico y Bootstrap para la Regresión P-spline lineal, con errores no normales.



## 5. CONCLUSIÓN

Las Regresiones Splines Penalizadas constituyen una herramienta poderosa a la hora de describir la relación entre dos variables, cuando esto no puede ser realizado mediante los modelos simples de regresión. Sin embargo, al igual que ocurre con los modelos clásicos de regresión, se suele pasar por alto la verificación del cumplimiento de los supuestos.

En esta aplicación, se observó el impacto de la violación de algún supuesto en los intervalos de confianza del ajuste de la regresión p-spline.

Tanto los intervalos clásicos como los intervalos Bootstrap empírico y paramétrico ofrecen comportamientos similares, incluso en las situaciones de incumplimiento de alguno de los supuestos. En cambio, los intervalos Bootstrap Wild sugieren intervalos de mayor amplitud, en especial, cuando no se cumple el supuesto de homogeneidad de variancia de los errores.

### REFERENCIAS BIBLIOGRÁFICAS

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

Kauermann, G., Claeskens, G. and Opsomer, J. D. (2006). *Bootstrapping for Penalized Spline Regression*. Journal of Computational and GraphicalStatistics, 18, 126-146.

Ngo, L. and Wand M. (2004). *Smoothing with Mixed Models Software*. Journal of Statistical Software Volume 09, Issue 01.

Ruppert, D. (2002). Selecting the Number of Knots for Penalized Splines. Journal of Computational and Graphical Statistics. Volume 11(4):735-757.

Ruppert, D., Wand, M. P. and Carrol, R. J. (2003). Semiparametric Regression. Cambridge University Press. New York.