Marí, Gonzalo
Zino, Nicolás
Instituto de Investigaciones Teóricas y Aplicadas en Estadística (IITAE)

INTERVALOS DE CONFIANZA BOOTSTRAP EN REGRESIÓN P-SPLINE CON ERRORES AUTOCORRELACIONADOS

1. Introducción

Los modelos de regresión constituyen una herramienta muy útil y potente cuando el objetivo es relacionar una variable respuesta con una o más variables explicativas. Sin embargo estos modelos tienen una estructura rígida (lineal, cuadrática, etc., o aún no lineal), que a menudo no logra "captar" el comportamiento de los datos en todo el campo de variación de las variables explicativas.

Cuando el diagrama de dispersión revela una fuerte asociación entre la variable respuesta y la explicativa pero esta relación presenta cambios de pendiente, picos y valles a lo largo del campo de variación de la variable explicativa, o bien cuando, debido a un tamaño de muestra grande y a una gran variabilidad en los datos, no es posible explicitar de antemano un modelo de regresión polinómica o de regresión no lineal, debe recurrirse a otras herramientas que permitan representar adecuadamente los datos.

Comprendidas dentro del conjunto de técnicas de suavizado, las Regresiones Splines Penalizadas (P-splines) han recibido durante las últimas décadas una renovada atención, convirtiéndose en una herramienta simple pero efectiva, para describir relaciones entre variables que no pueden ser debidamente alcanzadas por los métodos de regresión usuales. Su principal virtud radica en la flexibilidad para adaptarse a la forma original de la relación, permitiendo obtener buenos resultados con polinomios de bajo grado.

En definitiva, las regresiones P-splines corresponden a una regresión por partes, donde cada una de ellas es una región del campo de variación de la variable explicativa en la que se ajusta un modelo de regresión polinómica (en general de bajo orden) y que están unidas en los extremos ("nodos") para dar continuidad a la curva. Dado que dependen en gran medida de la cantidad de regiones que se consideren y de su amplitud, es necesario definir un número grande de regiones y ponderar la importancia de las mismas.

Como todo modelo de regresión, las regresiones P-splines están sujetas al cumplimiento de los supuestos, los cuales junto a los métodos de estimación, caracterizan a los resultados alcanzados por la regresión. El cumplimiento de los mismos permite asegurar la estabilidad de las estimaciones obtenidas y de esa manera, confirmar la validez de las inferencias que se realicen con posterioridad.

En consecuencia, cuando tales supuestos son violados, las propiedades de los estimadores es no se cumplen e impactan directamente en los coeficientes, las pruebas de hipótesis, los intervalos de confianza, las predicciones, etc.

Existen diversas formas de construir intervalos de confianza. La más conocida es que proviene de la teoría clásica. Por otro lado, la técnica Bootstrap es un método de replicación que consiste en la reutilización de los datos muestrales y que permite obtener estimaciones de los parámetros de interés aplicando el mismo estimador a cada una de las muestras generadas a partir de la original. Dentro de estos métodos, se pueden mencionar el Bootstrap paramétrico, empírico y Wild. La dificultad que tienen todos estos métodos es que se consideran a los errores como independientes. En este trabajo se presenta una adaptación de los

métodos Bootstrap que contempla errores autocorrelacionados. A partir de un estudio de simulación, se evalúa la cobertura de los métodos considerados.

2. Modelos de Regresión P-spline

Consideremos el modelo

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

donde $\boldsymbol{\varepsilon}$ representa a los errores aleatorios y $\mathbf{\mu(x)}$ se asume una función suave, no especificada. Resulta conveniente descomponer a $\mathbf{\mu(x)}$ en una matriz de baja dimensión **X** que puede contener una forma polinómica y en una componente de alta dimensión **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:

$$\mathbf{y = x\beta + zu + \varepsilon = C\theta + \varepsilon} \tag{2}$$

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

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

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

donde **A** es una matriz de penalidad adecuadamente elegida y λ es el parámetro de suavizado. Para el caso de bases truncadas, resulta conveniente elegir a **A** como la matriz identidad.

La estimación de $\boldsymbol{\theta}$ resulta entonces:

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

con $\mathbf{D} = diag(\mathbf{0}, \mathbf{A}) = diag(\mathbf{O_2}, \mathbf{I_k})$

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 (BLUE). 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

Se exige que los errores asociados a cada variable explicativa no estén relacionados en el tiempo. Cuando se viola este supuesto, surge el problema de autocorrelación.

Como los errores no son observables, la práctica habitual es suponer que han sido generados por algún mecanismo. El mecanismo comúnmente adoptado es el esquema autorregresivo de primer orden, que supone que el error en un determinado instante de tiempo está linealmente relacionado con el término del error en el tiempo anterior y cuya medida de esta interdependencia está dada por el coeficiente de autocorrelación, o sea,

$$\varepsilon_t = \rho\varepsilon_{t-1} + v_t$$

siendo v_t ruido blanco y $-1 < \rho < 1$.

La variancia del estimador bajo autocorrelación podrá ser mayor o menor que la de mínimos cuadrados dependiendo del valor de ρ. Si ρ es positivo, se sobreestima la variancia mientras que un valor negativo no define el sentido del sesgo.

En consecuencia, como en el caso de heterocedasticidad, en presencia de autocorrelación de los errores, los estimadores continúan siendo lineales, insesgados y consistentes pero dejan de ser los mejores, es decir, dejan de tener variancia mínima.

3. Intervalos de confianza para modelos de efectos fijos

3.1 Intervalos de confianza clásicos

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

$$\mathbf{C}\hat{\boldsymbol{\theta}} \pm z_{1-\alpha/2}\hat{S}(\mathbf{C}\hat{\boldsymbol{\theta}})$$

donde

$$\hat{S}(\mathbf{C}\hat{\boldsymbol{\theta}}) = \sigma_\varepsilon \sqrt{\mathbf{C_x}(\mathbf{C'C} + \lambda\mathbf{D})^{-1}\mathbf{C'C}(\mathbf{C'C} + \lambda\mathbf{D})^{-1}\mathbf{C'_x}}$$

con $\mathbf{D} = diag(\mathbf{O_2}, \mathbf{I_k})$.

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. Dado que esto métodos se basan en los supuestos antes mencionados, se presenta una modificación de los mismos que considera el caso de errores autocorrelacionados y se

lo denomina bootstrap correlacionado. 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 σ_ε^2
b. Calcular $\mathbf{y}^* = \mathbf{C}\hat{\boldsymbol{\theta}} + \boldsymbol{\varepsilon}^*$, donde $\boldsymbol{\varepsilon}^*$ es generado de una distribución $N(\mathbf{0}, \hat{\sigma}_\varepsilon^2 \mathbf{I})$
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 $\hat{\mathbf{y}}^* / x$

3.2.2. Bootstrap Empírico

a. Estimar los residuales a partir del modelo (2)
b. Obtener $\boldsymbol{\varepsilon}^* = \varepsilon_i^*{}_{i=1,\ldots,n}$, una muestra aleatoria con reemplazo de tamaño n de los residuales obtenidos en a.
c. Calcular $\mathbf{y}^* = \mathbf{C}\hat{\boldsymbol{\theta}} + \boldsymbol{\varepsilon}^*$, 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 $\hat{\mathbf{y}}^* / x$

3.2.3. Bootstrap Wild

a. Estimar los residuales a partir del modelo (2)
b. Obtener $\boldsymbol{\varepsilon}^* = \varepsilon_i^*{}_{i=1,\ldots,n}$, de una distribución de 2 puntos con masa $a_i = \frac{\hat{\varepsilon}_i(1-\sqrt{5})}{2}$ y $a_i = \frac{\hat{\varepsilon}_i(1+\sqrt{5})}{2}$ y probabilidad muestral $P(\hat{\varepsilon}_i = a_i) = \frac{5+\sqrt{5}}{10}$
c. Calcular $\mathbf{y}^* = \mathbf{C}\hat{\boldsymbol{\theta}} + \boldsymbol{\varepsilon}^*$, 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 $\hat{\mathbf{y}}^* / x$

3.2.4. Bootstrap Correlacionado

a. Estimar los residuales a partir del modelo (2)
b. A partir de los residuales de a) obtener los valores de $v_t = e_t - \hat{\rho} e_{t-1}$.
c. Sea $e_1^* = v^* / \left(1-\hat{\rho}\right)^{0.5}$ con v^* seleccionado de $v_t{}_{t=1,\ldots,n}$. Generar $e_t^* = \hat{\rho} e_{t-1} + v_t^*$ para $t = 2,\ldots,n$ con v_t^* seleccionado de $v_t{}_{t=1,\ldots,n}$ con reemplazo
d. Calcular $\mathbf{y}^* = \mathbf{C}\hat{\boldsymbol{\theta}} + \boldsymbol{\varepsilon}^*$, donde los $\boldsymbol{\varepsilon}^*$ son los obtenidos en c.
e. Postular el modelo (2) con $(\mathbf{x}, \mathbf{y}^*)$ y obtener $(\mathbf{x}, \hat{\mathbf{y}}^*)$
f. Repetir los pasos b. hasta e. B veces
g. Para cada valor de x obtener los percentiles 2.5% y 97.5% de la distribución de $\hat{\mathbf{y}}^* / x$

4. Estudio por simulación

Con el objetivo de evaluar el comportamiento de los intervalos de confianza asociados al ajuste de una regresión spline penalizada, se lleva a cabo un estudio por simulación a partir del cual se obtienen 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 ε_i provienen de una serie autorregresiva de orden 1 $\varepsilon_t = \rho\varepsilon_{t-1} + v_t$, con $v_t \sim N(0,1)$.

La formulación de este modelo teórico no es ingenua, debido a que se conoce con antelación que la forma del conjunto de datos generado difícilmente pueda ser captada por los modelos de regresión usuales, lo que origina la necesidad de ajustar un modelo de regresión p-spline.

Se considera un tamaño de muestra 100, y se seleccionan 200 muestras del modelo planteado para distintos valores de ρ: 0.2, 0.4, 0.6, y 0.8. Para cada uno de los valores de x's, se obtiene $\hat{y}$, y se construye en cada una de las 200 repeticiones el intervalo de confianza por los cinco métodos presentados en el punto 3: clásico, Bootstrap Paramétrico, Bootstrap Empírico, Bootstrap Wild, y Bootstrap Correlacionado. Se define la cobertura de cada uno de los métodos como el porcentaje de intervalos que cubren al verdadero valor que surge del modelo teórico.

A continuación se presentan los gráficos con los distintos niveles de cobertura obtenido en cada uno de los métodos.

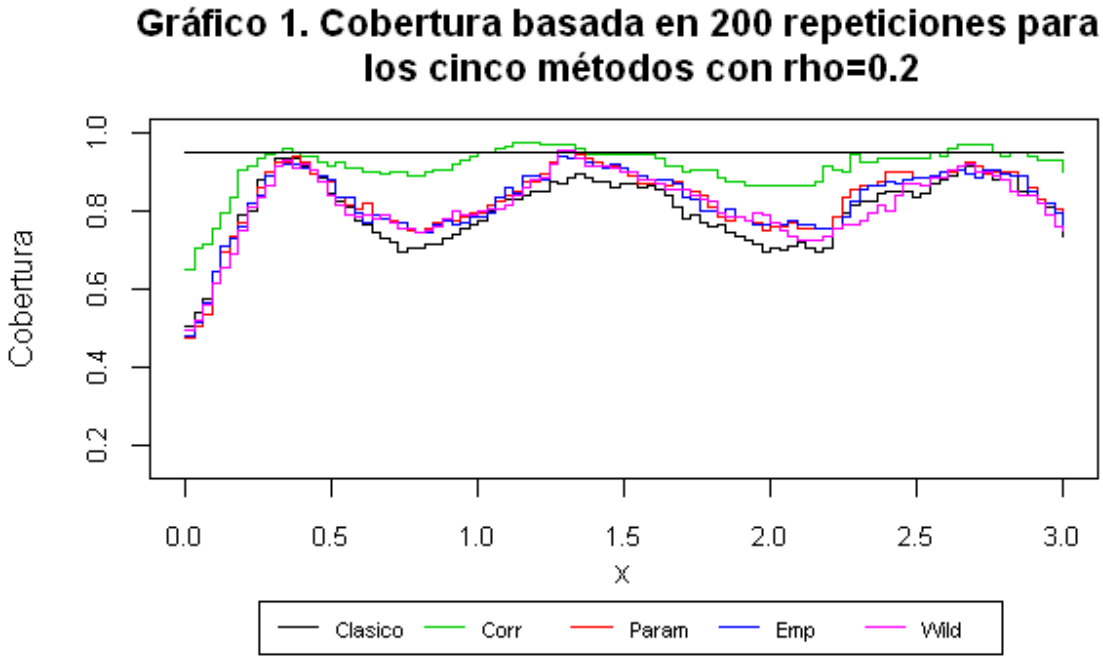


Gráfico 1. Cobertura basada en 200 repeticiones para los cinco métodos con rho=0.2
Cobertura
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
X
Clasico
Corr
Param
Emp
Wild

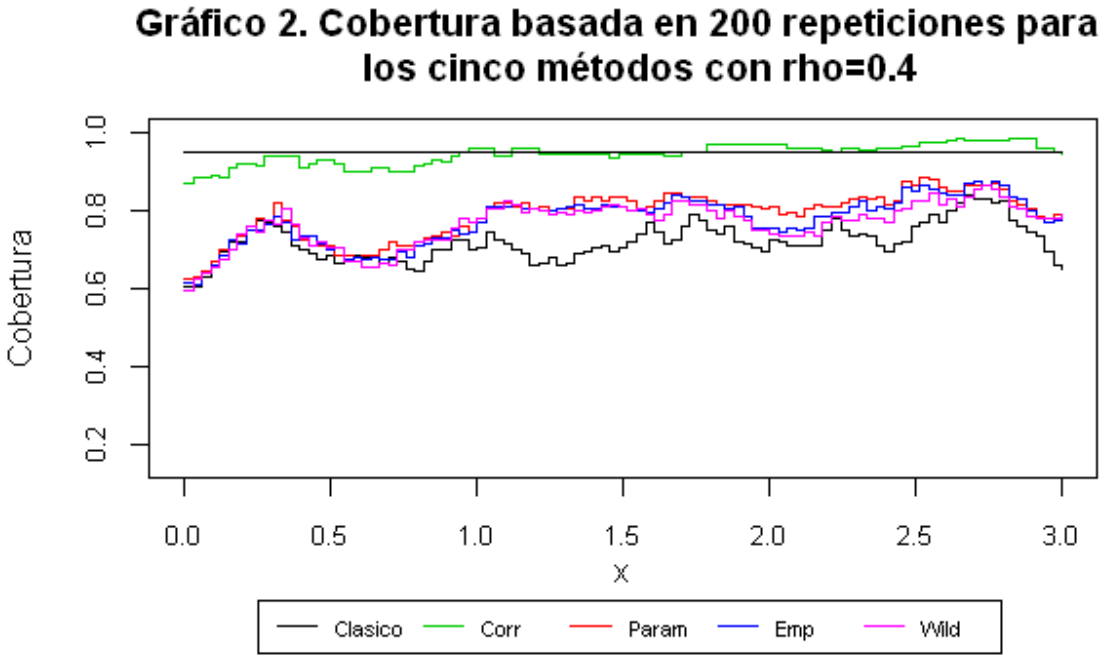
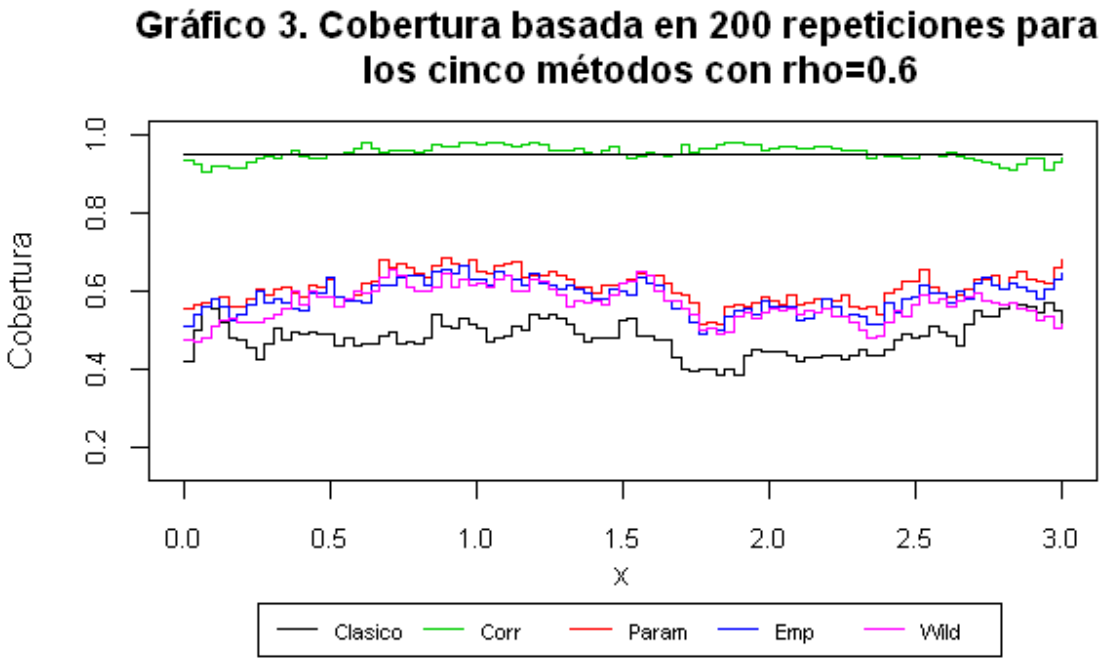
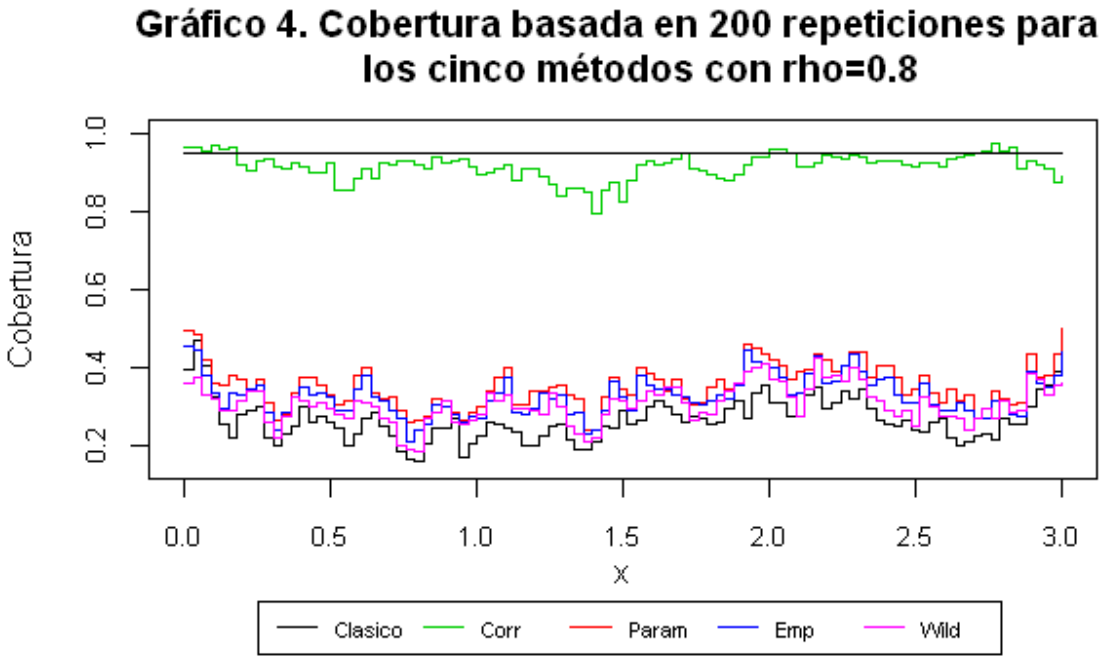


Gráfico 2. Cobertura basada en 200 repeticiones para los cinco métodos con rho=0.4
Cobertura
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
X
Clasico
Corr
Param
Emp
Wild





Por otro lado se presenta en la siguiente tabla, el promedio de las coberturas observadas en las 200 repeticiones para los 100 valores de x's,.

Tabla 1. Cobertura promedio para los cinco métodos y los cuatro valores de ρ

Método	ρ			
	0.2	0.4	0.6	0.8
Clásico	0.802	0.719	0.481	0.2671
Bootstrap Correlacionado	0.915	0.945	0.953	0.9161
Bootstrap Paramétrico	0.831	0.789	0.609	0.3536
Bootstrap Empírico	0.827	0.776	0.586	0.3286
Bootstrap Wild	0.817	0.767	0.566	0.3066

Puede observarse que los métodos que asumen independencia poseen un nivel de cobertura alejado del valor nominal planteado cuando los errores se encuentran autocorrelacionados. El Bootstrap Correlacionado es el único método que brinda cobertura más cercanas al 95%, siendo alcanzado (y superado) ese valor para algunos valores de x's. Los intervalos Bootstrap restantes presentan un resultado similar, apenas superior al que se observa para el método clásico, que es el que presenta el nivel de cobertura observado, basado en las simulaciones, más alejado del valor nominal. Esto se observa para todos los valores de ρ, aun para el más cercano a 0.

Estas diferencias se van acentuando a medida que el valor ρ aumenta, llegando a tener los intervalos que asumen independencia coberturas observadas muy alejadas del 95% en el cual se fijó el valor nominal.

5. Conclusiones

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 del supuesto de independencia de los errores en los intervalos de confianza del ajuste de la regresión p-spline.

Tanto los intervalos clásicos como los intervalos Bootstrap Empírico, Paramétrico y Wild ofrecen comportamientos similares, observándose mejor cobertura muy leve para los intervalos Bootstrap. En cambio, los intervalos Bootstrap Correlacionados sugieren intervalos de con un mayor poder de cobertura lo que habla a las claras que el supuesto de independencia que es a menudo ignorado, puede tener consecuencias indeseables cuando se utilizan métodos que no suponen un ajuste causado por el incumplimiento del mismo.

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 Graphical Statistics*, 18, 126-146.

Kim, J. (2005). Bias-Corrected Bootstrap Inference for Regression Models with Autocorrelated Errorrs". *Econometrics Bulletin*, Vol. 3, N° 44, pp. 1-8.

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

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

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

Shao, J., Tu, D. (1995). *The Jackknife and the Bootstrap*. Springer-Verlag.