Cristina Cuesta

María Luz Armida

Evangelina Lupachini

Instituto de Investigaciones Teóricas y Aplicadas en Estadística (IITAE)

SESGO DE ESTIMACIÓN EN REGRESIONES P-SPLINE BAJO MODELOS MIXTOS: UN ESTUDIO POR SIMULACIÓN.

Introducción

Las regresiones spline Penalizadas (P-spline) forman parte del conjunto de técnicas de suavizado que permiten relacionar una variable respuesta con una explicativa sin hacer supuestos a priori de la forma funcional que las liga. En particular, los métodos de regresión spline dividen el campo de variación de la variable explicativa en regiones dentro de las cuales se ajusta un modelo de regresión de bajo orden (lineal o cuadrático) que luego son unidos en los puntos extremos (llamados nodos). Por tratarse de un modelo de regresión, es simple estimar la variable respuesta e incluso construir intervalos de confianza y de predicción para dichas estimaciones. El mayor inconveniente que presenta esta técnica es que suele ser vulnerable a la cantidad de nodos que se empleen y a la ubicación de los mismos. Una opción para eludir este problema es ponderar los parámetros asociados a los nodos, recurriendo así a los modelos de regresión spline penalizada. Estos últimos modelos dependen del parámetro de suavizado que es quien regula la penalización impuesta a los parámetros. No siempre es simple decidir cual es el valor del parámetro de suavizado que provoca un mejor ajuste de los datos. Ruppert (2003) ha mostrado que esta problemática queda resuelta al reformular estos modelos de regresión P-spline dentro del marco de los modelos mixtos. Sin embargo, al momento de construir intervalos de confianza se suelen cometer sesgos, debidos a la nueva componente aleatoria que interviene en el modelo mixto, y que frecuentemente no se tiene en cuenta al momento de hacer inferencias.

En este trabajo se presenta la comparación entre los intervalos de confianza obtenidos teniendo en cuenta dichos sesgos y sin tenerlos en cuenta. Para ello se simulan situaciones teóricas bajo diferentes escenarios en función de distinta cantidad de nodos y variabilidad de la variable respuesta. A partir de los resultados obtenidos puede comprenderse mejor bajo qué situaciones es más necesario tener en cuenta el sesgo producido por utilizar un modelo mixto y en cuales otras dicho sesgo es despreciable.

Metodología

La relación entre una variable explicativa (o predictora) y una respuesta no es siempre simple de describir. A menudo no es suficiente con un polinomio de bajo orden o con un modelo de regresión no lineal en los parámetros. Esta situación puede subsanarse construyendo un modelo de regresión por partes, es decir construyendo modelos de regresión líneal (o cuadrática) en cada una de las regiones pre-definidas de la variable explicativa y uniéndolos en los extremos. Hay diferentes modelos que pueden construirse bajo estas premisas. Así por ejemplo el modelo de regresión *spline* de bases truncadas cumple con dichas características.

La expresión general de un modelo de regresión es: $y_i = f(x_i) + \varepsilon_i$. En particular si se trata de un modelo de regresión Spline lineal de base truncada:

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

Es decir,

$$f(x_i) = \begin{cases} \beta_0 + \beta_1 x_i & ; x_i < N_1 \\ \beta_0 + \beta_1 x_i + \beta_{11}(x_i - N_1)_+ & ; N_1 \le x_i < N_2 \\ \ldots & \\ \beta_0 + \beta_1 x_i + \beta_{11}(x_i - N_1)_+ + \beta_{12}(x_i - N_2)_+ + \ldots + \beta_{1K}(x_i - N_K)_+ & ; x_i \ge N_K \end{cases}$$

donde, N_k representa el nodo k-ésimo y $a < N_1 < N_2 < \ldots < N_K < b$, siendo $[a,b]$ el intervalo de interés de la variable explicativa que se divide en $K+1$ subintervalos. La combinación lineal $1, x, (x-N_1)_+, \cdots, (x-N_K)_+$ logra un ajuste de funciones lineales en cada subintervalo que son unidas en los nodos $N_1, N_2, \cdots, N_K$.

β_0 y β_1 son los coeficientes de la regresión lineal de base y β_{1k} son los coeficientes de los efectos asociados a los nodos, también fijos.

La influencia de los parámetros asociados a los nodos puede ser controlada utilizando alguna restricción (por ejemplo $\sum_k \beta_{1k}^2 < C$). De forma tal que la cantidad de nodos y su ubicación pasa a ser un tema de menor relevancia. Bajo esta restricción, la estimación del vector de parámetros viene dada por la minimización de: $\|\mathbf{y} - \mathbf{X\beta}\|^2 + \lambda^{2p} \boldsymbol{\beta}' \mathbf{D} \boldsymbol{\beta}$. Por lo tanto el nuevo desafío es encontrar el parámetro de suavizado (λ).

Más aún, el modelo anterior puede reescribirse como un modelo mixto. En efecto, el modelo puede reformularse como sigue:

$$y_i = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} u_k (x_i - \kappa_k)_+ + \varepsilon_i, \tag{2}$$

donde β_0 y β_1 son los coeficiente fijos de la regresión y u_k son los coeficientes de los efectos aleatorios asociados a los nodos , $\varepsilon_i \underset{iid}{\sim} N(0, \sigma_\varepsilon^2)$ y $u_k \underset{iid}{\sim} N(0, \sigma_u^2)$ $k = 1, \ldots, K$ siendo ε_i y u_k independientes. Matricialmente: $\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon}$, donde

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad \mathbf{Z} = \begin{pmatrix} (x_1 - N_1)_+ & \cdots & (x_1 - N_K)_+ \\ \vdots & \ddots & \vdots \\ (x_n - N_1)_+ & \cdots & (x_1 - N_K)_+ \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \mathbf{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_K \end{pmatrix} = \begin{pmatrix} \beta_{11} \\ \vdots \\ \beta_{1K} \end{pmatrix},$$

con $\boldsymbol{\varepsilon} \to N(0, \mathbf{R} = \sigma_\varepsilon^2 \mathbf{I})$ y $\mathbf{u} \to N(0, \mathbf{G} = \sigma_u^2 \mathbf{I})$. En efecto, el criterio de ajuste de la Regresión Spline Penalizada dividido por σ_ε^2 resulta:

$$\frac{1}{\sigma_\varepsilon^2}\left\|\mathbf{y}-\mathbf{X\beta}-\mathbf{Zu}\right\|^2+\frac{\lambda^2}{\sigma_\varepsilon^2}\left\|\mathbf{u}\right\|^2$$

y, por su parte, el estimador BLUP del modelo (2) minimiza el criterio

$$\mathbf{(y - X\beta - Zu)'R^{-1}(y - X\beta - Zu) + u'G^{-1}u}\,,$$

con lo que queda claramente mostrado que estas expresiones coinciden cuando $\sigma_u^2=\sigma_\varepsilon^2/\lambda^2$. Es decir, la estimación de "y" será la misma bajo el modelo (1) y el modelo (2) cuando $\hat{\lambda}=\frac{\hat{\sigma}_\varepsilon}{\hat{\sigma}_u}$. Se recomienda ampliamente el uso del modelo mixto ya que con él no es necesario "preocuparse" por determinar el mejor valor para λ. Hay muchas otras ventajas que sugieren el uso del modelo mixto, por ejemplo, que pueden usarse los programas de computación usuales para modelos mixtos, se pueden agregar nuevas covariables al modelo, cambiar estructuras de correlación de los errores, etc.

Si se trabaja con la estimación BLUP de $f(x)$ entonces resulta, $\hat{f}\left(x\right)=\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{Z}\hat{\mathbf{u}}$, donde:

$$\begin{aligned} EBLUP(\boldsymbol{\beta}) &\equiv \hat{\boldsymbol{\beta}}=\left(\mathbf{X'}\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X'}\hat{\mathbf{V}}^{-1}\mathbf{y} \\ EBLUP(\mathbf{u}) &\equiv \hat{\mathbf{u}}=\hat{\mathbf{G}}\mathbf{Z'}\hat{\mathbf{V}}^{-1}\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right) \end{aligned},$$

siendo $\hat{\mathbf{V}}=\hat{\mathbf{Z}}'\hat{\mathbf{G}}\hat{\mathbf{Z}}+\hat{\mathbf{R}}$.

Para poder obtener las estimaciones EBLUP descriptas, es necesario previamente tener estimaciones de las componentes de variancia, para ello, se sugiere utilizar el procedimiento de maxima-verosimilitud o maxima-verosimilitud restringida.

A fin de acompañar la estimación puntual de $f(x)$ con sus intervalos de confianza, es necesario mencionar que la construcción de dichos intervalos bajo el modelo (2) difiere si se tiene o no en cuenta la variabilidad atribuida a la componente aleatoria asociada a los nodos. Bajo el argumento que la aleatoriedad de u se utiliza como un artificio para modelar la curvatura mientras que ε tiene en cuenta la variabilidad alrededor de la curva, el desvío estándar de $\hat{f}(x)$ se puede hacer condicional a u. En tal caso:

$$\hat{st.dev}\left(\hat{f}\left(x\right)\middle|\mathbf{u}\right)=\sigma_\varepsilon^2\sqrt{\mathbf{C_x}\left(\mathbf{C'C}+\frac{\sigma_\varepsilon^2}{\sigma_u^2}\mathbf{D}\right)^{-1}\mathbf{C'C}\left(\mathbf{C'C}+\frac{\sigma_\varepsilon^2}{\sigma_u^2}\mathbf{D}\right)^{-1}\mathbf{C_x'}}\,,$$

donde $\mathbf{C}=\left[\mathbf{X}\ \mathbf{Z}\right]$ y el intervalo de confianza aproximado del 100(1-α)% para $E\left\{\tilde{f}\left(x\right)\middle|\mathbf{u}\right\}$ es

$$\hat{f}(x_i)\pm z(1-\alpha/2)\ \hat{\sigma}_\varepsilon\ \hat{st.dev}(\hat{f}(x\ )/\mathbf{u})\,.$$

Si el sesgo es despreciable entonces $E\left[\tilde{f}\left(x\right)\middle|\mathbf{u}\right]\approx f\left(x\right)$, y el intervalo anterior puede interpretarse como un intervalo de confianza para $f\left(x\right)$.

Sin embargo, si el sesgo no es despreciable y se quiere tener en cuenta, entonces el intervalo de confianza del 100(1-α)% resulta:

$$\hat{f}(x_i)\pm z(1-\alpha/2)\ \hat{\sigma}_\varepsilon\ \hat{st.dev}(\hat{f}(x\ )-f(x))\ ,$$

donde: $\hat{st.dev}(\hat{f}(x) - f(x)) = \hat{\sigma}_\varepsilon^2 \sqrt{\mathbf{C}_x(\mathbf{C'C} + \frac{\sigma_\varepsilon^2}{\sigma_u^2}\mathbf{D})^{-1}\mathbf{C}_x'}$.

Este intervalo es un poco más amplio que el anterior ya que tiene en cuenta ambas componentes de error (el sesgo y la variancia) mientras que el primero sólo tiene en cuenta la variancia y cubre a $E(\tilde{f}(x)/\mathbf{u})$ y no a $f(x)$.

La diferencia entre ambos se nota especialmente en las regiones donde hay mayor curvatura, ya que aquí es donde el sesgo es mayor.

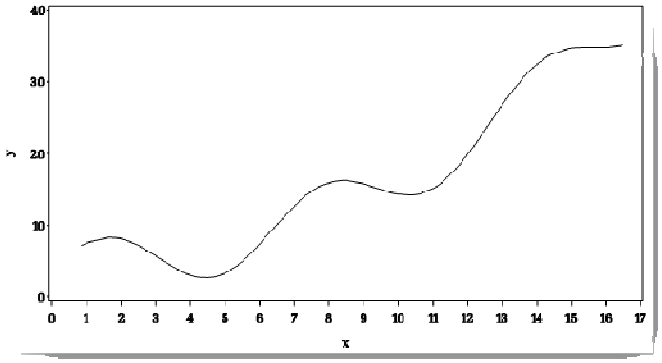
En general es preferible construir un intervalo para $f(x)$, sin embargo a menudo el IC utilizado cubre a $E(\tilde{f}(x)/\mathbf{u})$.

Se hace notar que ninguno de estos intervalos tiene en cuenta la variabilidad debido a la estimación de los parámetros de suavizado, es decir a la variabilidad causada por estimar σ_ε^2 y σ_u^2. Esto es lo más complejo de tener en cuenta, en la práctica se ignora esta fuente de variabilidad suponiendo que si el tamaño de muestra es suficientemente grande esta variabilidad extra es despreciable.

Estudio por simulación

Se lleva a cabo un estudio por simulación bajo el siguiente modelo de generación de datos $y_{ij} = \frac{x_i}{8} + 4sen(x_i) + 4 + r_{ij}$, donde $r_i \underset{iid}{\sim} N(0, \sigma_r^2)$ y $x_i \in [-\pi + 4; 4\pi + 4]$. La elección de este modelo no es ingenua, de hecho su formulación se debe a que se conoce a priori que su forma difícilmente pueda ser captada por un modelo lineal tradicional. El Gráfico 3.1 muestra el modelo teórico planteado.

Gráfico 3.1. Modelo teórico de generación de datos



Bajo el modelo propuesto se generan 100 pares de datos (x,y) para cada uno de 3 escenarios diferentes. Estos escenarios se diferencian en el valor de σ_r^2 utilizado (es decir, presentan diferente variabilidad en la variable respuesta). Los escenarios corresponden a las situaciones $\sigma_r^2 = 1, 4$ y 25.

Para cada uno de estos tres escenarios se ajustan tres modelos spline penalizados bajo el enfoque de los modelos mixtos . La diferencia entre ellos radica en la cantidad de nodos utilizados para su construcción. Se utilizan 4, 6 y 12 nodos respectivamente.

Las ubicaciones de los nodos para cada caso son:

4 nodos en (4.00; 7.14; 10.28; 13.46)

6 nodos en (0.86; 4.00; 7.14; 10.28; 13.46; 16.6)

12 nodos en (0.86; 2.43; 4.00; 5.56; 7.14; 8.71; 10.28; 11.85; 13.46; 14.9; 16.6; 18.13)

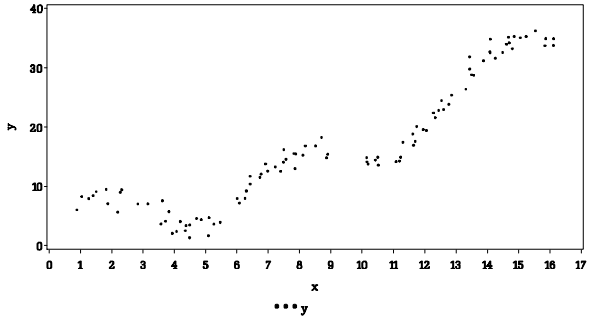
Luego se calculan los intervalos de confianza teniendo en cuenta y no teniendo en cuenta la aleatoriedad de u. Es decir se obtienen intervalos para $f(x/u)$ donde no se tiene en cuenta el sesgo y para $f(x)$ donde sí se lo tiene en cuenta.

Resultados

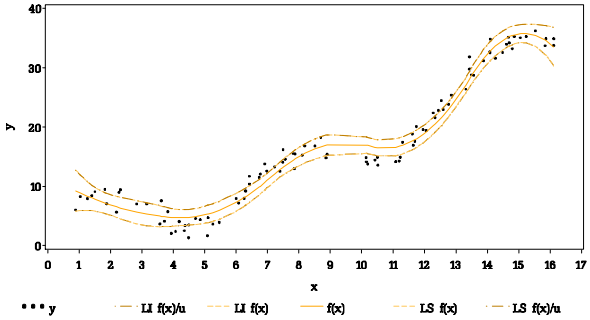
Es necesario notar que, en este trabajo, se presenta una sola simulación de cada modelo y ajuste ya que el objetivo es mostrar simplemente cada situación. A continuación se grafica cada uno de los escenarios y ajustes postulados.

Caso 1: $\sigma_r^2 = 1$

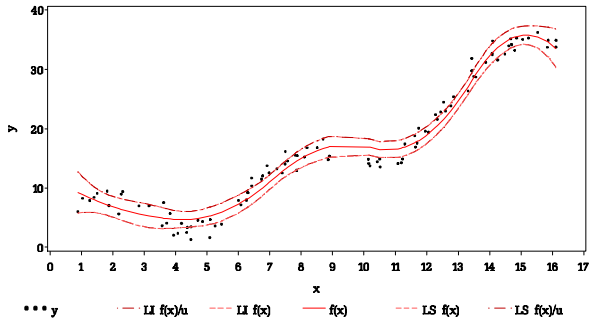
Gráfico de dispersión



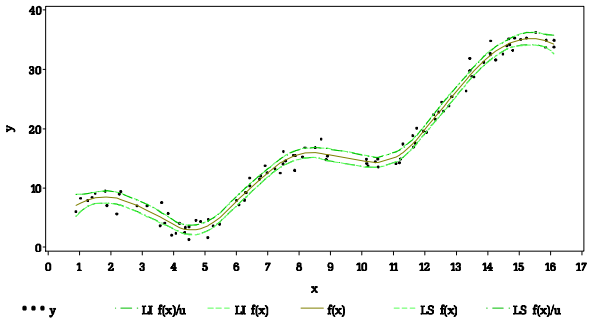
4 nodos



6 nodos

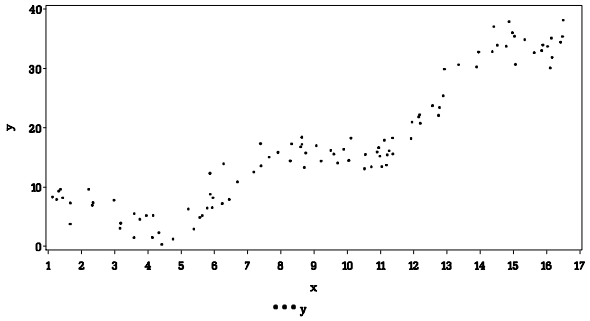


12 nodos

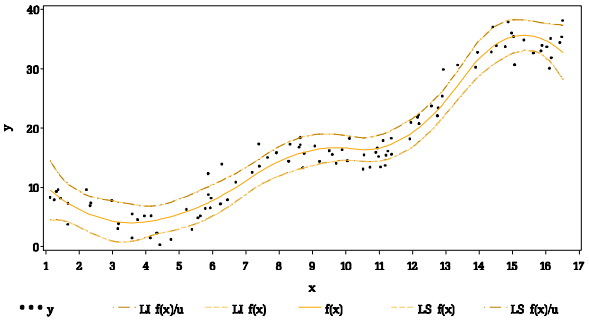


Caso 2: $\sigma_r^2 = 4$

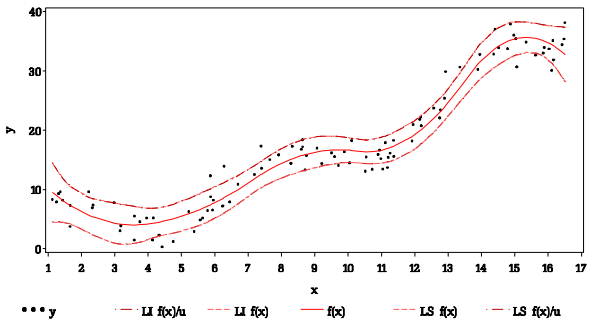
Gráfico de dispersión



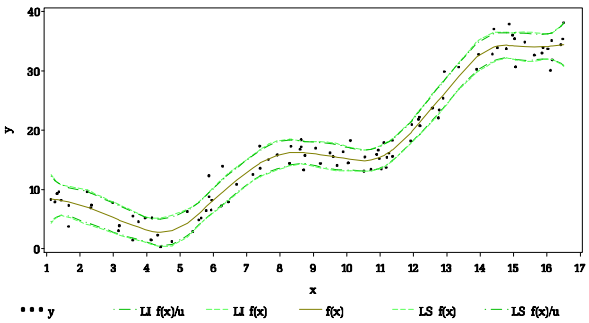
4 nodos



6 nodos

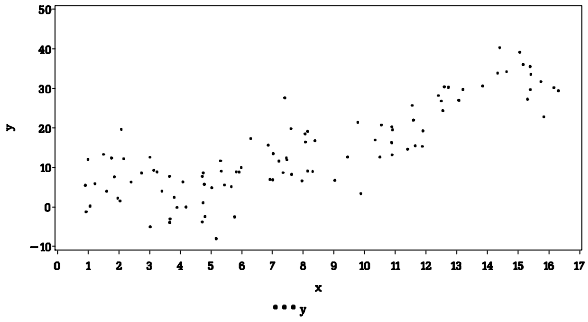


12 nodos

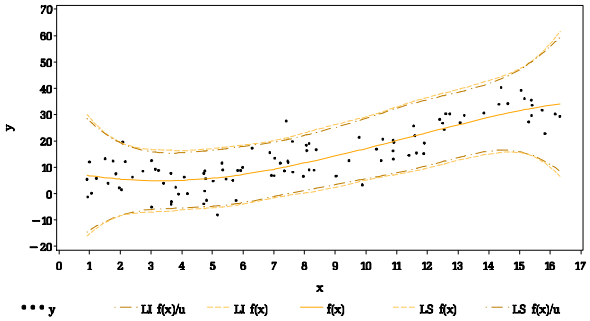


Caso 3: $\sigma_r^2 = 25$

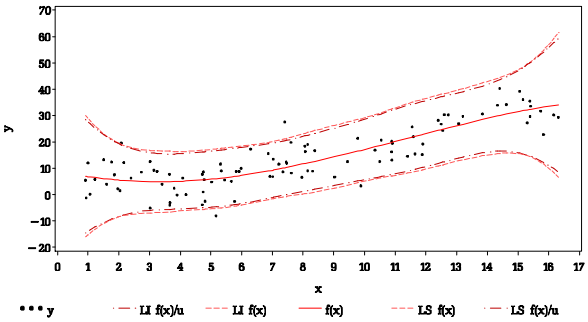
Gráfico de dispersión



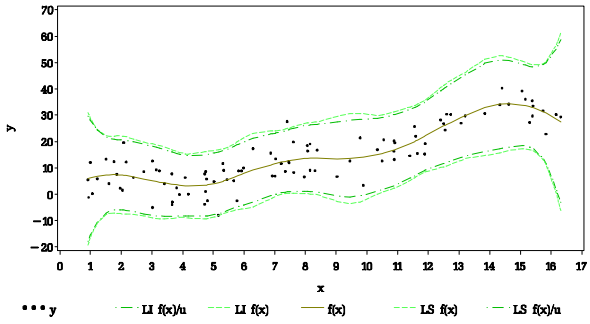
4 nodos



6 nodos



12 nodos



Las diferencias entre ambos intervalos son poco notorias en los casos $\sigma_r^2 = 1$ y $\sigma_r^2 = 4$. En el caso $\sigma_r^2 = 25$ las diferencias se pueden visualizar y son más importantes cuando se ajusta un modelo con más cantidad de nodos.

A fin de poder dimensionar numéricamente las diferencias entre ambos intervalos se calculó una estimación del sesgo debido a no tener en cuenta $\mathbf{u}$. Para ello se promedió de la diferencia entre las semi-amplitudes de los intervalos (obviando $z(1-\alpha/2)$ y σ_ε^2), a través de todas las observaciones. Es decir se calculó:

$$\hat{sesgo} = \frac{1}{n}\sum_{j=1}^{n}\left(\sqrt{\mathbf{C}_x(\mathbf{C}'\mathbf{C}+\frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_u^2}\mathbf{D})^{-1}\mathbf{C}'_x} - \sqrt{\mathbf{C}_x(\mathbf{C}'\mathbf{C}+\frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_u^2}\mathbf{D})^{-1}\mathbf{C}'\mathbf{C}(\mathbf{C}'\mathbf{C}+\frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_u^2}\mathbf{D})^{-1}\mathbf{C}'_x}\right)$$

Los resultados para los distintos escenarios se presentan en la Tabla 1.

Tabla1. Estimación del sesgo, variancias y parámetro de suavizado para cada uno de los escenarios y ajustes

σ_r^2	Nodos	sesgo	$\hat{\sigma}_u^2$	$\hat{\sigma}_\varepsilon^2$	$\hat{\lambda}$
1	4	0,00197	7,02	3,14	0,45
	6	0,00197	7,02	3,14	0,45
	12	0,02131	5,04	1,37	0,27
4	4	0,00272	4,85	5,35	1,10
	6	0,00272	4,85	5,35	1,10
	12	0,02274	6,36	3,67	0,58
25	4	0,03070	0,09	33,86	395,70
	6	0,03070	0,09	33,86	395,70
	12	0,04705	5,58	27,43	4,91

Comparando los distintos escenarios se observa, tanto en los gráficos como en la Tabla 1 que:

* con 4 y 6 nodos no hay diferencias en todos los casos las estimaciones con 12 nodos tienen un sesgo mayor y menor variancia,

* a medida que aumenta la variabilidad ven los datos, la estimación del sesgo es mayor,

* el modelo teórico es fácilmente descripto con pocos nodos cuando la variabilidad de los datos es baja. Sin embargo, si la variable respuesta presenta una gran variabilidad, son necesarios más nodos para poder representar a relación subyacente,

* tal como es de esperar, a medida que $\lambda \to \infty$ la curva suavizada tiende a ser una recta,

* a pesar de que gráficamente no se notan mayores diferencias entre los dos tipos de intervalos en los casos con $\sigma_r^2 = 1$ y $\sigma_r^2 = 4$, sin embargo, el sesgo está presente y su cuantificación se observa en la Tabla1. Las mayores diferencias se observan en el caso de mayor variabilidad en los datos y mayor cantidad de nodos,

* en todos los casos se observa que a mayor sesgo, menor variancia σ_ε^2 ya que gran parte de la variabilidad está "captada" por la variabilidad asociada a los nodos (σ_u^2).

Consideraciones finales

Las regresiones spline penalizadas son herramientas muy eficaces para describir la relación entre dos variables, su utilidad se remarca en los casos en que la naturaleza de la verdadera relación entre las variables no se puede describir con un modelo de regresión simple. Cuando este método de suavizado se circunscribe a los modelos mixtos, se pueden aprovechar muchas ventajas de estos últimos, sin embargo es necesario ser cautelosos al momento de realizar estimaciones por intervalos ya que debe tenerse en cuenta el sesgo provocado por considerar aleatorios a los parámetros asociados a los nodos. Este hecho no siempre se respeta y en este trabajo se mostraron algunas situaciones donde el sesgo puede ser mas o menos importante.

En este trabajo se mostró una única simulación para cada escenario. Debido a la actualidad y utilidad, este tema merece seguir siendo estudiado y por ello se está trabajando para producir más simulaciones de cada escenario y así obtener estimaciones más estables acerca del sesgo planteado.

Asimismo es importante remarcar que en este trabajo no se tuvo en cuenta la variabilidad causada por la estimación de las componentes de variancia ya que se espera que para muestras grandes esta variabilidad sea despreciable.

Finalmente vale aclarar que los intervalos de confianza construidos son individuales y no pretenden ser mostrados como bandas de confianza ya que en ese caso deberían tenerse en cuenta las distribuciones conjuntas. Por lo que no es correcto sacar conclusiones acerca de la forma de la relación en función de estos intervalos presentados.

Referencias Bibliográficas

Bowman, A. and Azzalini, A. (1997). Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustration. Oxford, UK: University Press

Demidenko E. (2004). Mixed Models. Theory and Applications. New York: Wiley

Eilers, P.H., Marx B.D. (1996).Flexible Smoothing with B-Splines and Penalties. Statistics Science, 11,89-102

Eubank, R. (1999). Nonparametric Regression and Spline Smoothing. New York. Marcel Dekker.

Fox, J. (2000). Nonparametric Simple Regression: Smoothing Scaterplots. Thousand Oaks, CA: Sage

Gurrin L., Scurrah K. Hazelton M. (2005). Tutorial in biostatistics : Spline Smoothing with Linear Mixed Models. Statistics in Medicine 24:3361-3381

Härdle, W. (1991). Smoothing Techniques, with Implementation in S. New York. Springer - Verlag

Littell R., Milliken G., Stroup W., Wolfinger R. (1996). SAS System for Mixed Models. Cary, NC:SAS Institute Inc.

McCulloch, C. E.; Searle S. R. (2001). Generalized, Linear and Mixed models. New York: Wiley

Ngo L., Wand M. (2004). Smoothing with Mixed Models Software. Journal of Statistical Software Volume 09, Issue 01. URL: http:/www.jstatsoft.org/

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., Carroll R. (2003). Semiparametric Regression. Cambridge University Press

Searle, S.R.; Casella, G.; McCulloch, C.E. (1992). Variance Component. New York: Wiley

Simonoff, J.S. (1996). Smoothing Methods in Statistics. New York. Springer -Verlag

Smith P. (1979). Spline as a Useful and Convenient Statistical Tool. The American Statistician Volume 33, No 2.