Gac Sanit. 2017;31(3):235–237

Nota metodológica

Multivariate Adaptative Regression Splines (MARS), una alternativa para el análisis de series de tiempo

Jairo Vanegas [a,*] y Fabián Vásquez [b]

[a] *Facultad de Ciencias Médicas, Escuela de Obstetricia y Puericultura, Universidad de Santiago de Chile, Santiago de Chile, Chile*
[b] *Instituto de Nutrición y Tecnología de los Alimentos (INTA), Universidad de Chile, Santiago de Chile, Chile*

INFORMACIÓN DEL ARTÍCULO

Historia del artículo:
Recibido el 19 de marzo de 2016
Aceptado el 6 de octubre de 2016
On-line el 19 de diciembre de 2016

Palabras clave:
Metodología
Estadística no paramétrica
Predicción

RESUMEN

Multivariate Adaptative Regression Splines (MARS) es un método de modelación no paramétrico que extiende el modelo lineal incorporando no linealidades e interacciones de variables. Es una herramienta flexible que automatiza la construcción de modelos de predicción, seleccionando variables relevantes, transformando las variables predictoras, tratando valores perdidos y previniendo sobreajustes mediante un autotest. También permite predecir tomando en cuenta factores estructurales que pudieran tener influencia sobre la variable respuesta, generando modelos hipotéticos. El resultado final serviría para identificar puntos de corte relevantes en series de datos. En el área de la salud es poco utilizado, por lo que se propone como una herramienta más para la evaluación de indicadores relevantes en salud pública. Para efectos demostrativos se utilizaron series de datos de mortalidad de menores de 5 años de Costa Rica en el periodo 1978-2008.

Multivariate Adaptive Regression Splines (MARS), an alternative for the analysis of time series

Keywords:
Methods
Non-parametric statistics
Forecasting

ABSTRACT

Multivariate Adaptive Regression Splines (MARS) is a non-parametric modelling method that extends the linear model, incorporating nonlinearities and interactions between variables. It is a flexible tool that automates the construction of predictive models: selecting relevant variables, transforming the predictor variables, processing missing values and preventing overshooting using a self-test. It is also able to predict, taking into account structural factors that might influence the outcome variable, thereby generating hypothetical models. The end result could identify relevant cut-off points in data series. It is rarely used in health, so it is proposed as a tool for the evaluation of relevant public health indicators. For demonstrative purposes, data series regarding the mortality of children under 5 years of age in Costa Rica were used, comprising the period 1978–2008.

Introducción

El desarrollo de un buen modelo de regresión requiere tiempo y una considerable experiencia de modelización. Sin embargo, con el advenimiento de la *Multivariate Adaptative Regression Splines* (MARS) los modelos de regresión pueden ser desarrollados de manera sistemática y automática sin la limitación de los supuestos que deben cumplir los modelos de regresión tradicionales[1].

MARS es una herramienta flexible que automatiza la construcción de modelos de predicción, permitiendo la selección de variables relevantes, la transformación de las variables predictoras, establecer las interacciones de las variables predictoras, el tratamiento de los valores perdidos y un autotest para protegerse del sobreajuste. Finalmente, puede revelar patrones y relaciones que es difícil, si no imposible, que otros métodos puedan revelar[1,2].

Esta herramienta aún es poco conocida en el ámbito de la salud y podría ser de gran utilidad para la evaluación de datos, incluyendo aquellos que se encuentran en forma agregada, como es el caso de la mayoría de los datos relacionados con la salud[3,4].

Exposición del método

MARS es un método de modelación no paramétrico que extiende el modelo lineal incorporando no linealidades e interacciones de variables. Es una generalización de la *Recursive Partitioning Regression* (RPR), que divide el espacio de las variables predictoras en diferentes subregiones[5,6].

El modelo puede escribirse como:

$$y_t = f(x_t) = \beta_0 + \sum_{i=1}^{k} \beta_i B(x_{it})$$

* Autor para correspondencia.
Correo electrónico: jairo.vanegas.l@usach.cl (J. Vanegas).

http://dx.doi.org/10.1016/j.gaceta.2016.10.003

Tabla 1
Resumen de modelos de predicción obtenidos a través de modelos MARS. Costa Rica, 1978-2008

Modelos de regresión MARS	Años	Valor observado	Valor predicho	R^2 ajustado	GCV[a]	mpab[b] (%)
Tasa = 3,74–0,11 * max(0; años–1988) + 0,31 * max(0; 1988 - años)–0,01 * max(0; 86–vacunados sarampión) + 0,0003 * max(0; PIB per cápita–3116)–0,03 * max(0; pobreza–16,7) + 0,02 * max(0; 198,86–gasto social educación per cápita)–0,08 * max(0; tasa global fecundidad–3,2) * max(0; pobreza–16,7)	2006 2007 2008	2,19 2,40 2,09	2,19 2,29 2,15	98	0,062	2

GCV: criterio de validación cruzada; mpab: proporción media de error absoluto; PIB: producto interior bruto.
Fuente: elaboración propia a partir de datos de CEPAL, 1978-2008.

donde y_t es la variable respuesta en el instante t y β_i son los parámetros del modelo para las respectivas variables x_{it}, que van de $i = 1, \ldots, k$. El valor β_0 representa al intercepto, las funciones bases $B(x_{it})$ son funciones que dependen de las respectivas variables x_{it}, en donde cada $B(x_{it})$ puede escribirse como $B(x_{it}) = max(0, x_{it} - c)$ o $B(x_{it}) = max(0, c - x_{it})$, c es un valor umbral y k representa el número de explicativas, que incluye interacciones de las variables predictoras. Los puntos de partición del espacio y los parámetros de los modelos se obtienen a partir de los datos analizados. El número de funciones base resultantes indica la complejidad del modelo[7].

MARS genera puntos de corte para las diferentes variables. Los puntos son identificados a través de las funciones basales, las que indican el inicio y el término de una región. En cada región en que se va dividiendo el espacio se ajusta una función base de una variable, la cual es lineal. El modelo final se constituye como una combinación de las funciones base generadas. Para determinar estos puntos de corte usa un algoritmo *forward/backward stepwise* por etapas. Primero, mediante el algoritmo *forward stepwise* se genera un modelo sobreestimado con un gran número de funciones base; posteriormente, mediante el algoritmo *backward stepwise*, se eliminan los nodos que menos contribuyen al ajuste global. El algoritmo se detiene cuando la aproximación construida incluye un número máximo de funciones fijadas por el investigador.

Cuando un investigador tiene varios modelos teóricos y se desea identificar el mejor, pueden utilizarse los siguientes estadísticos[8–10]:

- Criterio de validación cruzada (GCV), propuesto por Friedman[6], medida de ajuste a los datos y penalización, debido a la complejidad del modelo y el aumento de la varianza. Según este criterio, un modelo más sencillo puede ser preferido frente a otro más complejo.
- Coeficiente de determinación (R^2 ajustado) entre el valor observado y el predicho, el cual permite la adecuación del modelo utilizado para las predicciones.
- La proporción media de error absoluto (mpab), que viene dada por los valores observados y el valor predicho:

$$\sum_{i=1}^{n} \left\{ \left[\left| (valor\ \ observado_{-i} - valor\ predicho_{-i}) \right| / valor\ observado_{-i} \right] \right\} / n.$$

Muestra el porcentaje de error en que se incurre en la predicción en comparación con los datos observados, considerándose que el mejor modelo es aquel que tiene el porcentaje de error más bajo.

Entre los *softwares* estadísticos disponibles para aplicar MARS cabe destacar el paquete estadístico R, Matlab, Python, Salford Predictive Modeler (SPM 8), Statitstica Data Miner- StatSoft y Adaptivreg para SAS.

Aplicación práctica

Para ejemplificar el modelo MARS se utilizan las tasas de mortalidad en menores de 5 años de Costa Rica y las variables relacionadas con indicadores sociales y económicos, entre 1978 y 2008 (año, tasa global de fecundidad, partos hospitalarios, población menor de 1 año vacunada contra el sarampión, analfabetismo femenino mayor de 15 años, producto interior bruto [PIB] per cápita, gasto social en salud, gasto social en educación, agua potable, pobreza y extrema pobreza, tasa de desempleo y control prenatal); variables no estacionarias, pero cointegradas.

De la serie de 31 años (1978-2008) se extrajeron los últimos 3 años para conformar una muestra de entrenamiento y una de validación, esta última compuesta por valores observados y predichos.

La tabla 1 presenta el modelo explícito y las funciones base seleccionadas. El modelo selecciona las variables más relevantes: año, niños vacunados contra el sarampión, PIB per cápita, porcentaje de pobreza, gasto social en educación per cápita y tasa global de fecundidad. También se identifican un punto de corte que corresponde al año 1988 y una interacción entre tasa global de fecundidad y pobreza.

Las funciones base $-0{,}11 * \max(0; \text{años} - 1988) + 0{,}31 * \max(0; 1988 - \text{años})$ identifican el punto de corte en el año 1988. La interpretación sugiere que a partir de este año el comportamiento de las tasas de mortalidad en menores de 5 años se modifican en su velocidad de descenso (fig. 1).

En cada una de las siguientes funciones base se observan los valores de las variables estructurales relacionados con los años de la serie de tiempo (1978-2008).

La función $-0{,}01 * \max(0; 86 - \text{vacunados sarampión})$ identifica un punto de corte en 1996, con un reporte del 86% de inmunizados. Por debajo de este porcentaje se generarían bolsones de población susceptibles a una epidemia.

La función $0{,}02 * \max(0; 198{,}86 - \text{gasto social en educación per cápita})$ establece un punto de corte en 1994 correspondiente a US$ 198 per cápita. Valores superiores al punto de corte tienen efectos positivos en la reducción de la tasa de mortalidad.

La función $0{,}0003 * \max(0; \text{PIB per cápita} - 3116)$ presenta un punto de corte en US$ 3116 dólares per cápita reportado para el

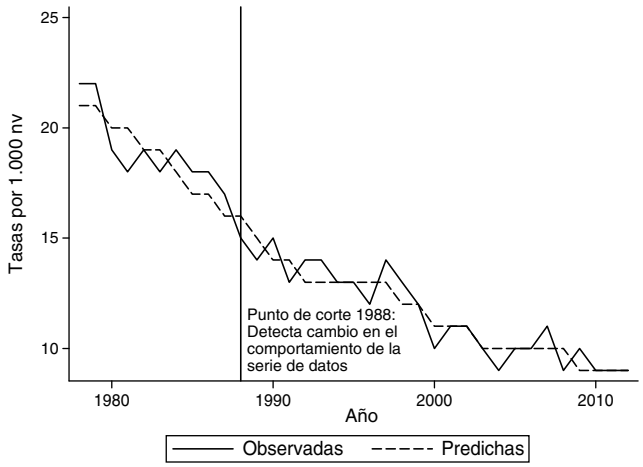

Figura 1. Tasa de mortalidad en menores de 5 años. Tasas observadas, predichas y punto de corte detectado con MARS. Costa Rica, 1978-2008.

año 1991. Anterior al punto de corte, el PIB per cápita tiene poco efecto sobre la tasa de mortalidad.

Finalmente, la interacción −0,08 * max(0; tasa global fecundidad − 3,2) * max(0; pobreza − 16,7), sugiere efectos sobre la reducción de las tasas de mortalidad entre los años 1990 y 2007 con 3,2 hijos y un 16,7% de pobreza. En este periodo se inicia una reducción de las tasas de fecundidad en un contexto de reducción del porcentaje de la pobreza. La tasa de mortalidad pasa de 16,2 a 10 por 1000 nacidos vivos.

Conclusiones

MARS permite automatizar los aspectos de modelación de la regresión clásica, seleccionando las variables predictoras, estimando los valores perdidos, transformando variables, detectando interacciones, contrastando y asegurando la correcta construcción del modelo, y permitiendo resultados más exactos y completos.

En el ejemplo, MARS permite configurar modelos hipotéticos y escenarios de carácter predictivos tomando en cuenta los factores estructurales que pudieran estar influyendo sobre la velocidad de descenso de la tasa de mortalidad. El resultado final permitiría identificar hitos relevantes, como el impacto de una política pública sobre el tiempo, deducir hasta dónde esta logra influenciar positivamente en la variable respuesta y cuándo empieza a perder influencia.

Editor responsable del artículo

Miguel Ángel Negrín Hernández.

Declaración de transparencia

El autor principal (garante responsable del manuscrito) afirma que este manuscrito es un reporte honesto, preciso y transparente del estudio que se remite a Gaceta Sanitaria, que no se han omitido aspectos importantes del estudio, y que las discrepancias del estudio según lo previsto (y, si son relevantes, registradas) se han explicado.

Contribuciones de autoría

J. Vanegas: investigador principal, recolección de datos, análisis de datos y redacción del documento. F. Vásquez: coinvestigador, contribución en el análisis de los datos, corrección del manuscrito y redacción final.

Financiación

Comisión Nacional Científica y Tecnológica (CONICYT). Programa de inserción de capital humano avanzado. Proyectos N° 791220020.

Programa FONDECYT de Postdoctorado. Proyecto 3140344.

Conflicto de intereses

Ninguno.

Bibliografía

1. Alvarado S, Silva C, Cáceres D. Modelación de episodios críticos de contaminación por material particulado (PM10) en Santiago de Chile. Comparación de la eficiencia predictiva de los modelos paramétricos y no paramétricos. Gac Sanit. 2010;24:466–72.
2. Bonilla M, Olmeda I, Puertas R. Modelos paramétricos y no paramétricos en problemas de credit scoring. Rev Esp Financ Contab. 2003;118:833–69.
3. Hawkley LC, Hughes ME, Waite LJ, et al. From social structural factors to perceptions of relationship quality and loneliness: the Chicago Health, Aging, and Social Relations Study. J Gerontol B Psychol Sci Soc Sci. 2008;63B:S375–84.
4. Bello L, Martínez S. Una metodología de series de tiempo para el área de la salud; caso práctico. Rev Fac Nac Salud Pública. 2007;25:118–22.
5. Montero R. Variables no estacionarias y cointegración. Documentos de Trabajo en Economía Aplicada. España: Universidad de Granada; 2013.
6. Friedman J. Multivariate Adaptive Regression Splines. Ann Statist. 1991;19:1–141.
7. Salford Systems Multivariate Adaptive Regression Splines (MARS): user guide. Chapter 3. MARS Basics - Smoothing, splines and knot selection; 2001. p. 9-34.
8. Lewis P, Stevens J. No linear modeling of time series using multivariante adaptive regression splines (MARS). J Am Stat Assoc. 1991;86:1–36.
9. Lin C, Chen F, Lee S. Forecasting tourism demand using time series, artificial neural networks and multivariate adaptive regression splines: evidence from Taiwan. Int J Bus Adm Manag Res. 2011;2:14–24.
10. Silva C, Alvarado S, Montaño R, et al. Modelamiento de la contaminación atmosférica por partículas: comparación de cuatro procedimientos predictivos en Santiago. Chile. Biomatemática XIII. 2003:113–27.