Utilización de métodos no paramétricos para el control de variables de confusión no observadas en estudios ecológicos de series temporales

Sr. Director:

En la mayoría de los diseños epidemiológicos para controlar la confusión es necesario que las variables confusoras sean observables y cuantificadas. No controlar la confusión conduce a un importante error de especificación, la omisión de variables explicativas relevantes, causando sobredispersión, por la que las varianzas reales son mayores que las teóricas. Cuando la respuesta es normal, el efecto de la sobredispersión es el de sesgar los errores estándar de los estimadores. Afortunadamente, este sesgo puede corregirse mediante la aproximación propuesta por McCullagh y Nelder[1]. En epidemiología, sin embargo, dominan los análisis en los que la variable dependiente no sigue una distribución normal: regresiones de Poisson, logísticas o logísticas para datos censurados. En estos casos no sólo los errores estándar, sino también las estimaciones de los parámetros de interés —el riesgo relativo (RR), por ejemplo— resultan sesgadas.

La gran ventaja de los diseños de series temporales en estudios ecológicos es que la variable tiempo y sus transformadas (cuadráticas y funciones sinusoidales [seno o coseno] de distinta frecuencia) pueden ser utilizadas para controlar el efecto que tienen los confusores no observables (como la concentración de polen) sobre la variable dependiente (mortalidad o morbilidad). Estas transformaciones de la variable tiempo tienen el objetivo de «recoger» el efecto que aquellas variables no observadas tienen sobre la variable dependiente. Sin embargo, si algunas de estas varialbes no observadas siguen un componente cíclico de frecuencia y amplitud variable (como podría ser la concentración de polen), las funciones paramétricas del tiempo o de sus transformadas sinusoidales no pueden «adaptarse» a estos cambios. Como consecuencia se produce «confusión residual»[2] que puede conducir a estimaciones sesgadas en la estimación del riesgo relativo (RR) del factor de riesgo[2] (el contaminante o la variable meteorológica).

Analizado desde el punto de vista estadístico, si las transformadas del tiempo no son capaces de captar el componente cíclico de las variables observadas, quedará una varianza no explicada que se correlaciona con alguno de los regresores (en este caso con el tiempo pues se trata de una regresión dinámica). Para paliar estos errores de especificación, se recurre a la introducción de términos autorregresivos, pero no siempre son capaces de eliminarlos.

Estas limitaciones del análisis clásico de los estudios ecológicos de series temporales se pueden minimizar aplicando modelos aditivos generalizados (GAM)[3] que utilizan funciones suaves que se estiman localmente[4]. En esta moderna regresión, el investigador no asume ninguna relación paramétrica previa entre la variable dependiente y la independiente («Let the data show us the appropriate functional form»[3]). Esto es particularmente útil para controlar el efecto de variables no medidas. Así, si utilizamos funciones no paramétricas de la variable tiempo[5,6], estas funciones se «adaptarán» de una forma flexible a los componentes cíclicos de las variables no medidas, sin importar los cambios de frecuencia o de amplitud a lo largo del período estudiado.

Aunque en los modelos GAM el investigador no debe de especificar la función a la que han de adaptarse los datos, sí necesita especificar «a priori» el número de grados de libertad (gl) de la función no paramétrica suavizada. Los gl determinan el grado de flexibilidad de la función: a mayor gl mayor flexibilidad. Simplificando, los gl podrían ser considerados como el número de tramos en los que se divide la variable explicativa[6]. Así, un gl implicaría una función lineal y cuatro gl la consideración de cuatro tramos. El problema surge en cómo determinar el número de gl de la variable tiempo. Si el número de gl es demasiado bajo, se puede producir varianza no explicada que dependa del regresor tiempo (sobredispersión) y, por tanto, existe riesgo de estimaciones sesgadas del RR. Si el número de gl es demasiado elevado, se producirán estimaciones ineficientes (aumenta el error estándar) de los RR, al aumentar el número de parámetros a estimar. Algunos autores[5] proponen un número de gl aproximadamente igual al número de meses del período estudiado. También se han propuesto procedimientos automáticos de la selección del número de gl basándose en la minimización del error de predicción teórico aplicando criterios como el Akaike Information Criteria (AIC). Sin embargo, recientes estudios de simulación han concluido que la aplicación del AIC a modelos no paramétricos tiende a sobreestimar los gl, y se ha propuesto un nuevo AIC adaptado a los modelos no paramétricos, aunque hasta ahora sólo se ha validado en respuesta gaussiana[7]. Queda por determinar su validez en modelos de regresión de la familia poisson.

Finalmente, aún con la aplicación de métodos no paramétricos puede existir cierta sobredispersión[5]. Esto se debe a que si las variables no medidas tienen una periodicidad inferior al cociente entre el número de días de la serie y el número de gl, éstas no pueden ser «captadas» por la función no paramétrica de la variable tiempo. En este caso, utilizar términos autorregresivos de la variable dependiente permite disminuir la sobredispersión[8].

En resumen, la utilización de métodos no paramétricos

tipo GAM está siendo cada vez más frecuente en el campo de la epidemiología, debido a los avances en el software, a su mayor flexibilidad y parsimonia, con respecto a las aproximaciones paramétricas y, sobre todo, a la mayor posibilidad de control de la confusión.

A. Figueiras[1], M. Saez[2] y A. Tobias[3]
[1]Área de Medicina Preventiva y Salud Pública. Universidad de Santiago.
[2]Departament d'Economia, Universitat de Girona.
[3]Unitat de Recerca Respiratòria i Ambiental. Institut Municipal d'Investigació Médica (IMIM). Barcelona.

Bibliografía

1. McCullagh P, Nelder JA. Generalized Linear Models. New York: Chapman and Hall, 1989.
2. Brenner H, Blettner M. Controlling for continuous confounders in epidemiologic research. Epidemiology 1997;8:429-34.
3. Hastie TJ, Tibshirani RJ. Generalized additive models. New York. NY: Chapman and Hall, 1990.
4. Sánchez-Cantalejo E, Ocaña-Riola R. Actualizaciones en regresión: suavizando las relaciones. Gac Sanit 1998;12:223-8.
5. Kelsall JE, Samet JM, Zeger SL, Xu J. Air pollution and mortality in Philadelphia, 1974-1988. Am J Epidemiol 1997;146:750-62.
6. Tobías A, Sunyer J, Samoli E, Katsouyanni K. Contaminación ambiental y mortalidad. Análisis de sensibilidad. Gac Sanit 1999; 13:73.
7. Hurvich CM, Simonoff JS, Tsai CL. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. J R Stat Society 1998; Series B, 60: 271-93.
8. Saez M, Figueiras A. Re: Air pollution and mortality in Philadelphia, 1974-1988. Am J Epidemiol 1999 (en prensa).