Regresión ecológica mediante modelos jerárquicos bayesianos con valores ausentes en la covariable

C. Abellán[1] y A. López-Quílez[2]
[1]Àrea d'Epidemiologia. Direcció General de Salut Pública. Conselleria de Sanitat. Generalitat Valenciana
[2]Departament d'Estadística i Investigació Operativa. Universitat de València

Introducción

Este trabajo se enmarca en el creciente interés por el estudio de la repercusión en la Salud Pública de los problemas medioambientales. La regresión ecológica se ha convertido en una herramienta esencial para llevar a cabo este tipo de análisis ya que permite la inclusión de factores de riesgo como covariables así como la de componentes aleatorias que modelicen posibles factores ocultos.

Sin embargo, la resolución de este tipo de problemas resulta muy compleja empleando procedimientos habituales cuando tenemos valores ausentes en la covariable.

Nuestro objetivo es abordar el problema de trabajar con datos incompletos en el marco de los modelos lineales generalizados mixtos ayudándonos de los avances en análisis bayesiano vía métodos Markov Chain Monte Carlo (MCMC).

Regresión ecológica espacial

Recientes desarrollos en Estadística Espacial permiten obtener mejores resultados en el estudio de la correlación entre factores de riesgo e indicadores de salud. Modelos estadísticos como la regresión de Poisson así como otros modelos lineales generalizados (McCullagh y Nelder, 1989) han contribuido mucho a este desarrollo incluyendo la información que sobre los factores de riesgo nos proporcionan las unidades próximas geográficamente. Es más, la inclusión de factores aleatorios en el modelo lineal generalizado nos permite tener en cuenta la sobredispersión que generalmente aparece en datos geográficos de salud.

En particular, dentro de los análisis estadísticos espaciales los relacionados con el efecto de factores medioambientales y más concretamente el estudio de los posibles efectos que sobre la salud tiene la calidad del agua potable han sido importantes últimamente debido a su gran interés desde el punto de vista social y de la salud pública (Ferrándiz et al., 1999, 2003).

Nuestro interés se centra en analizar la posible relación entre las concentraciones

de nitrato presentes en el agua potable y la mortalidad por cáncer de estómago en los municipios de la Comunidad Valenciana teniendo en cuenta el patrón geográfico.

Al abordar un estudio de estas características, es lógico pensar en las observaciones en cada unidad geográfica como datos correlados debido a la existencia de factores ocultos. Además, podemos considerar el origen de parte de estos factores ocultos en fenómenos medioambientales, lo que produciría observaciones correladas en unidades geográficas vecinas.

Estos factores ocultos suelen ser modelizados con dos efectos aleatorios, heterogeneidad y espacial (Besag et al., 1991). La componente de heterogeneidad permite ajustar la influencia de factores de riesgo ocultos locales independientemente en cada unidad geográfica. Esto se consigue generalmente considerando cada componente como una observación aleatoria de una distribución normal debidamente centrada.

La componente espacial recoge factores de riesgo geográficos cuya influencia afecta no solo a una única unidad geográfica o que posee una suave variación geográfica mostrando niveles de riesgo similares en observaciones vecinas. Este efecto espacial es generalmente modelizado mediante una distribución condicional autoregresiva normal, denominada CARNormal (Besag y Kooperberg, 1995) que permite introducir la relación de vecindad entre observaciones definiendo la distribución de cada observación condicional al resto de observaciones como la distribución condicionada únicamente a las observaciones vecinas. El empleo de la distribución CAR para modelizar la similitud entre observaciones próximas geográficamente es habitual en el contexto de la suavización de razones de mortalidad (Ferrándiz et al., 2002).

La perspectiva bayesiana nos permite modelizar este tipo de estudios geográficos de un modo sencillo y natural mediante el empleo de modelos jerárquicos. Además recurriendo a métodos MCMC evitamos las dificultades derivadas de intentar obtener la distribución posterior exacta de los parámetros de interés. Este procedimiento es utilizado con frecuencia en diferentes ámbitos permitiendo realizar una amplia variedad de estudios como análisis espacio temporales de riesgos de enfermedad (Bernardinelli et al., 1995; Waller et al., 1997) o análisis de regresión con covariables (Ferrándiz et al., 2003).

Valores ausentes

Sin embargo, en nuestro estudio el principal problema es que los datos de concentraciones de nitratos están incompletos debido a la existencia de valores ausentes. Este problema es bastante común en datos medioambientales y provoca que sea muy complicado llevar a cabo análisis con estas covariables. Además las técnicas estadísticas diseñadas para trabajar con valores ausentes suelen ser complejas, muy costosas a nivel computacional y específicas de cada aplicación. Por tanto, para facilitar el análisis mediante las técnicas estadísticas habituales incluidas en los paquetes esta-

dísticos comunes, es preferible completar la covariable estimando los valores ausentes, es decir, imputándolos. La imputación de datos se convierte en necesaria cuando la covariable va a ser utilizada en diferentes ámbitos de análisis y para diferentes propósitos. Además, generalmente la fuente de los datos y el usuario de éstos suelen estar en diferentes entidades lo que reafirma aún más la necesidad de imputar la información ausente (Rubin, 1987).

Datos

Como hemos mencionado antes, la calidad del agua potable es un importante factor de vigilancia en salud pública y, por tanto, es regularmente analizado por la Agencia de Medioambiente de la Comunidad Valenciana desde 1991. Concretamente, se han medido anualmente las concentraciones de nitratos en cada uno de los 540 municipios de la Comunidad Valenciana durante el periodo 1991-2000.

Estos datos presentan un alto porcentaje de valores ausentes (alrededor del 30 %), haciendo la imputación necesaria.

Por otro lado, la mortalidad por cáncer de estómago en los municipios de la Comunidad Valenciana ha sido registrada anualmente para el periodo 1991-2000 y proporcionada por el Registro de Mortalidad de la Comunidad Valenciana.

Modelización

El hecho de que los datos de concentraciones de nitratos haya sido registrado por municipio nos induce a pensar en una posible correlación geográfica. Por tanto, en el modelo de imputación, incluimos una componente espacial aleatoria para cada municipio que recoja la relación espacial entre observaciones vecinas. Esta componente además presenta la ventaja de proporcionar imputaciones fiables para aquellos municipios vecinos con un alto porcentaje de valores ausentes a lo largo del periodo de estudio.

Además de esta correlación espacial, como los datos han sido medidos anualmente en el periodo de 1991 a 2000, cabe pensar en una tendencia temporal independiente para cada municipio en las concentraciones de nitratos. Debido a esto, incluimos una componente temporal aleatoria diferente para cada municipio y año con la intención de ajustar dicha tendencia temporal.

Como el principal interés está en el estudio de cómo afectan las concentraciones de nitratos a la mortalidad por cáncer de estómago mediante una regresión espacial, incluimos la imputación de la covariable en el propio estudio. Es decir, planteamos un modelo jerárquico bayesiano que incluye en un primer nivel las concentraciones de nitratos como covariable y las componentes de heterogeneidad y espacial mencionadas antes, teniendo en cuenta la correlación espacial de los datos de mortalidad. En un

segundo nivel modelizamos la imputación de la covariable introduciendo la estructura espaciotemporal descrita. En los siguientes niveles especificamos las distribuciones de las componentes que forman parte de la modelización así como las distribuciones previas de los parámetros del modelo.

Procediendo de esta forma en lugar de imputar la covariable primero y realizar la regresión ecológica después, conseguimos no perder la variabilidad de la imputación incluyéndola en el estudio de la relación entre las concentraciones de nitratos y la mortalidad por cancer de estómago.

Resultados

Con esta compleja modelización es muy difícil trabajar a nivel analítico. Sin embargo, su análisis es posible desde la perspectiva bayesiana empleando métodos de simulación basados en técnicas MCMC para obtener las distribuciones finales de las cantidades desconocidas: los parámetros del modelo y las estimaciones de los valores ausentes.

El procedimiento de simulación consiste, de forma resumida, en simular primero observaciones de la covariable para aquellos valores que están ausentes (imputación), con esto se completa la covariable, y después simular valores de los parámetros de la regresión.

Para el proceso de simulación hemos hecho uso del software WinBUGS. Este programa de libre distribución implementa técnicas de simulación mediante Gibbs Sampling (Green, 2001) y facilita el análisis bayesiano, teniendo las cautelas necesarias.

Como resultados preliminares, encontramos que el modelo de imputación ajusta adecuadamente la estructura espacio temporal de los datos de concentración de nitratos, proporcionando mejores imputaciones que otros modelos con los que hemos probado. Además, a pesar de ser difícil detectar relación de factores de riesgo medioambientales en estudios de salud debido a que los efectos que éstos tienen sobre la salud son generalmente muy débiles, observamos mayor riesgo en aquellas zonas con valores altos de concentración de nitratos.

Agradecimientos

Este trabajo ha sido parcialmente financiado por la Dirección General de Salud Pública de la Generalitat Valenciana mediante un acuerdo de colaboración con la Universitat de València.

Referencias

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C. y Ghislandi, M. (1995). Bayesian analysis of space-time variation in disease risk. *Statistics in Medicine*

14:2433–2443.

Besag, J. y Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* 82:733–746.

Besag, J., York, J. y Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* 43:1–59.

Ferrándiz, J., Abellán, J., López, A., Sanmartín, P., Vanaclocha, H., Zurriaga, O., Martínez-Beneito, M., Melchor, I. y Calabuig, J. (2002). Geographical distribution of the cardiovascular mortality in Comunidad Valenciana (Spain). D. Briggs, P. Forer, L. Jarup y R. Stern. (eds.), *GIS for Emergency Preparedness and Health Risk Reduction*, 267–282. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Ferrándiz, J., López, A., Gómez-Rubio, V., Sanmartín, P., Martínez-Beneito, M. A., Melchor, I., Vanaclocha, H., Zurriaga, O., Ballester, F., Gil, J. M., Pérez-Hoyos, S. y Abellán, J. J. (2003). Statistical relationship between hardness of drinking water and cerebrovascular mortality in Valencia: a comparison of spatiotemporal models. *Environmetrics* 14:491–510.

Ferrándiz, J., López, A. y Sanmartín, P. (1999). Spatial regression models in epidemiological studies. A. Lawson, A. Biggeri, D. Böhning, E. Lesaffre, J. Viel y R. Bertolini (eds.), *Disease Mapping and Risk Assessment for Public Health*, 203–215. Chichester, U.K.: Wiley.

Green, P. (2001). A primer of Markov Chain Monte Carlo. O. E. Barndoff-Nielsen, D. R. Cox y C. Klüppelberg (eds.), *Complex Stochastic Systems*, 1–62. London: Chapman and Hall / CRC.

McCullagh, P. y Nelder, J. (1989). *Generalized linear models*. London: Chapman and Hall, Second Edition.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Waller, L. A., Carlin, B. P., Xia, H. y Gelfand, A. E. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association* 92:607–617.