Annals of Epidemiology 26 (2016) 43–49

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

Original article

Bayesian model selection methods in modeling small area colon cancer incidence

Rachel Carroll PhD [a,*], Andrew B. Lawson PhD [a], Christel Faes PhD [b], Russell S. Kirby PhD, MS, FACE [c], Mehreteab Aregay PhD [a], Kevin Watjou MS [b]

[a] Department of Public Health, Medical University of South Carolina, Charleston
[b] Interuniversity Institute for Statistics and Statistical Bioinformatics, Hasselt University, Diepenbeek, Belgium
[c] Department of Community and Family Health, University of South Florida, Tampa

ARTICLE INFO

Article history:
Received 14 July 2015
Accepted 25 October 2015
Available online 14 November 2015

Keywords:
Colon cancer
Bayesian model averaging
Bayesian model selection
Spatial regression
MCMC

ABSTRACT

Purpose: Many types of cancer have an underlying spatial incidence distribution. Spatial model selection methods can be useful when determining the linear predictor that best describes incidence outcomes.
Methods: In this article, we examine the applications and benefits of using two different types of spatial model selection techniques, Bayesian model selection and Bayesian model averaging, in relation to colon cancer incidence in the state of Georgia, United States.
Results: Both methods produce useful results that lead to the determination that median household income and percent African American population are important predictors of colon cancer incidence in the Northern counties of the state, whereas percent persons below poverty level and percent African American population are important in the Southern counties.
Conclusions: Of the two presented methods, Bayesian model selection appears to provide more succinct results, but applying the two in combination offers even more useful information into the spatial preferences of the alternative linear predictors.

Introduction

Colon cancer (*International Classification of Diseases, Ninth Revision, Clinical Modification* code: 153), accompanied by rectum cancer (*International Classification of Diseases, Ninth Revision, Clinical Modification* code: 154.1), is ranked as the third most common tumor type in the United States, with colon cancer being the more frequent of the two. Routine screening for this cancer, particularly after the age of 50 years, is encouraged because a good prognosis typically accompanies an early diagnosis. Important risk factors of colon cancer include nutritional inclinations, age, smoking status, inflammatory bowel diseases, previous incidence of malignant disease, and some genetic traits [1–3]. Research examining the geography of some of these risk factors suggests that there may be an underlying spatial structure to the incidence of colon cancer [4,5].

The data of interest in this study are the 2003 colon cancer incidence for the 159 counties in the state of Georgia, United States. The Area Health Resource Files [6] data set provides ecological predictors useful for explaining the variation in this outcome. The chosen predictors are as follows: median household income (in thousands of dollars), percent persons below poverty level (PPBPL), unemployment rate of those aged 16 years or greater (UER), and percent African American (AA) population. Other studies indicate that poverty and race are associated with colon cancer incidence [1,7]. Of the chosen variables, there is evidence to suggest that median income and PPBPL may be correlated (see section titled Data and Linear Predictor Alternatives). This same evidence could also be an indicator of the underlying spatial effect that we believe may play a role in colon cancer incidence. The age cutoff associated with the unemployment variable may be criticized as much of the younger population in this age range may not hold steady jobs as they are full-time students. In the individual level data used to create this county-level variable, “student” is an option as an employment status.

Selecting appropriate linear predictors is one of the most important aspects of data analysis, and this can become very challenging when spatial structures are present in the data. Many methods, such as variable selection, transformation selection, model selection, model averaging, and other model uncertainty

* Corresponding author. Department of Public Health, Medical University of South Carolina, 135 Cannon St, Charleston, SC 29425. Tel.: +1 8433197125.
E-mail address: mosra@musc.edu (R. Carroll).

http://dx.doi.org/10.1016/j.annepidem.2015.10.011
1047-2797/

methods, have been proposed and explored to achieve these goals [8–12]. In this article, we discuss the application of two types of spatial model selection techniques, Bayesian model selection (BMS) and Bayesian model averaging (BMA) [12–14], in modeling small area cancer incidence. This is achieved by assigning prior probability distributions to each of the possible linear predictors. For BMS, we simply choose the linear predictor associated with the largest posterior probability as the true model. This type of inference works well when a single model stands out, but if that is not the case, BMA is a more appropriate alternative method that can produce a model that blends the alternative linear predictors. In the BMA method, an average posterior mean and variance are calculated based on the posterior model probabilities. However, this posterior mean and variance can be quite difficult to interpret [15]. An additional statistical issue involving these types of models revolves around the correlated spatial effect, and there have been several studies examining the issues related to this [16–18]. Our models, however, do not involve the correlated spatial random effect in this same way. Rather than using the effect as add additive component in the separate linear predictors, we only use this element as a structure within the model weights, and probabilities produced with the model selection techniques.

This article is developed as follows. First, we describe the available data and the linear predictors of interest. Second, we explain the BMS and BMA methods to be applied. Next, we display the results of using these methods to the colon cancer data using these different model selection techniques. Finally, we discuss the results and draw conclusions.

Materials and methods

Our data for this study involve measures of incidence of colon cancer for each of the 159 counties in the state of Georgia, United States and predictors from the Area Health Resource Files data set. As our outcome of interest is the incidence of colon cancer, a conditionally independent Poisson distribution is a reasonable model for these data. This is a commonly assumed model for small area counts in disease mapping [19] and is appropriate because the Poisson distribution is a discrete frequency distribution that provides the probability of events occurring in a given area.

Data and linear predictor alternatives

The colon cancer data come from the online analytical statistical information system (Oasis) of the Georgia Department of Public Health. For the 1332 diagnosed colon cancers across the state in the year 2003, there was approximately a mean incidence of 8.38 cases per county where the minimum county level value was 0 and the maximum value was 102. In these data, there are no missing values at the county level.

The geographical distributions of the chosen predictors are displayed in Figure 1 and suggest some spatial clustering. An additional indicator of the underlying spatial structure is made evident by the pattern of standardized incidence ratios displayed in Figure 2. The standardized incidence ratio is calculated as the ratio of the observed colon cancer incidences to the expected rates for each of the 159 counties and can be useful as a first step in data analysis [20]. Qualitatively, for these data, there does appear some spatial structure.

Based on the chosen predictors (median income in thousands—x_1, PPBPL—x_2, UER—x_3, and percent AA population—x_4), we have used three possible linear predictors for use with both the BMS and BMA methods. Table 1 displays these alternative predictor options. The first linear predictor (Alt1) includes all the covariates. The second (Alt2) includes only income and percent AA population. The third and final linear predictor (Alt3) includes PPBPL and percent AA population. Note that all our possible linear predictors contain an uncorrelated random effect to aid in accounting for any uncontrolled for parameters or extra noise present in the data, and they differ by the predictors included. Additionally, for all these linear predictor alternatives, the prior distributions are such that:

$$u_{id} \sim Norm(0, \tau_u),\ \tau_u \sim Gam(1, 0.5), \text{and } \alpha_{jd} \sim Norm(0, 1)$$

Where $i = 1,\ldots, 159$, $d = 1,\ldots, D$ such that D is the number of linear predictors to be selected between, and $j = 0,\ldots, J$ such that J is the number of predictors for the dth model.

We alternate income and PPBPL in the second two linear predictors because there is evidence to suggest that they may be correlated. This is not an uncommon assumption as, typically, when income is higher, poverty is lower, as shown in Table 2. This table illustrates, through individual Poisson model fits, that median income and PPBPL are collinear with respect to the incidence of colon cancer outcome because PPBPL becomes well estimated when median income is removed from the model. We also note some changes in percent AA population when PPBPL is used in place of median income. These individual model fits were performed using Bayesian approximation techniques by way of the R package INLA [21,22].

In addition to collinearity, the changes seen in the parameter estimates could also indicate that some of these predictors may be more important in certain regions of the county map. This indication will be further explored with the application of the BMS and BMA techniques. The covariates were standardized before fitting the models.

Statistical methods

In what follows, we describe the methodology associated with the BMS and BMA techniques which are implemented using the R package BRugs which calls OpenBUGS [23,24].

Bayesian model selection

To evaluate a number of alternative linear predictor models, we adopt a method which fits a variety of models, and the selection of weights allows each model to be evaluated for its appropriateness. In general, for $d = 1, \ldots, D$ models, the following structure applies:

$$y_i|\mu_i \sim Poi(\mu_i)$$

$$\mu_i = e_i\theta_i$$

$$\log(\theta_i) = \sum_d w_{id}\varphi_{id}$$

$$\text{logit}(p_{id}) \sim Norm\left(\frac{1}{n_i}\sum_{i\sim l}\text{logit}(p_{idl}), \frac{1}{n_i\tau_{id}}\right); \tau_{id}^{-1/2} \sim Unif(0, 10)$$

$$w_{id} \sim Bern(p_{id})$$

where φ_{id} is our dth model's suggested linear predictor for the ith county complimented with a possible uncorrelated random effect. In general, we write φ_{id} as $x_i^T\beta_d\psi_{dj} + u_i\psi_{d.\,J+1}$ with x_i^T, the vector of J possible covariates ($j = 1, \ldots, J$), and ψ_{dj} an indicator for if the jth predictor or random effect is to be included in the linear predictor of the dth model. Hence, for a variable not included in the dth model, ψ_{dj} would be zero, otherwise it would be one. Further, w_d is a model indicator, equal to 1 if the dth model is selected and zero otherwise. The model selection probability for the dth model in the ith county is given by the probability p_{id}. Additionally, in the