Computational Statistics and Data Analysis 56 (2012) 1256–1274

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Bayesian model selection for logistic regression models with random intercept☆

Helga Wagner*, Christine Duller

Department of Applied Statistics and Econometrics, Johannes Kepler Universität Altenbergerstrasse 69, 4040 Linz, Austria

ARTICLE INFO

Article history:
Available online 6 July 2011

Keywords:
Variable selection
Variance selection
MCMC
Auxiliary mixture sampling
Normal scale mixtures
Spike and slab priors

ABSTRACT

Data, collected to model risk of an interesting event, often have a multilevel structure as patients are clustered within larger units, e.g. clinical centers. Risk of the event is usually modeled using a logistic regression model, with a random intercept to control for heterogeneity among clusters. Model specification requires to decide which regressors have a non-negligible effect, and hence, should be included in the final model and whether risk is actually heterogeneous among centers, i.e. whether the model should include a random intercept or not. In a Bayesian approach, these questions can be answered by combining variable selection with variance selection of the random intercept. Bayesian model selection is performed for a reparameterized version of the logistic random intercept model using spike and slab priors on the parameters subject to selection. Different specifications for these priors are compared on simulated data as well as on a data set where the goal is to identify risk factors for complications after endoscopic retrograde cholangiopancreatography (ERCP).

1. Introduction

Medical studies often are carried out with the goal to identify factors affecting risk of a disease or an adverse treatment effect. In this paper, we analyze data from an Austrian benchmarking project where information on routinely applied endoscopic retrograde cholangiopancreatographies was collected in 29 Austrian centers in 2006 and 2007. ERCP is an X-ray examination of pancreatic and bile ducts using a contrast medium and a special kind of endoscope. Moreover, ERCP is often used for medical treatments, e.g. for removal of gallstones. The procedure entails risk of several complications, e.g. post-ERCP pancreatitis, cholangitis, perforation, bleeding, and very rarely even procedure-related death (Kapral et al., 2008).

The aim of our analysis was identification of patient- and procedure-related risk factors for bleeding after routinely applied ERCPs. A further issue was to assess whether risk is homogeneous among clinical centers after controlling for covariates. We model the risk of bleeding after ERCP using a logistic model with a large set of potential risk factors as covariates and include a random intercept to account for clustering of observations within centers. The data set comprises data on 3143 patients and contains one continuous covariate (age) and 36 binary covariates, which indicate the presence or absence of a factor considered to affect risk. Risk of bleeding is small as it occurs only for 118 (3.8%) patients.

Inference for this data set is challenging as incidence is rare for some of the potential risk factors, entailing rare joint incidence of bleeding and these risk factors. Even though the sample size is not particularly small, no case of bleeding was observed in the following groups of patients: patients with previous gastric surgery, patients whose indication for ERCP was pancreatic duct stone, and patients for whom the oxygenation of hemoglobin was controlled by pulse oximetry.

☆ The code used in this paper and a file describing its usage can be found as a supplementary material of the electronic version of the paper.

* Corresponding author. Tel.: +43 2468 5883; fax: +43 2468 9846.
E-mail addresses: helga.wagner@jku.at (H. Wagner), christine.duller@jku.at (C. Duller).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.06.033

Each of these three potential risk factors separates cases and non-cases. Separation, which arises when the outcome can be predicted perfectly by a linear combination of the predictors (Albert and Anderson, 1984), is a problem for binary response data as it causes monotonicity of the logistic likelihood. If separation occurs for a binary covariate, the ML estimate of the corresponding regression effect does not exist and classical variable selection procedures, e.g. based on Wald tests, fail (Heinze and Schemper, 2002; Heinze, 2006). Thus, separation can lead to the paradoxical situation that, based on standard tests, covariates which are highly correlated with the binary response are not selected in the final model. In a Bayesian approach, which we take here, separation is no problem, if the prior distributions are chosen appropriately to achieve regularization. A frequentist alternative would be inference based on penalized likelihood, see Heinze and Schemper (2002).

Bayesian variable selection methods for logistic regression models have been considered by many authors (Figueiredo, 2003; Genkin et al., 2007; Gelman et al., 2008) with the goal to identify relevant regressors. To control for heterogeneity among centers, we will include a center-specific random intercept in the logistic regression model. Combining variable selection with variance selection of the random intercept as in Tüchler (2008) and Chen and Dunson (2003) allows full model specification search to determine not only which covariates but also whether a random intercept should be included in the final model: if the variance of the random intercept is zero, risk is homogeneous among clinical centers and the model comprises only fixed effects. If, however, the random intercept variance is positive, the resulting model is a model with center-specific random intercepts, i.e. a model where risk is heterogeneous among centers.

Many Bayesian variable selection methods use spike and slab priors for the regression coefficients. These priors are mixtures of two components: a spike component concentrated around zero to allow shrinkage of small regression effects to zero and a flat slab component to prevent heavy shrinkage of larger effects. We follow Frühwirth-Schnatter and Tüchler (2008) and Tüchler (2008) in using a non-centered parameterization of the random intercepts, as this parameterization allows to specify spike and slab priors also for an appropriately defined parameter of the random intercept variance. We consider spike and slab priors with two different specifications for the spike: absolutely continuous spikes and Dirac spikes. For both types of spikes, the posterior inclusion probability of an effect is defined as the probability of belonging to the slab component. These posterior inclusion probabilities are estimated using MCMC methods, but sampling schemes differ depending on the type of the spike. We are interested in the performance of both implementations with regard to correct model selection and MCMC sampling efficiency of posterior inclusion probabilities.

With a Dirac spike, the marginal likelihoods of different models have to be computed in each MCMC iteration, requiring integration over the parameters subject to selection. As a closed formula for the marginal likelihood is available only for Gaussian and partially Gaussian models, we make use of a data augmentation scheme in Frühwirth-Schnatter and Frühwirth (2010) to obtain a representation of the original model as a Gaussian model in auxiliary variables. If the spike is specified by an absolutely continuous distribution, posterior inclusion probabilities can be computed conditional on the effects. We expect draws of the posterior probabilities for continuous spikes to show higher autocorrelation than under a Dirac spike, where the effects subject to selection are marginalized out. It is, however, not obvious which implementation will have higher computational cost in CPU time: specifying a continuous spike will save CPU time as no marginal likelihoods have to be computed; at the same time under the Dirac spike, coefficients assigned to the spike component are exactly zero and therefore only a reduced model has to be fitted in each MCMC iteration, whereas for a continuous spike, the dimension of the model is not reduced during MCMC.

The rest of the paper is structured as follows. Section 2 describes the project, where data were collected. Section 3 develops the Bayesian study model, which comprises the observation model, a logistic regression model with center-specific random intercepts, and the prior distributions for all parameters. Different spike and slab priors which can be used for selection of covariates and the random intercepts are introduced. Section 4 outlines the implementation of the MCMC sampling schemes. Simulation studies to assess the performance of the different MCMC implementations are presented in Section 5. In Section 6, the data from the ERCP study are analyzed, and finally, Section 7 summarizes the results.

2. The ERCP project

In Austria, 140 registered sites perform about 15,000 ERCP procedures per year. All of the 140 sites were invited to participate voluntarily in the nation-wide 'Benchmarking ERCP' project. 29 centers registered and reported 4846 procedures in 2006 and 2007. The data from the participating sites as well as patient data were transmitted pseudonymously, and the endoscopists remained anonymous. The participants were urged to report each performed ERCP.

The online questionnaire covered the following indicators (Kapral et al., 2008):

- identification of the endoscopic center (pseudonymous) and the endoscopist (anonymous),
- patient-related factors: indication for ERCP, gender, age, significant co-morbidities, anticoagulation,
- general information on the examination: date of ERCP, number of ERCPs the patient has undergone during the present hospitalization, emergency or scheduled procedure during or outside regular working hours, kind of sedation, additional medication, general setup,
- technical feasibility and therapeutic target: previous surgery, achievement of therapeutic target, visualization and cannulation of the demanded duct,
- intervention: sphincterotomy (previous and kind of sphincterotomy); stent removal; insertion, exchange (kind and localization), or extraction of concretion; other interventions (dilation, nasobiliary tube, papillectomy, etc.),