



ELSEVIER

Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Computational Statistics & Data Analysis 48 (2005) 857–868

COMPUTATIONAL
STATISTICS
& DATA ANALYSIS

www.elsevier.com/locate/csda

Bayesian computation for logistic regression

Pieter C.N. Groenewald^{a,*}, Lucky Mokgathe^b

^a*Mathematical Statistics, University of the Free State, Bloemfontein 9300, South Africa*

^b*University of Botswana, Gaborone, Botswana*

Received 28 October 2003; received in revised form 15 April 2004; accepted 15 April 2004

Abstract

A method for the simulation of samples from the exact posterior distributions of the parameters in logistic regression is proposed. It is based on the principle of data augmentation and a latent variable is introduced, similar to the approach of Albert and Chib (J. Am. Stat. Assoc. 88 (1993) 669), who applied it to the probit model. In general, the full conditional distributions are intractable, but with the introductions of the latent variable all conditional distributions are uniform, and the Gibbs sampler is easily applicable. Marginal likelihoods for model selection can be obtained at the expense of additional Gibbs cycles. The technique is extended and can be applied with nominal or ordinal polychotomous data.

© 2004 Elsevier B.V. All rights reserved.

Keywords: Data augmentation; Bayes factors; Gibbs sampling; Logit model; Ordinal data; Polychotomous response

1. Introduction

When modelling binary data, the outcome variable Y has a Bernoulli distribution with probability of success π . If the probability of success depends on a set of covariates, then we have a distinct probability π_i , specific to the i th observation, Y_i . The probability π_i is regressed on the covariates through a link function that preserves the properties of probability. So $\pi_i = H(\beta \mathbf{x}_i)$ where \mathbf{x}_i is the vector of covariates associated

* Corresponding author. Tel.: 27514012609; fax: 27514442024.

E-mail address: groenepc@sci.uovs.ac.za (P.C.N. Groenewald).

with the i th observation, $0 \leq H(\cdot) \leq 1$, and $H(\cdot)$ is a continuous non-decreasing function. Usually the link function is taken as the cumulative distribution function (CDF) of some continuous random variable, defined on the whole real line. The two link functions in common use are the CDF of the standard normal distribution, the probit model, and the CDF of the logistic distribution, the logit model. These kinds of models are described in detail in a number of books. See, for example, Cox (1971) or Maddala (1983). For a sample of n observations, the likelihood function is given by

$$L(\boldsymbol{\beta}|\text{data}) \propto \prod_{i=1}^n H(\boldsymbol{\beta}\mathbf{x}_i)^{y_i} (1 - H(\boldsymbol{\beta}\mathbf{x}_i))^{1-y_i}. \quad (1.1)$$

When using maximum likelihood estimation, inferences about the model are usually based on asymptotic theory. Griffiths et al. (1987) found that the MLEs have significant bias for small samples. With the Bayesian approach and prior $\pi(\boldsymbol{\beta})$, the posterior of $\boldsymbol{\beta}$ is given by

$$\pi(\boldsymbol{\beta}|\text{data}) \propto \pi(\boldsymbol{\beta})L(\boldsymbol{\beta}|\text{data}), \quad (1.2)$$

which is intractable in the case of the probit and logit models. In the past, asymptotic normal approximations were used for the posterior of $\boldsymbol{\beta}$. Zellner and Rossi (1984) used numerical integration when the number of parameters is small. Albert and Chib (1993) introduced a simulation-based approach for the computation of the exact posterior distribution of $\boldsymbol{\beta}$ in the case of the probit model. The approach is based on the idea of data augmentation (Tanner and Wong, 1987), where a normally distributed latent variable is introduced into the problem. This approach also enables them to model binary data using a t link function.

In this paper we apply the data augmentation approach of Albert and Chib (1993) to the logit model. This enables us to use Gibbs sampling to obtain samples from the posterior distribution of $\boldsymbol{\beta}$, drawing only from uniform distributions. The technique is extended in Section 3 to multiple response categories, and in Section 4 applied to ordinal responses where the thresholds, or cut off points, must also be estimated. Again, only simulation from uniform distributions is required to obtain marginal posterior distributions.

Gibbs sampling is a simplified version of the Metropolis–Hastings algorithm (Metropolis et al., 1953; Hastings, 1970), and applicable when it is possible to sample directly from all conditional distributions. The Metropolis–Hastings algorithm is usually employed in the case of logistic regression. Other Markov chain Monte Carlo techniques in use are adaptive rejection sampling (ARS), which is used in the WinBugs software, and adaptive rejection metropolis sampling (ARMS).

While marginal posterior distributions of parameters in logistic regression can be obtained using WinBugs, it cannot provide marginal likelihoods. In Section 5 the data augmentation technique is applied to model selection via Bayes factors. Based on a method proposed by Chib (1995), the marginal likelihood under a particular model can be calculated by running additional Gibbs cycles, one for each parameter in the model. In Section 6 the technique is illustrated by two applications.

متن کامل مقاله

دریافت فوری ←

ISIArticles

مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات