Bayesian computation for logistic regression

Pieter C.N. Groenewald[a,*], Lucky Mokgatlhe[b]

[a] *Mathematical Statistics, University of the Free State, Bloemfontein 9300, South Africa*
[b] *University of Botswana, Gabarone, 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.

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 ith observation, Y_i. The probability π_i is regressed on the covariates through a link function that preserves the properties of probability. So $\pi_i = H(\boldsymbol{\beta}\mathbf{x}_i)$ where $\mathbf{x}_i$ is the vector of covariates associated with the ith observation, $0 \leqslant H(.) \leqslant 1$, and $H(.)$ 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

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

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

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

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.

2. Dichotomous response variable

Let

$$Y_i = \begin{cases} 1 & \textit{with probability } \pi_i \\ 0 & \textit{with probability } 1-\pi_i \end{cases} \qquad i = 1, 2, \ldots, n, \tag{2.1}$$

where

$$\log \frac{\pi_i}{1-\pi_i} = \boldsymbol{\beta}\mathbf{x}_i, \tag{2.2}$$

i.e. the log-odds for the ith sampling unit is a linear function of the observed covariates $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})'$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ is a row vector of regression coefficients. Then

$$\pi_i = \frac{\exp(\boldsymbol{\beta}\mathbf{x}_i)}{1+\exp(\boldsymbol{\beta}\mathbf{x}_i)} = F_Z(\boldsymbol{\beta}\mathbf{x}_i) \tag{2.3}$$

where $F_Z(.)$ is the CDF of the logistic random variable Z, with probability density function

$$f_Z(z) = \frac{\exp(z)}{(1+\exp(z))^2}, \quad -\infty < z < \infty.$$

So

$$\begin{aligned} \pi_i &= \int_{-\infty}^{\boldsymbol{\beta}\mathbf{x}_i} \frac{\exp(z)}{(1+\exp(z))^2}\,\mathrm{d}z, \\ &= P\left(U < \frac{\exp(\boldsymbol{\beta}\mathbf{x}_i)}{1+\exp(\boldsymbol{\beta}\mathbf{x}_i)}\right), \end{aligned} \tag{2.4}$$

where U has a Uniform (0,1) distribution. Introducing the independent latent variables $\mathbf{u} = (u_1, u_2, \ldots, u_n)$, the joint posterior density of $\boldsymbol{\beta}$ and $\mathbf{u}$ given the data $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ is then given by

$$\begin{aligned} \pi(\boldsymbol{\beta}, \mathbf{u} \mid \mathbf{y}) \propto \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \Bigg\{ & I\left(u_i \leqslant \frac{\exp(\boldsymbol{\beta}\mathbf{x}_i)}{1+\exp(\boldsymbol{\beta}\mathbf{x}_i)}\right) I(y_i = 1) \\ & + I\left(u_i > \frac{\exp(\boldsymbol{\beta}\mathbf{x}_i)}{1+\exp(\boldsymbol{\beta}\mathbf{x}_i)}\right) I(y_i = 0) \Bigg\} I(0 \leqslant u_i \leqslant 1) \end{aligned} \tag{2.5}$$

$$\begin{aligned} \propto \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \Bigg\{ & I\left(\boldsymbol{\beta}\mathbf{x}_i \geqslant \log\left(\frac{u_i}{1-u_i}\right)\right) I(y_i = 1) \\ & + I\left(\boldsymbol{\beta}\mathbf{x}_i < \log\left(\frac{u_i}{1-u_i}\right)\right) I(y_i = 0) \Bigg\} I(0 \leqslant u_i \leqslant 1), \end{aligned} \tag{2.6}$$

where $\pi(\boldsymbol{\beta})$ is the prior and $I(X \in A)$ is the indicator function that is equal to 1 if $X \in A$, and zero otherwise.

It is clear from (2.5) that, given $\boldsymbol{\beta}$ and $\mathbf{y}$, u_i has a uniform distribution, such that the conditional distribution of u_i is given by

$$u_i|\boldsymbol{\beta},\mathbf{y} \sim \begin{cases} \text{Uniform} \left(0, \dfrac{\exp(\boldsymbol{\beta}\mathbf{x}_i)}{1+\exp(\boldsymbol{\beta}\mathbf{x}_i)}\right) & \text{if } y_i = 1, \\ \text{Uniform} \left(\dfrac{\exp(\boldsymbol{\beta}\mathbf{x}_i)}{1+\exp(\boldsymbol{\beta}\mathbf{x}_i)}, 1\right) & \text{if } y_i = 0, \end{cases} \qquad i = 1,2,\ldots,n. \tag{2.7}$$

From (2.6) we have that

$$\sum_{j=0}^{p} \beta_j x_{ij} \geqslant \log \frac{u_i}{1-u_i} \quad \text{if } y_i = 1,$$

so

$$\beta_k \geqslant \frac{1}{x_{ik}} \left(\log \frac{u_i}{1-u_i} - \sum_{j \neq k}^{n} \beta x_{ij} \right)$$

for all i for which $y_i = 1$ and $x_{ik} > 0$, as well as for all i for which $y_i = 0$ and $x_{ik} < 0$. Similarly,

$$\beta_k < \frac{1}{x_{ik}} \left(\log \frac{u_i}{1-u_i} - \sum_{j \neq k}^{n} \beta x_{ij} \right)$$

for all i for which $y_i = 0$ and $x_{ik} > 0$, as well as for all i for which $y_i = 1$ and $x_{ik} < 0$, assuming $x_{ik} \neq 0$. Let A_k and B_k be the sets

$$A_k = \{i : ((y_i = 1) \cap (x_{ik} > 0)) \cup ((y_i = 0) \cap (x_{ik} < 0))\},$$

$$B_k = \{i : ((y_i = 0) \cap (x_{ik} > 0)) \cup ((y_i = 1) \cap (x_{ik} < 0))\}.$$

Then, assuming the diffuse prior, $\pi(\boldsymbol{\beta}) \propto 1$, for $\boldsymbol{\beta}$, the conditional distribution of β_k, given all the other β's and $\mathbf{u}$, is the uniform distribution;

$$\beta_k|\boldsymbol{\beta}_{(-k)},\mathbf{u},\mathbf{y} \sim \text{Uniform}(a_k, b_k), \quad k = 0,1,2,\ldots,p, \tag{2.8}$$

where

$$a_k = \max_{i \in A_k} \left[\frac{1}{x_{ik}} \left(\log \frac{u_i}{1-u_i} - \sum_{j \neq k}^{n} \beta x_{ij} \right) \right], \tag{2.9}$$

and

$$b_k = \min_{i \in B_k} \left[\frac{1}{x_{ik}} \left(\log \frac{u_i}{1-u_i} - \sum_{j \neq k}^{n} \beta x_{ij} \right) \right]. \tag{2.10}$$

The Gibbs sampler is now easy to implement by drawing from uniform distributions. The n values of $\mathbf{u}$ can be drawn in one block through (2.7), while the elements of $\boldsymbol{\beta}$ are drawn successively by using (2.8). If the β's are independently distributed apriori with any prior $\pi(\beta_k)$ on β_k, then the conditional distribution is simply $\pi(\beta_k)$, truncated between a_k and b_k.

3. Polychotomous response variable

Consider now the case where Y has more than two response categories and assume independence among repeated trials. This results in a data set of n observations whose distribution is multinomial with r categories. Let

$$\pi_{ij} = P(Y_i = j),$$

and assume the logit link function,

$$\log\left(\frac{\pi_{ij}}{\pi_{ir}}\right) = \boldsymbol{\beta}_j \mathbf{x}_i, \quad j = 1, 2, \ldots, r. \tag{3.1}$$

Then

$$\begin{aligned}\pi_{ij} &= \frac{\exp(\boldsymbol{\beta}_j \mathbf{x}_i)}{1 + \sum_{s=1}^{r-1} \exp(\boldsymbol{\beta}_s \mathbf{x}_i)}, \quad j = 1, 2, \ldots, r. \\ &= P\left(U < \frac{\exp(\boldsymbol{\beta}_j \mathbf{x}_i)}{1 + \sum_{s=1}^{r-1} \exp(\boldsymbol{\beta}_s \mathbf{x}_i)}\right),\end{aligned} \tag{3.2}$$

where $U \sim$ Uniform $(0, 1)$. Note that the rth category is the baseline category and $\boldsymbol{\beta}_r = \mathbf{0}$. The joint posterior distribution of $\boldsymbol{\beta} = \{\beta_{jk}\}((r-1) \times (p+1))$ and $\mathbf{U} = \{u_{ij}\}(n \times (r-1))$, given the observed data, is

$$\pi(\boldsymbol{\beta}, \mathbf{U}|y) \propto \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \sum_{j=1}^{r-1} \left[I\left(u_{ij} < \frac{\exp(\boldsymbol{\beta}_j \mathbf{x}_i)}{1 + \sum_{s=1}^{r-1} \exp(\boldsymbol{\beta}_s \mathbf{x}_i)}\right) I(y_i = j) \right] I(0 \leqslant u_{ij} \leqslant 1), \tag{3.3}$$

so that the conditional distribution of u_{ij} is given by

$$u_{ij}|\boldsymbol{\beta}, \mathbf{y} \sim \begin{cases} \text{Uniform } (0, \pi_{ij}) & \text{if } y_i = j \\ \text{Uniform } (\pi_{ij}, 1) & \text{if } y_i \neq j \end{cases}, \quad i = 1, 2, \ldots, n; \; j = 1, 2, \ldots, r-1,$$

with π_{ij} given in (3.2). From (3.3) it follows that

$$u_{ij}\left(1 + \sum_{s=1}^{r-1} \exp\left(\sum_{k=0}^{p} \beta_{sk} x_{ik}\right)\right) < \exp\left(\sum_{k=0}^{p} \beta_{jk} x_{ik}\right)$$

$$u_{ij}\left(1 + \sum_{s \neq j}^{r-j} \exp\left(\sum_{k=0}^{p} \beta_{sk} x_{ik}\right)\right) < \exp\left(\sum_{k=0}^{p} \beta_{jk} x_{ik}\right) - u_{ij} \exp\left(\sum_{k=0}^{p} \beta_{jk} x_{ik}\right)$$

$$\log\left[\frac{u_{ij}}{1 - u_{ij}} \left(1 + \sum_{s \neq j}^{r-1} \exp\left(\sum_{k=0}^{p} \beta_{sk} x_{ik}\right)\right)\right] < \sum_{k=0}^{p} \beta_{jk} x_{ik}.$$

So

$$\beta_{jt} > \Lambda_{ijt},$$

where

$$\Lambda_{ijt} = \frac{1}{x_{it}} \left\{ \log \left[\frac{u_{ij}}{1-u_{ij}} \left(1 + \sum_{s \neq j}^{r-1} \exp\left(\sum_{k=0}^{p} \beta_{sk} x_{ik} \right) \right) \right] - \sum_{k \neq t}^{p} \beta_{jk} x_{ik} \right\}. \tag{3.4}$$

The above inequality must hold for all i for which $y_i = j$ and $x_{it} > 0$ as well as for all i for which $y_i \neq j$ and $x_{it} < 0$. So let

$$A_{jt} = \{i : ((y_i = j) \cap (x_{it} > 0)) \cup ((y_i \neq j) \cap (x_{it} < 0))\}$$

$$B_{jt} = \{i : ((y_i = j) \cap (x_{it} < 0)) \cup ((y_i \neq j) \cap (x_{it} > 0))\},$$

then the conditional distribution of β_{jt} is Uniform (a_{jt}, b_{jt}), where

$$a_{jt} = \max_{i \in A_{jt}} \Lambda_{ijt} \quad \text{and} \quad b_{jt} = \min_{i \in B_{jt}} \Lambda_{ijt}.$$

4. Ordinal responses

Suppose Y_i can take one of r ordered categories, $j=1,2,\ldots,r$, so that $P(Y_i=j)=\pi_{ij}$, and the cumulative probabilities are $\eta_{ij}=\sum_{k=1}^{j} \pi_{ik}=P(Y_i \leqslant j)$. Introduce the continuous latent variable U, uniformly distributed over [0,1], such that

$$\eta_{ij} = P\left(U_i < \frac{\exp(\alpha_j + \boldsymbol{\beta}\mathbf{x}_i)}{1+\exp(\alpha_j + \boldsymbol{\beta}\mathbf{x}_i)} \right) = \frac{\exp(\alpha_j + \boldsymbol{\beta}\mathbf{x}_i)}{1+\exp(\alpha_j + \boldsymbol{\beta}\mathbf{x}_i)}, \tag{4.1}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ are the regression coefficients and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_r)$ are the cut-off points of the intervals, such that $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_r = \infty$.

The joint posterior distribution of $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and $\mathbf{u}$, given the response $\mathbf{y}$, is then given by

$$\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{u} \mid \mathbf{y}) \propto \pi(\boldsymbol{\alpha}, \boldsymbol{\beta}) \prod_{i=1}^{n} \left\{ \sum_{j=1}^{r-1} I(y_i = j) I(\eta_{i,j-1} < u_i \leqslant \eta_{ij}) \right\} I(0 \leqslant u_i \leqslant 1). \tag{4.2}$$

Assuming diffuse prior distributions for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the conditional distribution of the latent variable u_i is given by

$$u_i | \boldsymbol{\alpha}, \boldsymbol{\beta}, y_i = j \sim \text{Uniform}(\eta_{i,j-1}, \eta_{ij}), \quad i = 1, 2, \ldots, n, \tag{4.3}$$

where η_{ij} is given in (4.1).

Let

$$H_{ijt} = \frac{1}{x_{it}} \left[\log \frac{u_i}{1-u_i} - \alpha_j - \sum_{s \neq t}^{p} \beta_s x_{is} \right]. \tag{4.4}$$

Then it follows from (4.2) that $H_{ijt} \leqslant \beta_t < H_{i,j-1,t}$ $(H_{i,j-1,t} \leqslant \beta_t < H_{ijt})$ for all i for which $y_i = j$ and for which $x_{it} > 0$ $(x_{it} < 0)$. But (4.4) must also hold for all $j = 1, 2, \ldots, r$, so let $A_j = \{i : y_i = j\}$, then

$$\beta_t \mid \boldsymbol{\beta}_{(-t)}, \boldsymbol{\alpha}, \mathbf{u}, \mathbf{y} \sim \text{Uniform}(a_t, b_t), \quad t = 1, 2, \ldots, p, \tag{4.5}$$

where $a_t < b_t$ always and

$$a_t = \max_j \left\{ \max_{i \in A_j} [\min(H_{i,j-1,t}, H_{ijt})] \right\},$$

$$b_t = \min_j \left\{ \min_{i \in A_j} [\max(H_{i,j-1,t}, H_{ijt})] \right\}. \tag{4.6}$$

To find the conditional distribution of the cut-off points, $\alpha_1, \ldots, \alpha_{r-1}$, we have the condition that $u_i \leqslant \eta_{ij}$ for all $i \in A_j$ and $u_i > \eta_{ij}$ for all $i \in A_{j+1}$. Also $\alpha_{j-1} < \alpha_j < \alpha_{j+1}$, so that

$$\alpha_j \mid \boldsymbol{\alpha}_{(-j)}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{y} \sim \text{Uniform}(c_j, d_j), \quad j = 1, 2 \ldots, r-1, \tag{4.7}$$

where

$$c_j = \max_{i \in A_{j+1}} \left[\max \left\{ \log \frac{u_i}{1-u_i} - \boldsymbol{\beta}\mathbf{x}_i, \alpha_{j-1} \right\} \right],$$

$$d_j = \min_{i \in A_j} \left[\min \left\{ \log \frac{u_i}{1-u_i} - \boldsymbol{\beta}\mathbf{x}_i, \alpha_{j+1} \right\} \right]. \tag{4.8}$$

5. Model selection

Bayesian model selection, or variable selection, is usually based on the Bayes factor, which is the ratio of the marginal likelihoods of two competing models. Priors in general should be proper, so in the context of the dichotomous model of Section 2, exchangeable logistic priors with mean zero and scale parameter σ are assumed for the elements of $\boldsymbol{\beta}_t$, the set of regression coefficients under model M_t. Let p_t be the number of covariates included under model M_t, then

$$\pi(\beta_j \mid \sigma, M_t) = \frac{\exp(\beta_j/\sigma)}{\sigma(1+\exp(\beta_j/\sigma))^2}, \quad j = 0, 1, \ldots, p_t. \tag{5.1}$$

Priors on nuisance parameters that appear in all models, and with the same interpretation in all models, does not greatly affect the results (Kass and Raftery, 1995), and improper priors are being used in these situations. So let $\pi(\sigma) \propto 1/\sigma$ for all models, and to justify a common scale parameter for all regression coefficients, the covariates are standardised.

The marginal likelihood under model M_t can be written in terms of the posterior distribution as

$$m(\mathbf{y} \mid M_t) = \frac{L(\boldsymbol{\beta}_t \mid \mathbf{y}, M_t)\pi(\boldsymbol{\beta}_t \mid \sigma, M_t)\pi(\sigma)}{\pi(\boldsymbol{\beta}_t, \sigma \mid \mathbf{y}, M_t)}. \tag{5.2}$$

The above identity holds for any parameter value. The numerator can be directly evaluated at a given point, say $(\boldsymbol{\beta}_t^*, \sigma^*)$. However, the posterior, $\pi(\boldsymbol{\beta}_t, \sigma \mid \mathbf{y}, M_t)$, is not available in closed form, and the initial Gibbs sampling does not provide an estimate of the posterior value at $(\boldsymbol{\beta}_t^*, \sigma^*)$. Chib (1995) developed an approach to obtain an estimate of the posterior ordinate by performing additional Gibbs cycles. Using basic probability rules, the posterior density ordinate, $\pi(\boldsymbol{\beta}_t^*, \sigma^* | \mathbf{y}, M_t)$, can be expressed as

$$\pi(\boldsymbol{\beta}_t^*, \sigma^* \mid \mathbf{y}, M_t) = \pi(\beta_0^* \mid \mathbf{y}, M_t)\pi(\beta_1^* \mid \beta_0^*, \mathbf{y}, M_t) \ldots\ldots \pi(\sigma^* \mid \boldsymbol{\beta}_t^*, \mathbf{y}, M_t). \tag{5.3}$$

Each term on the right-hand side of (5.2) can now be evaluated by Gibbs sampling. Consider the first term, which is the ordinate of the marginal posterior of β_0 at β_0^* under model M_t. During each cycle the latent variables, $u_i, i=1,2,\ldots,n$, are drawn according to (2.7) and a_j and b_j calculated from (2.10) for each parameter β_j. A variable v_j is drawn where

$$v_j \mid a_j, b_j, \sigma \sim \text{Uniform}\left(\frac{\exp(a_j/\sigma)}{(1+\exp(a_j/\sigma))}, \frac{\exp(b_j/\sigma)}{(1+\exp(b_j/\sigma))}\right), \tag{5.4}$$

and β_j follows as

$$\beta_j = -\sigma \ln\left(\frac{1-v_j}{v_j}\right), \quad j=0,1,\ldots,p_t. \tag{5.5}$$

The conditional distribution of σ follows as

$$\pi(\sigma|\boldsymbol{\beta}_t) \propto \frac{1}{\sigma^{p_t+2}} \frac{\exp(\sum \beta_j/\sigma)}{\prod(1+\exp(\beta_j/\sigma))^2}, \tag{5.6}$$

from which σ can be simulated using the Metropolis–Hastings algorithm, or, as done in the applications, by discretisizing the distribution. After the mth cycle the conditional ordinate at β_0^* is given by

$$\pi(\beta_0^* \mid \beta_1^{(m)},\ldots,\beta_{p_t}^{(m)},\mathbf{u}^{(m)},\mathbf{y}) = \begin{cases} C_0^{(m)} \dfrac{\exp(\beta_0^*/\sigma)}{\sigma(1+\exp(\beta_0^*/\sigma))^2} & \text{for } a_0^{(m)} < \beta_0^* < b_0^{(m)} \\ 0 & \text{otherwise,} \end{cases} \tag{5.7}$$

where

$$C_0 = \frac{(1+\exp(a_0/\sigma))(1+\exp(b_0/\sigma))}{\exp(b_0/\sigma)-\exp(a_0/\sigma)}.$$

After L cycles the ordinate $\pi(\beta_0^* \mid \mathbf{y}, M_t)$ is obtained by averaging the values from (5.7). Similarly for the rest of the parameters. For each parameter a new Gibbs chain is run, keeping the values of parameters from previous runs fixed. Finally, the last term on the right-hand side of (5.3), $\pi(\sigma^* \mid \boldsymbol{\beta}_t^*, \mathbf{y}, M_t)$, is just Eq. (5.6), evaluated at $\boldsymbol{\beta}_t^*$ and σ^*.

Although this procedure leads to an increase in the number of iterations, it requires little new programming beyond what is needed for estimation and is straightforward to implement. Eq. (5.2) holds for any parameter value, but the procedure is more stable if high density points are used, and in practice points close to the posterior mode are preferable. See Chib (1995) for more details and properties. The procedure can similarly be applied to the polychotomous and ordinal response models.

6. Applications

6.1. Application 1

Piegorsch (1992) analysed data on the analgesic effect of iontophoretic treatment with the chemical vincristine on elderly patients complaining of postherpetic neuralgia. Eighteen patients were interviewed 6 weeks after undergoing treatment to determine if any improvement in the neuralgia was evident. The response variable Y is 1 if an improvement was recorded, and 0 otherwise. The four covariates are X_1; treatment (1 or 0), X_2; age, X_3; sex (1 for male), X_4; pre-treatment duration of symptoms. In our analysis the covariates are standardised. The data is also reproduced in Hand et al. (1994).

The marginal likelihoods of all possible models were calculated, using the procedure of Section 5, with 20,000 iterations during each cycle. The conditional distribution of σ (Eq. (5.6)) was discretisized with steps 0.2(0.2)3. For the model with no covariates, $\sigma = 1$ was assumed, corresponding to a uniform prior on $\pi = P(Y_i = 1)$. Table 1 lists the models with the highest Bayes factors relative to the simple model with no covariates.

According to the Bayes factors, the model with the two covariates, treatment and sex, is the best, with the first eight models showing a marked improvement on the simple model. All these models include X_1 (treatment), and the Bayes factor for the model with X_1 alone is 1.35 when compared to the full model. Piegorsch (1992) used the data to illustrate complementary log regression and concluded that the three additional variables, age, sex and duration, do not significantly improve the fit after inclusion of the treatment term in the model. Adopting the model with covariates X_1 and X_3, the posterior distributions of the parameters are shown in Fig. 1, calculated from 30,000 iterations. The model with improper priors from Section 2 gives essentially the same results for the three regression parameters.

From the Gibbs output also follows other measures of interest. For example, the mean posterior probability of a female on the treatment experiencing an improvement is 0.742 with 90% HPD interval (0.518–0.987), while the same probability for a male without treatment is only 0.089 with 90% HPD interval (0–0.279).

6.2. Application 2

In a study on the job expectations of students at the University of Regensburg (Fahrmeir and Tutz, 2001), psychology students were asked if they expect to find

Table 1
Bayes factors of models including given covariates against the simple model

Covariates	X_1,X_3	X_1	X_1,X_4	X_1,X_2,X_4	X_1,X_2,X_3,X_4	X_1,X_2
Bayes factor	121.1	65.8	56.3	55.1	48.8	42.9
Covariates	X_1,X_2,X_3	X_1,X_3,X_4	X_2	X_4	X_2,X_3,X_4	X_2,X_3
Bayes factor	21.1	20.4	2.5	1.7	1.5	1.5

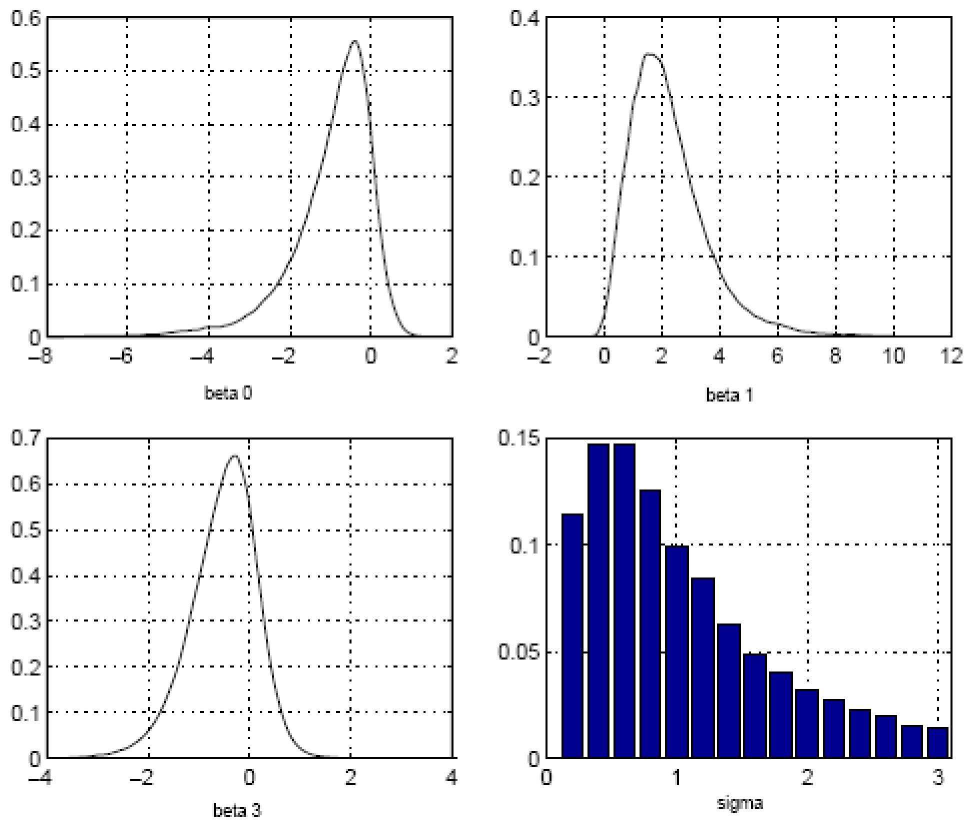

Fig. 1. Posterior distributions of parameters under model with covariates X_1 (treatment) and X_3 (sex).

Table 2
Posterior means and 95% HPD intervals of parameters for employment data

Parameter	Post. Mean	95% HPD Interval	F&T
α_1	14.925	13.38–16.47	14.987
α_2	18.101	16.55–19.66	18.149
β	−5.390	−5.86–−4.90	−5.402

employment after getting their degree, and the response categories were ordered according to their expectation. The three categories were 1 (do not expect adequate employment), 2 (not sure), and 3 (expect immediate employment). The only covariate is age which ranges from 19 to 34. There are 102 observations and the data can be found in Fahrmeir and Tutz (2001). The response can be considered an assessed ordered variable and the theory in Section 4 applied with two cut-off points, α_1 and α_2, and the regression coefficient β of $x_i \equiv \log(\text{age})$. Vague uniform

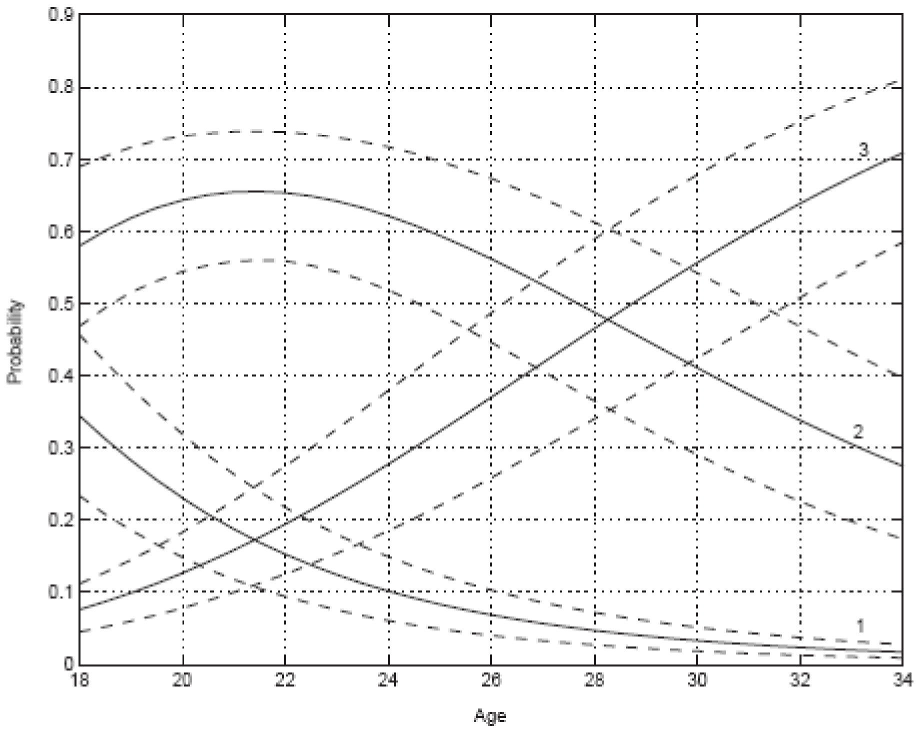

Fig. 2. Posterior probabilities for the three categories as a function of age, with 95% HPD intervals.

priors was used for all parameters. The Gibbs cycle was run 50,000 times and the Bayes estimators are given in Table 2 with the 95% HPD intervals. The maximum likelihood estimators obtained by Fahrmeir and Tutz (2001) (F&T) are also shown.

The posterior means correspond closely with the estimators of F&T. Fig. 2 shows the probabilities for the three categories as a function of age, with the 95% HPD intervals. It is clear that optimism about employment increases with age.

7. Conclusion

The main purpose of this paper is to illustrate a relatively simple method of simulating values from the marginal posterior distributions of the parameters in a logit model using the Gibbs sampler. This model is also very suitable for calculating marginal likelihoods and thus Bayes factors when comparing competing models. As the full conditional distributions of the parameters are intractable, Bayesian analyses usually employed the Metropolis–Hastings algorithm to obtain posterior distributions. The Gibbs sampler is much easier to apply, and, from experience, converges quickly. The method is based on the data augmentation approach of Tanner and Wong (1987), which is applied by Albert and Chib (1993) to the probit model, where a normally distributed latent variable is introduced. In the case of the logit model a logistic variable is transformed so that we work with a uniformly distributed latent variable.

In the applications of Section 6 the marginal posterior of a parameter was obtained by the "Rao-Blackwell" method (see Gelfand and Smith, 1990, Section 2.6) of averaging over the conditional distributions, given the generated values of the other parameters. Because of the nature of the conditional distributions (uniform) it needs a relatively large number of cycles to obtain a smooth marginal posterior distribution. In the case of several covariates it may be more efficient to approximate the posteriors by smoothed histograms of the generated values.

In this paper we used exchangeable logistic priors for all regression parameters when calculating Bayes factors. This was done for convenience (still drawing from uniform distributions), and because the logistic distribution is close to the normal distribution in shape. The differences in posteriors are negligible if the prior variances are not too small. However, this can be generalised at a slight expense of computational effort. If the β's are assumed to have some exchangeable proper priors, then the full conditional distribution of a particular β will be the prior distribution, truncated at the same end-points as the uniform distribution derived in this paper. The conditional distribution of the latent variable $\mathbf{u}$ remains unaffected.

References

Albert, J.H., Chib, S., 1993. Bayesian analysis of binary and polychotomous response data. J. Amer. Statist. Assoc. 88, 669–679.

Chib, S., 1995. Marginal likelihood from the Gibbs output. J. Amer. Statist. Assoc. 90, 1313–1321.

Cox, D.R., 1971. The Analysis of Binary Data. Methuen, London.

Fahrmeir, L., Tutz, G., 2001. Multivariate Statistical Modelling Based on Generalized Linear Models, 2nd Edition. Springer, New York.

Gelfand, A.E., Smith, A.F.M., 1990. Sampling-based approaches to calculating marginal densities. J. Amer. Statist. Assoc. 85, 398–409.

Griffiths, W.E., Hill, R.C., Pope, P.J., 1987. Small sample properties of probit model estimators. J. Amer. Statist. Assoc. 82, 929–937.

Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J., Ostrowski, E., 1994. A Handbook of Small Data Sets. Chapman & Hall, London.

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chain and their applications. Biometrika 87, 97–109.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Amer. Statist. Assoc. 90, 773–795.

Maddala, G.S., 1983. Limited Dependent and Qualitative Variables in Econometrics. Cambridge University Press, New York.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E., 1953. Equations of state calculations by fast computing machines. J. Chem. Phys. 21, 1087–1091.

Piegorsch, W.W., 1992. Complimentary log regression for generalized linear models. Amer. Statistician 46, 94–99.

Tanner, T.A., Wong, W.H., 1987. The calculation of posterior distributions by data augmentation. J. Amer. Statist. Assoc. 82, 528–549.

Zellner, A., Rossi, P.E., 1984. Bayesian analysis of dichotomous quantal response models. J. Econometrics 25, 365–393.