Journal of the Korean Statistical Society (2000), **29**: *4, pp 407–422*

A Bayesian Method for Narrowing the Scope of Variable Selection in Binary Response t-Link Regression

Hea-Jung Kim [1]

Abstract

This article is concerned with selecting predictor variables to be included in building a class of binary response t-link regression models where both probit and logistic regression models can be approximately taken as members of the class. It is based on a modification of the stochastic search variable selection method (SSVS), intended to propose and develop a Bayesian procedure that uses probabilistic considerations for selecting promising subsets of predictor variables. The procedure reformulates the binary response t-link regression setup in a hierarchical truncated normal mixture model by introducing a set of hyperparameters that will be used to identify subset choices. In this setup, the most promising subset of predictors can be identified as that with highest posterior probability in the marginal posterior distribution of the hyperparameters. To highlight the merit of the procedure, an illustrative numerical example is given.

Keywords: Binary response t-link regression; Variable selection; Hierarchical normal mixture model; Data augmentation; Gibbs sampler; High frequency model

1. Introduction

The theory of binary dependent variable regression has its genesis in bioassay(see, for example, Finney, 1971). In that context, Y_i denotes a Bernoulli random variable with the probability of success p_i, $i = 1, \cdots, n$, where p_i is related to a set of predictors which may be continuous or discrete, and the binary response regression model is defined as

$$Y_i = H(X_i'\beta) + \varepsilon_i, \quad i = 1, \cdots, n, \tag{1}$$

[1] Department of Statistics, Dongguk University, Seoul 100-715, Korea.

The author wishes to acknowledge the financial support of the Korea Research Foundation made in program year of 1999.

where the ε_i is uncorrelated random error with $E\varepsilon_i = 0$, X_i is a vector of, say, k fixed predictors, β is a $k \times 1$ vector of unknown coefficients, and $H(\cdot)$ is a known cdf linking the probability p_i with linear structure $X_i'\beta$ so that $p_i = H(X_i'\beta)$. Suppose we specify the model (1) by choosing the link cdf H to be the family of t distributions, the resulting model is called binary response t-link regression model. A special feature of this model is that probit model is a member of the family for $t_\infty = N(0,1)$ and logistic model can be approximately viewed as a t_8 link model (cf. Albert and Chib, 1993 and Soofi, Ebrahimi and Habibullah, 1995).

At some point during the analysis with the binary response t-link regression model, one may wish to delete some predictors from the model. The search for the best submodel is called variable selection or subset selection. In recent years, the use of MCMC simulation techniques leads to develop various Bayesian procedures to handle variable selection for regression problems. Those procedures are designed to search for promising subsets of predictors stochastically, so that they avoid overwhelming burden of calculation involved in comparison of all 2^k possible subsets, and allow for variable selection based on practical significance rather than statistical significance. Among them, stochastic search variable selection (SSVS) introduced by George and McCulloch (1993), conditional Bayes factor method by Geweke (1996) and reversible-jump Metropolis-Hasting algorithm of Green (1994) are prominent. See Dellaportas, Foster and Ntzoufras (1997) and references therein for other methods. For the selection in binary response regression models, George and McCulloch (1996) and George, McCulloch and Tsay (1996) discussed an implementation of SSVS using the adaptive rejection sampling method of Gilks and Wild (1992) and Kuo and Mallick (1997) suggested implementation of a Metropolis algorithm.

The purpose of this paper is to develop and suggest yet another procedure fitted to variable selection for the binary response t-link regression that allows sign and interval constraints on regression coefficients. The procedure is an extension of SSVS of George and McCulloch (1996) in three directions: (i) It reformulates binary response t-link regression setup in a hierarchical truncated normal mixture model so that it enables implementation of the Gibbs sampling algorithm. This is achieved by assuming a latent continuous response with a t-distribution and determining an observed binary response by a cutpoint. (ii) It achieves variables selection not only for a class of t-link models, but for logistic and probit link models. (iii) It accounts for sign and interval constraints on regression coefficients as considered by Geweke (1996) in the Gaussian regression.

In Section 2 we define and motivate the hierarchical framework for the t-link regression model that serves as the basis for the stochastic search method for variable selection. Section 3 show how the hierarchical model can be used to identify highly promising subsets of the predictor variables not only for the t-link regression models, but for the probit and the logistic regression models. A posterior distribution derived from a data augmentation scheme and a computational algorithm are also outlined in this section. In Section 4 we illustrate the suggested procedure on simulated data sets. The last section summarizes and discusses some possible extensions of this work.

2. Hierarchical Model For t Link Regression

Suppose that we have n binary response observations Y_i, $i = 1, \ldots, n$, where $E(Y_i) = p_i$ which is the success probability corresponding to the i-th observation. If we set $H(\cdot)$ in (1) as $T_\nu(\cdot)$, the binary response t-link regression model for the dependence of p_i on k explanatory variables vector, $X_i = (x_{1i}, x_{2i}, \ldots, x_{ki})'$, is

$$T_\nu^{-1}(p_i) \;=\; \beta' X_i, \;\; i = 1, \ldots, n. \tag{2}$$

where $\beta = (\beta_1, \ldots, \beta_k)'$ is an unknown coefficient vector and $T_\nu(\cdot)$ is cdf of t-distribution with fixed ν degrees of freedom. As a result of some arrangement,

$$p_i \;=\; T_\nu(\beta' X_i) \;=\; (\pi\nu)^{-1/2} \frac{\Gamma[(\nu+1)/2]}{\Gamma[\nu/2]} \int_{-\infty}^{\beta' X_i} (1 + u^2/\nu)^{-(\nu+1)/2} du. \tag{3}$$

Since Y_i is an observation from a Bernoulli distribution with mean p_i, corresponding model for the expected value of Y_i is $E(Y_i) = T_\nu(\beta' X_i)$. For the model (2), selecting a subset of predictors is equivalent to setting to 0 those β_i's corresponding to the unselected predictors. If an intercept was to be included in the variable selection (as is usually the case), then one should set $x_{1i} = 1$, $i = 1, \ldots, n$.

The likelihood function of the model (2) is given by

$$L(\beta) \;=\; \prod_{i=1}^{n} p_i^{Y_i} (1 - p_i)^{1 - Y_i}, \tag{4}$$

where p_i is defined by (3). The likelihood depends on the unknown success probabilities p_i, which in turn depends on the β through (3), so that the likelihood function may be regarded as a function of β.

To extract information relevant to variable selection, we consider the following hierarchical model structure (cf. Bernardo and Smith, 1994). In conventional terminology, the first stage of the hierarchy relates data to parameters via (4). The key feature of the hierarchical model is that each component of β is modeled as having come from a mixture of two normal distributions with a truncated interval. This is done by introducing a set of latent variables based on the data augmentation idea of Tanner and Wong (1987). The second stage models can be simply expressed via the introduction of a set of distinct latent variables $\{\alpha_j = 0 \text{ or } 1,\ j = 1, \ldots, k\}$, so that $\beta_j' s$ are independent and random samples from normal mixtures represented by

$$\beta_j|\alpha_j \sim (1-\alpha_j)TN_{\{a_j \le \beta_j \le b_j\}}(0, \sigma_j^2) + \alpha_j TN_{\{a_j \le \beta_j \le b_j\}}(0, c_j^2\sigma_j^2),\ j = 1, \ldots, k, \tag{5}$$

where $p(\alpha_j = 1) = 1 - p(\alpha_j = 0) = q_j$ and hyperparameters σ_j, q_j, a_j b_j and c_j are fixed and $TN_{\{a_j \le \beta_j \le b_j\}}(0, \psi)$ denotes the normal distribution $N(0, \psi)$ truncated to interval $a_j \le \beta_j \le b_j$. In case β_j is not truncated in priori, we may simply set $a_j = -\infty$ and $b_j = \infty$.

The above formulation shows that, for $\alpha_j = 0$, $\beta_j \sim TN_{\{a_j \le \beta_j \le b_j\}}(0, \sigma_j^2)$, and for $\alpha_j = 1$, $\beta_j \sim TN_{\{a_j \le \beta_j \le b_j\}}(0, c_j^2\sigma_j^2)$. It may be interpreted as follows: (i) In case $\alpha_j = 0$, our choice of small $\sigma_j (> 0)$ implies that β_j is likely to be so small that it could have zero estimate in the constrained estimation space. (ii) In case $\alpha_j = 1$, our prior judgment about non-zero estimate of β_j being more likely than zero estimate is captured by choosing $c_j (> 1)$ to be large. Based on this interpretation, q_j may be thought of as the prior probability that β_j has a non-zero estimate satisfying the constrained interval, or equivalently jth predictor should be included in the final model. So that if we have a prior belief that β_j lies in an interval $[a_j,\ b_j]$ not including zero, then we may simply set $q_j = 1$.

A similar setup (hereafter referred to as GM) in this context was considered by George and McCulloch (1993). GM can be viewed as the priors (5) with $a_j = -\infty$ and $b_j = \infty$. However, their priors differ from (5) in three respects. First, the present paper employs no presumption of conjugacy in the priors and propose an independent prior distribution for each regression coefficient that is a mixture of two univariate normal distribution, while GM prior distributions that are natural conjugate or near natural conjugate and permits prior dependence across coefficients. Second, (5) allows sign constraints for particular coefficients by differing the truncated intervals (for example, one may set $a_j = 0$ and $b_j = \infty$ for the positive β_j), but GM does not permit sign constraints. Third, if one has priori belief that $\beta_j \in [a_j, b_j]$, where a_j and b_j have the same sign, then the prior distributions used in this paper include the case that j-th predictor variable is included in every possible model, whereas GM for technical reasons does not allow the case.

For choosing $c_j(> 1)$ and σ_j in (5), a useful guide is the following. The density of $TN_{\{a_j \le \beta_j \le b_j\}}(0, c_j^2\sigma_j^2)$ is larger than that of $TN_{\{a_j \le \beta_j \le b_j\}}(0, \sigma_j^2)$ iff $|\beta_j| > \delta(c_j)\sigma_j$, where $\delta(c_j) = (2\ln(\eta_j c_j)c_j^2/(c_j^2 - 1))^{1/2}$ and

$$\eta_j = \frac{\Phi(b_j/(c_j\sigma_j)) - \Phi(a_j/(c_j\sigma_j))}{\Phi(b_j/\sigma_j) - \Phi(a_j/\sigma_j)}.$$

It may be also useful to note that $\eta_j c_j$ is the ratio of the heights of $TN_{\{a_j \le \beta_j \le b_j\}}(0, c_j^2\sigma_j^2)$ and $TN_{\{a_j \le \beta_j \le b_j\}}(0, \sigma_j^2)$ at 0, indicating the prior odds of excluding x_j when β_j is very close to 0.

The third, and final, stage specifies beliefs about α_j's. This can be done via a reasonable choice of the prior density for $\alpha = (\alpha_1, \dots, \alpha_k)'$;

$$p(\alpha) \ = \ \prod_{j=1}^{k} q_j^{\alpha_j}(1 - q_j)^{(1-\alpha_j)}.$$

Therefore, the complete model structure of the hierarchy has the form.

$$\begin{aligned}
p(Y|\beta) &= \prod_{i=1}^{n} p_i^{Y_i}(1 - p_i)^{1-Y_i}, \\
p(\beta|\alpha) &= \prod_{j=1}^{k} \left[\frac{(2\pi\sigma_j^2)^{-1/2}}{\Phi(b_j/\sigma_j) - \Phi(a_j/\sigma_j)} \exp\left\{-\frac{\beta_j^2}{2\sigma_j^2}\right\} [I(\alpha_j = 0)I(a_j \le \beta_j \le b_j)] \right. \\
&+ \left. \frac{(2\pi c_j^2\sigma_j^2)^{-1/2}}{\Phi(b_j/(c_j\sigma_j)) - \Phi(a_j/(c_j\sigma_j))} \exp\left\{-\frac{\beta_j^2}{2c_j^2\sigma_j^2}\right\} [I(\alpha_j = 1)I(a_j \le \beta_j \le b_j)] \right], \\
p(\alpha) &= \prod_{j=1}^{k} q_j^{\alpha_j}(1 - q_j)^{(1-\alpha_j)},
\end{aligned}$$

where $I(A)$ is an indicator function of the event A.

In many applications, it may be of interest to make inferences both about the unit characteristics, the β_j's, and the population characteristics, the α_j's. In either case, straightforward probability manipulations involving Bayes' theorem provide the required joint posterior density of β and α from which one can make the inference of interest:

$$f(\beta, \alpha|Y) =$$

$$C \prod_{j=1}^{k} \left[\frac{(2\pi\sigma_j^2)^{-1/2}}{\Phi(b_j/\sigma_j) - \Phi(a_j/\sigma_j)} \exp\left\{-\frac{\beta_j^2}{2\sigma_j^2}\right\} [I(\alpha_j = 0)I(a_j \le \beta_j \le b_j)] \right.$$

$$+ \quad \frac{(2\pi c_j^2\sigma_j^2)^{-1/2}}{\Phi(b_j/(c_j\sigma_j)) - \Phi(a_j/(c_j\sigma_j))} \exp\left\{-\frac{\beta_j^2}{2c_j^2\sigma_j^2}\right\} [I(\alpha_j = 1)I(a_j \le \beta_j \le b_j)]\Bigg],$$

$$\times \quad \prod_{j=1}^{k} q_j^{\alpha_j}(1-q_j)^{(1-\alpha_j)} \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i}, \tag{6}$$

where C in the above equation is a generic proportionality constant.

Our main reason for embedding the t-link model (2) in the above hierarchical mixture model is to obtain the marginal posterior distribution $h(\alpha|Y) \propto f(Y|\alpha)p(\alpha)$, which contains the information relevant to variable selection. However, it is easily seen that the problem of analytically calculating the marginal from (6) is a challenging one. Fortunately, recent developments of a MCMC method, say the Gibbs sampler, provides a method that directly addresses simulation based calculation of the marginal posterior (cf. Gelfand and Dey 1994).

3. Data Augmentation and the Gibbs Sampler

3.1. Data Augmentation

To allow the possibility that the posterior simulation requires data augmentation (cf. Albert and Chib 1993), we introduce a set of latent variables $\{Z_i,\ i = 1,\dots,n\}$, where the Z_i are independently distributed as a t with location parameter $X_i'\beta$, scale parameter 1, and degrees of freedom ν, so that

$$Z_i \ \sim\ t_\nu(X_i'\beta, 1) \ \text{ and } \ Y_i = I(Z_i > 0), \ \ i = 1,\dots,n, \tag{7}$$

where $I(A)$ is an indicator function of the event A. The above specification in fact defines the t_ν link model for $P(Z_i > 0) = T_\nu(X_i'\beta)$. Let us introduce the additional independent random variables λ_i, and write the distribution of Z_i as the following scale mixture of normal distribution:

$$Z_i|\lambda_i \ \sim\ N(X_i'\beta, \lambda_i^{-1}) \ \text{ and } \ \lambda_i \ \sim\ Gamma(\nu/2, 2/\nu), \ \ i = 1,\dots,n, \tag{8}$$

so that $Z_i \ \sim\ t_\nu(X_i'\beta, 1)$.

Under the hierarchical model, the above data augmentation scheme leads to the joint posterior density of the unobservables β, α, $\lambda = (\lambda_1,\dots,\lambda_n)'$ and $Z = (Z_1,\dots,Z_n)'$, given the data $Y = (Y_1,\dots,Y_n)'$:

$$f(\beta,\alpha,\lambda,Z|Y) \ =$$

$$C\prod_{j=1}^{k}\left[\frac{(2\pi\sigma_j^2)^{-1/2}}{\Phi(b_j/\sigma_j) - \Phi(a_j/\sigma_j)} \exp\left\{-\frac{\beta_j^2}{2\sigma_j^2}\right\} [I(\alpha_j = 0)I(a_j \le \beta_j \le b_j)]\right.$$

$$\begin{aligned} &+ \frac{(2\pi c_j^2\sigma_j^2)^{-1/2}}{\Phi(b_j/(c_j\sigma_j)) - \Phi(a_j/(c_j\sigma_j))} \exp\left\{-\frac{\beta_j^2}{2c_j^2\sigma_j^2}\right\}[I(\alpha_j = 1)I(a_j \leq \beta_j \leq b_j)]\Bigg], \\ &\times \prod_{j=1}^{k} q_j^{\alpha_j}(1-q_j)^{(1-\alpha_j)} \prod_{i=1}^{n}[\{I(Z_i > 0)I(Y_i = 1) + I(Z_i \leq 0)I(Y_i = 0)\} \\ &\times \phi(Z_i; X_i'\beta, \lambda_i^{-1})\delta(\nu)\lambda_i^{\nu/2-1}e^{-\nu\lambda_i/2}], \end{aligned} \tag{9}$$

where C here is a generic proportionality constant and $\delta(\nu) = [\Gamma(\nu/2)(2/\nu)^{\nu/2}]^{-1}$, $\phi(\cdot\,; a, b)$ is the $N(a, b)$ pdf and $\Phi(\cdot)$ is cdf of the standard normal distribution.

3.2. The Gibbs Sampler

Note that the joint posterior distribution (9) is complicated in the sense that it is difficult to normalize and directly sample from. But computation of respective marginal posterior distributions of β, Z_i's, λ_i's and α_j's using the Gibbs sampling algorithm requires only fully conditional distributions of them. The conditional distributions of β_j's involved in the algorithm are simple.

Given β_ℓ $(\ell \neq j)$, Z_i and λ_i, define

$$W_i = Z_i - \sum_{\ell \neq j} \beta_\ell x_{i\ell} \quad i = 1, \ldots, n. \tag{10}$$

Then the conditional distribution of β_j is the usual posterior density for the regression parameter in the normal linear model

$$W_i = \beta_j x_{ij} + e_i, \quad \text{where } e_i \overset{ind}{\sim} N(0, \lambda_i^{-1}), \tag{11}$$

obtained from proper $TN_{\{a_j \leq \beta_j \leq b_j\}}(0, \sigma_j^2)$ and $TN_{\{a_j \leq \beta_j \leq b_j\}}(0, c_j^2\sigma_j^2)$ priors of β_j for $\alpha_j = 0$ and $\alpha_j = 1$, respectively.

For $\alpha_j = 0$, the corresponding full conditional posterior kernel is

$$\begin{aligned} &\exp\left\{-\frac{\sum_{i=1}^n \lambda_i(W_i - \beta_j x_{ij})^2}{2} - \frac{\beta_j^2}{2\sigma_j^2}\right\} I(a_j \leq \beta_j \leq b_j) \\ \propto\ &\exp\left\{-\frac{(\beta_j - \tilde{\beta}_j)^2}{2\tilde{\sigma}_j^2}\right\} I(a_j \leq \beta_j \leq b_j), \end{aligned} \tag{12}$$

where

$$\tilde{\sigma}_j^2 = (1/\sigma_j^2 + \sum_{i=1}^n \lambda_i x_{ij}^2)^{-1} \text{ and } \tilde{\beta}_j = \tilde{\sigma}_j^2 \sum_{i=1}^n \lambda_i W_i x_{ij}.$$

This gives full conditional posterior distribution of β_j,

$$\beta_j | Y, Z, \lambda, \alpha, \beta_{(j)} \sim TN_{\{a_j \leq \beta_j \leq b_j\}}(\tilde{\beta}_j, \tilde{\sigma}_j^2), \text{ if } \alpha_j = 0;\ j = 1, \ldots, k, \tag{13}$$

where $\beta_{(j)} = (\beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_k)$.

Similarly, for $\alpha = 1$, we have

$$\beta_j | Y, Z, \lambda, \alpha, \beta_{(j)} \sim TN_{\{a_j \le \beta_j \le b_j\}}(\tilde{\beta}_j^*, \tilde{\sigma}_j^{*2}), \text{ if } \alpha_j = 1;\ j = 1, \ldots, k, \tag{14}$$

where

$$\tilde{\sigma}_j^{*2} = (1/(c_j^2\sigma_j^2) + \sum_{i=1}^n \lambda_i x_{ij}^2)^{-1} \text{ and } \tilde{\beta}_j^* = \tilde{\sigma}_j^{*2} \sum_{i=1}^n \lambda_i W_i x_{ij}.$$

If the prior distribution for β_j is not truncated (i.e. $a_j = -\infty,\ b_j = \infty$) then conditional on Z, λ, α, and $\beta_{(j)}$, $\beta_j \sim N(\tilde{\beta}_j, \tilde{\sigma}_j)$ for $\alpha_j = 0$ and $\beta_j \sim N(\tilde{\beta}_j^*, \tilde{\sigma}_j^{*2})$ for $\alpha_j = 1,\ j = 1, \ldots, k$.

Full conditional distributions of $Z_1, \ldots, Z_n$ are independently distributed as truncated normal distributions :

$$\begin{aligned} Z_i | Y, \beta, \lambda, \alpha &\sim TN_{\{Z_i > 0\}}(X_i'\beta, \lambda_i^{-1}), \text{ if } Y_i = 1, \\ Z_i | Y, \beta, \lambda, \alpha &\sim TN_{\{Z_i \le 0\}}(X_i'\beta, \lambda_i^{-1}), \text{ if } Y_i = 0, \end{aligned} \tag{15}$$

where $TN_{\{A\}}(X_i'\beta, \lambda_i^{-1})$ is the normal distribution $N(X_i'\beta, \lambda_i^{-1})$ truncated to the interval event A.

$\lambda_1, \ldots, \lambda_n$ are independent with

$$\lambda_i | \beta, Z, \alpha, Y \sim Gamma\left(\frac{\nu+1}{2}, \frac{2}{\nu + (Z_i - X_i'\beta)^2}\right). \tag{16}$$

Additional variables $\alpha_1, \ldots, \alpha_k$ are independently distributed as

$$\alpha_j | Y, Z, \beta, \lambda, \alpha_{(j)} \sim Be\left(\frac{f_j}{f_j + d_j}\right), \tag{17}$$

where $\alpha_{(j)} = (\alpha_1, \ldots, \alpha_{j-1}, \alpha_{j+1}, \ldots, \alpha_k)$, $Be(\gamma)$ denotes a Bernoulli distribution with parameter γ and

$$\frac{f_j}{f_j + d_j} = \frac{\exp\{-\beta_j^2/(2c_j^2\sigma_j^2)\}q_j}{\exp\{-\beta_j^2/(2c_j^2\sigma_j^2)\}q_j + \eta_j c_j \exp\{-\beta_j^2/(2\sigma_j^2)\}(1 - q_j)},$$

where

$$\eta_j = \frac{\Phi(b_j/(c_j\sigma_j)) - \Phi(a_j/(c_j\sigma_j))}{\Phi(b_j/\sigma_j) - \Phi(a_j/\sigma_j)},\ j = 1, \ldots, k.$$

Remark 1. $0 < \eta_j \le 1$ for $c_j > 1$, where the equality holds for $(a_j = -\infty, b_j = \infty)$, $(a_j = -\infty, b_j = 0)$ and $(a_j = 0, b_j = \infty)$. This is an effect of the truncated prior distribution in (5) that puts more possibility of j-th predictor variable being included in the variable selection. This effect tends to be more evident when a_j and b_j ($a_j \leq 0 \leq b_j$) take values near 0.

Remark 2. Form (8), it is easily seen that, as $\nu \to \infty$, $P(\lambda_i = 1) = 1, i = 1, \ldots, n$, because, for $\nu \to \infty$, the limit of the moment generating function of $Gamma(\nu/2, 2/\nu)$ is e^t. Thus, by fixing all λ_i's equal to 1, we can use the above Gibbs sampler for the variable selection in the probit regression.

Remark 3. Since t_8 random variable is approximately .634 times a logistic random variable, the above Gibbs sampler with $\nu = 8$ can be used for the variable selection in the logistic regression (cf. Albert and Chib 1993).

3.3. Subset Selection Scheme

The hierarchical nature of the model gives relatively straightforward implementation of the Gibbs sampling scheme as practiced by Geman and Geman (1984). A possible complication could be the simulation from truncated normal distribution. This can be easily resolved by the algorithm of Devroye (1986).

The Gibbs sample of α can be used to compute an empirical distribution which converges to the actual marginal posterior $h(\alpha|Y)$ (cf. Casella and George 1992 and Tierney 1994). In particular, the empirical distribution of the α would have following implications:

(i) The distribution corresponding to the most promising subsets of $x_1, \ldots, x_k$ will appear with the highest frequency, because it is just those values which have largest probability under $h(\alpha|Y)$.

(ii) The low-frequency or zero-frequency values of α may simply be ignored, because these correspond to the least promising models.

(iii) Even when no high-frequency values of α appeared in the empirical distribution, the marginal distribution of α may contain useful information for model selection. The marginal distribution may clearly show us that some predictor variables are inactive.

These imply that a simple tabulation of the high-frequency values of α can be used to identify the corresponding subsets of predictors as potentially promising.

4. Numerical Examples

The objectives in these examples are to demonstrate a convenient method for the formulation of priors, illustrate favorable performance of the procedure, and study the relation between prior and posterior distributions for the coefficients of some predictor variables.

This example treats problems involving k=12 potential predictors with constrained coefficients. The predictors were obtained as independent standard normal variables, $x_1, \ldots, x_{12} \overset{iid}{\sim} N(0,1)$, so that they were practically uncorrelated. The dependent variables were generated according to a probit model and a logistic model:

$$p_i = P(Y_i = 1) = \Phi(\beta_1 x_{i1} + \beta_1 x_{i1} + \beta_{12} x_{i12}), \qquad (18)$$

$$p_i = P(Y_i = 1) = \frac{\exp\{\beta_1 x_{i1} + \beta_1 x_{i1} + \beta_{12} x_{i12}\}}{1 + \exp\{\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i12}\}}. \qquad (19)$$

For the example, two cases of coefficient values and constraints are considered.

Case 1: $\beta_1 = -2$, $\beta_2 = -2$ and $\beta_{12} = 0.1$ with constraints $\beta_1 \leq 0$, $\beta_2 \leq 0$ and $0.01 \leq \beta_{12} \leq 0.5$.

Case 2: Case 1: $\beta_1 = 4$, $\beta_2 = 4$ and $\beta_{12} = 0.1$ with constraints $\beta_1 \geq 0$, $\beta_2 \geq 0$ and $0.01 \leq \beta_{12} \leq 0.5$.

Thus $\beta = (\beta_1, \beta_2, 0, 0, 0, 0, 0, 0, 0, 0, 0, \beta_{12})'$. We applied the suggested variable selection method with the indifference priors for the second and third hierarchy of the model suggested in Section 2. The indifference priors are constructed as follows

$$P(\alpha_j = 1) = q_j = q, \quad \text{for } j = 1, \ldots, 11,$$

$$\sigma_j = \sigma, \ c_j = c, \quad \text{for } j = 1, \ldots, 12.$$

We set $q_{12} = 1$ to reflect the constraint $0.01 \leq \beta_{12} \leq 0.5$ in the prior for the second and third hierarchy of the model. Different prior beliefs will, of course, lead to other choices for q_j, σ_j and c_j. For instance, it is thought that a certain predictor may not be enter the model at all, the corresponding σ_j and q_j would be smaller, while c_j would be larger and their values may be set employing the same kind of reasoning about marginal effects. We considered various choices of the hyperparameters for the indifference priors. For each σ_j, we considered the low and high settings, $\sigma_j = .3$ and $\sigma_j = .5$. For each c_j we considered the low and high settings, $c_j = 4$ and $c_j = 9$. These choices provided substantial separation between the two mixture components in (5) while still allowing for plausible values of β_j when $\alpha_j = 1$. As a base probability that each predictor is included in the model, we took $q_j = .2$ except for $q_{12} = 1$. To study the relation between the prior and posterior distribution of α, we also considered $q_j = .1$ and $q_j = .5$ (setting $q_{12} = 1$). Thus we set up following twelve priors for the example.

Table 1. Twelve Priors

prior	1	2	3	4	5	6	7	8	9	10	11	12
q	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.5	0.5	0.5	0.5
σ	0.3	0.3	0.5	0.5	0.3	0.3	0.5	0.5	0.3	0.3	0.5	0.5
c	4	9	4	9	4	9	4	9	4	9	4	9

Using SAS/IML we generated two artificial data sets of each size 50 from the models (19) and (20) with given values of β_j's, and ran twelve parallel chains of the Gibbs sampler for t_{100} link model and t_8 model (formulated by using each prior in Table 1), respectively. The parallel chains were obtained by differing starting points overdispersed to provide good coverage of the posterior. Twelve sets of starting points considered for each model were combinations of following parameter values:

(i) β_j, $j = 1, \ldots, 12$: mle of β_j, mle + (s.d. of mle), mle − (s.d. of mle);

(ii) $(\alpha_1, \ldots, \alpha_{12})$: (0, ..., 0), (0, 1, 0, 1, ..., 0, 1), (1, 0, 1, 0, ..., 1, 0), (1, ..., 1);

(iii) $(\lambda_1, \ldots, \lambda_{12})$: (1, ..., 1).

Here mle denotes the unconstrained maximum likelihood estimate of β_j for the probit model. We obtained 24 ((2 different values of (β_1, β_2, β_{12})) × 12 (priors)) sets of the twelve parallel chains from running the Gibbs sampler. By use of plots option of the "CODA Output Analysis Menu" by Best et al. (1996), we got result of the diagnostic checks (outlined in Cowles and Carlin, 1996) to each set of parallel chains. This gave a clear indication that convergence was achieved within 1000 iterations for all the models.

Using the same artificial data set of size n=50, a Gibbs sample of m=1000 observations from the Gibbs sequence was obtained from each Gibbs sampler. The sampling scheme adopted here was to allow initial 1000 iterations for "burn-in" and then to pick up every 10th observation until Gibbs sample of size m=1000 was collected. For each Gibbs sampling, we used corresponding unconstrained mle for $\beta^{(0)}$, $\alpha_j^{(0)} = 1$ and $\lambda_j^{(0)} = 1$, $j = 1, \ldots, 12$, as starting values.

The variable selection results of the t link model for each artificial data set (generated from (19) and (20) are noted in Table 2 and Table 3. They display respective three high-frequency probit and logistic models corresponding to the frequencies of $\alpha = (\alpha_1, \ldots, \alpha_{12})'$ that appeared for each prior. In each case of the priors, the true model is included in the first three high-frequency values among 2^{12} different frequency values of α, suggesting reasonable robustness with respect to prior specifications. Aside from the robustness, the tables note the following implications: (i) They show how the suggested variable selection method successful in identifying several promising models rather than the single best model. This feature is similar to the way in which stepwise methods are used to narrow the scope of model selection. (ii) For every prior, the true model is included in three most probable models selected. However, under the same hyperparameters given, algorithm for the probit model seemed to favor more saturated models than that for the logistic model. This fact coincides with Remark 3. (iii) Although all the frequencies of the 2^{12} possible models are not presented in the tables, it is seen that, for fixed c_j and σ_j, q_j get smaller, the Gibbs sampler tends to select smaller model than larger q_j does. On the other hand, for fixed σ_j and q_j, high setting of c_j uniformly yields higher frequency for the true model than low setting of c_j does.

5. Concluding Remarks

The present paper has developed and illustrated a Bayesian approach to narrow the scope of possible models in the variable selection for a class of the binary response t link regression models. Though the suggested approach would not directly lead to a single best fitting model, it is demonstrated as a way to save the overwhelming job of comparing all the 2^k possible submodels for the t link regression model having k predictor variables. Thus, as an alternative to usual optimal subset selection procedure (involving the overwhelming comparisons of all 2^k possible subset models), a two-stage variable selection procedure can be constructed: First, select $m << 2^k$ promising subset models via the suggested approach. In the second stage, choose a best fitting model by means of usual variable selection criteria such as AIC, BIC, the deviance criterion (Collett 1991) and the marginal likelihood by Chip (1995). For the full Bayesian two-stage procedure, we may adopt the marginal likelihood criterion in the second stage.

Table 2. High Frequency Probit Models (Approximation by t_{100} Link Model)

Probit Model	Case 1		Case 2	
	Selected variables	prop. (%)	Selected variables	prop. (%)
prior 1	$x_1x_2x_{12}$	58.2	$x_1x_2x_{12}$	54.4
	x_1x_{12}	6.8	x_1x_{12}	11.8
	$x_1x_2x_8x_{12}$	5.3	$x_1x_2x_4x_{12}$	4.6
prior 2	$x_1x_2x_{12}$	71.0	$x_1x_2x_{12}$	71.8
	$x_1x_2x_8x_{12}$	10.0	$x_1x_2x_8x_{12}$	6.1
	$x_1x_2x_{11}x_{12}$	2.4	$x_1x_2x_6x_{12}$	4.5
prior 3	$x_1x_2x_{12}$	61.9	$x_1x_2x_{12}$	50.5
	x_1x_{12}	5.3	x_1x_{12}	14.4
	$x_1x_2x_8x_{12}$	5.0	$x_1x_2x_4x_{12}$	5.8
prior 4	$x_1x_2x_{12}$	75.5	$x_1x_2x_{12}$	56.7
	$x_1x_2x_8x_{12}$	8.3	x_1x_{12}	22.3
	$x_1x_2x_5x_{12}$	2.0	$x_1x_2x_4x_{12}$	2.3
prior 5	$x_1x_2x_{12}$	33.4	$x_1x_2x_{12}$	30.9
	$x_1x_2x_8x_{12}$	8.3	$x_1x_2x_5x_{12}$	5.9
	$x_1x_2x_{10}x_{12}$	4.0	$x_1x_2x_8x_{12}$	5.5
prior 6	$x_1x_2x_{12}$	54.9	$x_1x_2x_{12}$	42.5
	$x_1x_2x_8x_{12}$	11.0	$x_1x_2x_8x_{12}$	9.5
	$x_1x_2x_4x_{12}$	3.8	$x_1x_2x_5x_{12}$	5.3
prior 7	$x_1x_2x_{12}$	38.6	$x_1x_2x_{12}$	33.4
	$x_1x_2x_8x_{12}$	7.4	$x_1x_2x_4x_{12}$	7.7
	$x_1x_2x_9x_{12}$	4.0	$x_1x_2x_8x_{12}$	5.6
prior 8	$x_1x_2x_{12}$	54.2	$x_1x_2x_{12}$	54.6
	$x_1x_2x_8x_{12}$	13.1	$x_1x_2x_6x_{12}$	8.6
	$x_1x_2x_5x_{12}$	4.1	$x_1x_2x_8x_{12}$	4.3
prior 9	$x_1x_2x_{12}$	6.3	$x_1x_2x_{12}$	4.3
	$x_1x_2x_8x_{12}$	2.5	$x_1x_2x_4x_{12}$	3.0
	$x_1x_2x_5x_{12}$	1.6	$x_1x_2x_8x_{12}$	2.1
prior 10	$x_1x_2x_{12}$	6.3	$x_1x_2x_{12}$	8.6
	$x_1x_2x_8x_{12}$	5.4	$x_1x_2x_8x_{12}$	5.3
	$x_1x_2x_5x_{12}$	3.4	$x_1x_2x_4x_{12}$	4.5
prior 11	$x_1x_2x_{12}$	4.0	$x_1x_2x_{12}$	3.5
	$x_1x_2x_8x_{12}$	3.1	$x_1x_2x_4x_{12}$	3.1
	$x_1x_2x_5x_{12}$	2.6	$x_1x_2x_5x_{12}$	2.8
prior 12	$x_1x_2x_{12}$	7.5	$x_1x_2x_{12}$	6.8
	$x_1x_2x_8x_{12}$	7.3	$x_1x_2x_5x_8x_{12}$	5.6
	$x_1x_2x_4x_8x_{12}$	5.8	$x_1x_2x_4x_6x_8x_{12}$	5.2

Table 3. High Frequency Logistic Models (Approximation by t_8 Link Model)

Logistic Model	Case 1		Case 2	
	Selected variables	prop. (%)	Selected variables	prop. (%)
prior 1	$x_1x_2x_{12}$	34.0	$x_1x_2x_{12}$	54.8
	x_1x_{12}	24.6	x_1x_{12}	9.0
	x_1	3.9	$x_1x_2x_6x_{12}$	4.5
prior 2	$x_1x_2x_{12}$	59.9	$x_1x_2x_{12}$	74.6
	x_1x_{12}	17.4	$x_1x_2x_{11}x_{12}$	3.2
	x_{12}	3.6	$x_1x_2x_5x_{12}$	3.1
prior 3	$x_1x_2x_{12}$	34.9	$x_1x_2x_{12}$	62.2
	x_{12}	24.2	x_1x_{12}	7.2
	x_1x_{12}	9.0	$x_1x_2x_5x_{12}$	4.2
prior 4	$x_1x_2x_{12}$	56.0	$x_1x_2x_{12}$	79.3
	x_1x_{12}	23.2	$x_1x_2x_5x_{12}$	3.6
	x_{12}	3.9	$x_1x_2x_4x_{12}$	3.2
prior 5	$x_1x_2x_{12}$	17.5	$x_1x_2x_{12}$	37.3
	x_1x_{12}	16.7	$x_1x_2x_5x_{12}$	13.1
	$x_1x_3x_{12}$	5.0	$x_1x_2x_4x_{12}$	4.4
prior 6	$x_1x_2x_{12}$	34.3	$x_1x_2x_{12}$	47.9
	x_1x_{12}	23.7	$x_1x_2x_5x_{12}$	13.1
	$x_1x_3x_{12}$	5.0	$x_1x_2x_4x_{12}$	4.4
prior 7	$x_1x_2x_{12}$	22.9	$x_1x_2x_{12}$	35.1
	x_1x_{12}	16.2	$x_1x_2x_6x_{12}$	6.9
	x_{12}	5.8	$x_1x_2x_5x_{12}$	6.3
prior 8	$x_1x_2x_{12}$	33.3	$x_1x_2x_3$	52.1
	x_{12}	20.6	$x_1x_2x_5x_{12}$	8.9
	x_1x_{12}	13.3	$x_1x_2x_6x_{12}$	7.8
prior 9	$x_1x_2x_{12}$	3.8	$x_1x_2x_{12}$	3.4
	$x_1x_2x_5x_{12}$	2.3	$x_1x_2x_5x_{12}$	2.6
	$x_1x_2x_3x_{12}$	2.2	$x_1x_2x_{10}x_{12}$	1.8
prior 10	$x_1x_2x_{12}$	12.5	$x_1x_2x_{12}$	10.6
	$x_1x_2x_3x_{12}$	4.7	$x_1x_2x_{11}x_{12}$	5.2
	$x_1x_2x_4x_{12}$	4.3	$x_1x_2x_5x_{12}$	4.6
prior 11	$x_1x_2x_{12}$	3.8	$x_1x_2x_{12}$	3.6
	$x_1x_2x_3x_{12}$	2.5	$x_1x_2x_8x_{12}$	2.8
	$x_1x_2x_4x_{12}$	2.0	$x_1x_2x_5x_{12}$	2.5
prior 12	$x_1x_2x_{12}$	14.8	$x_1x_2x_{12}$	12.0
	x_1x_{12}	7.8	$x_1x_2x_5x_{12}$	6.3
	$x_1x_2x_4x_{12}$	4.3	$x_1x_2x_6x_{12}$	6.2

REFERENCES

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data, *Journal of the American Statistical Association,* **88**, 669-679.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian theory,* Wiley, New York.

Best, N., Cowles, M. K., and Vines, K. (1996). *CODA; Convergence diagnosis and output analysis software for Gibbs sampling output version 0.30,* MRC Biostatistics Unit, Cambridge.

Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler, *The American Statistician,* **46**, 167-174.

Chib, S. (1995) Marginal likelihood from the Gibbs output, *Journal of the American Statistical Association,* **90**, 1313-1321.

Collett, D. (1991). *Modelling binary data,* Chapman and Hall, New York.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review, *Journal of the American Statistical Association,* **91**, 883-904.

Dellaportas, P., Forster, J. J. and Ntzoufras, I. (1997). On Bayesian model and variable selection using MCMC, available on the MCMC preprint server.

Devroye, L. (1986). *Non-uniform random generation,* Springer Verlag, New York.

Finney, D. J. (1971). *Probit analysis,* 3rd ed., Cambridge University Press, Cambridge.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Transactions; Pattern Analysis and Machine Intelligence,* **6**, 721-741.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations, *Journal of the Royal Statistical Society,* **B 56**, 501-514.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion), *Statistical Science,* **7**, 457-511.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling, *Journal of the American Statistical Association,* **88**, 881-889.

George, E. I. and McCulloch, R. E. (1996). Stochastic search variable selection, In *Markov Chain Monte Carlo in Practice,* Eds. W. R. Gilks, S. Richardson and D. S. Spiegelhalter, Chapman and Hall, New York, 203-214.

George, E. I. and McCulloch, R. E., and Tsay, R. (1996). Two approaches to Bayesian model selection with applications, In *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner,* Eds. D. A. Berry, K. M. Chaloner an J. K. Geweke, Wiley, New York.

Geweke, J. (1996). Variable selection and model comparison in regression, In *Bayesian Statistics 5,* Eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, Oxford: University Press, 609-620.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling, *Applied Statistics,* **41**, 337-348.

Green, P. J. (1994). Reversible jump MCMC computation and Bayesian model determination, *Technical Report 5-94-03,* University of Bristol.

Kuo, L. and Mallick, B. (1997). Variable selection for regression models, *Sankhya,* to appear, available on the MCMC preprint server.

Soofi, S. S., Ebrahimi, N., and Habibullah, M. (1995). Information distinguishability with application to analysis of failure data, *Journal of the American Statistical Association,* **90**, 657-668.

Tanner, T. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation, *Journal of the American Statistical Association,* **82**, 528-549.

Tierney, L. (1994). Markov chains for exploring posterior distributions, *Annals of Statistics,* **22**, 1701-1762.