Bayesian Analysis of Binary and Polychotomous Response Data

JAMES H. ALBERT and SIDDHARTHA CHIB*

A vast literature in statistics, biometrics, and econometrics is concerned with the analysis of binary and polychotomous response data. The classical approach fits a categorical response regression model using maximum likelihood, and inferences about the model are based on the associated asymptotic theory. The accuracy of classical confidence statements is questionable for small sample sizes. In this article, exact Bayesian methods for modeling categorical response data are developed using the idea of data augmentation. The general approach can be summarized as follows. The probit regression model for binary outcomes is seen to have an underlying normal regression structure on latent continuous data. Values of the latent data can be simulated from suitable truncated normal distributions. If the latent data are known, then the posterior distribution of the parameters can be computed using standard results for normal linear models. Draws from this posterior are used to sample new latent data, and the process is iterated with Gibbs sampling. This data augmentation approach provides a general framework for analyzing binary regression models. It leads to the same simplification achieved earlier for censored regression models. Under the proposed framework, the class of probit regression models can be enlarged by using mixtures of normal distributions to model the latent data. In this normal mixture class, one can investigate the sensitivity of the parameter estimates to the choice of "link function," which relates the linear regression estimate to the fitted probabilities. In addition, this approach allows one to easily fit Bayesian hierarchical models. One specific model considered here reflects the belief that the vector of regression coefficients lies on a smaller dimension linear subspace. The methods can also be generalized to multinomial response models with $J > 2$ categories. In the ordered multinomial model, the J categories are ordered and a model is written linking the cumulative response probabilities with the linear regression structure. In the unordered multinomial model, the latent variables have a multivariate normal distribution with unknown variance–covariance matrix. For both multinomial models, the data augmentation method combined with Gibbs sampling is outlined. This approach is especially attractive for the multivariate probit model, where calculating the likelihood can be difficult.

KEY WORDS: Binary probit; Data augmentation; Gibbs sampling; Hierarchical Bayes modeling; Latent data; Logit model; Multinomial probit; Residual analysis; Student-t link function.

1. INTRODUCTION

Suppose that N independent binary random variables $Y_1, \ldots, Y_N$ are observed, where Y_i is distributed Bernoulli with probability of success p_i. The p_i are related to a set of covariates that may be continuous or discrete. Define the binary regression model as $p_i = H(\mathbf{x}_i^T\beta)$, $i = 1, \ldots, N$, where β is a $k \times 1$ vector of unknown parameters, $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ik})$ is a vector of known covariates, and $H(\ \ )$ is a known cdf linking the probabilities p_i with the linear structure $\mathbf{x}_i^T\beta$. The probit model is obtained if H is the standard Gaussian cdf, whereas the logit model is obtained if H is the logistic cdf. (For general discussions of this class of models, see Cox 1971, Finney 1947, Nelder and McCullagh 1989, and Maddala 1983.)

Let $\pi(\beta)$, a proper or improper prior density, summarize our prior information about β. Then the posterior density of β is given by

$$\pi(\beta \mid \text{data}) = \frac{\pi(\beta)\prod_{i=1}^{N} H(\mathbf{x}_i^T\beta)^{y_i}(1 - H(\mathbf{x}_i^T\beta))^{1-y_i}}{\int \pi(\beta)\prod_{i=1}^{N} H(\mathbf{x}_i^T\beta)^{y_i}(1 - H(\mathbf{x}_i^T\beta))^{1-y_i}\, d\beta}, \quad (1)$$

which is largely intractable. Letting $N_k(\mu, \Sigma)$ denote the k-variate multivariate normal distribution with mean μ and variance-covariance matrix Σ, the usual asymptotic approximation is that β is distributed $N_k(\hat{\beta}, \mathbf{I}(\hat{\beta})^{-1})$, where $\hat{\beta}$ is the posterior mode and $\mathbf{I}(\hat{\beta})$ is the negative of the second derivative matrix evaluated at the mode. When a uniform prior is chosen for β, $\hat{\beta}$ is the maximum likelihood estimate (MLE) and $\mathbf{I}(\ \ )$ is the observed information matrix. From a non-Bayesian viewpoint, Griffiths, Hill, and Pope (1987) found the MLE to have significant bias for small samples. Zellner and Rossi (1984), from a Bayesian approach, also commented on the inaccuracy of the normal approximation for small N. For a small number of parameters, they summarized the posterior using numerical integration. For large models (k large), they computed posterior moments by Monte Carlo integration with a multivariate Student's t importance function.

In this article we introduce a simulation-based approach for computing the exact posterior distribution of β. Suppose that the link function H is the standard Gaussian cdf (the probit case). The key idea is to introduce N independent latent variables $Z_1, \ldots, Z_N$ into the problem, where Z_i is distributed $N(\mathbf{x}_i^T\beta, 1)$, and define $Y_i = 1$ if $Z_i > 0$ and $Y_i = 0$ if $Z_i \le 0$. Observe that if the Z_i are known and a multivariate normal prior is chosen for β, then the posterior distribution for β can be derived using standard normal linear model results. The Z_i are of course unknown; however, given the data Y_i, the distribution of Z_i follows a truncated normal distribution. These principal observations, combined with the tool of Gibbs sampling, allow us to simulate from the exact posterior distribution of β. This approach is very similar to the data augmentation/Gibbs sampling framework used in censored regression models (Chib 1992; Wei and Tanner 1990).

* James H. Albert is Professor, Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH 43403. Siddhartha Chib is Associate Professor of Economics at the Olin School of Business, Washington University, St. Louis, MO 63130. The work was completed while the second author held a joint visiting appointment with the Economics Department and the Olin School of Business at Washington University. The authors thank the editor, the associate editor, and two referees for many helpful comments.

This approach connects the probit binary regression model on the Y_i with a normal linear regression model on the latent data Z_i. This framework also permits us to elaborate the probit model by using mixtures of normal distributions to model the latent variables. In Section 3 we use this device to model binary data using a t link function. Application of this link family in an example in Section 5.2 suggests that inferences can be sensitive to the choice of link function. (See Carlin and Polson 1991 for a similar use of mixtures of normal distributions in modeling the sampling density.) Our framework also makes it easy to model uncertainty about a particular probit model by means of a hierarchical normal linear structure on β. In the example of Section 5.2, we use this model to check the hypothesis that some covariates may be removed from the model with little change in the fit.

The sampling approach allows one to compute marginal posterior distributions of many parameters of interest. As an illustration, in Section 5.2 we compute the set of posterior distributions of the residuals $y_i - \mathbf{x}_i^T\beta$. For binary data, the usual frequentist definition of residual can take on only two possible values. In contrast, the Bayesian residual has a continuous distribution on an interval and thus can be more informative than the usual frequentist residual in detecting outliers.

In the preceding discussion we restrict the observation Y_i to two values. Suppose that Y_i has $J > 2$ categories. In Section 4 we show how the preceding data augmentation/Gibbs sampling strategy can be generalized to handle multinomial data. In the first situation the categories are assumed ordered, and the linear regression structure is linked to the cumulative response probabilities. In the second case the categories are unordered and linked to a multivariate normal structure on the latent data. In each situation the Gibbs sampling algorithm can be generalized straightforwardly to simulate from the posterior distribution of the regression parameter of interest.

This article is organized as follows. Section 2 outlines the Gibbs sampling algorithm used in simulating the posterior distributions. Section 3 discusses the binary regression models based on normal and mixtures of normal linear models on the latent data. Section 4 presents some analogous models for multinomial response data. Section 5.1 uses the data set of Finney (1947), with three covariates, as a benchmark to compare the posterior densities with those computed using numerical integration. Section 5.2 contains a detailed illustration with a seven-covariate data set, which illustrates the generalizations of the usual probit model. Section 5.3 summarizes Bayesian calculations for the trivariate probit model of Daganzo (1979). Finally, Section 6 presents some concluding remarks.

2. THE GIBBS SAMPLER

In this section we review the Gibbs sampling algorithm (Gelfand and Smith 1990) with a focus on its implementation in the binary and polychotomous response models. One is interested in simulating from the posterior distribution of θ partitioned into the vector components $\theta = (\theta_1, \ldots, \theta_p)$. Although it may be difficult to sample from the joint posterior, suppose that it is easy to simulate from the fully conditional distributions $\pi(\theta_k | \{\theta_j, j \neq k\})$. To implement the Gibbs sampler, one starts with initial guesses of the θ_i—call them $\theta_1^{(0)}, \ldots, \theta_p^{(0)}$—and then simulates in turn

$$\begin{aligned} &\theta_1^{(1)} \quad \text{from} \quad \pi(\theta_1 | \{\theta_j^{(0)}, j \neq 1\}) \\ &\theta_2^{(1)} \quad \text{from} \quad \pi(\theta_2 | \theta_1^{(1)}, \{\theta_j^{(0)}, j > 2\}) \\ &\vdots \\ &\theta_p^{(1)} \quad \text{from} \quad \pi(\theta_p | \{\theta_j^{(1)}, j < p\}). \end{aligned} \tag{2}$$

The cycle (2) is iterated t times, generating the sample $\theta^{(t)} = (\theta_1^{(t)}, \ldots \theta_p^{(t)})$. As t approaches infinity, the joint distribution of $\theta^{(t)}$ can be shown to approach the joint distribution of θ. So for sufficiently large t, say t^*, $\theta^{(t^*)}$ can be regarded as one simulated value from the posterior of θ. Replicating this process m times gives the sample $\{(\theta_{1j}^{(t^*)}, \theta_{2j}^{(t^*)}, \ldots, \theta_{pj}^{(t^*)}), j = 1, \ldots, m\}$, which can be used to compute posterior moments and density estimates.

There are two practical drawbacks to this replication approach. First, the method is inefficient, because the samples $\{\theta_j^{(t)}\}$, for $t < t^*$ are discarded. Second, after the initial run it may be necessary to repeat the simulation with a larger number of replications to get accurate density estimates. This is unsatisfactory in that the observations in the initial run are discarded.

Here we propose a "one-run" Gibbs sampling scheme, suggested by Zeger and Karim (1991), that is efficient in that few observations are discarded. Only one replication is used, and the cycle (2) is run a large number of times, with the sequence extended until convergence. It is difficult to make general comments on the convergence behavior of this Gibbs sample because, from our experience, it appears that the rate of convergence depends on the particular application. But we can make some comments on monitoring this one-run method for the applications described in this article.

First, the general objective of the sampling is to collect a sufficiently large number of values from the joint posterior of θ to obtain accurate estimates of marginal posterior densities of subsets of θ and posterior moments. One should collect the values starting at the cycle t where one believes that $\theta^{(t)}$ is approximately a simulated value from the posterior of θ. In the examples presented here, the value of t is small (say 10–40) relative to the total number of values collected. Thus one would get similar convergent behavior by collecting all of the simulated values.

Second, there are typically strong positive correlations between the values $\theta^{(t)}$ and $\theta^{(t+1)}$. If one wishes to obtain an approximate independent sample of the θ, the simulated values of θ could be collected at cycles t, $t + n_1$, $t + 2n_1$, and so on, where n_1 is the spacing between cycles where $\theta^{(t)}$ and $\theta^{(t+n_1)}$ are believed to be approximately independent. But it is not necessary to obtain an independent sample of θ to obtain, say, a marginal posterior density estimate of θ_k.

One goal of this article is to obtain estimates of the densities of the individual parameters or their functionals. Suppose that the function $g(\theta_k)$ is of interest. One can estimate the density of this function using a kernel density estimate of the simulated values of $g(\theta_k)\{g(\theta_k^{(i)}), i = 1, \ldots, m\}$. Gelfand and Smith (1990) argued that a slightly preferable estimate of this marginal posterior density is given by $\hat{\pi}(g(\theta_k)) \approx (1/m) \sum_{i=1}^{m} \pi(g(\theta_k) | \{\theta_r^{(i)}, r \neq k\}))$. To detect convergence of the Gibbs sample in practice, we collect values

of θ in batches of 100–200 until all the marginal density estimates for the components of θ stabilize.

A second goal is estimation of posterior expectations. Suppose that the expectation of interest is $E[g(\theta_k)]$. From the simulated sample, this posterior expectation can be estimated by either the sample mean of $\{g(\theta_k^{(i)})\}$ or the sample mean of $\{E[g(\theta_k)|\{\theta_r^{(i)}, r \neq k\}]\}$. In either case it will be of interest to compute a simulation standard error for the sample mean estimate. To compute this standard error from this correlated simulation sample, we apply the well-known batch means method (see, for example, Bratley, Fox, and Schrage 1987). We batch or section the sample into subsamples of equal size. When the lag one autocorrelation of the batch means is under .05, the simulation standard error is computed as the standard deviation of the batch means divided by the square root of the number of batches.

The Gibbs sampler requires simulation from the p fully conditional posterior distributions (2). In the applications described here, these distributions are not all of standard functional forms (e.g., multivariate normal, gamma) and can be difficult to simulate. Devroye (1986) described some general acceptance algorithms for sampling from nonconjugate distributions. In the examples presented here, we use two schemes that are easier to implement than those algorithms and appear to work well in practice. Instead of sampling from a continuous posterior density $\pi(\theta)$, π is discretized into k mass points of interest, and then (using inversion techniques) the discretized version of π is sampled. By choosing k sufficiently large, one can adequately approximate many continuous densities. Alternately, one can approximate π by a matching normal distribution with the same mode and curvature at mode as π and then sample from this normal distribution. This particular method may perform poorly if π has tails significantly flatter than a normal distribution. In this case one can approximate π by a matching t distribution or mixture of t distributions. (See West 1992 for a discussion of using mixtures of t distributions in this context.)

3. DATA AUGMENTATION AND GIBBS SAMPLING FOR BINARY DATA

3.1 Introduction

To introduce the data augmentation approach (Tanner and Wong 1987), let $H = \Phi$, leading to the probit model. Introduce N latent variables $Z_1, \ldots, Z_N$, where the Z_i are independent $N(\mathbf{x}_i^T\beta, 1)$, and define $Y_i = 1$ if $Z_i > 0$ and $Y_i = 0$ otherwise. It can be easily shown that the Y_i are independent Bernoulli random variables with $p_i = P(Y_i = 1) = \Phi(\mathbf{x}_i^T\beta)$.

The joint posterior density of the unobservables β and $\mathbf{Z} = (Z_1, \ldots, Z_N)$ given the data $\mathbf{y} = (y_1, \ldots, y_N)$ is given by

$$\pi(\beta, \mathbf{Z}|\mathbf{y}) = C\pi(\beta)\prod_{i=1}^{N}\{1(Z_i > 0)1(y_i = 1) + 1(Z_i \leq 0)1(y_i = 0)\} \times \phi(Z_i; \mathbf{x}_i^T\beta, 1). \quad (3)$$

In (3), $\phi(\ ; \mu, \sigma^2)$ is the $N(\mu, \sigma^2)$ pdf, $1(X \in A)$ is the indicator function that is equal to 1 if the random variable X is contained in the set A, and C here and henceforth is a generic proportionality constant. Note that this joint distribution is complicated in the sense that it is difficult to normalize and sample from directly. But computation of the marginal posterior distribution of β using the Gibbs sampling algorithm requires only the posterior distribution of β conditional on $\mathbf{Z}$ and the posterior distribution of $\mathbf{Z}$ conditional on β, and these fully conditional distributions are of standard forms. First, note from (3) that the posterior density of β given $\mathbf{Z}$ is given by

$$\pi(\beta|\mathbf{y}, \mathbf{Z}) = C\pi(\beta)\prod_{i=1}^{N}\phi(Z_i; \mathbf{x}_i^T\beta, 1). \quad (4)$$

This fully conditional posterior density is the usual posterior density for the regression parameter in the normal linear model $\mathbf{Z} = \mathbf{X}\beta + \varepsilon$, where $\mathbf{X} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_N^T)^T$ and ε is distributed $N_N(0, \mathbf{I})$, where $\mathbf{I}$ is the identity matrix. Using standard linear model results, if a priori the distribution of β is diffuse, then

$$\beta|\mathbf{y}, \mathbf{Z} \text{ is distributed } N_k(\hat{\beta}_Z, (\mathbf{X}^T\mathbf{X})^{-1}), \quad (5)$$

where $\hat{\beta}_Z = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Z})$. If β is assigned the proper conjugate $N(\beta^*, \mathbf{B}^*)$ prior, then the posterior distribution of β given $\mathbf{Z}$ is $N_k(\tilde{\beta}, \tilde{\mathbf{B}})$, where the posterior mean and covariance matrix are given by $\tilde{\beta} = (\mathbf{B}^{*-1} + \mathbf{X}^T\mathbf{X})^{-1}(\mathbf{B}^{*-1}\beta^* + \mathbf{X}^T\mathbf{Z})$ and $\tilde{\mathbf{B}} = (\mathbf{B}^{*-1} + \mathbf{X}^T\mathbf{X})^{-1}$. Next, note from (3) that the posterior distribution of $\mathbf{Z}$, conditional on β, also has a simple form. The random variables $Z_1, \ldots, Z_N$ are independent with

$$Z_i|\mathbf{y}, \beta \text{ distributed } N(\mathbf{x}_i^T\beta, 1) \text{ truncated at the left by 0} \quad \text{if } y_i = 1$$

$$Z_i|\mathbf{y}, \beta \text{ distributed } N(\mathbf{x}_i^T\beta, 1) \text{ truncated at the right by 0} \quad \text{if } y_i = 0. \quad (6)$$

In practice it is customary to assign a flat noninformative prior to β. Given a previous value of β, one cycle of the Gibbs algorithm would produce $\mathbf{Z}$ and β from the distributions (6) and (5). The starting value of β, $\beta^{(0)}$ may be taken to be the maximum likelihood (ML) estimate, or alternatively the least squares (LS) estimate $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Note that it is computationally easy to simulate from both the multivariate normal distribution (5) and the truncated normal distributions in (6) (see Devroye 1986 for simulation algorithms).

3.2 The t Link

By introducing the Z_i's into the model, the probit regression model on the Bernoulli observations $\mathbf{Y}$ is seen to have an underlying normal regression on $\mathbf{Z}$. Since the posterior distribution of β given $\mathbf{Z}$ is multivariate normal, it is possible to generalize this model by applying suitable mixtures of normal distributions.

For example, one can generalize the probit link by choosing the link cdf H to be the family of t distributions. This generalization allows one to investigate the sensitivity of the fitted probabilities to the choice of link function. In addition, one can see which value of the t degrees of freedom parameter

is best supported by the data. The most popular link function for binary data is the logit, which corresponds to a choice of a logistic distribution for H. Figure 1 plots quantiles of the logistic distribution against quantiles of a t distribution for various degrees of freedom. Note that, for probabilities between .001 and .999, logistic quantiles are approximately a linear function of $t(8)$ quantiles. This statement is consistent with Mudholkar and George's (1978) result that the logistic distribution has the same kurtosis as a t distribution with 9 df. Thus, approximately, one can view the logistic distribution as a member of the t family. The connection between the logistic and t links will be further explored in the example in Section 5.2.

Let the Z_i be independently distributed from t distributions with locations $\mathbf{x}_i^T\beta$, scale parameter 1, and degrees of freedom ν. Equivalently, with the introduction of the additional random variable λ_i, we write the distribution of Z_i as the following scale mixture of a normal distribution: $Z_i \mid \lambda_i$ is distributed $N(\mathbf{x}_i^T\beta, \lambda_i^{-1})$ and λ_i is distributed Gamma $(\nu/2, 2/\nu)$ with pdf proportional to $\lambda_i^{\nu/2-1}\exp(-\nu\lambda_i/2)$.

Suppose a uniform prior is chosen for the regression parameter β. Let $\lambda = (\lambda_1, \ldots, \lambda_N)$ be the vector of scale parameters. Then the posterior density for $\mathbf{Z}$, λ, β, and ν is given by

$$\pi(\mathbf{Z}, \lambda, \beta, \nu \mid \mathbf{y}) = C\pi(\nu)\prod_{i=1}^{N}(\{1(Z_i > 0)1(Y_i = 1) + 1(Z_i \le 0)1(Y_i = 0)\}\sqrt{\lambda_i/2\pi} \times \exp(-\lambda_i/2(Z_i - \mathbf{x}_i^T\beta)^2)c(\nu)\lambda_i^{\nu/2-1}e^{-\nu\lambda_i/2}), \quad (7)$$

where $c(\nu) = [\Gamma(\nu/2)(\nu/2)^{(\nu/2)}]^{-1}$ and $\pi(\nu)$ is the prior on ν. In this case the unknown vector $\theta = (\mathbf{Z}, \lambda, \beta, \nu)$. The fully conditional distributions of β, $\mathbf{Z}$, λ and ν are given below:

- $\beta \mid \mathbf{y}, \mathbf{Z}, \lambda, \nu$ is distributed $N_k(\hat{\beta}_{Z,\lambda}, (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1})$, where
$$\hat{\beta}_{Z,\lambda} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Z} \quad \text{and} \quad \mathbf{W} = \text{diag}(\lambda_i). \quad (8)$$
- The fully conditionally distributions of $Z_1, \ldots, Z_N$ are independent with
$$\mathbf{Z}_i \mid \mathbf{y}, \beta, \lambda, \nu \text{ distributed } N(\mathbf{x}_i^T\beta, \lambda_i^{-1}) \text{ truncated at the left by 0 if } y_i = 1$$
$$\mathbf{Z}_i \mid \mathbf{y}, \beta, \lambda, \nu \text{ distributed } N(\mathbf{x}_i^T\beta, \lambda_i^{-1}) \text{ truncated at the right by 0 if } y_i = 0 \quad (9)$$
- $\lambda_1, \ldots, \lambda_N \mid \mathbf{y}, \mathbf{Z}, \beta, \nu$ are independent with
$$\lambda_i \text{ distributed Gamma}\left(\frac{\nu+1}{2}, \frac{2}{\nu + (Z_i - \mathbf{x}_i^T\beta)^2}\right). \quad (10)$$
- $\nu \mid \mathbf{y}, \mathbf{Z}, \beta, \lambda$ is distributed according to the pdf proportional to
$$\pi(\nu)\prod_{i=1}^{N}(c(\nu)\lambda_i^{\nu/2-1}e^{-\nu\lambda_i/2}). \quad (11)$$

To implement the Gibbs sampler, we start with β equal to the least squares estimate under the probit model, set $\lambda_i = 1$ for all i, and cycle through the conditional distributions (9), (8), (10), and (11), in that order. The only difficult distribution to simulate from is the fully conditional distribution of the degrees of freedom ν. But because we are interested in the posterior probabilities for ν in a finite set, it is then easy to simulate from the discrete distribution (11).

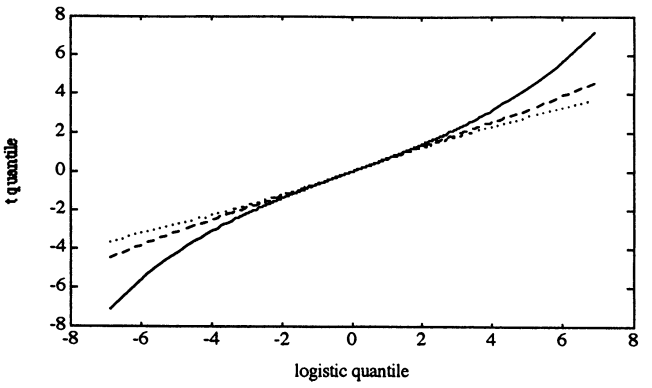

Figure 1. Plot of Logistic Quantiles Against t Quantiles for Probabilities Between .001 and .999. The solid line is $\nu = 4$, the dashed line is $\nu = 8$, and the dotted line is $\nu = 16$.

Suppose that one is interested in making inferences about the regression vector β and the probabilities $\{p_k\}$. The posterior for β is approximated by $\hat{\pi}(\beta) \approx (1/m) \times \sum_{i=1}^{m}\pi(\beta \mid Z^{(i)}, \lambda^{(i)})$, where $\pi(\beta \mid \mathbf{Z}, \lambda)$ is the multivariate normal posterior density (8). To obtain a posterior density estimate for the probability p_k, first note that $p_k = \Phi(\lambda_k^{1/2}\mathbf{x}_k^T\beta)$. Then, by a transformation of the conditional density of β, the density estimate of the probability is given by $\hat{\pi}(p_k) = (1/m)\sum_{i=1}^{m}\phi(\Phi^{-1}(p_k); \mu, \sigma^2)/\phi(\Phi^{-1}(p_k); 0, 1)$, where $\phi(\ ; \mu, \sigma^2)$ is the $N(\mu, \sigma^2)$ pdf, $\mu = \sqrt{\lambda_k^{(i)}}\mathbf{x}_k^T\hat{\beta}_{Z,\lambda}^{mu3(i)}$ and $\sigma^2 = \lambda_k^{(i)}\mathbf{x}_k^T(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{x}_k$.

3.3 Hierarchical Analysis

The normal regression structure on $\mathbf{Z}$ also motivates the consideration of normal hierarchical models as presented in Lindley and Smith (1972). Given a particular probit model and regression parameter β of dimension k, one may suspect that β lies on a linear subspace $\mathbf{A}\beta^0$, where β^0 is p-dimensional, where $p < k$. (Alternately, one may believe that the regression parameter satisfies the $(k-p)$-dimensional subspace restriction $\Omega\beta = 0$. This prior belief may be reexpressed in the above form.) This prior belief can be modeled by the hierarchical model

(1) $\mathbf{Z}$ is distributed $N(\mathbf{X}\beta, \mathbf{I})$,
(2) β is distributed $N(\mathbf{A}\beta^0, \sigma^2\mathbf{I})$, and
(3) (β^0, σ^2) is distributed according to the prior density $\pi(\beta^0, \sigma^2)$. (12)

In usual practice, the hyperparameters β^0 and σ^2 are assumed independent with β^0 assigned a uniform prior and σ^2 given a noninformative prior (in Section 5.2, we assume that $\log \sigma^2$ is uniform distributed). The focus of posterior inferences is on the prior variance σ^2 and the regression vector β. Note that σ^2 reflects the precision of the prior belief that β lies on the linear subspace. After data is observed, the posterior distribution of σ^2 is informative about the goodness

of fit of the reduced model. The posterior density of the regression vector β compromises between least squares estimates from the "full" k-dimensional model β and the "reduced" p-dimensional model where $\beta = \mathbf{A}\beta^0$.

The fully conditional distribution of the latent data $\mathbf{Z}$ is given by (6). One can use standard theory for the normal hierarchical model (Lindley and Smith 1972) to obtain the posterior distributions of β and σ^2 conditional on the latent data $\mathbf{Z}$. (Note that these distributions are marginal posterior distributions with the hyperparameter β^0 integrated out.) Specifically, we have that

- $\beta|\mathbf{Z}, \sigma^2$ is distributed $N_k(\mu, \mathbf{V})$, where

$$\mu = \mathbf{W}_1\hat{\theta}_1 + (\mathbf{I} - \mathbf{W}_1)\mathbf{A}\hat{\theta}_2$$

$$\hat{\theta}_1 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}$$

$$\hat{\theta}_2 = [\mathbf{A}^T\mathbf{X}^T(\mathbf{I} + \mathbf{X}\mathbf{X}^T\sigma^2)^{-1}\mathbf{X}\mathbf{A}]^{-1}\mathbf{A}^T\mathbf{X}^T \times (\mathbf{I} + \mathbf{X}\mathbf{X}^T\sigma^2)^{-1}\mathbf{Z}$$

$$\mathbf{W}_1 = [\mathbf{X}^T\mathbf{X} + \mathbf{I}/\sigma^2]^{-1}\mathbf{X}^T\mathbf{X}$$

$$\mathbf{V} = ((\mathbf{I} - \mathbf{W}_1)\mathbf{A})[\mathbf{A}^T\mathbf{X}^T(\mathbf{I} + \mathbf{X}\mathbf{X}^T\sigma^2)^{-1}\mathbf{X}\mathbf{A}]^{-1} \times ((\mathbf{I} - \mathbf{W}_1)\mathbf{A})^T + [\mathbf{X}^T\mathbf{X} + \mathbf{I}/\sigma^2]^{-1} \quad (13)$$

- $\sigma^2|\mathbf{Z}$ is distributed according to the density proportional to

$$c(\mathbf{Z})\frac{|(\mathbf{I} + \mathbf{X}\mathbf{X}^T\sigma^2)^{-1}|^{1/2}}{|\mathbf{A}^T\mathbf{X}^T(I + \mathbf{X}\mathbf{X}^T\sigma^2)^{-1}\mathbf{X}\mathbf{A}|^{1/2}} \times \exp\left\{-\frac{1}{2}Q(\mathbf{Z}, \mathbf{X}\mathbf{A}\hat{\theta}_2, (\mathbf{I} + \mathbf{X}\mathbf{X}^T\sigma^2)\right\}\pi(\sigma^2), \quad (14)$$

where $Q(\mathbf{Z}, \mu, \Sigma) = (\mathbf{Z} - \mu)^T\Sigma^{-1}(\mathbf{Z} - \mu)$ and $c(\mathbf{Z})$ is a proportionality constant.

To implement the Gibbs sampler, one starts with initial guesses at β and σ^2, simulates the Z_i from (6), and then simulates β and σ^2 from the distributions (13) and (14). The fully conditional posterior density of σ^2 is not of a convenient form for simulation. However, as in the t-link example, if we place all of the prior probability on a grid of σ^2 values, it is easy to simulate from the discrete posterior density (14).

4. GENERALIZATIONS TO A MULTINOMIAL RESPONSE

4.1 Ordered Categories

Suppose that $Y_1, \ldots, Y_N$ are observed, where Y_i takes one of J ordered categories, $1, \ldots, J$. Letting $p_{ij} = P[Y_i = j]$, we define the cumulative probabilities $\eta_{ij} = \sum_{k=1}^{j} p_{ij}$, $j = 1, \ldots, J-1$. Then one popular regression model for the $\{p_{ij}\}$ (Aitchison and Silvey 1957; Gurland et al. 1960; McCullagh 1980; McKelvey and Zavoina 1975) is given by $\eta_{ij} = \Phi(\gamma_j - \mathbf{x}_i^T\beta)$, $i = 1, \ldots, N, j = 1, \ldots, J-1$. One can motivate this model by assuming that there exists a latent continuous random variable Z_i distributed $N(\mathbf{x}_i^T\beta, 1)$, and we observe Y_i, where $Y_i = j$ if $\gamma_{j-1} < Z_i \leq \gamma_j$ (we define $\gamma_0 = -\infty$ and $\gamma_J = \infty$). This problem is a normal regression problem where the response is in the form of grouped data.

In the preceding model, the regression vector β and the bin boundaries $\gamma_1, \ldots, \gamma_{J-1}$ are unknown. To ensure that the parameters are identifiable, it is necessary to impose one restriction on the bin boundaries; without loss of generality, we take $\gamma_1 = 0$. The joint posterior density of β and $\gamma = (\gamma_2, \ldots, \gamma_{J-1})$ is then given by

$$\pi(\beta, \gamma|\mathbf{y}) = C\pi(\beta, \gamma)\prod_{i=1}^{N}\sum_{j=1}^{J} 1(y_i = j) \times [\Phi(\gamma_j - \mathbf{x}_i^T\beta) - \Phi(\gamma_{j-1} - \mathbf{x}_i^T\beta)], \quad (15)$$

where $\pi(\beta, \gamma)$ is the prior. As in the two category case discussed earlier, it is straightforward to find the posterior mode of (β, γ) using Newton–Raphson and to obtain approximate posterior standard deviations of (β, γ) using the second derivative matrix of the log posterior evaluated at the mode.

The Gibbs algorithm for the binary case described in Section 3.1 can be generalized to this situation. We introduce the unobserved latent variables $Z_1, \ldots, Z_N$ defined previously and simulate values from the joint posterior distribution of $(\beta, \gamma, \mathbf{Z})$. If we assign a diffuse prior for (β, γ), then this joint posterior density is given by

$$\pi(\beta, \gamma, \mathbf{Z}|\mathbf{y}) = C\prod_{i=1}^{N}\left[\sqrt{1/2\pi}\exp(-(Z_i - \mathbf{x}_i^T\beta)^2/2) \times \left\{\sum_{j=1}^{J} 1(Y_i = j)1(\gamma_{j-1} < Z_i < \gamma_j)\right\}\right]. \quad (16)$$

The posterior distribution of β conditional on $\mathbf{y}$ and $\mathbf{Z}$ is given by the multivariate normal form (5). The fully conditional posterior distributions of $Z_1, \ldots, Z_N$ are independent with

$$Z_i|\beta, \gamma, y_i = j \text{ distributed } N(\mathbf{x}_i^T\beta, 1) \text{ truncated at the left (right) by } \gamma_{j-1}(\gamma_j). \quad (17)$$

Finally, the fully conditional density of γ_j given $\mathbf{Z}$, $\mathbf{y}$, β and $\{\gamma_k, k \neq j\}$ is given (up to a proportionality constant) by

$$\prod_{i=1}^{N}[1(Y_i = j)1(\gamma_{j-1} < Z_i < \gamma_j) + 1(Y_i = j+1)1(\gamma_j < Z_i < \gamma_{j+1})]. \quad (18)$$

This conditional distribution can be seen to be uniform on the interval $[\max\{\max\{Z_i: Y_i = j\}, \gamma_{j-1}\}, \min\{\min\{Z_i: Y_i = j+1\}, \gamma_{j+1}\}]$. To implement the Gibbs sampler here, start with (β, γ) set equal to the MLE and simulate from the distributions (18), (17), and (5), in that order.

4.2 Unordered Categories With a Latent Multinormal Distribution

The Gibbs sampling approach can also be applied to the multinomial probit model introduced by Aitchison and Bennett (1970); (also see Hausman and Wise 1978, Daganzo 1979, Amemiya 1985). For illustrative purposes, we focus on one particular version of the model. First, we introduce independent unobservable latent variables $Z_1, \ldots, Z_N$, where $Z_i = (Z_{i1}, \ldots, Z_{iJ})$ $(J > 2)$, and define $Z_{ij} = \mathbf{x}_{ij}^T\beta + \varepsilon_{ij}$, $i = 1, \ldots, N$, $j = 1, 2, \ldots, J$, where $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iJ})^T$ is distributed $N_J(\mathbf{0}, \Sigma)$ and Σ is a $J \times J$ matrix that is parameterized (for identifiability reasons) in

terms of a parameter vector θ of dimension not exceeding $J(J-1)/2$. It is helpful to think of i as the index of experimental units and j as the index of categories. On unit i we observe one of J possible outcomes with respective probabilities $p_{i1}, \ldots, p_{iJ}$. Category j is observed if $Z_{ij} > Z_{ik}$ for all $k \neq j$. (McFadden [1974] has shown that the multinomial logit model can be derived in this setup if and only if the errors $\{\varepsilon_{ij}\}$ are a random sample from a Type I extreme value distribution.) The multinomial probabilities are given by $p_{ij} = P[\mathbf{x}_{ij}^T\beta + \varepsilon_{ij} > \mathbf{x}_{ik}^T\beta + \varepsilon_{ik}$, for all $k \neq j]$. Note that computation of these probabilities entails calculation of multiple integrals of the multivariate normal density; thus maximum likelihood estimation is very difficult to perform for large J.

The computation of the multinomial probabilities can be avoided by the following Gibbs sampling approach. As in Section 4.1, denote the vector of observed categories as $\mathbf{Y} = (y_1, \ldots, y_N)$, where $y_i \in \{1, \ldots, J\}$. Letting $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iJ})^T$, the preceding model can be rewritten as

$$\begin{bmatrix} \mathbf{Z}_1 \\ \vdots \\ \mathbf{Z}_N \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_N \end{bmatrix} \beta + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{bmatrix} \tag{19}$$

or as $\mathbf{Z} = \mathbf{X}\beta + \varepsilon$, where $\varepsilon = (\varepsilon_1^T, \ldots, \varepsilon_N^T)^T$ is distributed $N_{NJ}(\mathbf{0}, \Omega = \mathbf{I}_N \otimes \Sigma)$. To implement the Gibbs sampler, we require samples from the following conditional distributions:

$$\begin{aligned} &\beta | \mathbf{Y}, \mathbf{Z}_1, \ldots, \mathbf{Z}_N, \theta \\ &\mathbf{Z}_1, \ldots, \mathbf{Z}_N | \mathbf{Y}, \beta, \theta \\ &\theta | \mathbf{Y}, \mathbf{Z}_1, \ldots, \mathbf{Z}_N, \beta. \end{aligned} \tag{20}$$

From the representation (19), if a diffuse prior is placed on β, then standard multivariate normal theory yields that $\beta | \mathbf{Z}_1, \ldots, \mathbf{Z}_N, \mathbf{Y}, \theta$ is distributed $N_k(\hat{\beta}_\mathbf{Z}, (\mathbf{X}^T\Omega^{-1}\mathbf{X})^{-1})$, where $\hat{\beta}_\mathbf{Z} = (\mathbf{X}^T\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}^T\Omega^{-1}\mathbf{Z}$. Note that computation of the parameters of the latter distribution is easy, because Ω^{-1} is a block diagonal matrix with Σ^{-1} as the typical block. Next, given $\mathbf{Y}, \beta, \theta, \{\mathbf{Z}_i\}$ is an independent collection with $\mathbf{Z}_i | \mathbf{Y}, \beta, \theta$ distributed $N(\mathbf{x}_i\beta, \Sigma)$, $i = 1, \ldots, N$, such that the y_ith component of $\mathbf{Z}_i$ is the maximum. This may be simulated by drawing a sample from $N(\mathbf{x}_i\beta, \Sigma)$ and accepting the draw if the condition is satisfied. An alternative method of performing this draw is outlined in McCulloch and Rossi (1991) (see also Geweke, 1991). Finally, consider the sampling of $\theta | \mathbf{Z}_1, \ldots, \mathbf{Z}_N, \mathbf{Y}, \beta$. Using a prior $\pi(\theta)$ on θ, the density of this distribution is proportional to

$$\pi(\theta)|\Omega(\theta)|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{Z} - \mathbf{X}\beta)^T\Omega^{-1}(\theta)(\mathbf{Z} - \mathbf{X}\beta)\right\}. \tag{21}$$

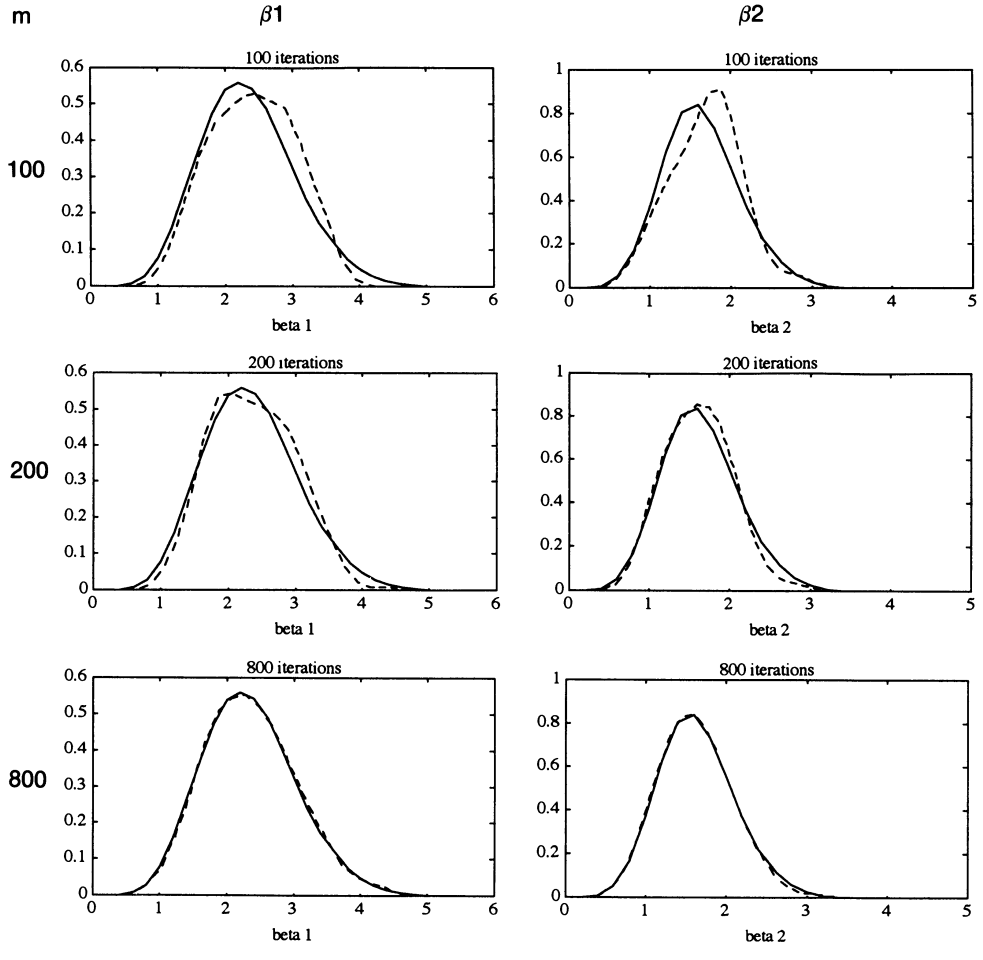

Figure 2. Estimated Posterior Densities of Regression Parameters for Finney Data for Different Number of Iterations m. The solid line is the "exact" density computed using adaptive quadrature. The dashed line represents approximation using Gibbs sampling.

This distribution is not a member of any familiar parametric family and is relatively difficult to simulate. As in Section 4.1, however, one can simplify this sampling by drawing from a normal distribution with matching mode and curvature.

5. EXAMPLES

5.1 Finney Data

We illustrate the sampling method on data first analyzed by Finney (1947); see also Myers (1990, pp. 330–332). One probit model of interest is

$$\Phi^{-1}(p_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \; i = 1, \ldots, 39, \quad (22)$$

where x_{1i} is the volume of air inspired, x_{2i} is the rate of air inspired, and the binary outcome observed is the occurrence or nonoccurrence on a transient vasorestriction on the skin of the digits. In the following posterior analysis, a uniform prior is placed on the regression parameter β.

In Figure 2 (p. 674) density estimates for β_1 and β_2 are plotted for simulated samples of size 100, 200, and 800. In each plot the solid line represents the "exact" posterior density computed using the adaptive quadrature scheme of Naylor and Smith (1982), and the dashed line is the Gibbs sampling approximation. Note that the accuracy of the Gibbs estimates increases for larger sample sizes, and the two density curves are indistinguishable for a simulation sample of 800.

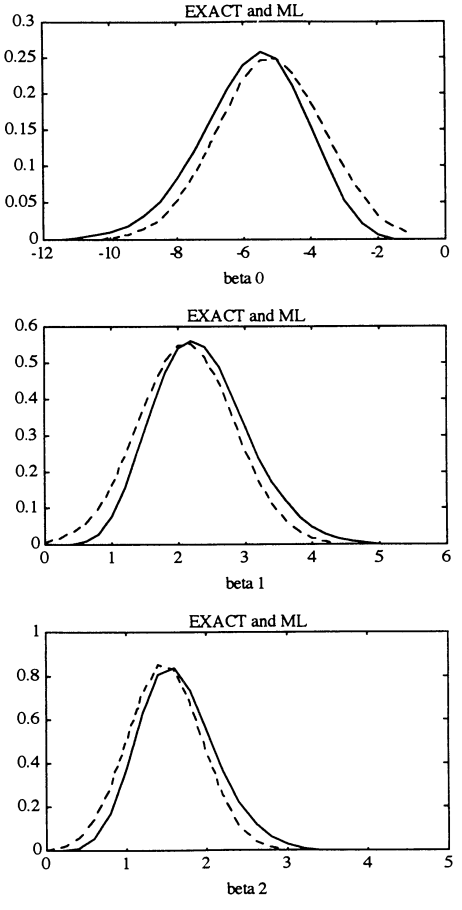

Figure 3. Exact (Solid Line) and Maximum Likelihood Approximate (Dashed Line) Posterior Densities For the Regression Coefficients for Finney's Data.

Table 1. Maximum Likelihood Parameter Estimates and Associated Standard Errors for Election Data Using Probit Link

Variable	Coefficient	Standard error
Constant	6.153E+01	3.008E+01
Income	−8.926E−06	2.373E−04
School	−5.085E+00	2.583E+00
Urban	3.181E−02	1.539E−02
SE	**	**
MW	−9.026E−02	5.831E−01
WE	−5.928E−01	7.360E−01

NOTE: Nonexistence of the estimate is indicated by two asterisks.

Figure 3 compares these final Gibbs density estimates against the approximate densities based on the ML normal approximation. In this example the exact densities exhibit some skewness, although the locations of the densities are similar.

5.2 Election Data

In this section we illustrate extensions of the probit model for election data discussed in Green (1990, p. 671). The problem is to predict the Carter/Ford vote in the 1976 Presidential election using six socioeconomic and regional variables. We begin by presenting in Table 1 the ML parameter estimates and associated standard errors using a probit link. The first observation based on this table is that the MLE for the SE regional variables does not converge. Closer examination of the data reveals that the SE variable equals 1 only for observations 1, 10, 11, 19, 25, 34, and 41 and that for these particular observations, the fitted probabilities are all equal to 1. Second, by comparing the size of the estimates with their standard errors, it appears that only the school and urban variables are important in predicting the Carter/Ford vote.

Is the ML fit sensitive to the choice of link function? To help answer this question, Figure 4 compares ML fitted probabilities using probit and $t(4)$-link functions. For ease of comparison, both sets of probabilities are transformed to the logit ($\log(p/(1-p))$) scale. On the graph, the probit-fitted probability is plotted against logit($t(4)$ prob.fit) − logit(probit prob.fit). This figure demonstrates that the fit-

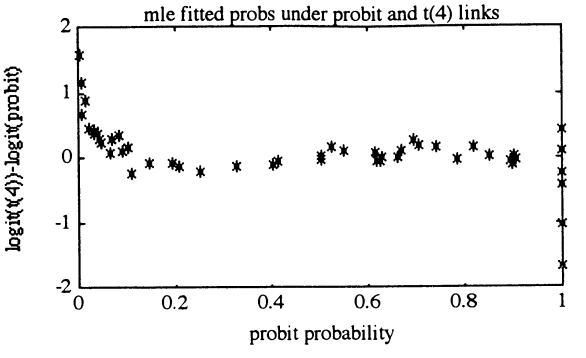

Figure 4. A Comparison of MLE Fitted Probabilities Using Probit and t(4) Link Functions.

ted probabilities for the two models can be significantly different. The most noticeable differences here correspond to small values of p where the t-fitted probabilities are significantly larger. This brief analysis suggests that the choice of link can make a difference and that it is worthwhile to consider a general form for the link function.

Before we apply some Bayesian models to this data, we must attend to the problem of nonconvergence of the MLE of β_5 in this example. When one explores the likelihood function, one discovers that when the other parameters are held fixed, the likelihood of β_5 approaches a constant value as the parameter approaches infinity. Thus to obtain a proper posterior distribution for β, one must assign a proper prior distribution to β_5.

Because the logistic model is a popular model for this data, we first investigate the connection between the logistic model and the generalized probit models described in this article. In Section 3.2 it was stated that the logistic link function appears approximately equivalent to a t-link function with 8 df. To confirm this observation, we compare the posterior analyses for the following two models:

- *Logistic model:* $p_i = \exp(\mathbf{x}_i^T\beta)/(1 + \exp(\mathbf{x}_i^T\beta))$, $(\beta_5, \{\beta_j, j \neq 5\})$ independent with β_5 distributed $N(0, 1)$ and $\{\beta_j, j \neq 5\}$ distributed uniform on R^6
- *T(8) model:* $p_i = F_{T(8)}(\mathbf{x}_i^T\beta)$, $(\beta_5, \{\beta_j, j \neq 5\})$ independent with β_5 distributed $N(0, .4)$ and $\{\beta_j, j \neq 5\}$ distributed uniform on R^6.

Note that for each model, a proper prior distribution is assigned to β_5 to ensure that the posterior distribution will be proper.

Second, because a $t(8)$ random variable is approximately .634 times a logistic random variable, the logistic regression parameter β_5 distributed $N(0, 1)$ is approximately equivalent to a $t(8)$ regression parameter β_5 distributed $N(0, .4)$. Thus the significant prior information in the two models is approximately matched.

The Gibbs sampler procedure described in Section 3.2 was used to obtain a simulated sample β of size 2,000 for the $t(8)$ link model. Dellaportas and Smith (in press) described the use of Gibbs sampling and an adaptive rejection algorithm to simulate the regression parameter for a logistic model. Their procedure was used to simulate a sample of size 2,000 for the logistic model. Normal kernel density estimates of the simulated draws of the seven regression parameters are presented in Figure 5. (The $t(8)$ parameters were modified by a factor of .634 to make them comparable to the logistic parameters.) Note that the two sets of marginal posterior density estimates are very similar, supporting the claim made earlier that the logistic and $t(8)$ models are approximately equivalent.

The preceding analysis considered the use of a t link function with known df. Next, suppose that one wanted to use a t link where the df was unknown from the set {4, 8, 16, 32}. One may be uncertain about the most likely value of degrees of freedom, and so a prior is used that assigns equal

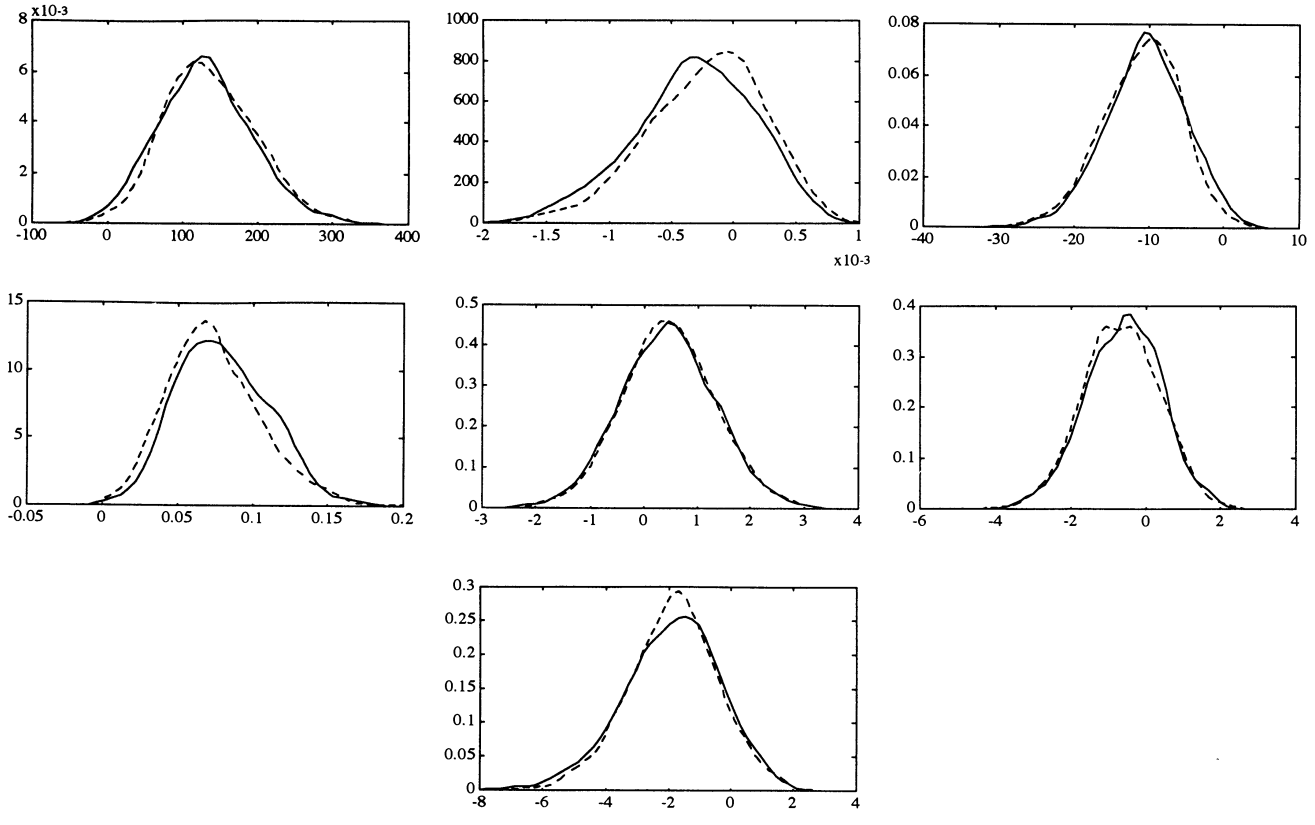

Figure 5. Posterior Densities of Regression Coefficients for t8 (Solid Lines) and Logistic (Dashed Lines) Models. The top row corresponds to β_1, β_2, β_3; the middle row to β_4, β_5, β_6; and the bottom row to β_7.

Table 2. Posterior Moments of Regression Coefficients for t-Link Analysis With Unknown Degrees of Freedom

Variable	Posterior mean	Posterior standard deviation	Lag one correlation	Numerical standard error
Constant	89.3	43.8	.863	1.7
Income	−1.95E−4	3.62E−4	.914	.20E−4*
School	−7.18	3.76	.867	.14
Urban	.0511	.0245	.919	.0017*
SE	.256	0.587	.502	.010
MW	−.467	0.728	.715	.019
WE	−1.38	1.17	.861	.06*

NOTE: Uncertain values of numerical standard errors due to strong lag correlation are indicated by an asterisk.

probabilities to ν in the finite set. As in the previous example, we assign β_5 a $N(0, 1)$ prior. The Gibbs sampler was run with $m = 10{,}000$ cycles. The main conclusions of this analysis are as follows:

1. In the Gibbs run the posterior probabilities of the degrees of freedom values 4, 8, 16, and 32 were given by .52, .21, .14, and .11. Thus the best fit appears to be a t-link with the smallest value of ν.

2. A table of posterior standard moments for the seven regression coefficients is given in Table 2. Because these moments are based on a single run of the Gibbs sampler, two columns of this table give summary statistics for this run that are helpful in diagnosing convergence. The "lag one correlation" column gives lag one autocorrelations of the sample, and the "numerical standard errors" column give numerical standard errors for the posterior means based on the batch means method described in Section 2.3. Batches of increasing size were collected until the lag one correlation of the batch means was under .05. The numerical standard error is then the standard deviation of the batch means divided by the square root of the number of batches. An asterisk indicates that the lag one correlation of the batch means was not under .05; for these three coefficients the lag correlation of the batch means was in the .1–.2 range and the corresponding numerical standard errors are probably slightly understated.

3. The posterior distributions of the probabilities p_i using the t-link (df unknown) were noticeably different from the posterior distributions for p_i using the probit link. This is demonstrated in Figure 6, which plots the difference in the logits of the posterior means of p_i (against the observation number) for the probit and t-links. Note that the graph's snake-type appearance is qualitatively similar to the MLE comparison in Figure 4.

4. Figure 7 plots the normalized posterior distribution for the residuals $y_i - p_i = y_i - H(\mathbf{x}_i^T\boldsymbol{\beta})$ (corresponding to the first 12 observations) using the t-link model. In the usual classical analysis, it is difficult to interpret the size of the residual $y_i - \hat{p}_i$, because it takes on only the two values—$\hat{p}_i$ and $1 - \hat{p}_i$. In contrast the Bayesian residual $y_i - p_i$ is real valued on the interval $[y_i - 1, y_i]$. As in Chaloner and Brant (1988), one can determine whether a particular observation is an outlier by the posterior computation $P[|y_i - p_i| > k]$ for a particular constant k. In this example one can informally check for outliers by looking for the distributions concentrated away from 0. For the 12 observations presented here, one observes that the distributions of residuals of observations 7, 8, and 12 are concentrated away from 0.

In the preceding ML fit with seven covariates, one concern is the effect of the SE variable on the analysis. One may suspect that this term (and the other regional effects) may be removed from the model with little change in the overall

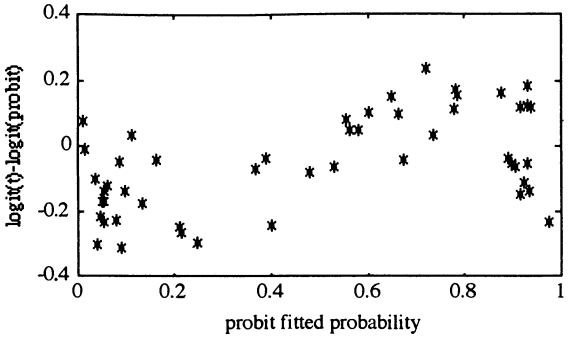

Figure 6. Posterior Means of the Fitted Probabilities p_i Using Probit and t-Links.

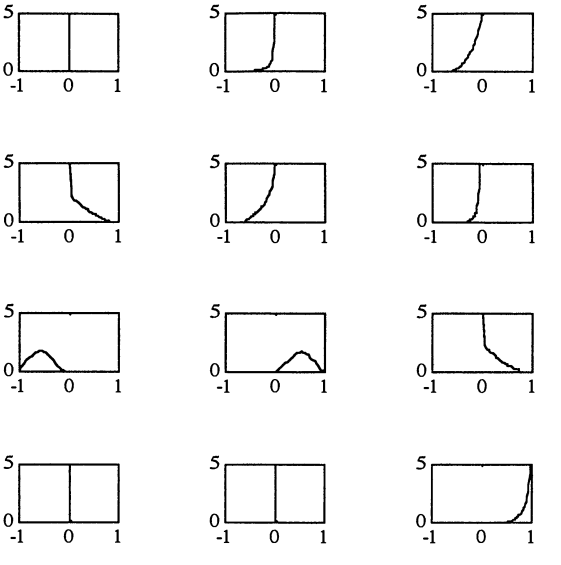

Figure 7. Posterior Densities Using the t-Link of Residuals $y_i - p_i$ for the First 12 Observations for the Voting Data. Observations 1–3 are in the first row, 4–6 are in the second row, 7–9 are in the third row, and 10–12 are in the fourth row.

Table 3. Maximum Likelihood Estimates Under Full and Reduced Models and Hierarchical Posterior Means

Beta index	1	2	3	4	5	6	7
Full model	61.5	−9E−6	−5.08	.0318	**	−.0952	−.593
	(30.1)	(2.37E−4)	(2.58)	(.0154)	(**)	(.583)	(.736)
Reduced model	87.3	2.37E−5	−7.20	.0331	0	0	0
	(24.4)	(1.92E−4)	(2.09)	(.0142)	(0)	(0)	(0)
Hierarchical	94.9	−2.0E−5	−7.79	.0398	.0997	−.0137	−.0642
	(27.6)	(1.95E−4)	(2.24)	(.0166)	(.3553)	(.214)	(.264)
Numerical standard error	1.2	.85E−5	.10	.0007	.014	.0083	.0086

NOTE: Associated standard errors are in parentheses. Numerical standard errors of the posterior means are also given. Nonexistence of the mle is indicated by an asterisk.

fit. To reflect this belief, we can apply the hierarchical model (14) with $\mathbf{A} = \begin{bmatrix} \mathbf{I}_4 \\ 0 \end{bmatrix}$, where $\mathbf{I}_4$ is the identity matrix of dimension 4.

In this example we placed uniform prior probability for $\log \sigma^2$ on the set $\{-8, -7, \ldots, 4\}$ and ran the Gibbs sampler with $m = 1{,}000$. (Here simulated values were collected using a spacing of five cycles.) Table 3 presents posterior means and standard deviations for the components of β. (As in Table 2, numerical standard errors of the posterior means are given.) To help understand these posterior moments, this table also gives MLE's for β under the full model and the reduced model, where the last three components are equal to 0. Note that the posterior moments of the last three components are all near 0. In addition, the posterior density for σ^2 is concentrated towards 0. From these observations, it appears that the regional effects are not significant.

5.3 A Trivariate Probit Example

To illustrate the calculations for the unordered multinomial setting of Section 4.2, consider the analysis of the trivariate probit model discussed by Daganzo (1979, chap. 2). The model for the latent variable for the ith subject is given by

$$\begin{bmatrix} Z_{i1} \\ Z_{i2} \\ Z_{i3} \end{bmatrix} \text{ is distributed } N\left(\begin{bmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \end{bmatrix}\beta, \begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right), \quad (23)$$

and the observation y_i is the index of the maximum of (Z_{i1}, Z_{i2}, Z_{i3}). In the example discussed by Daganzo, Z_{ik} represents the perceived attractiveness by subject i of the kth mode of transportation, x_{ik} denotes the travel time by mode k, and the scalar β represents the attractiveness of the value of travel time. Modes 1 and 2 are public transportation modes that are believed related, with unknown correlation ρ. The subject bases the choice y_i of transportation mode on the maximum of the perceived attractiveness values $\{Z_{ik}\}$.

In this setting the fully conditional posterior of θ (21) takes a simple form. If one defines $w_{ik} = Z_{ik} - x_{ik}\beta$, $k = 1, 2$, then the fully conditional posterior density of the correlation coefficient ρ is proportional to

$$\pi(\rho)(1-\rho^2)^{-N/2} \times \exp\left\{-\frac{1}{2(1-\rho^2)}(S_{w1} - 2\rho S_{w1w2} + S_{w2})\right\},$$

where S_{w1}, S_{w2}, and S_{w1w2} are the usual sum of squares of the paired data $\{(w_{i1}, w_{i2}), i = 1, \ldots, N\}$. Using well-known approximations to this posterior (see, for example, Lee 1989, chap. 6), one can simulate from this density. To implement the Gibbs sampler, we initialize $\rho = 0$ and then cycle through the conditional distributions $\mathbf{Z}$, β and ρ, in that order.

Daganzo (1979) analyzed a hypothetical data set of 50 observations. The MLE of (β, ρ) was found to be (.238, .475) with standard errors (.144, .316). The Gibbs sampler was run for $m = 12{,}000$ cycles with a uniform prior on (β, ρ). The estimated posterior means and standard deviations of (β, ρ) were given by (.234, .291) and (.0475, .340). (By the computation of simulation standard errors, the posterior means appear to be accurate to the second significant digit.) Note that the posterior standard deviation of β is significantly smaller than the MLE standard deviation. In addition, due to the strong left skewness of the marginal posterior distribution of ρ, the posterior mean is significantly smaller than the MLE. This example further illustrates that the information provided by the exact posterior and ML can be different.

6. CONCLUDING REMARKS

The main point of this article is that by introducing latent data into the problem, the probit model on the binary response is connected with the normal linear model on the continuous latent data response. This approach has a number of advantages. First, it allows one to perform exact inference for binary regression models; this likely will be preferable to ML methods for small samples. The approach is especially attractive in the multinomial setup, where it can be difficult to evaluate the likelihood function. Second, applying this approach using Gibbs sampling requires simulation mainly from standard distributions such as the multivariate normal and, therefore, is easy to implement in many statistical computer languages. Finally, one can easily elaborate the probit model by using suitable mixtures of normal distributions to model the latent data. This approach was illustrated by consideration of t-link and hierarchical models.

One caution in the use of Gibbs sampling is that, by simulation, one is introducing extra randomness into the estimation procedure, and it is important to understand when a particular simulation process has converged. Some methods for the diagnosis of convergence have been discussed here—namely, the "settling down" of marginal posterior density estimates and the use of batching to obtain simulation standard errors for estimated posterior moments. Because the relative simplicity of this simulation method in this appli-

cation, we believe that there will be future research into the automation of this algorithm so that it can be incorporated into standard statistical software.

[*Received April 1991. Revised June 1992.*]

REFERENCES

Aitchison, J., and Bennett, J. A. (1970), "Polychotomous Quantal Response by Maximum Indicant," *Biometrika,* 57, 253–262.

Aitchison, J., and Silvey, S. D. (1957), "The Generalization of Probit Analysis to the Case of Multiple Responses," *Biometrika,* 44, 131–140.

Amemiya, T. (1985), *Advanced Econometrics,* Cambridge, MA: Harvard University Press.

Bratley, P., Fox, B., and Schrage, L. (1987), *A Guide to Simulation,* New York: Springer-Verlag.

Carlin, B. P., and Polson, N. G. (1991), "Inference for Nonconjugate Bayesian Methods Using the Gibbs Sampler," *Canadian Journal of Statistics,* 4, 399–405.

Chaloner, K., and Brant, R. (1988), "A Bayesian Approach to Outlier Detection and Residual Analysis," *Biometrika,* 75, 651–659.

Chib, S. (1992), "Bayes Inference in the Tobit Censored Regression Model," *Journal of Econometrics,* 51, 79–99.

Cox, D. R. (1971), *The Analysis of Binary Data,* London: Methuen.

Daganzo, C. (1979), *Multinomial Probit,* New York: Academic Press.

Dellaportas, P., and Smith, A. F. M. (in press), "Bayesian Inference for Generalised Linear Models Via Gibbs Sampling," *Applied Statistics.*

Devroye, L. (1986), *Non-Uniform Random Variate Generation,* New York: Springer-Verlag.

Finney, D. J. (1947), "The Estimation From Individual Records of the Relationship Between Dose and Quantal Response," *Biometrika,* 34, 320–334.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. M. (1990), "Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling," *Journal of the American Statistical Association,* 85, 972–985.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association,* 85, 398–409.

Geweke, J. (1991), "Efficient Simulation From the Multivariate Normal and Student-T Distributions Subject to Linear Constraints, in *Computing Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface,* Alexandria, VA: American Statistical Association, pp. 571–578.

Greene, W. H. (1990), *Econometric Analysis,* New York: Macmillan.

Griffiths, W. E., Hill, R. C., and Pope, P. J. (1987), "Small Sample Properties of Probit Model Estimators," *Journal of the American Statistical Association,* 82, 929–937.

Gurland, J., Lee, I., and Dahm, P. A. (1960), "Polychotomous Quantal Response in Biological Assay," *Biometrics,* 16, 382–398.

Hausman, J. A., and Wise, D. A. (1978), "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences," *Econometrica,* 46, 403–426.

Lee, P. M. (1989), *Bayesian Statistics,* New York: Oxford University Press.

Lindley, D. V., and Smith, A. F. M. (1972), "Bayes Estimates for the Linear Model," *Journal of the Royal Statistical Society,* Ser. B, 135, 370–384.

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

McCullagh, P. (1980), "Regression Models for Ordinal Data" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 109–142.

McCulloch, R. and Rossi, P. E. (1991), "A Bayesian Analysis of the Multinomial Probit Model," technical report, University of Chicago Graduate School of Business.

McFadden, D. (1974), "Conditional Logit Analysis of Qualitative Choice Behaviour," in *Frontiers of Econometrics,* ed. P. Zarembka, New York: Academic Press, pp. 105–142.

McKelvey, R., and Zavoina, W. (1975), "A Statistical Model for the Analysis of Ordinal Level Dependent Variables," *Journal of Mathematical Sociology,* 4, 103–120.

Mudholkar, G. S., and George, E. O. (1978), "A Remark on the Shape of the Logistic Distribution," *Biometrika,* 65, 667–668.

Myers, R. H. (1990), *Classical and Modern Regression With Applications,* Boston: PWS-Kent.

Naylor, J. C., and Smith, A. F. M. (1982), "Applications of a Method for the Efficient Computation of Posterior Distributions," *Applied Statistics,* 31, 214–225.

Nelder, J. A., and McCullagh, P. (1989), *Generalized Linear Models,* New York: Chapman and Hall.

Prentice, R. L. (1976), "A Generalization of the Probit and Logit Model for Dose–Response Curves," *Biometrics,* 32, 761–768.

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.

Wei, G. C. G., and Tanner, M. A. (1990), "Posterior Computations for Censored Regression Data," *Journal of the American Statistical Association,* 85, 829–839.

West, M. (1992), "Modeling With Heavy Tails," in *Proceedings of the 4th Valencia Meeting on Bayesian Statistics.*

Zeger, S. L., and Karim, M. R. (1991), "Generalized Linear Models With Random Efects: A Gibbs Sampling Approach," *Journal of the American Statistical Association,* 86, 79–86.

Zellner, A. (1975), "Bayesian Analysis of Regression Error Terms," *Journal of the American Statistical Association,* 70, 138–144.

Zellner, A., and Rossi, P. E. (1984), "Bayesian Analysis of Dichotomous Quantal Response Models," *Journal of Econometrics,* 25, 365–393.