Generalized Linear Models With Random Effects; A Gibbs Sampling Approach

SCOTT L. ZEGER and M. REZAUL KARIM*

Generalized linear models have unified the approach to regression for a wide variety of discrete, continuous, and censored response variables that can be assumed to be independent across experimental units. In applications such as longitudinal studies, genetic studies of families, and survey sampling, observations may be obtained in clusters. Responses from the same cluster cannot be assumed to be independent. With linear models, correlation has been effectively modeled by assuming there are cluster-specific random effects that derive from an underlying mixing distribution. Extensions of generalized linear models to include random effects has, thus far, been hampered by the need for numerical integration to evaluate likelihoods. In this article, we cast the generalized linear random effects model in a Bayesian framework and use a Monte Carlo method, the Gibbs sampler, to overcome the current computational limitations. The resulting algorithm is flexible to easily accommodate changes in the number of random effects and in their assumed distribution when warranted. The methodology is illustrated through a simulation study and an analysis of infectious disease data.

KEY WORDS: Bayesian; Correlation; Heterogeneity; Logistic regression; Monte Carlo; Overdispersion; Regression.

1. INTRODUCTION

Generalized linear models (McCullagh and Nelder 1989; Nelder and Wedderburn 1972) have unified regression methodology for a wide variety of discrete, continuous, and censored responses that can be assumed to be independent. In many problems, however, responses are clustered. For example, in longitudinal research, repeated observations on each subject are unlikely to be independent. In genetic epidemiology, observations on members of one family will be correlated. In sample surveys, responses from members of the same village are likely to be correlated. This dependence must be taken into account to correctly assess the relationship of the response Y with explanatory variables X.

With Gaussian responses, random effects models (e.g., Laird and Ware 1982, Lindstrom and Bates 1988) have been widely used to account for the dependence within clusters. The linear model has the general form

$$y_{ij} = x_{ij}'\beta + z_{ij}'b_i + \varepsilon_{ij}, \qquad j = 1, \ldots, n_i, i = 1, \ldots, I,$$

where y_{ij} is the response for the jth observation in cluster i; x_{ij} is a $p \times 1$ vector of covariates associated with that response; β is the vector of regression coefficients that are of scientific interest; z_{ij} is a $q \times 1$ subset of x_{ij} with random coefficients; b_i is a $q \times 1$ vector of random effects assumed to follow a Gaussian distribution with mean 0 and unknown variance D; and ε_{ij} is an independent Gaussian error with mean 0 and variance σ_ε^2. As a simple example, if $z_{ij} = 1$ ($q = 1$), the responses from cluster i are correlated by virtue of their sharing a common intercept $\beta_0 + b_i$. Here the correlation of y_{ij} and y_{ik} is $D/(D + \sigma_\varepsilon^2)$. In general, random effects models reflect heterogeneity across clusters in the regression coefficients, causing observations from the same cluster to be associated.

Several authors have investigated the extension of random effects models to the generalized linear model (GLM) family. The beta-binomial (Williams 1982) and Poisson-gamma models (Breslow 1984) were among the earliest. Here covariates cannot vary within a cluster, that is, $x_{ij} = x_i$ ($j = 1, \ldots, n_i$). Stiratelli, Laird, and Ware (1984) and Anderson and Aitkin (1985) considered general covariates in logistic regression with a Gaussian random intercept using expectation maximization (EM) and Newton–Raphson algorithms, respectively. Ochi and Prentice (1984) and Gilmour, Anderson, and Rae (1985) discussed probit-Gaussian models. Harville and Mee (1984) studied random effects models for ordered categorical data. With count data, Breslow (1984), Crowder (1985) and Tsutakawa (1988) have investigated log-linear models with random effects. Related models are used in Bayesian analysis of contingency tables, for example, by Leonard (1975).

There are serious limitations to the methods for fitting random effects GLM's when $x_{ij} \neq x_i$ because of the need for numerical integration of order q to evaluate the likelihood, except in the linear model. The computational burden has limited data analysis in several ways. First, investigators have largely restricted their attention to random intercept models ($q = 1$) to avoid higher dimensional numerical integration. Second, specialized software is required and is typically optimized for a particular random effects distribution (e.g., the Gaussian). Third, inferences about the regression coefficients have often been made conditional upon the estimated random effects variance, again to avoid difficult integrations. In the linear model, $\hat{\beta}$ and $\hat{D}$ are asymptotically orthogonal so that $\operatorname{var}(\hat{\beta} \mid \hat{D})$ converges to $\operatorname{var}(\hat{\beta} \mid D)$. In nonlinear models, $\hat{\beta}$ and $\hat{D}$ are asymptotically correlated, and hence inferences about β must take into account the uncertainty in $\hat{D}$.

An alternative approach to the analysis of clustered responses that avoids these numerical problems has been pro-

* Scott L. Zeger is Associate Professor and M. Rezaul Karim is Graduate Student, Department of Biostatistics, Johns Hopkins University, Baltimore, MD. We gratefully acknowledge support from the Burroughs-Wellcome Fund and from NIH Grant 1-R29-AI25529-02. This research was partially conducted while Scott L. Zeger was a visitor at the Medical Research Council Biostatistics Unit in Cambridge, England, whose hospitality is also acknowledged. This is Technical Report Number P691 from the Department of Biostatistics, Johns Hopkins University School of Hygiene and Public Health. The authors thank David Clayton, Duncan Thomas, and Jill Kirby for helpful discussions, Alfred Sommer and Joanne Katz for use of their data, and anonymous referees and the editor for their constructive reviews.

Journal of the American Statistical Association
March 1991, Vol. 86, No. 413, Theory and Methods

posed by Liang and Zeger (1986) and further discussed by Zeger, Liang, and Albert (1988). They modeled the marginal expectation of the response rather than the conditional expectation given a cluster-specific random effect. A regression coefficient in this marginal model is interpreted as the change in the "population-averaged" response rather than the change in any one cluster's expected response with X. This methodology is a multivariate analog of quasi-likelihood modeling (McCullagh and Nelder 1989; Wedderburn 1974). See Zeger et al. (1988) for a more detailed discussion. Prentice (1988) and Zhao and Prentice (1990) have considered various properties of these marginal models.

This article focuses on the random effects model. As Nelder (1972) has pointed out in his discussion of Lindley and Smith's (1972) article on Bayesian methods in regression, there is a strong connection between the random effects and Bayesian regression models. We exploit this relationship and adopt a Monte Carlo method, the Gibbs sampler (Gelfand and Smith 1990; Geman and Geman 1984), to overcome the difficulties mentioned previously. The resulting method is simple to implement, if computationally intensive and flexible to accommodate changes in the choice of explanatory variables, random effects variables, and random effects distribution. The Gibbs sampler has already been successfully applied in the analysis of frailty models by Clayton (1989), for problems in genetic linkage by Thomas (1989), and in linear variance components models by Gelfand, Hills, Racine-Poon, and Smith (1990).

Section 2 presents the GLM with random effects and the intractable likelihood function. In Section 3 the model is cast as a problem in Bayesian inference. Section 4 briefly reviews the Gibbs sampler and its application to the random effects GLM. In Section 5, the necessary conditional distributions for use of the Gibbs sampler are given. The algorithm for obtaining the joint posterior distribution of the regression coefficients and random effects parameters is detailed. Section 6 illustrates the methodology with a brief simulation study and an application to infectious disease data.

2. RANDOM EFFECTS GENERALIZED LINEAR MODEL

The data set we intend to model is composed of a response y_{ij} and a vector of p predictors x_{ij} for observations $j = 1, 2, \ldots, n_i$ within clusters $i = 1, \ldots, I$. In most applications, I is large relative to the n_i's.

The scientific objective is to characterize the dependence of y_{ij} on x_{ij}. However, the observations $y_{i1}, \ldots, y_{in_i}$ in cluster i are likely to be correlated. This must be taken into account to draw correct inferences.

In the random effects GLM, the within-cluster correlation arises from heterogeneity among clusters in the coefficients for a subset z_{ij} of the covariates x_{ij}. Conditional on a random variable b_i, we have a GLM (McCullagh and Nelder 1989), where y_{ij} follows an exponential family distribution of the form

$$f(y_{ij} \mid b_i) = \exp\{[y_{ij}\theta_{ij} - a(\theta_{ij}) + b(y_{ij})]/\phi\}. \quad (2.1)$$

The conditional moments $u_{ij} = E(y_{ij} \mid b_i) = a'(\theta_{ij})$ and $v_{ij} = \operatorname{var}(y_{ij} \mid b_i) = a''(\theta_{ij})\phi$ are assumed to satisfy

$$h(u_{ij}) = \eta_{ij} = x_{ij}'\beta + z_{ij}' b_i \quad (2.2)$$

and

$$v_{ij} = g(u_{ij})\phi,$$

where h and g are known link and variance functions, respectively. The specification is completed by assuming b_i ($i = 1, \ldots, I$) are independent observations following the parametric distribution $F(\theta)$ with $E(b_i) = 0$. In what follows, we will assume b_i is multivariate Gaussian with mean 0 and variance D and let $g(b_i \mid D)$ be the corresponding Gaussian density. Alternative choices for $F(\theta)$ can also be handled easily with large data sets using the Bayesian approach described in the next two sections.

The likelihood function for the parameters β and D has the form

$$L(\beta, D, y) \propto \prod_{i=1}^{I} \int \prod_{j=1}^{n_i} f(y_{ij} \mid b_i)|D|^{-1/2} \times \exp\left(-\frac{1}{2} b_i'D^{-1}b_i\right) db_i. \quad (2.3)$$

Except in the linear model with Gaussian errors, the integral above does not have an analytic solution. Hence likelihood inference requires numerical evaluation. The integral's dimension is equal to the number of random effects.

In the remainder of the article, we let $y_i = (y_{i1}, \ldots, y_{in_i})'$, $X_i = (x_{i1}', \ldots, x_{in_i}')$, $Z_i = (z_{i1}', \ldots, z_{in_i}')$, $u_i = (u_{i1}, \ldots, u_{in_i})$, $\eta_i = (\eta_{i1}, \ldots, \eta_{in_i})$, and $V_i = \operatorname{diag}(v_{i1}, \ldots, v_{in_i})$.

3. BAYESIAN FORMULATION

In a Bayesian approach to analyzing the random effects GLM, the parameters β and D are random variables and treated symmetrically with the observed y's and unobserved b's. Let $p(\beta, D)$ represent the joint prior distribution for β and D. The first objective of our analysis is to derive the posterior distribution $f(\beta, D \mid y)$ given by

$$f(\beta, D \mid y) = \frac{\prod_{i=1}^{I} \int f(y_i \mid b_i, \beta) g(b_i \mid D) p(\beta, D)\, db_i}{\int \prod_{i=1}^{I} \int f(y_i \mid b_i, \beta) g(b_i \mid D) p(\beta, D)\, db_i\, d\beta\, dD}. \quad (3.1)$$

Note that the denominator is a normalizing constant independent of β, D so that estimators, such as the posterior mode, can be derived from the numerator alone. Also, note that if $p(\beta, D)$ is a constant, the numerator is just the likelihood function. However, a Bayesian flavor remains since the likelihood of the observed data y is analogous to a posterior obtained from the joint likelihood for y and b and from the "prior" or mixing distribution for b.

In some problems, the posterior distributions $f(b_i \mid y)$ ($i = 1, \ldots, I$) may also be important. For example, Zeger, See, and Diggle (1988) estimated subgroup specific growth rates for the AIDS epidemic using random effects log-linear models in which $E(b_i \mid y)$ is the amount by which the ith risk group rate is greater (less) than the national average. $f(b_i \mid y)$ is given by

$$f(b_i \mid y) = \frac{\int f(y_i \mid b_i, \beta) g(b_i \mid D) p(\beta, D)\, d\beta\, dD}{\int f(y_i \mid b_i, \beta) g(b_i \mid D) p(\beta, D)\, db_i\, d\beta\, dD}. \quad (3.2)$$

Numerical evaluation of either $f(\beta, D \mid y)$ or $f(b_i \mid y)$ is typically intractable.

4. GIBBS SAMPLER

The Gibbs sampler is a Monte Carlo method for estimating the desired posterior distributions. Its greatest advantage is its ease of implementation, which is attained at the cost of computational efficiency. The method is an adaptation of the Metropolis algorithm (Hastings 1970; Li 1988; Metropolis, Rosenbluth, Rosenbluth, and Teller et al. 1953) and was discussed in detail by Geman and Geman (1984) in the context of spatial processes. Gelfand and Smith (1990) gave an excellent overview and develop the connections between Gibbs sampling and the closely related imputation posterior algorithm of Tanner and Wong (1987) and importance sampling introduced by Rubin (1987).

To review the method, consider three variables U, V, and W and suppose the conditional distribution of each, given the remainder has a simple form while the joint distribution is more complicated. Denote the conditional distributions by $[U \mid V, W]$, $[V \mid U, W]$, and $[W \mid U, V]$ and the joint distribution by $[U, V, W]$. Suppose that the joint distribution $[U, V, W]$ is positive over its entire domain to ensure that the joint distribution is fully determined by the three conditionals (Besag 1974). Then, the Gibbs sampler is a method for generating a random variate from $[U, V, W]$ as follows. Given arbitrary starting values, $U^{(0)}$, $V^{(0)}$, $W^{(0)}$, draw $U^{(1)}$ from $[U \mid V^{(0)}, W^{(0)}]$, then draw $V^{(1)}$ from $[V \mid U^{(1)}, W^{(0)}]$, and finally, complete the first iteration by drawing $W^{(1)}$ from $[W \mid U^{(1)}, V^{(1)}]$. After a large number, B, of iterations, we obtain $(U^{(B)}, V^{(B)}, W^{(B)})$. Geman and Geman (1984) have shown that under mild conditions, the joint distribution of $(U^{(B)}, V^{(B)}, W^{(B)})$ converges at an exponential rate to $[U, V, W]$ as $B \to \infty$.

The desired joint distribution $[U, V, W]$ can be approximated by the empirical distribution of the M values $(U^{(k)}, V^{(k)}, W^{(k)})$ $(k = B + 1, B + M)$, where B is large enough so that the Gibbs sampler has converged and M is chosen to give sufficient precision to the empirical distribution of interest. Gelfand and Smith (1990) advocated using every tth ($t \approx 50$) value in this sequence to have more nearly independent contributions, but this may not be necessary, as will be discussed in more detail.

When a lower dimensional marginal distribution is of interest, for example, $[U]$, a more precise estimate than the empirical distribution of the $U^{(k)}$'s can be obtained if $[U \mid V, W]$ has a closed form. It is more efficient to approximate $[U]$ by

$$\frac{1}{M} \sum_{k=1}^{M} [U \mid V^{(B+k)}, W^{(B+k)}]$$

as we make use of the known form of the conditional distribution. This alternative is particularly desirable for estimating tail probabilities of $[U]$. This estimate can also be used to build up higher order marginals, for example, $[U, V]$, by combining it with another closed-form conditional $[V \mid U]$.

In the random effects GLM, we seek the joint distribution $[\beta, D, b \mid y]$ and its marginals $[\beta, D \mid y]$ and $[b_i \mid y]$. By the preceding arguments they can be obtained by sampling from the conditional distributions: $[\beta \mid D^{(k)}, b^{(k)}, y]$, $[D \mid \beta^{(k)}, b^{(k)}, y]$, and $[b \mid \beta^{(k)}, D^{(k)}, y]$. The elegance of this approach is that it is relatively easy to simulate from each of these conditionals. The next section gives the details.

5. CONDITIONAL DISTRIBUTIONS

The conditional distributions from which simulated values are to be drawn simplify because the random effects GLM is an example of a hierarchical Bayes model. The conditional $[\beta \mid D, b, y]$ is independent of D, that is, $[\beta \mid D, b, y] = [\beta \mid b, y]$ as long as $p(\beta, D) = p(\beta)p(D)$. Similarly, $[D \mid \beta, b, y] = [D \mid b]$. The conditional $[b \mid \beta, D, y]$ does not simplify. The simulation method from each conditional is now specified.

5.1 $[\beta \mid b^{(k)}, y]$

Given the $b^{(k)}$'s, the random effects model reduces to a generalized linear model with offset $z_{ij} b_i^{(k)}$ for each observation. Assuming a flat prior for β, $[\beta \mid b^{(k)}, y]$ is proportional to the likelihood function $\Pi_{ij} f(y_{ij} \mid b_i^{(k)})$. In larger samples, this can be closely approximated by a Gaussian distribution with mean $\hat{\beta}^{(k)}$, the maximum likelihood estimator, and variance $V_{\hat{\beta}}^{(k)}$, the inverse of the Fisher information. That is, to sample from $[\beta \mid b^{(k)}, y]$, we find $\hat{\beta}^{(k)}$ and $V_{\hat{\beta}}^{(k)}$ by performing GLM regression of y_{ij} on x_{ij} using the simulated values $z_{ij} b_i^{(k)}$'s as offsets and then generate a random variate $\beta^{(k+1)}$ from a multivariate Gaussian distribution, $N(\hat{\beta}^{(k)}, V_{\hat{\beta}}^{(k)})$.

The preceding Gaussian approximation may not be adequate in smaller samples. A sample from the exact distribution can be obtained with little additional effort using "rejection sampling" (Ripley 1987). Denote the Gaussian density by $g(\beta)$ and the true density by $f(\beta)$. To perform rejection sampling, a constant $c \geq 1$ is chosen so that $c \cdot g(\beta) \geq f(\beta)$ over the range of β. The following steps result in a random variate $\beta^{(k+1)}$ with density $f(\beta)$:

1. Generate a random variate β^* from $g(\beta)$
2. Generate a uniform (0, 1) variate u
3. If $f(\beta^*)/(c \cdot g(\beta^*)) < u$, $\beta^{(k+1)} = \beta^*$, otherwise return to step 1

That $\beta^{(k+1)}$ has density $f(\beta)$ is shown in Ripley (1987). Note the additional computation is only to evaluate the likelihood function at one or a few β^*'s. The choice of c involves a tradeoff of accuracy and computational effort. If c is close to 1, $c \cdot g(\beta)$ may be less than $f(\beta)$ and the method will fail. If c is large, a large fraction of β^*'s will be rejected and more computation will be necessary. In practice, better approximations to $f(\beta)$ have been obtained by using $c_1 N(\hat{\beta}^{(k)}, c_2 V_{\hat{\beta}}^{(k)})$, that is, by inflating the variance by a second constant c_2. In the simulations described later, we choose c_1 so that the ordinates at the common mode of f and g are equal and let $c_2 = 2$ as discussed further in Section 5.3.

It is difficult a priori to decide whether a Gaussian approximation to $[\beta \mid b, y]$ is adequate or whether rejection sampling is needed. In practice, therefore, we always use rejection sampling. By using a Gaussian $g(\beta)$ and optimizing the values of c_1 and c_2, few rejections will occur and little computing effort is wasted when the Gaussian approximation is adequate.

5.2 $[D \mid b^{(k)}]$

We have assumed the b_i's are independent Gaussian (0, D) random variables. The standard noninformative prior for D (Box and Tiao 1973) is $P(D) \propto \mid D \mid ^{-(q+1)/2}$. Then, the posterior distribution of D^{-1} follows a Wishart distribution with parameters

$$S^{(k)} = \sum_{i=1}^{I} b_i^{(k)}\, b_i^{(k)\prime}$$

and $(I - q + 1)$ df.

A random matrix D^* can be obtained by generating W^*, a standardized Wishart variate with $I - q + 1$ df, using the algorithm of Odell and Feiveson (1966). D^* is then given by $D^* = (H^{(k)'}\, W^* H^{(k)})^{-1}$, where $S^{(k)^{-1}} = H^{(k)\prime}H^{(k)}$.

5.3 $[b \mid \beta^{(k)}, D^{(k)}, y]$

Generating $b_i^{(k+1)}$'s from $[b_i \mid \beta^{(k)}, D^{(k)}, y]$ is the most time-consuming step. In the linear model, $[b_i \mid \beta, D, y]$ is a multivariate Gaussian distribution with mean $(Z_i'\, Z_i + \sigma^2 D^{-1})^{-1} Z_i'(y_i - X_i\beta)$ and covariance matrix $(Z_i'Z_i + \sigma^2 D^{-1})^{-1}$. $E(b_i \mid \beta, D, y)$ is a compromise between the least squares estimator based only upon the ith cluster data and 0 where the extent of shrinkage toward 0 is governed by the random effects variance D. Unfortunately, the conditional distribution $[b_i \mid \beta, D, y]$ does not have a closed form for the entire GLM family and must usually be evaluated by numerical techniques; its density is given by

$$f(b_i \mid \beta, D, y) = \frac{f(y_i \mid b_i, \beta) g(b_i \mid D) p(\beta, D)}{\int f(y_i \mid b_i, \beta) g(b_i \mid D) p(\beta, D) db_i}. \qquad (5.1)$$

The numerator is easily evaluated, but the scale factor in the denominator involves the same integral with respect to b_i we avoided in the likelihood analysis. But, in Gibbs sampling, only a simulated value from $[b_i \mid \beta^{(k)}, D^{(k)}, y]$ is needed. It can be obtained using rejection sampling without evaluating the integral in the denominator. The idea is to find the mode and curvature of the numerator of (5.1), call it $p(b)$, and to match a Gaussian kernel $g(b)$ to $p(b)$. This automatically incorporates the intractable denominator into the constants c_1 and c_2. If we let y_i^* be the linear approximation to $h(y_i)$ given by

$$y_i^* = \eta_i + \left(\frac{\partial u_i}{\partial \eta_i}\right)^{-1} (y_i - u_i),$$

where $\eta_i = X_i\beta + Z_i b_i$, then the maximum value of $p(b_i)$ occurs at $\hat{b}_i = (Z_i'V_iZ_i + D^{-1})^{-1}Z_i'V_i(y_i^* - X_i\beta)$, and its curvature is $\hat{v}_i = (Z_i'V_iZ_i + D^{-1})^{-1}$. Note, y_i^* and v_i depend on b_i, and hence the actual mode and curvature must be obtained by iterating the equations for $\hat{b}_i$ and $\hat{v}_i$. Finding the mode of $p(b_i)$ is solving a ridged GLM by iterative weighted least squares. The rejection sampling now has the following steps:

1. Generate b_i^* from a Gaussian distribution with mean $\hat{b}_i$ and variance $c_2\hat{v}_i$. Call this Gaussian density $g(b_i)$. We use $c_2 = 2$ in the simulations below but this can be optimized and should depend on the dimension of b_i.
2. Calculate $c_{1i} = p(\hat{b}_i)/g(\hat{b}_i)$ so that the scaled Gaussian kernel and the numerator $p(b_i)$ have equal ordinates at their common mode $\hat{b}_i$.
3. Generate a uniform (0, 1) u and let $b_i^{(k+1)} = b_i^*$ if $p(b_i^*)/(c_{1i}\, g(b_i^*)) < u$ otherwise return to step 1.

In the GLM with Gaussian random effects, covering $p(b_i)$ with a scaled Gaussian distribution works well. If b was assumed to follow a longer-tailed distribution such as a multivariate t, another kernel, perhaps from the t family would be necessary. Geweke (1989) has discussed split-t and split-normal alternatives.

5.4 Optimizing the Algorithm

The Gibbs sampler will produce simulated realizations of β, D, and b from their joint posterior distribution as long as each conditional distribution is sampled sufficiently often. The order and relative frequency of sampling from the conditionals can be adjusted to speed convergence. With the random effects GLM, it is inefficient to sample from β, D, and b equally often because this leads to successive simulated values of D or b that are highly autocorrelated and to successive $\beta^{(k)}$'s that are more nearly independent. These autocorrelations reflect the higher correlation in the joint posterior distribution between D and b than between D and β or b and β. The precise relationship between the moments of the posterior distribution and the autocorrelation function of the sequence of simulated values $\{\beta^{(k)}, D^{(k)}, b^{(k)}\}$ $(k = 1, \ldots, B)$ will be discussed elsewhere. In short, it is more efficient to iteratively update b and D many times for each update of β.

When each random effect has a substantial variance (i.e., $D_{ii} >> 0$, $i = 1, \ldots, q$), the sequence $\{D_{ii}^{(k)}, k = 1, \ldots, B\}$ has the appearance of a realization from a stationary process, and the posterior distributions can be estimated as described in Section 4. However, if D_{ii} is small, the sequences $\{D_{ii}^{(k)}, k = 1, \ldots, B\}$ will have extreme long-term dependence and often will be "trapped" near $D_{ii} = 0$. This is because $D_{ii} = 0$ is an absorbing state for the Gibbs sampler Markov chain. In this situation it is preferable to reinitialize the chain, run t ($\approx$50) steps to obtain one set of simulated parameter values, and then repeat this process M times.

A third issue in implementing the algorithm is determining when the sampler has converged. Gelfand and Smith (1990) recommend the use of Q–Q plots in which the last M simulated values of each parameter are plotted against the preceding M values. In the random effects GLM problem, the parameters D are by far slowest to converge. We have chosen to monitor them by comparing estimates of the posterior distribution of D for increasing values of t.

Finally, we discuss estimation of the one-dimensional marginal posterior distributions. Neither $[\beta \mid D, b, y]$ nor $b \mid \beta, D, y]$ has closed forms, and hence the empirical distribution of the $\beta^{(k)}$'s and $b^{(k)}$'s must be used to estimate $[\beta \mid y]$ and $[b \mid y]$, respectively. With large data sets, it may be reasonable to approximate $[\beta \mid D, b, y]$ by a Gaussian distribution, as discussed in Section 5.1, in which case a better estimate of $[\beta \mid y]$ is given by $(1/M) \sum_{k=1}^{M} [\beta \mid D^{(B+k)}, b^{(B+k)}, y]$. The conditional distribution $[D^{-1} \mid b]$ is Wishart. The posterior $[D^{-1} \mid y]$ can therefore be estimated as the Wishart mixture $1/M \sum_{k=1}^{M} [D^{-1} \mid S^{(k)}, y]$, where $S^{(k)}$ is the sample covariance matrix of the $b^{(k)}$'s. This mixture is more precise than the empirical distribution of $D^{(k)^{-1}}$.

6. SIMULATION AND EXAMPLE

The Gibbs sampler for a GLM with quite general crossed or nested random effects structure has been implemented in FORTRAN-77 on an IBM 3841 computer. Uniform, Gaussian, and chi-squared random numbers are generated using IMSL routines GGUBFS, GGNML, and GGCHS, respectively. We now report results of a brief simulation study and an example to illustrate the methodology.

In the simulation, we consider the following logistic analog of an analysis of covariance model

$$\text{logit} \Pr(y_{it} = 1 \mid b_i)$$
$$= \beta_0 + \beta_1 t + \beta_2 x_i + \beta_3 (t \cdot x_i) + b_{0_i} + b_{1_i} t,$$

where $x_i = 0$ for half the population and 1 for the remainder and $t = -3, -2, -1, 0, 1, 2, 3$. Each data set was comprised of $I = 100$ clusters of size $n_i = 7$. The fixed effects coefficients were set at $\beta_0 = -2.5$, $\beta_1 = 1.0$, $\beta_2 = -1.0$ and $\beta_3 = -.5$ so that when $x_i = 0$ and $b_i = 0$, the probability of a positive response ranged from .0041 to .62, and when $x_i = 1$, it ranges from .047 to .50. Two random effects distributions were simulated: (1) $\text{var}(b_i) = D = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, (2) $D = \begin{pmatrix} .49 & .0 \\ .0 & .25 \end{pmatrix}$. For case (1), 95% of subjects with $x_i = 0$ at time 0 will have a probability of positive response between .01 and .38 with the same slope. In case (2), these probabilities will be between .02 and .25 with slopes between .0 and 2.0 on a logit scale. For each data set, the Gibbs sampler was run for 2,000 iterations. One hundred data sets were generated and analyzed for each of the two random effects models. For each data set, we calculated: the mean $\hat{\theta}$, variance $V_{\hat{\theta}}$ and the 5th and 95th percentiles of the one-dimensional marginal posterior distributions. Table 1 lists the following: the true parameter θ; the sample mean $\bar{\hat{\theta}}$ and standard deviation $s_{\hat{\theta}}$ of the parameter estimate $\hat{\theta}$ over the 100 trials; the square root of the average variance $\bar{V}_{\hat{\theta}}^{1/2}$; and the actual coverage of the nominal 90% Bayes interval from the 5th to 95th percentile of the posterior, that is, the fraction of such intervals that contained the true parameter value.

The results in Table 1 indicate that the Gibbs sampler gives reasonable inferences in this finite sample case. The intercept estimate $\hat{\beta}_0$ has slight negative bias, while the other fixed effect coefficients are approximately unbiased. The posterior means for the random effects variances are also positively biased by 20%–30%. This bias is alleviated by using posterior modes or geometric means rather than means. The positive bias in the mean results from the long right tail of the posterior. The variance estimates $V_{\hat{\theta}}$ are also approximately unbiased leading to consistent inferences. The actual coverage probabilities of the nominal 90% Bayes posterior intervals range from 80% to 100%. With 100 trials, a 90% interval for the coverage rate is plus or minus 6%.

Table 1. Results of Simulation Study. For Each Case, 100 Data Sets With 100 Clusters of Size Seven Were Generated

	β_0	β_1	β_2	β_3	D_{11}	D_{12}	D_{22}
Case 1							
θ	−2.5	1.0	−1.0	0.5	1.0	—	—
$\bar{\hat{\theta}}$	−2.67	1.07	−0.96	0.49	1.21		
$s_{\hat{\theta}}$	0.40	0.16	0.61	0.25	0.63		
$(\bar{V}_{\hat{\theta}})^{1/2}$	0.36	0.15	0.56	0.24	0.60		
Coverage of nominal 90% interval	84	88	86	87	86		
Case 2							
θ	−2.5	1.0	−1.0	0.5	0.49	0.0	0.25
$\bar{\hat{\theta}}$	−2.74	1.10	−1.14	0.57	0.83	−0.04	0.37
$s_{\hat{\theta}}$	0.38	0.17	0.54	0.31	0.50	0.12	0.14
$(\bar{V}_{\hat{\theta}})^{1/2}$	0.37	0.19	0.57	0.28	0.59	0.21	0.22
Coverage of nominal 90% interval	80	87	89	87	88	100	95

The Gibbs sampler has been used to fit a logistic-normal random effects model to infectious disease data on 250 Indonesian children, a subset of the cohort studied by Sommer, Katz, and Tarwotjo (1983). The preschool children were examined up to six consecutive quarters for the presence of respiratory infection. There were 1,200 observations in total. The covariates of interest include: age in months (centered at 36); presence/absence of xerophthalmia, an ocular manifestation of chronic vitamin A deficiency; cosine and sine terms for the annual cycle; gender; height for age, as a percent of the National Center for Health Statistics (NCHS) standard (centered at 90%), which indicates longer-term nutritional status; and presence of stunting, defined as being below 85% in height for age. The intercept is assumed to be a random effect with a Gaussian distribution.

Table 2 lists the estimated posterior mean, mode, standard deviation, and 90% confidence interval for each parameter. These statistics were estimated from 1,000 simulated values. Figure 1 displays the estimated marginal posterior densities for each of the nine parameters.

The results indicate that respiratory infection is strongly related to age, gender, season, and height for age. The risk of infection decreases with age approximately 4% per month for children from one to five years (90% CI, −2.3 to 4.7%). Infection is less prevalent in better nourished children. The relative odds of disease for a child at 95% of the NCHS height-for-age standard relative to a child at 100% is about 1.3 (90% CI, 1.06 to 1.64). There is no evidence that the height-for-age relationship deviates from the logistic-linear model for the most undernourished children, as the stunting coefficient has 90% posterior interval from −.63 to .91. The xerophthalmia coefficient is .57, indicating a 70% in-

Table 2. Results for a Random Effects Logistic Model of Respiratory Infection Regressed on Indicated Explanatory Variables With a Gaussian Random Intercept and for a Marginal Logistic Model. The Data Set Includes 1,200 Observations on 250 Children

	Random effects model posterior characteristics				*Marginal model*	
Explanatory variable	*Mean*	*Mode*	*St. dev*	*90% CI*	*Estimate*	*St. dev.*
Intercept	−2.74	−2.72	.23	−3.10, −2.34	−2.41	.18
Age (months)	− .035	− .035	.0073	−.047, −.023	− .032	.006
Xerophthalmia (0-no; 1-yes)	.64	.64	.51	−.31, 1.34	.59	.44
Sex (0-male, 1-female)	− .61	− .57	.18	−.91, −.33	− .57	.17
Seasonal cosine	− .17	− .16	.17	−.45, .131	− .16	.15
Seasonal sine	− .45	− .44	.26	−.89, −.03	− .42	.24
Height for age (% of NCHS standard)	− .051	− .049	.028	−.099, −.006	− .049	.03
Stunted (<85% height for age; 0-no, 1-yes)	.18	.24	.47	−.63, .91	.15	.41
D	.80	.78	.40	.25, 1.54	—	—

crease in the odds of respiratory infection among vitamin A deficient children. However, there are only 55 cases of xerophthalmia in the 1,200 visits, so this coefficient is not statistically significant. There is strong evidence of heterogeneity among children in the propensity for respiratory infection. The random effects variance is estimated to be .80 by the posterior mean, with 90% interval .25 to 1.54. Note that in Figure 1 the posterior distributions for the fixed effects are reasonably symmetric while that for D is skewed to the right.

Figure 2 displays the bivariate distribution of the intercept and the random effects variance to illustrate that higher dimensional marginal posterior distributions can easily be estimated using Gibbs sampling. Note the negative correlation, as previously discussed by Zeger et al. (1988).

A marginal model with equal correlation for every pair of responses within a subject was also fit to these data for comparison. Here, the marginal mean is modeled by

$$\operatorname{logit} \Pr(Y_{it} = 1) = X_{it}\beta^*.$$

The results are in Table 2. The coefficients β^* from the marginal model are slightly attenuated relative to β. However, qualitatively, the random effects and marginal models give similar findings for the regression coefficient.

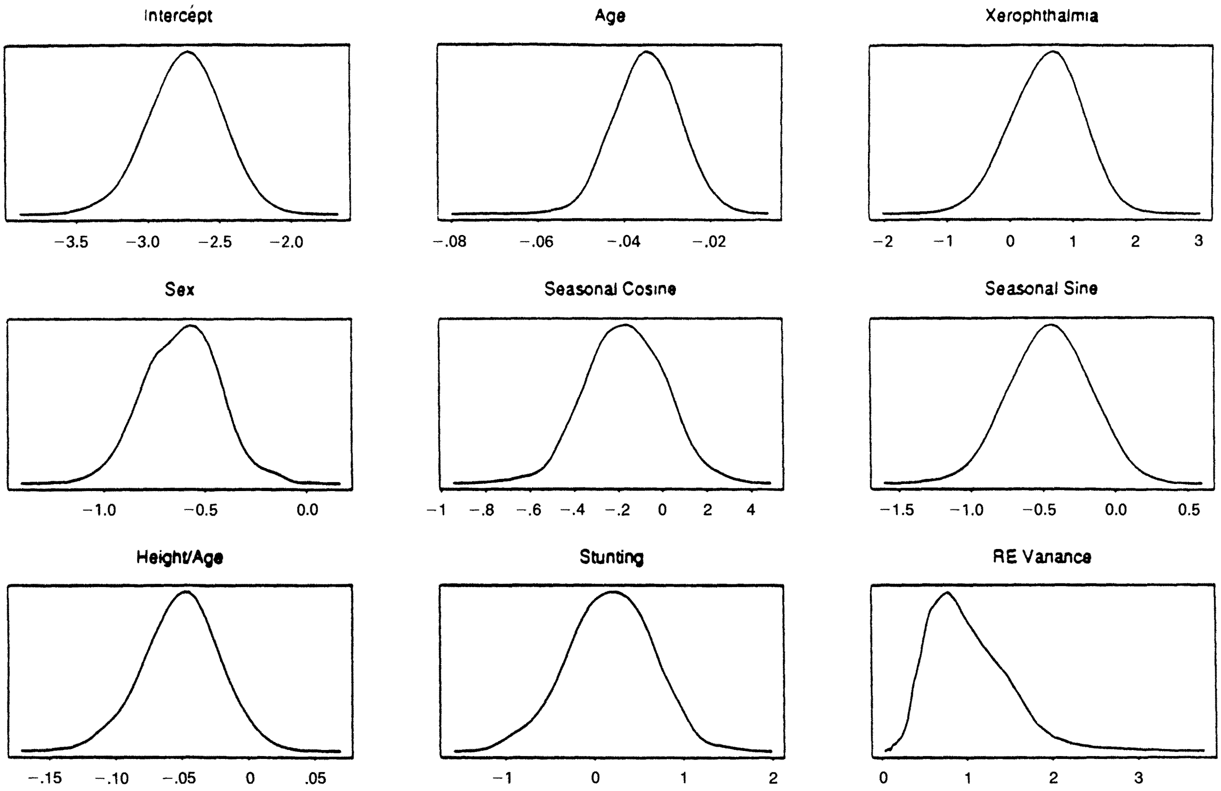

Figure 1. Estimated Marginal Posterior Distributions for Each of the Unknown Parameters in the Indonesian Children's Example Obtained by Applying a Gaussian Kernel Estimate to the Simulated Realizations.

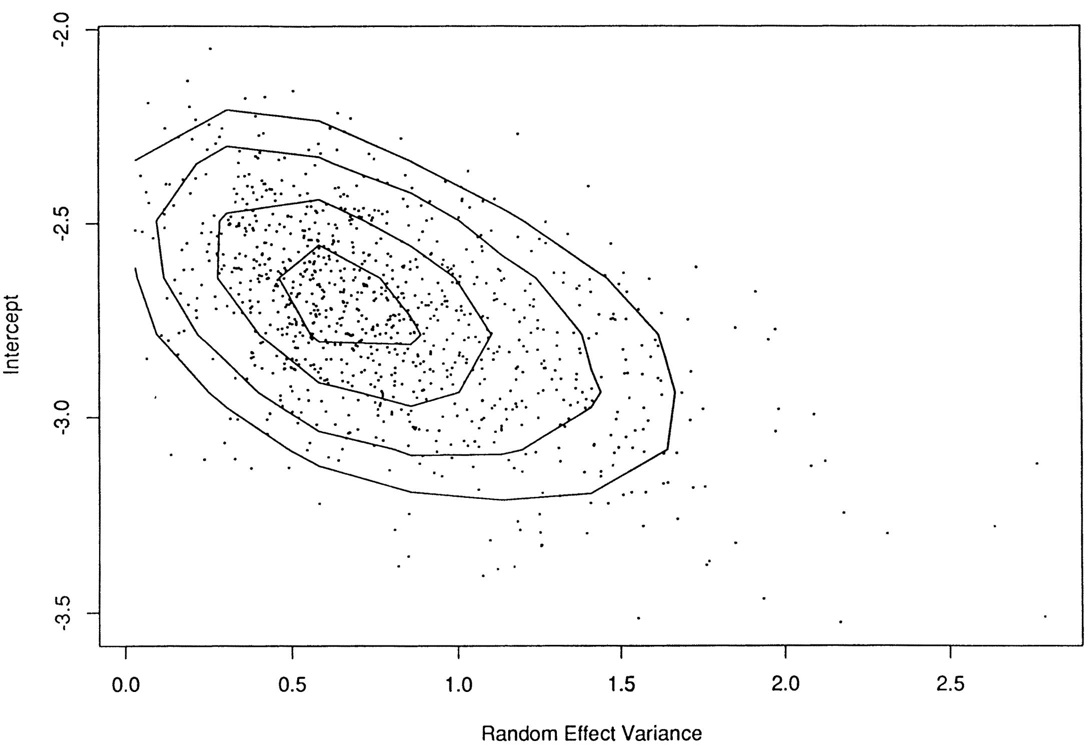

Figure 2. Simulated Values for Random Effects Variance and Intercept, With Contours of Estimated Bivariate Posterior Distribution.

7. DISCUSSION

We have illustrated the Gibbs sampler for estimating parameters in the generalized linear model with Gaussian random effects. We have focused on the logistic-Gaussian case because it is the most common example and poses numerical difficulties (Zeger et al., 1989). In models with only a random intercept (e.g., Anderson and Aitkin 1985; Stiratelli, Laird, and Ware 1984), likelihood evaluation by numerical integration is a competitive alternative to Gibbs sampling from a computational viewpoint. However, the strength of the Gibbs sampling approach is its extensibility to multivariate and non–Gaussian random effects. We have illustrated the multivariate Gaussian case. With large data sets, an effective alternate model for the random effects distribution $F(\theta)$ is a finite mixture of Gaussian distributions (Everitt and Hand 1981), which can represent long-tailed, skewed and bimodal distributions.

The methodology has broad application to other statistical problems with latent or imperfectly observed variables. Specific examples include errors-in-variables regression (e.g., Stefanski 1985), latent class and factor analytic models (Bartholomew 1987) common in psychometric research, and non-Gaussian dynamic time series models (West, Harrison, and Migon 1985; Zeger 1989).

REFERENCES

Anderson, D. A., and Aitkin, M. (1985), "Variance Component Models With Binary Responses: Interviewer Variability," *Journal of the Royal Statistical Society,* Ser. B, 47, 203–210.

Bartholomew, D. J. (1987), *Latent Variable Models and Factor Analysis.* New York: Oxford University Press.

Besag, J. (1974), "Spatial Interaction and the Statistical Analysis of Lattice Systems" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 36, 192–326.

Box, G. E. P., and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis,* Reading, MA: Addison-Wesley.

Breslow, N. E. (1984), "Extra-Poisson Variation in Log-Linear Models," *Applied Statistics,* 33, 38–44.

Clayton, D. G. (1989), "A Monte Carlo Method for Bayesian Inference in Frailty Models," University of Leicester Department of Community Health Technical Report, Leicester, U. K.

Crowder, M. J. (1985), "Gaussian Estimation for Correlated Binary Data," *Journal of the Royal Statistical Society,* Ser. B, 47, 229–237.

Diggle, P. J., and Gratton, R. J., "Monte Carlo Methods of Inference for Implicit Statistical Models" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 46, 193–227.

Everitt, B. S., and Hand, D. J. (1981), *Finite Mixture Distributions.* London: Chapman & Hall.

Gelfand, A. E., Hills, S. I., 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,* 90, 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.

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 6, 721–741.

Geweke, J. (1989), "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica,* 57, 1317–1339.

Gilmour, A. R., Anderson, R. D., and Rae, A. L. (1985), "The Analysis of Binomial Data by a Generalized Linear Mixture Model," *Biometrika,* 72, 593–599.

Harville, D. A., and Mee, R. W. (1984), "A Mixed-Model Procedure for Analyzing Ordered Categorical Data," *Biometrics,* 40, 393–408.

Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika,* 57, 97–109.

Laird, N. M., and Ware, J. H. (1982), "Random Effects Models for Longitudinal Data," *Biometrics,* 38, 963–974.

Leonard, T. (1975), "Bayesian Estimation Methods for Two-Way Contingency Tables," *Journal of the Royal Statistical Society,* Ser. B, 37, 23-37.

Li, K-H. (1988), "Imputation Using Markov Chains," *Journal of Statistical Computing and Simulation,* 30, 57–79.

Liang, K. Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika,* 73, 13–22.

Lindley, D. V., and Smith, A. F. M. (1972), "Bayes Estimates for the Linear Model (with discussion)," *Journal of the Royal Statistical Society,* Ser. B, 34, 1–41.

Lindstrom, M. J., and Bates, D. M. (1988), "Newton–Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated Measure Data," *Journal of the American Statistical Association,* 83, 1014–1022.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman & Hall.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., and Teller, A. H. (1953), "Equations of State Calculations by Fast Computing Machines," *Journal of Chemical Physics,* 21, 1087–1091.

Nelder, J. A. (1972), "Discussion of Paper by Lindley and Smith," *Journal of the Royal Statistical Society,* Ser. B, 24, 1–41.

Nelder, J. A., and Wedderburn, R. W. M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society,* Ser. A, 135, 370–384.

Ochi, Y., and Prentice, R. L. (1984), "Likelihood Inference in a Correlated Probit Regression Model," *Biometrika,* 71, 531–543.

Odell, P. L., and Feiveson, A. H. (1966), "A Numerical Procedure to Generate a Sample Covariance Matrix," *Journal of the American Statistical Association,* 61, 198–203.

Prentice, R. L. (1988), "Correlated Binary Regression With Covariates Specific to Each Binary Observation," *Biometrics,* 44, 1033–1048.

Ripley, B. (1987), *Stochastic Simulation,* New York: John Wiley.

Rubin, D. (1987), "Comment on paper by Tanner and Wong," *Journal of the American Statistical Association,* 82, 543–546.

Sommer, A., Katz, J., and Tarwotjo, I. (1983), "Increased Mortality in Children With Mild Vitamin A Deficiency," *American Journal of Clinical Nutrition,* 40, 1090–1095.

Stefanski, L. A. (1985), "The Effect of Measurement Error on Parameter Estimation," *Biometrika,* 72, 583–592.

Stiratelli, R., Laird, N. M., and Ware, J. H. (1984), "Random-Effects Model for Several Observations With Binary Response," *Biometrics,* 40, 961–971.

Tanner, M., and Wong, W. (1987), "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association,* 82, 528–550.

Thomas, D. C. (1989), "A Monte Carlo Method for Genetic Linkage Analysis," technical report, University of Southern California, Dept. of Preventive Medicine.

Tsutakawa, R. K. (1988), "Mixed Models for Analyzing Geographic Variability in Mortality Rates," *Journal of the American Statistical Association,* 83, 37–42.

Wedderburn, R. W. M. (1974), "Quasi-Likelihood Functions, Generalized Linear Models and the Gauss–Newton Method," *Biometrika,* 61, 439–447.

West, M., Harrison, J. P., and Migon, H. S. (1985), "Dynamic Generalized Linear Models and Bayesian Forecasting" (with discussion), *Journal of the American Statistical Association,* 80, 73–96.

Williams, D. A. (1982), "Extra-Binomial Variation in Logistic Linear Models," *Applied Statistics,* 31, 144–148.

Zeger, S. L., (1988), "A regression Model for Time Series of Counts," *Biometrika*, 75, 621–629.

Zeger, S. L., Liang, K. Y., and Albert, P. S. (1988), "Models for Longitudinal Data: A Generalized Estimating Equation Approach," *Biometrics,* 44, 1049–1060.

Zeger, S. L., See, L. C., and Diggle, P. J. (1988), "Statistical Methods for Monitoring the AIDS Epidemic," *Statistics in Medicine,* 8, 3–22.

Zhao, L. P., and Prentice, R. L. (1989), "Correlated Binary Regression Using a Quadratic Exponential Model," *Biometrika,* 77, 642–648.