A Bayesian Hierarchical Model for Categorical Data with Nonignorable Nonresponse

Paul E. Green
Department of Epidemiology, University of Michigan, Ann Arbor, Michigan 48105, U.S.A.
email: pgreen@umich.edu

and

Taesung Park
Department of Statistics, Seoul National University, Seoul, Korea 151-742
email: tspark@stats.snu.ac.kr

Summary. Log-linear models have been shown to be useful for smoothing contingency tables when categorical outcomes are subject to nonignorable nonresponse. A log-linear model can be fit to an augmented data table that includes an indicator variable designating whether subjects are respondents or nonrespondents. Maximum likelihood estimates calculated from the augmented data table are known to suffer from instability due to boundary solutions. Park and Brown (1994, Journal of the American Statistical Association 89, 44–52) and Park (1998, Biometrics 54, 1579–1590) developed empirical Bayes models that tend to smooth estimates away from the boundary. In those approaches, estimates for nonrespondents were calculated using an EM algorithm by maximizing a posterior distribution. As an extension of their earlier work, we develop a Bayesian hierarchical model that incorporates a log-linear model in the prior specification. In addition, due to uncertainty in the variable selection process associated with just one log-linear model, we simultaneously consider a finite number of models using a stochastic search variable selection (SSVS) procedure due to George and McCulloch (1997, Statistica Sinica 7, 339–373). The integration of the SSVS procedure into a Markov chain Monte Carlo (MCMC) sampler is straightforward, and leads to estimates of cell frequencies for the nonrespondents that are averages resulting from several log-linear models. The methods are demonstrated with a data example involving serum creatinine levels of patients who survived renal transplants. A simulation study is conducted to investigate properties of the model.

Key words: MCMC simulation; Nonignorable missing data; Variable selection.

1. Introduction

When categorical data are collected, it is possible that there will be missing observations with respect to the response variable. To make meaningful inference, it is necessary to adjust for nonresponse by allocating the nonrespondents to the categories of the response variable. For example, the data in Table 1, reported by Sung et al. (1998) and later analyzed by Park (1998), are taken from a clinical trial designed to investigate chronic renal allograft dysfunction in renal transplant patients. Data were available from 109 renal transplant patients who survived more than 4 years after transplant.

The table is cross-classified according to the three variables Gender, Year 1, and Year 4. The variables Year 1 and Year 4 represent serum creatinine levels at years 1 and 4, respectively. After transplant, the serum creatinine levels are expected to decrease over time, showing that the renal function improves. If the creatinine levels do not decrease over time, then there might be some possibility of kidney allograft failure. An estimate of the proportion of patients whose creatinine levels are high (“High”) at Year 4 among those who were low (“Low”) at Year 1 could be a useful indicator for assessing the kidney transplantation procedure.

The data table is incomplete, however, since data are missing on variable Year 4. If we regard Year 4 as the response, note that the respondents are fully categorized with respect to all variables, but nonrespondents are only partially categorized, since their serum creatinine levels are not known at Year 4. If we can allocate the nonrespondents to the High and Low categories of the response variable, the proportion of patients who are High at Year 4 among those who were Low at Year 1 can be estimated.

Handling categorical data with nonresponse has been the subject of much research activity. Nonresponse mechanisms are called ignorable for likelihood-based inference when the response mechanism is independent of the subject’s unobserved response, and nonignorable when the probability of being a nonrespondent depends on the unobserved response. These terms for missing data mechanisms follow the conventions established by Little and Rubin (1987). Many Bayesian approaches for incomplete frequency tables have assumed
ignorable response mechanisms (see, for example, Kaufman and King, 1973; Basu and Pereira, 1983; Gunel, 1984; Albert, 1985; Smith, Choi, and Gunel, 1985; Chiu and Sedransk, 1986). In a simulation study, Park and Brown (1994) showed that it is important to decide whether the underlying response mechanism is ignorable or nonignorable. Estimating cell frequencies using the incorrect underlying model can produce large biases and mean square errors compared to using the correct model.

In the case of nonignorable nonresponse mechanisms for categorical data, Fay (1986) and Baker and Laird (1988) proposed a class of log-linear models for which the data table is augmented by a latent indicator variable that designates whether subjects are respondents or nonrespondents. A log-linear model is then fit to the augmented data table to adjust for nonresponse. Baker and Laird (1988), Conaway et al. (1992), Chambers and Welsh (1993), and Park (1998) demonstrated that maximum likelihood estimations (MLEs) based on different nonignorable models can lead to unstable boundary estimates. In addition, Conaway et al. (1992) discussed issues of model selection and interpretation, along with the effect of discarding nonresponses.

The issue of boundary problems under MLEs produces zero estimates for some of the cells. As a result, the estimate of the log-linear model parameter which represents the nonresponse mechanism is undefined (−∞ or ∞). The explicit condition for the existence of boundary solutions was given in Baker and Laird (1988) and Park and Brown (1997). Park and Brown (1994) demonstrated the instability of maximum likelihood (ML) estimation by showing that a small shift of the nonrespondents can result in large changes in the MLEs of the expected cell frequencies.

For full data problems, Bayesian hierarchical models for categorical and generalized linear models have been proposed and are well established (see, for example, Albert and Chib, 1997). Following their approach, we develop a Bayesian model for incomplete frequency tables with nonignorable nonresponse. In our approach, problems related to estimation and model selection are conveniently incorporated into the Bayesian framework by adopting a hierarchical model with a suitable prior specification. At the first prior, a log-linear model is induced which, when combined with the likelihood, tends to smooth estimates away from the boundary. At the second prior, a Bayesian variable selection procedure, due to George and McCulloch (1997), is introduced; it averages a finite number of models simultaneously, thereby reducing the uncertainty of smoothing estimates towards just one log-linear model.

Until now, most attempts at analyzing categorical data with nonignorable nonresponse have relied on the EM algorithm for estimation. We build on the ideas of Park (1998) by developing a hierarchical Bayesian model that produces posterior estimates of cell means using Markov chain Monte Carlo (MCMC) simulation. By using MCMC, the uncertainty in the unobserved data for nonrespondents is taken into account by generating the missing data at each iteration of the simulation. Section 2 describes the likelihood and prior distributions. In Section 3, the stochastic search variable selection (SSVS) procedure is outlined. The conditional distributions and methods needed for implementation of an MCMC sampler are provided in Sections 4 and 5. In Section 6, the methods are illustrated with an analysis of the data presented in Table 1. In Section 7, a detailed simulation study is presented to investigate the properties of the proposed model. Section 8 is a discussion with some concluding remarks.

2. A Bayesian Model

A Bayesian probability model is developed for contingency tables when the outcome is subject to nonignorable nonresponse. The goal is to estimate cell means for which data are missing so that nonrespondents can be allocated to the categories of the response variable. This is accomplished by expressing prior belief in a log-linear model where the data table is augmented by a latent indicator variable that designates whether outcomes are associated with respondents or nonrespondents. In addition, due to uncertainty in the model selection process for smoothing the contingency table based on a single log-linear model, the prior distribution includes a variable selection procedure for considering a finite number of models simultaneously. This procedure is based on the (SSVS) method developed by George and McCulloch (1997).

Due to the categorical nature of the problem being considered here, and in the context of a missing data problem, the following notation and terms are defined. Denote by \( X = (X_1, X_2, \ldots, X_s) \) an \( s \)-dimensional explanatory variable that is always observed and indexed by \( i = (i_1, i_2, \ldots, i_s) \). Let \( Y \) be the response variable, indexed by \( j \), that may be missing, and let \( R \) be a latent indicator variable that augments the data table. The variable \( R \) is indexed by \( k \) and corresponds with whether a subject is a respondent or a nonrespondent (\( k = 1 \) corresponds to a response, and \( k = 2 \) corresponds to no response). The levels (total number of categories) of \( Y \) and \( X \) are denoted by \( J \) and \( I = I_1 \times \cdots \times I_s \), respectively. Let \( N = (I \times J \times 2) \) be the total number of cells in the augmented contingency table. As an example of this notation, see the

<table>
<thead>
<tr>
<th>Gender</th>
<th>Year 1</th>
<th>Year 2</th>
<th>R</th>
<th>Count</th>
<th>( y_{4,2} )</th>
<th>Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_1 )</td>
<td>( X_2 )</td>
<td>( Y )</td>
<td>( R )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1.82</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>4.26</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>5.19</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>11</td>
<td>10.94</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4.11</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>10</td>
<td>9.95</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>25</td>
<td>24.56</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>?</td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>?</td>
<td>0.61</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>7</td>
<td>7.47</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>7</td>
<td>12</td>
<td>4.17</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>7</td>
<td>12</td>
<td>10.57</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>7</td>
<td>14</td>
<td>3.33</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>?</td>
<td>11.05</td>
<td>9.03</td>
</tr>
</tbody>
</table>

Note: \( X_1 = 1: \) Male, \( X_1 = 2: \) Female, \( X_2 = 1: \) High, \( X_2 = 2: \) Low, \( Y = 1: \) High, \( Y = 2: \) Low, \( R = 1: \) Respondents, \( R = 2: \) Nonrespondents \( y_{4,2} = \) Marginal totals for Year 4.
data in Table 1. The presentation represents a familiar way in which categorical data are coded and submitted to various statistical software programs. Note that in this example, \(I = (I_1 \times I_2) = (2 \times 2)\) and \(J = 2\). In general, there can be more than two explanatory variables, each with more than two levels and, additionally, the response variable can have more than two levels. Regarding the latent data \(R\), a nonignorable model is defined to be a model that includes a \(Y-R\) interaction term, since the probability of response is associated with the outcome variable, and an ignorable model does not contain a \(Y-R\) interaction term. This terminology follows the framework established by Little and Rubin (1987) for missing data mechanisms.

The model is developed in a Bayesian framework and is comprised of a likelihood and a prior distribution. The observed cell frequencies for the respondents for which \(\eta_{ij2}\) is normalized. Conditional on marginal totals \(y_{i2}\) summed over the nonresponses, the unobserved cell frequencies for which \(\eta_{ij1}\) is Poisson, such that the log of the Poisson mean is \(\eta_{ij1}\).

\[
y_{ij1} | \eta_{ij1} \sim \text{Poisson} (\exp (\eta_{ij1}))
\]

In the data example, the Poisson counts refer to the first 8 observations in Table 1 in which the data are fully categorized. Conditional on marginal totals \(y_{i2}\) summed over the nonresponses, the unobserved cell frequencies for which \(k = 2\), denoted by \(y_{ij2}\), are multinomial

\[
y_{i2}, \ldots, y_{i2} | \sum_j y_{ij2} = y_{i2}, \pi_{i2}, \ldots, \pi_{i2}\sim \text{multinomial} (y_{i2}; \pi_{i2}, \ldots, \pi_{i2})
\]

under the usual constraints for multinomial sampling

\[
\sum_j \pi_{ij2} = 1 \quad \text{and} \quad \pi_{ij2} = \frac{\exp (\eta_{ij2})}{\sum_j \exp (\eta_{ij2})}.
\]

Note that the multinomial probabilities \(\pi_{ij2}\) are expressed in terms of \(\eta_{ij2}\). As shown in Table 1, the \(y_{ij2}\) are not observed, but the marginal totals \(y_{i2}\) are known. In the data example \(J = 2\) and the \(y_{ij2}\) are actually binomial at each level of \(i\).

At the first prior, a log-linear model is induced by allowing the log of expected cell frequencies \(\eta_{ijk}\) over the entire table to be normal

\[
\eta | \beta, \Sigma \sim N (Z \beta, \Sigma),
\]

where \(\eta\) is \(N \times 1\), \(\beta\) is \(p \times 1\), \(\Sigma\) is an \(N \times N\) diagonal covariance matrix, and \(Z\) is the \(N \times p\) design matrix. The covariance matrix \(\Sigma\) has only two parameters, \(\sigma^2_1\) and \(\sigma^2_2\), corresponding to \(k = 1\) for the respondents and \(k = 2\) for the nonrespondents. If \(N_1 + N_2 = N\) with \(N_1 = N_2\), then the first \(N_1\) diagonal elements of \(\Sigma\) are \(\sigma^2_1\), and the remaining \(N_2\) diagonal elements are \(\sigma^2_2\). We parameterize \(\Sigma\) in this manner, to reflect the prior belief that the variances, with \(\sigma^2_1\) corresponding to the respondents and \(\sigma^2_2\) for the nonrespondents, may not be equal. Where convenient, \(\eta\) will be expressed in terms of two individual independent components for respondents and nonrespondents

\[
\eta_1 | \beta, \sigma^2_1 \sim N (Z_1 \beta, \sigma^2_1 I) \quad \text{and} \quad \eta_2 | \beta, \sigma^2_2 \sim N (Z_2 \beta, \sigma^2_2 I)
\]

where

\[
Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}
\]

and \(I\) is the \(N_1 \times N_1\) (or \(N_2 \times N_2\)) identity matrix.

As a consequence of missing data on \(y_{ij2}\), a feature of this model that distinguishes it from other Bayesian models is that \(y_{ij2}\) is designated as likelihood, but it is not observed. In the posterior, it will be treated as other unknown parameters and estimated by MCMC simulation. This suggests that it may be suitable to entertain a prior for \(y_{ij2}\). By not assigning a prior, a discrete uniform prior over all combinations of \((y_{i21}, \ldots, y_{i2J})\) is implied. We expect that this model could be sensitive to more informative priors, especially in the presence of missing data. If information were available regarding the missing data mechanism, an informative prior could be used.

At the second stage, the hyperparameters \(\beta, \sigma^2_1\) and \(\sigma^2_2\) are assigned vague, but proper, priors. The regression parameters \(\beta\) are modeled as \(p\)-variate normal

\[
\beta \sim N_p (0, \Theta^0).
\]

The value \(10^{6}\) provides for a sufficiently flat and vague prior, while retaining the properties of a proper probability model. The parameters \(\sigma^2_1\) and \(\sigma^2_2\) are modeled as conjugate inverse gamma priors

\[
\sigma^2_1 \sim \text{Inv-Gamma}(\nu_1/2, \lambda_1/2) \quad \text{and} \quad \sigma^2_2 \sim \text{Inv-Gamma}(\nu_2/2, \lambda_2/2)
\]

where \(\nu_1, \lambda_1, \nu_2, \text{and} \lambda_2\) are known and fixed in advance.

3. A Bayesian Variable Selection Procedure

The prior for \(\beta\) can be modified to accommodate a Bayesian variable selection procedure for entertaining a finite number of log-linear models simultaneously. Under the SSVS method developed by George and McCulloch (1997), additional binary latent variables can be included as part of the MCMC simulation, such that, as the simulation progresses, the most promising models are visited most frequently. In particular, \(\beta\) is modeled as \(p\)-variate normal

\[
\beta | \gamma \sim N_p (0, D^{\gamma})
\]

where \(D^{\gamma}\) is a diagonal matrix. Let \(\beta_l, l = 1, \ldots, p\) be the independent components of \(\beta\) and let \(m = 1, \ldots, M \leq p\) be the indexed subset of \(\{1, 2, \ldots, p\}\) for which testing \(\beta_m = 0\) is of interest. The vector \(\gamma\) is a sequence of ones and zeroes of length \(M\). Then, the diagonal elements of \(D^{\gamma}\) are, say, \(10^3\), which represents a vague prior when \(\beta_l = 0\) is not being tested, and \([1 - \gamma_m]\tau_m + \gamma_m c_m \tau_m\) when \(\beta_m = 0\) is being tested. The idea behind this strategy is that \(\beta_m | \gamma_m\) can be viewed as a mixture of two normals

\[
(1 - \gamma_m) N \left(0, \tau_m^2 \right) + \gamma_m N \left(0, (c_m \tau_m)^2 \right)
\]

such that the real scalar \(\tau_m\) is chosen small, and the real scalar \(c_m\) is chosen large. Figure 1 provides a visual interpretation...
The two normals intersect at $\delta$.

Figure 1. The two normals intersect at $\delta$.

of this strategy. When $\gamma_m = 1$, the regression coefficient $\beta_m$ (the prior shown by the dashed line) should be included in the model, and when $\gamma_m = 0$, the regression coefficient $\beta_m$ (the prior shown by the solid line) should be excluded from the model. Note that the dimension of the parameter space $p$ does not change when it is determined that certain elements of $\beta$ are sufficiently close to zero. This is an important consideration when implementing MCMC sampling.

George and McCulloch (1997) present guidelines and discuss criteria for choosing $\tau_m$ and $c_m$ based on Figure 1. The first step is to specify a value $\delta_m$ such that $|\beta_m| < \delta_m$. The value $\delta_m$ represents an a priori significance level. The second step is to choose a value $c_m$ in the range 20–100. Empirical evidence has shown that this range of values for $c_m$ tends to work well when implementing MCMC. Finally, $\tau_m = [2c_m\log(c_m)/(c_m^2 - 1)]^{-0.5} \delta_m$ can be calculated using the relationship shown in Figure 1. We have found that results based on SSVS can be sensitive to choices of $\delta_m$ and $c_m$. However, this is not unusual, for example, in much the same way as variable selection procedures in classical data analysis are sensitive to commonly used specified significance levels, such as $\alpha = 0.05$ or $\alpha = 0.15$.

Alternative methods for model selection exist and have been considered. Reversible jump MCMC is a viable alternative for the present model and is discussed at great length in Green (1995). In addition, Richardson and Green (1997) explain in great detail and provide many excellent examples of how reversible jump MCMC allows a complete change of parameter dimension in Bayesian model selection and mixture modeling. The main reason for our choice of SSVS is based on its compatibility and natural incorporation into the model.

As an illustration of SSVS, we may be interested in fitting a model to the data in Table 1, which can be written using standard notation in categorical data analysis as

$$(X_1X_2Y, X_1R, X_2R, YR),$$
or, equivalently, by including the main effects, some two-way interactions, and one three-way interaction as

$$\eta_{ijk} = \mu + \lambda_{X_1}^{X_1i} + \lambda_{X_2}^{X_2j} + \lambda_{Y}^{Y} + \lambda_{R}^{R} + \lambda_{ij}^{ij} + \lambda_{i}^{i} + \lambda_{j}^{j} + \lambda_{ij}^{ij} + \lambda_{i}^{i} + \lambda_{j}^{j} + \lambda_{ij}^{ij} + \lambda_{i}^{i} + \lambda_{j}^{j},$$

where, in our notation, $\beta^R = (\mu, \lambda_{X_1}^{X_1i}, \lambda_{X_2}^{X_2j}, \ldots)$ denotes the model parameters. But there is uncertainty in considering just one model, so we may be interested in fitting several models simultaneously. Suppose we are interested in the following four models

$$M1 \ (X_1X_2Y, X_1R, X_2R, YR)$$
$$M2 \ (X_1X_2Y, X_1R, YR)$$
$$M3 \ (X_1X_2Y, X_2R, YR)$$
$$M4 \ (X_1X_2Y, YR).$$

Note that these models are not nested. Then we can define a vector, say, $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_M)$, where $\gamma_m$, $m = 1, \ldots, M$ is defined

$$\gamma_m = \begin{cases} 1 & \text{include parameter} \\ 0 & \text{exclude parameter} \end{cases}$$

so that $\gamma$ is associated with the four models by the relationship

$$M1 \ (X_1X_2Y, X_1R, X_2R, YR) \ \gamma = (11)$$
$$M2 \ (X_1X_2Y, X_1R, YR) \ \gamma = (10)$$
$$M3 \ (X_1X_2Y, X_2R, YR) \ \gamma = (01)$$
$$M4 \ (X_1X_2Y, YR) \ \gamma = (00).$$

Then, as part of the MCMC simulation, a sequence is generated

$$\gamma^1, \gamma^2, \gamma^3, \ldots ,$$
such that the most promising models are visited most frequently. Note that this is a sequence of four nonignorable models, since each contains the $Y$-$R$ interaction term. In addition, in this example, $M = 2$ because only two terms are being considered for inclusion or exclusion.

At the third and final stage of the hierarchical model, the $\gamma_m$’s are modeled as independent Bernoulli priors

$$p(\gamma) = \prod_{m=1}^{M} w_m^{\gamma_m} (1 - w_m)^{(1-\gamma_m)}.$$

In practice, one convenient parameterization is to let $w_m = 0.5$, resulting in $p(\gamma) = (0.5)^M$. This prior gives no a priori preference for inclusion or exclusion of $\beta_m$. Although some interest is associated with estimates for $\beta, \sigma_1^2, \sigma_2^2,$ and $\gamma$, the real focus of this model is on estimating the cell means for which data are missing, so that nonrespondents can be allocated to the categories of the response variable $Y$. 

A Bayesian Hierarchical Model for Categorical Data with Nonignorable Nonresponse 889
4. Conditional Distributions for MCMC Sampling

The joint posterior distribution of all unknown parameters including the unobserved \( y_{i2j} \) can be written as

\[
p(\eta, \beta, \sigma_1^2, \sigma_2^2, \gamma, y_{i1j}, y_{i2j}) = \prod_{i} \left\{ \prod_{j} \exp \left[ -\frac{1}{2} \sum_{m}^M \left( \frac{1}{\sigma_1^2} y_{i1j} - \eta_{i1j} \right)^2 \right] \right\} \\
\times \prod_{i} \left\{ \prod_{j} \exp \left[ -\frac{1}{2} \sum_{m}^M \left( \frac{1}{\sigma_2^2} y_{i2j} - \eta_{i2j} \right)^2 \right] \right\}.
\]

The first line in (1) consists of the Poisson and multinomial likelihoods. The second line contains the normal priors for \( \eta \) and \( \beta \). The third line is a product of the inverse gamma priors for \( \sigma_1^2 \) and \( \sigma_2^2 \), respectively. The constant prior for \( \gamma \), outlined in Section 2, is used and therefore omitted. Based on (1), we now present various conditional distributions to implement MCMC sampling. The conditional for \( \beta \) is normal

\[
\beta \mid \eta, \sigma_1^2, \sigma_2^2, \gamma \sim N_p \left( \left( Z^T \Sigma_{\gamma}^{-1} Z + D_{\gamma}^{-2} \right)^{-1} Z^T \Sigma_{\gamma}^{-1} \eta, \left( Z^T \Sigma_{\gamma}^{-1} Z + D_{\gamma}^{-2} \right)^{-1}. \right)
\]

The conditionals for \( \sigma_1^2 \) and \( \sigma_2^2 \) are both inverse gamma distributions:

\[
\sigma_1^2 \mid \eta_1, \beta \sim \text{Inv-Gamma} \left( \frac{N_1 + \nu_1}{2}, \frac{\nu_1 \lambda_1 + (\eta_1 - Z_1 \beta)^T (\eta_1 - Z_1 \beta)}{2} \right),
\]

\[
\sigma_2^2 \mid \eta_2, \beta \sim \text{Inv-Gamma} \left( \frac{N_2 + \nu_2}{2}, \frac{\nu_2 \lambda_2 + (\eta_2 - Z_2 \beta)^T (\eta_2 - Z_2 \beta)}{2} \right).
\]

The distribution for \( y_{i1j}, \ldots, y_{iLj} \), conditional on \( y_{i2j} \) being observed, is multinomial

\[
y_{i1j}, \ldots, y_{iLj} \mid \sum_{j} y_{i2j}, y_{i1j}, \ldots, y_{iLj} \sim \text{multinomial} \left( y_{i1j}, \ldots, y_{iLj} \right)
\]

where \( \pi_{i2j} \) can be expressed in terms of \( \eta_{i2j} \), as shown in Section 2. The conditionals for \( \gamma_m, m = 1, \ldots, M \) are independent Bernoullis

\[
\gamma_m \mid \beta_m \sim \text{Bernoulli} \left( p_m \right),
\]

which by application of Bayes’ formula

\[
p_m = P(\gamma_m = 1 \mid \beta_m) = \frac{P(\beta_m \mid \gamma_m = 1) P(\gamma_m = 1)}{P(\beta_m \mid \gamma_m = 1) P(\gamma_m = 1) + P(\beta_m \mid \gamma_m = 0) P(\gamma_m = 0)}
= \frac{1}{c_m} \exp \left( -\frac{\beta_m^2}{2(c_m \gamma_m)^2} \right)
= \frac{1}{c_m} \exp \left( -\frac{\beta_m^2}{2(c_m \gamma_m)^2} \right) + \exp \left( -\frac{\beta_m^2}{2(c_m \gamma_m)^2} \right).
\]

All the conditionals except \( \eta \) are known distributions and can be sampled from directly. The focus of the next section is devoted to sampling from \( \eta \).

5. Sampling from the Conditional of \( \eta \)

In short, we use an independence sampler with a normal candidate to sample from the conditional of \( \eta \). (See, for example, Tierney (1994) for details concerning the independence sampler.) Some asymptotic results are provided to justify this choice. The parameter \( \eta \) appears in the posterior distribution (1) as a function of the Poisson likelihood for \( y_{i1j} \), the multinomial likelihood for \( y_{i2j} \), which can be viewed as latent data, and in the normal prior. The conditional for \( \eta \) depends on \( \beta, \sigma_1^2 \), and the entire data vector, \( y = (y_{i1j}, y_{i2j}) \).

Since this conditional depends on the entire data vector, we can think in terms of a full data problem, and the conditional can be sampled from without regard to any missing data. In addition, as far as estimation of \( \eta \) is concerned, Poisson and multinomial sampling are equivalent, so the conditional portion can be treated as Poisson and combined into a single likelihood. As a result, the conditional distribution for \( \eta \) for a single observation \( \eta_{ijk} \) is

\[
p(\eta_{ijk} \mid \beta, \sigma_k^2, y_{ij}) \propto \exp \left[ -\frac{1}{2 \sigma_k^2} (\eta_{ijk} - z_{ijk} \beta)^2 + y_{ijk} \eta_{ijk} - e^{\eta_{ijk}} \right],
\]

where \( z_{ijk}^T \) is the \( ijk \)th row of \( Z \). Inspection of the conditional distribution in (2), however, reveals that it is a product of two separate distributions. The first part resembles the kernel of a normal distribution with mean \( z_{ijk}^T \beta \) and variance \( \sigma_k^2 \), and the second part resembles the kernel of a one-parameter log-gamma density indexed by \( y_{ijk} \). Thus, we expect the conditional in (2) to be quite close to normal.

The properties of the log-gamma density are well understood (see, for example, Lawless (1982)) and including the normalizing constant, it is

\[
p(\eta_{ijk} \mid y_{ijk}) = \frac{\exp(y_{ijk} \eta_{ijk} - e^{\eta_{ijk}})}{\Gamma(y_{ijk})},
-\infty < \eta_{ijk} < \infty, \quad y_{ijk} > 0.
\]

In particular, it is known that as \( y_{ijk} \to \infty \)

\[
\sqrt{y_{ijk}}(\eta_{ijk} - \log y_{ijk}) \overset{d}{\to} N(0,1),
\]

and that this distribution is quite close to normal, even for \( y_{ijk} \) as small as 5. In fact, application of Laplace’s method, by expanding \( \log p(\eta_{ijk} \mid y_{ijk}) \) in a second-order Taylor series about the value \( \bar{\eta}_{ijk} \) that maximizes \( \log p(\eta_{ijk} \mid y_{ijk}) \), leads to the same result, namely,

\[
\log p(\eta_{ijk} \mid y_{ijk}) \approx -\frac{1}{2} y_{ijk} (\eta_{ijk} - \log y_{ijk})^2.
\]

Substitution in (2) gives a normal approximation

\[
p(\eta_{ijk} \mid \beta, \sigma_k^2, y_{ij}) \propto \exp \left[ -\frac{1}{2} \left( \frac{(\eta_{ijk} - z_{ijk}^T \beta)^2}{\sigma_k^2} + \frac{(\eta_{ijk} - \log y_{ijk})^2}{1/y_{ijk}} \right) \right].
\]
The conditional distribution now appears as a product of two normals and we can recognize that
\[ \eta_{ijk} | \beta, \sigma_k^2, y_{ijk} \sim N\left( \theta_{ijk}, \delta_{ijk}^2 \right), \]
where
\[ \theta_{ijk} = \frac{1/\sigma_k^2 x_{ijk}^T \beta + y_{ijk} \log y_{ijk}}{1/\sigma_k^2 + y_{ijk}} \]
and
\[ \delta_{ijk}^2 = \left( \frac{1}{\sigma_k^2} + y_{ijk} \right)^{-1}. \]

As usual, in Bayesian data analysis, the posterior mean is a weighted average of the data and the prior mean. In this case, the approximate mean \( \theta_{ijk} \) is a weighted average of the log-linear regression function \( x_{ijk}^T \beta \) and the data \( \log y_{ijk} \), with weights determined by the variances \( 1/\sigma_k^2 \) and \( y_{ijk} \). Thus, \( 1/\sigma_k^2 \) can be expected to be approximately the same order of magnitude as the data values \( y_{ijk} \).

The normal approximation is derived merely for the purpose of providing guidelines for choosing a candidate distribution for implementing MCMC. It provides approximations to the first two moments of the conditional distribution, and this information is valuable for MCMC design. For example, in our initial attempts, we tried a Metropolis sampler with a normal candidate and fixed variance \((2)\). In practice, however, (1995) and the references therein concerning guidelines for constructing efficient MCMC samplers. In practice, however, by trial and error, we found that an independence sampler (see, for example, Tierney (1994)) performs well using a normal candidate centered at \( \log y_{i1} \) for the respondents, and an estimate for \( \log y_{i2} \) for the nonrespondents, with scale \( 2.4 \delta_{ijk} \).

We base these judgments largely on MCMC diagnostic tools, such as index plots and acceptance rates. Note that \( y_{i1} \) for the respondents remains fixed, whereas \( y_{i2} \) will change at each iteration, since these values are being generated from a multinomial for the nonrespondents. Values for \( \log y_{i2} \) for use in the independence sampler can be approximated from \( y_{i2} \).

On the log scale, these values tend to be quite reasonable. If any \( y_{ijk} = 0 \), the candidate can be centered at \( 0.5 \).

In using an independence sampler, we are sensitive to the situation that it may be possible that the candidate distribution does not adequately cover the true underlying distribution, even after scaling the variance. We have found that this is an important consideration, and examination of diagnostic tools from several MCMC runs helps to alleviate this problem. In addition, a detailed simulation study, provided in Section 7, was designed to test this specific situation. Various properties of the model were learned from the simulation study. In some cases, the scale of the candidate distribution needs to be inflated to improve MCMC coverage.

The log-linear model helps to smooth the contingency table. An independence sampler tends to perform well in this case, because estimates based on the log-linear model should not be very different from the data, and a good candidate distribution can be determined in advance.

6. Data Example

In this example, we analyze the data shown in Table 1. As stated in the introduction, the goal of this analysis is to allocate the nonrespondents to the High and Low categories corresponding to the response variable Year 4, and then estimate the proportion of patients who are High at Year 4 among those who were Low at Year 1.

Park (1998) fit several ignorable and nonignorable models to these data and found that the nonignorable models consistently provided a better fit. From among a group of four nonignorable models, Park chose one that appeared to give the best fit based on a likelihood ratio statistic and degrees of freedom. In this data example, MCMC simulation is used to average over the same four models, labeled M1, M2, M3, and M4, and described in Section 2. The MCMC simulation is partially based on the SSVS procedure developed by George and McCulloch (1997).

The output generated from this MCMC sampler exhibits more autocorrelation for the unobserved nonresponse observations than for the observed responses, where data is complete. For this reason, we ran a fairly long simulation based on a chain of size 60,000, discarding the initial 10,000 simulated values as burn-in. Once we have an approximate posterior sample \( (\eta^1, \ldots, \eta^{50,000}) \), we may undertake various posterior tasks, as usual. For example, for some real valued function \( f \), posterior estimates can be evaluated via

\[ \hat{E}[f(\eta)] = \frac{1}{50,000} \sum_{i=1}^{50,000} f(\eta^i). \]

Setting \( f(\eta) = \exp(\eta) \) in the above formula gives estimates based on the four models given in the right portion of Table 1. When \( R = 1 \), it is possible to compare observed and fitted values.

To estimate the proportion of patients who are ranked High at Year 4 among those who were Low at Year 1, once again it is necessary to average the appropriate function over the sample of 50,000 simulated values. In this case, our estimate is 43.3%. This estimate is less than Park’s estimates of 44.7% and 45.2%, which are based on two different methods using model M3 alone.

In Figures 2 and 3, we provide kernel density plots based on the 50,000 sampled values of \( \eta \). The left side of each figure displays plots for the full data, and the right side displays plots associated with the nonrespondents where data is missing. Note that when there is no missing data, the plots are fairly well-behaved, and even close to normal. However, when missing data are present, the plots are often skewed and non-normal. These plots demonstrate the difficulty with estimation involving the nonrespondents.

Since this is a categorical data problem, it is necessary to place constraints on parameters so that the model is identifiable and the design matrix \( Z \) is not singular. In regression problems, MCMC runs more efficiently when covariates are centered (see, for example, Gelfand, Sahu, and Carlin (1995)). We constrain model parameters to sum to zero, resulting in a design matrix consisting of only the two values \(-1\) and \(1\).

For the SSVS method, we set \( c_m = 50 \) and \( \tau_m = 0.002 \), \( m = 1, 2 \). These settings resulted in all four models receiving
approximately equal weight. Finally, we set \( \nu_1 = 4 \), \( \lambda_1 = 0.22 \), \( \nu_2 = 4 \), and \( \lambda_2 = 0.60 \). These settings are based on the large sample results presented in Section 5, and the prior belief that the variance for nonrespondent missing data is greater than the variance when data is complete.

7. Simulation Study
We performed simulation studies to compare the proposed Bayesian method with previous methods of Park and Brown (1994), Park (1998), and ML estimation (Baker and Laird, 1988). For simplicity, we restricted attention to \( 2 \times 2 \times 2 \)
tables. As a criterion for comparison, the mean square errors (MSE) of the estimates of the expected cell frequencies were used.

We assume that the true underlying model is \((XY, YR)\). This model introduces three main-effect parameters and two interaction parameters. These five parameters uniquely determine the cell probability vector \(\pi = (\pi_{111}, \pi_{121}, \pi_{211}, \pi_{221}; \pi_{112}, \pi_{122}, \pi_{212}, \pi_{222})^T\). To determine the three main-effect parameters, we use the following three parameters, which represent the ratio of the marginal probabilities of the first level and those of the second level: (i) for \(X\), \(\psi_X = \pi_{2+}/\pi_{1+}\), (ii) for \(Y\), \(\psi_Y = \pi_{+2}/\pi_{+1}\), and (iii) for \(R\), \(\psi_R = \pi_{+2}/\pi_{+1}\). The

Figure 3. Kernel density plots for logs of cell means 5–8, 13–16.
remaining two interaction parameters are determined by the odds ratios, \( \psi^Y \) and \( \psi^R \), where \( \beta^Y = \log \psi^Y / 4 \) and \( \beta^R = \log \psi^R / 4 \). A further detailed description on determining cell probabilities is given in Park (1998).

For a given set of cell probabilities, eight cell frequencies \( y_{ijk} \) \((i = 1, 2; j = 1, 2; k = 1, 2)\) were generated from the multinomial distribution, with parameters \( N \) and \( \pi \). We chose the total sample size, \( N \), as 200 and fixed the number of replications at 500. After the cell frequencies were generated, the vector of observed cell frequencies \( \hat{y}_{ijk} \) was obtained by setting \( \hat{y}_{ijk} = y_{ijk} + \epsilon_{ijk} \) for \( i = 1, 2 \), \( j = 1, 2 \), and \( k = 1, 2 \). The nonignorable nonresponse model \((X, Y, R)\) was fitted to the simulated data first by ML, then by Park and Brown’s method (PB), and Park’s method (P). Also, for the purpose of comparison, the same model was fitted to the complete data using all eight cells. For the Bayesian method (B), the SSVS procedure was implemented by considering two models, \((X, Y, R)\), and \((X, R, Y)\), so that a univariate \( \gamma \) was either 1 or 0, depending on whether the term \( X \) was included or excluded. SSVS parameters were set such that the posterior mean of \( \gamma \) was approximately 0.1 in all cases.

We summarize the results for four nonrespondent cells \( y_{i+2} \) and four marginal sums \( y_{i+j} \). Table 2(a) shows the results for the case when \( (\psi^X, \psi^Y, \psi^R, \psi^{XY}, \psi^{YR}) = (1.00, 1.00, 0.25, 0.25, 0.25) \). Table 2(b) shows the results for the case when \( (\psi^X, \psi^Y, \psi^R, \psi^{XY}, \psi^{YR}) = (1.00, 1.00, 0.25, 0.25, 4.00) \). The first four rows summarize the results of \( y_{i+2} \) and the next four rows summarize the results of \( y_{i+j} \).

8. Concluding Remarks
In the past, most attempts to estimate cell frequencies in contingency tables with missing data have relied on the EM algorithm. In this work, the missing data is generated at each iteration of MCMC simulation. Thus, the missing data changes at each iteration. This forms the basis for our belief that there should be two sources of variation, namely, \( \sigma_1^2 \) for the respondents, and \( \sigma_2^2 \) for the nonrespondents. The priors for \( \sigma_1^2 \) and \( \sigma_2^2 \) are fairly strong, but we have found that this is sometimes necessary when dealing with categorical data with missing

Table 2
MSEs for the Nonignorable Nonresponse Model \((X, Y, R)\) in \(2 \times 2 \times 2\) tables with \( N = 200\)

<table>
<thead>
<tr>
<th>Cell ((i, j, k))</th>
<th>MLE (R_{PB})</th>
<th>Park &amp; Brown (R_B)</th>
<th>Park (R_P)</th>
<th>Bayesian method (R_B)</th>
<th>Complete MLE (R_C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1, 1, 2)</td>
<td>40.078</td>
<td>35.152</td>
<td>0.877</td>
<td>31.350</td>
<td>0.782</td>
</tr>
<tr>
<td>(1, 2, 2)</td>
<td>11.692</td>
<td>11.437</td>
<td>0.978</td>
<td>9.402</td>
<td>0.804</td>
</tr>
<tr>
<td>(2, 1, 2)</td>
<td>16.536</td>
<td>11.902</td>
<td>0.720</td>
<td>10.844</td>
<td>0.656</td>
</tr>
<tr>
<td>(2, 2, 2)</td>
<td>36.596</td>
<td>33.021</td>
<td>0.903</td>
<td>27.915</td>
<td>0.763</td>
</tr>
<tr>
<td>(1, 1, +)</td>
<td>61.797</td>
<td>65.065</td>
<td>1.053</td>
<td>59.831</td>
<td>0.968</td>
</tr>
<tr>
<td>(1, 2, +)</td>
<td>44.018</td>
<td>49.983</td>
<td>1.136</td>
<td>43.352</td>
<td>0.985</td>
</tr>
<tr>
<td>(2, 1, +)</td>
<td>51.510</td>
<td>44.793</td>
<td>0.869</td>
<td>41.921</td>
<td>0.814</td>
</tr>
<tr>
<td>(2, 2, +)</td>
<td>67.439</td>
<td>65.799</td>
<td>0.976</td>
<td>60.871</td>
<td>0.903</td>
</tr>
<tr>
<td>(1, 1, 2)</td>
<td>40.092</td>
<td>34.394</td>
<td>0.858</td>
<td>30.592</td>
<td>0.763</td>
</tr>
<tr>
<td>(1, 2, 2)</td>
<td>17.527</td>
<td>13.278</td>
<td>0.758</td>
<td>12.121</td>
<td>0.692</td>
</tr>
<tr>
<td>(2, 1, 2)</td>
<td>12.667</td>
<td>11.786</td>
<td>0.930</td>
<td>10.366</td>
<td>0.792</td>
</tr>
<tr>
<td>(2, 2, 2)</td>
<td>41.094</td>
<td>35.945</td>
<td>0.875</td>
<td>32.234</td>
<td>0.784</td>
</tr>
<tr>
<td>(1, 1, +)</td>
<td>74.755</td>
<td>69.239</td>
<td>0.926</td>
<td>65.915</td>
<td>0.882</td>
</tr>
<tr>
<td>(1, 2, +)</td>
<td>59.158</td>
<td>53.350</td>
<td>0.902</td>
<td>50.170</td>
<td>0.848</td>
</tr>
<tr>
<td>(2, 1, +)</td>
<td>40.092</td>
<td>34.394</td>
<td>0.858</td>
<td>30.592</td>
<td>0.763</td>
</tr>
<tr>
<td>(2, 2, +)</td>
<td>11.692</td>
<td>11.902</td>
<td>0.720</td>
<td>10.844</td>
<td>0.656</td>
</tr>
</tbody>
</table>
data. When dealing with a continuous response, it is sometimes possible to relax these restrictions.

Even though this model allows for different variances for respondents and nonrespondents, it might be possible to consider a covariance matrix for Σ other than a diagonal one, especially if we have some prior knowledge of the missing data mechanism or the covariance structure between the two groups. For example, specific covariance structures that depend on several unknown parameters could be incorporated into the model. It would be possible to generalize the independent inverse gamma priors to an inverse Wishart prior by considering a covariance structure such as \( \Sigma^{1/2} R \Sigma^{1/2} \), where Σ is a diagonal matrix, as in our model, and \( R \) is a correlation matrix. Various forms of \( R \) can be entertained. One idea is to allow all off-diagonal elements of \( R \) to equal a single parameter \( \rho \). Certain band or block diagonal correlation matrices could also be considered. At present, we consider a diagonal covariance matrix and plan to investigate the effects of a more general covariance structure in future work.

All the inference presented is based on the output from MCMC simulation and is therefore restricted to sample means. No attempt is made to estimate standard errors for cell means using, for example, the method of batched means. No attempt is made to estimate standard errors for categorical variables with outcome subject to nonignorable nonresponse. Journal of the American Statistical Association 83, 62–69.


