Bayesian Posterior Estimation of Logit Parameters with small Samples

Galindo-Garre, F.; Vermunt, J.K.; Bergsma, W.P.

Published in:
Sociological Methods and Research

Publication date:
2004

Document Version
Peer reviewed version

Bayesian posterior estimation of logit parameters with small samples

Francisca Galindo-Garre (1), Jeroen K. Vermunt (2)
and Wicher P. Bergsma (3)

Department of Methodology and Statistics

Tilburg University

P.O.Box 90153

5000 LE Tilburg

Tel: +31.13.466.2325

Fax: +31.13.466.3002

E-mail: (1) f.galindogarre@uvt.nl, (2) j.k.vermunt@uvt.nl
(3) w.p.bergsma@uvt.nl

Bayesian posterior estimation of logit parameters with small samples

Abstract

When the sample size is small compared to the number of cells in a contingency table, it is well known that maximum likelihood estimates of logit parameters and their associated standard errors may not exist or may be highly biased. An increasingly popular method of dealing with this problem is to "smooth" the estimates by assuming a certain prior distribution for the logit parameters. For example, a default option in SPSS 10 is to add the constant 0.5 to the cell frequencies in saturated models, which corresponds to using a Dirichlet prior with parameter 1.5 for the cell probabilities. The aim of this paper is to investigate the performance of point and interval estimates obtained by assuming a variety of prior distributions for model parameters. A small-scale simulation study is done to investigate the bias and median squared error of the posterior mode and the posterior mean Bayesian point estimators, and the coverage probabilities and median width of associated confidence intervals. We focus on two logit parameters of a $2 \times 2 \times 2$ table: (i) the logit interaction effect of two predictors on a response variable, and (ii) the logit main effect of one of two predictors on a response variable, under the assumption that the logit interaction effect is zero. The results indicate a clear superiority of the posterior mode to the posterior mean. We argue that the most reasonable priors are the Jeffreys' and a prior introduced by Clogg and Eliason (1987).

INTRODUCTION

When the sample size is small in comparison with the number of cells in the contingency table, there may be a number of cells that contain few or no observations. In such sparse tables, standard statistical procedures based on large-sample assumptions do not work as well as we would like. Maximum likelihood (ML) estimates of certain log-linear parameters may

not exist or may be on the boundary of the parameters space. Clogg, Rubin, Schenker, Schultz, and Widman (1991) report the difficulties with zero cells when standard log-linear analysis software is used.

Adding a small constant, generally 0.5, to every cell of the observed table has been a common recommendation in some standard references; for example, Goodman (1970) recommended this practice for saturated log-linear models. Adding 0.5 yields good results in terms of bias reduction for log-linear parameter estimates under the saturated log-linear model. Because of this, it has also become the default option in the log-linear analysis routine of SPSS 10.0 for saturated models.

Usually, we want to have confidence intervals as well as point estimates for the unknown parameters. In interval estimation, it is common to assume that ML estimates are approximately normally distributed and to apply the delta method to derive the standard errors. However, the delta method is based on the asymptotic properties of the ML estimates and works poorly for small samples (for example, see Agresti 2002). In contingency tables with empty cells, adding a constant has become a common way to improve the performance of confidence intervals. For example, Agresti (2002) proposed adding a constant that smooths toward the model of independence to construct logit confidence intervals for odds ratios. Chosen for its simplicity and good performance, this method has been used successfully in investigating a binomial proportion (see also, Brown, Cai, and DasGupta 2001, and Agresti and Coull 1998).

From a Bayesian point of view, adding 0.5 to each cell entry is equivalent to using a Dirichlet prior for the cell probabilities with all parameters equal to 1.5 (see, for example, Gelman, Carlin, Stern, and Rubin 1995, pp. 398-399). This is, however, just one of the many possible ways of introducing prior information on the parameter values. Another option is to use a different type of Dirichlet distribution, smoothing the parameter to a specific model (see Bishop, Fienberg, and Holland 1975, and Clogg, Rubin, Schenker, Schultz, and Widman, 1991). It is also possible to work with priors that have

different distributional forms than Dirichlet. Two such priors, which have become popular in logit modeling, are normal priors (Congdon 2001, Koop and Poirier 1995, and Weiss, Berk, Li, and Farrell-Ross 1999) and the Jeffreys' prior (Ibrahim and Laud 1991). For instance, many of the log-linear and logit modelling examples from the BUGS computer program manual (Gilks et al. 1994) make use of normal priors, and Congdon (2001) also suggests using normal priors with mean zero and large variance when estimating binomial logit regression coefficients in the absence of prior information.

Since Bayesian methods are often used in applied research, more research should be done to investigate whether Bayesian estimates have better properties than ML estimates, and whether some prior distributions produce better estimates than others. In the present work, we deal with the problem of parameter estimation in sparse tables using a Bayesian approach. In a $2 \times 2 \times 2$ contingency table, the estimation of two parameters was examined. First, we explored the interaction parameter of a saturated logit model. Second, we examined an effect parameter of a no-interaction logit model. We computed two commonly used Bayesian point estimators – posterior mode or modal a-posteriori (MAP) and posterior mean or expected a-posteriori (EAP) – and their confidence intervals under several prior distributions. A simulation experiment was performed to determine which Bayesian estimation method produces the best estimates. The quality of the point estimates was measured by the medians and the median squared errors, and the quality of the interval estimates was determined by the coverage probabilities and the median widths of the confidence intervals.

The remainder of this paper is divided into four sections. The first section illustrates the two parameters investigated by mean of examples that illustrate the zero cells problems treated in each case. In the second section, the Bayesian estimation methods used are described. Next, the results of the simulation study are presented and discussed. The paper ends with some conclusions and some recommendations.

TWO EXAMPLES

A three-way contingency table used by Agresti (2002, Table 2.6) in the textbook "Categorical Data Analysis" to explain certain concepts of the analysis of contingency tables is presented in Table 1. The example deals with the effect of the racial characteristics of defendants and victims on whether individuals convicted of homicide receive the death penalty. The variables in Table 1 are "death penalty verdict," with the categories (yes=1, no=2), and "defendant's race" (X_1) and "victim's race" (X_2), with the categories (white=1,black=2).

[Table 1 about here]

Suppose we would like to test the hypothesis as to whether the effect of the defendant's race depends on the victim's race regarding the giving of the death penalty. This implies that we have to estimate a saturated logit model of the form

$$\log\left(\frac{\pi_{1|jk}}{\pi_{2|jk}}\right) = \alpha + \beta_j^{X_1} + \beta_k^{X_2} + \beta_{jk}^{X_1X_2}. \tag{1}$$

Here, j and k denote categories of X_1 and X_2, respectively, $\pi_{i|jk}$ represents the probability of giving "response" i on the dependent variable given predictor values j and k, α is the model constant, and the β terms are the logit effect parameters. When effect coding is used, the interaction term $\beta_{jk}^{X_1X_2}$ is directly related to the three-variable-log-odds ratio by

$$\beta_{11}^{X_1X_2} = \frac{1}{4}(\log(or_1) - \log(or_2)),$$

where $\log(or_1)$ represents the effect of the defendant's race on death penalty for white victims, and is formulated as

$$\log(or_1) = \log\left(\frac{\pi_{1|11}}{\pi_{2|11}}\right) - \log\left(\frac{\pi_{1|21}}{\pi_{2|21}}\right), \tag{2}$$

and $\log(or_2)$ the same effect for black defendants,

$$\log(or_2) = \log\left(\frac{\pi_{1|12}}{\pi_{2|12}}\right) - \log\left(\frac{\pi_{1|22}}{\pi_{2|22}}\right). \tag{3}$$

The ML estimate $\widehat{\beta}_{jk}^{X_1X_2}$ is obtained by replacing the expected probabilities $\pi_{i|jk}$ by the corresponding observed probabilities $p_{i|jk}$. The confidence interval for the estimated logit interaction parameter $\widehat{\beta}_{jk}^{X_1X_2}$ is calculated by the formula

$$[\widehat{\beta}_{jk}^{X_1X_2} - z_{\alpha/2}\widehat{\sigma}(\widehat{\beta}_{jk}^{X_1X_2}), \widehat{\beta}_{jk}^{X_1X_2} + z_{\alpha/2}\widehat{\sigma}(\widehat{\beta}_{jk}^{X_1X_2})], \tag{4}$$

which is based on the asymptotic normality of $\widehat{\beta}_{jk}^{X_1X_2}$. The estimated asymptotic standard error $\widehat{\sigma}(\widehat{\beta}_{jk}^{X_1X_2})$ equals the square root of the diagonal elements of the Hessian matrix (for computational details, see the section on estimation methods and algorithms below).

In Table 1, the $(1, 1, 2)$ observed frequency is equal to zero. This implies that the ML estimates of the logit parameters do not exist because one of the sufficient statistics is zero. In addition, the confidence interval (4) is not defined. Agresti (2002, pp. 397-398) proposes to add 0.5 to each cell before computing the parameters of the saturated model and their standard errors.

[Table 2 about here]

Point and interval estimates for the logit interaction parameter obtained by estimating the saturated model with and without adding the constant 0.5 to each cell are presented in Table 2. We only consider $\widehat{\beta}_{11}^{X_1X_2}$ because the rest of the interaction parameters can be obtained from this one. As can be seen in Table 2, if we do not add the constant, $\widehat{\beta}_{11}^{X_1X_2}$ and $\widehat{\sigma}(\widehat{\beta}_{11}^{X_1X_2})$ are infinity so that the lower bound of the confidence interval cannot be determined using (4). On the other hand, if we add the constant 0.5, $\widehat{\beta}_{11}^{X_1X_2}$ and its confidence interval can be computed.

[Table 3 about here]

In Table 3, the hypothetical contingency table constructed by Clogg et al. (1991) to illustrate another zero-cells problem is presented. As can be seen, the table contains two sampling zeros. The purpose of this example is to

predict a dichotomous outcome variable Y using two dichotomous predictors, X_1 and X_2. We are interested in the logit main effect parameters of the model without a three-variable-interaction effect,

$$\log\left(\frac{\pi_{1|jk}}{\pi_{2|jk}}\right) = \alpha + \beta_j^{X_1} + \beta_k^{X_2}. \tag{5}$$

When effect coding is used for the predictors, the logit main effect parameter $\beta_j^{X_1}$ equals one half times the conditional log-odds ratios given in equations (2) and (3):

$$\beta_1^{X_1} = \frac{1}{2}\log(or_1) = \frac{1}{2}\log(or_2).$$

Similar expressions could be given for $\beta_j^{X_2}$.

Though all the two-way marginal totals are greater than zero, ML estimates of the logit parameters do not exist. In this case, the ML estimates of the probabilities reproduce the observed frequencies, and, therefore, two estimated frequencies equal zero. For this reason, the logit main effect parameters are plus or minus infinity. For more details on the existence of ML estimates see Haberman (1973).

[Table 4 about here]

Since the same result can be observed in both main effect parameters, it suffices to focus on $\widehat{\beta}_1^{X_1}$. In Table 4, we see the point estimates and confidence intervals for $\widehat{\beta}_1^{X_1}$ computed without and with adding 0.5 to each cell. If we do not add the constant, the logit parameter $\beta_1^{X_1}$ is minus infinity, and its standard error is infinity. As a consequence, the lower bound of the confidence interval is minus infinity and the upper bound is not defined. A problem here is that, though the parameter can be estimated if we add 0.5 to each cell, there is no theoretical justification for adding 0.5 in this case because the model is not saturated.

The examples described above represent the two parameters that were investigated in detail, and that are described in the sequel:

- CASE 1 refers to the estimation of the logit interaction parameter of a saturated model, $\beta_{11}^{X_1X_2}$, which has been examined in the first example. Here, the parameter cannot be estimated as a result of the fact that ML estimates do not exist when at least one sufficient statistic equals zero.

- CASE 2 refers to the estimation of the logit main effect parameter $\beta_1^{X_1}$ under the no three-variable-interaction model. Here, ML estimates of the logit effect parameters do not exist even though all two-way marginal totals are larger than zero.

The Bayesian approach described in the next section may resolve the problems associated with these two cases by introducing a certain amount of prior information on the parameters.

BAYESIAN ESTIMATION

Let $\boldsymbol{\beta}$ be the vector of unknown parameters and $\mathbf{y}$ the observed data. The most important difference between classical and Bayesian approaches is that, while the former assumes that parameters have unknown values that have to be estimated, the Bayesian approach treats unknown parameters as random variables. The posterior distribution $p(\boldsymbol{\beta}|\mathbf{y})$ is obtained by combining the likelihood function $p(\mathbf{y}|\boldsymbol{\beta})$ with a prior distribution, $p(\boldsymbol{\beta})$, and subsequently applying the Bayes rule,

$$p(\boldsymbol{\beta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\beta})\, p(\boldsymbol{\beta})}{\int p(\mathbf{y}|\boldsymbol{\beta})\, p(\boldsymbol{\beta}) d\boldsymbol{\beta}} \propto p(\mathbf{y}|\boldsymbol{\beta})\, p(\boldsymbol{\beta}),$$

where "$\propto$" stands for "is proportional to". As is explained in more detail below, different types of point estimators can be constructed using the posterior distribution function, two of which are the posterior mode, which represents the maximum of the posterior distribution, and the posterior mean.

The likelihood function we worked with is a (product) multinomial density function; that is,

$$p(\mathbf{y}|\boldsymbol{\beta}) \propto \prod_{p=1}^{P} \prod_{i=1}^{I} \pi_{i|p}^{\ n_{ip}}. \tag{6}$$

Here, n_{ip} denotes the observed number of cases with covariate pattern p that gives response i to the dependent variable. The number of covariate patterns and the number of possible responses are denoted P and I, respectively. The model probabilities that are functions of the unknown parameters $\boldsymbol{\beta}$ are denoted by $\pi_{i|p}$. Below, we will use N_p to denote the total number of cases with covariate pattern p; that is, $N_p = \sum_{i=1}^{I} n_{ip}$.

Three types of priors for Bayesian estimation of logit models were investigated here: natural conjugate priors, normal priors, and the Jeffreys' prior. It is typical of a Dirichlet prior, which is the conjugate prior of the multinomial likelihood, as well as of a normal prior that one has to define the values of one or more (hyper) parameters. In contrast, given the form of the likelihood, there is only one Jeffreys' prior because this is calculated using a standard formula.

JEFFREYS' PRIOR

A commonly used prior in Bayesian analysis is Jeffreys' prior (Jeffreys 1961). This prior is obtained by applying Jeffreys' rule, which means taking the prior density to be proportional to the square root of the determinant of the Fisher information matrix; that is,

$$p(\boldsymbol{\beta}) \propto |I(\boldsymbol{\beta})|^{\frac{1}{2}} .$$

Here, $|\cdot|$ denotes the determinant and $I(\boldsymbol{\beta})$ is the Fisher information matrix, which equals the expected value of the second derivatives of the log-likelihood function; that is, $I(\boldsymbol{\beta}) = -E\left(\frac{\partial^2 \ln p(\mathbf{y}|\boldsymbol{\beta})}{(\partial \boldsymbol{\beta})^2}\right)$. When the second derivatives matrix does not depend on the data, as with the multinomial distribution, the information matrix simplifies to $I(\boldsymbol{\beta}) = -\frac{\partial^2 \ln p(\mathbf{y}|\ \boldsymbol{\beta})}{(\partial \boldsymbol{\beta})^2}$. An important property of Jeffreys' prior is its invariance under scale transformations of the parameters. This means, for example, that it does not make a difference whether the prior is specified for the log-linear or the multiplicative parameters of the logit model, or whether we use dummy or effect coding.

We applied Jeffreys' prior in the two cases described above. CASE 1 is a special situation because the Jeffreys' prior has a very simple form in

saturated models, i.e.,

$$p(\boldsymbol{\beta}) \propto \sum_{p=1}^{P} \prod_{i=1}^{I} {\pi_{i|p}}^{0.5},$$

which yields a posterior that amounts to using $n_{ip} + 0.5$ as data. In other words, in saturated models, using a Jeffreys' prior for the log-linear parameter means adding 0.5 to each cell entry.

However, in non-saturated models (CASE 2), the Jeffreys' prior is computationally more complicated. Let L denote the total number of parameters, β_ℓ a particular parameter, and $x_{ip\ell}$ an element of the design matrix. The elements of first column of the design matrix (x_{ip1}) will usually be equal to one to obtain an intercept. For a logit model of the form

$$\log\left(\frac{\pi_{1|p}}{\pi_{2|p}}\right) = \sum_{\ell=1}^{L} \beta_\ell \cdot x_{ip\ell},$$

and the (product) multinomial likelihood defined in Equation (6), element (ℓ, m) of the information matrix is obtained as follows:

$$I_{\ell m}(\beta) = \sum_{p=1}^{P} \sum_{i=1}^{I} N_p \pi_{i|p} \left(x_{ip\ell} - \overline{x}_{p\ell}\right) \cdot \left(x_{ipm} - \overline{x}_{pm}\right),$$

with $\overline{x}_{p\ell} = \sum_{i=1}^{I} \pi_{i|p} x_{ip\ell}$. Ibrahim and Laud (1991) give a general theoretical justification for using Jeffreys' prior with exponential family distributions by showing that proper posterior distributions are obtained.

UNIVARIATE NORMAL PRIOR

It is also possible to work with other types of prior distribution for the log-linear parameters. Assuming that no information about the dependence between parameters is available, it is convenient to adopt a set of univariate normal priors. For instance, Congdon (2001) suggested that, in absence of prior expectation about the direction or size of covariate effects, flat priors may be approximated in BUGS by taking univariate normal distributions with mean zero and large variance.

The effect of using normal priors with means of 0 is that parameter estimates are smoothed towards zero. However, since the variance determines

the amount of prior information that is added, we can decrease the smoothing effect by increasing the variances.

DIRICHLET PRIOR

In contrast to the normal prior presented above, which are based on the logit parameters ($\boldsymbol{\beta}$), the Dirichlet prior is based on the model probabilities ($\boldsymbol{\pi}$). As the conjugate prior of the multinomial distribution, the Dirichlet prior belongs to the family of functions whose densities have the same functional form as the likelihood (Schafer 1997, p.306; Gelman et al. 1995, p.76). The Dirichlet distribution is defined as

$$p(\boldsymbol{\pi}) \propto \prod_{p=1}^{P} \prod_{i=1}^{I} \pi_{i|p}^{(\alpha_{ip}-1)}, \tag{7}$$

where the α_{ip} terms are the (hyper) parameters of the prior. In the saturated model (CASE 1), the posterior distribution is a Dirichlet distribution with parameters $n_{ip}+\alpha_{ip}-1$. In CASE 2, however, the probabilities are restricted functions of the β parameters. Schafer (1997, p.306) referred to a prior of this form as a constrained Dirichlet prior. Gelman et al. (1995, pp.398-399) also used such a prior in the Bayesian estimation of log-linear models.

When using a Dirichlet prior, one has to specify the α_{ik} parameters. If there is no information on the values of $\boldsymbol{\beta}$, it is a common practice to take a common value for the α_{ik} parameters. Using a common value larger than 1 has the effect that the estimated probabilities are smoothed towards a table in which all cell probabilities are equal. Schafer (1997, p.253) called such a constant a flattening prior. Note that adding 0.5 to each cell amounts to setting $\alpha_{ik} = 1.5$.

It is not always desirable to smooth the data towards the equal probability model. It is, however, also possible to work with cell specific α_{ik} parameters that are in agreement with a particular log-linear model. Bishop, Fienberg, and Holland (1975) proposed a prior, called pseudo-Bayes prior, in which α_{il} parameters smooth the data toward the model of independence.

For logit models, Clogg and Eliason (1987) and Clogg et al. (1991) proposed using a Dirichlet prior that, on the one hand, preserves the marginal distribution of the dependent variable and, on the other hand, takes into account the number of parameters to be estimated. It is obtained as follows:

$$\alpha_{ip} = 1 + \left(\frac{\sum_{p=1}^{P} n_{ip}}{\sum_{p=1}^{P} N_p} \right) \left(\frac{L}{P} \right)$$

Here, L denotes the number of unknown logit parameters. Note that the value of α_{ip} does not depend on p. We will refer to this prior as the Clogg-Eliason (C-E) prior.

ESTIMATION METHODS AND ALGORITHMS

Various types of point estimators for the unknown parameters can be used within a Bayesian context. Three of them are posterior mode, posterior mean, and posterior median estimates. In the simulation study reported in the next section, we worked with posterior mode and posterior mean estimators, which are the most commonly used in practice.

Posterior mode estimation of logit coefficients is similar to applying maximum likelihood estimation, assuming that the posterior distribution has a unique mode. If this is not the case, the solution corresponding to the global maximum of the posterior function is taken. This estimator is called maximum a-posteriori (MAP). It should be noted that, with Dirichlet priors, standard algorithms for ML estimation, such as iterative proportional fitting (IPF) and Newton-Raphson (NR) can be used to obtain MAP estimates (Gelman et al. 1995, pp.399-400; Schafer 1997, pp.307-308). However, when a normal prior or the Jeffreys' priors for non-saturated models are used, the posterior distribution does not have an analytically tractable form. For these cases, we implemented a modified NR algorithm to obtain MAP estimates. The algorithm uses numerical derivatives instead of analytical ones (see, Gelman et al. 1995, p. 273), and it was applied on the log-posterior density,

$$L(\boldsymbol{\beta}) = \log(p(\boldsymbol{\beta}|\mathbf{y})) \propto \log(p(\mathbf{y}|\boldsymbol{\beta})) + \log(p(\boldsymbol{\beta})), \tag{8}$$

which combined the log-likelihood function $\log(p(\mathbf{y}|\boldsymbol{\beta}))$ with the logarithm of the prior distribution $\log(p(\boldsymbol{\beta}))$. The Newton-Raphson algorithm proceeds as follows:

1. Choose a set of starting values $\boldsymbol{\beta}^{(0)}$.
2. For each iteration, $s = 1, 2, 3, ...$
   - Compute the vector of first derivatives and the matrix of second derivatives with respect to $\boldsymbol{\beta}$, denoted by L' and L'', evaluated at the parameter values $\boldsymbol{\beta}^{(s-1)}$.
   - Calculate the new $\boldsymbol{\beta}^{(s)}$ by

$$\boldsymbol{\beta}^{(s)} = \boldsymbol{\beta}^{(s-1)} - [L''(\boldsymbol{\beta}^{(s-1)})]^{-1} L'(\boldsymbol{\beta}^{(s-1)}).$$

   - Compute the value of the posterior distribution using the new $\boldsymbol{\beta}^{(s)}$.
   - Stop the iterations if the increase of $L(\boldsymbol{\beta})$ between subsequent iterations is smaller than 10^{-8}.

The expected a-posteriori (EAP) or posterior mean estimator is defined as follows:

$$E(\boldsymbol{\beta}|\mathbf{y}) = \int \boldsymbol{\beta}\, p(\boldsymbol{\beta}|\mathbf{y})\, d\boldsymbol{\beta},$$

A problem in the computation of EAP estimates is that there is no analytical solution for the integral at the right-hand side of the above equation. Markov chain Monte Carlo (MCMC) methods can, however, be used to obtain samples from the posterior distribution $p(\boldsymbol{\beta}|\mathbf{y})$ (Gelman et. al. 1995, Chapter 11). Suppose we sampled T sets of parameters, where $\boldsymbol{\beta}^t$ denotes one of these sets. The Monte Carlo approximation of the posterior expectation is

$$E(\boldsymbol{\beta}|\mathbf{y}) \approx \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{\beta}^t.$$

Gelman et al. (1995, pp.400-403) and Schafer (1997, pp.308-320) showed that, with Dirichlet priors, it is possible to adapt the IPF algorithm to obtain posterior mean estimates. This MCMC variant of IPF is called Bayesian IPF. With other priors, no such simple algorithm is available.

We drew samples from the posterior distribution using a random-walk Metropolis algorithm with a univariate normal jumping distribution for each parameter. For each logit parameter β_ℓ, at iteration s, one samples a value β_ℓ^* from a univariate normal distribution with a mean equal to the current value β_ℓ^{s-1} and a variance equal to σ_ℓ^2; that is, $\beta_\ell^* \tilde{} N(\beta_\ell^{s-1}, \sigma_\ell^2)$. The new set of parameters $\boldsymbol{\beta}^*$ that is obtained in this way is accepted with probability

$$r = \min\left(1, \frac{p(\boldsymbol{\beta}^*|\mathbf{y})}{p(\boldsymbol{\beta}^{s-1}|\mathbf{y})}\right);$$

that is, $\boldsymbol{\beta}^s = \boldsymbol{\beta}^*$ with probability r and otherwise $\boldsymbol{\beta}^s = \boldsymbol{\beta}^{s-1}$. In other words, if the posterior associated with $\boldsymbol{\beta}^*$ is larger than the one with $\boldsymbol{\beta}^{s-1}$, we take the new values $\boldsymbol{\beta}^*$, and otherwise we take the new values with a probability equal to the ratio of the "new" and current posterior.

The exact implementation of our Metropolis algorithm is:

1. We retained each tenth sample for the computation of posterior means and posterior standard errors.

2. The iterations started with 1000 burning in samples, with σ_ℓ^2 being the inverse of the square of the number of parameters. Then, we performed another 1000 burning in iterations, with σ_ℓ^2 equated to the estimated variance from the first samples divided by the square of the number of parameters. The σ_ℓ^2 for the subsequent iterations was equated to the estimated variance from the second set of burning in samples divided by the square of the number of parameters. This method yielded acceptance rates of around 0.5 for all situations that we investigated.

3. The convergence of the algorithm was determined using the $\sqrt{\widehat{R}}$ criterion described in Gelman et al. (1995, Section 11.4). For this purpose,

three independent parallel sequences were generated. Convergence was reached when the $\sqrt{\widehat{R}}$ values was smaller than 1.001 for each parameter, which is a extremely precise converge criterion. This convergence was checked at each 25,000th iteration. The maximum number of iterations was set equal to 1,000,000.

In order to obtain interval estimators, it is assumed that the marginal posterior distribution of the parameters is approximately a normal distribution, and confidence intervals are computed following equation (4). The standard error of the posterior mode is the square root of the diagonal elements of the second derivatives matrix of the posterior distribution, and the standard errors of the posterior mean is the square root of the variance of the samples retained in the Metropolis algorithm (see Step 1 of the Metropolis algorithm.)

SIMULATION STUDY

In this section, we present the results of the two simulation experiments we conducted to evaluate the performance of point estimators and confidence intervals based on different prior distributions.

In CASE 1, investigating the logit interaction parameter of a saturated model, we generated data from a multinomial distribution whose parameters satisfied a logit model where the effect parameters $\beta_1^{X_1}$ and $\beta_1^{X_2}$ were fixed at zero, and the logit interaction parameter took the values 0, representing the uniform model in which all the probabilities are equal, 1, representing an intermediate interaction between the predictors, and 2, representing a strong interaction between the predictors. In CASE 2, investigating a logit main effect parameter of a no-interaction model, we generated data from a multinomial distribution whose parameters satisfied a logit model with effect parameters equal to each other and taking values of 0, representing the uniform model, 1, representing intermediate effects, 2, representing strong effects. In each case, we varied the sample size by taking samples of 20 and 100 units. 5000 samples were drawn from each condition. The data were

simulated using Vermunt's (1997) program LEM. The input files used can be found in the Appendix.

MAP and EAP estimates were obtained using the Newton-Raphson algorithm and the Metropolis algorithm described in the previous section. The priors used were the Jeffreys' prior, three types of Dirichlet prior with constant α_{ik} parameters, the prior defined by Clogg and Eliason (1987; C-E), and three types of normal prior. For Jeffreys' and the C-E prior, no additional parameters needed to be specified. The parameters of the three Dirichlet distributions were: $\alpha_{ik} = 1.5$ [Dir(1.5)], $\alpha_{ik} = 1.333$ [Dir(1.333)], and $\alpha_{ik} = 1.1$ [Dir(1.1)], where the first represents the standard practice of adding 0.5 to each cell entry, and the second and third are examples of situations in which a somewhat smaller number is added in order to make the prior less informative. The three types of normal priors were: N(0,4), N(0,10), and N(0,25). In order to approximate the ML estimates, we used MAP under a Dirichlet prior distribution with $\alpha_{ik} = 1.001$. Using this prior distribution, we prevented numerical problems in the estimation, especially with the small sample of size 20.

In order to summarize the results obtained, we report the median of the MAP and EAP estimates, and the median of the root of the squared errors (RMdSE), that is, the square root of the median of $(\widehat{\beta} - \beta)^2$. Though mean squared errors are the most common statistics used to measure the quality of the point estimates, we used the median squared error instead to avoid the effect that extreme values have on the mean. For the confidence intervals, we report the coverage probabilities, which represent the proportion of times that the simulated intervals contain the population parameter, and the median widths of the intervals.

By definition, a 95% confidence interval should have a coverage probability of at least 0.95. However, even if the true coverage probability equals 95%, the coverage probabilities coming from the simulation experiment will not be exactly equal to 0.95 because of the Monte Carlo error. This error tends to zero when the number of replications tends to infinity. Since we

worked with 5000 replications, the Monte Carlo standard error was equal to $\sqrt{\frac{0.95 \cdot 0.05}{5000}} = 0.0031$, which means that coverage probabilities between 0.946 and 0.9531 are in agreement with the nominal level of 95%.

[Table 5 and 6 about here]

Tables 5 and 6 summarize the results for CASE 1 when the sample sizes are $N = 20$ and $N = 100$, respectively. It should noted that the results from the Jeffrey's prior were omitted because, in this case, they were equal to the results from the Dir(1.5) prior. From Tables 5 and 6, we can see that, when $\beta_{11}^{X_1X_2}$ equals zero, there are not many differences between the results obtained under different prior distributions for both point estimates. However, the differences increase with higher values of $\beta_{11}^{X_1X_2}$. If we compare MAP and EAP, we can see that the values of the EAP are always more extreme than the values of the MAP. Also, in terms of RMdSE, it can be seen that MAP gives better results than EAP. If we compare the various prior distribution, we see that C-E, Dir(1.5), and, therefore, also Jeffreys' produce the smallest RMdSE for $\beta_{11}^{X_1X_2}$ equal to zero or one. However, if $\beta_{11}^{X_1X_2}$ equals two, the medians are considerably smaller than the population values. In that case, Dir(1.1) and N(0,4) give more accurate results. For the interval estimators, the conclusions are similar: again C-E, Dir(1.5), and Jeffreys' priors show the smallest median widths. As far as the coverage probabilities is concerned, only the intervals for the EAP under normal priors present coverage probabilities below the 95% nominal level.

[Table 7 and 8 about here]

Tables 7 and 8 summarize the results for CASE 2 when the sample sizes are $N = 20$ and $N = 100$, respectively. In terms of point estimators, the results are similar to the results obtained in CASE 1. Again, the smoothing effect of C-E and Dir(1.5) priors seems to be too extreme when the parameter value is high ($\beta_1^{X_1} = 2$). The medians are smaller than the population value, and, for $N = 20$, the RMdSE of the Dir(1.5) is higher than the RMdSE

obtained under the other prior distributions. Also in terms of coverage probabilities, C-E prior and Dir(1.5) yield values lower than the nominal level. The Jeffreys' prior is a better option in CASE 2 because it produces a lower RMdSE and a smaller median width along all the degrees of association. The smoothing effect of Dir(1.1) and the normal priors is small, but the confidence intervals tend to be huge. This means that the population parameter is included in the interval only because of the extreme width, but not because of the accuracy of the estimates.

CONCLUSIONS AND RECOMMENDATIONS

Bayesian estimation of two logit parameters in a $2 \times 2 \times 2$ table has been investigated. Firstly, estimation of $\beta_{11}^{X_1X_2}$, which measures the logit interaction effect of X_1 and X_2 on a response variable Y under the saturated model. Secondly, estimation of $\beta_1^{X_1}$, which measures the logit main effect of X_1 on the response variable Y, under the assumption that $\beta_{11}^{X_1X_2} = 0$. The performance of both point estimates and confidence intervals has been evaluated. A good point estimator has small bias (defined here as the deviation of the median estimate from the true value) and small residual median square error (RMdSE). A good confidence interval has small median width under the condition that its coverage probability is at least 0.95.

Using these criteria, the simulation results can be summarized as follows:

1. All of the prior distributions studied yield better point estimates and confidence intervals than maximum likelihood.

2. In almost all cases, the bias and RMdSE of the MAP estimates are smaller than those of the EAP estimates. In all cases, the median width of the MAP confidence intervals is smaller than the median width of the EAP confidence intervals. Furthermore, in several cases with normal priors, coverage probabilities are unacceptably low for the EAP confidence intervals.

3. Among the three normal priors studied, the one with variance 4, N$(0, 4)$, performs best and the one with variance 25, N$(0, 25)$, performs worst.

4. Among the three Dirichlet priors studied, the one with parameter 1.33, Dir(1.33), appears to be the most reasonable; a Dir(1.5) seems to oversmooth the data, and a Dir(1.1) does not sufficiently smooth them. Moreover, the Dir(1.5) gives unacceptably low coverage probabilities for the confidence interval for $\beta_1^{X_1}$ when its true value equals 2.

Concluding, among the procedures studied, the most reasonable ones seem to be MAP estimation with a Jeffreys', C-E, Dir(1.33), or N(0,4) prior. EAP estimation, which is commonly recommended in textbooks, appears to perform badly under the criteria we have used. Congdon's recommendation to use a normal distribution with large variance as a "noninformative"prior should not be followed when the sample size is small. The parameters of the Dir(1.33) and N$(0, 4)$ have no particular theoretical justification, and therefore it is not clear how they might perform in other (logit or loglinear) estimation problems. Because the Jeffreys' and C-E priors do have a good theoretical justification, and because they perform reasonably well in the present simulations, these may be the most recommendable in general settings.

A program to do the estimation studied in this paper is available from the first author. Estimation with Dirichlet priors can be done in standard statistical software packages by adding a constant to the observed frequencies and then doing ML estimation. Estimation with normal priors can be done with the BUGS program. Unfortunately, no standard software is available yet for estimation with the recommended Jeffreys or C-E priors.

APPENDIX

The LEM input file used to simulate the data used in investigating the three-variable interaction parameter was

```
man 3
dim 2 2 2
```

```
lab A B C
mod ABC cov(ABC,1)
des [1 0 -1 0 -1 0 1 0]
sim 20 baysim.dat
sim 100 baysim.dat

sta log cov(ABC) [0]
sta log cov(ABC)[1]
sta log cov(ABC) [2]
```

The LEM input file used to simulate the data used in investigating the logit parameter was:

```
man 3
dim 2 2 2
lab A B C
mod ABC cov(AC,1),cov(BC,1)
des [1 0 -1 0 1 0 -1 0]
sim 20 baysim.dat
sim 100 baysim.dat

sta log cov(AC) [0]
sta log cov(BC) [0]
sta log cov(AC) [1]
sta log cov(BC) [1]
sta log cov(AC) [2]
sta log cov(BC) [2]
```

The asterisk means that the program does not read what is written behind.

REFERENCES

Agresti, Alan. 2002. Categorical data analysis. New York: Wiley.

Agresti, Alan. 1999. "On Logit Confidence Intervals for the Odds Ratio with Small Samples." Biometrics 55:597-602.

Agresti, Alan and Brent A. Coull. 1998. "Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions." The American Statistician 52:119-126.

Anscombe, F.J. 1956. "On estimating binomial response relations." Biometrika 43:461-464.

Bishop, Yvonne M.M., Stephen E. Fienberg and Paul W. Holland. 1975. Discrete Multivariate Analysis Cambridge.Massachusetts:MIT Press.

Brown, Lawrence D., Tony T. Cai and Anirban DasGupta. 2001. "Interval Estimation for a Binomial Proportion." Statistical Science 16:101-133.

Clogg, Clifford C. and Scott R. Eliason. 1987. "Some common problems in log-linear analysis." Sociological Methods and Research 16:8-44.

Clogg, Clifford C., Donald B. Rubin, Nathaniel Schenker, Bradley Schultz and Lynn Widman. 1991. "Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression." Journal of the American Statistical Association 86:68-78.

Congdon, Peter. 2001. Bayesian Statistical Modelling Chichester:Wiley.

Gart, J.J. and J.R. Zweifel. 1967. "On the bias of various estimators of the logit and its variance with application to quantal bioassay." Biometrika 54:181-187.

Gelman, Andrew, John B. Carlin, Hál S. Stern and Donald B. Rubin. 1995. Bayesian data analysis. London: Chapman and Hall.

Gilks, Wally R., Andrew Thomas and David Spiegelhalter. 1994. "A language and program for complex Bayesian modeling." The Statistician 43:169-177.

Goodman, Leo A. 1970. "The multivariate analysis of qualitative data: Interactions among multiple classifications." Journal of the American Statistical Association 65: 225-256.

Haberman, Shelby J. 1973. "Sufficient statistics and likelihood equations." Annals of Statistics 1:617-632.

Ibrahim, Joseph G. and Purushottam W. Laud. 1991. "On Bayesian analysis of general linear models using Jeffreys'prior." Journal of the American Statistical Association 86:981-986.

Jeffreys, H. 1961. Theory of probability, (3rd ed.) Data. Oxford: Oxford University Press.

Kass, Robert E. and Larry Wasserman. 1996. "Formal Rules for Selecting Prior Distributions." Journal of the American Statistical Association 91: 1343-1370.

Koop, Gary and Dale J. Poirier. 1993. "Bayesian analysis of logit models using natural conjugate priors." Journal of Econometrics 56:323-340.

Koop, Gary and Dale J. Poirier. 1995. "An Empirical Investigation of Wagner's Hypothesis by using a Model Occurrence Framework." Journal of the Royal Statistical Society A 58:123-141.

Price, Robert M. and Douglas G. Bonett. 2000. "Estimating the ratio of two Poisson rates." Computational Statistics & Data Analysis 34:345-356.

Poirier, Dale J. 1994. "Jeffreys'prior for logit models." Journal of Econometrics 63:327-339.

Schafer, Joseph L. 1997. Analysis of Incomplete Multivariate Data. London: Chapman and Hall.

Vermunt, Jeroen K. 1997. LEM: A general program for the analysis of categorical data. Tilburg, The Netherlands: Tilburg University.

Weiss, Robert, Richard Berk, Wenzhi Li and Margaret Farrell-Ross. 1999. "Death Penalty Charging in Los Angeles County: An Illustrative Data Analysis Using Skeptical Priors." Socilogical Methods and Research 28:91-115.

Zellner, Arnold and P.E. Rossi. 1984. "Bayesian analysis of dichotomous quantal response models." Journal of Econometrics 25:365-393.