The authors develop binomial-beta hierarchical models for ecological inference using insights from the literature on hierarchical models based on Markov chain Monte Carlo algorithms and King's ecological inference model. The new approach reveals some features of the data that King's approach does not, can be easily generalized to more complicated problems such as general $R \times C$ tables, allows the data analyst to adjust for covariates, and provides a formal evaluation of the significance of the covariates. It may also be better suited to cases in which the observed aggregate cells are estimated from very few observations or have some forms of measurement error. This article also provides an example of a hierarchical model in which the statistical idea of "borrowing strength" is used not merely to increase the efficiency of the estimates but to enable the data analyst to obtain estimates.

Binomial-Beta Hierarchical Models for Ecological Inference

GARY KING
Harvard University
ORI ROSEN
University of Pittsburgh
MARTIN A. TANNER
Northwestern University

1. INTRODUCTION

Ecological inference is the process of learning about discrete individual-level behavior by analyzing data on groups. In this article, we develop binomial-beta hierarchical models for this problem using insights from King's (1997) ecological inference model and the literature on hierarchical models based on Markov chain Monte Carlo (MCMC) algorithms (Tanner 1996). For many of the applications we have studied, our approach provides empirical results similar to King's. However, as illustrated in our first example, the present model can reveal some features of the data that King's model does not—at

AUTHORS' NOTE: *Our thanks go to the National Institutes of Health for Grant CA35464 to Tanner, and the National Science Foundation (SBR-9729884), the Centers for Disease Control and Prevention (Division of Diabetes Translation), the National Institutes of Aging, the World Health Organization, and the Global Forum for Health Research for research support to King.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 28 No. 1, August 1999 61-90

the price of increased computation. Because some individual-level information is lost in the aggregation process, any single approach to the ecological inference problem will by necessity require a set of modeling assumptions, and the success of the endeavor will depend on these assumptions. It is therefore of value to the data analyst to have a variety of models with which to explore the data. In one scenario, different models will yield qualitatively similar conclusions and the results will be robust to the different sets of assumptions. In another scenario, the models will yield different conclusions, prompting the data analyst to examine the impact of the various assumptions on these conclusions. Thus, the hierarchical models presented in this article provide helpful data analytic checks on King's model. In addition, this MCMC-based approach has several other advantages: It can be easily generalized to more complicated ecological inference problems such as $R \times C$ tables (see King, Rosen, and Tanner 1999), it enables the data analyst to adjust for a covariate and provides a formal evaluation of the significance of this covariate, and it is better suited to data in which the observed aggregate variables are estimated from very few observations or have some form of measurement error. This article also provides an example of a hierarchical model in which the statistical idea of "borrowing strength" is used not merely to increase the efficiency of the estimates but to enable the data analyst to obtain estimates.

We introduce the ecological inference problem and our notation in Section 2 and summarize King's model in Section3. Section 4 gives a brief introduction to the concept of hierarchical models. We then introduce our binomial-beta hierarchical model for the situation with no covariates in Section 5 and the corresponding model for the case with covariates in Section 6. All methods are illustrated with examples. Section 7 concludes by outlining future work in this field, some of which is currently under investigation.

2. THE PROBLEM

We introduce the ecological inference problem in this section with the notation and an example from King (1997, chap. 2). For expository purposes, we discuss only a special case of the problem and

TABLE 1: Notation for Precinct *i*

	Voting Decision		
Race of Voting Age Person	*Vote*	*No Vote*	
Black	β_i^b	$1-\beta_i^b$	X_i
White	β_i^w	$1-\beta_i^w$	$1-X_i$
	T_i	$1-T_i$	

NOTE: The goal is to estimate the quantities of interest, β_i^b (the fraction of blacks who vote) and β_i^w (the fraction of whites who vote), from the aggregate variables X_i (the fraction of voting age people who are black) and T_i (the fraction of people who vote), along with N_i (the known number of voting-age people).

save discussion of the more general case for the concluding section. The basic problem has two observed variables (T_i and X_i) and two unobserved quantities of interest (β_i^b and β_i^w) for each of p observations. Observations represent aggregate units, such as geographic areas, and the unobserved individual- level variables being aggregated are dichotomous.

To be more specific, in Table 1, we observe for each electoral precinct i ($i = 1, \ldots, p$) the fraction of voting-age people who turn out to vote (T_i) and who are black (X_i), along with the number of voting-age people (N_i). The quantities of interest, which remain unobserved because of the secret ballot, are the fractions of blacks who vote (β_i^b) and whites who vote (β_i^w). The proportions β_i^b and β_i^w are not observed because T_i and X_i are from different data sources (electoral results and census data, respectively), and so the cross tabulation cannot be computed.

3. A SUMMARY OF KING'S MODEL

The ecological inference literature before King (1997) was bifurcated between supporters of the method of bounds, originally proposed by Duncan and Davis (1953), and supporters of statistical approaches, proposed by Ogburn and Goltra (1919) but first formalized into a coherent statistical model by Goodman (1953, 1959).[1] Although these authors moved on to other interests following their seminal contributions, most of the ecological inference literature since 1953 has been an ongoing, and not always polite, war between these two key approaches.

The purpose of the method of bounds and its generalizations is to extract deterministic information about the problem. For example, if a precinct contained 150 African Americans and 87 people in the precinct voted, then the number of African American voters must lie between 0 and 87. The statistical approach examines variation in the marginals (X_i and T_i) over the precincts to attempt to reason back to the district-wide fractions of blacks and whites who vote (the average over i of β_i^b and of β_i^w weighted by the number of blacks and whites per precinct, respectively). The problem with the method of bounds approach used in isolation is that it yields only a range of possible answers. The problem with the statistical approach is that (as Goodman made clear) if the assumptions are wrong, the answers can be far off. For example, if T_i is low when X_i is high, one might infer that blacks vote less frequently than whites, but it could equally be true that whites who happen to live in heavily black precincts are those who vote less frequently, yielding the opposite ecological inference to the individual-level truth.

A key point of King's approach that we draw on is that the insights from these two literatures do not conflict with each other; the sources of information are largely distinct and can be combined to improve inference overall. Thus, we too combine the information from the bounds, applied to both quantities of interest for each and every precinct, with a statistical approach for extracting information within the bounds. The amount of information in the bounds depends on the data set, but for many data sets, it can be considerable. For example, if precincts are spread uniformly over a scatter plot of X_i by T_i, the average bounds on β_i^b and β_i^w are narrowed from [0,1] to less than half of that range (hence eliminating half of the problem with certainty). This additional information also helps make the statistical portion of the model far more robust than previous statistical methods, which exclude the bounds.

To illustrate these points, we first present all the information available without making any assumptions, thus extending the bounds approach as far as possible. As a starting point, the left graph in Figure 1 provides a scatter plot of a sample data set as observed, X_i horizontally by T_i vertically. Each point in this figure corresponds to one precinct, for which we would like to estimate the unknowns. We display the unknowns in the right-hand graph of the same figure; any point in

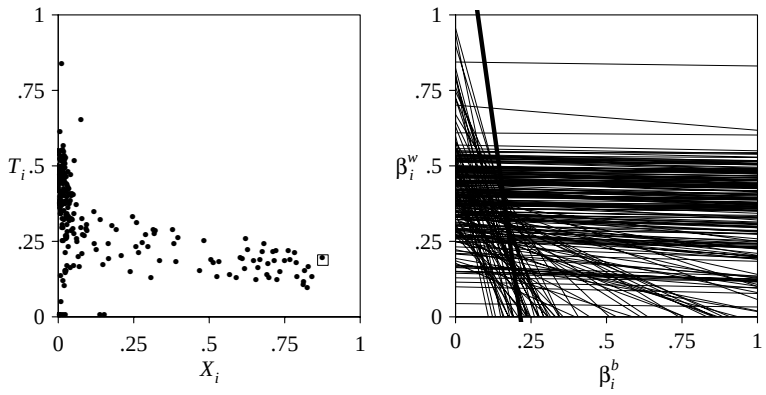

Figure 1: Two Views of the Same Data
NOTE: The left-hand graph is a scatter plot of the observables, X_i by T_i. The right-hand graph displays this same information as a tomography plot of the quantities of interest, β_i^b by β_i^w. Each precinct i that appears as a point in the left-hand graph is a line (rather than a point because of information lost due to aggregation) in the right-hand graph. For example, precinct 52 appears as the dot with a little square around it in the left-hand graph and the dark line in the right-hand graph. The data are from King (1997, Figs. 5.1, 5.5).

that graph portrays values of the two unknowns, β_i^b, which is plotted horizontally, and β_i^w, which is plotted vertically. Ecological inference involves locating, for each precinct, the one point in this unit square corresponding to the true values of β_i^b and β_i^w, since values outside the square are logically impossible.

To map the knowns onto the unknowns, we begin with this accounting identity:

$$T_i = X_i\beta_i^b + (1 - X_i)\beta_i^w. \tag{1}$$

This identity holds exactly; it is not a regression equation and has no error term. From this equation, we solve for one unknown in terms of the other:

$$\beta_i^w = \left(\frac{T_i}{1 - X_i}\right) - \left(\frac{X_i}{1 - X_i}\right)\beta_i^b. \tag{2}$$

Equation (2) shows that β_i^w is a *linear* function of β_i^b, where the intercept and slope are known (since they are functions of X_i and T_i).

We now map the knowns from the left-hand graph onto the right-hand graph by using the linear relationship in equation (2). A key point is that each dot on the left-hand graph can be expressed, without assumptions or loss of information, as a (tomography) line within the unit square in the right-hand graph.[2] It is precisely the information lost due to aggregation that causes us to have to plot an entire line (on which the true point must fall) rather than the goal of one point on the right-hand graph. In fact, the information lost can be thought of as equivalent to having a graph of the β_i^b by β_i^w points but having the ink smear, making the points into lines and partly obscuring the correct positions of the (β_i^b, β_i^w) points.

What does a tomography line tell us? Before we know anything, we know that the true (β_i^b, β_i^w) point must lie somewhere within the unit square. After X_i and T_i are observed for a precinct, we also know that the true point must fall on a specific line represented by equation (2) and appearing in the tomography plot in Figure 1. In many cases, narrowing the region to be searched for the true point from the entire square to one line in the square can provide a significant amount of information. To see this, consider the point enclosed in a box in the left-hand graph and the corresponding dark line in the right-hand graph. This precinct, number 52, has observed values of $X_{52} = 0.88$ and $T_{52} = 0.19$. As a result, substituting into equation (2) gives $\beta_i^w = 1.58 - 7.33\beta_i^b$, which, when plotted, appears as the dark line on the right-hand graph. This particular line tells us that in our search for the true $\beta_{52}^b, \beta_{52}^w$ point on the right-hand graph, we can eliminate with certainty all area in the unit square except that on the line, which is clearly an advance over not having the data. Translated into the quantities of interest, this line tells us (by projecting the line downward to the horizontal axis) that wherever the true point falls on the line, β_{52}^b must fall in the relatively narrow bounds of [0.07, 0.21]. Unfortunately, in this case, β_i^w can be bounded (by projecting to the left) only to somewhere within the entire unit interval. More generally, lines that are relatively steep like this one tell us a great deal about β_i^b and little about β_i^w. Tomography lines that are relatively flat give narrow bounds on β_i^w and wide bounds on β_i^b. Lines that cut off

the bottom left (or top right) of the figure give narrow bounds on both quantities of interest.

If the only information available to learn about the unknowns in precinct i is X_i and T_i, a tomography line in Figure 1 exhausts all this available information. This line immediately tells us the known bounds on each of the parameters, along with the precise relationship between the two unknowns, but it is not sufficient to narrow in on the correct answer any further. Fortunately, additional information exists in the other observations in the same data set (X_j and T_j for all $i \neq j$) that, under the right assumptions, can be used to learn more about β_i^b and β_i^w in our precinct of interest.

In order to borrow statistical strength from all the precincts to learn about β_i^b and β_i^w in precinct i, some assumptions are necessary. The simplest version (i.e., the one most useful for expository purposes) of King's model requires three assumptions, each of which can be relaxed in different ways. First, the set of (β_i^b, β_i^w) points must fall in a single cluster within the unit square. The cluster can fall anywhere within the square. The cluster can be widely or narrowly dispersed or highly variable in one unknown and narrow in the other, and the two unknowns can be positively, negatively, or not at all correlated over i. An example that would violate this assumption would be two or more distinct clusters of (β_i^b, β_i^w) points, as might result from subsets of observations with fundamentally different data generation processes (such as from markedly different regions). The specific mathematical version of this one-cluster assumption is that β_i^b and β_i^w follow a truncated bivariate normal distribution, although Monte Carlo experiments indicate that the main assumption here is that of a distribution with a single mode. The second assumption is the absence of spatial autocorrelation: Conditional on X_i, T_i and T_j are independent. The final assumption is that X_i is independent of β_i^b and β_i^w.

These three assumptions—one cluster, no spatial autocorrelation, and no correlation between the regressor and the unknowns—enable one to compute a posterior (or sampling) distribution of the two unknowns in each precinct. Extensive Monte Carlo evidence (King 1997) demonstrates that most features of the model are highly robust to violations of the first two assumptions. In cases where the bounds are sufficiently narrow for many of the precincts (an observation that

can be made from the aggregate data), the model is also robust to violations of the third assumption.

One key generalization of the model, which we will also consider in Section 6, allows covariates to be included to control for the correlation between X_i and the unknowns, to allow for multiple clusters, or to model spatial autocorrelation. Because the bounds, which differ in width and information content for each i, generally provide substantial information, even X_i can be used as a covariate. In previous approaches, which do not include the information in the bounds, including X_i leads to models that are unidentified.

The model assumptions are especially important given the loss of information due to aggregation. In fact, this loss of information can be expressed by noting that the joint distribution of β_i^b and β_i^w cannot be fully identified from the data without some untestable assumptions. To be precise, distributions with positive mass over *any* curve that connects the bottom left point ($\beta_i^b = 0$, $\beta_i^w = 0$) to the top right point ($\beta_i^b = 1$, $\beta_i^w = 1$) of a tomography plot cannot be rejected by the data (King 1997:191). Other features of the distribution are estimable. This fundamental indeterminacy is of course a problem because it prevents pinning down the quantities of interest with certainty, but it can also be something of an opportunity because different distributional assumptions can lead to the same estimates, especially since only those pieces of the distributions above the tomography lines are used in the final analysis. Further details with regard to inference for this model can be found in King (1997).

4. WHAT ARE HIERARCHICAL MODELS?

In the context of meta-analysis (Morris and Normand 1992), one attempts to combine data from related, but statistically independent, studies to summarize information about possible treatment effects. In the context of small-area estimation (Ghosh, Natarajan, Stroud, and Carlin 1998), one attempts to pool data across geographic regions or local areas. In both of these cases, the expectation is that by "borrowing strength" from the other cells, an efficiency is obtained by reducing the standard error of the estimate of each particular study

or region. By the early 1960s, it was known in simple situations (James and Stein 1960) that this borrowing or "shrinkage" results in an estimator that dominates the unpooled analogue. The basic tool for facilitating this pooling has been the hierarchical model.

The fundamental idea behind hierarchical models is as follows. In standard, nonhierarchical models, the procedure is to specify at the outset the full distribution for an outcome variable; for example, $Y_i \sim p(y|\theta)$. From this assumption, the likelihood (or, by adding priors, the posterior) is formed and analyzed directly. This nonhierarchical approach is of course time honored, enormously useful, and indeed can even be thought of as encompassing hierarchical models as a special case. The difficulty in nonhierarchical modeling is the specification of the full distribution, $p(y|\theta)$, since it is difficult to conceptualize complicated multidimensional densities, and since distribution theory has not given us models that are sufficiently flexible for many types of data.

Hierarchical models construct the same required density in separate steps. For example, we might begin with an assumption that $Y_i|\beta \sim p_1(y|\beta)$ and then recognize that β is not constant over i. We would then add to this a second step in the hierarchy by assuming that β has a distribution, such as $\beta \sim p_2(\beta|\theta)$. The two distributions can be combined by the usual rules of probability to give the same density as could have been specified at the outset:

$$p(y|\theta) = \int_{-\infty}^{\infty} p_1(y|\beta)p_2(\beta|\theta)d\beta. \tag{3}$$

In other words, the product $p_1(y|\beta)p_2(\beta|\theta)$ gives us the joint distribution $p_3(y, \beta|\theta)$. Then, the integral in equation (3) collapses this joint distribution over the unknown β parameter to yield $p(y|\theta)$. A third and also equivalent way to understand this equation is that by averaging $p_1(y|\beta)$ over the uncertainty in the unit-specific effects—that is, $p_2(\beta|\theta)$—we recover the distribution of interest. Thus, even though $p(y|\theta)$ may have such a complicated form that a researcher would not be able to intuit it directly, it can still be constructed from simpler components.

The idea of building distributions hierarchically in this way has been known almost as long as probability theory, but the difficulty of

computing the integral in equation (3) has prevented many from carrying out the strategy in practice in most cases. However, although integrals are often difficult or impossible to compute, drawing random samples is often much easier. Thus, Monte Carlo simulation is a practical solution to this problem, since it enables a researcher to approximate $p(y|\theta)$ to any degree of accuracy by substituting computing cycles for analytical calculations that may not be possible. To solve the problem in equation (3), we merely need to draw random samples of $\tilde{\beta}$ from $p_2(\beta|\theta)$ and then, conditional on these samples, draw y randomly from $p_1(y|\tilde{\beta})$. A histogram of the draws of y approximates $p(y|\theta)$.

One unusual aspect of hierarchical modeling works is that the ultimate distribution of the outcome variable, $p(y|\theta)$, is not typically written down. In many cases, of course, it would not be possible to do so. Fortunately, the hierarchical structure is typically much easier to interpret and can be made to follow, in many cases, the hierarchical structure of the data generation process.

The recent dramatic increases in computing speed have greatly facilitated simulation-based hierarchical modeling. Another important development has been iterative simulation methods, such as MCMC methods, which have made the technique of simulation much more widely applicable (Tanner 1996).

In the present context, we also use hierarchical models—not simply to decrease variation of the parameter estimates but to obtain estimates of the unobserved quantities β_i^b and β_i^w. Like King's model, ours also includes the information in the bounds and the application of distributional assumptions to borrow statistical strength across precincts to model information within the bounds. In this article, we consider an alternative distributional structure to provide a data analytic check on King's model. In addition, in Section 6, we consider the incorporation of covariates into the model and provide a means to assess the significance of a given covariate.

5. THE BINOMIAL-BETA MODEL: NO COVARIATES

In this section, we present our first alternative hierarchical model for ecological inference, with no covariates. In Section 6, we present

a hierarchy to allow for the incorporation of covariates into the model. Our hierarchical models use MCMC methods, specifically the Gibbs sampler (see Tanner 1996).[3]

Following Section 2, suppose that there are p precincts. Let T_i' denote the number of voting-age people who turn out to vote. At the top level of the hierarchy, we assume that T_i' follows a binomial distribution with probability equal to $\theta_i = X_i\beta_i^b + (1 - X_i)\beta_i^w$ and count N_i. Note that at this level, it is assumed that the *expectation* of T_i, rather than T_i, is equal to $X_i\beta_i^b + (1 - X_i)\beta_i^w$. It therefore follows that the contribution of the data of precinct i to the likelihood is

$$\left(X_i\beta_i^b + (1 - X_i)\beta_i^w\right)^{T_i'} \left(1 - X_i\beta_i^b - (1 - X_i)\beta_i^w\right)^{(N_i - T_i')}. \quad (4)$$

By taking the logarithm of this contribution to the likelihood and differentiating with respect to the parameters of interest β_i^b and β_i^w, it can be shown that the maximum of (4) is not a unique point but rather a line whose equation is given by the tomography line:

$$\beta_i^w = \left(\frac{T_i}{1 - X_i}\right) - \left(\frac{X_i}{1 - X_i}\right)\beta_i^b,$$

where T_i is the fraction of voting-age people who turn out to vote. Thus, the log likelihood for precinct i looks like two playing cards leaning against each other. Furthermore, the derivative in the direction of steepest ascent at the point $(\beta_i^b, \beta_i^w) = (0.5, 0.5)$ is equal[4] to $2N_i|1 - 2T_i|\sqrt{2X_i^2 - 2X_i + 1}$. As long as T_i is fixed and bounded away from 0.5 (and X_i is a fixed known value between 0 and 1), the derivative at this point is seen to increase with N_i; that is, the pitch of the playing cards increases with the sample size. In other words, for large N_i, the log likelihood for precinct i degenerates from a surface defined over the unit square into a single playing card standing perpendicular to the unit square and oriented along the corresponding tomography line.

At the second level of the hierarchical model, we assume that β_i^b is sampled from a beta distribution with parameters c_b and d_b and that β_i^w is sampled independently from a beta distribution with parameters c_w and d_w. The beta family of distributions, defined over the interval [0,1], is quite a rich family, providing shapes ranging from flat, to U shaped, to bell shaped, to skewed exponential (see Lee 1997:78-79).

As we will see in an example later in this section, this flexibility allows us to relax the single-cluster assumption of the truncated bivariate normal. Although β_i^b and β_i^w are taken to be a priori independent, we will see from the full conditionals of the Gibbs sampler that they are a posteriori dependent.

At the third and final level of the hierarchical model, we assume that the unknown parameters c_b, d_b, c_w, and d_w follow an exponential distribution with a large mean. In the examples in this article, we take the mean to be $1/\lambda = 2$ (i.e., a fairly noninformative distribution at the final level).

By Bayes' theorem, the posterior distribution is proportional to the likelihood times the prior. Thus, given this three-stage model, it then follows that the posterior distribution for the parameters is proportional to

$$
\begin{aligned}
&p(\text{data}\,|(\beta_i^b, \beta_i^w), i = 1, \ldots, p) \times p((\beta_i^b, \beta_i^w), \\
&\qquad i = 1, \ldots, p|c_b, d_b, c_w, d_w) \times p(c_b, c_w, d_b, d_w) \\
&= \prod_{i=1}^{p} \left(X_i\beta_i^b + (1 - X_i)\beta_i^w\right)^{T_i'} \left(1 - X_i\beta_i^b - (1 - X_i)\beta_i^w\right)^{(N_i - T_i')} \\
&\times \prod_{i=1}^{p} \frac{\Gamma(c_b + d_b)}{\Gamma(c_b)\Gamma(d_b)} (\beta_i^b)^{c_b - 1}(1 - \beta_i^b)^{d_b - 1} \prod_{i=1}^{p} \frac{\Gamma(c_w + d_w)}{\Gamma(c_w)\Gamma(d_w)} \\
&\qquad (\beta_i^w)^{c_w - 1}(1 - \beta_i^w)^{d_w - 1} \\
&\times \exp(-\lambda c_b) \times \exp(-\lambda c_w) \times \exp(-\lambda d_b) \times \exp(-\lambda d_w)\ .
\end{aligned}
$$

Obtaining the marginals of this posterior distribution using high-dimensional numerical integration is not feasible. Instead, we use the Gibbs sampler (Tanner 1996). To implement the Gibbs sampler, we need the following full conditional distributions; that is, we need the distribution of each unknown parameter conditional on the full set of the remaining parameters:

$$
\begin{aligned}
p(\beta_i^b|\beta_i^w, c_b, d_b) \propto &\left(X_i\beta_i^b + (1 - X_i)\beta_i^w\right)^{T_i'} \\
&\times \left(1 - X_i\beta_i^b - (1 - X_i)\beta_i^w\right)^{(N_i - T_i')} \\
&\times (\beta_i^b)^{c_b - 1}(1 - \beta_i^b)^{d_b - 1}
\end{aligned}
$$

$$p(\beta_i^w|\beta_i^b, c_w, d_w) \propto \left(X_i\beta_i^b + (1 - X_i)\beta_i^w\right)^{T_i'} \times \left(1 - X_i\beta_i^b - (1 - X_i)\beta_i^w\right)^{(N_i - T_i')} \times (\beta_i^w)^{c_w - 1}(1 - \beta_i^w)^{d_w - 1}$$

$$p(c_b|\beta_i^b, i = 1, \ldots, p, d_b) \propto \left(\frac{\Gamma(c_b + d_b)}{\Gamma(c_b)}\right)^p \exp[(\sum_{i=1}^{p} \log \beta_i^b - \lambda)c_b]$$

$$p(d_b|\beta_i^b, i = 1, \ldots, p, c_b) \propto \left(\frac{\Gamma(c_b + d_b)}{\Gamma(d_b)}\right)^p \exp[(\sum_{i=1}^{p} \log(1 - \beta_i^b) - \lambda)d_b]$$

$$p(c_w|\beta_i^w, i = 1, \ldots, p, d_w) \propto \left(\frac{\Gamma(c_w + d_w)}{\Gamma(c_w)}\right)^p \exp[(\sum_{i=1}^{p} \log \beta_i^w - \lambda)c_w]$$

$$p(d_w|\beta_i^w, i = 1, \ldots, p, c_w) \propto \left(\frac{\Gamma(c_w + d_w)}{\Gamma(d_w)}\right)^p \exp[(\sum_{i=1}^{p} \log(1 - \beta_i^w) - \lambda)d_w],$$

where the a priori independence assumptions cause some of the conditioning parameters to drop out of some equations.

To generate a Gibbs sampler (Markov) chain, one draws a random deviate from each of these full conditionals, in turn updating the value of the variable after each draw. Unfortunately, none of these distributions are standard distributions (e.g., normal, gamma, etc.), for which prewritten sampling subroutines are available. For this reason, we use the Metropolis algorithm to sample from each of these distributions. Thus, to sample a value for c_b, d_b, c_w, or d_w, a candidate value for the next point in the Metropolis chain is drawn from the univariate normal distribution with mean equal to the current sample

value and variance sufficiently large to allow for variation around the current sample value. To sample a value for β_i^b or β_i^w, a candidate value for the next point in the Metropolis chain is drawn from the uniform distribution. The candidate value is then accepted or rejected according to the Metropolis scheme of evaluating the ratio of the full conditional at the candidate value to the full conditional evaluated at the current point in the chain. If this ratio is greater than or equal to unity, the candidate value is accepted. If the ratio is less than unity, the candidate value is accepted with probability given by this ratio (see Tanner 1996). The Metropolis algorithm is iterated, and the final value in this chain is treated as a deviate from the full conditional distribution. In the examples considered in this article, we iterated the Metropolis algorithm 25 times to yield a deviate. A rigorous theory for the convergence of the Gibbs sampler and other MCMC methods is given in Tierney (1994).

A variety of methods are available for assessing convergence for a given data set. A critical review of these methods is presented in Cowles and Carlin (1996). A very popular method presented in Gelman and Rubin (1992) is based on comparing the between-chain variation (among multiple chains) to the within-chain variation. Clearly, if the between-chain variation is much larger than the within-chain variation, further iteration is required. Although this approach can fail (see Tanner 1996), it generally works well in practice and is fairly simple to implement. For the examples considered in this article, the outputs of three chains were compared. Having considered sufficiently long chains, there was very little difference across these three runs given the different starting values. All examples in this article were run on a Hewlett-Packard J210 workstation running FORTRAN, with IMSL supplying the pseudorandom deviates.

5.1. EXAMPLE 1

The data considered in this example are taken from King (1997, chap. 10). The data include voter registration and racial background information of people in 275 counties in four U.S. states: Florida, Louisiana, North Carolina, and South Carolina. The data from each county include the total voting-age population (N_i), the proportion who are black (X_i), and the total number registered (T_i') in 1968. The

goal of this analysis is to estimate the fraction of blacks registered and the fraction of whites registered in each county. The data also include information from public records on the true fraction of blacks (β_i^b) and whites (β_i^w) who are registered in each county. We chose these data in part because the (known) low correlation between X_i and (β_i^b, β_i^w) simplifies the analysis. Although this relationship would not generally be known in real applications, the simplification helps us put aside one important problem while improving other features of the statistical model.

The Gibbs sampler chains for this data set were run for 600,000 iterations. The results presented in the figures in this section are based on the final 300,000 iterations. The reason for this run length is discussed below.

Figure 2 presents the posterior distribution of the second-stage mean for blacks ($c_b/(c_b + d_b)$) and the second-stage mean for whites ($c_w/(c_w+d_w)$). Using the final 300,000 iterates, we estimate the mean of the posterior distribution for blacks to be 0.60, whereas the corresponding value for whites is 0.85. These values compare favorably with the corresponding true values for this data set (i.e., the fraction of registered blacks and the fraction of registered whites in all counties) of 0.56 and 0.85, as well as with the figures quoted by King (1997) of 0.62 and 0.83 based on the truncated bivariate normal. The posterior standard deviations of the second-stage mean in the present context are 0.04 (blacks) and 0.02 (whites) and are congruent with the values quoted by King (1997) of 0.04 and 0.01.

Figure 3 presents the posterior distribution of the fraction of whites registered and the fraction of blacks registered in county 50 (i.e., β_{50}^b and β_{50}^w). The posterior distribution of β_{50}^w indicates that although the distribution is skewed, a high percentage of whites in this county are registered. The posterior distribution of β_{50}^b, which is defined over a much larger region, is also skewed and indicative of a lower registration rate for blacks. The posterior means of 0.73 and 0.98 for blacks and whites, respectively, are similar to the true values for this county of 0.63 and 1.00. *Within* county, the present approach can detect possible bimodality of the distribution of the parameters. For example, with regard to county 150 (Figure 4), we see that the posterior distribution of β_{150}^b not only has significant positive mass over the entire interval [0.0,1.0] but actually appears to be bimodal—an

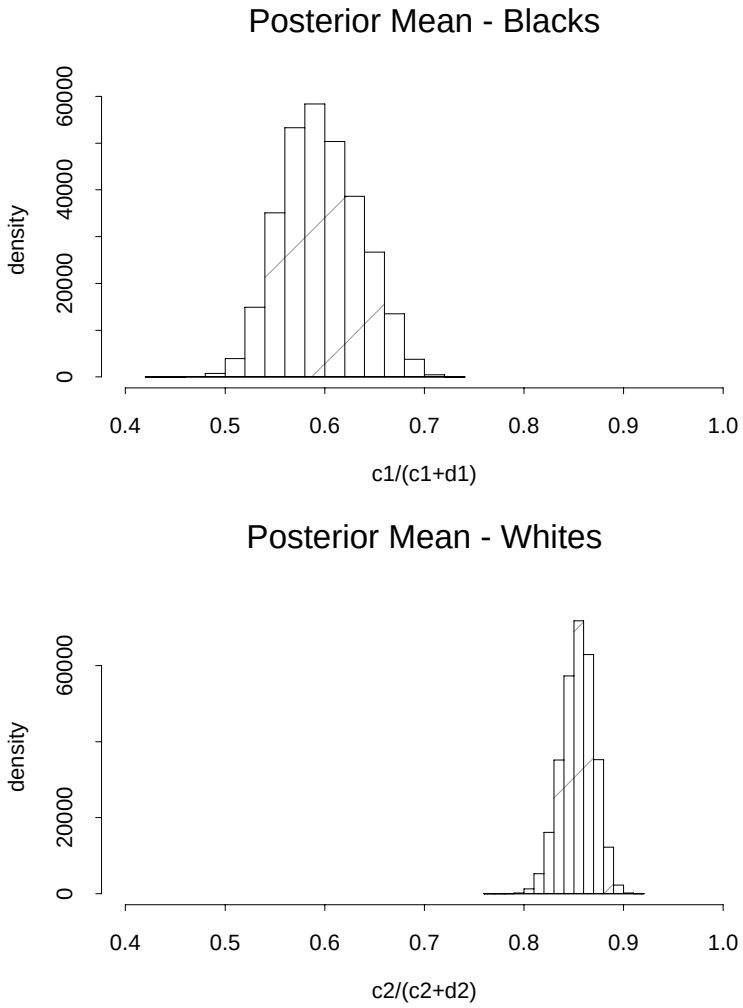

Figure 2: Posterior Distribution of $c_b/(c_b + d_b)$ and $c_w/(c_w + d_w)$
NOTE: The mean of a beta (a, b) distribution is $a/(a + b)$. These figures present the posterior distribution of the second-stage means for blacks and for whites.

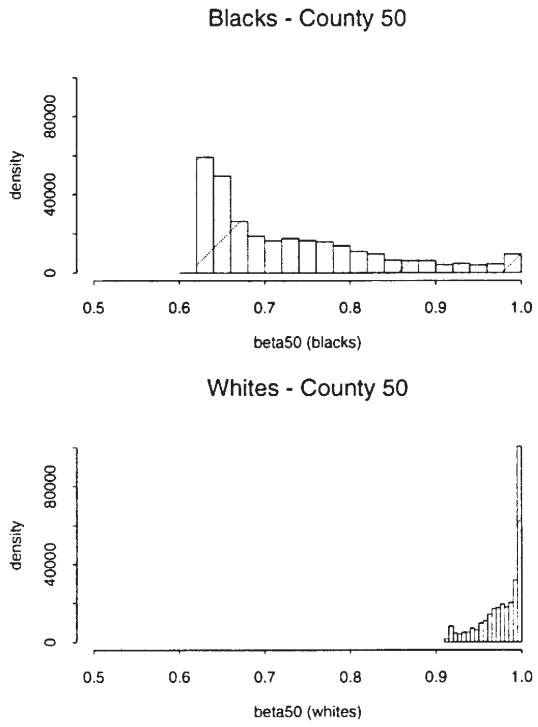

Figure 3: Posterior Distribution of β_{50}^{b} and β_{50}^{w}

observation that was not detected with the truncated bivariate normal model.[5] The corresponding distribution for whites is less diffuse. The posterior means of 0.48 and 0.58 for blacks and whites, respectively, are similar to the true values for this county of 0.42 and 0.60.

5.2. EXAMPLE 2

In the previous subsection, we noticed that the hierarchical model can detect bimodality within precincts. It is important to note that both the present hierarchical model and the model in King (1997) can detect

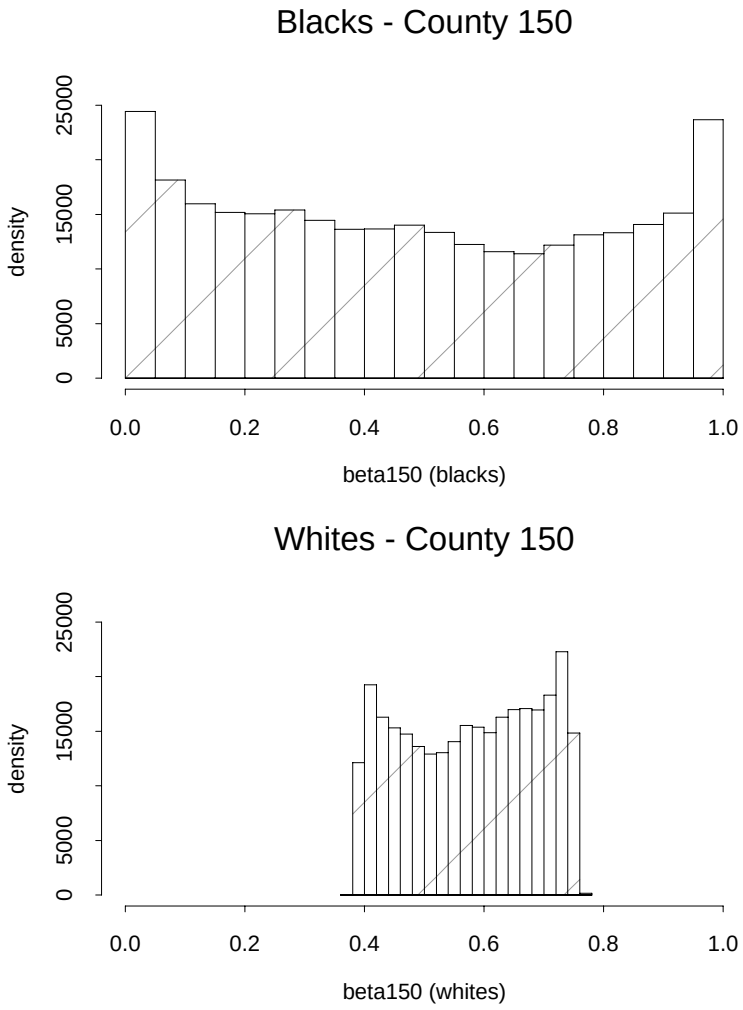

Figure 4: Posterior Distribution of β_{150}^{b} and β_{150}^{w}

bimodality *across* precincts, even without introducing covariates. To illustrate this point, we generated data corresponding to 100 precincts from a bimodal truncated normal distribution—50 precincts from a

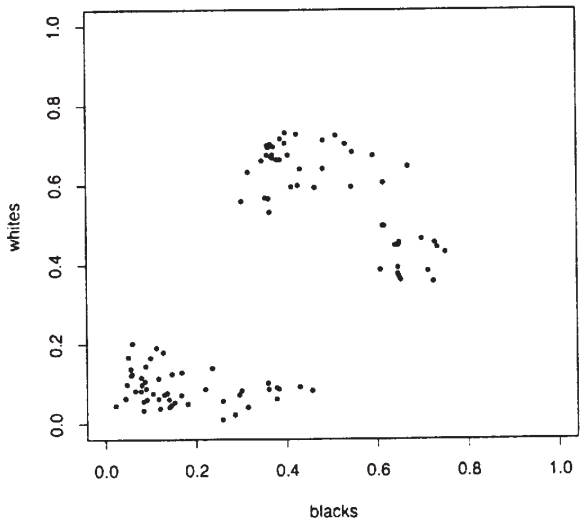

Figure 5: $\hat{E}(\beta_i^w)$ **Versus** $\hat{E}(\beta_i^b)$, $i = 1, \ldots, 100$

truncated normal centered at (0.1, 0.1) and 50 precincts from a truncated normal centered at (0.6, 0.6). The hierarchical model was then applied to these data, with the Gibbs sampler run for 100,000 iterations. The values from the final 40,000 iterations were then analyzed. In particular, for each precinct we averaged the simulated β_i^b and β_i^w values to obtain $\hat{E}(\beta_i^b)$ and $\hat{E}(\beta_i^w)$. Figure 5 presents a scatter plot of these 100 points. Clearly, this methodology was able to recover the underlying bimodal structure of the data. (Parallel results, not presented, were also obtained from King 1997.)

In this example, both the hierarchical model and the King (1997) model are able to detect the bimodality due to the contribution of the data to the likelihood (see equation (4)). Corresponding to each mode, the tomography lines crisscross and bunch together. Heuristically,

when we compute the mean of the "projection" of either the truncated normal or the beta distributions along each line, we are able to detect the underlying bimodal nature of the data.

6. THE BINOMIAL-BETA MODEL: WITH COVARIATES

A key point of King (1997) is the importance of bringing in and representing formally the normally vast array of nonquantified knowledge to which researchers generally have access, and which is not represented in T_i, X_i, and N_i. Only by supplementing the formal data set with this qualitative knowledge is it possible to begin to fill in the missing information lost to aggregation and reach reliable ecological inferences. This approach, which we capitalize on and advance further, is to provide a rich family of models from which the data analyst can choose. Our model without covariates allows for a posteriori dependence between X_i and β_i^b, β_i^w, even though it assumes a priori independence (King's model has the same property). Nonetheless, we expand the model presented in the previous section by allowing the parameters to vary as a function of additional measured covariates. Covariates allow the distribution to be more flexible, effectively allowing more complicated shapes of densities. By conditioning on X_i, or correlates of it, one can begin to model the relationship between this information and β_i^b and β_i^w rather than assume they are a priori independent. Moreover, our Bayesian methodology provides a formal approach for assessing the significance of a covariate.

Following the notation of Section 2, let Z_i denote a covariate value associated with precinct i. In this article, we assume that Z_i is a scalar for simplicity of presentation—the generalization to a vector is straightforward.

As in the previous section, we will approach our analysis of this problem using a hierarchical model. At the first stage of the hierarchy, we again assume that T_i' follows a binomial distribution, although in the present model the probability equals $\theta_i^{Z_i} = X_i\beta_i^{b(Z_i)} + (1 - X_i)\beta_i^{w(Z_i)}$, with count N_i. Note that in contrast to the model of Section 5, here both $\beta_i^{b(Z_i)}$ and $\beta_i^{w(Z_i)}$ depend on the covariate Z_i, with the dependency on Z_i to be specified at the second stage. To simplify

notation, we will suppress the dependency on Z_i in the remainder of this section.

At the second stage of the hierarchical model, we assume that β_i^b is sampled from a beta distribution with parameters $d_b \exp(\alpha + \beta Z_i)$ and d_b, whereas β_i^w is sampled from a beta distribution with parameters $d_w \exp(\gamma + \delta Z_i)$ and d_w. Recall that the mean of the beta distribution with parameters (a, b) is $a/(a + b)$. Thus, the second-stage mean of $\beta_i^b = E(\beta_i^b)$ is

$$\frac{d_b \exp(\alpha + \beta Z_i)}{d_b + d_b \exp(\alpha + \beta Z_i)} = \frac{\exp(\alpha + \beta Z_i)}{1 + \exp(\alpha + \beta Z_i)},$$

which implies that

$$\log\left(\frac{E(\beta_i^b)}{1 - E(\beta_i^b)}\right) = \alpha + \beta Z_i.$$

In other words, the log odds depend linearly on the covariate Z_i. Similarly, the second stage of the hierarchical model implies that

$$\log\left(\frac{E(\beta_i^w)}{1 - E(\beta_i^w)}\right) = \gamma + \delta Z_i.$$

At the third and final stage, we follow standard Bayesian practice and treat the regression parameters to be a priori independent, putting a flat prior on these regression parameters (α, β, γ, and δ). The parameters d_b and d_w are assumed to follow an exponential distribution with mean λ. In the examples in this section, we take $1/\lambda = 2$ (i.e., a fairly noninformative prior).

To implement the Gibbs sampler, we require the full conditionals, which are given as

$$\begin{aligned} p(\beta_i^b | \beta_i^w, \alpha, \beta, d_b) \propto & \left(X_i \beta_i^b + (1 - X_i)\beta_i^w\right)^{T_i'} \\ & \times \left(1 - X_i \beta_i^b - (1 - X_i)\beta_i^w\right)^{(N_i - T_i')} \\ & \times (\beta_i^b)^{d_b \exp(\alpha + \beta Z_i) - 1} (1 - \beta_i^b)^{d_b - 1} \end{aligned}$$

$$p(\beta_i^w|\beta_i^b,\gamma,\delta,d_w) \propto \left(X_i\beta_i^b + (1-X_i)\beta_i^w\right)^{T_i'} \times \left(1 - X_i\beta_i^b - (1-X_i)\beta_i^w\right)^{(N_i - T_i')} \times (\beta_i^w)^{d_w \exp(\gamma+\delta Z_i)-1}(1-\beta_i^w)^{d_w-1}$$

$$p(d_b|\beta_i^b, i = 1,\ldots,p,\alpha,\beta) \propto \left(\prod_{i=1}^{p} \frac{\Gamma(d_b(1+\exp(\alpha+\beta Z_i)))}{\Gamma(d_b)\Gamma(d_b\exp(\alpha+\beta Z_i))}(\beta_i^b)^{d_b\exp(\alpha+\beta Z_i)}(1-\beta_i^b)^{d_b}\right) \times \exp(-\lambda d_b)$$

$$p(d_w|\beta_i^w, i = 1,\ldots,p,\gamma,\delta) \propto \left(\prod_{i=1}^{p} \frac{\Gamma(d_w(1+\exp(\gamma+\delta Z_i)))}{\Gamma(d_w)\Gamma(d_w\exp(\gamma+\delta Z_i))}(\beta_i^w)^{d_w\exp(\gamma+\delta Z_i)}(1-\beta_i^w)^{d_w}\right) \times \exp(-\lambda d_w)$$

$$p(\alpha|\beta_i^b, i = 1,\ldots,p,\beta,d_b) \propto \prod_{i=1}^{p} \frac{\Gamma(d_b(1+\exp(\alpha+\beta Z_i)))}{\Gamma(d_b\exp(\alpha+\beta Z_i))} \times (\beta_i^b)^{d_b\exp(\alpha+\beta Z_i)}$$

$$p(\beta|\beta_i^b, i = 1,\ldots,p,\alpha,d_b) \propto \prod_{i=1}^{p} \frac{\Gamma(d_b(1+\exp(\alpha+\beta Z_i)))}{\Gamma(d_b\exp(\alpha+\beta Z_i))} \times (\beta_i^b)^{d_b\exp(\alpha+\beta Z_i)}$$

$$p(\gamma|\beta_i^w, i = 1,\ldots,p,\delta,d_w) \propto \prod_{i=1}^{p} \frac{\Gamma(d_w(1+\exp(\gamma+\delta Z_i)))}{\Gamma(d_w\exp(\gamma+\delta Z_i))} \times (\beta_i^w)^{d_w\exp(\gamma+\delta Z_i)}$$

$$p(\delta|\beta_i^w, i = 1,\ldots,p,\gamma,d_w) \propto \prod_{i=1}^{p} \frac{\Gamma(d_w(1+\exp(\gamma+\delta Z_i)))}{\Gamma(d_w\exp(\gamma+\delta Z_i))} \times (\beta_i^w)^{d_w\exp(\gamma+\delta Z_i)}.$$

As was the situation in Section 5, none of these distributions is a standard distribution (e.g., normal, gamma, etc.), for which prewritten sampling subroutines are available. For this reason, we again use the Metropolis algorithm to sample from each of these distributions.

Thus, to sample a value for d_b, d_w, α, β, γ, or δ, a candidate value for the next point in the Metropolis chain is drawn from the univariate normal distribution with mean equal to the current sample value and variance sufficiently large to allow for variation around the current sample value. To sample a value for β_i^b or β_i^w, a candidate value for the next point in the Metropolis chain is drawn from the uniform distribution. As in the example of Section 5.1, we iterated the Metropolis algorithm 25 times. The candidate value is then accepted or rejected according to the standard Metropolis scheme (Tanner 1996).

6.1. EXAMPLES

To illustrate the methodology of incorporating covariates into the hierarchical framework, we consider two examples. In the first example, data from 200 precincts were simulated assuming the truncated normal distribution presented in King (1997). In addition, an independent normal random deviate was generated for each precinct corresponding to white noise. Clearly, in such a situation, one would expect the methodology to recognize that the covariate information is irrelevant. In addition, one would expect this binomial-beta model to give similar results to those of King's truncated normal model, since the data were generated according to this model.

Figure 6 presents the posterior distribution of β—the slope parameter for regressing the log odds for blacks on the independent normal deviates. In this example, the algorithm converged much quicker than for the data in Section 5.1. Here, the chains were iterated 25,000 times, with the presented results based on the final 10,000 iterates. For this marginal, the 90% credible interval (obtained by locating the 5th and 95th percentiles of the simulated values) is $(-0.31, 0.08)$. The analogous 95% credible interval is given by $(-0.35, 0.11)$. Because zero is located in both these intervals, zero is a plausible value for the regression parameter, and our analysis indicates (as expected) that there is little evidence to suggest a regression effect.

Figure 7 presents the corresponding posterior distribution of δ—the slope parameter for regressing the log odds for whites on the independent normal deviates. Here, the 90% credible interval is $(-0.17, 0.15)$, whereas the 95% credible interval is $(-0.20, 0.19)$. Again,

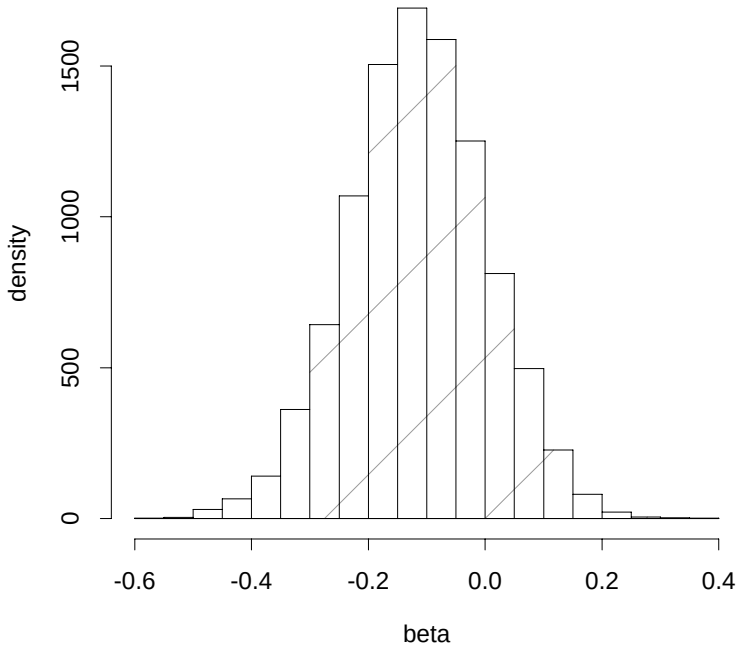

Figure 6: Posterior Distribution of β—The Slope Parameter for Regressing the Log Odds for Blacks on the Covariate

there is little evidence to suggest a regression effect, since zero is located in both of these ranges of plausible values.

Figure 8 presents the posterior distribution of β_1^b and β_1^w. The mean of these distributions (0.14 and 0.07 for blacks and whites, respectively), as well as the standard deviations of these distributions (0.10 and 0.03 for blacks and whites, respectively), are congruent with the results based on the truncated normal model of 0.14 and 0.07 for the means and 0.09 and 0.03 for the standard deviations. Similar

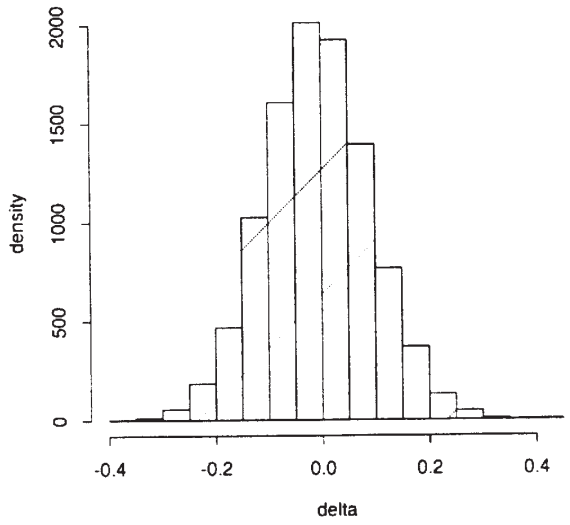

Figure 7: Posterior Distribution of δ—The Slope Parameter for Regressing the Log Odds for Whites on the Covariate

results are obtained for other precincts. In this context, where the true model is the truncated normal, the binomial-beta hierarchical model is capable of recovering that structure.

As a second example of the incorporation of covariates into the hierarchical model, we consider a situation in which the covariate is informative. For this example, the β_i^b and β_i^w are again generated from a truncated bivariate normal distribution. However, in contrast to the previous example, β_i^b is then perturbed by adding a multiple of X_i, whereas β_i^w is then perturbed by subtracting a multiple of X_i. Can the binomial-beta model recognize this dependency on the covariate?

Figure 9 presents the marginal posterior distribution of δ (the slope parameter for whites) based on iterations 20,000 through 40,000. The 90% and 95% credible intervals for this marginal are $(-4.88, -1.22)$

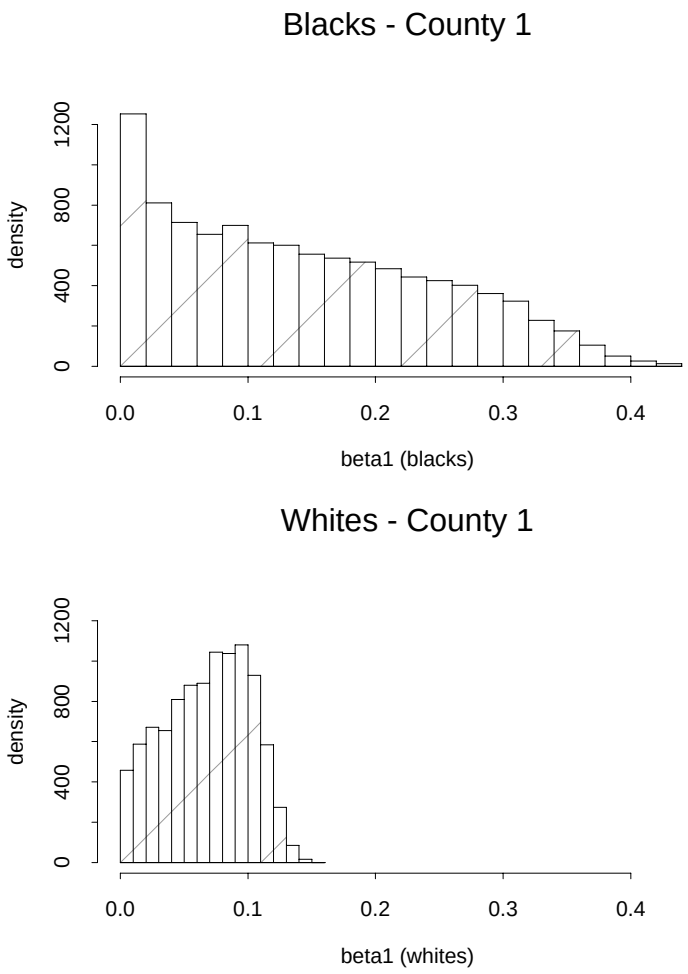

Figure 8: Posterior Distribution of β_1^b and β_1^w

and $(-5.12, -1.00)$, respectively. Because zero is in neither range of plausible values for δ, there does seem to be some evidence of a dependency of β_i^w on X_i. In fact, from the negative sign of the slope parameter, one can conclude that the fraction of whites registered decreases as X_i increases.

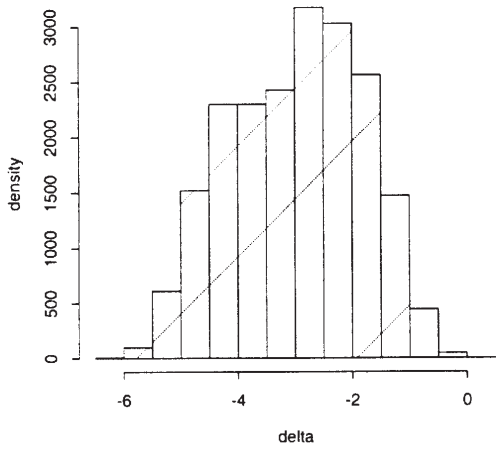

Figure 9: Posterior Distribution of δ—The Slope Parameter for Regressing the Log Odds for Whites on the Covariate

Figure 10 presents the corresponding marginal posterior distribution of β (the slope parameter for blacks), also based on iterations 20,000 through 40,000. The 90% and 95% credible intervals for this marginal are (0.68, 4.52) and (0.41, 4.74), respectively. Thus, as was the case for whites, zero is not a plausible value providing evidence to suggest a dependency of β_i^b on X_i. From the positive sign of the slope parameter, one can conclude that the fraction of blacks registered increases as X_i increases.

7. CONCLUDING REMARKS

Modeling uncertainty in T_i and X_i, as done here, has the potential to expand significantly the range of applications of reliable models of ecological inference. The model can be used to represent sampling variability if the observed variables are estimated from sample surveys

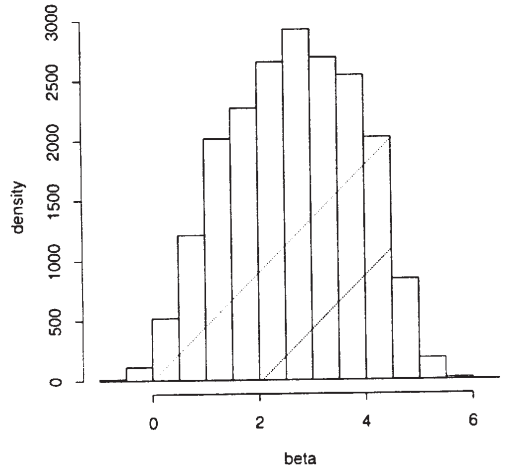

Figure 10: Posterior Distribution of β—The Slope Parameter for Regressing the Log Odds for Blacks on the Covariate

instead of assumed known. One interesting application is using ecological inference methods to study individual-level change between two independent cross-sectional surveys broken into profiles defined by demographic variables common to both surveys (as in Penubarti and Schuessler 1998). Our model may also be useful for more traditional ecological inference applications where N_i is very small, as in mortality data, and so the tomography line in Figure 1 becomes a dotted line. The model is also useful if T_i and X_i are directly observed without sampling but with random measurement error.

The focus of this article has been on hierarchical models based on the beta distribution. Alternative hierarchical models can be based on bivariate extensions of the beta distribution (e.g., Gupta and Wong 1985), as well as by reparameterizing and placing a Dirichlet distribution on the four (unobserved) cell probabilities of the 2×2 table. By casting the ecological inference problem in terms of a hierarchical model, we have opened up a wealth of new tools for the analysis of

ecological correlation data. One goal of future work will be to understand the operating characteristics of these various hierarchical models, as well as compare and contrast their strengths and weaknesses.

NOTES

1. For the historians of science among us: Despite the fact that these two monumental articles were written by two colleagues and friends in the same year and in the same department and university (the Department of Sociology at the University of Chicago), the principals did not discuss their work prior to completion. The Duncan and Goodman articles are each brilliant contributions to social science methodology, and even judging by today's standards, nearly a half century after their publication, the articles still are models of clarity and creativity.

2. King (1997) also showed that the ecological inference problem is mathematically equivalent to the tomography problem of many medical imaging procedures (such as CAT and PET scans), where one attempts to reconstruct the inside of an object by passing X rays through it and gathering information only from the outside. Because the line sketched out by an X ray is closely analogous to equation (2), King labeled the latter a *tomography line* and the corresponding graph a *tomography graph*.

3. The goal of the Gibbs sampler is to draw random values from a joint distribution—for example, $p(x, y)$—which may be difficult to accomplish directly. Instead, we analytically compute the full conditionals and then draw x from $p(x|y)$ given a starting value for y, y from $p(y|x)$ given the simulated value of x, and x from $p(x|y)$ given the simulated value of y; we then iterate until stochastic convergence. After convergence, subsequent draws from this sequence are equivalent to drawing from $p(x, y)$ directly.

4. This result is obtained by computing the length of the gradient vector of the log likelihood for precinct i at the point (0.5, 0.5) (see Marsden and Hoffman 1993:350).

5. This bimodality explains to some degree the slow convergence of the chain in this instance. Typically, when the underlying posterior has bimodality or multimodality, the corresponding chain will tend to wander about a given mode, then migrate to the other mode and visit that portion of the space, before migrating to another mode or returning to the first mode.

REFERENCES

Cowles, M. K. and B. Carlin. 1996. "Markov Chain Monte Carlo Diagnostics: A Comparative Review." *Journal of the American Statistical Association* 91:883-904.

Duncan, O. D. and B. Davis. 1953. "An Alternative to Ecological Correlation." *American Sociological Review* 18:665-66.

Gelman, A. and D. B. Rubin. 1992. "Inference From Iterative Simulation Using Multiple Sequences." *Statistical Science* 7:457-72.

Ghosh, M., K. Natarajan, T.W.F. Stroud, and B. Carlin. 1998. "Generalized Linear Models for Small-Area Estimation." *Journal of the American Statistical Association* 93:273-82.

Goodman, L. A. 1953. "Ecological Regressions and the Behavior of Individuals." *American Sociological Review* 18:663-66.

Goodman, L. A. 1959. "Some Alternatives to Ecological Correlation." *American Journal of Sociology* 64:610-24.

Gupta, A. K. and C. F. Wong. 1985. "On Three and Five Parameter Bivariate Beta Distributions." *Metrika* 32:85-91.

James, W. and C. Stein. 1960. "Estimation With Quadratic Loss." Pp. 361-79 in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press.

King, G. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton, NJ: Princeton University Press.

King, G., O. Rosen, and M. A. Tanner. 1999. "Who Voted for Hitler?" (in preparation).

Lee, Peter M. 1997. *Bayesian Statistics*. 2d ed. New York: John Wiley.

Marsden, J. E. and M. J. Hoffman. 1993. *Elementary Classical Analysis*. New York: W. H. Freeman.

Morris, C. N. and S. L. Normand. 1992. "Hierarchical Models for Combining Information and for Meta-Analyses." Pp. 321-44 in *Bayesian Statistics*, edited by J. M. Bernardo, J. O. Berger, A. P. Dawid, and A.F.M. Smith. Oxford: Oxford University Press.

Ogburn, W. F. and I. Goltra. 1919. "How Women Vote: A Study of an Election in Portland, Oregon." *Political Science Quarterly* 34:413-33.

Penubarti, M. and A. Schuessler. 1998. "Inferring Micro- from Macrolevel Change." Unpublished manuscript, Department of Political Science, University of California, Los Angeles.

Tanner, M. A. 1996. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. 3d ed. New York: Springer- Verlag.

Tierney, L. 1994. "Markov Chains for Exploring Posterior Distributions." *Annals of Statistics* 22:1701-62.

Gary King is a professor of government in the Faculty of Arts and Sciences at Harvard University and director of the Harvard-MIT Data Center.

Ori Rosen is an assistant professor in the Department of Statistics at the University of Pittsburgh.

Martin A. Tanner is a professor in the Department of Statistics at Northwestern University.