Statistica Neerlandica (2001) Vol. 55, nr. 2, pp. 134–156

Bayesian and frequentist inference for ecological inference: the $R \times C$ case

Ori Rosen*

Department of Statistics, University of Pittsburgh, 2702 Cathedral of Learning, Pittsburgh, PA 15260, USA

Wenxin Jiang

Department of Statistics, Northwestern University, USA

Gary King

Department of Government, Harvard University, USA

Martin A. Tanner

Department of Statistics, Northwestern University, USA

In this paper we propose Bayesian and frequentist approaches to ecological inference, based on $R \times C$ contingency tables, including a covariate. The proposed Bayesian model extends the binomial-beta hierarchical model developed by KING, ROSEN and TANNER (1999) from the 2×2 case to the $R \times C$ case. As in the 2×2 case, the inferential procedure employs Markov chain Monte Carlo (MCMC) methods. As such, the resulting MCMC analysis is rich but computationally intensive. The frequentist approach, based on first moments rather than on the entire likelihood, provides quick inference via nonlinear least-squares, while retaining good frequentist properties. The two approaches are illustrated with simulated data, as well as with real data on voting patterns in Weimar Germany. In the final section of the paper we provide an overview of a range of alternative inferential approaches which trade-off computational intensity for statistical efficiency.

Key Words and Phrases: ecological inference, Bayesian inference, frequentist inference, voting patterns.

1 Introduction to ecological inference

Ecological inference is the process of inferring discrete individual behavior from data available on groups of individuals – or, more precisely, learning about the cells of p cross-tabulations from the observed marginal totals in each. Ecological inference is a special, and somewhat easier, case than general issues of aggregation, which can span

* ori@stat.pitt.edu

For research assistance, we thank Lara Birk, Albert Diversé, Kira Foerster, and Ethan Katz.

continuous individual variables, time series problems, and many other issues. However, in all aggregation problems, including ecological inference, assumptions (and when possible inferences) about the process by which data get aggregated substitute for information lost in the aggregation process.

When a valid sample of individuals is available, direct inference about the cells of the population cross-tabulation is possible and so ecological inference is unnecessary (and of course when a census is available, direct calculation makes any type of inference unnecessary). Unfortunately, in many academic fields and areas of public policy, the only available information is ecological. For example, political scientists have sought for half a century to learn how and why the democratic institutions of Weimar Germany led to the popular election of one of the most destructive regimes in history. Since no surveys were conducted, a key part of this inquiry involves using ecological data to make inferences about who voted for the Nazi Party. We analyze these data in this paper.

For another example, in applying the U.S. Voting Rights Act, the law requires information about whether African Americans and other ethnic minorities vote differently than the white majority. Since political polls about racially divisive contests are notoriously untrustworthy, even when they are available, one must draw inferences about quantities like the fraction of African Americans voting for the Democrats from separate Census data on the numbers of blacks and electoral data on the number of people voting for the Democrats. The secret ballot prevents one from computing the cross-tabulation directly.

The inherent difficulties of making ecological inferences, combined with their necessity in these fields, and others such as epidemiology, marketing, campaign targeting, education, sociology, and political science, should keep ecological inference on the agenda for some time. The statistical field dates to OGBURN and GOLTRA (1919) and GEHLKE (1917), who, in the context of studies of political behavior, first recognized the ecological inference problem, and ROBINSON (1950), who first formalized it. GOODMAN (1953, 1959) proposed the first statistical approach to the problem; DUNCAN and DAVIS (1953) offered the first formal deterministic method; and KING (1997) gave the first model that combined available statistical and deterministic information from both approaches. KING, ROSEN, and TANNER (1999), who also combined statistical and deterministic information, offered the first hierarchical Bayesian approach to ecological inference.

The ecological inference problem has been seen as difficult enough that most analysts have focused on the 2×2 case. Of the methods that have been proposed for larger tables, only direct generalizations of GOODMAN (1953, 1959) have been used much in real applications. Unfortunately, Goodman's statistical approach in the $R \times C$ case usually gives at least some out of bounds point estimates, and always implies a posterior distribution with support over impossible parts of the parameter space.

2 Hierarchical models for the 2 × 2 case

KING, ROSEN and TANNER's (1999) model uses Markov chain Monte Carlo (MCMC) methods. They describe the basic probem in terms of an example from KING (1997). Specifically, for each of p electoral precincts, two variables are observed, the fractions of voting-age people who turn out to vote (T_i) and the fraction of voting-age people who are black (X_i), along with the number of voting age people (N_i). In addition, there are two unobserved quantities of interest, the fraction of blacks who vote (β_i^b) and the fraction of whites who vote (β_i^w). Denoting the number of voting-age people who turn out to vote by T_i', KING *et al.* (1999) propose hierarchical Bayesian models, at the top level of which T_i' is assumed to follow a binomial distribution with probability equal to $\theta_i = X_i\beta_i^b + (1 - X_i)\beta_i^w$ and count N_i. At the second level of the hierarchy, it is assumed 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. Using the beta family of distributions which is quite rich allows one to relax the single-cluser assumption in KING (1997). At the third and final level, KING *et al.* (1999) assume that the unknown parameters c_b, d_b, c_w and d_w follow exponential distributions. Given this three-stage model, these authors construct the posterior distribution of all the parameters. Using MCMC methods, they estimate the unknown quantities, based on samples drawn from the posterior distribution.

KING *et al.* (1999) expand the above-described hierarchical model by incorporating covariates (Z_i). In particular, at the second stage of the model, they 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 sample independently from a beta distribution with parameters $d_w \exp(\gamma + \delta Z_i)$ and d_w. This assumption implies that the log odds for blacks depend linearly on the covariate Z_i, with intercept α and slope β. Analogously, the log odds for whites depend on Z_i with intercept γ and slope δ. Incorporating covariates allows the distribution to be more flexible by allowing more complicated shapes of densities. At the third level of the hierarchical model with covariates, the regression parameters are assumed to be a priori independent with a flat prior on these parameters, whereas d_b and d_w are assumed to follow exponential distributions. Having drawn samples from the posterior distribution of the parameters, KING *et al* (1999) are able to assess the significance of the covariate.

3 Hierarchical models for the $R \times C$ case – Bayesian inference

3.1 *The multinomial-Dirichlet hierarchical model*

The binomial-beta model with covariates, proposed by KING *et al.* (1999) and summarized in Section 2, is extended in this section to a multinomial-Dirichlet model, allowing for analysis of $R \times C$ tables ($C > 2$ and $R \geqslant 2$), with covariates. We introduce the ecological inference problem in the $R \times C$ case with an example on voting patterns in Weimar Germany (see Table 1). (This example will be considered in detail in Section 5.) The data for this example consist of a 4 × 4 table for each of

Table 1. Notation for Kreis i in a 4 × 4 table

Social Class	Voting Decision				
	Far-Left or Left	Far-Right	NSDAP	Other	
White Collar	β^i_{11}	β^i_{12}	β^i_{13}	$1-\sum_{c=1}^{3}\beta^i_{1c}$	X_{1i}
Blue Collar	β^i_{21}	β^i_{22}	β^i_{23}	$1-\sum_{c=1}^{3}\beta^i_{2c}$	X_{2i}
Unemployed	β^i_{31}	β^i_{32}	β^i_{33}	$1-\sum_{c=1}^{3}\beta^i_{3c}$	X_{3i}
Others	β^i_{41}	β^i_{42}	β^i_{43}	$1-\sum_{c=1}^{3}\beta^i_{4c}$	$1-\sum_{r=1}^{3}X_{ri}$
	T_{1i}	T_{2i}	T_{3i}	$1-\sum_{c=1}^{3}T_{ci}$	

$p = 695$ German Kreise (electoral precincts). For Kreis (the singular of Kreise) $i (i = 1, \ldots, p)$, we observe the fractions of voting-age people who turn out to vote for specific parties $(T_{1i}, \ldots, T_{C,i})$ and the fractions of voting-age people in different social classes $(X_{1i}, \ldots, X_{R,i})$. The unobserved quantities $(\beta^i_{rc}, r = 1, \ldots, R, c = 1, \ldots, C-1)$ are the fractions of people in social class r, who vote for party c.

To describe the multinomial-Dirichlet model, suppose there are p Kreise. Let $T'_i = (T'_{1i}, T'_{2i}, \ldots, T'_{Ci})$ be the numbers of voting-age people who turn out to vote for the different parties. At the first stage of the hierarchy we assume that $\mathbf{T}'_i$ follows a multinomial distribution with parameter vector $\boldsymbol{\theta}_i = (\theta_{1i}, \theta_{2i}, \ldots, \theta_{C,i})^t$ and count N_i, where θ_{ci} equals $\sum_{r=1}^{R}\beta^i_{rc}X_{ri}$ for $c = 1, \ldots, C$, under the constraint that $\sum_{c=1}^{C}\theta_{ci} = 1$. Note that the θ_{ci}'s and the β^i_{rc}'s depend on a covariate Z_i in a manner to be specified at the second stage. It therefore follows that the contribution of the data of Kreis i to the likelihood is

$$\theta_{1i}^{T'_{1i}} \times \ldots \times \theta_{C-1,i}^{T'_{C-1,i}} \times (1 - \sum_{c=1}^{C-1}\theta_{ci})^{N_i - \sum_{c=1}^{C-1} T'_{ci}}.$$

At the second stage of the hierarchical model, we assume that the vectors which we denote $\boldsymbol{\beta}^i_r = (\beta_{r1}, \beta_{r2}, \ldots, \beta_{r,C-1})^t$ $(i = 1, \ldots p, r = 1, \ldots, R)$ follow independent Dirichlet distributions with parameters $(d_r \exp(\gamma_{r1} + \delta_{r1}Z_i), \ldots, d_r \exp(\gamma_{r,C-1} + \delta_{r,C-1}Z_i), d_r)$. Thus, the second-stage means of the β^i_{rc}'s are

$$\frac{d_r \exp(\gamma_{rc} + \delta_{rc}Z_i)}{d_r\left(1 + \sum_{j=1}^{C-1}\exp(\gamma_{rj} + \delta_{rj}Z_i)\right)} = \frac{\exp(\gamma_{rc} + \delta_{rc}Z_i)}{1 + \sum_{j=1}^{C-1}\exp(\gamma_{rj} + \delta_{rj}Z_i)}, \tag{1}$$

for $i = 1, \ldots, p$, $r = 1, \ldots, R$ and $c = 1, \ldots, C-1$, which implies that

$$\log\frac{E(\beta^i_{rc})}{E(\beta^i_{rC})} = \gamma_{rc} + \delta_{rc}Z_i. \tag{2}$$

In other words, the log odds depend linearly on the covariate Z_i. We will refer to $E(\beta_{rc})/E(\beta_{rC})$ as the odds, although this term usually describes the ratio of the fractions β_{rc}/β_{rC}, and not the ratio of the expected fractions.

At the third and final stage, we treat the regression parameters (the γ_{rc}'s and the δ_{rc}'s) to be a priori independent, putting a flat prior on these regression parameters. The parameters d_r, $r = 1, \ldots, R$, are assumed to follow exponential distributions with means $1/\lambda$.

3.2 *Inference via Markov Chain Monte Carlo Methods*

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

$$\begin{aligned} & p(\text{data}|\boldsymbol{\beta}_i,\ i = 1, \ldots, p) \times p(\boldsymbol{\beta}_i,\ i = 1, \ldots, p|\boldsymbol{\delta}) \times p(\boldsymbol{\delta}) \\ & \quad = \prod_{i=1}^{p}\prod_{c=1}^{C}\theta_{ci}^{T'_{ci}} \times \prod_{i=1}^{p}\prod_{r=1}^{R} \\ & \qquad \times \left\{ \frac{\Gamma\left(d_r \sum_{c=1}^{C} \exp(\gamma_{rc} + \delta_{rc} Z_i)\right)}{\prod_{c=1}^{C} \Gamma(d_r \exp(\gamma_{rc} + \delta_{rc} Z_i))} \prod_{c=1}^{C} {\beta_{rc}^{i}}^{d_r \exp(\gamma_{rc} + \delta_{rc} Z_i) - 1} \right\} \\ & \qquad \times \exp\left(-\lambda \sum_{r=1}^{R} d_r\right), \end{aligned} \tag{3}$$

where $\boldsymbol{\beta}_i$ denotes all the β_{rc}^{i}'s in the ith Kreis and $\boldsymbol{\delta} = (\gamma_{rc},\ \delta_{rc},\ d_r)_{r,c=1,1}^{R,C-1}$. Obtaining the marginals of this posterior distribution using high-dimensional numerical integration is infeasible. Instead we use the Gibbs sampler (TANNER, 1996). To implement the Gibbs sampler, we need the following conditional distributions; that is, we need the distribution of each unknown parameter conditional on the full set of the remaining parameters

$$p(\beta_{rc}^{i}|\{\beta_{jk}^{i}\}_{j\neq r}^{k\neq c},\ d_r,\ \gamma_{rc},\ \delta_{rc}) \propto \theta_{ci}^{T'_{ci}} \times \theta_{Ci}^{T'_{Ci}} \times {\beta_{rc}^{i}}^{d_r \exp(\gamma_{rc} + \delta_{rc} Z_i) - 1} \times {\beta_{rC}^{i}}^{d_r - 1}$$

$$\begin{aligned} & p(\gamma_{rc}|\{\beta_{rc}^{i}\}_{i=1}^{p},\ d_r,\ \{\delta_{rc}\}_{c=1}^{C-1},\ \{\gamma_{rj}\}_{j\neq c}) \\ & \quad \propto \prod_{i=1}^{p} \frac{\Gamma\left(d_r \sum_{c=1}^{C} \exp(\gamma_{rc} + \delta_{rc} Z_i)\right)}{\Gamma(d_r \exp(\gamma_{rc} + \delta_{rc} Z_i))} {\beta_{rc}^{i}}^{d_r \exp(\gamma_{rc} + \delta_{rc} Z_i)} \end{aligned}$$

$$\begin{aligned} & p(\delta_{rc}|\{\beta_{rc}^{i}\}_{i=1}^{p},\ d_r,\ \{\delta_{rj}\}_{j\neq c},\ \{\gamma_{rc}\}_{c=1}^{C-1}) \\ & \quad \propto \prod_{i=1}^{p} \frac{\Gamma\left(d_r \sum_{c=1}^{C} \exp(\gamma_{rc} + \delta_{rc} Z_i)\right)}{\Gamma(d_r \exp(\gamma_{rc} + \delta_{rc} Z_i))} {\beta_{rc}^{i}}^{d_r \exp(\gamma_{rc} + \delta_{rc} Z_i)} \end{aligned}$$

$$
\begin{aligned}
&p(d_r | \{\beta^i_{rc}\}_{(c,i)=(1,1)}^{(C-1,p)}, \{\gamma_{rc}\}_{c=1}^{C-1}, \{\delta_{rc}\}_{c=1}^{C-1}) \\
&\quad \propto \prod_{i=1}^{p} \left\{ \frac{\Gamma(d_r \sum_{c=1}^{C} \exp(\gamma_{rc} + \delta_{rc} Z_i))}{\prod_{c=1}^{C} \Gamma(d_r \exp(\gamma_{rc} + \delta_{rc} Z_i))} \prod_{c=1}^{C} \beta^{i}_{rc}{}^{d_r \exp(\gamma_{rc}+\delta_{rc} Z_i)} \right\} \\
&\quad \times \exp(-\lambda d_r).
\end{aligned}
$$

To generate a Gibbs sampler (Markov) chain, one draws a random deviate from each of these full conditionals, in turn updating the values of the variable after each draw. Unfortunately none of these distributions are standard distributions for which prewritten sampling subroutines are available. For this reason, we use the Metropolis algorithm (Metropolis *et al.*, 1953) to sample from each of these distributions. Thus, to sample a value for γ_{rc}, δ_{rc} or d_r, 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_{rc}, a candidate value for the next point in the Metropolis chain is drawn from the uniform distribution. Note that since $\sum_{c=1}^{C} \beta^i_{rc} = 1$, for each r, a value β^i_{rc} is sampled from a uniform distribution over the interval $[0, 1 - \sum_{j \neq c} \beta^i_{rj}]$. The candidate value is then accepted or rejected according to the Metropolis scheme of evaluating $\pi(y^*)/\pi(y)$, the ratio of the full conditional at the candidate value to the full conditional evaluated at the current point in the chain. With probability $\alpha(y, y^*) = \min\{\pi(y^*)/\pi(y), 1\}$ the candidate value is accepted; otherwise it is rejected (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 three 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.

3.3 *Simulated examples*

To illustrate the Bayesian methodology, we now consider two simple examples. In the first example, data from 150 precincts consisting of 3 × 3 tables were simulated from the hierarchical model. In all examples in this section λ was taken equal to 0.5. In addition, an independent normal random deviate was generated for each precinct with true δ_{rc}'s equal to zero. Clearly, in such a situation, one would expect the methodolgy to recognize that the covariate information is irrelevant. Figure 1 presents the posterior distributions of the δ_{rc}'s, the slope parameters for regressing

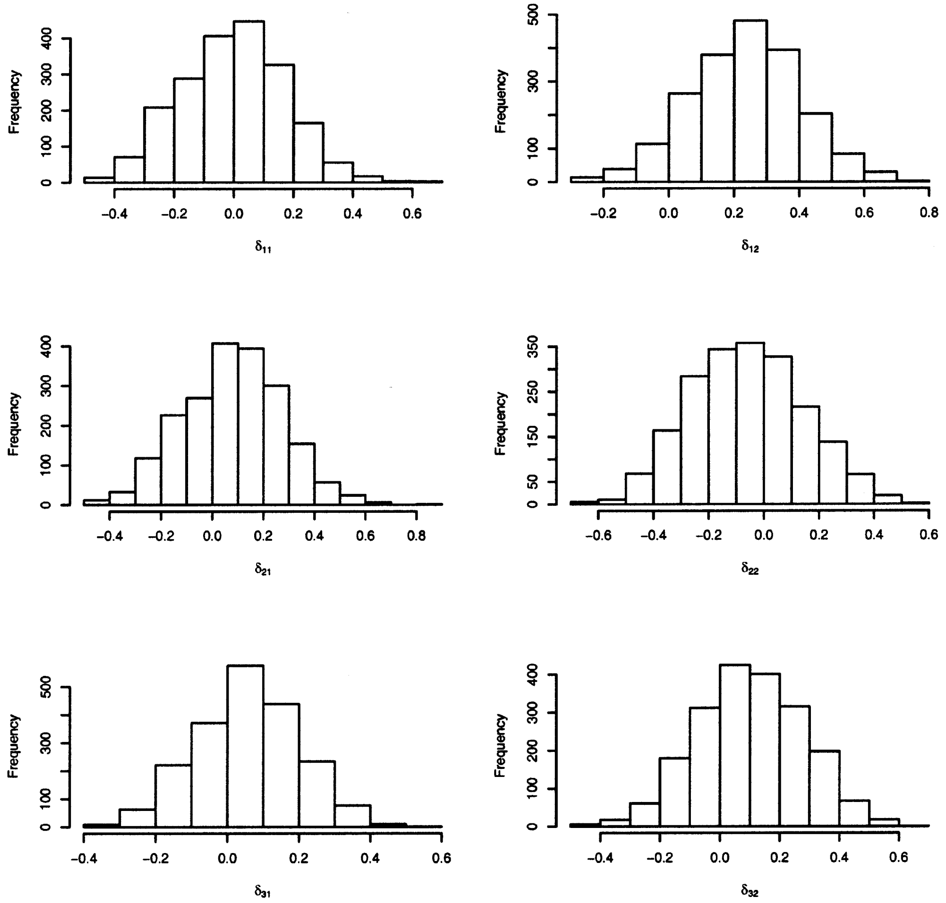

Fig. 1. Posterior distributions of the slope parameters (δ_{rc}'s) with an unimportant covariate

the log odds (2) on the covariate. The chains were iterated 300,000 times, with the results based on the final 200,000 iterates. It is seen that zero is a plausible value in all of the histograms. Table 2 presents the posterior mean and standard deviation for each of the regression parameters.

As a second example of using Bayesian inference, we consider a situation in which the covariate is important. The data were generated from the hierarchical model, but this time the covariate is related to the log odds according to equation (2). Figure 2 shows the posterior distributions of the slope parameters (δ_{rc}'s). For this example, it is seen that the bulk of the mass of the marginal posterior distributions are removed from zero indicating the importance of the covariate. Table 3 presents the posterior means and standard deviations of the regression parameters.

Table 2. Posterior means and standard deviations of the regression parameters for the simulated data with an unimportant covariate

Parameter	Posterior		Parameter	Posterior	
	Mean	SD		Mean	SD
γ_{11}	−0.227	0.284	δ_{11}	−0.002	0.174
γ_{12}	0.508	0.270	δ_{12}	0.235	0.169
γ_{21}	−0.307	0.261	δ_{21}	0.079	0.194
γ_{22}	0.649	0.282	δ_{22}	−0.061	0.203
γ_{31}	−0.750	0.209	δ_{31}	0.057	0.142
γ_{32}	−0.761	0.232	δ_{32}	0.102	0.178

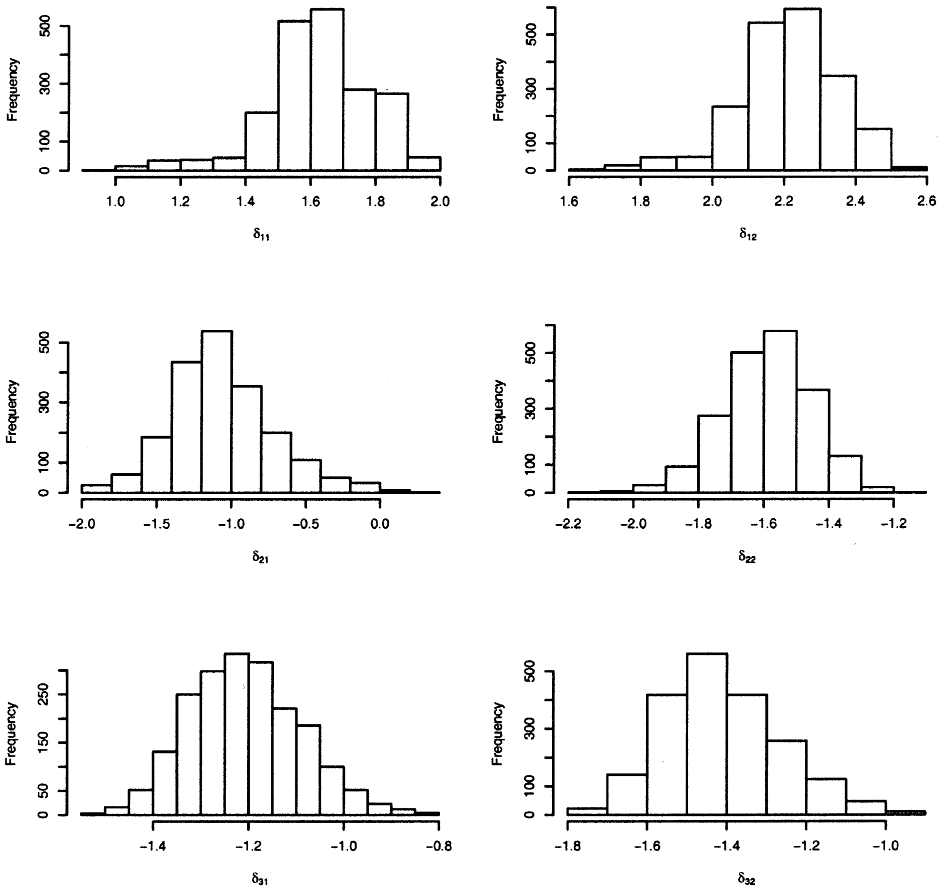

Fig. 2. Posterior distributions of the slope parameters (δ_{rc}'s) with an important covariate

Table 3. Posterior means and standard deviations of the regression parameters for the simulated data with an important covariate

Parameter	Posterior		Parameter	Posterior	
	Mean	SD		Mean	SD
γ_{11}	0.288	0.182	δ_{11}	1.623	0.161
γ_{12}	0.421	0.163	δ_{12}	2.211	0.140
γ_{21}	−1.549	0.677	δ_{21}	−1.058	0.348
γ_{22}	0.923	0.161	δ_{22}	−1.590	0.134
γ_{31}	−0.768	0.155	δ_{31}	−1.207	0.117
γ_{32}	−1.466	0.207	δ_{32}	−1.410	0.146

4 Hierarchical models for the $R\times C$ case- nonlinear least-squares estimation

4.1 *Nonlinear least-squares estimation*

In the previous section, we considered full Bayesian inference. This approach based on MCMC methods yields a rich analysis of the data. Unfortunately, the full Bayesian approach can be quite computationally intensive, and for complex models the assessment of convergence may not be straightforward. For these reasons, we now consider a simpler nonlinear least-squares approach which provides quick inference, while retaining good frequentist properties. We do not advocate one method above another. Rather, as discussed in Section 6, we feel that these methods are complementary and important tools in the analyses of ecological inference data. Although GOODMAN's (1953, 1959) method is also a least squares estimator, our nonlinear least-squares approach is an approximation to our MCMC method, not Goodman's. In particular, even though our method, like Goodman's, only gives estimates of the moments at present, it, unlike Goodman's, is consistent with a posterior that has positive density only in regions that are logically possible.

We introduce nonlinear least-squares estimation by first considering the general non-linear regression model, expressed as

$$Y = f(\mathrm{x}, \boldsymbol{\theta}) + \epsilon, \tag{4}$$

where Y is an observed response or dependent variable, $\mathbf{x}$ is a vector of independent variables, ϵ is an unobserved error with mean zero, and $\boldsymbol{\theta}$ is a vector of unknown parameters. The nonlinear function f is assumed to be known. Upon observing $(\mathrm{x}_1, Y_1), \ldots, (\mathrm{x}_n, Y_n)$, nonlinear least-squares estimation of $\boldsymbol{\theta}$ consists of minimizing the sum of squared deviations

$$\mathrm{SSE}(\boldsymbol{\theta}) = \sum_{i=1}^{n}(Y_i - f(\mathrm{x}_i, \boldsymbol{\theta}))^2 \tag{5}$$

over all possible $\boldsymbol{\theta}$'s. Note that unlike the case in linear regression, minimizing (5) requires iterative procedures. In (5), $f(\mathrm{x}_i, \boldsymbol{\theta})$ is the mean or the first moment of Y_i, according to model (4), and $(Y_i - f(\mathrm{x}_i, \boldsymbol{\theta}))^2$ is the squared Euclidean distance between Y_i and $f(\mathrm{x}_i, \boldsymbol{\theta})$. The parameter vector $\boldsymbol{\theta}$ characterizes the mean function. In the next section, this nonlinear least-squares methodology is used to etimate the mean parameters (the γ_{rc}'s and the δ_{rc}'s) and to test the significance of the covariate. This method does not require simulation and is therefore faster than MCMC methods. A disadvantage of this method is that it does not yield estimates of the β^i_{rc}'s, however one can obtain an estimate of the mean of β^i_{rc} conditional on the value of a covariate.

4.2 *Nonlinear least squares for ecological inference*

The model considered in this section consists of the first two levels of the hierarchical model described in Section 3.1. At the first stage of the hierarchy we assume that $\mathbf{T}'_i$ follows a multinomial distribution with parameter vector $\boldsymbol{\theta}_i = (\theta_{1i}, \theta_{2i}, \ldots, \theta_{C,i})^t$ and count N_i, where θ_{ci} equals $\sum_{r=1}^{R} \beta^i_{rc} X_{ri}$ for $c = 1, \ldots, C$, under the constraint that $\sum_{c=1}^{C} \theta_{ci} = 1$. At the second stage of the hierarchical model, we assume that the mean of β^i_{rc} is equal to $\exp(\gamma_{rc} + \delta_{rc} Z_i)/(1 + \sum_{j=1}^{C-1} \exp(\gamma_{rj} + \delta_{rj} Z_i))$, for $r = 1, \ldots, R$, $c = 1, \ldots, C-1$ and $i = 1, \ldots, p$.

For each table $i = 1, \ldots, p$, define $T_{ci} = T'_{ci}/N_i$, for $c = 1, \ldots, C$, and $\beta_i = (\beta^i_{rc})^{R,C}_{r,c=1,1}$ (subject to $\sum_{c=1}^{C} \beta^i_{rc} = 1$ for each r). Define also $\boldsymbol{\eta} = (\gamma_{rc}, \delta_{rc})^{R,C}_{r,c=1,1}$ (with constraints $\gamma_{rC} = \delta_{rC} = 0$). From the first level of the hierarchical model, it follows that $\mathbf{T}'_i | \boldsymbol{\beta}_i$ is Multinomial $\{N_i;\ E(T_{ci}|\boldsymbol{\beta}_i)^C_{c=1}\}$, where

$$E(T_{ci}|\boldsymbol{\beta}_i) = \sum_{r=1}^{R} X_{ri}\beta^i_{rc} \equiv \sum_{r=1}^{R} (X'_{ri}/N_i)\beta^i_{rc}.$$

However, to facilitate inference we want the expectation conditional on $\boldsymbol{\eta}$.

By iterating the expectation we have:

$$\begin{aligned} m^i_c(\boldsymbol{\eta}) &= E(T_{ci}|\boldsymbol{\eta}) \\ &= E\{E(T_{ci}|\boldsymbol{\beta}_i, \boldsymbol{\eta})|\boldsymbol{\eta}\} \\ &= E\left(\sum_{r=1}^{R} X_{ri}\beta^i_{rc}|\boldsymbol{\eta}\right) \\ &= \sum_{r=1}^{R} X_{ri} E(\beta^i_{rc}|\boldsymbol{\eta}) \end{aligned}$$

for $c = 1, \ldots, C-1$ and $i = 1, \ldots, p$. The $m^i_c(\boldsymbol{\eta})$'s ($c = 1, \ldots, C-1$) are mean functions, and $E(\beta^i_{rc}|\boldsymbol{\eta})$ is given in (1). The least-squares approach consists of solving

$$\min_{\boldsymbol{\eta}} \sum_{i=1}^{p} \sum_{c=1}^{C-1} (T_{ci} - m^i_c(\boldsymbol{\eta}))^2 \tag{6}$$

to obtain the estimate of $\boldsymbol{\eta}$. The method of inference adopted in this section is based on first moments. As seen in equation (1), while the d_r's are used to model variance they do not appear in the mean function. As such, they are eliminated in this first moment inferential procedure. By including second order moment conditions in the estimation procedure, the efficiency may be increased and the d_r's estimated as a byproduct – see Section 6.

To obtain the standard errors of the parameter estimates, let $SS = \sum_{i=1}^{p}\sum_{c=1}^{C-1}(T_{ci} - m_c^i(\boldsymbol{\eta}))^2 = \sum_{i=1}^{p} ss_i$ and define $\mathbf{G} = \nabla(SS)$, where ∇ is the gradient with respect to $\boldsymbol{\eta}$. It then follows that the asymptotic variance of $\hat{\boldsymbol{\eta}}$ is

$$avar(\hat{\boldsymbol{\eta}}) = \{\nabla\mathbf{G}(\hat{\boldsymbol{\eta}})^t\hat{V}^{-1}\nabla\mathbf{G}(\hat{\boldsymbol{\eta}})\}^{-1}, \tag{7}$$

where $\hat{V} = \sum_{i=1}^{p}\mathbf{g}_i(\hat{\boldsymbol{\eta}})\mathbf{g}_i(\hat{\boldsymbol{\eta}})^t$. The vector $\mathbf{g}_i$ is the ith summand of G, that is, $\mathbf{g}_i = \nabla(ss_i)$. In Section 6, we will see that formula (7) is a general formula that can be used in the context of other moment-based estimation methods. The explicit formulas for the first and second derivatives are presented in Appendix 1.

Starting values for $\boldsymbol{\eta}$ can be obtained by recognizing that the mean function of T_{ci} forms a linear model with respect to the composite parameter $E(\beta_{rc}^i|\boldsymbol{\eta})$. To estimate the composite parameters, the covariate Z_i is first discretized into a number of strata, and the T_{ci}'s are then regressed on the predictors X_{ri}, $r = 1, \ldots, R$, within each stratum. A good set of starting values can then be obtained by solving for $\boldsymbol{\eta}$ from the composite parameter estimates. An alternative method (JENNRICH, 1969, p. 642) involves a random search over p values of $\boldsymbol{\eta}$ and uses the one that minimizes the least squares objective function, as the starting value.

The following proposition provides regularity conditions under which the least-squares estimator $\hat{\boldsymbol{\eta}}$ is consistent for $\boldsymbol{\eta}$ and asymptotically normal with mean $\boldsymbol{\eta}$ and variance-covariance given by (7). The result is most conveniently formulated by treating $(\mathbf{T}_i^t, \mathbf{X}_i^t, Z_i)_{i=1}^p$ as independent and identically distributed random vectors, where $\mathbf{T}_i = (T_{ci})_{c=1}^C$ and $\mathbf{X}_i = (X_{ri})_{r=1}^R$. Denote $\mathbf{T}_i^{red} = (T_{ci})_{c=1}^{C-1}$.

PROPOSITION 1 *Suppose that* $var(\mathbf{T}_i^{red}|\mathbf{X}_i Z_i)$ *is nonsingular for all* $\mathbf{X}_i$ *and* Z_i. *Suppose* $var(Z_i) > 0$. *Then, for almost all true parameters* $\boldsymbol{\eta}$ *in the Lebesgue sense, there exists a sequence of local least square estimators* $\hat{\boldsymbol{\eta}}_p$ *such that*

$$avar(\hat{\boldsymbol{\eta}}_p)^{1/2}(\hat{\boldsymbol{\eta}}_p - \boldsymbol{\eta}) \xrightarrow{d} N(\mathbf{0}, I)$$

as $p \to \infty$.

PROOF. This proposition is proved in a similar manner to the consistency and asymptotic normality of the local maximum likelihood estimates [see, e.g., SERFLING (1980, Theorem 4.2.2, use the multi-dimensional generalization), or REDNER and WALKER (Theorem 3.1, 1984)], in which regularity conditions of three types are involved: smoothness conditions, integrability conditions, and local identification. The only nontrivial condition in the current context is the local identification-type condition, as ensured by the nonsingularity of the expected Hessian $\mathscr{I}^*(\boldsymbol{\eta}) = -E\nabla\nabla(SS)$, together with the nonsingularity of the variance of the 'score vector'

$\mathcal{V}(\boldsymbol{\eta}) = var\{\nabla(SS)\}$. (Note that unlike the likelihood situation, here $\mathscr{T}^*(\boldsymbol{\eta})$ is not proportional to $\mathcal{V}(\boldsymbol{\eta})$ in general.) Using the analyticity of $\mathscr{T}^*(\cdot)$ it is easy to show that as a consequence of a nonzero $var(Z_i)$, the set A must have zero Lebesgue measure, where A is the set of $\boldsymbol{\eta}$ giving a zero $\det\{\mathscr{T}^*(\boldsymbol{\eta})\}$. This together with the nonsingularity of $var(\mathbf{T}_i^{red}|\mathbf{X}_i, Z_i)$ also ensure the nonsingularity of $\mathcal{V}(\boldsymbol{\eta})$ for almost all $\boldsymbol{\eta}$. □

Comments:

1. The nonsingularity of $var(\mathbf{T}_i^{red}|Z_i, \mathbf{X}_i)$ implies that there should not exist any further constraints on the components of $\mathbf{T}_i$, except that all its components add to one. This condition is needed to guarantee the nonsingularity of the asymptotic variance.
2. The condition $var(Z_i) > 0$ is also necessary, since without this condition the regression coefficients would not be identifiable. It is noted in the 2×2 case when $Z = X$ that when the tomography lines have very similar slopes, by the formulation of the model, X is a near constant. In such a case, the slopes and intercepts of the model will be poorly identified.
3. An interesting situation is when a component of $\mathbf{X}_i$ is taken as a covariate. For example, in the 2×2 case let $Z = X$. It then can be shown that the parameters are not locally identifiable (with a singular $\mathscr{T}^*(\boldsymbol{\eta})$) on a set of measure zero of true parameters. Specific elements of this set include: (i) $\delta_{11} = \delta_{21} = 0$, and (ii) $(\gamma_{11}, \delta_{11}) = (\gamma_{21}, \delta_{21})$. In these cases the asymptotic normality results fail. Note that this is a clear advantage of the model logit $E(\beta_{r1}|X, \boldsymbol{\eta}) = \gamma_{r1} + \delta_{r1}X$, over the model $E(\beta_{r1}|X, \boldsymbol{\eta}) = \gamma_{r1} + \delta_{r1}X$ – the latter would have nonidentifiable parameter values *everywhere*, as is easy to see from the nonidentifiability of the parameters in the mean function $E(T_1|X, \boldsymbol{\eta})$, or by checking the Hessian $\mathscr{T}^*(\boldsymbol{\eta})$. For example, note that
$$\begin{aligned} E(T_1|X, \boldsymbol{\eta}) &= X(\gamma_{11} + \delta_{11}X) + (1 - X)(\gamma_{21} + \delta_{21}X) \\ &= \gamma_{21} + (\gamma_{11} - \gamma_{21} + \delta_{21})X + (\delta_{11} - \delta_{21})X^2, \end{aligned}$$
which involves three coefficients of the powers of X but *four* parameters; and that $\mathscr{T}^*(\boldsymbol{\eta}) \propto E(vv^t)$ where $v = [X, X^2, 1 - X, (1 - X)X]^t$ is a set of linearly dependent random vectors.
4. As an implicaton of the previous comment, the power of a test against the null hypothesis of a zero covariate effect ($\delta_{11} = \delta_{21} = 0$) will be unity in the large sample (p) limit, for almost all true parameters, in the present approach.
5. For the purpose of testing the null hypothesis of no covariate effect, we suggest using the working model logit $E(\beta_{r1}|X, \boldsymbol{\eta}) = \gamma_{r1} + \delta_{r1}h(X)$ and test $\delta_{r1} = 0$, by, for example, the Wald statistic. Here $h(X)$ is a nonlinear function of X, which could be logit$(X) \equiv \log(X/(1 - X))$, or $I[X > 0.5]$. This corresponds to using the transformed covariate $Z = h(X)$, which makes the expected Hessian $\mathscr{T}^*(\boldsymbol{\eta})$

nonsingular. This model does not suffer from nonidentifiability at the null value, guarantees the asymptotic normality under the null hypothesis, and therefore also guarantees the correct asymptotic type-one error rate.

6. It is noted that the local nonidentifiability is associated with the use of the mean function alone in the least-squares procedure. It is potentially possible to avoid such problems by augmenting the least-squares method and modeling the second moment of β_i or $\mathbf{T}_i$ in addition to the first moment.

4.2 *Simulated examples*

In this section we again use the simulated data sets described in Section 3.3, to demonstrate the nonlinear least-squares methodology. The point estimates along with approximate standard errors, computed according to (7), are presented in Table 4 and Table 5 for the unimportant and important covariate, respectively. Note that with the exception of δ_{21}, the regression parameters in Table 4 are not significantly different from zero. The least-squares standard errors are larger than the corresponding values obtained in the full Bayesian approach (see Table 2). This is expected – the data were generated according to the hierarchical model and the full Bayesian approach which assumes this model will be more efficient than the first moment approach. As discussed further in Section 6, the first moment approach is expected to be more robust against departures from the full hierarchical model. By modeling the variance

Table 4. Nonlinear least-squares point estimates and standard errors of the regression parameters for the simulated data with an unimportant covariate

Parameter	Point Estimate	SE	Parameter	Point Estimate	SE
γ_{11}	0.581	0.522	δ_{11}	−0.271	0.462
γ_{12}	0.937	0.450	δ_{12}	0.065	0.379
γ_{21}	−1.815	1.391	δ_{21}	1.774	0.821
γ_{22}	0.849	0.413	δ_{22}	0.410	0.353
γ_{31}	−0.972	0.387	δ_{31}	0.103	0.375
γ_{32}	−1.200	0.363	δ_{32}	0.221	0.442

Table 5. Nonlinear least-squares point estimates and standard errors of the regression parameters for the simulated data with an important covariate

Parameter	Point Estimate	SE	Parameter	Point Estimate	SE
γ_{11}	0.466	0.180	δ_{11}	1.563	0.211
γ_{12}	0.347	0.210	δ_{12}	2.259	0.179
γ_{21}	−2.788	0.874	δ_{21}	−2.539	0.587
γ_{22}	1.011	0.182	δ_{22}	−1.625	0.169
γ_{31}	−0.805	0.152	δ_{31}	−1.011	0.181
γ_{32}	−1.598	0.285	δ_{32}	−1.470	0.282

one can potentially realize a more efficient estimator, i.e. smaller standard error (see Section 6).

Examining Table 5, evidence of a covariate effect is seen. Again, the standard errors are usually larger than the corresponding values in the full Bayesian case. Interestingly, comparing the least-squares estimates of the slopes and the corresponding full Bayesian values to the true values of $(\delta_{11}, \delta_{12}, \delta_{21}, \delta_{22}, \delta_{21}, \delta_{32}) = (1.5, 2.0, -1.3, -1.5, -1.0, -1.2)$, neither approach is seen to dominate uniformly.

5 Applications to voting patterns in Weimar Germany

5.1 *Overview of the data*

We began with aggregate data from FALTER and HÄNISCH (1989; see also HÄNISCH, 1989) on 1932 electoral results in voting for the German Reichstag, and census data on social occupations and religious denominations, in the Weimar Republic. The observations include 1246 contiguous geographic units, with many changes over time, resulting in only about 1000 towards the end. We aggregated these units into 695 Kreise that, to the extent possible, remained stable over the entire period. This facilitated comparisons over time, made possible the comparison of the electoral data with social data from German census, and enabled exploratory analyses via geographic mapping with available computerized boundary files. Our resulting 695 observations tile the country (with the minor exceptions of one tiny area in Prussia due to the absence of data, and the 'Saarland', becaue it was occupied by France.) We aggregated the data by hand with the help of OSS map 6289 'Greater Germany – Kreis Boundaries July 1, 1944,' and *Germany* by C. S. HAMMOND & CO. (N.Y. 1924), and by studying population changes over time.

The political parties that ran for office in the 1932 election include the *Far Left* (KPD, the communist party); the parties in the *Government*, which we grouped together, including primarily the Left (SPD, the Social Democrats) and the Catholic and Center parties (Zentrum, BVP, and others); the *Far Right* (DNVP, a small party preferring a return to monarchy), and the *Nazi Party* (NSDAP, the National Socialists). We grouped several very small parties (such as DVP, DSTP) together with *nonvoters*. Alternative analyses could be based on different groupings.

From the German Census, our occupational groups included the self-employed, white collar, blue collar, and the unemployed. We also included 'Others', which includes those outside the regular labor force, including domestic employees, helping family members, and the nonworking population, most of whom were on fixed incomes. We also have data on the fraction of the population in each Kreis that is Protestant (all but a few percent of non-Protestants are Catholic).

5.2 *Overview of the literature*

Most prior literature on Nazi voting behavior assumes or argues that the electorate was organized on the basis of social class. (This is no longer the view of modern voting behavior outside of this historical case, in part because of changes in electorates around the world, but mostly because of better information, largely from survey research, as well as from more successful theories.) Prior to the 1980s, the dominant view was that the economic depression led to 'middle class panic', and the Middle Class was the group that most believed gave the Nazis their strongest support. Then, in a major book, HAMILTON (1982) conducted a very careful analysis of about two dozen areas that gave nearly homogeneous voting support to the Nazis, examined the housing stock in each, and concluded that the upper classes (White Collar and Self-Employed) constituted the core elctoral support of the NSDAP. (Hamilton's 'homogeneous precinct analysis' is common in U.S. Voting Rights Act litigation and is a special case of DUNCAN and DAVIS' (1953) method of bounds.) Hamilton's major academic adversary is CHILDERS (1983), who ran GOODMAN's (1953) regression, and backed up his statistical analysis with archival research and summaries of contemporary accounts of the electoral campaign. Childers concluded that the main class basis of Nazi support came from the working classes (Blue Collar workers and the Unemployed).

Although Hamilton and Childers, and numerous other writers, have considered many other variables, such as religion, and numerous special cases and exceptions, most of the literature is focused on providing a single, uniform, class-based explanation for all of Germany.

5.3 *Empirical analyses*

The models we have developed here combine and generalize the statistical methods used by HAMILTON (1982) and CHILDERS (1983). Despite the goal of the literature, we see no reason to assume that the basis of Nazi voting was uniform across the entire country. Thus, we apply our model separately in Protestant areas (where the Nazis mainly campaigned) and Catholic areas (where the Church was dominant), and we also allow each area's results to differ by the degree of unemployment in the area by including unemployment as a covariate. Further modeling could incorporate spatial effects across Germany. The analyses presented below are not meant to represent a complete analysis of the data – they are to illustrate the proposed methodology.

We began by analyzing a reduced 4×4 version of the voting data using the Bayesian model, as well as the frequentist approach, and confirmed that both gave substantively similar results. In this reduced analysis, there seemed to be especially strong agreement between the predicted voting fractions corresponding to the two approaches. For this reason we only use the least-squares approach to analyze the more substantively meaningful 5×5 case, avoiding the intensive computations required in the MCMC analysis.

We now discuss the substantive results from our least squares analysis in the 5×5 case. We believe these to be consistent with much qualitative evidence about the Weimar Republic, and considerable prior research on voting behavior outside Germany, although they clearly contradict some of the central points in the literature on Nazi voting behavior. Table 6 and Figure 3 present predicted values of the voting

Table 6. Predicted voting fractions for each party, for low and high protestant percentages, conditional on social class, corresponding to covariate values Z (% unemployment): less than 10% (marked 10%), between 10 and 20% (marked 20%), and greater than 20% (marked 30%), based on least-squares analysis

	Catholic Areas					Protestant Areas				
	Far-Left	Govt.	Far-Right	NSDAP	No vote	Far-Left	Govt.	Far-Right	NSDAP	No vote
	$Z = 10\%$									
Self-employed	0.083	0.000	0.013	0.597	0.307	0.000	0.331	0.000	0.000	0.669
White Collar	0.019	0.415	0.064	0.501	0.002	0.188	0.082	0.149	0.349	0.232
Blue Collar	0.081	0.331	0.042	0.081	0.464	0.116	0.132	0.185	0.362	0.205
Unemployed	0.000	0.936	0.064	0.000	0.000	0.141	0.859	0.000	0.000	0.000
Others	0.000	0.761	0.000	0.000	0.239	0.000	0.090	0.021	0.888	0.002
	$Z = 20\%$									
Self-employed	0.000	0.000	0.000	0.666	0.334	0.000	0.145	0.048	0.433	0.375
White Collar	0.000	0.586	0.047	0.351	0.016	0.000	0.171	0.195	0.385	0.249
Blue Collar	0.131	0.550	0.064	0.125	0.130	0.182	0.167	0.039	0.393	0.220
Unemployed	0.318	0.241	0.033	0.018	0.390	0.232	0.768	0.000	0.000	0.000
Others	0.000	0.791	0.000	0.000	0.209	0.000	0.260	0.064	0.662	0.014
	$Z = 30\%$									
Self-employed	0.000	0.000	0.000	0.673	0.327	0.000	0.000	0.000	0.999	0.000
White Collar	0.000	0.680	0.028	0.202	0.090	0.000	0.272	0.196	0.327	0.205
Blue Collar	0.145	0.630	0.067	0.133	0.025	0.245	0.180	0.007	0.366	0.203
Unemployed	0.445	0.000	0.000	0.120	0.435	0.263	0.474	0.000	0.023	0.240
Others	0.000	0.755	0.077	0.000	0.169	0.000	0.486	0.129	0.318	0.066

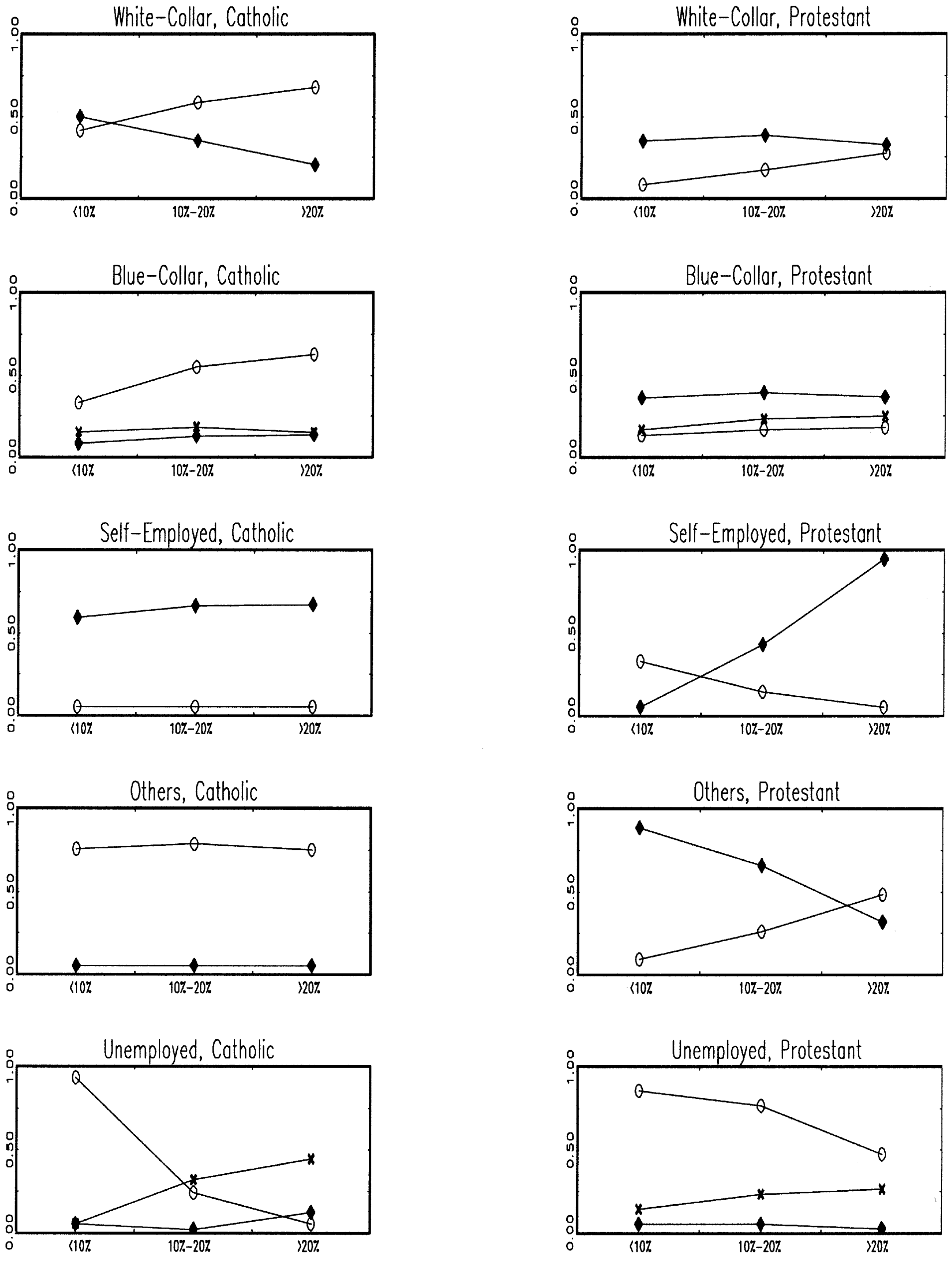

Fig. 3. Graphic representation of voting support from social groups for Major Parties in Table 6. The vertical axis of each graph is the proportion of the designated social group voting for the government parties (marked with an oval), the Nazi party (a solid diamond), and, where nontrivial, the Far Left (a cross). Nonvoters and smaller parties have been deleted, and lines were added to connect the symbols, for visual clarity. The horizontal axis in each graph is the percent unemployment in the Kreis (our covariate). The left column are for Catholic areas, and the right are for Protestant areas.

fractions for each party, conditional on the social class, corresponding to the trichotomized values of our covariate, percent unemployment (less than 10%, between 10 and 20% and greater than 20%). The predicted values were computed by plugging the least-squares estimates and the covariate values into (1). Note that the predicted values are not restricted to being a monotonic function of the covariate.

Although we find that there was indeed a coherent structure to voting behavior in Weimar Germany, the claim that the class structure had a uniform effect across the entire nation is unambiguously rejected. Most striking is the massive difference between the columns of Figure 6, representing Catholic (on the left) and Protestant (on the right) areas. Contrary to previous literature (although see FALTER, 1991), we find that voting behavior is not constant across these different regions for *any* social group.

Moreover, the results for each social group do not differ haphazardly, but rather are quite systematic and consistent with much qualitative evidence. They are not, however, consistent with either Hamilton's or Childers' analyses. We now discuss these results for each of the social groups.

With one exception, White and Blue Collar workers in Catholic areas form the core supporters of the government parties. These parties included the Catholic parties and so they had the advantages of both representing the policy interests of these constituents and providing the selective incentives from the Church's organizational efforts and social activities. The Nazi Party was also avowedly anti-Catholic, which also contributed to these results. The exception is the high White Collar support for the Nazis in low unemployment areas. We believe this anomaly is explained by a number of local political entrepreneurs who developed Nazi political organizations in Catholic areas to oppose the Catholic Church; HAMILTON's (1982, pp. 383–385) historical analysis reveals several significant examples of this, all of which we note were in areas of low unemployment. In Protestant areas, White and Blue Collar workers gave plurality support to the Nazis. This seems due to Nazi campaign strategy, which focused on these areas almost exclusively, and the absence of Church organizational efforts and selective incentives. Not surprisingly, Blue Collar workers also gave some support to the Far Laft party.

The self-employed (the third row in Figure 3) voted more for the NSDAP in areas of higher unemployement, although in Catholic areas with a much higher intercept (and always above 60%) and in Protestant areas with a much steeper slope (covering near the entire unit interval). Self-employed workers in low unemployment Protestant areas did not vote much for the Nazis, but they did not vote much for anyone (67% were nonvoters, according to Table 6). The self-employed were mostly small shop owners who feared the large department stores coming into town and taking over their business: 'In Nazi appeals to [self-employed] artisans and merchants, Jews were identified with those aspects of modern capitalism most repugnant to the old middle class – big business, the banks, and of course the department stores' CHILDERS (1983, p. 68). Self-employed support for the Nazis was also based on deep dislike of the government who they felt was taxing them disproportionately to pay for war reparations and social welfare benefits, and their belief that the electoral system should be organized by occupation, so they would feel better represented, rather than by geographic area. Furthermore, our auxiliary analyses (not shown) suggest that many self-employed in Catholic areas were Protestant.

The fourth row in Figure 3 portrays the votes of a category marked ‘Others’, which mostly includes people who depend on a fixed income; three-quarters of these people were receiving pensions or rent, more than half were older, and most were women (CHILDERS, 1983, p. 277). This group had little to fear of unemployemnt, and although they might fear inflation (as present day Social Security recipients in the U.S.), deflation characterized 1932 Germany. Given these powerful incentives to keep things as they are, these voters supported the status quo and hence the Government parties. The only exception was in Protestant areas with low unemployment, which was mostly agricultural east Prussia where the government only months before the election had ended agricultural subsidies and had ended the rule of the local Prussian state. Although the Nazis did not offer to reinstate the subsidies, they promised that no one would lose their land as a result.

The final category includes the unemployed, which are displayed in the last row of the figure. Like in most elections, the unemployed often did not vote, especially in areas of high unemployment. In this election, when they did vote, their choice seemed to be between the government and the Far Left parties, with the Nazis picking up only a small fraction of the vote. In both Catholic and Protestant areas, the vote for the government parties was lower, and for the Far Left was higher, in areas with higher unemployment. The unemployed in many societies have more important things to focus on than revolutionary change; they need a job. In Weimar Germany, when unemployed voters decided there was a need for a change, they usually cast their vote for the party designed for them, the Communists.

6 Discussion

In this paper we have used MCMC methods, as well as nonlinear least-squares for ecological inference. Inference via MCMC methods is based on a posterior distribution, whereas the nonlinear least-squares method is based on moments only. As we have seen, MCMC techniques are simulation-based methods and as such, are computationally intensive. Non-linear least-squares, on the other hand, do not involve simulation and therefore are faster. Another estimation method that competes with the full Bayesian approach is maximum likelihood. In the present situation, the likelihood function can be expressed as

$$L(\mathbf{T}_i'|\boldsymbol{\delta}) = \prod_{i=1}^{p} \int p(\mathbf{T}_i'|\boldsymbol{\beta}_i) p(\boldsymbol{\beta}_i|\boldsymbol{\delta}) d\boldsymbol{\beta}_i,$$

where $p(\mathbf{T}_i'|\boldsymbol{\beta}_i)$ is the multinomial distribution, and $p(\boldsymbol{\beta}_i|\boldsymbol{\delta})$ is the Dirichlet distribution. Thus, in the present situation, maximum likelihood requires integrating out the unobserved $\boldsymbol{\beta}_i$'s. Numerical integration is infeasible due to the high-dimensional integrals. Instead, one could sample from the Dirichlet distribution and use a mixture of multinomial distributions to approximate the integrals. Again, this method would be computationally intensive, though possibly less so than the fully Bayesian approach – especially if one used a simpler estimator (e.g. least-squares) as a good starting value for one- or two-step iteration of the maximum likelihood approach.

The least-squares approach is one representative of a whole host of moment-based methods. Since in our application, the moments are available in closed form, a large amount of computation is avoided. Other examples of moment-based methods include the generalized method of moments, popular in the econometrics literature (see MÁTYÁS, 1999), and the generalized estimating equations approach (see DIGGLE, LIANG and ZEGER, 1994), commonly used in biostatistics. All these methods are based on unbiased estimating functions, that is, estimating functions **G**(data; parameter) such that

$$E_{\text{data}|\text{parameter}}\mathbf{G}(\text{data};\ \ \text{parameter}) = \mathbf{0}$$

for all parameter values. Parameter estimates are obtained, in some sense, by minimizing a distance between **G** and **0**. In the special case, where the dimension of the estimating function matches the dimension of the parameter, the minimization is equivalent to solving the 'unbiased estimating equation' $\mathbf{G} = \mathbf{0}$ if a root exists. This special case has been widely used in the biostatistics literature (see DIGGLE *et al.*, 1994 or CARROLL, RUPPERT and STEFANSKI, 1995). There are many examples of such estimating equations: the moment equation for the method of moments, the normal equation for the method of least squares, the quasilikelihood equation (MCCULLAGH and NELDER 1989), the generalized estimating equations (DIGGLE *et al.*, 1994), or even the score equation for the maximum likelihood method. In general, the solution of this estimating equation is consistent and asymptotically normal, with mean equal to the true parameter and with variance typically equal to the 'sandwich formula' $\{\nabla\mathbf{G}^t var(\mathbf{G})^{-1}\nabla\mathbf{G}\}^{-1}$ (see CARROLL *et al.*, 1995).

The reason why estimating equation methods may reduce computation is that the estimating functions sometimes involve only analytically computable moments rather than the full probability function. These methods are also attractive due to their robustness against model misspecification. Often, inferential results such as consistency and asymptotic normality remain valid even with a misspecified probability model *p*(data|parameter), as long as the low-order moments are correctly modeled to make the estimating function unbiased. For illustration purposes we have presented an analysis based on nonlinear least squares, where the estimating functions are obtained from the gradients or normal equations. Alternatively, other methods could have been used, such as the generalized estimating equations approach or quasilikelihood or weighted least-squares, which would require a correct model for the variance/correlation structure to improve efficiency (DIGGLE *et al.*, 1994). These approaches can be viewed as intermediate approaches between the simple least-squares approach and the full Bayesian approach or even the maximum likelihood approach. All of these methods share similar theoretical properties, and the asymptotic variances can all be obtained from the same 'sandwich formula', with the corresponding estimating functions. In summary, the least-squares approach (as well as other higher moment-based approaches) has the following advantages:

1. It saves a large amount of computation, as there is no need to compute or simulate integrals.
2. The results are robust against departures from distributional assumptions.

The likelihood and the Bayesian methods, on the other hand, have the following advantages:

1. They may achieve higher efficiency (narrower frequentist or Bayesian confidence intervals) if the distributional assumptions are correctly specified.
2. They are required (either via empirical Bayes or via hierarchical Bayes) at the precinct level for computing the posterior probabilities for the β^i_{rc}'s.

The likelihood-based approaches and the moment-based approaches complement each other – the strengths of one are the weaknesses of the other, and vice versa.

Appendix 1: Calculation of derivatives

In this Appendix, we present the explicit formulas for the first and second derivatives required for the computation of the asymptotic variance of $\boldsymbol{\eta}$.

$$\nabla(ss_i) = -2\sum_{c=1}^{C-1}(T_{ci} - m^i_c(\boldsymbol{\eta}))\frac{\partial m^i_c(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}$$

$$\nabla\nabla(ss_i) = 2\sum_{c=1}^{C-1}\left[\frac{\partial m^i_c(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}\frac{\partial m^i_c(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^t} - (T_{ci} - m^i_c(\boldsymbol{\eta}))\frac{\partial^2 m^i_c(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}^t}\right]$$

$$\frac{\partial m^i_c(\boldsymbol{\eta})}{\partial a_{rk}} = \begin{cases} C^a_i X_{ri}\dfrac{\exp(\gamma_{rc} + \delta_{rc}Z_i)(1 + \sum_{j\neq c,C}\exp(\gamma_{rj} + \delta_{rj}Z_i))}{(1 + \sum_{j=1}^{C-1}\exp(\gamma_{rj} + \delta_{rj}Z_i))^2} & k = c \\ -C^a_i X_{ri}\dfrac{\exp(\gamma_{rc} + \delta_{rc}Z_i)\exp(\gamma_{rk} + \delta_{rk}Z_i)}{(1 + \sum_{j=1}^{C-1}\exp(\gamma_{rj} + \delta_{rj}Z_i))^2} & k \neq c, \end{cases}$$

where

$$C^a_i = \begin{cases} 1 & \text{if } a = \gamma \\ Z_i & \text{if } a = \delta. \end{cases}$$

Note that the hessian of $m^i_c(\boldsymbol{\eta})$ is dropped in the calculation of $\nabla\nabla(ss_i)$, since it is multiplied by $T_{ci} - m^i_c(\boldsymbol{\eta})$, which has a zero mean.

References

CARROLL, R. J., D. RUPPERT and L. A. STEFANSKI (1995), *Measurement error in nonlinear models*, Chapman and Hall, New York.

CHILDERS, T. (1983), *The Nazi voter: the social foundations of fascism in Germany, 1919–1933*, University of North Carolina Press, Chapel Hill.

COWLES, M. K. and B. CARLIN (1996), Markov chain Monte Carlo diagnostics: a comparative review, *Journal of the American Statistical Association* **91**, 883–904.

DIGGLE, P. J., K.-Y. LIANG and S. L. ZEGER (1994), *Analysis of longitudinal data*, Oxford University Press, New York.

DUNCAN, O. D. and B. DAVIS (1953), An alternative to ecological correlation, *American Sociological Review* **18**, 665–6.

FALTER, J. W. (1991), *Hitlers Wähler*, Beck, Müchen.

FALTER, J. W. and D. HÄNISCH (1989), Election and social data of the districts and municipalities of the German empire from 1920 to 1933, (Zentralarchiv study number 8013), http://.www.za.uni-koeln.de/.

GALLANT, A. R. and H. WHITE (1988), *A unified theory of estimation and inference for nonlinear dynamic models*, Basil Blackwell, Oxford, UK.

GEHLKE, C. E. (1917), On the correlation between the vote for suffrage and the vote on the liquor question. A preliminary study, *Publications of the American Satistical Association*, **15**, 524–532.

GELMAN, A., J. B. CARLIN, H. S. STERN and D. B. RUBIN (1995), *Bayesian data analysis*, Chapman and Hall, London.

GOODMAN, L. (1953), Ecological regressions and the behavior of individuals, *American Sociological Review* **18**, 663–666.

GOODMAN, L. (1959), Some alternatives to ecological correlation, *American Journal of Sociology* **64**, 610–24.

HAMILTON, R. F. (1982), *Who voted for Hitler?*, Princeton University Press, Princeton.

HÄNISCH, D. (1989), Inhalt und Struktur der Datenbank 'Wahl- und Sozialdaten der Kreise und Gemeinden des Deutschen Reiches von 1920 bis 1933', *Historical Social Research/Historische Sozialforschung* **14**, 39–67.

JENNRICH, R. I. (1969), Asymptotic properties of nonlinear least squares estimators, *Annals of Mathematical Statistics* **40**, 633–643.

KING, G. (1997), *A solution to the ecological inference problem: reconstructing individual behavior from aggregate data*, Princeton University Press, Princeton, NJ.

KING, G., O. ROSEN and M. A. TANNER (1999), Binomial-Beta hierarchical models for ecological inference, *Sociological Methods and Research* **28**, 61–90.

MÁTYÁS, L. (1999), *Generalized method of moments estimation*, Cambridge University Press. Cambridge.

MCCULLAGH, P. and J. A. NELDER (1989), *Generalized linear models*, 2nd edn Chapman and Hall, New York.

METROPOLIS, N., A. W. ROSENBLUTH, M. N. ROSENBLUTH, A. H. TELLER and E. TELLER (1953). Equation of state calculations by fast computing machines, *Journal of Chemical Physics* **21**, 1087–1092.

OGBURN, W. F. and I. GOLTRA (1919), 'How women vote: a study of an election in Portland, Oregon', *Political Science Quarterly* **3**, XXXIV 413–433.

REDNER, R. A. and H. F. WALKER (1984), Mixture of densities, maximum likelihood and the EM algorithm, *SIAM Review* **26** 195–202.

ROBINSON, W. S. (1950), 'Ecological correlation and the behavior of individuals', *American Sociological Review* **15** 351–57.

SERFLING, R. J. (1980), *Approximation theorems of mathematical statistics*, John Wiley & Sons, New York.

TANNER, M. A. (1996), *Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions*, 3rd edn Springer, New York.

TIERNEY, L. (1994), Markov chains for exploring posterior distributions, *Annals of Statistics* **22**, 1701–1762.

WHITE, H. (1994), *Estimation, inference and specification analysis*, Cambridge University Press, Cambridge, England.

Received: September 2000. Revised: January 2001.