

A Missing Data Approach to the Ecological Inference (EI) Problem

(The following is essentially a modification of the poster that the authors presented at the 2000 political science methodology conference in LA, California. We apologize for the lack of formality and confusion that certainly might be in this presentation. Thanks.)

Q: What is the model?

A: A hierarchical, fully Bayesian model within the framework of modern missing data theory.

Q: Does this model have any advantages?

A: There are many:

1. Most importantly, this model explicitly formulates EI as a missing data problem. Therefore, modern missing data techniques such as Markov Chain Monte Carlo (MCMC), especially the Gibbs sampler and the Metropolis-Hastings algorithm can be used to handle the high-dimensional posterior sampling
2. Nice interpretability
3. Capacity to take into account covariates
4. Flexibility in incorporating additional information such as survey data
5. No essential difficulty with handling $R \times C$ tables
6. Ability to detect aggregation bias when there is limited information
7. Robustness to apparent outliers
8. Easy-to-use software available upon request to the authors

Q: "Impressive"! But does it take long to run the program?

A: Not at all. Even if you are a poor graduate student without a fancy computer, it would typically take only several hours to run the program. Obviously, we don't know how long it would take for a famous professor

Q: Good. What is the ecological inference problem anyway?

A: ...

The ecological inference problem

Suppose you have data for the number of blacks and non-blacks as well as the number of people registered to vote in each of the precincts of a certain district. The task is to fill in the true numbers of blacks who are registered.

In other words, you have a number of 2×2 (or other $R \times C$ dimensions) tables with only the marginals observed and you want to do inference about the missing cells.

If this is not very clear, your best bet is to refer to Prof. Gary King's [A Solution to the Ecological Inference Problem](#) (1997)

Q: So, can you tell me a little about your model?

A:

A sketch of the model

(E.g., the 3 by 3 case)

For the k th precinct:

	Dem.	Rep.	No vote	Total
Black	Y_{11k}	Y_{12k}	Y_{13k}	X_{1k}
White	Y_{21k}	Y_{22k}	Y_{23k}	X_{2k}
Other	Y_{31k}	Y_{32k}	Y_{33k}	X_{3k}
Total	T_{1k}	T_{2k}	T_{3k}	N_k

The observed data: T_{1k}, T_{2k}, T_{3k} (the total number of people in each response category: Democrat, Republican, No vote), and X_{1k}, X_{2k}, X_{3k} (the total number of people who belong to each of the categories: black, white, other, respectively)

The missing data: Y_{11k} (the number of people who are black and who vote for Democrat), etc.

Level 1:

For the k -th precinct:

$$(Y_{i1k}, Y_{i2k}, Y_{i3k}) | \theta_{ijk} \sim \text{Multinomial}(X_{ik}, \theta_{i1k}, \theta_{i2k}, \theta_{i3k}),$$

where the parameters θ_{ijk} are the propensities for voter type i to have voting behavior type j . They are not assumed to be the same for every precinct. Rather, these propensities are assumed to follow some hyper level distribution, which makes the model a hierarchical one.

Level 2:

E.g.,

$$(\log(\theta_{i1k}/\theta_{i3k}), \log(\theta_{i2k}/\theta_{i3k})) \sim t_v(\mu_{ik}, \Sigma_i)$$

where v is the degree of freedom of the (bivariate) t distribution for the logit propensities. In other words, the log odds of Democrat versus No vote and the log odds of Republican versus No vote are assumed to be drawn from a suitably heavy-tailed distribution. The choice of the t distribution rather than the normal, which people generally use, is explained later.

For the scale matrix of the t distribution, a typical assumption is to choose a diagonal matrix Σ_i . More generally, however, we assume Σ_i being an arbitrary symmetric, positive definite matrix and put an inverse Wishart prior on it (a uniform prior on Σ_i can be regarded as an improper inverse Wishart prior). We call Σ_i the variance matrix or scale matrix for the fixed effects.

Denote the vector of $(\mu_{i1}, \mu_{i2}, \dots)$ by μ_i . μ_i follows the following level 3 hyper prior distribution:

Level 3:

$$\mu_i = \beta_i M_i + b_i Z_i$$

$$b_i \sim \text{Normal}(0, \tau_i^2 I)$$

where M_i and Z_i (as matrices) are called the design matrices for fixed effects and random effects respectively. These design matrices incorporate possible important covariates. Therefore the hyper level of the model is a mixed effects regression model.

The parameters β_i (a vector), and b_i (a vector) are the regression coefficients of the fixed and random effects, respectively.

τ_i^2 is the variance parameter for the random effects (It is still possible to have more than one variance parameter for the random effects). I is an identity square matrix whose size is the number of random effects.

Q: Isn't something missing? I don't think you included the priors for the hyper-parameters.

A: In general we use very diffuse priors. For the regression coefficients and variance parameters of the fixed effects we use flat priors; for the variance parameters of the random effects we recommend slightly more informative priors because the random effects are used to detect possible aggregation bias and inference of this type is not practical without some relatively informative priors. Technically there are no natural "non-informative" priors for the variance parameters of the random effects in this particular problem. Still, these priors are made fairly diffuse. Please refer to the priors in our examples.

Q: Hmm... After you set up the model, how do you do the inference?

Posterior sampling

A: Bayesian inference is based on the posteriors of the quantities of interest. In EI the primary quantity of interest is (e.g.) the proportion of blacks who are registered to vote, which is the sum of the number of blacks who are registered divided by the total number of blacks. In order to obtain the posterior distribution of this quantity we need the joint posterior of all the parameters given the observed quantities. And to achieve that we resort to Monte Carlo methods. Suppose we need to calculate the posterior of a particular quantity of interest, simply draw a large enough sample from the joint posterior, then for each draw, calculate the quantity of interest and this gives a sample from the correct marginal posterior of the quantity of interest.

In the missing data scenario, the joint posterior includes not only the parameters but also the missing data. In fact, the parameters and the missing data are treated in the same way: we sample from the joint posterior of the parameters and the missing data given the observed data. In our treatment of EI problems, the missing data are the missing cell counts, the parameters include various prior and hyper prior parameters, and the observed data are the observed marginals. A sample from such a high-dimensional posterior distribution amounts to a

possible representation of truth predicted by the model based on the observed data.

Q: The dimension of the joint posterior sounds pretty high. How can you deal with that?

A: This is done by MCMC techniques. Specifically we use a Gibbs sampler with certain Metropolis-Hastings steps.

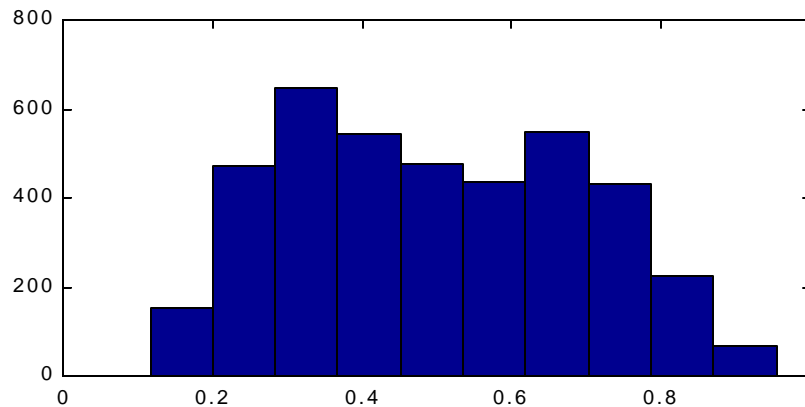
For example, the first step in the Gibbs sampler could be to sample for the true numbers of people in each cell (Y_{11k} , etc.) given their respective propensities θ_{ijk} and hyper level parameters. (Notice that given the propensities θ_{ijk} , the distribution of the numbers of people, e.g., Y_{11k} , does not depend on the hyper level parameters, e.g., μ_i . Conditional independence of this type makes it very appealing to apply Gibbs sampling.) We design a routine (a Metropolis-Hastings step) to do that. Another step is to sample for θ_{ijk} given Y_{ijk} and the hyper level parameters. Another Metropolis-Hastings step is used. Furthermore, yet another step could be to sample for the hyper level parameters given θ_{ijk} and Y_{ijk} , and this turns out to be just like sampling from the posteriors of the regression coefficients and variance parameters of a standard linear regression model.

The sampling algorithms used in our examples are slightly different from and more efficient than the abovementioned but the essential idea is the same. For simplicity, we could not talk at length about all these technicalities. The idea is that with well-designed MCMC algorithms we can sample from very complicated high-dimensional distributions.

Q: Are there any convergence issues?

A: The algorithms that we use are reasonably fast. Generally we run multiple chains (3 to 5) from diverse starting values (generally randomly selected starting values), comparing the between-chain variability and within-chain variability of the quantities of interest, a method proposed by Gelman and Rubin. (Please refer to Gelman et al. 1995.) In the examples that are presented here, it generally takes several thousand iterations per chain for the 2 by 2 case to converge,

although for the 6 by 6 case that follows it takes 30,000 iterations per chain to get fairly reasonable mixing. It is not true that we can just take any data set and apply the algorithm and wait for it to converge. In an artificial example, we use a data set that has 100 2 by 2 tables with every marginal being 100, i.e., there are 100 precincts and for each one of them, there are exactly 200 people: 100 blacks, 100 whites; 100 voters, 100 non-voters. The algorithm has been running for a considerable amount of time, and it turns out that different chains converge to different areas in the sample space. In real data sets, however, we have not yet encountered any serious convergence problem. As an illustration of the efficiency of the algorithm, The model is fit to the black registration in Kentucky data. (Please refer to King, 1997, Chapter 12 for more details about this data set. The authors thank Professor King for allowing us to use this data set as well as the voting registration data for southern states below.) From the posterior distribution of the proportion of blacks who are registered, it is clear that this posterior is bimodal. (Notice that it is not the effect of different chains converging to different places. We create a histogram of the posterior based on each of the chains and the results are roughly the same) Therefore our algorithm is capable of jumping between modes in this particular example.



Marginal posterior of the black proportion of registration. Notice the bimodality.

Q: Fine. I don't think I am interested in listening to an MCMC lecture...

An example and results

A: We have more detailed results for another data set. The data include voter registration and racial background of people from 275 counties in four Southern U.S. States: Florida, Louisiana, North Carolina, and South Carolina. The data from each county include the total voting age population and the proportions of this population who are black, white, and registered in 1968. For more details about this data set, please refer to King, 1997.

Two models are fit to the data set: a model with only one fixed effect for blacks and whites each (the coefficient β_i , $i=1, 2$, for this fixed effect is simply the grand mean of the logit propensities), and a model with one fixed effect and three random effects (again for blacks and whites each) designed to reflect possible aggregation bias.

A normal prior with mean 0 and variance 100 (which practically amounts to a flat prior) is used for the mean logit propensities (for blacks and whites each) in the fixed effects model as well as the grand mean logit propensities in the random effects model. In both the fixed effects and random effects models, the variance parameters for the fixed effects have uniform priors from 0.000001 to 100. In the random effects model the square roots of the variance parameters for the random effects have uniform priors from 0.001 to 5.

Q: How do you design the model to handle aggregation bias again?

A: In the fixed effects model it is assumed that (e.g.) the logit propensities of the blacks follow a single t distribution with a common mean across counties. In the random effects model, however, we divide the counties into three groups: the first group consists of the counties with black population less than 3000; the second, the counties with black population between 3000 and 8000; the third, other counties. For each county in a particular group, the logit propensity (for registering to vote) for the blacks follows a t distribution with a group mean. And these three group means are assumed to have a normal distribution with a common grand mean. So if there is any aggregation

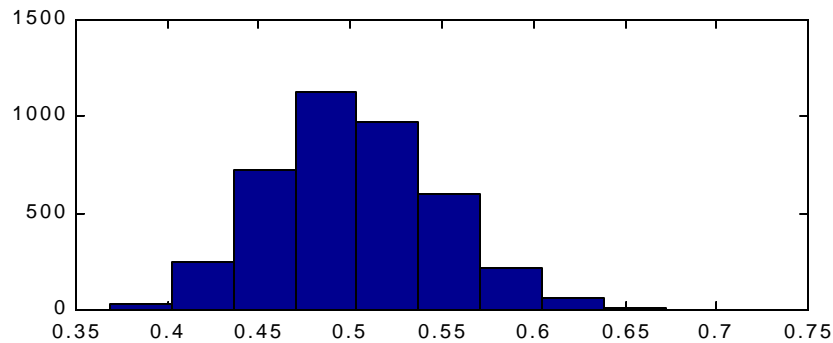
bias it is reflected in the differences between these three group means.

Q: I am still not convinced that this can really work. Do you have some reasonable results?

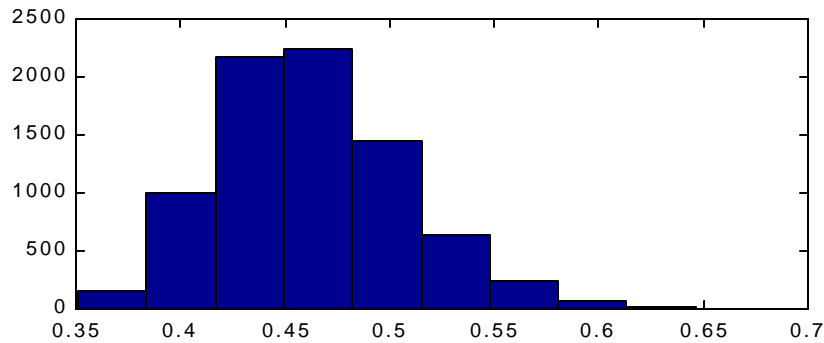
A: We do have some results. Before we get to the detailed results, however, the basic conclusion is that based on our analysis, the proportion of blacks who are registered lies roughly between 0.4 and 0.6, at any rate clearly less than the proportion for the whites. This agrees with King's analysis (King 1997).

There are some minor differences between the two models proposed, and the random effects model slightly under-estimates the proportion of blacks who are registered compared with the simple fixed effects model. Furthermore, from the results of the random effects model, there is some evidence of moderate aggregation bias. From the posterior samples of the coefficients for the random effects, the group mean logit propensity for the first group for the blacks is slightly higher than for the second and third groups. The same pattern is found for the whites (we design Z_2 , the design matrix for the random effects for the whites, in the same manner as we design Z_1 , which is for the blacks). In other words, counties with lower black population tend to be slightly higher in voting registration rates for both blacks and whites. Something to be careful here: when a random effects model like this is fit, essentially it is assumed that the three groups are different to some extent. After the analysis we do see some differences. The problem is that it is hard to tell whether these differences are truly due to the intrinsic nature of the data or to the prior/model specification. In other words, it is possible that the data alone cannot distinguish between aggregation bias and no aggregation bias. Thus models involving the issue of aggregation bias depend to various extents on the prior/model specification.

Histograms of the marginal posteriors for the black proportion of voting registration:



Fixed effects model



Random effects model

Q: Just a couple more questions. First, why should you use a t distribution instead of a normal distribution in your model? Are you just showing us that you can design complicated models that have no relevance to the real problem?

A: Not at all. The idea is that in such a subtle problem as ecological inference, results generally depend considerably on model assumptions. One serious concern with statistical models is robustness. A t error structure is more robust than a normal, i.e., roughly speaking, it is less likely to be influenced by extreme observations and apparent outliers. In this particular data set, there are quite a few apparent extreme observations: in seven of the counties all people are registered, which means that the logit propensities for these counties tend to be really high. By using a t distribution, these apparent outliers will not have a huge effect on the inference. In all of the

examples we use t distributions with 10 degrees of freedom. Using t distributions with a small number of degrees of freedom, such as 4 or 5 is certainly possible, but the resulting posterior distributions might have many modes and the sampling algorithm might get stuck. There is also the issue of robustness versus efficiency: t distributions with smaller degrees of freedom will generally induce more robust procedures but with some loss of efficiency in estimating the quantities of interest when the model actually fits well.

Q: Another question is about incorporating survey data. You did not say much about this, did you?

A: If survey data is available, e.g., for some of the precincts, only level 1 of the model would need to be changed: Instead of saying that the cell counts are all missing now each cell consists of one missing part (due to aggregation) and one observed part (the survey). Of course, the observed part will have a considerable effect on the inference because the missing parts alone cannot really tell much information.

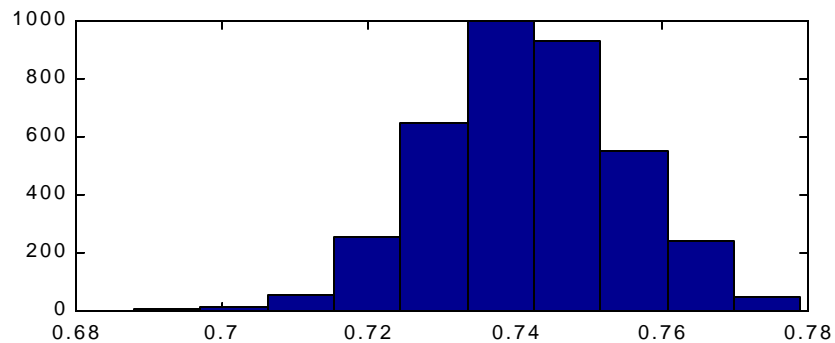
In the implementation of the Gibbs sampler, just one of the steps needs to be changed, which is the step where you sample for the propensities given the numbers of people in each cell. What needs to be done is just to add the numbers in the survey data to the corresponding cells. It is conceptually clear and easy. In practice, if sampling algorithms other than the abovementioned are used, as we do in our examples for more efficiency, incorporating survey information can make the sampler more complicated.

We have another example using voting data for East Germany in 1998. Germany has two votes. The first vote is for a candidate in a first-past-the-post system, and the second vote is the party vote. If a party gets more than 5% it wins seats in parliament. The data set consists of two parts, aggregate data for the first and second votes for the 72 electoral districts in East Germany, and individual level survey (post-election study 1998) data of about 1000 people taken from a simple random sample in East Germany. (The authors thank Thomas Gschwend of the State University of New York at Stony Brook for this data set.) The question of primary interest is voter transition patterns. For example, we are interested in the number of people who

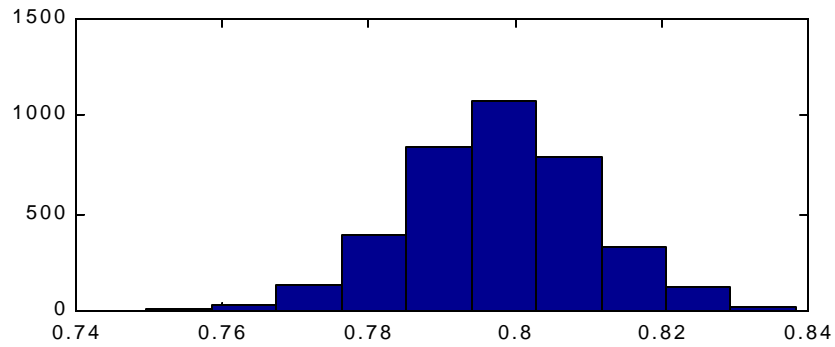
voted for the Green Party in the first vote and PDS for the second, etc.. We formulate the problem as an EI problem with the number of voters for each party (5 large parties and others, so it is 6 by 6 case) in the first vote as the explanatory variables, and formulate the second vote as the response variables. We fit the same model with the same prior specifications as in the fixed effects model in the previous example. A sample from the posterior distribution of the proportion of people in each voting category is tabulated below. Also included are posteriors of the proportions of people who stayed with their original voting decisions (i.e., e.g., the proportion of people who vote for Green Party in both the first and the second votes among the people who vote for Green Party in the first vote).

	SPD2	CDU2	FDP2	Green2	PDS2	Other2
SPD1	0.740385	0.0466968	0.0117538	0.0269973	0.133293	0.0408747
CDU1	0.0467698	0.800532	0.0390297	0.00792498	0.0151862	0.090557
FDP1	0.123307	0.17833	0.385825	0.079406	0.125265	0.107868
Green1	0.131577	0.087251	0.0686003	0.465848	0.15545	0.0912739
PDS1	0.197749	0.0201034	0.0042898	0.0331723	0.663005	0.0816806
Other1	0.0919161	0.105228	0.0442493	0.0602559	0.0761088	0.622242

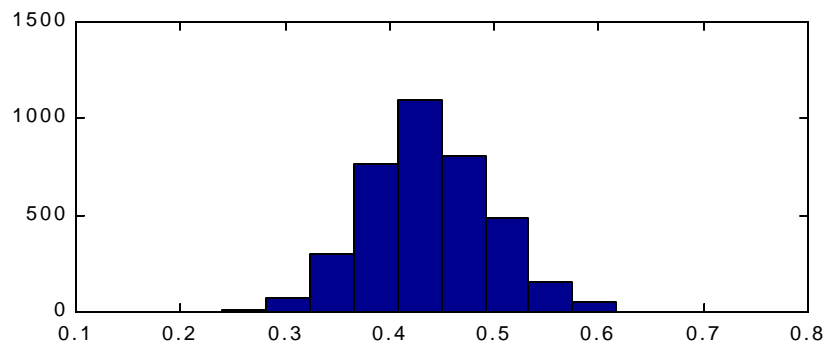
A sample of the proportion of people in each voting category: e.g., the 0.740385 in the upper left corner means that roughly 74 percent of the people who voted for SPD in the first vote again voted for SPD in the second. Notice that each row in the above table adds up to 1.0.



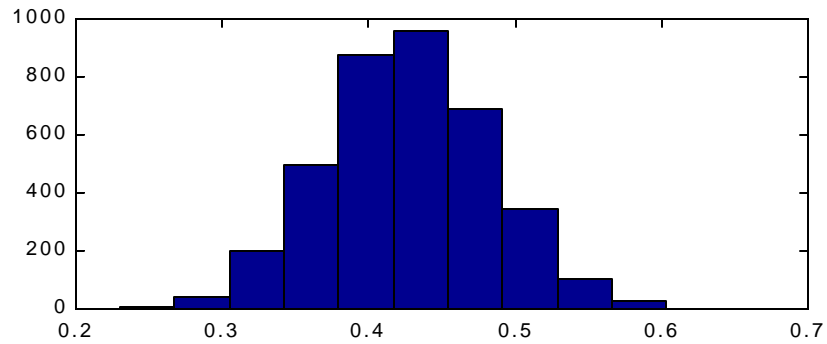
Histogram of the marginal posterior for the proportion of people who voted for SPD in the first vote and voted for SPD in the second.



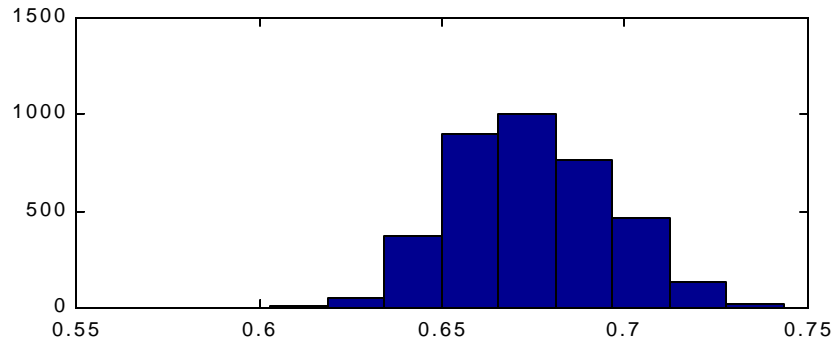
Histogram of the marginal posterior for the proportion of people who voted for CDU in the first vote voted for CDU in the second.



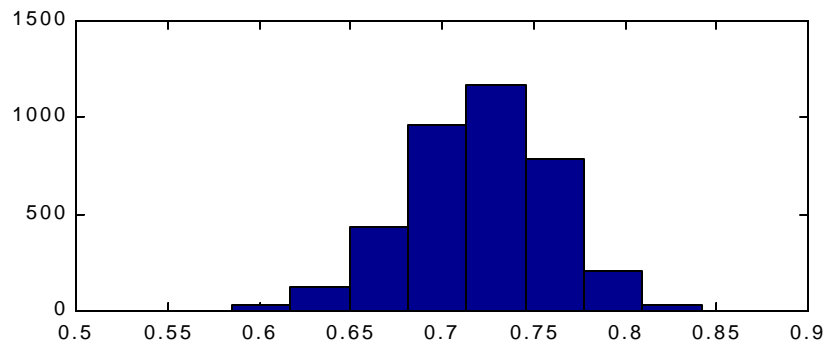
Histogram of the marginal posterior for the proportion of people who voted for FDP in the first vote voted for FDP in the second.



Histogram of the marginal posterior for the proportion of people who voted for Green in the first vote voted for Green in the second.



Histogram of the marginal posterior for the proportion of people who voted for PDS in the first vote voted for PDS in the second.



Histogram of the marginal posterior for the proportion of people who voted for Other in the first vote voted for Other in the second.

Possible extensions and directions for further research

1. Explore the time and space dependence structures for EI problems
2. Sensitivity analysis with respect to model specification, prior specification, etc.. Model check and goodness-of-fit diagnostics

Reference

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 1995. *Bayesian Data Analysis*. London: Chapman and Hall.

Goodman, Leo. 1953a. "Ecological Regressions and the Behavior of Individuals." *American Sociological Review* 18: 663-666.

Goodman, Leo. 1959. "Some Alternatives to Ecological Correlation." *American Journal of Sociology* 64: 610-624.

King, Gary. 1997. *A Solution to the Ecological Inference Problem*. Princeton University Press.

King, Gary, Ori Rosen, and Martin Tanner. 1999. "Binomial-Beta Hierarchical Models for Ecological Inference." *Sociological Methods and Research* Vol. 28, No. 1: 61-90.