This article provides an applied introduction to Bayesian statistics for sociologists. Unlike frequentist statistics, which attaches repeated-sampling frequencies to test statistics, Bayesian statistics directly describes uncertainty about unknown statistical parameters with a probability distribution. With this foundation, much of Bayesian statistics follows from basic rules of probability theory. Three areas of Bayesian statistics are especially relevant for sociologists. First, hierarchical regression models allow several levels of uncertainty into an analysis. Second, Bayes factors provide a useful approach to the problems of model selection, model averaging, and posterior inference about model indexes. Third, recent breakthroughs in estimation methods offer valuable new tools for analysis of Bayesian models that were previously intractable.

Bayesian Analysis for Sociologists

An Introduction

BRUCE WESTERN
Princeton University

Interest in Bayesian analysis has grown tremendously in the field of statistics. Theoretical and applied articles are now common in statistical journals, whereas 20 years ago Bayesian statistics was a much more specialized interest. A number of books have recently been published providing practical guides to Bayesian analysis (e.g., Gelman et al. 1995; Carlin and Louis 1996). Furthermore, Bayesian methods are increasingly on the agenda in discussions of statistical education (Berry 1997; Albert 1997; Moore 1997). Despite a burgeoning research program among professional methodologists, however, Bayesian analysis has made little headway in quantitative sociology.

In part, this reflects the different priorities of methodologists and applied researchers. The Bayesian paradigm has powerful theoretical appeal among statisticians. From a technical point of view, Bayesian methods provide a basis for coherent decision making that is consistent with precepts of rational choice under uncertainty. The Bayesian

AUTHOR'S NOTE: *I thank German Rodríguez and Chris Winship, who provided excellent comments on an earlier draft.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 28 No. 1, August 1999 7-34

approach is also consistent with the likelihood principle that holds that only observed, not hypothetical, data should influence inferences. The attraction of decision-theoretic coherence or the likelihood principle carries little weight among empirical sociologists. If its key attractions are at the foundations of statistics, the Bayesian approach has little to offer applied researchers. This raises the basic question: If you were a Bayesian, would you do your empirical research differently?

I try to answer this question by providing an introduction to Bayesian methods that focuses not on theoretical foundations but on the applied problems of sociologists. I argue that Bayesian statistics provides both a distinctive attitude to data analysis and a novel set of techniques. I also argue that Bayesian analysis is not only different from conventional approaches but, in many cases, better. In general, the Bayesian analysis will be better because it allows a more complete account of all the sources of uncertainty on which a data analysis is based. Consequently, Bayesian conclusions can be more uncertain, "more conservative" than usual.

This article is partly an introduction to basic concepts and partly a survey of recent developments important to sociology. Bayesian methods are introduced with a brief account of the conceptual machinery and an illustration with the linear regression model. After covering the basics, the article reviews three areas of application: (1) hierarchical models, (2) the analysis of multiple models with Bayes factors, and (3) simulation methods for estimation. Because of the applied orientation of the discussion, little technical detail is included (a more thorough treatment is found in the references).

BASIC CONCEPTS

Bayesian statistics rests on interpretative and analytical foundations. For Bayesians, statistical parameters are interpreted probabilistically. This means that the value of parameters is described by probability distributions. Analytically, these probability distributions are found by implementing the conditional probability rule, also known as Bayes rule.

INTERPRETATIVE FOUNDATIONS

Conventional methods for statistical inference are not intuitive. Say I run a regression and find that a t ratio for a coefficient, β, is significant on a two-tailed test at the .05 level. Does this mean that there is just a 5 percent chance that the null hypothesis ($\beta = 0$) is true? No. Well, what if I calculate a 95 percent confidence interval for β. The true coefficient lies inside the interval with 95 percent certainty, right? Wrong. Classical inference does not attach probabilities to coefficients. Instead, inference describes the behavior of test statistics and confidence intervals under hypothetical repeated sampling from an underlying population. Given a statistically significant coefficient, we conclude that if the null hypothesis were true, we would obtain a test statistic at least as large as that actually calculated just 5 percent of the time, under repeated sampling from the population (for a description of repeated-sampling inference, see Barnett 1982, chap. 5).

Unlike repeated-sampling inference that is typically used in quantitative sociology, Bayesian statistics draws a probability distribution over statistical parameters. This probability distribution quantifies the researcher's uncertainty about a parameter's value. In practice, sociologists often interpret p values and confidence intervals as descriptions of the probable value of parameters. Consequently, most sociologists act like Bayesians even if they do not perform explicitly Bayesian calculations. By assigning probabilities to the values of parameters, Bayesian inference is strongly sociologically intuitive.

In the Bayesian framework, a probabilistic interpretation of parameters takes two steps. First, researchers use a subjective probability concept. Second, subjective probability distributions, called prior distributions, are specified for parameters before the data are observed. Probability in repeated-sampling inference is defined by the long-run relative frequency of a statistic under infinite hypothetical samples from an underlying population. Here, probability describes an external, or objective, feature of a sampling procedure. Subjective probability, on the other hand, quantifies a researcher's uncertainty or degree of belief. The subjective approach to probability was forcefully advocated by DeFinetti. DeFinetti ([1970] 1991) provides a colorful

plea for subjective probability in the introduction to his *Theory of Probability*:

> My thesis, paradoxically, and a little provocatively, but nonetheless genuinely, is simply this:
>
> PROBABILITY DOES NOT EXIST
>
> The abandonment of superstitious beliefs about the existence of Phlogiston, the Cosmic Ether, Absolute Space and Time,... or Fairies and Witches, was an essential step along the road to scientific thinking. Probability, too, if regarded as something endowed with some kind of objective existence is no less a misleading misconception, an illusory attempt to exteriorize or materialize our true probabilistic events...
>
> Probabilistic reasoning—always to be understood as subjective—merely stems from our being uncertain about something. It makes no difference whether the uncertainty relates to an unforeseeable future, or to an unnoticed past, or to a past doubtfully reported; it may even relate to something more or less knowable ... but for which we are not willing or able to make the effort and so on... The only relevant thing is uncertainty—the extent of our own knowledge and ignorance. The actual fact of whether or not the events considered are in some sense *determined*, or known by other people, and so on, is of no consequence. (P. x)

At the extremes, a subjective probability of one indicates a researcher's certain belief about the true value of a quantity, whereas a subjective probability of zero indicates an equally certain belief about what the true value is not. Bayesians thus have three states of mind that exhaustively describe all scientific conjectures. The conjectures are either true (having a probability of one), false (having a probability of zero), or uncertain (having a probability between zero and one).

Bayesian analysis begins with an assessment of uncertainty about a parameter before the data are observed. This uncertainty assessment is represented by a probability distribution. Say I am interested in finding the mean income in the United States using survey data from a sample of households. Before I analyze the sample, I describe what I think mean income might be. My uncertainty about this quantity is expressed in the variance of my prior probability distribution. I might think that mean income is most likely to be $40,000. I am not very sure that this is right, so it may be as low as $10,000 or as high as $70,000.

This belief can be represented by a normal distribution with a mean of 40,000 and standard deviation of 15,000. The interior four standard deviations cover 95 percent of the normal distribution. This indicates that I am virtually certain that mean income is between $10,000 and $70,000. This prior distribution is then used in combination with the survey data to estimate the mean income.

It may sound preposterous that these thoroughly speculative guesses are somehow incorporated into a scientific analysis. Everyone could write down their own prior distribution and draw their own conclusion. Indeed, this is the thrust of most criticisms of Bayesian analysis (e.g., Efron 1986). Not only is the basic probability concept subjective in the sense of describing the researcher's state of mind, the whole analysis is subjective because the choice of prior distribution is completely whimsical. If prior distributions are the price paid for probabilistic interpretations of parameters, the cost may be too high.

Although this criticism is common, counterarguments are just as familiar. First, all analysis is subjective, Bayesian or not. In classical inference, α levels for significance testing, the null hypothesis, and estimation methods must all be subjectively selected. These subjective components of classical inference do not appear in Bayesian analysis. Furthermore, the subjective part of Bayesian analysis—prior information— is transparent. If conclusions are based on implausible priors, we have an explicit basis for finding the conclusions implausible. There is no method for incorporating prior information in classical inference, so the subjective dimension of analysis remains hidden and unaccounted for. Second, priors can often be written in a way that allows the sample data to dominate the final results. Such priors are often called noninformative or diffuse and are sometimes interpreted as describing prior ignorance. Third, the influence of prior information on conclusions declines as sample size grows. Asymptotically, Bayesian analysis yields identical answers to maximum likelihood analysis that contains no prior information. Fourth, Bayesian analysis can also consider the sensitivity of the results to the choice of prior distribution. Strong results should be relatively insensitive to choice of priors. All these defenses of Bayesian prior information underscore an important insight that would seem desirable for any statistical philosophy: Conclusions are only as plausible as the subjective foundations on which they are based.

ANALYTICAL FOUNDATIONS

The analytical foundations of Bayesian statistics are rooted in basic rules of probability. The probability of a proposition A can be written $p(A)$. If we have two propositions, A and B, the probability of both propositions is written $p(AB)$ and is called the joint probability of A and B. The conditional probability describes the probability of A given that B is true, $p(A|B)$. Each of these different types of probabilities can be expressed as functions of the others. These expressions provide the three basic probability rules that form the basis for Bayesian inference. The first rule of probability, called the *product rule*, relates the joint to the conditional probability:

$$p(AB) = p(A|B)p(B).$$

The *sum rule* expresses the marginal probability as a weighted sum of conditional probabilities:

$$p(A) = \sum_B p(AB) = \sum_B p(A|B)p(B).$$

The summation sign indicates that we are summing over different values of B. Say $B = 1$ or 0; then, the marginal probability of A is given by

$$p(A) = p(A|B = 0)p(B = 0) + p(A|B = 1)p(B = 1).$$

Finally, the conditional probability of A given B can be calculated with *Bayes rule*:

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}.$$

Bayes rule takes its name from Reverend Thomas Bayes, who in 1763 was posthumously credited with "An Essay Towards Solving a Problem in the Doctrine of Chances." The essay proves the conditional probability rule using the sum and product rules. Bayes rule itself is universally accepted, and indeed trivial, in probability theory. It is the

application of Bayes rule with a subjective probability concept and nonsample information in the prior distribution that is controversial.

How do the three basic rules of probability relate to real statistical analyses? Start by replacing A and B with the two parts of a statistical model—the parameters, θ, and the data, y. Each probability rule then contributes a core element to Bayesian analysis. A full Bayesian probability model is written for the joint distribution of the data and the parameters. Using the product rule, the model can be written as

$$p(y, \theta) = p(y|\theta)p(\theta), \tag{1}$$

where $p(y|\theta)$ is the sampling distribution of the data and $p(\theta)$ is the prior distribution for the parameters. When the sampling distribution for y is written as a function of θ conditional on the data, the distribution is called the likelihood function, familiar from maximum likelihood estimation. The usual goal of a Bayesian analysis is to find the distribution for the parameters, given the data. This distribution, $p(\theta|y)$, is called the posterior. The posterior distribution is given by Bayes rule:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}. \tag{2}$$

Unlike conventional approaches to data analysis, Bayes rule provides a simple formal statement of how to learn from data. Given a prior opinion about the parameters, $p(\theta)$, we can obtain a posterior opinion by updating with sample data summarized by $p(y|\theta)$.

In a complicated model with many parameters, the marginal distribution of the data, $p(y)$, can be difficult to calculate. Consequently, a simpler version of Bayes rule is sometimes used in applications:

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

This version of Bayes rule drops the marginal distribution of the data, $p(y)$.

In principle, the marginal distribution of the data can be found using the sum rule. Typically, however, the parameter θ is continuous rather than discrete. Thus, finding the marginal distribution involves taking an integral rather than a sum:

$$p(y) = \int p(y|\theta)p(\theta)d\theta. \tag{3}$$

The marginal distribution can be viewed as a weighted average of all the distributions of y, given different values of θ. The weights depend on uncertainty about θ described by the prior, $p(\theta)$. Marginalization is a common operation in Bayesian statistics, particularly when interest focuses on a subset of all the parameters in the model. Those quantities, not of key interest, are called nuisance parameters. Obtaining unconditional inference about parameters of interest by integrating over nuisance parameters distinguishes the Bayesian approach from other methods. This provides an important tool for drawing conclusions about some parameters while taking account of the uncertainty of others.

In sum, equations (1), (2), and (3) show that Bayesian statistics simply implements the three basic rules of probability theory. These rules formalize a method of learning in which uncertain prior belief is updated by sample data to form posterior conclusions. Marginalization provides an additional tool to account for irrelevant uncertainty.

A SIMPLE ILLUSTRATION WITH REGRESSION

Bayesian calculations can be illustrated analytically and with simulation methods. In regression analysis, assuming simple distributional forms provides useful analytical expressions that illustrate how the posterior distribution combines information from the prior and the likelihood. In the linear regression of the dependent variable vector, y, on the matrix of covariates, X,

$$y = Xb + e,$$

where b is a vector of coefficients and e is an error term. Assume that the error variance, σ^2, is known and y is conditionally normal. A computationally simple prior for the coefficients is normal with mean, b_0, and variance, V_0. If we write the ordinary least squares (OLS) (maximum likelihood) estimates of the coefficients and their variance as b_1 and V_1, then the posterior distribution for the coefficients is also normal with mean

$$b^* = WV_0^{-1}b_0 + WV_1^{-1}b_1, \tag{4}$$

and variance $W = (V_0^{-1} + V_1^{-1})^{-1}$ (e.g., Leamer 1978:76-78). With a normal prior and normal likelihood, the Bayesian estimate of the regression coefficients is the (matrix) weighted sum of the prior mean, b_0, and the sample estimate, b_1. The weights depend on the least squares and prior variances.

The effect of marginalization over the error variance can also be illustrated with this prior. The model can be generalized by treating σ^{-2} as unknown, with a gamma prior distribution. With this prior, the posterior mean and variance of the coefficients are approximately the same as for the normal case, but the marginal posterior distribution for the coefficients will be t distributed rather than normal (Leamer 1978:78). Integrating over uncertainty about σ^2 thus yields a posterior distribution with fatter tails than the normal, reflecting additional uncertainty about the unknown variance parameter.

The analytical results are a little restrictive because they apply only to the special case of a normal prior with normal likelihood. Recent advances in computing enable simulation methods that can be applied much more generally. Simulation also usefully illustrates how prior and sample are combined with Bayes rule. The object of simulation is to obtain random draws from a posterior distribution. These draws can then be summarized to describe important features of the posterior, such as its mean and spread.

The sampling importance-resampling (SIR) algorithm provides an instructive (but inefficient) method of posterior simulation. The algorithm requires only random draws from the prior distribution, evaluation of the likelihood, and resampling from the prior. The SIR algorithm proceeds in three steps:

1. Generate m random draws from the prior distribution of parameters, $p(\theta)$. Call these draws $\theta^* = \{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(m)}\}$.
2. Calculate weights $w_j = L_j / \sum L_j$, $(j = 1, \ldots, m)$, where L_j is the value of the likelihood function evaluated for $\theta^{(j)}$, given the sample data.
3. Sample with replacement m times from θ^*, with sampling weights, w_j, to form θ^{**}.

The new set of parameters, θ^{**}, approximates a random draw from the posterior distribution, $p(\theta|y) \propto p(\theta)f(\theta|y)$. We can understand how SIR works if we think of θ^* as describing our prior beliefs about the value of a parameter. Before the data are observed, each element of θ^* has the same weight, $1/m$. The prior weights are revised by multiplication by the likelihood. The resulting sample, θ^{**}, thus reflects influence of data on our prior beliefs (Albert 1993:183).

The SIR algorithm is applied to a data set of 18 countries containing information about the average level of unionization between 1971 and 1980, U, and an index measuring the centralization of industrial relations, C (Table 1). Unionization is expressed as a proportion of the total number of workers in the labor force. The centralization index runs from zero to one. Interest centers on a regression of unionization on centralization:

$$U = b_0 + b_1 C + e.$$

With the current coding scheme, the intercept term expresses the average level of unionization among decentralized countries, whereas the slope coefficient describes the difference in unionization between the least centralized and most centralized countries. The regression model includes an additional parameter—the error variance—which I initially assume to be fixed at its least squares estimate.

The Bayesian analysis assigns prior distributions to the two unknowns, the slope and the intercept. Previous research suggests union density is low among the decentralized countries, perhaps as low as .20, but possibly as high as .50. For illustration, let the prior distribution for the intercept be uniform over this range. Previous research again suggests that unionization among the most centralized countries might be anywhere from 15 to 55 percentage points higher than in the least centralized countries. The prior for the slope is thus uniform between .15 and .55. Note that both priors are quite spread out, together accounting for more than the observed range of the dependent variable. The left-hand panel of Figure 1 reports a plot of 2,000 random draws from the prior distribution. Histograms on the horizontal and vertical axes reflect the shape of the underlying uniform distribution.

To summarize the data, we also need to write a likelihood function. I assume that the dependent variable is conditionally normal, so OLS is

TABLE 1: Data for Regression Analysis of Unionization, 18 OECD Countries

	Proportion Unionized	*Centralization of Industrial Relations*
Australia	0.52	.00
Austria	0.59	1.00
Belgium	0.63	.75
Canada	0.32	.00
Denmark	0.71	.75
Finland	0.74	.75
France	0.20	.00
Germany	0.74	1.00
Ireland	0.56	.00
Italy	0.49	.12
Japan	0.33	.38
Netherlands	0.39	.50
New Zealand	0.43	.12
Norway	0.59	1.00
Sweden	0.82	1.00
Switzerland	0.34	.25
United Kingdom	0.52	.00
United States	0.23	.00

SOURCE: Western (1997: 89).
NOTE: OECD = Organization for Economic Cooperation and Development.

maximum likelihood. With this likelihood, the log of the importance weights, w_j, is proportional to the sum of squared residuals obtained with randomly sampled slope and intercept coefficients from the prior distribution.

Weighted resampling from the prior provides an approximate random draw from the posterior distribution. This draw is shown in the right-hand panel of Figure 1. Although the plot of the prior showed an even distribution of coefficients over the parameter space, the introduction of the sample data has concentrated our beliefs about the value of the intercept and the slope in a much narrower area. The posterior can be understood as a combination of the uniform prior distribution with the normal likelihood. Because the prior was diffuse compared to the likelihood, the posterior resembles the normal distribution more than the uniform. Still, the prior does influence the shape of the posterior. Because the prior distributions are bounded, giving no prior weight to coefficients outside the specified range, the posterior distributions are also bounded. Commonly in Bayesian re-

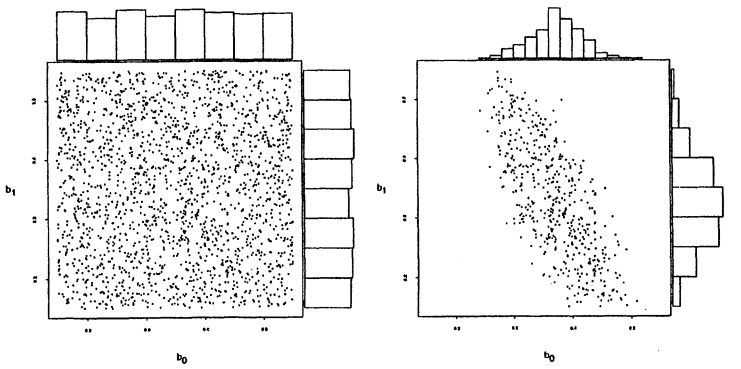

Figure 1: The Joint Prior Distribution of Intercept and Slope Coefficients for a Model of the Effect of Labor Market Centralization on Unionization

gression, unbounded distributions are often chosen to represent prior beliefs, removing arbitrary bounds on the posterior.

Table 2 reports an alternative summary of the results. The table describes the coefficient distributions with the mean and standard deviation. Maximum likelihood results are the OLS estimates. The table shows that the results from the Bayesian analysis, assuming σ^2 known, are almost identical to the OLS results, and quite different from the prior.

Additional insight is provided by repeating the analysis with an unknown error variance. In this case, a prior distribution must now be supplied for σ^2. The error variance describes the predictive accuracy of the unionization model. It is a bit simpler to think of the standard error of the residuals rather than the variance, so σ is given a diffuse and uniform prior, in this case over the interval [.05, .25]. With σ^2 unknown, there is an additional source of uncertainty in the analysis that increases posterior uncertainty about the coefficients. The resampling algorithm is instructive. If a prior sample of coefficients is quite different from the least squares estimate but the prior sampled σ^2 is large, the coefficients will have higher likelihood than if σ^2 were small. Consequently, these more dispersed coefficients receive rela-

TABLE 2: Summary of Means and Standard Deviations of Coefficient Distributions in a Bayesian Regression Analysis of Unionization

Distribution	*Intercept*	*Centralization*
Uniform prior	.20	.20
	(.115)	(.115)
Likelihood (OLS)	.373	.320
	(.044)	(.075)
Posterior with σ^2 known	.373	.322
	(.042)	(.073)
Posterior with σ^2 unknown	.373	.322
	(.048)	(.079)

NOTE: Standard deviations of distributions in parentheses.

tively more weight in resampling than if σ^2 were fixed at a relatively small value like its maximum likelihood estimate. Prior uncertainty about the error variance thus increases posterior uncertainty about the coefficients. In other words, marginal inference about the coefficients in the Bayesian approach averages over posterior uncertainty about the error variance. Thus, with σ^2 unknown, the posterior standard deviations for the two coefficients are slightly larger than with σ^2 known. Table 2 shows that marginalizing over σ^2 makes little difference. Sometimes, however, posterior uncertainty about nuisance parameters can contribute substantially to uncertainty in posterior inferences.

The analytical results and the simulation exercise help sharpen several intuitions about Bayesian data analysis:

- The Bayesian posterior distribution combines information from the prior and the likelihood distributions. If one distribution is quite concentrated relative to the other, the more concentrated distribution drives the final result.
- This implies that, as sample size gets bigger, the likelihood increasingly dominates the prior and the results move toward the maximum likelihood results.
- If the prior distribution is very diffuse, the sample data will dominate the prior even if sample size is small. This is the case in the current regression analysis of just 18 observations.

More generally, in a wide variety of statistical analyses, the Bayesian approach mechanically applies the axiomatic rules of probability

about conditioning and marginalization. Quantities that are uncertain a priori are given a prior probability distribution. Quantities that are known with certainty, such as the observed data, are conditioned on. Quantities that are uncertain, but uninteresting, are marginalized. This approach automatically takes into account all certain and uncertain information before drawing a conclusion about statistical parameters.

Bayesian methods also suggest a distinctive attitude to data analysis. Quantities not known with certainty, but which are treated as certain and conditioned on, produce unrealistically optimistic conclusions. Posterior inference overstates confidence in research conclusions as a consequence of ignoring prior uncertainty. This idea is critical because it provides a formal framework for describing real data analyses. In practice, prior uncertainty is rife. Not only are the parameters not known before the data are analyzed, coding decisions must be made, predictor variables are unknown, features of the data such as outliers or influential data points are not known a priori, and the shape of the likelihood function that generates the data is also uncertain. With conventional statistics, the only uncertainty admitted to the analysis is sampling uncertainty. The Bayesian approach offers guidance for dealing with the myriad sources of uncertainty faced by applied researchers in real analyses. Thus, although Bayesian statistics are appealing for those with an interest in statistical foundations, they also offer real solutions for real-world empirical research.

Because statistical texts are theoretically oriented and often quite technical, there are few good introductory sources suitable for sociologists interested in Bayesian analysis. Still, an accessible discussion of the foundations of different paradigms of statistical inference is provided by Howson and Urbach (1993). Berk, Western, and Weiss (1995) discuss subjective probability in a sociological context. Lee's (1997) textbook offers a straightforward treatment of basic concepts, with little calculus or linear algebra. More advanced statistical introductions with an applied orientation include Box and Tiao ([1973] 1992) and Gelman et al. (1995). On Bayesian regression, Leamer (1978:21-84) provides a compact summary of different models, as well as a clear introduction to Bayesian concepts more generally. Western and Jackman (1994) present a simple example of Bayesian regression with informative priors in a social scientific

context. A clear description of the SIR algorithm is given by Smith and Gelfand (1992).

APPLICATIONS

The basic concepts discussed above can be extended to produce a wide variety of distinctively Bayesian analyses. This section focuses on three areas. First, hierarchical modeling provides a type of Bayesian regression with prior information. Second, Bayesian marginalization provides an analysis of multiple statistical models. Third, new simulation methods facilitate estimation of a wide variety of previously intractable Bayesian models.

HIERARCHICAL MODELS

Hierarchical models, and hierarchical linear models in particular, are widely used in sociology (see the review by DiPrete and Forristal 1994). These specifications are sometimes called random effects or multilevel models. The models are typically used in cases where the data are collected in hierarchical fashion. In comparative research, for example, batches of observations are drawn from a sample of countries. Similarly, with survey panel data, repeated observations are made on respondents over time. Hierarchical regression models follow this structure by treating coefficients as if they were drawn from some population distribution. The parameters of this population distribution, called hyperparameters, are estimated with a combination of prior information and the sample data.

To see the advantages of the hierarchical model, say we have sampled respondents from N of the 50 U.S. states and we are interested in studying the relationship between education and earnings in each state. Our education variable is simply a dummy variable indicating those with college education. For each of the $j = 1, \ldots, N$ states, or, more generally, "clusters," we can estimate the regression coefficient θ_j that gives the earnings premium of the college educated. To estimate θ_j, we might pool all observations from all states, introduce $N - 1$ state dummies, and specify interactions between the state and

education variables. Alternatively, we might simply analyze the data from each cluster separately.

Although both methods seem reasonable, hierarchical modeling has advantages over each. First, consider pooled analysis with interactions. The interaction model ignores the structure of the data by treating all observations as independent when we really have N distinct samples, and respondents within each sample are related by their cluster membership. There are two sources of uncertainty in this example: the usual sampling uncertainty associated with the selection of respondents within clusters and uncertainty about the selection of clusters. By treating θ_j as a realization from a population distribution, the hierarchical model allows for uncertainty in addition to the usual error term that we find when the data have a multilevel structure.

Second, consider the unpooled analysis that models each cluster separately. Although this may allow for the variation across clusters that is not otherwise captured by covariates, it is inefficient. By estimating population parameters (hyperparameters) from the whole data set, and finding the θ_j given those parameters, information about all the clusters is used to produce an estimate for each. Pooling information in this way yields more precise estimates of θ_j than separate cluster-level analyses.

Hierarchical models have a natural Bayesian justification. Extending the above example, earnings, y, is written as a linear function of education, E, in each of our N clusters:

$$y_{ij} = b_{0j} + b_{1j}E_{ij} + e_{ij}, \quad i = 1, \ldots, \quad n_j, \quad j = 1, \ldots, N.$$

Subscripts for the slope and intercepts indicate that the regression coefficients vary across clusters. The population distribution for the coefficients can be viewed as a special type of prior distribution, where, commonly,

$$b_{kj} \sim N(\eta_k, \sigma_k^2), \qquad k = 1, 2.$$

Unlike the usual Bayesian regression model, the parameters of these distributions are not simply chosen a priori. Instead, they are determined chiefly by the sample data from all N clusters. Write the hyperparameters in matrix form, with the mean vector $\eta = \{\eta_k\}$ and

diagonal covariance matrix $\Sigma = \text{diag}\{\sigma_k^2\}$. Assuming that y follows a normal distribution, the mean of the posterior distribution for the regression coefficients is given by

$$b_j = W_j \Sigma^{-1} \eta + W_j V_j^{-1} \hat{b}_j, \tag{5}$$

where the coefficients are collected in the vector $b_j' = (b_{0j}, b_{1j})$, $\hat{b}_j$ is the least squares estimate, $W_j = (\Sigma^{-1} + V_j^{-1})^{-1}$, and V_j is the covariance matrix of the least squares estimate (Bryk and Raudenbush 1992:42).

Equation (5) has precisely the same form as the posterior mean for coefficients in the Bayesian linear regression with a normal prior (equation (4)). The coefficients of the hierarchical model are a (matrix) weighted average of the coefficients obtained from the sample data within a particular cluster and the population distribution based on data from all clusters. This setup, with no prior distribution for the hyperparameters, is called "empirical Bayes" because the sample data are used to form the "prior" or population distribution. A pure Bayes analysis is obtained by writing proper prior distributions for the hyperparameters. As we would expect, with diffuse priors on the hyperparameters, pure Bayes and empirical Bayes yield similar results. However, pure Bayes analysis—even with diffuse priors— provides hyperparameters with larger posterior variance than does empirical Bayes. This is due to explicit accounting for prior uncertainty in η and Σ.

Three main statistical advantages recommend hierarchical models for analysis of multilevel data. First, they tend to produce more accurate forecasts than rival models that take no account of the structure of multilevel data (Dempster, Rubin, and Tsutakawa 1981; Western 1998). Second, the hierarchical model provides more accurate estimates of cluster-level effects than does unpooled analysis. Although biased, like all Bayesian linear estimators, mean squared error for the hierarchical coefficients is lower than for corresponding least squares estimates (Lindley and Smith 1972). We can understand the high precision (low variance) of hierarchical coefficients in a particular cluster to result from the use of the population distribution, determined by data from all clusters. Third, because there are two sources of uncertainty, one within clusters and the other across clusters, the hierarchi-

cal model also provides a more realistic accounting of uncertainty for data with a multilevel structure.

There are several good textbook treatments of multilevel models. Bryk and Raudenbush (1992) provide an accessible introduction to hierarchical linear models for social scientists. Longford (1993) and Goldstein (1995) present non-Bayesian treatments of hierarchical models. Both books also describe hierarchical structures for the coefficients of generalized linear models. More technical, yet strongly applied, discussions from a Bayesian perspective are provided by Gelman et al. (1995) and Carlin and Louis (1996). These two books additionally offer good discussions of model evaluation and criticism.

MULTIPLE MODELS

In real data analysis, researchers typically examine many different models, experimenting with various subsets of predictor variables, transformations, interaction terms, alternative methods of estimation, and so on. Just as Bayes rule can be applied to find posterior probabilities for parameters that are a priori uncertain, the method can also be used to find posterior model probabilities when there is prior uncertainty about the model.

A statistical model consists of a set of assumptions, M, and parameters that arise under those assumptions, θ. The assumptions in M might include lists of independent variables in a regression or a shape for the sampling distribution of the data. Because statistical models in sociology are not known with certainty, prior distributions can be drawn over the model space and the parameters, $p(M, \theta) = p(\theta|M)p(M)$. The joint prior probability of the model assumptions and the parameters are factored, using the product rule, into a prior for the parameters conditional on the model and a prior for the marginal probability of the model.

To calculate posterior model probabilities, we introduce some sample data, y. The key quantity determining the probability of a model is the quality of its prediction of the observed data, expressed in the conditional distribution $p(y|M)$. To measure how well a model predicts the observed data, we simply integrate over the parameters, θ:

$$p(y|M) = \int p(y|\theta, M)p(\theta|M)p(M)d\theta.$$

This measure of the model predictions, $p(y|M)$, is sometimes called the marginal likelihood or the model likelihood. With this measure, posterior model probabilities can be found with Bayes rule:

$$\text{model posterior} \propto \text{model likelihood} \times \text{model prior}.$$

The ratio of two posterior model probabilities, for models M_1 and M_2, can then be written as

$$\frac{p(M_1|y)}{p(M_2|y)} = \frac{p(M_1)}{p(M_2)} \frac{p(y|M_1)}{p(y|M_2)}.$$

The first term on the right-hand side, $p(M_1)/p(M_2)$, describes the relative prior plausibility of model M_1 compared to M_2. If both models are equally plausible, the ratio equals one. The second term on the right-hand side quantifies the relative predictive performance of the two models. If M_1 predicts the observed data better than M_2, the ratio is greater than one. This ratio is called the Bayes factor, written B_{12}.

The Bayes factor, and statistics related to it, has generated growing interest in sociology. This is largely due to the work of Raftery (1986, 1995), who popularized the Bayesian information coefficient (BIC). The BIC statistic provides a very simple but fairly accurate approximation of two times the log Bayes factor, where the baseline for model comparison is a saturated model that fits the data perfectly. For linear regression, the BIC for model k is written as

$$\text{BIC}_\text{k} = n \log(1 - R_k^2) + p_k \log n,$$

where R_k^2 is the R^2 from the least squares fit and p_k is the number of coefficients in the model excluding the intercept. A negative BIC statistic indicates the superior predictions of model k in comparison to the saturated model. The saturated model is seldom of interest, but

statistics based on other model comparisons can be found by simply taking the difference of two BIC statistics:

$$2 \log B_{12} \approx \mathrm{BIC}_1 - \mathrm{BIC}_2.$$

Bayes factors and related statistics might be used in three different ways: (1) for model selection, (2) for model averaging, and (3) for posterior inference about model indexes. The Bayes factor is a valuable tool for model selection because the methodology automates a preference for parsimonious explanation. Recalling our intuitions about marginalization, model likelihoods tend to carry more uncertainty for complex models because they integrate over a larger parameter space than simple models. Greater prior parametric uncertainty contributes to higher predictive uncertainty. The principle of parsimony is reflected in the expression for the BIC where the usual goodness-of-fit statistic, R^2, is balanced by a penalty based on the number of parameters in the model, p. Because of its preference for parsimony, BIC is commonly used in the selection of log-linear models fit to large mobility tables. In this context, conventional likelihood ratio tests overwhelmingly favor the saturated model, since the sample size is so large. By emphasizing parsimonious explanation, the BIC often prefers nonsaturated models.

However, reporting a "best fitting" or "most probable" specification is unsatisfactory from a strict Bayesian viewpoint because posterior inference neglects prior uncertainty about the model (Gelman and Rubin 1995). Often, model searching is ad hoc, driven by preference for certain findings, and continues until those findings, or some approximation, are discovered. Results produced through such data mining can be viewed as summaries of posterior distributions conditional on a model that is assumed to be true but is really uncertain a priori. The Bayesian alternative acknowledges prior uncertainty about the model with a prior distribution and obtains unconditional inference by averaging over a variety of possible models. Say we have three regression models, M_i, $(i = 1, 2, 3)$. All models contain the predictor, x, plus other, possibly overlapping, sets of predictors. Posterior inference about the effect of x given the data, y, is a weighted average of conditional posterior distribution for the coefficient, β,

$$p(\beta|y) = \rho_1 p(\beta|y, M_1) + \rho_2 p(\beta|y, M_2) + \rho_3 p(\beta|y, M_3),$$

where the weights are posterior model probabilities, ρ_i. Approximate posterior model probabilities can be calculated from the BIC statistic:

$$\rho_j = \frac{\exp(\text{BIC}_j)}{\sum_i \exp(\text{BIC}_i)}.$$

Like most marginalization methods, model averaging drives inferences in a conservative direction by taking account of prior uncertainty about unknowns. An important limitation of the method is that it can be applied only to quantities that share the same meaning across models. For instance, it makes little sense to average coefficients across regression models with logged and unlogged dependent variables, since the coefficients describe different statistical relationships in each model.

Sometimes, variation in model assumptions is described by an index that might also be subject to posterior inference. For example, robust regression can be performed by fitting a t distribution instead of a normal (Lange, Little, and Taylor 1989). Different t distributions are indexed by a degree-of-freedom parameter that expresses the resistance of coefficients to outliers. Highly resistant fits are obtained with degree of freedom close to zero. At the other extreme, OLS sets the degree of freedom parameter to infinity. Another example arises in the analysis of structural breaks in time series. Say we have a bivariate time series regression of y_t on x_t ($t = 1, \ldots, T$), but we believe that the regression relationship is not stable over time. We could fit the model:

$$y_t = b_0 + b_1 x_t + b_2 D_t + b_3 D_t x_t + e_t,$$

where D_t is a dummy variable that equals zero for all $t < S$ and one otherwise, and S is the time point marking the structural break in the series. Following the structural break, the effect of x_t switches from b_1 to $b_1 + b_3$.

In both these examples, the most probable model is not known a priori, but the model space is quite continuous. In the robust regression example, the degree of freedom parameter for the t distribution

indexes the range of alternative models. In the time series example, the models vary with the break point, S. Posterior inference for the indexes in both examples could proceed simply by calculating Bayes factors over a grid of points and plotting them. The resulting shape is proportional to the marginal posterior distribution of the index parameters, with a uniform prior distribution on the model index. This method seems particularly useful in the time series analysis where statistical relationships are often unstable in comparative and historical problems (Isaac and Griffin 1989). The Bayesian approach treats structural breaks in time series as parameters for estimation and inference.

Bayesian analysis of multiple models in sociology is comprehensively reviewed by Raftery (1995). His article provides an accessible derivation of the BIC, as well as computational formulas for a range of models. Social science examples of model averaging are reported by Western (1996) and Bartels (1997). Leamer (1983) provides an excellent and engaging discussion of the problem of model uncertainty. He also suggests a simple alternative method for propagating prior model uncertainty to posterior inferences about coefficients of a regression model. Western and Healy (forthcoming) illustrate the Bayesian approach to diagnosing a structural break in time series analysis. The utility of Bayes factors is somewhat controversial because the statistics are sensitive to the choice of prior distribution for parameters, and there are alternative methods for the evaluation of Bayesian models (Gelman and Rubin 1995; Gelman et al. 1995:162-83). The BIC approximation to the Bayes factor has also been criticized. The prior distribution implied by the BIC may be unrealistically diffuse, and the quality of the approximation may decline with skewed data (see Weakliem 1999 and accompanying papers). From this perspective, the BIC understates sample evidence for modest effects. More accurate Bayes factor approximations that allow informative priors may thus be preferable to the BIC (e.g., Raftery 1995, eqs. (14) and (15)).

ESTIMATION

Numerical methods based on computer simulation have opened an exciting new area in Bayesian statistics. Historically, Bayesian methods offered little to applied researchers because complex inte-

gration problems that arise in marginal and posterior inference were numerically intractable. Recently, however, simulation methods have opened the way for practical Bayesian analysis of a wide range of models.

Simulation methods provide random draws from posterior distributions that can then be summarized with posterior means, medians, quantiles, standard errors, and so on. With simulation, difficult mathematical calculations are replaced by computing power. Thus, practical Bayesian inference may be computationally intensive but is now increasingly feasible as computing costs decline. The SIR algorithm, illustrated above, is one example of a simulation method for posterior inference. However, SIR is quite inefficient in the sense that accurate inferences are computationally very costly, particularly when there is strong disagreement between the prior and the sample.

Markov chain Monte Carlo (MCMC) methods provide a more efficient approach to posterior simulation. Whereas SIR generates independent draws from the posterior distribution, MCMC methods generate dependent draws from the posterior. With MCMC, simulation follows a Markov process in which current draws from the posterior depend probabilistically on prior draws in the chain.

The best-known MCMC method is the Gibbs sampler. Say we have two parameters, α and β. These might be the coefficients and error variance of a linear regression or the hierarchical coefficients and hyperparameters of a hierarchical model. In many cases, we have a known form for the distributions of subsets of parameters conditional on the data and all other parameters, $p(\beta|\alpha, y)$ and $p(\alpha|\beta, y)$. This is of limited immediate help because interest focuses on the marginal distributions, $p(\beta|y)$ and $p(\alpha|y)$. These marginal distributions are unknown. The Gibbs sampler uses the known conditional distributions to simulate from the joint posterior distribution:

1. Given a starting value, $\alpha^{(0)}$, and data, y, simulate $\beta^{(0)}$ by generating a random draw from the distribution $p(\beta|\alpha^{(0)}, y)$.
2. Now, simulate $\alpha^{(1)}$ by generating a random draw from $p(\alpha|\beta^{(0)}, y)$.

When the full conditional distributions have finite integrals, repeating steps 1 and 2 many times yields a sequence of random variables whose distributions converge to the joint posterior distribution $p(\alpha, \beta|y)$. Once the Gibbs sampler has converged, marginal inference

proceeds by summarizing the distribution of simulated parameters, $\alpha^{(i)}$ and $\beta^{(i)}$.

Gibbs sampling is especially useful when normal approximations to posterior distributions are inaccurate. This can occur in the analysis of small samples, nonlinear models, or models with nonstandard prior distributions. In particular, the Gibbs sampler can provide very accurate inferences about the parameters of hierarchical generalized linear models (Rodríguez and Goldman 1995). The Gibbs sampler has also been applied to selection models, latent class models, analyses with covariate measurement error, and missing data problems. Instead of presenting an illustration of the Gibbs sampler here, I refer readers to the other articles in this issue. Weiss et al. (1999) use the Gibbs sampler for analysis of a logistic regression model with prior information. King, Rosen, and Tanner (1999) and Lewis and Raftery (1999) fit hierarchical models using MCMC methods.

Implementation of the Gibbs sampler can be difficult, however. There are two main applied problems. Sometimes, the Gibbs chain moves very slowly over the parameter space. This can happen if the posterior distributions are not strongly determined by the data, if the parameters of the model are highly correlated, or if the posterior distribution has an irregular shape. When the Gibbs chain moves slowly over the parameter space, iterations of the chain are highly autocorrelated and convergence is slow. Second, convergence of the Gibbs chain can be difficult to diagnose. Although a variety of diagnostics are available, a recent review suggests that many of these diagnostics can fail even in the case of idealized, simplified problems (Cowles and Carlin 1996). In sociological applications, Gibbs sampling is likely to be used most in inference for hierarchical regression models. Some researchers claim reliable performance for the Gibbs sampler under certain standard specifications in this context (Kass et al. 1998).

Applied sociologists are likely to have little interest in the details of the Gibbs sampler. Still, because the behavior of the sampler must be closely monitored to ensure convergence, researchers should know something about the mechanics of the algorithm. An accessible introduction to the Gibbs sampler is provided by Casella and George (1992). Tanner (1993) provides an excellent discussion, with many examples, of methods for simulating posterior distributions. Practical strategies for Gibbs sampling are outlined in the roundtable discus-

sion moderated by Kass (Kass et al. 1998) and in the contributions to Gilks, Richardson, and Spiegelhalter (1996). In that volume, Clayton (1996) provides a useful discussion of Gibbs sampling for hierarchical generalized linear models. More general presentations of Gibbs sampling in hierarchical models are provided by Gelman et al. (1995) and Carlin and Louis (1996). Cowles and Carlin (1996) review and assess a wide variety of convergence diagnostics. Software for Gibbs sampling is freely available in the BUGS (Bayesian inference using Gibbs sampling) package (Spiegelhalter et al. 1997). This software is accompanied by extensive documentation, worked examples, and a suite of S-Plus functions for analysis of posteriors and convergence diagnostics.

CONCLUSION

Bayesian statistics offers practical methods for statistical inference that are rooted in the basic rules of probability theory. By using probability to describe uncertainty about parameters, the Bayesian approach converges with standard sociological understanding. Unlike conventional methods, Bayesian methods provide a formal rule for learning from data: Prior beliefs are updated with sample data using Bayes rule to obtain posterior probability statements about unknown quantities. With substantively uninteresting unknowns, the Bayesian prescribes marginalization. This way, inferences about quantities of interest also incorporate uncertainty about other parameters in the model. Although Bayesian prior distributions are often criticized, they can usually be specified to represent vague prior opinions about unknown quantities.

This basic framework motivates a wide variety of different applications. The discussion focused on three areas: hierarchical models, the analysis of multiple models, and methods for posterior simulation. Hierarchical models are valuable tools for analyzing clustered data. These models yield coefficients with smaller mean squared error and smaller forecast error than standard specifications. In the analysis of multiple models, the Bayesian approach accommodates prior uncertainty about the choice of model in final inferences. In addition, statistics for Bayesian modeling automatically incorporate the

principle of parsimony into the analysis. The explosion of Bayesian applications in statistics is due largely to breakthroughs in estimation methods. MCMC methods provide a computationally expensive but analytically cheap approach to studying Bayesian models that were previously intractable or only poorly approximated. Although MCMC necessitates closer scrutiny than standard estimation procedures, the rewards are substantial. In relatively familiar areas of application—such as hierarchical generalized linear models—there is no accurate alternative to MCMC.

This article began by asking: If you were a Bayesian, would you do your research differently? Most generally, the Bayesian approach allows a rich view of uncertainty in data analysis. Whereas conventional statistics admits only uncertainty due to sampling variability, virtually any assumption of a Bayesian model can, in principle, be treated as uncertain a priori. Bayesian methods thus offer sociologists a powerful and practical tool for data analysis. Realistically complex models can be developed that reflect a variety of sources of uncertainty faced by a researcher in the context of real empirical problems. In this article, uncertainty about model specifications and nuisance parameters was allowed for, driving inferences in a more conservative direction. In sociology, where sampling uncertainty constitutes only a fraction of all uncertainties in real data analysis, standard methods face a compelling Bayesian alternative.

REFERENCES

Albert, James H. 1993. "Teaching Bayesian Statistics Using Sampling Methods and MINITAB." *American Statistician* 47:182-91.

———. 1997. "Teaching Bayes Rule: A Data-Oriented Approach." *American Statistician* 51:247-53.

Barnett, Vic. 1982. *Comparative Statistical Inference*. 2d ed. New York: Wiley.

Bartels, Larry M. 1997. "Specification Uncertainty and Model Averaging." *American Journal of Political Science* 41:641-74.

Berk, Richard, Bruce Western, and Robert Weiss. 1995. "Bayesian Inference for Apparent Populations." *Sociological Methodology* 25:421-85.

Berry, Donald A. 1997. "Teaching Elementary Bayesian Statistics With Real Applications in Science." *American Statistician* 51:241-46.

Box, George E. P. and George C. Tiao. [1973] 1992. *Bayesian Inference in Statistical Analysis*. Reprint, New York: Wiley.

Bryk, Anthony S. and Stephen W. Raudenbush. 1992. *Hierarchical Linear Models: Applications and Data Analysis Methods.* Newbury Park, CA: Sage.

Carlin, Bradley P. and Thomas A. Louis. 1996. *Bayes and Empirical Bayes Methods for Data Analysis.* New York: Chapman and Hall.

Casella, George and Edward I. George. 1992. "Explaining the Gibbs Sampler." *American Statistician* 46:167-74.

Clayton, David G. 1996. "Generalized Linear Mixed Models." Pp. 275-319 in *Markov Chain Monte Carlo in Practice*, edited by W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. London: Chapman and Hall.

Cowles, Mary Kathryn and Bradley P. Carlin. 1996. "Markov Chain Monte Carlo Diagnostics: A Comparative Review." *Journal of the American Statistical Association* 91:883-904.

DeFinetti, Bruno. [1970] 1991. *Theory of Probability.* Reprint, New York: Wiley.

Dempster, A. P., D. B. Rubin, and R. K. Tsutakawa. 1981. "Estimation in Covariance Components Models." *Journal of the American Statistical Association* 76:341-53.

DiPrete, Thomas A. and Jerry D. Forristal. 1994. "Multilevel Models: Methods and Substance." *Annual Review of Sociology* 20:331-57.

Efron, Bradley. 1986. "Why Isn't Everyone a Bayesian?" (with discussion). *American Statistician* 40:1-11.

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

Gelman, Andrew and Donald A. Rubin. 1995. "Avoiding Model Selection in Bayesian Social Research." *Sociological Methodology* 25:165-74.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter, eds. 1996. *Markov Chain Monte Carlo in Practice.* London: Chapman and Hall.

Goldstein, Harvey. 1995. *Multilevel Statistical Models.* 2d ed. London: Arnold.

Howson, Colin and Peter Urbach. 1993. *Scientific Reasoning: The Bayesian Approach.* 2d ed. Chicago: Open Court.

Isaac, Larry and Larry J. Griffin. 1989. "Ahistoricism in Time Series Analyses of Historical Process: Critique, Redirection, and Illustrations from U.S. Labor History." *American Sociological Review* 54:873-90.

Kass, Robert E., Bradley P. Carlin, Andrew Gelman, and Radford Neal. 1998. "Markov Chain Monte Carlo in Practice: A Roundtable Discussion." *American Statistician* 52:93-100.

King, Gary, Ori Rosen, and Martin Tanner. 1999. "Binomial-Beta Hierarchical Models for Ecological Inference." *Sociological Methods & Research* 28:61-90.

Lange, Kenneth L., Roderick J. A. Little, and Jeremy M. G. Taylor. 1989. "Robust Statistical Modeling Using the *t* Distribution." *Journal of the American Statistical Association* 84:881-95.

Leamer, Edward E. 1978. *Specification Searches: Ad Hoc Inference With Nonexperimental Data.* New York: Wiley.

———. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review* 23:31-43.

Lee, Peter M. 1997. *Bayesian Statistics: An Introduction.* 2d ed. London: Arnold.

Lewis, Steven M. and Adrian E. Raftery. 1999. "Comparing Explanations of Fertility Decline Using Event History Models With Unobserved Heterogeneity." *Sociological Methods & Research* 28:35-60.

Lindley, D. V. and A.F.M. Smith. 1972. "Bayes Estimates for the Linear Model." *Journal of the Royal Statistical Society, Series B* 34:1-41.

Longford, Nicholas T. 1993. *Random Coefficient Models.* Oxford: Clarendon Press.

Moore, David. 1997. "Bayes for Beginners? Some Reasons to Hesitate." *American Statistician* 51:254-61.

Raftery, Adrian. 1986. "Choosing Models for Cross- Classifications." *American Sociological Review* 51:145-46.

———. 1995. "Bayesian Model Selection in Social Research." *Sociological Methodology* 25:111-64.

Rodríguez, Germán and Noreen Goldman. 1995. "An Assessment of Estimation Procedures for Multilevel Models With Binary Responses." *Journal of the Royal Statistical Society, Series A* 158:73-89.

Smith, A.F.M. and A. E. Gelfand. 1992. "Bayesian Statistics Without Tears: A Sampling-Resampling Perspective." *American Statistician* 46:84-88.

Spiegelhalter, David, Andrew Thomas, Nicky Best, and Wally Gilks. 1997. *BUGS 0.6: Bayesian Inference Using Gibbs Sampling.* Cambridge: MRC Biostatistics Unit.

Tanner, Martin A. 1993. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions.* 2d ed. New York: Springer-Verlag.

Weakliem, David. 1999. "A Critique of the Bayesian Information Criterion for Model Selection." *Sociological Methods & Research* 27:359-97.

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

Western, Bruce. 1996. "Model Uncertainty in Macrosociology." *Sociological Methodology* 26:165-92.

———. 1997. *Between Class and Market: Postwar Unionization in the Capitalist Democracies.* Princeton, NJ: Princeton University Press.

———. 1998. "Causal Heterogeneity in Comparative Research: A Bayesian Hierarchical Modelling Approach." *American Journal of Political Science* 42:1233-59.

Western, Bruce and Kieran Healy. Forthcoming. "Explaining the OECD Wage Slowdown: Recession or Labor Decline?" *European Sociological Review.*

Western, Bruce and Simon Jackman. 1994. "Bayesian Inference for Comparative Research." *American Political Science Review* 88:412-23.

Bruce Western is an associate professor of sociology at Princeton University. In addition to Bayesian statistics, his interests include methods for comparative research and sociological analysis of labor markets. His most recent book is Between Class and Market *(Princeton University Press, 1997).*