4

BAYESIAN MODEL SELECTION IN SOCIAL RESEARCH

*Adrian E. Raftery**

It is argued that *P*-values and the tests based upon them give unsatisfactory results, especially in large samples. It is shown that, in regression, when there are many candidate independent variables, standard variable selection procedures can give very misleading results. Also, by selecting a single model, they ignore model uncertainty and so underestimate the uncertainty about quantities of interest. The Bayesian approach to hypothesis testing, model selection, and accounting for model uncertainty is presented. Implementing this is straightforward through the use of the simple and accurate BIC approximation, and it can be done using the output from standard software. Specific results are presented for most of the types of model commonly used in sociology. It is shown that this approach overcomes the difficulties with *P*-values and standard model selection procedures based on them. It also allows easy comparison of nonnested models, and permits the quantification of the evidence *for* a null hypothesis of interest, such as a convergence theory or a hypothesis about societal norms.

This research was supported by NIH grant no. 5R01HD26330. I would like to thank Robert Hauser, Michael Hout, Steven Lewis, Scott Long, Diane Lye, Peter Marsden, Bruce Western, Yu Xie, and two anonymous reviewers for detailed comments on an earlier version. I am also grateful to Clem Brooks, Sir David Cox, Tom DiPrete, John Goldthorpe, David Grusky, Jennifer Hoeting, Robert Kass, David Madigan, Michael Sobel, and Chris Volinsky for helpful discussions and correspondence. I may be contacted by email at raftery@stat.washington.edu.

*University of Washington

1. INTRODUCTION

P-values and significance tests based on them have traditionally been used for statistical inference in the social sciences. In the past 15 years, however, some quantitative sociologists have been attaching less importance to *P*-values because of practical difficulties and counterintuitive results.

These difficulties are most apparent with large samples, where *P*-values tend to indicate rejection of the null hypothesis even when the null model seems reasonable theoretically and inspection of the data fails to reveal any striking discrepancies with it. Because much sociological research is based on survey data, often with thousands of cases, sociologists frequently come up against this problem. In the early 1980s, some sociologists dealt with this problem by ignoring the results of *P*-value-based tests when they seemed counterintuitive and by basing model selection instead on theoretical considerations and informal assessment of discrepancies between model and data (e.g., Fienberg and Mason 1979; Hout 1983, 1984; Grusky and Hauser 1984).

Then, in 1986, Bayesian hypothesis testing was brought to the attention of sociologists, particularly using the simple BIC approximation (Schwarz 1978; Raftery 1986*b*). This seemed to lead to intuitively reasonable results when *P*-values did not, and retrospectively validated some of the "common sense" decisions made in spite of *P*-values by the researchers mentioned above. As a result, BIC has become quite popular for model selection in sociology, particularly in log-linear and other models for categorical data.

Two other difficulties with the use of *P*-values for model selection are also prevalent in sociology, although they are less obvious. They arise when many statistical models are implicitly considered in the earlier stages of a data analysis. This happens when many possible control variables are measured, and one must decide which ones to include in the final model. Often this choice is made using a strategy that involves a collection or sequence of *P*-value-based significance tests, either informally by screening the *t*-values in the full model with all variables included and removing the least significant ones, or more formally by stepwise regression and its variants.

The first difficulty is that *P*-values based on a model selected from among a large set of possibilities no longer have the same interpretation that they did when only two models were ever considered (Miller 1984, 1990). Indeed, the use of *P*-values following

model selection can be dramatically misleading (Freedman 1983; Freedman, Navidi, and Peters 1988).

The second difficulty is that several different models may all seem reasonable given the data but nevertheless lead to different conclusions about questions of interest. This can happen even when the dataset is moderately large, and striking examples have been observed in educational stratification (Kass and Raftery 1995) and epidemiology (Raftery 1993*b*). In this situation, the standard approach of selecting a single model and basing inference on it underestimates uncertainty about quantities of interest because it ignores uncertainty about model form.

The Bayesian approach to model selection and accounting for model uncertainty overcomes these difficulties. It was first used in sociology in 1986 purely as a model selection criterion, and since then it has been widely applied. Here my aim is to give the rationale behind it, to show how it avoids the problems that plague *P*-values, to explain how it can be used to account for model uncertainty as well as to select a single "best" model, and to give some guidelines on its practical implementation for specific model classes.

In Section 2 I review some of the practical difficulties with *P*-values in empirical research and give examples. In Section 3 I give the basic ideas of Bayesian hypothesis testing and Bayes factors. In Section 4 I derive the BIC approximation and equivalent expressions useful for specific models used in social research. I discuss the interpretation of BIC and why it sometimes leads to different conclusions than *P*-values. In particular, BIC tends to favor simpler models and null hypotheses more than do *P*-values in large data sets. In Section 5 I show how the Bayesian approach can be used to account for model uncertainty, and in Section 6 how it resolves the difficulties with *P*-values discussed in Section 2. In Section 7 I discuss modeling strategies, and in the Appendix I describe some valuable software.

2. PRACTICAL DIFFICULTIES WITH *P*-VALUES

2.1. *P*-values

The standard statistical approach to hypothesis testing assumes that only two hypotheses, H_0 and H_1, are envisaged, and that one of these, the null hypothesis H_0, is nested within the other one. The alternative hypothesis H_1 is represented by a probability model with

TABLE 11
Fit of Models for the Four-Way Table of U.S. Mobility 1972–1985 (n = 9,227).

Model	Marginals Fitted	Deviance	d.f.	BIC
1 Table 4, model 3	$[SPO][SD]$	2653	1066	−7079
2 Table 4, model 10	$[SPO][SPD][OD]$	770	781	−6360
3 Table 5, SAT model	$[SP(SAT)]$	1167	990	− 7872

Note: O = origin occupation (17 categories); D = destination occupation (17 categories); S = gender; P = period (3 categories); (SAT) = $[OD]$ interaction parameterized using Hout's (1984) SAT model.

Source: From Hout (1988).

Thus Hout's (1988) iterative model search guided by BIC led to a model that fits better than others and is parsimonious, with each parameter being substantively interpretable. The parameter estimates (Table 5 of Hout [1988]) showed clearly how the associations between origins and destinations changed between 1972 and 1985. This clarity would have been harder to achieve with other, overparameterized, models considered.

8. DISCUSSION

In this chapter I have described the Bayesian approach to hypothesis testing, model selection, and accounting for model uncertainty. Some of the main points I have tried to argue are the following:

- Bayes factors provide a better assessment of the evidence for a hypothesis than P-values, particularly with large samples.
- Bayes factors allow the direct comparison of *nonnested* models, in a simple way.
- Bayes factors can quantify the evidence *for* a null hypothesis of interest (such as a convergence hypothesis or a theory about societal norms). They can distinguish between the situation where a null hypothesis is not rejected because there is not enough data, and that where the data provide evidence for the null hypothesis.
- BIC (or BIC′) provides a simple and accurate approximation to Bayes factors.
- When there are many candidate independent variables, standard model selection procedures are misleading and tend to find strong

evidence for effects that do not exist. By conditioning on a single model, they also ignore model uncertainty and so understate uncertainty about quantities of interest.

- Bayesian model averaging enables one to take into account model uncertainty and to avoid the difficulties with standard model selection procedures.
- The Occam's window algorithm is a manageable way to implement Bayesian model averaging, even with many models, and allows effective communication of model uncertainty.
- BIC can be used to guide an iterative model selection process.
- The methods described here can be implemented using only the output from standard statistical model-fitting software.
- Some software to implement Bayesian model averaging automatically is available.

I know of no non-Bayesian way of dealing with the model uncertainty problem. One proposal is to bootstrap the entire model-building process, including model selection. However, there is no theoretical justification for this, and Freedman, Navidi, and Peters (1988) have shown that it does not give satisfactory results. The same is true of the jackknife.

Bayesian model selection does not remove the need to check whether the models chosen fit the data. Even if many models are considered initially, they may *all* be bad! Thus diagnostic checking, residual analysis, graphical displays, and so on, all remain essential.

I have emphasized the difficulties with P-value-based tests in large samples, but there are difficulties also in small samples, such as arise especially in macrosociology. There, tests at a .05 level often fail to reveal any effects, which has been a source of frustration for those doing comparative and historical research (e.g., see Ragin 1987). The use of BIC corresponds to a particular sample-size-dependent choice of significance level and, as Table 9 shows, for samples sizes below about 50, that level is *greater* than .05. Thus with small samples BIC is actually *less* stringent than significance tests at a .05 level, and so BIC may provide a more satisfactory basis for the use of statistical models in comparative and historical research, as well as other areas with small samples.

BIC was introduced as a large-sample approximation to the

Bayes factor, and one may ask how large the sample has to be for it to be used.[12] That question remains to be answered, but in empirical investigations Raftery (1993*b*) found BIC to be quite accurate in examples with as few as about 40 observations. Small and unreported numerical experiments suggest it to be surprisingly accurate even for much smaller samples than that, but more research is needed on this issue. For generalized linear models, the much more accurate approximation of Raftery (1993*b*) can be used with small samples; this is implemented in the GLIB software described in the appendix to this chapter.

I have focused on the choice of independent variables in regression and related models in this chapter. However, model selection is much broader than this and also includes such modeling decisions as the coding of variables, the choice of functional forms and variable transformations, error distributions, and whether or not to remove outliers. The general framework of Bayesian model selection can be applied to these problems also. For a practical implementation of Bayesian model selection in linear regression to include the choice of independent variables, variable transformations and outlier removal, see Hoeting (1994).

What is the role of theory in all of this? Theory is essential and should be used to the greatest possible extent to define the model to be used. Indeed, the ideal situation is one in which there is no model uncertainty whatever. This ideal is sometimes approached, especially in the study of topics on which there has already been a great deal of research. Unfortunately, however, theory is often weak and vague, and does not fully specify which control variables should be included, which functional forms should be used, what the distribution of the error term is, and so on. This is often particularly the case when there has not been much previous research on the phenomenon under study. Statistical methods for model selection and accounting for model uncertainty should be used only to address issues left unresolved by theory. Bayesian model selection is not an all-purpose panacea: strong theory, clear conceptualization and careful measurement remain vital for successful social research.

[12]Bayesian model selection itself in its exact form places no restrictions on sample size, and can be used validly with even a single observation (although in that case it is unlikely to reveal much evidence for or against any model!).