Evaluating and Using Statistical Methods in the Social Sciences

A Discussion of "A Critique of the Bayesian Information Criterion for Model Selection"

ANDREW GELMAN
Columbia University
DONALD B. RUBIN
Harvard University

The Bayesian information criterion (BIC) can be a helpful statistical tool in sociology and elsewhere (for discussion and examples, see Raftery 1995; Kass and Raftery 1995). However, David L. Weakliem (1999 [this issue]) presents several powerful criticisms, both theoretical and applied, of the BIC, which are similar to critical issues discussed in Gelman and Rubin (1995).

Weakliem's article makes three main points. First, the BIC corresponds to a specific probability model that, in many examples, does not make good scientific sense. Second, the mathematical form of the BIC can lead to pathologies with respect to sample size, data grouping, and model formulation, as illustrated with several real examples in Weakliem's article. Third, in the case of the social mobility data used as an example by Raftery (1986, 1995), using the BIC to select a model leads to scientifically misleading conclusions. This third point is in fact a problem with *any* automatic model selection method and, as Weakliem notes, should be interpreted as a criticism not specifically of the BIC but rather of any blind data reduction technique or "Occam's razor" argument applied to data without consideration of the scientific context.

AUTHORS' NOTE: *This work was supported in part by the U.S. National Science Foundation Grants SBR-9708424, SBR-9709359, and Young Investigator Award DMS-9457824.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 27 No. 3, February 1999 403-410

EVALUATING THE APPLIED USE OF A STATISTICAL METHOD

In general, how should one evaluate the appropriateness of statistical methods for application? We consider the relevance of both analytical and empirical approaches to the original justifications of the BIC by Raftery and others, as well as to Weakliem's criticisms of the BIC.

One can generally use both analytical and empirical approaches to evaluate applied statistical methods, but, broadly speaking, empirical evaluations and narratives of applications are ultimately more important determinants of how useful we view a statistical method to be in a specific application context. Although the more theoretical considerations are also useful, theoretical beauty alone is certainly not enough to justify the use of a statistical method, and theoretical weaknesses alone do not make a method useless. Of course, the identification of theoretical flaws in a model implies that theoretical improvements are possible, which should lead to applied improvements as well, however minor.

ANALYTICAL APPROACHES

Evaluation of Implicit or Explicit Assumptions

One general approach to evaluating a statistical method is to determine a set of assumptions (explicit or implicit) that underlie it and then to assess the reasonableness of those assumptions in the context of the applications in which the method is used. In some of the physical sciences (but rarely, unfortunately, in social science), one can also check whether the assumptions are consistent with "known" physical laws.

Evaluations of assumptions are certainly relevant when considering the motivations for and criticisms of the BIC. To begin with, the method was motivated theoretically as Bayesian and thus consistent with a probability model (see Kass and Raftery 1995). Weakliem finds, however, that the BIC's implicit prior distributions often do not make sense in substantive contexts involving the analysis of contingency tables from sociological survey data; an important criticism, since the BIC is often applied in such problems.

Study of the Performance of the Method in Limiting or Idealized Theoretical Settings

More indirectly, one can consider the behavior of the method in various limiting situations (e.g., very large or very small sample sizes) where we have a clear idea of what is appropriate behavior. A related approach is to evaluate the accuracy of the method in theoretical settings that are comparable to the applications being considered. For example, one can evaluate expected squared prediction errors based on the assumption that the observed data are a simple random sample from a larger population, which might be approximated using a jackknife or bootstrap (see Efron and Tibshirani 1993) or explicit analytical distributions.

Theoretical study provided one important motivation of the BIC, in that a key weakness of χ^2 tests is that they always reject any imperfect yet parsimonious model in the limit of large sample sizes, no matter how tiny the imperfections. This theoretical point warns us that it cannot make sense to rely on significance testing *alone* to analyze contingency table data in social science when parsimonious models are sought. This conclusion motivates the use of alternative methods such as the BIC.

However, Weakliem points out that the BIC has some more subtly inappropriate behaviors in limiting situations; for example, in two-sample problems as one sample remains small while the other becomes large (for other theoretical criticisms of the BIC and Bayes factors in general, see Gelman and Rubin 1995; Gelman et al. 1995, sec. 6.5).

An additional theoretical claim raised by proponents of the BIC is that a model selected by the BIC should outperform models selected by other criteria (e.g., by χ^2 tests) in the sense of having smaller expected out-of-sample prediction errors. This result is correct (from Bayes's theorem) *if* the prior distribution is correct, but, as pointed out by Weakliem, the BIC corresponds to one particular choice of prior distribution, and there is no reason for this result to hold if the actual data come from a different model, as in fact appears to be the case with the social mobility data set. For example, a jackknife error analysis of the social mobility data would find that the BIC-selected quasi-symmetry model performs *worse* (in terms of mean squared prediction error) than the saturated model. This does not mean that the BIC is useless here, but it does mean that the claim of lower average prediction error for the BIC is inappropriate.

EMPIRICAL APPROACHES

There are other ways of evaluating a statistical method that are essentially empirical. Most directly, one can consider how the method works on various problems to which it might be applied when the correct answer is known (e.g., see Stigler 1977). Alternatively, one can consider case studies in which the method has been used or could be used and assess the method in the context of the story or narrative of the scientific problem, the statistical methods used, and the substantive conclusion. For example, a typical article in the Applications and Case Studies section of the *Journal of the American Statistical Association* does not merely try to show the reader *how* to use a new method, but also points out the benefits of the method within a (stylized) scientific narrative that typically follows the following steps:

1. Recognizing a scientific or engineering problem,
2. Gathering relevant data,
3. Analyzing the data, and
4. Constructing a new scientific understanding or engineering capability.

It is generally important in such articles to emphasize that if the new statistical method is not used in step 3 (or possibly step 2), then it is not possible to reach step 4.

Narratives of statistical application have played a central role in both the justifications and criticisms of the BIC. Raftery (1995), Kass and Raftery (1995), and others have provided several plausible statistical narratives in which the BIC has led to scientific understanding that might have been eluded had conventional methods such as classical significance tests been used. On the other hand, Gelman and Rubin (1995) and Weakliem find the BIC to be misleading in Raftery's motivating applications. In particular, in the social mobility example, Weakliem has provided a counternarrative in which the BIC misleads by causing the analysts to miss interesting patterns in the departures from quasi-symmetry.

To understand the role of the BIC in this application, we need to combine the two narratives and understand the implications of the combination. This prepares us for an improved role for the BIC, a role that makes allowances for its theoretical limitations.

UNDERSTANDING THE ROLE OF BIC IN THE SOCIAL MOBILITY EXAMPLE

For the social mobility example, a plausible combination of the Raftery and Weakliem narratives yields the following story. With this large data set, the χ^2 test rejects all models considered except for the saturated model. The BIC leads one to prefer the quasi-symmetry model, and choosing this model allows some sociologists to better understand the data. When the data are split by country, however, the BIC favors the saturated model over quasi-symmetry. The evidence against quasi-symmetry becomes even stronger after adjusting the sample size in the BIC to include only the data relevant to the test. Examination of the discrepancy of the data from quasi-symmetry, along with sociological understanding (e.g., comparing farming to other occupations), suggests some asymmetry models that outperform the quasi-symmetry models in terms of the BIC and other model selection rules. At this point, further data (perhaps occupational mobility tables with more than three classes) are needed to understand the implications of the new theory.

In this narrative, the BIC plays a useful role in selecting the quasi-symmetry model as a reasonable candidate (i.e., under certain theoretical assumptions, quasi-symmetry is the most probable model, given the data), but the BIC is harmful when it is used to eliminate further inquiry—the theoretical assumptions of the BIC are not in fact reasonable here, and so we should not take its probability claims too seriously. As Raftery (1995) and Weakliem both discuss, the BIC can be useful when analyzing large data sets because it gives the analyst one possible objective justification for choosing a parsimonious model. The next step is to recognize that the model selected by the BIC still needs to be evaluated on substantive terms, in particular, by evaluating the discrepancies between the data and the model (often using graphical methods, such as those described in Cleveland 1985; Tufte

1983; these methods can be formalized in a Bayesian context by posterior predictive checking, as is discussed by Rubin 1984; Gelman et al. 1995; Gelman, Meng, and Stern 1996).

THE ROLE OF THE SATURATED MODEL IN THE CONTINGENCY TABLE STORY

A paradox remains in this narrative, however. From our point of view, one of the most interesting actors in the contingency table story is the *saturated model*, which seems to appear only as an antagonist, despite its being the best fitting of all the available models in the social mobility example in terms of χ^2 tests (and in terms of minimizing expected jackknife or cross-validated prediction error).

In some sense, the saturated model is certainly *true*, since it allows the parameters to take on any possible values; its low BIC arises from the assumed vague prior distribution on all its parameters. Yet, as Weakliem notes, many sociologists do not take the saturated model seriously. That the saturated model fits the data the best is tautological and tells us essentially nothing of sociological interest, especially since we know that, with a large enough sample size, we will be able to reject any nonsaturated model in practice (since no simpler model is *exactly* true).

We can, however, make the saturated model more useful by estimating aspects of its prior distribution from the data; that is, by constructing a hierarchical model. This would ideally be defined as centered on a substantively reasonable parsimonious model; for example, an additive log linear model comprising the quasi-symmetry model plus independent error terms for the cross-diagonal parameters. An additional parameter—the variance of the discrepancies from quasi-symmetry—would need to be estimated from the data. This hierarchical model automatically reduces to quasi-symmetry when the variance of the discrepancies is zero. More generally, the new model provides a framework for systematically and smoothly exploring discrepancies from the parsimonious model. In addition, the hierarchical model, centered on a substantively interesting model such as quasi-symmetry, fits nicely into the general Bayesian model selection approach of Raftery (1995) and Kass and Raftery (1995).

INCORPORATING THE BIC INTO AN IMPROVED DATA ANALYTIC FRAMEWORK

Our final narrative then proceeds as follows. The BIC is used as a tool for identifying promising parsimonious models; when applied to the social mobility data in a particular way, it selects quasi-symmetry. A χ^2 test shows, however, that the lack of fit cannot be explained by chance, and a saturated hierarchical model is constructed centered on quasi-symmetry; that is, quasi-symmetry plus discrepancies. (Even without a "statistically significant" χ^2 test, there are reasons to construct this hierarchical model. For one thing, an omnibus test such as a χ^2 can miss important discrepancies in particular directions. In addition, a hierarchical model will often lead to more realistic modeling and predictions, especially when generalizing to other countries or conditions beyond the existing data set.) The discrepancies are analyzed using sociological understanding, and promising asymmetric models are identified that explain discrepancies of scientific interest. An updated saturated hierarchical model is constructed centered on this new asymmetric model. The new model can be used to make tests or predictions for new data.

It should be noted that our final narrative does not *yet* advance in sociological terms beyond Weakliem's. It has the advantages, however, of (1) having the potential for more accurate predictions by combining the virtues of the parsimonious and saturated models and (2) allowing one to examine discrepancies from the model more systematically. The BIC played an important role in this final narrative, but not the role of automatically selecting the final model.

REFERENCES

Cleveland, W. S. 1985. *The Elements of Graphing Data*. Monterey, CA: Wadsworth.

Efron, B. and R. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 1995. *Bayesian Data Analysis*. New York: Chapman & Hall.

Gelman, A., X. L. Meng, and H. S. Stern. 1996. "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies [with discussion]." *Statistica Sinica* 6:733-807.

Gelman, A., and D. B. Rubin. 1995. "Avoiding Model Selection in Bayesian Social Research." Pp. 165-74 in *Sociological Methodology 1995*, edited by P. V. Marsden. Oxford: Blackwell.

Kass, R. E., and A. E. Raftery. 1995. "Bayes Factors and Model Uncertainty." *Journal of the American Statistical Association* 90:773-95.

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

———. 1995. "Bayesian Model Selection in Social Research [with discussion]." Pp. 111-95 in *Sociological Methodology 1995*, edited by P. V. Marsden. Cambridge, MA: Blackwell.

Rubin, D. B. 1984. "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician." *Annals of Statistics* 12:1151-72.

Stigler, S. M. 1977. "Do Robust Estimators Work With Real Data [with discussion]." *Annals of Statistics* 5:1055-98.

Tufte, E. R. 1983. *The Visual Display of Scientific Information*. Cheshire, CT: Graphics Press.

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

Andrew Gelman is a professor of statistics at Columbia University. His research interests are in Bayesian statistics, decision analysis, political science, public policy, survey sampling, environmental statistics, statistical graphics, and computation.

Donald B. Rubin is a professor of statistics at Harvard University. His research interests are in causal inference in experiments and observational studies, inference in sample surveys with nonresponse, and in missing data problems, application of Bayesian techniques, and developing and applying statistical models to data.