Avoiding model selection in Bayesian social research*

Andrew Gelman
Department of Statistics
University of California
Berkeley, CA 94720

Donald B. Rubin
Department of Statistics
Harvard University
Cambridge, MA 02138

December 22, 1994

1 Introduction

Raftery's paper addresses two important problems in the statistical analysis of social science data: (1) choosing an appropriate model when so much data are available that standard P-values reject all parsimonious models; and (2) making estimates and predictions when there are not enough data available to fit the desired model using standard techniques.

For both problems, we agree with Raftery that classical frequentist methods fail and that Raftery's suggested methods based on BIC can point in better directions. Nevertheless, we disagree with his solutions because, in principle, they are still directed off-target and only by serendipity manage to hit the target in special circumstances. Our primary criticisms of Raftery's proposals are that (1) he promises the impossible: the selection of a model that is adequate for specific purposes without consideration of those purposes; and (2) he uses the same limited tool for model averaging as for model selection, thereby depriving himself of the benefits of the broad range of available Bayesian procedures.

Despite our criticisms, we applaud Raftery's desire to improve practice by providing methods and computer programs for all to use and applying these methods to real problems. We believe that his paper makes a positive contribution to social science, by focusing on hard problems where standard methods can fail and exposing failures of standard methods.

2 Too much data, model selection, and the example of the $3 \times 3 \times 16$ contingency table with 113,556 data points

There is no such thing as "too much data," but it is possible to have so much data that a test will reject every parsimonious model that is proposed. Raftery gives an example in his Section 2.2 of a

*Discussion of "Bayesian model selection in social research," by Adrian Raftery. To appear in *Sociological Methodology*. We thank Gary King for helpful comments and the National Science Foundation for grants SBR-9223637, SBR-9207456, DMS-9404305, and DMS-9457824.

$3 \times 3 \times 16$ contingency table with 113,556 observations to which several models are fit; all but the saturated model are soundly rejected by the χ^2 test. The real problem in this example is not that the simpler models are rejected—after all, one would not expect them to fit social reality *exactly*—but that χ^2 test results give no useful guidance for (a) selecting an incorrect but parsimonious model to convey sociological insight, and (b) deciding whether the lack of fit of a parsimonious model is a problem in practice.

In order to conduct social science, it is important to use real-world information in the form of (1) scientific theories, prior information, etc., and (2) knowledge of the purposes to which the analysis will be put. Although modeling data can usefully be done using only the first component, both components of information are needed to do model selection.

The issue is, as Raftery notes, the distinction between statistical and practical significance. There are two sources of predictive variability in any model: (a) inherent variance in the model (e.g., Poisson or multinomial variation in a contingency table model, or residual normal variance in a regression model), and (b) uncertainty due to estimation variability and inaccuracies of the model. If the first source of error is much larger than the second for the kind of predictions one has in mind, it can be acceptable to use a model even if one can detect that it does not fit the data. For example, we have no problem accepting Raftery's claim that the quasi-symmetry is useful to Grusky and Hauser because: (a) it "explains most (99.7%) of the deviance," (b) "the differences between observed and expected counts are a small proportion of the total," and (c) it "makes good theoretical sense." But this claim must be predicated on the uses to which the quasi-symmetry model will be put. For making global predictions, the quasi-symmetry model's lack of fit relative to the saturated model is swamped by inherent multinomial variation, and thus is arguably irrelevant.

In contrast, if, we were interested in the way that the countries differ from the typical pattern implied by quasi-symmetry, it would behoove us to move to a more complicated model that fits the data better. The rejection by the χ^2 test is telling us something: the quasi-symmetry model does *not* fit the data; the low P-value means that if the model were true, it would be extremely unlikely for such a poor fit to occur. If the χ^2 test did not have an extreme P-value (for example, if the deviance were 20—instead of 150—on 16 degrees of freedom), this would suggest that there is limited information in this dataset for measuring differences from the quasi-symmetry model.

A social scientist's happiness with quasi-symmetry follows from its pleasing theoretical properties and the realization that its lack of fit to the data is not *substantively* significant for a class of questions of interest (e.g., global predictions). The social scientist need not claim that quasi-symmetry is "better" than the saturated model; it is enough to say that quasi-symmetry explains 99.7% of the variation in the data that can be explained by the saturated model, and that the misfit 0.3% is not in a substantively important direction for a broad class of questions. Raftery writes in Section 6.1 that "Grusky and Hauser decided to ignore the P-value." Rather than ignoring the *reality* that the model does not fit the data, we would rather admit to using an inexactly-fitting model because of its convenience, scientific insights, and general explanatory power for questions we intend to address using it. Nothing needs to be ignored!

Raftery's BIC cannot work, *in principle*, because it purports to deliver the impossible: a rationale for selecting a model that does not fit the data (e.g., quasi-symmetry in the Grusky-Hauser example) over a model that does fit (e.g., the saturated model), *based on the data and theory alone*, without consideration of the questions the model will be used to address. This claim for BIC makes no logical sense, because it attempts to express a concept of "this model is acceptable for our present purposes" in terms of a single probability statement that is blind to what those purposes are.

Then what principled method could lead us to conclude that the quasi-symmetry model is adequate for the intended purposes of Grusky and Hauser? It would make sense to summarize the analysis using the quasi-symmetry results and also the residuals from quasi-symmetry. A decision that the quasi-symmetry model is acceptable for scientific purposes can be based on a scientific judgment that the residuals are small, relative to the size of effects of scientific interest. More generally, deviations from a model can be compared to their *posterior predictive distribution*, a Bayesian generalization of the reference distribution used for classical P-values (Rubin, 1981, 1984). Here, a Bayesian analysis of a posited model (e.g., quasi-symmetry) is used to generate hypothetical replicates of the data under their posterior predictive distribution. If the replicates are "close enough" to the actual data with respect to some measures of discrepancy that reflect the purposes of the analysis, then the posited model is adequate for these purposes. A general discussion of posterior predictive checks is given by Gelman, Meng, and Stern (1994), and applications to social science data include Rubin's (1981) analysis of educational testing experiments and Rubin and Stern's (1995) analysis of latent class models in psychology.

3 How can BIC select a model that does not fit the data over one that does?

In the last sentence of his Section 3, Raftery implies that the model with higher BIC will be expected to yield better out-of-sample predictions than any other model being compared. This implication is not generally true; there is no general result, either applied or theoretical, that implies this. For example, under Raftery's particular implicit assumptions, the quasi-symmetry model is *more probable*, but what does it mean for one model to be more probable than another, larger model when the data show that the smaller model is false? In this example, if predictive accuracy is measured by mean squared error, the saturated model is expected to predict slightly better than quasi-symmetry.

For a simpler example that conveys insight into the implicit assumptions underlying BIC, consider the problem of adding a single parameter to a normal linear model, as discussed in Raftery's Sections 4.3–4.4. For simplicity we consider the one-dimensional problem, with data $y_1, \ldots, y_n$, independent observations from a normal distribution with mean θ and variance 1. Consider the scenario with a large amount of data, $n = 100{,}000$, where the mean of the observations, $\overline{y}$, is 0.01—a small value, but 3.16 standard errors from zero (the standard error is $1/\sqrt{n}$). The P-value of this observation is 0.0016; it is extremely unlikely that data this or more extreme would be observed if $\theta = 0$. The value of the BIC, on the other hand (see Raftery's equation 27), is $3.16^2 - \log(100{,}000) = -1.51$, implying that the probability that $\theta = 0$ is $1/(1 + \exp(-1.51/2)) = 0.68$, despite the fact that the hypothesis $\theta = 0$ is contradicted by the data. In contrast, the correct inference with any relatively diffuse prior distribution on θ is that θ is small, but nonzero; more precisely, the 95% interval is $[0.01 \pm 1.96 \cdot 0.0032]$, which easily excludes zero. Recall the discussion of our Section 2: the simpler model may be acceptable if the deviation of the data from the model is small, but this does not mean that the simpler model is "true." How can BIC conclude that $\theta = 0$ is the better model? The answer lies in the implicit improper prior distribution on θ that is assumed by BIC—a mixture of a point mass at $\theta = 0$ and a uniform density on the real line. (We use the term *improper prior distribution* in its technical sense to refer to an unnormalized probability density that has an infinite integral.) When data contradict a Bayesian posterior distribution, there is something wrong with the modeling assumptions (or the data), and the posterior distribution should not be trusted.

4 Not enough data, model averaging, and the example of regression with 15 explanatory variables and 47 data points

When the number of parameters in a model is large relative to the number of data points, it is well known that Bayesian approaches, which assign a prior distribution to the parameters in the model, can yield parameter estimates and predictions that are better from the frequentist perspective (e.g., James and Stein, 1960, Efron and Morris, 1971, 1972). Different estimation procedures correspond to different prior distributions; for example, "ridge regression" corresponds to a normal prior distribution on the coefficients in a regression model. Raftery's "Occam's window" implicitly corresponds to a prior distribution for each regression coefficient that is a combination of a point mass at zero and a uniform prior distribution on $(-\infty, \infty)$—scientifically a peculiar model. This model will work well in situations in which this prior distribution is a good approximation to reality. For instance, the artificial examples discussed in Raftery's Section 2.3 and 6.2 are set up to be perfect matches for his prior distribution, with all or almost all the regression coefficients *defined* to be exactly zero, and consequently Raftery's method works well there.

More generally, realistic prior distributions in social science do not have a mass of probability at zero. For example, consider the real-data crime-rate example discussed in Sections 2.4 and 6.3. The difficulties in this example arise entirely from the small sample size. If we somehow had 100,000 data points and 15 predictors, there would be no question that we should include all 15 predictors in the regression model, because in this example, Raftery's goal is to produce accurate coefficient estimates, not a parsimonious model as in the Grusky-Hauser example. Any reasonable method, including stepwise regression, will ultimately include all the variables for such a problem if the sample size is large enough.

We agree with Raftery that, in this case, the scientific questions are answered by the estimated coefficients and their posterior distributions—not by their "statistical significance." We also agree that, if a discrete set of models is being fit to a dataset, it is better to average over the models than to pick just one; the latter procedure leads to confidence intervals that are consistently too narrow. For both these reasons, we find Raftery's analysis preferable to that obtained by stepwise regression. An even better approach would be to set up a more realistic model on the coefficients, which would be facilitated by transforming some of the predictors; for example, labor force participation rate could be per adult male under 65, police expenditures and GDP could be per capita, and the two unemployment rates could be recoded as an average and a difference. Moreover after such transformations, a hierarchical model might be more compelling. In his reliance on BIC, Raftery is limiting himself to a very narrow range of peculiar models.

But the largest gains in this example should come from elsewhere. It's not "cheating" to use real-world knowledge if you're actually interested in real-world answers. We see no reason to trust the results of *any* analysis of the 1960 Ehrlich data alone for any questions of long-term social interest. The right thing to do, obviously, is to obtain more data, especially in a problem such as this in which the cost of gathering data seems to be so little: why analyze data only from 1960 (an especially odd choice considering that Ehrlich's paper is dated 1973)? With data from several years, the difficulties of separately estimating 15 regression coefficients essentially vanish. For example, Campbell (1992) estimates a regression model for election forecasting that has over 15 predictor variables by using state-level data from several Presidential elections; also see Gelman and King (1993) for discussion of the political context and Boscardin and Gelman (1995) for a full Bayesian analysis of this example. With several years of data, regression coefficients can be pooled or partially pooled across years (in the same way that coefficients are partially pooled across schools in Rubin, 1980) using Bayesian methods. Other useful steps would be disaggregating the data (e.g., by race, sex, and age) and building an appropriate hierarchical model. Certainly, whether or not this extra information has been obtained, we would not want to restrict analyses to the particular model implied by BIC.

5 Conclusion

So far, we have said almost nothing about model selection, despite the title of Raftery's paper. That is because we believe model selection to be relatively unimportant compared to the task of constructing realistic models that agree with both theory and data. In most cases, we would prefer to fit a complicated model, probably using Bayesian methods—but not BIC—and then summarize it appropriately to answer the substantive questions of interest.

In addition to our disagreements with Raftery about model selection in applied social research, we have some specific theoretical criticisms about his presentation of BIC as "the Bayesian approach to hypothesis testing, model selection, and accounting for model uncertainty." The Bayesian approach is a general one, which we advocate (Gelman, Carlin, Stern, and Rubin, 1995), and it is important to recognize that there is no single Bayesian solution to a statistical problem. Bayesian approaches to the problems posed by multiple models include exact Bayes factors using proper prior distributions; embedding individual models in continuous parameterizations[1]; multilevel hierarchical modeling (see Bock, 1989, for some examples in educational research); and posterior predictive checks, in which models are compared not by posterior probabilities but rather by their predictive accuracy for intended purposes (see Rubin, 1984, and Gelman, Meng, and Stern, 1995).

Moreover, BIC cannot be construed as an approximation to any exact Bayesian solution, even a Bayes factor. In models with improper prior distributions (which include all the examples in Raftery's paper), the Bayes factor is, in fact, undefined! Equation (7) becomes 0/0. This is a serious problem, and it has attracted some interest in the theoretical Bayesian literature (see Spiegelhalter and Smith, 1982, for a discussion of the problem). In Raftery's presentation, this comes as a term of order 1 ("O(1)" in equation 16). It is implied that this is not a problem for large n, but there is another hidden assumption—that this is a fixed number, with some mathematical definition. The mathematics of Raftery's Section 4 obfuscate the key fact that BIC is not an approximation but a *definition*, which helps to explain why no *exact* Bayes factors are computed anywhere in Raftery's article. Raftery is too casual with the use of improper prior distributions across models of differing dimensions.

References

Bock, R. D., ed. (1989). *Multilevel Analysis of Educational Data.* New York: Academic Press.

Boscardin, W. J., and Gelman, A. (1995). Bayesian regression with parametric models for heteroscedasticity. *Advances in Econometrics*, to appear.

Campbell, J. E. (1992). Forecasting the Presidential vote in the states. *American Journal of Political Science* **36**, 386–407.

[1]To illustrate with a simple example from our own research, Boscardin and Gelman, 1995, fit a parametric model of heteroscedasticity that includes unweighted and weighted linear regression as two extreme special cases. A Bayesian analysis—with no model selection or BIC—averages over the continuous heteroscedasticity parameter, giving a fit that is better, and we believe is more accurate about uncertainties, than either of the two extreme models *or* any average of the two.

Efron, B., and Morris, C. (1971). Limiting the risk of Bayes and empirical Bayes estimators—Part I: the Bayes case. *Journal of the American Statistical Association* **66**, 807–815.

Efron, B., and Morris, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators—Part II: the empirical Bayes case. *Journal of the American Statistical Association* **67**, 130–139.

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

Gelman, A., and King, G. (1993). Why are American Presidential election campaign polls so variable when votes are so predictable? *British Journal of Political Science* **23**, 409–451.

Gelman, A., Meng, X. L., and Stern, H. S. (1995). Bayesian model checking using tail area probabilities. Under revision for *Statistica Sinica.*

James, W., and Stein, C. (1960). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium 1*, 361–380. Berkeley: University of California Press.

Rubin, D. B. (1980). Using empirical Bayes techniques in the law school validity studies (with discussion). *Journal of the American Statistical Association* **75**, 801–827.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics* **6**, 377–401.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics* **12**, 1151–1172.

Rubin, D. B., and Stern, H. S. (1995). Testing in latent class models using a posterior predictive check distribution. In *Analysis of Latent Variables in Developmental Research*, ed. A. Von Eye and C. Clogg. Sage Publications.

Spiegelhalter, D. J., and Smith, A. F. M. (1982). Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society B* **44**, 377–387.