Rejoinder: Model Selection is Unavoidable in Social Research

Adrian E. Raftery

Sociological Methodology, Vol. 25 (1995), 185-195.

Stable URL:
http://links.jstor.org/sici?sici=0081-1750%281995%2925%3C185%3ARMSIUI%3E2.0.CO%3B2-8

Sociological Methodology is currently published by American Sociological Association.

REJOINDER: MODEL SELECTION IS UNAVOIDABLE IN SOCIAL RESEARCH

*Adrian E. Raftery**

1. INTRODUCTION

I would like to thank Hauser and Gelman and Rubin for their thoughtful comments.

Hauser's discussion is very useful because it identifies new ways in which Bayesian model selection can shed light on scientific debates, and because it points to directions for further research.

Gelman and Rubin and I agree that classical methods fail, that Bayesian model selection can point in better directions, and that Bayesian model averaging is better than using a single model. We also have disagreements, however. I have found model selection to be an essential part of the task of building a realistic model in social research, while they view it as "relatively unimportant." We have different views of what it means for a model "not to fit the data." Also, Gelman and Rubin suggest in several places that BIC is based on a uniform, improper prior, but this is not the case. Several other points are discussed below.

2. RESPONSE TO HAUSER

Hauser's reanalysis of Rytina's (1992) occupational scoring system is striking because Bayesian model selection provides a simple resolution of the controversy, which was later borne out by a new data set. BIC did choose the model with better out-of-sample predictive performance.

I am grateful to Robert E. Kass and David Madigan for helpful discussions during the preparation of this rejoinder.

*University of Washington

NOTE: References to sections and tables are to my chapter unless otherwise stated.

A reanalysis of the Jencks, Perman and Rainwater (1988) Index of Job Desirability (IJD) along the lines that Hauser suggests would be worthwhile. There it is important to select only variables for which there is a good deal of evidence because of the costs of data collection.

Hauser points out that model selection is broader than the choice of independent variables in regression, and also includes choices of constraints on slopes, means, and variances in structural equation models. His own research has shown how these choices can correspond to important questions. He is right to say that the basic ideas of Bayesian model selection can be applied to those problems also: each combination of choices corresponds to a different model for the data, and these can be compared using the Bayesian ideas I have described.

Care should be taken with such extensions, however, as the theory behind them needs to be checked for each new class of models. It is always important to check that the model is regular; if not the BIC approximation can fail (e.g., Findley 1991). Also, the value of "n" to be used is not always clear, as shown by my own shifting recommendations for the structural equation model case.

Hauser also points out that with small samples the use of BIC corresponds to a *higher* significance level than conventional ones. This is the opposite of what happens with large samples; the crossover point for the 5-percent level is around $n = 50$. When $n < 50$ or so, Bayesian model selection is *more* likely to favor an alternative hypothesis than is a significance test at the 5-percent level.

As a result, Bayesian model selection may avoid some of the problems that standard statistical methods have when there are few cases, as often happens in comparative and historical analysis. Ragin (1987) has argued that standard statistical methods are not applicable in such areas, because the number of cases is too small to attain conventional levels of significance, and because such research often involves confronting competing theories that cannot be represented by nested statistical models. However, Bayes factors *can* quantify the evidence for one model against another even when there are few cases and the models are not nested. Thus the Bayesian approach may provide some relief to comparative and historical sociologists struggling to make inferences from small datasets. See Western (1994) for related discussion.

3. RESPONSE TO GELMAN AND RUBIN

3.1. *Points of Agreement*

Gelman and Rubin and I agree with one another (1) that classical frequentist methods fail and that my suggested methods based on BIC can point in better directions; and (2) that it is better to average over a discrete set of models than to pick just one. These points imply that standard significance tests, *P*-values, stepwise regression, and related model-building tools, which have long been basic to quantitative social research, are flawed. We further agree that Bayesian thinking leads to better methods that do not suffer from these flaws. After that, however, we part company.

3.2. *Can Model Selection Be Avoided?*

Gelman and Rubin write: "We believe model selection to be relatively unimportant compared to the task of constructing realistic models that agree with both theory and data." I agree that the task of constructing realistic models is primary, but in social research this task often involves many choices (e.g., control variables, error distributions, coding of variables, and other, more subtle, choices such as those mentioned by Hauser), each combination of which defines a different model. Thus model selection is an important and unavoidable part of the task. The way model selection is done can have a big effect on the conclusions reached, as illustrated in Table 5.

Sociologists do acknowledge the importance of model selection in their writings. Many articles in the *American Sociological Review,* for example, present several statistical models for their data. This gives the reader a sense of both the model-building process and of the model uncertainty that is present in the data. Also, research is often aimed at comparing rival theories, each of which may be represented by a statistical model. If that is the case, model comparison (and, if possible, model selection) is an intrinsic goal of the work.[1]

Gelman and Rubin say that in most cases they would prefer to fit a single complicated model. In the Grusky-Hauser example, they say: "If we were interested in the way that the countries differ from

[1]This argument is developed by Kass and Raftery (1995), who give five detailed examples from different areas of science.

the typical pattern implied by quasi-symmetry, it would behoove us to move to a more complicated model that fits the data better." However, this is just what was done in my chapter, following Grusky and Hauser (1984), who ended up adopting model 5 of Table 2, as discussed in Section 7. This is the model favored by BIC. The model-building process is well described in Hauser's discussion, which makes it clear that model selection was a vital part of it.

In the chosen model, the country-specific mobility parameters are allowed to vary systematically as functions of country-specific variables.[2] Even after fitting this model, the question of whether it is better than the quasi-symmetry model remains, and BIC provides one answer to it. Thus Grusky and Hauser (1984) provide a good example of the construction of a realistic and theory-consistent model, and it involves a good deal of model selection along the way.

3.3. *What Does It Mean to Say That a Model "Does Not Fit the Data"?*

Gelman and Rubin ask, "How can BIC select a model that does not fit the data over one that does?" To see whether this correctly describes what BIC does in the Grusky-Hauser example, we have to ask what they mean by saying that the quasi-symmetry model "does not fit the data." Gelman and Rubin use this phrase to mean that the classical frequentist likelihood ratio test for the quasi-symmetry model against the saturated model rejects the quasi-symmetry model. Given that Gelman and Rubin accept that classical frequentist methods fail, it seems incongruous that they would use results from them to decide that a model does not fit the data.

The likelihood ratio test in this example is subject to the criticisms summarized in Section 4.4. In particular, it asks whether data as extreme *or more so* would be likely to be observed if the quasi-symmetry model were true. It is the "*or more so*" that is the rub here. It is true that more extreme data were, on average, unlikely, but they did not occur, so their probability is irrelevant. Bayes factors, by contrast, are based on the probability of observing the data at hand; there is no "or more so." And that makes all the difference.

[2]Conceptually this model is more complicated than the quasi-symmetry one because it involves additional explanatory variables. However, it involves fewer parameters, and so could also be viewed as simpler.

This argument is made in detail by Berger and Sellke (1987), and pithily summarized in the short passage from Jeffreys (1980) that I quoted. Berger and Sellke identified this difference as the main source of the discrepancy between *P*-values and Bayes factors, and hence as the real answer to Gelman and Rubin's question. Any reader seeking to evaluate the arguments in Gelman and Rubin's Section 3 should read Berger and Sellke, who deal with the general issue and who also analyze in detail the normal example that Gelman and Rubin used.

Thus, if we are going to assess the "fit" of the quasi-symmetry model based on a comparison between it and the saturated model, I would argue that the comparison (and hence the assessment of "fit") should be based on Bayes factors and not on the classical test on which Gelman and Rubin based their statement. Given this, the "rejection" of the quasi-symmetry model by the likelihood ratio test does *not* imply that it does not fit the data. This is not to say, of course, that better models do not exist, and indeed an even better model was subsequently found (the model in line 5 of Table 2).

In spite of this, I do feel that *model checking* in the form of residual plots and other diagnostics is useful and indeed essential to identify model inadequacies and suggest improvements. Classical tests can be useful in this context as a rough way of calibrating the diagnostics so as to decide which ones to pay attention to. The posterior predictive checks of Rubin (1984) and Gelman, Meng and Stern (1995) are also useful.

But I do not feel that classical goodness-of-fit tests should be used to decide whether a model "fits the data." Rather, if diagnostics point to a new and possibly improved model, Bayes factors should then be used to decide whether or not to jettison the old model in its favor. Thus Bayes factors can be used to guide the iterative process of building a realistic model, which Gelman and Rubin rightly identify as the primary task.

3.4. *Bayesian Model Averaging Gives Better Out-of-Sample Predictions*

Gelman and Rubin called into question my statements about out-of-sample predictive performance. My main contention is that Bayesian model averaging (described in Section 5) has better out-of-sample

predictive performance than any one model that might reasonably have been selected. The justification for this is (1) a fairly general theoretical result; and (2) a series of empirical studies with broadly similar results. It is also supported by Hauser's Rytina example, and by the interpretation of the Bayes factor as a predictive score.

Madigan and Raftery (1994, eq. 4) gave a theoretical result showing that, on average over models and over datasets, Bayesian model averaging yields better out-of-sample predictions (as measured by Good's [1952] logarithmic predictive score) than any one model that might reasonably be selected.

This result was confirmed empirically in a series of studies using the kind of half-sample cross-validation mentioned by Hauser, for a range of datasets and of model classes: linear regression (Raftery, Madigan and Hoeting 1993), categorical data models (Madigan and Raftery 1994) and event history analysis (Raftery, Madigan and Volinsky 1995). The results from these studies are quite similar: in most cases, Bayesian model averaging improves out-of-sample predictive performance over the best single model, by about the same amount as would be achieved by increasing the sample size by 4 percent.

Does this imply that the single model chosen by Bayesian model selection is expected to have better out-of-sample performance than a model selected any other way? I know of no formal result to that effect as yet, but I would conjecture that something along those lines is true. Kass and Raftery (1995, sect. 3.2) showed that the Bayes factor can be interpreted as favoring the model with the better predictive score. Hauser's Rytina example provides striking support for this conjecture. The model that Rytina selected using standard criteria is very different from (and less parsimonious than) the model selected by BIC, and when the models were compared using *new* data, the BIC-best model did much better.

3.5. *BIC Does Not Correspond to an Improper Prior*

In several places, Gelman and Rubin say that BIC corresponds to a uniform, improper prior. (This is a prior distribution that does not integrate to one, and it can lead to all sorts of problems.) This is not the case. BIC is an approximation to the Bayes factor for a *proper* prior (i.e., one that does integrate to one), and it is especially accu-

rate for the prior of equation (17), as shown by Kass and Wasserman (1995). Equation (18) is the relevant one, rather than equation (16); the difference is the size of the error term. Thus Gelman and Rubin's last paragraph does not apply to the results in my chapter.

Gelman and Rubin take me to task for focusing on BIC and never calculating any exact Bayes factors. I did so because BIC is so easy to compute and is surprisingly accurate. Clearly, exact Bayes factors are preferable, but they are also harder to calculate. Elsewhere I have been involved in developing exact, or nearly exact, Bayes factors for generalized linear models (Raftery 1993), discrete graphical models (Madigan and Raftery 1994), and other models (Kass and Raftery 1995). Indeed, essentially exact Bayes factors for generalized linear models (including linear regression, logistic regression, and log-linear models) are available using the GLIB software described in the Appendix.

3.6. *Taking Account of the Purpose of Model Selection*

Gelman and Rubin say that the way we do model selection should depend on the purposes for which the model is selected. I certainly agree with this in principle. But the fact is that classical model selection methods do not take account of these purposes, and Gelman and Rubin propose no way of doing so either. And for good reason: it is really hard.

Examples where the model selection process has been guided, at least informally, by the purposes of the analysis include Carlin, Kass, Lerch and Huguenard (1992) and Raftery and Zeh (1993). In each of these cases the process was long and very demanding of the time and expertise of experienced statisticians; most social science research projects do not have access to such resources. Some general-purpose rough results are often useful even when the precise purposes of the model are left unstated, to guide early work and as a check on results.[3]

I know of only one systematic proposal for taking account formally of the purposes of the model, due to Kadane and Dickey (1980). This is that one specify the utility for each possible outcome and model selected, and choose the model that maximizes the ex-

[3]This paragraph is based on a personal communication from Rob Kass.

pected utility. There are various problems with this: it requires specification of these utilities, which may be an onerous task, and if the utilities are not fully known, sensitivity to the choice must be assessed. Thus this proposal demands a lot of the user and, perhaps as a result, has not been used very much.

Of course, if the purpose is well-defined, it should be taken into account. However, the use of purpose-specific utilities seems feasible only when (1) the researcher knows the purpose to which his or her results will be put; (2) the costs and benefits associated with plausible combinations of outcome and model selected can be assessed fairly accurately; and (3) the researcher has time to do the additional analysis.

The goal of research is often to *report* the extent to which data provide evidence in favor of or against particular hypotheses, information that is then used by others, perhaps for a variety of purposes. Bayes factors measure that evidence, and so provide a good general-purpose reporting tool. This is a modest but widespread need; surely meeting it is not promising the impossible!

3.7. *The Ehrlich Crime Example*

I am glad that Gelman and Rubin find my analysis of the Ehrlich example preferable to a standard one based on stepwise regression. They propose another alternative that they say would be "even better," including transforming some of the predictors. However, the transformations they suggest have already been done: labor force participation, police expenditures, and GDP have all been standardized by a relevant population (see Ehrlich 1973). They also suggest recoding the two unemployment rates as an average and a difference. This might be a good idea, but whether or not to do it is just one of the many modeling decisions that must be made. The decision is a model selection one that should be made based on the data.

Gelman and Rubin suggest that, rather than selecting a subset of variables, one include them all, using a hierarchical model. Of course this model is itself selected in some way: there are many measured characteristics of states, so why choose these particular 15? Thus, even with their approach, model selection is unavoidable.

I am skeptical that their approach would perform better than the one I have outlined. Raftery, Madigan, and Hoeting (1993) have

done a fully Bayesian analysis of these data, using Bayesian model selection based on exact Bayes factors rather than the BIC approximation. In that paper we assessed our method using half-sample cross-validation of the kind mentioned by Hauser, and found that Bayesian model averaging gave better out-of-sample predictive performance than any one model that could reasonably have been selected. Gelman and Rubin's approach could also be assessed this way and the results compared with ours; I look forward to such a comparison.

Of course, I agree that the real way to make progress in this example is to get more, better, and more recent data. Time-series data are probably needed to answer at least some of the questions being asked. However, often the best data to hand in social science are cross-sectional aggregate data, and it is important to analyze them while waiting for more data to come in.

3.8. *Other Points*

Gelman and Rubin dismiss the simulated examples of Section 2.3 and 6.2, saying that they were set up to be perfect matches for my prior distribution. However, this is not the case: my simulation experiment replicated the one done by Freedman (1983), a non-Bayesian who was not trying to match anyone's prior distribution! As I have already noted, Gelman and Rubin also misstated the prior distribution to which BIC corresponds. Indeed, I found it striking that Occam's window did so well here, *in spite of* the fact that the experiment did *not* correspond to the prior distribution underlying the method.

Gelman and Rubin say that "realistic prior distributions in social science do not have a mass of probability at zero." This is debatable: social scientists are prepared to act *as if* they had prior distributions with point masses at zero, as shown by their willingness to restrict attention *ab initio* to relevant variables and to remove nonsignificant variables from equations. Even if it is true, however, there is no denying that realistic prior distributions in social science often have a mass of probability *near* zero—i.e., social scientists often entertain the possibility that an effect is *small.*

Berger and Delampady (1987) have shown that the Bayes factor for $\beta = 0$, say, is a good approximation to the Bayes factor for β to be *near* zero, in the sense of $|\beta| < \delta$, where δ is fairly small, at

most about one-half of a standard error. Thus the convenient results available with point null hypotheses give a reasonable approximation to those with the interval null hypotheses that some may find more realistic.

REFERENCES

Berger, James O., and Mohan Delampady. 1987. "Testing Precise Hypotheses (with Discussion)." *Statistical Science* 3:317–52.

Berger, James O., and Thomas Sellke. 1987. "Testing a Point Null Hypothesis: The Irreconcilability of *P* Values and Evidence (with Discussion)." *Journal of the American Statistical Association* 82:112–22.

Carlin, Bradley P., Robert E. Kass, J. Lerch, and B. Huguenard. 1992. "Predicting Working Memory Failure: A Subjective Bayesian Approach to Model Selection." *Journal of the American Statistical Association* 87:319–27.

Ehrlich, Isaac. 1973. "Participation in Illegitimate Activities: A Theoretical and Empirical Investigation. *Journal of Political Economy* 81:521–65.

Findley, David F. 1991. "Counterexamples to Parsimony and BIC." *Annals of the Institute of Statistical Mathematics* 43:505–14.

Freedman, David A. 1983. "A Note on Screening Regression Equations." *The American Statistician* 37:152–55.

Gelman, Andrew, Xiao-Li Meng, and Hal S. Stern. 1995. "Bayesian Model Checking Using Tail-area Probabilities." *Statistica Sinica,* forthcoming.

Good, Irving John. 1952. "Rational Decisions." *Journal of the Royal Statistical Society, series B,* 14:107–14.

Grusky, David B., and Robert M. Hauser. 1984. "Comparative Social Mobility Revisited: Models of Convergence and Divergence in 16 Countries." *American Sociological Review* 49:19–38.

Jeffreys, Harold. 1980. "Some General Points in Probability Theory." In *Bayesian Analysis in Econometrics and Statistics,* edited by Arnold Zellner, 451–54. Amsterdam: North-Holland.

Jencks, Christopher S., Lauri Perman, and Lee Rainwater. 1988. "What Is a Good Job? A New Measure of Labor Market Success." *American Journal of Sociology* 93:1322–57.

Kadane, Joseph B., and Dickey, James M. 1980. "Bayesian Decision Theory and the Simplification of Models." In *Evaluation of Econometric Models,* edited by J. Kmenta and J. Ramsey, 245–68. New York: Academic Press.

Kass, Robert E., and Adrian E. Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90:377–95.

Kass, Robert E., and Larry Wasserman. 1995. "A Reference Bayesian Test for Nested Hypotheses with Large Samples." *Journal of the American Statistical Association,* forthcoming.

Madigan, David, and Adrian E. Raftery. 1994. "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window." *Journal of the American Statistical Association* 89:1535–46.

Raftery, Adrian E. 1993. "Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models." *Technical Report 255,* Department of Statistics, University of Washington.

Raftery, Adrian E., David Madigan, and Jennifer A. Hoeting. 1993. "Model Selection and Accounting for Model Uncertainty in Linear Regression Models." *Technical Report no. 262,* Department of Statistics, University of Washington.

Raftery, Adrian E., David Madigan, and Chris T. Volinsky. 1995. "Accounting for Model Uncertainty in Survival Analysis Improves Predictive Performance (with Discussion)." In *Bayesian Statistics 5,* edited by J. M. Bernardo et al. New York: Oxford University Press, forthcoming.

Raftery, Adrian E., and Judith E. Zeh. 1993. "Estimation of Bowhead Whale, *Balaena mysticetus,* population size (with Discussion)." In *Bayesian Statistics in Science and Technology: Case Studies,* edited by C. Gatsonis, J. S. Hodges, R. E. Kass and N. D. Singpurwalla, 160–240. New York: Springer-Verlag.

Ragin, Charles C. 1987. *The Comparative Method.* Berkeley: University of California Press.

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

Rytina, Steve. 1992. "Scaling the Intergenerational Continuity of Occupation: Is Occupational Inheritance Ascriptive After All?" *American Journal of Sociology* 97:1658–88.

Western, Bruce. 1994. "Vague Theory in Macrosociology." Paper presented at the Annual Meeting of the American Sociological Association, Los Angeles, August 1994.