Bayesian Model Selection in Social Research
(with Discussion by Andrew Gelman & Donald B. Rubin, and Robert M. Hauser, and a Rejoinder) [1]

Adrian E. Raftery
University of Washington

September 1994; revised November 1994

Abstract

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 using the simple and accurate BIC approximation, and 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 non-nested 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.

Contents

List of Tables

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 counter-intuitive 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 counter-intuitive, 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, 1986b). 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 as 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 data set is moderately large, and striking examples have been observed in educational stratification (Kass and Raftery, 1995) and epidemiology (Raftery, 1993b). 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 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 available 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 d_1 unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{d_1})$. H_0 can be represented by the same probability model as H_1, but with ν constraints imposed on $\boldsymbol{\theta}$, $g_i(\boldsymbol{\theta}) = 0$ $(i = 1, \ldots, \nu)$. H_0 can represent not only exclusion restrictions such as $\theta_1 = 0$, but also linear restrictions on the parameters of H_1, such as $\theta_1 - \theta_2 = 0$ or nonlinear restrictions such as $\theta_1^2 + \theta_2^2 = 1$ (restrictions such as the latter arise in association models for contingency tables).

A test statistic T is selected and calculated from the data at hand, D; its observed value is denoted by $t(D)$. The null hypothesis H_0 is rejected in favor of the alternative hypothesis H_1 if $t(D)$ is more extreme than would be expected if H_0 were true. This is implemented by choosing a significance level α (conventionally taken to be .05 or .01), and rejecting H_0 if the probability of T being greater than or equal to $t(D)$ is small (i.e. less than α), given that H_0 is true. More formally, H_0 is rejected if

$$P = \Pr[T \geq t(D)|H_0] < \alpha, \tag{1}$$

in which case H_1 is adopted. The quantity P is called the P-value and is often reported as an indication of the strength of the evidence against H_0.

This approach is so widely applied that it is often used without its basis being critically discussed. There are, however, several features worth noticing. A first point is that the significance level α has to be determined. It has become conventional to use $\alpha = .05$ or .01, based on Sir Ronald Fisher's experience with relatively small agricultural experiments (on the order of 30–200 plots). Subsequent advice has emphasized the need to take account of the power of the test against H_1

when setting α, and to balance power and significance in some appropriate way. However, a precise way of doing this is lacking, and this advice seems to boil down to a vague suggestion that α be lower for large sample sizes, a suggestion that is mostly ignored in practice. We will see that for the sample sizes often found in sociology, values of α much lower than the conventional ones can be appropriate.

A second point to note is that the whole standard hypothesis testing framework rests on the basic assumption that only two models are ever entertained. This is far from being the case in most sociological studies, which are often not experimental, and where a wide range of possible control variables must be considered. In practice a selection is made among the variables, and each possible choice represents a different model; with p possible variables the number of candidate models may reach 2^p, which can be huge (e.g. when $p = 15$, $2^p = 32{,}768$). We will see in Section 2.3 that failing to take account of the model selection process can yield very misleading results.

In the following sections I will outline some practical difficulties with P-value-based tests in sociological applications and give examples. I will return to the examples later in Section 6, after outlining the Bayesian approach to the problem.

2.2   Large Samples

Table 1 contains a three-way $3 \times 3 \times 16$ contingency table showing 3×3 social mobility tables for 16 countries, from Grusky and Hauser (1984)[1]. The total sample size is very large, $n = 113{,}556$.

Two hypotheses were of central interest in this study: the hypothesis that mobility flows are the *same* in all industrialized countries (Lipset and Zetterberg, 1959), and the hypothesis that the *patterns* of mobility (but not the actual amounts) are the same. This is the so-called FJH hypothesis (Featherman, Jones and Hauser, 1975), and the postulated common pattern is that of quasi-symmetry. Two other hypotheses are of interest as standards of comparison: the "baseline" hypothesis of independence between father's and son's occupation, and the hypothesis that there is no common pattern of mobility across countries.

Each of these four hypotheses can be represented by a log-linear model for the full three-way table, as explained by Grusky and Hauser (1984). The deviance and degrees of freedom for each model are shown in Table 2. Models 1, 3 and 4 form a nested sequence and so a test of one of these models against the next one takes the difference between their deviances and compares it with a χ^2 distribution with degrees of freedom equal to the difference between the degrees of freedom for the two models. Model 2 is also nested within model 3.[2]

It is clear that models 1 and 2 are unsatisfactory and should be rejected in favor of model 3.[3] By

---

[1]Strangely enough, although these data have been much analyzed in the literature, they have never been published in the open literature. They are provided here to facilitate reanalyses. They were first compiled by Hazelrigg and Garnier (1976), and have recently been reanalyzed by Xie (1992).

[2]The fifth model in Table 2 will be discussed below in Section 7.

[3]Strictly speaking, a test of the Lipset-Zetterberg hypothesis should involve only the nine industrialized countries in the sample, but imposing this restriction does not change the results.

Table 1: Social mobility tables for 16 countries, father's occupation by son's occupation. The categories are white collar, blue collar and farm. *Source:* Grusky and Hauser (1983).

Australia			Belgium			France			Hungary		
292	170	29	497	100	12	2085	1047	74	479	190	14
290	608	37	300	434	7	936	2367	57	1029	2615	347
81	171	175	102	101	129	592	1255	1587	516	3110	3751
Italy			Japan			Philippines			Spain		
233	75	10	465	122	21	239	110	76	7622	2124	379
104	291	23	159	258	20	91	292	111	3495	9072	597
71	212	320	285	307	333	317	527	3098	4597	8173	14833
United States			West Germany			West Malaysia			Yugoslavia		
1650	641	34	3634	850	270	406	235	144	61	24	7
1618	2692	70	1021	1694	306	176	369	183	37	92	13
694	1648	644	1068	1310	1927	315	578	2311	77	148	223
Denmark			Finland			Norway			Sweden		
79	34	2	39	29	2	90	29	5	89	30	0
55	119	8	24	115	10	72	89	11	81	142	3
25	48	84	40	66	79	41	47	47	27	48	29

Table 2: Fit of models to cross-national social mobility data (Grusky and Hauser, 1984). $n = 113,556$.

#	Model	In G&H	Deviance	d.f.	BIC
1	Independence	Table 5, model 1	42970	64	42227
2	Lipset-Zetterberg	Text, p. 22	18390	120	16997
3	Quasi-symmetry	Table 5, model 2	150	16	-36
4	Saturated	—	0	0	0
5	Explanatory	Table 5, model 4	490	46	-43

the standard test, model 3 should also be rejected, in favor of model 4, given the deviance difference of 150 on 16 degrees of freedom, corresponding to a P-value of about 10^{-120}. Grusky and Hauser (1984) nevertheless adopted model 3 because it explains most (99.7%) of the deviance under the baseline model of independence, fits well in the sense that the differences between observed and expected counts are a small proportion of the total, and makes good theoretical sense. This seems sensible, and yet is in dramatic conflict with the P-value-based test.

This type of conflict often arises in large samples, and hence is frequent in sociology with its survey data sets comprising thousands of cases. The main response to it has been to claim that there is a distinction between "statistical" and "substantive" significance, with differences that are statistically significant not necessarily being substantively important. I do not find this distinction to be a satisfactory panacea, and believe that in most cases where the conflict has arisen, including the Grusky-Hauser study, it is due to the miscalibration of statistical significance using P-values, rather than to any real conflict between statistical and substantive significance. When statistical significance is properly calibrated, I have found that such a conflict rarely arises.

2.3 Many Candidate Independent Variables

Most sociological studies are observational and aim to infer causal relationships between a dependent variable and independent variables of interest. To minimize the possibility of observed associations being due to other variables and hence spurious, other independent variables that could induce spurious associations if they were left out are also included in the regression-type models that are used. I will call these "control variables".

But which control variables should be included? Clearly this choice should be guided by theory as far as possible. However, the theory can be somewhat weak and often produces only a rather long "laundry list" of possible control variables suggested by various theoretical arguments. This is especially the case when the study of a social phenomenon is in its early stages and the theory is still weak. Later, when an area of study has matured, the theory tends to be stronger and knowledge of which to control for tends to be firmer, based on the accumulated research of a community of investigators.

Typically, some choice is made and results with one or more subsets of the laundry list are presented. One would like to make the choice on theoretical grounds, but there is usually little basis for this, as the theory or theories have already been used to establish the initial laundry list and often do not provide a basis for excluding variables from it. It is well known that including a control variable will not affect the estimation of the coefficient of the main independent variable of interest if the control variable is statistically independent of it or of the dependent variable. It would be nice to be able to use this fact to eliminate unnecessary control variables, but such independence usually is not known *a priori* and has to be assessed from the data.

We therefore have to fall back on statistical methods for choosing the control variables. Various methods are in common use. One is to always include the full laundry list. When this is long,

however, and includes many variables that have little or no effect, the precision of estimates of parameters of interest can be hurt (e.g. Bishop, Fienberg and Holland, 1975, pp. 310–315); see Section 2.4 for an example.

Another common approach is to first fit the full model, screen the t-statistics for the parameters, remove the variables for which these are small, and then reestimate the resulting, reduced, model. I will call this the "screening" method. A further method (included in many statistical software packages) is stepwise variable regression, in which variables are added one at a time starting from the null model (forward selection), eliminated one at a time starting from the full model (backwards elimination), or a mixture of the two, such as Efroymson's stepwise regression algorithm. Other methods include minimizing Mallows' C_p and maximizing the adjusted R^2; see Miller (1990) for an account of these and other variable selection methods in regression.

What these methods have in common is that they select one model out of the many possibilities, and then proceed as if that were the only model that had ever been considered. This can yield very misleading results, as pointed out by Freedman (1983), Freedman, Navidi and Peters (1988), Fenech and Westfall (1988) and Miller (1984, 1990). The reason is that by choosing among a large number of models one increases the probability of finding "significant" variables by chance alone. The sampling properties of these model selection methods (as distinct from those of the individual tests that make them up) are unknown in general, and there is little theoretical rationale for preferring one of the methods to the others, although they often give different answers to the questions of interest; see Section 2.4.

This is clearly illustrated by a simple simulation experiment of Freedman (1983), which is similar in several respects to typical sociological studies. In his words:

> A matrix was created with 100 rows (data points) and 51 columns (variables). All the entries in this matrix were independent observations drawn from the standard normal distribution. The 51st column was taken as the dependent variable Y in a regression equation; the first 50 columns were taken as the independent variables $X_1, \ldots, X_{50}$. By construction, then, Y was independent of the X's. Ideally, R^2 should have been insignificant, by the standard F test. Likewise, the regression coefficients should have been insignificant, by the standard t test.

I replicated his experiment and obtained similar results to his. The data were analyzed in two ways, representing perhaps the two most common approaches to variable selection in sociology. The first way consisted of two passes. In the first pass, Y was regressed on all 50 of the X's, with the following results:

- $R^2 = 0.60$, $P = 0.09$;

- 21 coefficients out of the 50 were significant at the .25 level (i.e. $|t| > 1.15$);

- 7 coefficients out of the 50 were significant at the .05 level (i.e. $|t| > 1.99$).

Only the 21 variables whose coefficients were significant at the .25 level were included in the second pass. The results were:

Table 3: Stepwise regression results for simulated noise.

Variable	Coefficient	t	P
Intercept	0.01	0.05	.956
X_8	0.30**	2.80	.006
X_{16}	−0.23*	−2.00	.049
X_{36}	−0.23*	−2.16	.034
X_{42}	0.34**	2.84	.006

* $P < .05$
** $P < .01$

- $R^2 = 0.50$; $P = 0.00001$;

- 20 coefficients out of the 21 were significant at the .25 level;

- 14 coefficients out of the 21 were significant at the .05 level;

- 6 coefficients out of the 21 were significant at the .01 level.

In addition, a battery of diagnostic displays and tests (e.g. Weisberg, 1985) showed no evidence of model inadequacy such as outliers, nonlinearity, heteroscedasticity or autocorrelation in the residuals.

In the words of Freedman, "the results from the second pass are misleading indeed, for they appear to demonstrate a definite relationship between Y and the X's, that is, between noise and noise." Nevertheless, this sort of procedure is often followed in sociology (and laundry lists of 50 variables are not atypical), and many a social researcher would feel confident about presenting such findings.

Stepwise regression does not help. Table 3 shows the results: a four-variable model with $R^2 = 0.18$ and $P = 10^{-6}$, and coefficients that are all significant at the .05 level (with two also significant at the .01 level). The minimum C_p and adjusted R^2 methods also lead to models with too many predictors and highly significant F statistics.

2.4 Model Uncertainty

When many models are initially considered, it often happens that several of them fit the data almost equally well, or that different models are arrived at by different model selection methods. It can then happen that different models, all of them defensible, lead to different answers to the main questions of interest.

The analyst then has three main options. The first is to pick one model and adopt the conclusions that flow from it rather than from the other defensible models; this is somewhat arbitrary. The second option is to present the analyses based on all the plausible models without choosing

between them; while not fully satisfactory, this seems better than the first option. The third possibility, which I will develop in later sections, is to take account explicitly of model uncertainty when drawing conclusions.

To show how the problem can arise, consider the criminological study by the economist Isaac Ehrlich (1973), which was one of the earliest systemmatic efforts to determine whether greater punishments reduce overall crime rates. Up to the 1960s, criminal behavior was traditionally viewed as deviant and linked to the offender's presumed exceptional psychological, social or family circumstances. Becker (1968) and Stigler (1970) argued, on the contrary, that the decision to engage in criminal activity is a rational choice determined by its costs and benefits relative to other (legitimate) opportunities. Ehrlich (1973) developed this argument theoretically, specified it mathematically, and tested it empirically using aggregate data from 47 U.S. states in 1960. Errors in Ehrlich's empirical analysis were corrected by Vandaele (1978) who gave the corrected data, which we use here[4].

Ehrlich's theory goes as follows. The costs of crime are related to the probability of imprisonment and the average time served in prison, which in turn are influenced by police expenditures, which may themselves have an independent deterrent effect. The benefits of crime are related to both the aggregate wealth and income inequality in the surrounding community. The expected net payoff from alternative legitimate activities is related to educational level and the availability of employment, the latter being measured by the unemployment and labor force participation rates. This payoff was expected to be lower (in 1960) for nonwhites and for young males than for others, so that states with high proportions of these were expected also to have higher crime rates. Vandaele (1978) also included an indicator variable for southern states, the sex ratio, and state population as control variables.

We thus have 15 candidate predictors of crime rate (Table 4), and so potentially $2^{15} = 32,768$ different models. As in the original analyses, all analyses were done in terms of the natural logarithms of the variables. Standard diagnostic checking did not reveal any striking violations of the assumptions underlying normal linear regression.

Interest focuses on the significance and size of the coefficients for variables 14 and 15, respectively the probability of imprisonment and the average time served in state prisons. Ehrlich (1973) did not use statistical model selection methods, but instead analyzed two regression models chosen in advance on theoretical grounds.

Table 5 shows results from six models selected using methods discussed so far. The statistically chosen models 2, 3 and 4 all give high and similar values of R^2 and share many of the same variables, while Ehrlich's theoretically chosen models 5 and 6 fit less well. There are striking differences, indeed conflicts, between the results from different models. Even the statistically chosen models, despite their superficial similarity, lead to conflicting conclusions about the main questions of interest.

---

[4]Ehrlich's study has been much criticized (e.g. Brier and Fienberg, 1980) and here I use it purely as an illustrative example. For economy of expression, I will use causal language and speak of "effects", even though the validity of this language for these data is dubious.

Table 4: Variables in crime data

#	Variable
1	percent of males 14–24
2	indicator variable for southern state
3	mean years of schooling
4	police expenditure in 1960
5	police expenditure in 1959
6	labor force participation rate
7	number of males per 1000 females
8	state population
9	number of nonwhites per 1000 people
10	unemployment rate of urban males 14-24
11	unemployment rate of urban males 35-39
12	GDP
13	income inequality
14	probability of imprisonment
15	average time served in state prisons

Table 5: Models selected for the crime data.

#	Method	Variables	R^2 (%)	# vars.	$\hat{\beta}_{14}$	$\hat{\beta}_{15}$	P_{15}
1	Full model	All	87	15	$-.30$	$-.27$	.133
2	Stepwise regression	1,3,4,9,11,13,14	83	7	$-.19$	—	—
3	Mallows' C_p	1,3,4,9,11,12,13,14,15	85	9	$-.30$	$-.30$	.050
4	Adjusted R^2	1,3,4,7,8,9,11,12,13,14,15	86	11	$-.30$	$-.25$	.129
5	Ehrlich model 1	9,12,13,14,15	66	5	$-.45$	$-.55$	.009
6	Ehrlich model 2	1,6,9,10,12,13,14,15	70	8	$-.43$	$-.53$	.011

NOTE: P_{15} is the P-value from a two-sided t-test for testing $\beta_{15} = 0$.

Consider first the effect of X_{14}, the probability of imprisonment, on the crime rate. All analyses and models concur in saying that this does have an effect, so interest focuses on estimating its size. To aid interpretation, recall that all variables have been logged, so that $\beta_{14} = -.30$ means roughly that a 10% increase in the probability of imprisonment produces a 3% reduction in the crime rate, all else being equal. The estimates of β_{14} fluctuate wildly between models. The stepwise regression model gives an estimate that is about one-third lower in absolute value than the full model, a difference that may be large enough to be of policy importance; this difference is equal to about 1.7 standard errors. The Ehrlich models give estimates that are about one-half higher than the full model, and more than twice as big as those from stepwise regression (in absolute value). There is clearly considerable model uncertainty about this parameter.

Another point of interest, not shown in Table 5, is that the standard error of $\hat{\beta}_{14}$ (and also of the other coefficients) is smaller for the more parsimonious models. For the full model it is .098, while for the stepwise regression model it is .066. Thus it could be argued that retaining the additional, non-significant, variables in the full model reduces the efficiency of estimation of β_{14} by a factor of $(.066/.098)^2 = .45$, and so is equivalent to throwing away more than half the data.

Now let us turn to β_{15}, the effect of the average time served in state prisons. Whether this is significant at all is not clear, and t-tests based on different models lead to different conclusions. In the full model it has a non-significant P-value of .133, while stepwise regression leads to a model that does not include the variable at all. On the other hand, Mallows' C_p leads to a model in which it is just significant at the .05 level, while with adjusted R^2 it is again not significant. In Ehrlich's models, by contrast, it is highly significant.

Together these results paint a confused picture about β_{15}, and there seem to be no frequentist results to help sort it out. I will argue that the confusion can be resolved by taking account explicitly of the model uncertainty.

2.5   Non-nested Hypotheses, and Evidence *for* the Null Hypothesis

Often, in sociology, competing hypotheses represent quite different views of the phenomenon being studied and cannot easily be neatly represented by nested statistical models. For instance, in the crime example of the preceding section, one hypothesis might be that criminal behavior is deviant and explainable by the criminal's own characteristics, while a competing hypothesis would be that it is a rational choice. Adjudicating between such hypotheses often involves comparing non-nested models, and so the standard theory of Section 2.1 breaks down.

One way around this has been proposed by Cox (1961, 1962); it has been applied to sociological problems by Weakliem (1992) and Halaby and Weakliem (1993). Cox's approach, which has spawned a large literature, tends to be cumbersome to implement and requires the often arbitrary designation of one of the two non-nested models as the null hypothesis. One way around this arbitrariness is to carry out two tests rather than one, with each model in turn as the null hypothesis. However, there is no guarantee of getting the standard kind of result of a test, namely rejection of

one model and non-rejection of the other. Both models may fail to be rejected, in which case it is not clear how to make inferences about quantities of interest, especially if the two models lead to different conclusions. Both models may be rejected (as often happens with large samples), in which case the tests do not provide a comparison between the two models.

Another difficulty is that standard significance tests allow one either to reject the null hypothesis or to fail to reject it, but they do not provide any measure of evidence *for* the null hypothesis. Sometimes, however, sociological theories specify that something is the *same* across different groups, and thus the null hypothesis is the hypothesis of interest. One example is the Lipset-Zetterberg hypothesis referred to earlier in Section 2.2, that social mobility flows are the same in all industrialized countries. Another is the hypothesis that all sections of U.S. society now obey a two-child norm, according to which most couples have two children and there is very little variation between socio-economic groups in average completed family size (among those who have any children) (Lye and Greek, 1994).

A standard test allows us only to say that the data have failed to reject our null hypothesis of interest, but gives no indication of whether the data support it or not. A test can fail to reject a null hypothesis either because there is not enough data, or because the data do support it, but it does not allow us to distinguish between these two different situations.

Difficulties with P-values and the associated significance tests have been much discussed in the scientific literature. The reader edited by Morrison and Henkel (1970) compiled about 30 important pre-1970 articles, the majority of them by sociologists; they are still worth reading. They referred a great deal to the problems with large samples, but talked very little about the other difficulties discussed here; they did not suggest alternatives that would seem fully satisfactory nowadays. Leamer (1978) was the first to discuss in depth the difficulties with empirical model-building using significance tests. Recent social science references include Johnstone (1990a, b).

3 Bayesian Hypothesis Testing

In this section, I first briefly review Bayesian statistical parameter estimation, and then introduce Bayes factors, which form the basis for Bayesian hypothesis testing.

3.1 Bayesian Estimation

Bayesian estimation expresses all uncertainty, including uncertainty about the unknown parameters of a model, in terms of probability, and it views unknown parameters as random variables. Thus all results in Bayesian statistics follow directly from elementary probability theory, notably the definition of conditional probability, Bayes' theorem, and the law of total probability.

We start with a probability model for the data D, which is specified by a vector of d unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$. Before any data are observed, our beliefs and uncertainty about $\boldsymbol{\theta}$ are represented by a prior probability density $p(\boldsymbol{\theta})$. The probability model is specified by the likelihood

$p(D|\boldsymbol{\theta})$, which is the probability of observing the data D given that $\boldsymbol{\theta}$ is the true parameter.

Having observed the data D, we update our beliefs about $\boldsymbol{\theta}$ using Bayes' theorem to obtain the posterior distribution of $\boldsymbol{\theta}$ given the data D, namely

$$p(\boldsymbol{\theta}|D) = p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})/p(D), \tag{2}$$

where $p(D) = \int p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$, by the law of total probability. For estimation purposes we need to know $p(\boldsymbol{\theta}|D)$ only up to a constant of proportionality, and since $p(D)$ does not involve $\boldsymbol{\theta}$ it can be omitted from equation (2), which is then written

$$p(\boldsymbol{\theta}|D) \propto p(D|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{3}$$

Thus the posterior distribution is proportional to the likelihood times the prior.

The posterior distribution, $p(\boldsymbol{\theta}|D)$, contains all the information needed to make inference about $\boldsymbol{\theta}$. The only question is how best to summarize and communicate that information. Often interest focuses on the individual parameters (i.e. the components of $\boldsymbol{\theta}$). The posterior distribution of a component of $\boldsymbol{\theta}$, say θ_1, follows from the law of total probability by *integrating* out the other components, so that

$$p(\theta_1|D) = \int p(\boldsymbol{\theta}|D)d\theta_2 d\theta_3 \ldots d\theta_d. \tag{4}$$

The univariate distribution (4) contains all the information needed to make inference about θ_1. It can be summarized in various ways. In my experience, the most useful summaries are the posterior mode (i.e. the value of θ_1 that maximizes $p(\theta_1|D)$ and so is the most likely value given the data), and the .025 and .975 quantiles, which define a 95% Bayesian confidence interval. The posterior standard deviation is also useful, as a Bayesian analogue of the standard error. The posterior mean is also often used, and is usually close to the posterior mode.

Bayesian inference has been controversial because it uses the prior distribution, $p(\boldsymbol{\theta})$, which is subjectively determined by the user. However, in large samples this has very little influence: its contribution to the posterior mean and variance is on the order of $(1/n)$-th of the total, where n is the sample size.

In large samples, the posterior mode is very close to the maximum likelihood estimator (MLE), and Bayesian confidence intervals are very similar to standard non-Bayesian confidence intervals. Asymptotically, in regular models[5], the posterior distribution is multivariate normal with mean at the MLE and variance matrix equal to the inverse (observed or, less accurately, expected) Fisher information matrix. Thus, for *estimation* in *regular models* with *large samples*, Bayesian and maximum likelihood methods give answers that are essentially the same. The answers can be

---

[5]A regular statistical model is one in which the MLE is asymptotically normal with mean at the true value and variance matrix equal to the inverse expected Fisher information matrix. A simple example of a non-regular model is that in which the data are independent and uniformly distributed between 0 and θ, and θ is unknown. Then the MLE of θ is equal to the largest observation and does not have the usual asymptotic distribution (Kotz and Johnson, 1985, p. 346).

different, however, for testing and model selection, for estimation in non-regular models, and with very small samples.

Edwards, Lindman and Savage (1963) gave what remains an excellent and delightfully written introduction to Bayesian statistics for a social science audience, while Press (1989) and Lee (1989) are accessible accounts in book form. For a more advanced and theoretical treatment, but one that is still practically motivated, see Bernardo and Smith (1994).

3.2   Bayes Factors

Suppose now that we want to use the data D to compare two competing hypotheses, which are represented by the statistical models M_1 and M_2, with parameter vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. They may be nested, but need not be. Then, by Bayes' theorem, the posterior probability that M_1 is the correct model (given that either M_1 or M_2 is) is

$$p(M_1|D) = \frac{p(D|M_1)p(M_1)}{p(D|M_1)p(M_1) + p(D|M_2)p(M_2)}, \tag{5}$$

where $p(D|M_k)$ is the (marginal) probability of the data given M_k (see below), and $p(M_k)$ is the prior probability of model M_k (k=1,2). A similar expression holds for $p(M_2|D)$ and, by construction, $p(M_1|D) + p(M_2|D) = 1$.

In equation (5), $p(D|M_1)$ is obtained by *integrating* (not maximizing) over $\boldsymbol{\theta}_1$, i.e.

$$\begin{aligned} p(D|M_1) &= \int p(D|\boldsymbol{\theta}_1, M_1)p(\boldsymbol{\theta}_1|M_1)d\boldsymbol{\theta}_1 \\ &= \int (\text{ likelihood } \times \text{ prior })d\boldsymbol{\theta}_1, \end{aligned} \tag{6}$$

where $p(D|\boldsymbol{\theta}_1, M_1)$ is the likelihood of $\boldsymbol{\theta}_1$ under model M_1. I will call this quantity, $p(D|M_1)$, the *integrated likelihood* for model M_1; it has also been called the marginal likelihood, the marginal probability of the data, and the predictive probability of the data.

The extent to which the data support M_2 over M_1 is measured by the *posterior odds* for M_2 against M_1, i.e. the ratio of their posterior probabilities. By equation (5), this is

$$\frac{p(M_2|D)}{p(M_1|D)} = \left[\frac{p(D|M_2)}{p(D|M_1)}\right]\left[\frac{p(M_2)}{p(M_1)}\right]. \tag{7}$$

The first factor on the right-hand side of equation (7) is the ratio of the integrated likelihoods of the two models and is called the *Bayes factor* for M_2 against M_1, denoted by B_{21}. The second factor on the right-hand side of (7) is the prior odds, and this will often be equal to 1, representing the absence of a prior preference for either model, i.e. $p(M_1) = p(M_2) = \frac{1}{2}$. Thus equation (7) can be written

$$\text{posterior odds } = \text{Bayes factor } \times \text{ prior odds.} \tag{8}$$

It follows that the Bayes factor is equal to the posterior odds when the prior odds are equal to 1.

When $B_{21} > 1$, the data favor M_2 over M_1, and when $B_{21} < 1$ the data favor M_1. The use of Bayes factors to compare scientific theories was first proposed by Jeffreys (1935), and in Jeffreys (1961, Appendix B) he proposed the following rules of thumb for interpreting B_{21}: when $1 \leq B_{21} \leq 3$, there is evidence for M_2, but it is "not worth more than a bare mention", when $3 \leq B_{21} \leq 10$ the evidence is positive, when $10 \leq B_{21} \leq 100$ it is strong, and when $B_{21} > 100$ it is decisive. Probability itself is a meaningful scale and so these categories are not a calibration of the Bayes factor but rather a rough descriptive statement about standards of evidence in scientific investigation. I will return to the issue of interpretation in Section 4.3 and suggest a slightly different scale for use in social research.

Evaluating the Bayes factor involves calculating the integrated likelihood (6), which can be a high-dimensional and intractable integral. Various analytic and numerical approximations have been proposed, and in Section 4 I will discuss the BIC approximation, which is both simple and accurate. The Bayes factor depends on the prior and, in principle, this should be carefully specified and sensitivity to it should be carefully assessed. However, as we will see in Section 4.1, the BIC approximation corresponds rather closely to a particular choice of prior that seems reasonable for many practical purposes.

These and other aspects of Bayes factors are reviewed in detail by Kass and Raftery (1995), who give many references. One point they make is that the logarithm of the integrated likelihood may also be viewed as a predictive score for the model (Kass and Raftery, 1995, Section 3.2). This is of interest because it leads to an interpretation of the Bayes factor that does not depend on viewing one of the models as "true". In this view, the Bayes factor is designed to choose the model that will, on average, give better out-of-sample predictions.

4 The BIC Approximation

In this section, I will introduce the BIC (*Bayesian Information Criterion*) approximation to the Bayes factor by deriving it heuristically, giving explicit expressions for it in various model classes, and finally discussing its interpretation and its relation to P-values.

4.1 Derivation

The key quantity underlying the Bayes factor is the integrated likelihood for a model, given by equation (6). I will first derive a simple approximation to this quantity, and then show how it leads to approximate Bayes factors and to the BIC criterion for assessing models. This subsection is fairly technical. The key result is equation (20) and, if you are not interested in the derivation of BIC, you can now skip to that point and still be able to follow the rest of the chapter.

For the moment I will concentrate on approximating the the integrated likelihood for a single model, and for simplicity I will simplify notation by not mentioning the model, so that equation

(6) will be rewritten

$$p(D) = \int p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{9}$$

For ease of exposition, I will consider the case where the data D consist of n independent and identically distributed observations, $y_1, \ldots, y_n$, each of which may be a vector. The results apply much more widely than this, however, and in essence are valid for any regular statistical model. This includes many time series models for data that are not independent, and also models for data that are not identically distributed. For example, it includes most common models for event history data.

The derivation proceeds by considering a Taylor series expansion of $g(\boldsymbol{\theta}) = \log\{p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})\}$ about $\tilde{\boldsymbol{\theta}}$, the value of $\boldsymbol{\theta}$ that maximizes $g(\boldsymbol{\theta})$, i.e. the posterior mode. The expansion is

$$g(\boldsymbol{\theta}) = g(\tilde{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^T g'(\tilde{\boldsymbol{\theta}}) + \tfrac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^T g''(\tilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) + o(\| \boldsymbol{\theta} - \tilde{\boldsymbol{\theta}} \|^2), \tag{10}$$

where the superscript T denotes matrix transpose, $g'(\boldsymbol{\theta}) = \left(\frac{\partial g(\boldsymbol{\theta})}{\partial \theta_1}, \ldots, \frac{\partial g(\boldsymbol{\theta})}{\partial \theta_d} \right)^T$ is the vector of first partial derivatives of $g(\boldsymbol{\theta})$, and $g''(\boldsymbol{\theta})$ is the Hessian matrix of second partial derivatives of $g(\boldsymbol{\theta})$ whose (i,j) element is $\frac{\partial^2 g(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}$. Now $g'(\tilde{\boldsymbol{\theta}}) = 0$ because $g(\boldsymbol{\theta})$ reaches a maximum at $\tilde{\boldsymbol{\theta}}$ and so its first derivative is equal to zero at that point. Thus

$$g(\boldsymbol{\theta}) \approx g(\tilde{\boldsymbol{\theta}}) + \tfrac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^T g''(\tilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}). \tag{11}$$

The approximation in equation (11) is not sure to be good unless $\boldsymbol{\theta}$ is close to $\tilde{\boldsymbol{\theta}}$. However, when n is large the likelihood $p(D|\boldsymbol{\theta})$ is concentrated about its maximum and declines fast as one moves away from $\tilde{\boldsymbol{\theta}}$, so that only values of $\boldsymbol{\theta}$ close to $\tilde{\boldsymbol{\theta}}$ will contribute much to the integral (9) defining $p(D)$. For a formalization of this argument see Tierney and Kadane (1986).

It follows that

$$\begin{aligned} p(D) &= \int \exp[g(\boldsymbol{\theta})]d\boldsymbol{\theta} \\ &\approx \exp[g(\tilde{\boldsymbol{\theta}})] \int \exp\left[\tfrac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^T g''(\tilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) \right] d\boldsymbol{\theta}, \end{aligned} \tag{12}$$

by equation (11). Recognizing the integrand in equation (12) as proportional to a multivariate normal density gives

$$p(D) \approx \exp[g(\tilde{\boldsymbol{\theta}})](2\pi)^{d/2}|A|^{-\frac{1}{2}}, \tag{13}$$

where d is the number of parameters in the model and $A = -g''(\tilde{\boldsymbol{\theta}})$. The use of equation (13) is called the *Laplace method for integrals*. The error in equation (13) is $O(n^{-1})$ (Tierney and Kadane, 1986), and so

$$\log p(D) = \log p(D|\tilde{\boldsymbol{\theta}}) + \log p(\tilde{\boldsymbol{\theta}}) + (d/2)\log(2\pi) - \tfrac{1}{2}\log|A| + O(n^{-1}), \tag{14}$$

where $O(n^{-1})$ represents any quantity such that $nO(n^{-1}) \to$ a constant as $n \to \infty$.

Now in large samples, $\tilde{\boldsymbol{\theta}} \approx \hat{\boldsymbol{\theta}}$ where $\hat{\boldsymbol{\theta}}$ is the MLE, and $A \approx n\mathbf{i}$, where $\mathbf{i}$ is the expected Fisher information matrix for one observation. This is a $(d \times d)$ matrix whose (i, j) element is $-E\left[\frac{\partial^2 \log p(y_1|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\mid_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right]$, the expectation being taken over values of y_1, with $\boldsymbol{\theta}$ held fixed. Thus $|A| \approx n^d|\mathbf{i}|$. These two approximations introduce an $O(n^{-\frac{1}{2}})$ error into equation (14), which becomes

$$\log p(D) = \log p(D|\hat{\boldsymbol{\theta}}) + \log p(\hat{\boldsymbol{\theta}}) + (d/2)\log(2\pi) - (d/2)\log n - \tfrac{1}{2}\log|\mathbf{i}| + O(n^{-\frac{1}{2}}). \quad (15)$$

Now the first term on the right-hand side of equation (15) is of order $O(n)$, the fourth term is of order $O(\log n)$, while the other four terms are of order $O(1)$ or less. Removing the terms of order $O(1)$ or less thus gives

$$\log p(D) = \log p(D|\hat{\boldsymbol{\theta}}) - (d/2)\log n + O(1). \quad (16)$$

Equation (16) says that the log-integrated likelihood, $\log p(D)$, is equal to the maximized log-likelihood, $\log p(D|\hat{\boldsymbol{\theta}})$, minus a correction term.

Equation (16) is the approximation on which BIC is based, and its $O(1)$ error means that, in general, the error in it does not vanish even with an infinite amount of data. This is not as bad as it sounds, however, because the other terms on the right-hand side of (16) tend to infinity as n does, and so will eventually dominate. Thus the error in (16) will tend towards zero as a *proportion* of $\log p(D)$, ensuring that the error will not affect the conclusion reached, given enough data. Nevertheless, the $O(1)$ error does suggest the approximation to be somewhat crude.

Empirical experience has found (16) to be more accurate in practice than the $O(1)$ error term would suggest (e.g. Raftery, 1993b). In fact, the error is of a much smaller order of magnitude for a particular, reasonable, choice of prior distribution. Suppose that the prior $p(\boldsymbol{\theta})$ is multivariate normal with mean $\hat{\boldsymbol{\theta}}$ and variance matrix $\mathbf{i}^{-1}$. Thus, roughly speaking, the prior distribution contains the same amount of information as would, on average, a single observation. This seems to be a reasonable representation of the common situation where there is a little, but not much, prior information. Then

$$\log p(\hat{\boldsymbol{\theta}}) = -(d/2)\log(2\pi) + \tfrac{1}{2}\log|\mathbf{i}|, \quad (17)$$

and substituting (17) into (15) gives

$$\log p(D) = \log p(D|\hat{\boldsymbol{\theta}}) - (d/2)\log n + O(n^{-\frac{1}{2}}). \quad (18)$$

Thus for the particular prior mentioned, the error in the approximation (16) is $O(n^{-\frac{1}{2}})$ rather than $O(1)$, which is much smaller for moderate to large sample sizes, and which does tend to zero as n tends to infinity.

The approximation (18) can be used to approximate the Bayes factor $B_{21} = p(D|M_2)/p(D|M_1)$. This is most conveniently written on the scale of twice the logarithm, as follows:

$$2\log B_{21} = 2\left(\log p(D|\hat{\boldsymbol{\theta}}_2, M_2) - \log p(D|\hat{\boldsymbol{\theta}}_1, M_1)\right) - (d_2 - d_1)\log n + O(n^{-\frac{1}{2}}). \quad (19)$$

If M_1 is nested within M_2, equation (19) can be rewritten

$$2 \log B_{21} \approx \chi_{21}^2 - \mathrm{df}_{21} \log n, \tag{20}$$

where χ_{21}^2 is the standard likelihood ratio test (LRT) statistic for testing M_1 against M_2, and $\mathrm{df}_{21} = d_2 - d_1$ is the number of degrees of freedom associated with the test.

The Laplace method for integrals was introduced into statistics by Tierney and Kadane (1986) and seems first to have been used for Bayes factors by Raftery (1988). Equation (15) goes back to Jeffreys (1961), while equation (16) is due to Schwarz (1978) and equation (18) was pointed out by Kass and Wasserman (1992). For other references, see Kass and Raftery (1995).

4.2 BIC for Specific Models

4.2.1 General form

When several models are being considered, it is useful to compare each of them in turn with a baseline model, usually either a null model (M_0) with no independent variables, or a saturated model (M_S) in which each data point is fit exactly.

When the baseline model is a saturated model, M_S, the LRT statistic in equation (20) is often called the *deviance*. The value of BIC for model M_k, denoted by BIC_k, is the approximation to $2 \log B_{Sk}$ given by (20), where B_{Sk} is the Bayes factor for model M_S against model M_k. This is

$$\mathrm{BIC}_k = L_k^2 - \mathrm{df}_k \log n, \tag{21}$$

where $L_k^2 = \chi_{Sk}^2$ is the deviance for model M_k and df_k is the corresponding number of degrees of freedom. Then BIC_S, the BIC value for the saturated model, is zero, and the saturated model is preferred to M_k if $\mathrm{BIC}_k > 0$, in which case M_k can be considered not to fit the data well. When $\mathrm{BIC}_k < 0$, M_k is preferred to the saturated model, and the smaller BIC_k is (i.e. the more negative), the better the fit of M_k.

When comparing two models, M_j and M_k, we note that

$$
\begin{aligned}
B_{jk} &= p(D|M_j)/p(D|M_k) \\
&= \left[\frac{p(D|M_S)}{p(D|M_k)} \right] \Big/ \left[\frac{p(D|M_S)}{p(D|M_j)} \right] \\
&= B_{Sk}/B_{Sj}, \quad \text{and so}
\end{aligned}
$$

$$
\begin{aligned}
2 \log B_{jk} &= 2 \log B_{Sk} - 2 \log B_{Sj} \\
&\approx \mathrm{BIC}_k - \mathrm{BIC}_j. \tag{22}
\end{aligned}
$$

Thus two models can be compared by taking the difference of their BIC values, with the model having the smaller (i.e. more negative) BIC value being preferred. I will discuss the interpretation

17

of the size of the difference in Section 4.3. Note that M_j and M_k do not have to be nested for equation (22) to be applicable.

When the baseline model is the null model, M_0, with no independent variables, then BIC_k is replaced by BIC'_k, the approximation (20) to $2 \log B_{0k}$, where B_{0k} is the Bayes factor for the null model M_0 against the model of interest M_k. This is

$$\text{BIC}'_k = -\chi^2_{k0} + p_k \log n, \tag{23}$$

where χ^2_{k0} is the LRT statistic for testing M_0 against M_k, and p_k is the number of degrees of freedom associated with that test. In regression-type models, p_k will usually be the number of independent variables in M_k.

BIC'_0, the BIC' value for the null model, is zero. Thus if BIC'_k is positive, the null model M_0 is preferred to M_k, indicating that M_k is overparameterized, containing parameters (and hence probably variables) for which the data provide little support. In that case, a submodel of M_k (containing some but not all of the variables in M_k) may well fit better than either M_0 or M_k. For examples of this, see Section 7. If BIC'_k is negative, then M_k is preferred to M_0 and the smaller (i.e. the more negative) BIC'_k is, the more M_k is supported by the data. For comparing two models, BIC' differences can be used in the same way as BIC differences, and equation (22) is still valid if BIC is replaced by BIC'.

Which of BIC_k or BIC'_k should be used? For any one model they will be numerically different, but for *comparing* any two given models, M_j and M_k, they are equivalent, in the sense that the BIC difference is the same as the BIC' difference, i.e.

$$\text{BIC}_k - \text{BIC}_j = \text{BIC}'_k - \text{BIC}'_j. \tag{24}$$

The two measures, BIC and BIC', differ only by a constant which is the same for all models; this constant is equal to both BIC_0 and $-\text{BIC}'_S$, which are in turn equal to one another. Thus

$$\text{BIC}_k - \text{BIC}'_k = c \tag{25}$$

for all models M_k, where $c = \text{BIC}_0 = -\text{BIC}'_S$.

In practice, which of BIC or BIC' is used will depend on whether the software that estimates the models provides the deviances or the LRT statistic against the null model. If the software yields the deviance, then BIC will be used, and if instead it reports the LRT statistic, then BIC' will be used. If both the deviance and the LRT statistic are available, either BIC or BIC' can be used, or both. Although equivalent for testing and model selection purposes, they do each provide some different information. BIC_k can be viewed as a measure of overall model fit[6], while BIC'_k provides an assessment of whether M_k is explaining enough of the variation in the data to justify the number of parameters it uses.

---

[6]This is true only in models for which goodness-of-fit statistics can be used for this purpose, such as models for categorical data.

There is one important ambiguity in equations (21) and (23), namely the definition of n, the "sample size". What this should be is clear in some situations but not in others. As a general rule, the definition of n should be the one that makes the approximation $|A| \approx n^d |\mathbf{i}|$ used in the derivation of (15) most accurate. More precise suggestions for specific model classes will be given in the following subsections.

4.2.2 Linear regression and analysis of variance

For linear regression with normal errors, the most convenient form is BIC$'$, and it can be shown that this has the simple form

$$\text{BIC}'_k = n \log(1 - R_k^2) + p_k \log n, \tag{26}$$

where R_k^2 is the value of R^2 for model M_k and p_k is the number of independent variables (not including the intercept).

Note that standard analysis of variance for designed experiments can be recast in terms of linear regression by using sets of dummy variables to represent the different factors and interactions, and then equation (26) can be used in that context also. In particular, simple problems like testing for a difference between two means can be solved using (26) in this way.

The sample size n will usually be just the number of cases. This will not be true, however, if responses with the same values of the independent variables have been grouped into a single case with the average response as dependent variable, and weighted regression carried out, with weights proportional to the number of individuals in the group. This often happens in the analysis of designed experiments, when individuals are grouped into "cells". Then n should be the actual number of individuals rather than the number of cases or cells. When the data have been collected using a complex survey design with resulting weights, it is not yet clear what n should be, and this issue awaits further study. However, it seems reasonable that if the model is based on an assumption of simple random sampling but the sampling design is less efficient, then n should be reduced to reflect the efficiency of the sampling design relative to simple random sampling.

4.2.3 Logistic regression

Some logistic regression software produces the deviance, some the LRT statistic, and some both. Thus BIC and BIC$'$ may both be used, depending on the software, and equations (21) and (26) apply directly. The same is true for other binary response models, such as those with the probit or complementary log-log link.

What should n be? When each individual is a separate case, then n should be simply the sample size. In logistic regression, however, responses with the same values of the independent variables are often grouped together into a single case for which the dependent variable is the number of positive responses, which has a binomial distribution. In that situation, the number of cases is not

the same as the number of individuals. Then n should be the number of individuals, i.e. the sum of the binomial denominators, and not the number of cases in the regression.

4.2.4 Log-linear modeling

Software that estimates log-linear models for contingency tables usually gives the deviance rather than the LRT statistic against a null model. Thus it is most natural to use BIC rather than BIC'.

What should n be? Once again, it is best to use the actual number of individuals, i.e. the sum of the cell counts, and *not* the number of cells (Raftery, 1986a).

4.2.5 Event history analysis

Most event history analysis software reports the LRT statistic against the null model with no independent variables, and so BIC' is the more convenient measure to use. For fully parametric event history models, the theory of Section 4.1 provides a direct justification for the use of BIC'. However, event history analysis is often based on the Cox proportional hazards model, and there there is a complication: it is not fully parametric because the baseline hazard rate is unspecified. The regression part *is* parametric, however, and this is a case of a semi-parametric model. In spite of this, BIC' may still be validly used for the Cox model (Raftery, Madigan and Volinsky, 1995). The number of degrees of freedom, p_k, is then just the number of independent variables.

What should n be? Should it be the number of individuals, the number of events, or the number of spells (including censored spells)? It seems best to use the number of events rather than either of the other two possibilities (Raftery, Madigan and Volinsky, 1995).

For discrete time event history analysis, the same choice has been made (Xie, 1994), while the total number of exposure time units has also been used, for consistency with logistic regression (Raftery, Lewis, Aghajanian and Kahn, 1993; Raftery, Lewis and Aghajanian, 1994). The latter choice is more conservative and seems safer in the absence of a definitive result about which is more appropriate. More research is needed on this matter, and I conjecture that the less conservative choice of Xie (1994) will eventually be shown to be the more appropriate one.

4.2.6 Structural equation models

In this subsection I will use the notation of Bollen (1989, Table 2.2), so that N is the number of individuals, p is the number of indicators of the independent variables, q is the number of indicators of the dependent variables, and ν_k is the number of independent parameters fitted in model M_k.

Software for estimating structural equation models, such as LISREL or EQS, tends to give the deviance (i.e. the LRT statistic against the "saturated" model in which each *covariance* is fit exactly), rather than the LRT statistic against a null model. Thus BIC rather than BIC' is the more convenient measure and equation (21) is the one to use. There df_k is the number of covariances minus the number of independent parameters fitted, i.e. $\mathrm{df}_k = \frac{1}{2}(p+q)(p+q+1) - \nu_k$.

When one is comparing two models M_k and M_{k-1} where M_{k-1} is nested within M_k and M_k has one more parameter than M_{k-1}, then, approximately, $L_{k-1}^2 - L_k^2 = t^2$, where L_k^2 is the deviance for model M_k and t is the t test statistic for testing the additional parameter. Thus

$$\text{BIC}_{k-1} - BIC_k \approx t^2 - \log n. \tag{27}$$

If this is positive, the larger model M_k will be preferred.

When one is comparing M_k with a bigger model M_{k+1} within which it is nested and which has one *more* parameter than M_k, then, approximately, $L_k^2 - L_{k+1}^2 = W$, the Lagrange multiplier test statistic or modification index, and so

$$\text{BIC}_k - \text{BIC}_{k+1} \approx W - \log n. \tag{28}$$

Again, if this is positive, the larger model M_{k+1} will be preferred.

Equations (27) and (28) are useful for model-building with BIC in structural equation models, because most software for estimating these models returns both t statistics and modification indices. Thus by fitting a single model, one can compute approximate BIC values for it, all the models that have one parameter less than it, and all the models that have one parameter more than it. For an example of a model search that exploits this fact, see Raftery (1993a).

What should n be? I recommend using $n = N$, the number of individuals. Raftery (1993a) used $n = N(p+q)$, but the derivation of equation (19) (which was not known when Raftery (1993a) was written) suggests that $n = N$ would be more accurate. Note, however, that equation (16) is valid for both of these definitions of n.

4.3   Interpretation

In Section 3.2 I gave the rules of thumb of Jeffreys (1961) for interpreting Bayes factors and, hence, between-model differences in BIC or BIC′. I find a slightly modified version more appropriate. I prefer to define "strong" evidence as corresponding to posterior odds of 20:1 rather than 10:1 (by analogy with the intention behind the standard .05 significance level), and to use the term "very strong" rather than "decisive" for the evidence implied by very high posterior odds. Jeffreys put the boundary for this at 100:1, corresponding to a BIC difference of $2 \log 100 = 9.2$, but I prefer to round this up to the slightly more conservative value of 10, corresponding to posterior odds of about 150:1. This yields the scheme shown in Table 6.

t statistics and their associated P-values can be converted to approximate BIC differences by noting that when $\text{df}_{21} = 1$ in equation (20), then, approximately in regular models, $\chi_{21}^2 \approx t^2$, where t is the usual t statistic for testing the significance of the parameter of M_2 that is set equal to zero in M_1. Then (20) becomes

$$2 \log B_{21} \approx t^2 - \log n \approx \text{BIC}_1 - \text{BIC}_2. \tag{29}$$

Table 6: Grades of evidence corresponding to values of the Bayes factor for M_2 against M_1, the BIC difference and the posterior probability of M_2.

BIC difference	Bayes factor	$p(M_2|D)(\%)$	Evidence
0–2	1–3	50–75	Weak
2-6	3–20	75–95	Positive
6–10	20–150	95–99	Strong
> 10	> 150	> 99	Very strong

Table 7: Approximate minimum t values corresponding to different grades of evidence.

Evidence	Minimum BIC difference	n					
		30	50	100	1,000	10,000	100,000
Weak	0	1.84	1.98	2.15	2.63	3.03	3.39
Positive	2	2.32	2.43	2.57	2.98	3.35	3.68
Strong	6	3.07	3.15	3.26	3.59	3.90	4.18
Very strong	10	3.66	3.73	3.82	4.11	4.38	4.64

(Note that the middle expression in equation (29) is only an approximation to the difference of BIC values, i.e. an approximation to an approximation.) It follows that one can translate t values roughly into BIC values and hence into grades of evidence such as those of Table 6. In particular, $|t| > \sqrt{\log n}$ is required for there to be even weak evidence for the additional parameter in M_2, while $|t| > \sqrt{\log n + 6}$ corresponds to strong evidence on this scale.

Table 7 shows the minimum t values required for various grades of evidence and sample sizes. The sample sizes are chosen to represent roughly the sample sizes that arise in various kinds of sociological study. The first three sample sizes are in the range of those that arise in aggregate studies and in quantitative macrosociology: very roughly, there are about 30 industrialized countries, 50 states of the U.S. and 100 U.S. SMSAs in a typical study. The last three sample sizes are more typical of individual-level survey and census data: there might be 1,000 cases in a small survey, 10,000 in a big one, and 100,000 in a census subsample, a large event history database, or a cross-national collection of surveys. The minimum t values in Table 7 are for the most part larger than 2, suggesting that the common rule of viewing t values greater than 2 as "significant" overstates the evidence that they imply.

In the context of linear regression, equation (26) indicates that the evidence for an additional independent variable can be measured by

$$\text{BIC}'_{k+1} - \text{BIC}'_k = n \log\{(1 - R^2_{k+1})/(1 - R^2_k)\} + \log n, \qquad (30)$$

where M_k is nested within M_{k+1}, which contains one additional variable. For there to be any evidence in favor of the new variable, the right-hand side of (30) should be negative. Thus for a

Table 8: Minimum percent reduction in the residual sum of squares required for different grades of evidence in favor of one additional variable in linear regression. When R^2 is small, the is roughly equal to the required increase in R^2.

Evidence	Minimum BIC difference	n					
		30	50	100	1,000	10,000	100,000
Weak	0	10.7	7.5	4.5	0.7	.09	.012
Positive	2	16.5	11.2	6.4	0.9	.11	.014
Strong	6	26.9	18.0	10.1	1.3	.15	.018
Very strong	10	36.0	24.3	13.6	1.7	.19	.022

BIC' change of more than $\nabla\text{BIC}'$, we would need to have

$$\text{RED}_{k,k+1} > 1 - \exp[-(\nabla\text{BIC}' + \log n)/n], \tag{31}$$

where $\text{RED}_{k,k+1} = 1 - (1 - R_{k+1}^2)/(1 - R_k^2)$ is the proportional reduction in residual sum of squares due to the additional variable. When R_k^2 is small, then $\text{RED}_{k,k+1} \approx R_{k+1}^2 - R_k^2$, which is the increase in R^2 due to the additional variable, and so equation (31) becomes

$$\text{Increase in } R^2 > 1 - \exp[-(\nabla\text{BIC}' + \log n)/n]. \tag{32}$$

Note that equation (32) is valid only when R_k^2 is small, and should be a reasonable approximation for, say, $R_k^2 < .30$. The values of (31) or (32) corresponding to various grades of evidence for different sample sizes are shown in Table 8.

4.4   BIC and P-values

The P-values corresponding to the t statistics in Table 7 are shown in Table 9. These are rather different from the commonly used .05 and .01 cutoffs, and in most cases are smaller. For sample sizes in the 30–50 range they are in rough agreement with conventional rules, but for larger sample sizes, much smaller P-values are required to imply that the data provide evidence for the effect of interest. Conventional advice has been that the significance level should decline as sample size increases, but how this should be done has not been spelled out. Table 9 provides precise guidelines for doing so, and reveals that, for large samples of the sizes that sociologists routinely work with, significance levels need to be lowered more drastically than one would perhaps have expected.

It is important to note that Table 9 is valid only for tests involving one additional parameter (i.e. one degree of freedom). Equivalent tables could be constructed for tests with more than one degree of freedom; typically the deviation from conventional values would be even greater than where there is one degree of freedom, especially for the larger sample sizes.

In fact, the use of Bayes factors can be viewed as a precise way of implementing the advice of Neyman and Pearson (1933) that power and significance be balanced when setting the significance

Table 9: Approximate two-sided P-values corresponding to different grades of evidence in favor of one additional parameter.

Evidence	Minimum BIC difference	n					
		30	50	100	1,000	10,000	100,000
Weak	0	.076	.053	.032	.009	.002	.0007
Positive	2	.028	.019	.010	.003	.0008	.0002
Strong	6	.005	.003	.001	.0003	.0001	.00003
Very strong	10	.001	.0005	.0001	.00004	.00001	.000004

level, in the following sense. Suppose that half the time the null hypothesis, M_1, is true and that half the time it is false, in which case the alternative hypothesis, M_2, is true. Then the overall error rate (total of Type I and Type II errors) is minimized when the testing rule is to reject the null hypothesis whenever the Bayes factor favors the alternative, i.e. whenever $B_{21} > 1$, or, approximately equivalently, when $\text{BIC}_2 < \text{BIC}_1$ or $\text{BIC}_2' < \text{BIC}_1'$. This was shown by Jeffreys (1961, pp. 396–397), as was pointed out by Kass (1991) using more modern terminology.

It is clear from Table 9 that naive interpretations of P-values such as "$P = .001$ means that the null hypothesis is false with probability .999" are wrong. To be fair, arguments for P-values do not claim that such an interpretation is valid, but it may be a surprise that with a large enough sample ($n = 100,000$) $P = .001$ actually corresponds to evidence *for* the null hypothesis.

There is no real conflict between Bayes factors and significance tests: Bayes factors can be viewed as a way of setting the significance level in the test. With large samples, the appropriate level can be well below conventional levels such as .05 or .01, as Table 9 shows. However, there is a conflict between Bayes factors and significance testing at pre-determined levels such as .05 or .01. There seem to be two reasons for this conflict. The first is the nature of the question posed by the P-value-based test:

What is wrong with the likelihood ratio test?

The aim of much social research is to describe the main features of selected aspects of social reality and is necessarily to some extent approximate. The LRT, in common with other significance tests, is designed to detect *any* discrepancies between model and reality. Such discrepancies do exist, by definition, although if the model is satisfactory, they should be small. With a large enough sample, the LRT will find them and reject even a good model.

In the contingency table case, the LRT tests a model M_0 say, against the saturated model M_1. Assume for the moment that no other models are being considered. Rejection of M_0 then implies acceptance of M_1, which says that each cell is a special case. This does constitute a statement about the underlying social reality and may, indeed, itself be a model of interest. Rejection of M_0 does not imply that M_1 provides a better description. The point is that we should be *comparing* the models, not just looking for possibly minor discrepancies between one of them and the data.

The question to which we really want an answer can perhaps often best be expressed as follows: which model better describes the main features of social reality as reflected in the data? A closely related and more precise question is: given the data, which of M_0 and M_1 is more

likely to be the true model.

The latter question can be answered by calculating the posterior odds for M_0 against M_1. (Raftery, 1986b).

The second reason relates to the nature of the conditioning in the two procedures. A standard test rejects H_0 if equation (1) holds, i.e. if the probability under H_0 of observing a value of the test statistic as extreme *or more so* is small. Thus the standard test conditions on the event $\{T \geq t(D)\}$, i.e. the event that the test statistic was as extreme as the value observed, or more so. However, what *actually* happened was the event $\{T = t(D)\}$, which is less surprising under H_0 (because less extreme), and hence casts less doubt on H_0. Bayesian model selection conditions on what actually happened, namely $\{T = t(D)\}$, suggesting the data to be less surprising under H_0 than does the standard test. Thus the Bayesian method tends to be less likely to reject a null hypothesis. Jeffreys (1980) wrote:

> I have always considered the arguments for the use of P absurd. They amount to saying that a hypothesis that may or may not be true is rejected because a greater departure from the trial value was improbable; that is, that it has not predicted something that has not happened.

Berger and Sellke (1987) gave the following simple illustration of the distinction[7]:

> Suppose that X is measured by a weighing scale that occasionally "sticks" (to the accompaniment of a flashing light). When the scale sticks at 100 (recognizable from the flashing light) one knows only that the true value x was, say, greater than 100. If large X casts doubt on H_0, occurrence of a "stick" at 100 should certainly be greater evidence that H_0 is false than should a true reading of $x = 100$. Thus there should be no surprise that the P-value might cause a substantial overevaluation of the evidence against H_0.

In this situation, the P-value will be the same whether or not the light is flashing, which seems counter-intuitive: it is clear that there is more evidence against H_0 when the light is flashing than when it is not. In a sense, P-value-based tests *always* proceed as if the light were flashing, and that is one reason why they overestimate the evidence against H_0 in the more usual situation where the data are fully observed (or, equivalently, where the light is not flashing). By contrast, the Bayes factor for H_1 against H_0 will be greater when the light is flashing than when it is not, in agreement with intuition.

The arguments are well summarized by Berger and Sellke (1987) and Berger and Delampady (1987) and the discussants of these papers, which I recommend to the reader.

5 Model Uncertainty and Occam's Window

I now turn to the situation where there are many models, $\{M_1, \ldots, M_K\}$, and no longer just two. Suppose that Δ is a quantity of interest such as a parameter of main interest or a future observation to be predicted. Then Bayesian inference about Δ is based on its posterior distribution, which is

$$p(\Delta|D) = \sum_{k=1}^{K} p(\Delta|D, M_k)p(M_k|D), \tag{33}$$

---

[7]The quotation has been slightly paraphrased.

by the law of total probability (Leamer, 1978, p. 117). Thus the full posterior distribution of Δ is a weighted average of its posterior distributions under each of the models, where the weights are the posterior model probabilities, $p(M_k|D)$. Equation (33) provides inference about Δ that takes full account of model uncertainty.

In equation (33) the posterior model probabilities $p(M_k|D)$ are obtained by Bayes' theorem, as follows:

$$p(M_k|D) = \frac{p(D|M_k)p(M_k)}{\sum_{\ell=1}^{K} p(D|M_\ell)p(M_\ell)}, \tag{34}$$

which is a direct generalization of equation (5) from two models to K of them. Often all the models will be on an equal footing *a priori*, so that $p(M_1) = \ldots = p(M_K) = 1/K$. By the results in Section 4.1, approximately, $p(D|M_k) \propto \exp(-\frac{1}{2}\text{BIC}_k)$ or $\exp(-\frac{1}{2}\text{BIC}'_k)$. Thus

$$p(M_k|D) \approx \exp(-\tfrac{1}{2}\text{BIC}_k)/\sum_{\ell=1}^{K} \exp(-\tfrac{1}{2}\text{BIC}_\ell). \tag{35}$$

Equation (35) still holds if BIC is replaced by BIC$'$.

I will now consider in more detail the situation where the quantity of interest is one of the regression parameters, β_1, say. Typically some of the models specify $\beta_1 = 0$, and so the posterior probability that $\beta_1 = 0$, $\Pr[\beta_1 = 0|D]$, will be non-zero. Of particular interest is $\Pr[\beta_1 \neq 0|D]$, the posterior probability that β_1 is in the model, which is just

$$\Pr[\beta_1 \neq 0|D] = \sum_{A_1} p(M_k|D), \tag{36}$$

where $A_1 = \{M_k : k = 1, \ldots, K; \beta_1 \neq 0\}$, i.e. the set of models that include β_1.

The probability that β_1 is in the model, $\Pr[\beta_1 \neq 0|D]$, can be converted to the odds scale using the relationship

$$\text{odds} = \text{probability} / (1 - \text{probability}),$$

and interpreted using rules of thumb such as those in Table 6. The breakpoints for weak, positive, strong and very strong evidence are then about .50, .75, .95 and .99 on the probability scale.

Of interest also is the size of the effect, given that it is non-zero. The posterior distribution of this is

$$p(\beta_1|D, \beta_1 \neq 0) = \sum_{A_1} p(\beta_1|D, M_k)p(M_k|D). \tag{37}$$

This can be summarized by its posterior mean and standard deviation, which may be viewed as, respectively, a Bayesian point estimator and a Bayesian analogue of the standard error. Convenient approximations to these are:

$$E[\beta_1|D, \beta_1 \neq 0] \approx \sum_{A_1} \hat{\beta}_1(k)p(M_k|D), \tag{38}$$

$$\text{SD}^2[\beta_1|D, \beta_1 \neq 0] \approx \sum_{A_1} \left[\text{se}_1^2(k) + \hat{\beta}_1(k)^2\right] p(M_k|D) - E[\beta_1|D, \beta_1 \neq 0]^2, \tag{39}$$

26

where $\hat{\beta}_1(k)$ and $\mathrm{se}_1(k)$ are respectively the MLE and standard error of β_1 under model M_k (Leamer, 1978, p. 118; Raftery, 1993a).

The main practical problem with putting this scheme into practice is that the number of models, K, may be so large that direct evaluation of the sums over all models is not feasible. For instance, in the crime example of Section 2.4, $K = 2^{15} = 32,768$, and so a literal implementation of the scheme would involve fitting all 32,768 regression models.

To get around this, Madigan and Raftery (1994) argued that one should exclude from the sum in (33) (a) models that are much less likely than the most likely model — say 20 times less likely, corresponding to a BIC (or BIC$'$) difference of 6; and (optionally) (b) models containing effects for which there is no evidence, that is models that have more likely submodels nested within them. The models that are left are said to belong to *Occam's window*, a generalization of the famous Occam's razor, or principle of parsimony in scientific explanation. When both (a) and (b) are used, Occam's window is said to be *strict*, and when only (a) is used it is said to be *symmetric*.

Both variants of Occam's window reduce the number of models enormously, while encompassing the essential model uncertainty present. In the crime example, there are $K = 32,768$ models to start with, while the symmetric Occam's window has 51, and the strict Occam's window has only 14. This is quite typical of experience to date.

A series of studies, summarized by Raftery, Madigan and Volinsky (1995), has shown that in a range of model classes and with a variety of data sets, *taking account of model uncertainty yields better out-of-sample predictive performance than any one model that might reasonably have been selected*. This is true whether one averages across all models, or uses Occam's window in either its strict or symmetric forms. But which of these three model averaging methods is the best? The studies to date suggest that the symmetric Occam's window has predictive performance as good as that of averaging over all models, while the strict Occam's window does slightly less well predictively, but is more useful for *reporting* model uncertainty, because it involves far fewer models, and these are the most important ones. In Section 6 we report only results from the strict Occam's window.

How can we find the models in Occam's window when the initial set of models is huge? It is not feasible to proceed directly by checking each model to see whether or not it is excluded, because the number of models is too large. For the special case of linear regression one can use the leaps and bounds algorithm of Furnival and Wilson (1974) to select a reduced set of good models, and then apply rules (a) and (b) directly to this reduced set. This is the basis for the BICREG software described in the Appendix. This has been adapted for logistic regression in the BIC.LOGIT software, which is also described in the Appendix. A more general tree-based algorithm is described by Madigan and Raftery (1994); this is applicable to a wide range of model classes.

The Bayesian approach to model uncertainty was introduced by Leamer (1978). For reviews of the work since then, see Draper (1995) and Kass and Raftery (1995).

6   Difficulties Resolved

I now return to the practical difficulties with P-value-based tests discussed in Section 2 and describe how they are dealt with by Bayes factors, BIC, and the Bayesian approach to model uncertainty.

6.1   Large Samples

The BIC values for the models proposed for the large cross-national social mobility data set of Section 2.2 are shown in Table 2. The Lipset-Zetterberg hypothesis (model 2) is indeed overwhelmingly rejected given its very large positive BIC value[8]. However, the quasi-symmetry model (model 3) is strongly preferred by BIC to the saturated model (model 4).

This agrees with the intuition of Grusky and Hauser (1984) and with the decision they made, and yet is in dramatic conflict with the result based on P-values. Thus in this case BIC gives a result that is in agreement with the scientific judgement of knowledgeable investigators, while P-values give a result that is directly opposed to it. It is interesting to note that when Grusky and Hauser decided to ignore the P-value, because they felt that it clearly did not make scientific sense, they did not know about BIC and so did not have any formal statistical justification for their decision. This was the original example of BIC for log-linear models (Raftery, 1986b). The fifth model in Table 2 is discussed below in Section 7.

6.2   Many Candidate Independent Variables

It was shown in Section 2.3 that when there are many candidate independent variables, statistical conclusions based on the selected model can be very misleading. They tend to identify seemingly strong relationships when, in fact, none exist. This was most strikingly illustrated by Freedman's (1983) simulation of 50 independent variables with 100 cases all consisting of pure noise unrelated to the dependent variable. In my replication of this, stepwise regression led to a highly significant and apparently satisfactory model with four independent variables (Table 3).

When the strict Occam's window was applied to these simulated data, it found five almost equiprobable models including the null model itself. When $\Pr[\beta_j \neq 0 | D]$ was calculated for each variable, it was found to be zero for 44 of the 50 variables, below $1/2$ for a further four, while for the remaining two it was 0.70 and 0.73. Even for these last two the evidence for an effect is weak on the scale of Table 6, with posterior odds of 2.3 and 2.7. Thus the conclusion from Occam's window would be that there is at most weak evidence for the inclusion of any variable, and that the null model itself is a plausible candidate. Unlike the conclusions that follow from screening methods and stepwise regression, this is not a misleading conclusion. Thus Occam's window seems to resolve the dilemma posed by Freedman's result.

It might be objected that Occam's window (and methods based on Bayes factors and BIC more generally) tends to favor parsimony to such an extent that it might find no signal even when there

---

[8]The same result holds when only the nine industrialized countries are included.

was one. To check whether this was so, I did two further small simulation experiments, using the same X matrix as that reported in Section 2.3. In both experiments, instead of Y being noise, Y was allowed to depend only on X_1: Y was simulated as $Y = \beta X_1 + \varepsilon$, where $\varepsilon \sim N(0, 1 - \beta^2)$, so that the "true" R^2 is β^2.

In the first experiment, $\beta = .45$ so that $R^2 = .20$. There Occam's window contained just one model: the correct one with X_1 only. Thus the correct conclusion was drawn by Occam's window in this case without any ambiguity or uncertainty. By contrast, the screening method described in Section 2.3 (screening out clearly nonsignificant variables from the full equation) yielded a model with 10 variables of which three were significant at the .05 level, and a P-value of 3×10^{-6}. Stepwise regression yielded a model with two variables (including X_1), both of them significant at the .05 level.

In the second experiment, $\beta = .32$, so that the true R^2 was only .10. Occam's window yielded two models with almost equal probabilities, one containing only X_1, and the other consisting of (X_1, X_{10}). Thus $\Pr[\beta_1 \neq 0 | D] = 1$ and $\Pr[\beta_{10} \neq 0 | D] = .52$, while $\Pr[\beta_j \neq 0 | D] = 0$ for all other 48 coefficients. Thus Occam's window would lead us to conclude that X_1 certainly has an effect, that there is some very weak evidence for X_{10} having an effect, while there is no evidence that any of the other 48 variables has an effect. This is strikingly faithful to the reality, especially given the low "true" R^2 (.10), the relatively small sample size (100), and the large number of irrelevant variables (49).

By contrast, the screening method gave a model with 11 variables of which four were significant at the .05 level, while stepwise regression gave a model with two variables (including X_1) both significant at the .05 level. Once again, standard variable selection strategies misleadingly detected evidence for effects of variables that were in fact not at all associated with the dependent variable.

6.3 Model Uncertainty

I now return to the crime example of Section 2.4, in which there was clear model uncertainty. Different variable selection methods gave quite different models. Also, in terms of the main questions of interest, different models selected gave very different estimates of β_{14}, the effect of probability of imprisonment, and also yielded different conclusions about whether X_{15}, the average time spent in state prisons, has an effect.

The Occam's window analysis of the crime data is shown in Table 10. There are 14 models, between them giving a picture of the model uncertainty in the data. Ehrlich's models do not fit well enough to be included in Occam's window, and they have BIC$'$ values that are far worse than the best model, by 25 and 30 points respectively. The theory on which Ehrlich's models are based would have to be very solid indeed to justify their being used as the basis for conclusions.

For X_{14}, the probability of imprisonment, the probability of an effect is high at 98% and the point estimate taking account of model uncertainty is -0.24. Interestingly, this is about half-way between the value from stepwise regression (-0.19) and those from the full model and the models

Table 10: Occam's window analysis of the crime data

#	Variable	Model 1	2	3	4	5	6	7	8	9	10	11	12	13	14	Prob (%)	Post. mean	Post. SD
1	% young males	•	•	•	•	•	•	•		•	•		•	•	•	94	1.40	0.50
2	South															0	—	—
3	Education	•	•	•	•	•	•	•	•	•	•	•	•	•	•	100	2.12	0.50
4	Police 1960	•	•		•	•		•	•	•			•		•	76	0.95	0.20
5	Police 1959			•			•				•	•		•		24	0.97	0.19
6	Labor part.															0	—	—
7	Sex ratio															0	—	—
8	Population						•		•			•				12	-0.08	0.04
9	Nonwhites	•	•	•	•		•			•	•	•	•			83	0.10	0.04
10	Unemp. 14-24															0	—	—
11	Unemp. 35-39	•	•	•		•		•						•		68	0.32	0.13
12	GDP															0	—	—
13	Inequality	•	•	•	•	•	•	•	•	•	•	•	•	•	•	100	1.33	0.32
14	Prob. prison	•	•	•	•	•	•	•	•	•	•	•	•		•	98	-0.24	0.10
15	Prison time	•			•						•					35	-0.30	0.15
	R^2 (%)	84	83	82	82	80	82	80	80	80	81	79	79	78	78			
	# vars.	8	7	7	7	6	7	6	6	6	7	6	6	5	5			
	BIC$'$ (+50)	-5.9	-5.4	-4.4	-3.8	-3.6	-3.1	-2.7	-2.4	-2.4	-1.5	-1.3	-1.2	-0.9	-0.9			
	PMP (%)	24	18	11	8	8	6	5	4	4	3	2	2	2	2			

NOTES: For fuller definitions of the independent variables, see Table 4.
"Prob" denotes $\Pr[\beta_j \neq 0|D]$ and is given by equation (36).
The posterior mean and SD are given by equations (38) and (39).
PMP denotes "posterior model probability" and is given by equation (35).

chosen by C_p and adjusted R^2 (-0.30) in Table 5. The posterior standard deviation of β_{14} is 0.10, while for the stepwise regression model the standard error was 0.07; the difference is due to model uncertainty. The one-model standard error underestimates uncertainty, because it ignores the component due to model uncertainty.

For X_{15}, the average time spent in state prisons, the overall posterior probability that it has an effect is 0.35. Thus the data provide no evidence for this variable to have an effect, but they do not exclude this possibility either.

As for the other variables, there is very strong evidence that education and income inequality are associated with higher crime rates (each with "crime elasticities" greater than 1), positive but not strong evidence for effects of the proportions of young males and of nonwhites, and weak evidence for an effect of unemployment among males aged 35–39.

The case of police expenditures is interesting. This has been measured in two successive years and the measures are very highly correlated ($r = .993$). The data show clearly that the 1960 crime rate is associated with police expenditures, and that only one of the two measures (X_4 and X_5) is needed, but they do not say for sure which measure should be used. Each model in Occam's window contains one measure or the other, but not both. And we have $\Pr[\beta_4 \neq 0|D] + \Pr[\beta_5 \neq 0|D] = 1$, so that the data provide very strong evidence for an association with police expenditures.

The coefficient for police expenditures is positive, which may be contrary to expectations. It does indicate that increased police expenditures are not associated with lower crime rates, and hence that police expenditure is not a confounding variable for inference about the effect of X_{14}

on Y, for example, at least not in the way one might expect.[9] A simple way of dealing with this is to exclude from Occam's window models with any coefficient in the wrong direction; here this would amount to excluding X_4 and X_5 and redoing the analysis.[10] Note that if the purpose of the modeling exercise is solely to *predict* crime rates (for example in the three states not included in the data), rather than to make inference about causal mechanisms, then models with X_4 and X_5 should be included, even if the coefficients have the "wrong" sign.

There is no evidence for an effect of any of the other variables, and in the case of five of them (those for which $\Pr[\beta_j \neq 0 | D] = 0$), there is evidence *against* an effect.

A more exact Bayesian analysis of these data that does not rely on the BIC$'$ approximation was done by Raftery, Madigan and Hoeting (1993).

7   Model Building Strategy

One apparent difficulty with the approach outlined here is that when a parsimonious but ill-fitting model M_1 is compared with a highly over-parameterized model M_2, BIC often prefers the more parsimonious model, even though it may be clearly sociologically unacceptable. When forced to choose between two unsatisfactory models, BIC tends to choose the one with less parameters. This has led some researchers to worry that BIC is biased in favor of parsimony over fit.

Formally speaking, this worry is unfounded, given that one ever considers only M_1 and M_2. Bayes factors are designed to choose the model that provides better out-of-sample predictions on average (Kass and Raftery, 1995, Section 3.2), and their use as a significance test minimizes the total error rate. In practice, however, when this occurs it can be an indication that neither M_1 nor M_2 is a very good model, in that M_1 may be missing an important aspect of the underlying phenomenon, while M_2 may be using too many parameters to represent it, for several of which there is no evidence.

A reasonable course of action when this happens is to search for a further model, M_3 say, which achieves most of the improvement in deviance or maximized likelihood in going from M_1 to M_2, but uses fewer parameters to do it. One way of doing this is to ask why M_2 should fit better than M_1, and then build a model that has one parameter (or so) for each reason or mechanism given. Another, complementary, approach is to inspect the residuals from M_1 to see if there is a pattern or if they can be predicted by other variables not in M_1. The resulting model, M_3, or some variant of it, may well have a better BIC value than either M_1 or M_2. Thus BIC can be used to guide an iterative model-building process.

This is well illustrated by the cross-national social mobility data set of Sections 2.2 and 6.1. Grusky and Hauser (1984) noted that the quasi-symmetry model was preferable to the saturated

---

[9] One possible explanation is that increases in the crime rate lead to increased police expenditure. Time series data would be needed to address the issue properly.

[10] This is roughly equivalent to the more sophisticated Bayesian approach of using a prior distribution for β_4 and β_5 that excludes positive values.

Table 11: Fit of models for the four-way table of US mobility 1972–1985, from Hout (1988). ($n = 9,227$).

#	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.

model which asserts that the mobility regime in each country is different. They nevertheless searched for systematic patterns in cross-national differences between mobility regimes, explained by characteristics of the countries studied that might be expected to affect social mobility.

This led to model 5 of Table 2 above, in which the country-specific mobility parameters are allowed to vary systematically as functions of industrialization, educational participation, social democracy and inequality, with a dummy variable for Hungary. By a conventional P-value-based significance test, this model would be strongly rejected in favor of the quasi-symmetry model (and the saturated model also), but Grusky and Hauser (1984) used it and claimed that its good fit provides evidence of systematic cross-national variation in mobility parameters. Once again, their intuitively based support of this model was (retrospectively) validated by BIC, which supports this model over the quasi-symmetry model; see Table 2.[11]

A second illustration, also from the area of social mobility, is provided by the model selection process in Hout (1988), part of which is shown in Table 11. Hout's article is about gender differences and changes over time in social mobility in the U.S. over the period 1972–1985. His starting point was the four-way $2 \times 3 \times 17 \times 17$ cross-classification of gender (S) $\times$ period (P) $\times$ father's occupation (O) $\times$ current occupation (D), and he used log-linear models.

Model 1 in Table 11 can be viewed as a kind of baseline model; it does not contain the $[OD]$ interaction and so would not be sociologically acceptable. Model 2 in Table 11 does include the $[OD]$ association, but uses no fewer than $16 \times 16 = 256$ parameters to represent it. The result is a decrease in deviance that is substantial, but not enough to justify the large number of parameters used to achieve it, according to BIC.

The surprising fact that BIC prefers model 1 to model 2 in Table 11 led Hout to ask how the $[OD]$ association in model 2 (which was responsible for most of the 1883 points decrease in

---

[11] I have not discussed the possible presence of overdispersion in these data. Given the sample design, it is hard to see what the source of substantial overdispersion would be. In any event, if overdispersion were explicitly taken into account using standard methods (McCullagh and Nelder, 1989), the deviances would be deflated and the evidence for the more parsimonious models would be stronger. Among the models considered here, the choices made would be unaffected.

deviance) could be more parsimoniously and interpretably represented. The answer was that the occupations of fathers and sons are associated because they have similar statuses, levels of on-the-job autonomy and job-specific training. Using these ideas, the $[OD]$ interaction can be represented using far fewer than 256 parameters, each of which has a direct interpretation. This is achieved using Hout's own (1984) status-autonomy-training (SAT) model. The result was model 3 in Table 11, which parsimoniously represents the full four-way $[SPOD]$ interaction and has a much better BIC value than either model 1 or model 2.

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:

- 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 *non-nested* 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, from 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 account of 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; see, e.g., Ragin, 1987. The use of BIC corresponds to a particular sample-size-dependent choice of significance level and, as Table 9 shows, for sample 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 (1993b) 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 (1993b) can be used with small samples; this is implemented in the GLIB software described in the Appendix.

In this chapter, I have focused on the choice of independent variables in regression and related models. 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

---

[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!).

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, what 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.

References

Becker, Gary S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy*, **76**, 526–536.

Berger, James O. and Mohan Delampady (1987). Testing precise hypotheses (with Discussion). *Statistical Science*, **3**, 317–352.

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**, 112–122.

Bernardo, José M. and Adrian F.M. Smith (1994). *Bayesian Theory.* New York: Wiley.

Bishop, Yvonne M.M., Steven E. Fienberg and Paul W. Holland (1975). *Discrete Multivariate Analysis.* Cambridge, Mass.: MIT Press.

Bollen, Kenneth A. (1989). *Structural Equations with Latent Variables.* New York: Wiley.

Brier, S.S. and Steven E. Fienberg (1980). Recent econometric modeling of crime and punishment: Support for the deterrence hypothesis? *Evaluation Review*, **4**, 147–191.

Cox, David R. (1961). Tests of separate families of hypotheses. *Proceedings of the 4th Berkeley Symposium*, **1**, 105–123.

Cox, David R. (1962). Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society, series B*, **24**, 406–424.

Draper, David (1995). Assessment and propagation of model uncertainty (with Discussion). *Journal of the Royal Statistical Society, series B*, to appear.

Edwards, Ward, Harold Lindman and Leonard J. Savage (1963). Bayesian statistical inference for psychological research. *Psychological Review*, **70**, 193–242.

Ehrlich, Isaac (1973). Participation in illegitimate activities: a theoretical and empirical investigation. *Journal of Political Economy*, **81**, 521–565.

Featherman, David L., Frank L. Jones and Robert M. Hauser (1975). Assumptions of mobility research in the United States: the case of occupational status. *Social Science Research*, **4**, 329–360.

Fenech, A. and Peter Westfall (1988). The power function of conditional log-linear model tests. *Journal of the American Statistical Association*, **83**, 198–203.

Fienberg, Steven E. and William M.M. Mason (1979). Identification and estimation of age-period-

cohort effects in the analysis of discrete archival data. *Sociological Methodology 1979*, 1–67.

Freedman, David A. (1983). A note on screening regression equations. *The American Statistician*, **37**, No. 2, 152–155.

Freedman, David A., W.C. Navidi and Steven C. Peters (1988). On the impact of variable selection in fitting regression equations. In *On Model Uncertainty and its Statistical Implications* (T.K. Dijkstra, ed.), Berlin: Springer-Verlag, pp. 1–16.

Furnival, G.M. and R.W. Wilson, Jr. (1974). Regression by leaps and bounds. *Technometrics*, **16**, 499–511.

Grusky, David B. and Robert M. Hauser (1983). Comparative social mobility revisited: Models of convergence and divergence in 16 countries. Working Paper, Center for Demography and Ecology, University of Wisconsin.

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.

Halaby, Charles L. and David L. Weakliem (1993). Ownership and authority in the earnings function: Nonnested tests of alternative specifications. *American Sociological Review* 58: 16–30.

Hazelrigg, Lawrence E. and Maurice A. Garnier (1976). Occupational mobility in industrial societies: A comparative analysis of differential access to occupational ranks in seventeen countries. *American Sociological Review* 41: 498–511.

Hoeting, Jennifer A. (1994). *Accounting for Model Uncertainty in Linear Regression.* Ph.D. dissertation, Department of Statistics, University of Washington.

Hout, Michael (1983). *Mobility Tables.* Beverly Hills, Calif.: Sage.

Hout, Michael (1984). Status, autonomy and training in occupational mobility. *American Journal of Sociology*, **89**, 1379–1409.

Hout, Michael (1988). More universalism, less structural mobility: The American occupational structure in the 1980s. *American Journal of Sociology*, **93**, 1358–1400.

Jeffreys, Harold (1935). Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophical Society*, **31**, 203–222.

Jeffreys, Harold (1961). *Theory of Probability*, (3rd ed.), Oxford University Press.

Jeffreys, Harold (1980). Some general points in probability theory. In *Bayesian Analysis in Econometrics and Statistics*, edited by A. Zellner, 451–454, Amsterdam: North-Holland.

Johnstone, David (1990a). Interpreting statistical insignificance: A Bayesian perspective. *Psychological Reports* 66: 115–121.

Johnstone, David (1990b). Sample size and the strength of evidence. *Abacus* 26: 17–35.

Kass, Robert E. (1991). About *Theory of Probability. Chance*, **4**, 13.

Kass, Robert E. and Adrian E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association*, to appear.

Kass, Robert E. and Larry Wasserman (1992). A reference Bayesian test for nested hypotheses with large samples. *Technical Report no. 567*, Department of Statistics, Carnegie Mellon University.

Kotz, Samuel and Norman L. Johnson (eds.) (1985). *Encyclopaedia of Statistical Sciences, vol. 5.* New York: Wiley.

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

Lee, Peter M. (1989). *Bayesian Statistics: An Introduction.* Oxford University Press.

Lipset, Seymour M. and H.L. Zetterberg (1959). Social mobility in industrial societies. In *Social Mobility in Industrial Society* (S.M. Lipset and R. Bendix, eds.), Berkeley, Calif.: University of California Press, pp. 11–75.

Lye, Diane and April Greek (1994). The emerging two-child norm in America. Working Paper no. 95-1, Center for Studies in Demography and Ecology, University of Washington.

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,* to appear.

McCullagh, Peter and John A. Nelder (1989). *Generalized Linear Models.* London: Chapman and Hall.

Miller, Alan J. (1984). Selection of subsets of regression variables (with Discussion). *Journal of the Royal Statistical Society (Series A),* **147**, 389–425.

Miller, Alan J. (1990). *Subset Selection in Regression.* New York: Chapman-Hall.

Morrison, Denton E. and Ramon E. Henkel (eds.) (1970). *The Significance Test Controversy.* Chicago: Aldine.

Neyman, Jerzy and Egon S. Pearson (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society,* **A 231**, 289–337.

Press, S. James (1989). *Bayesian Statistics: Principles, Models and Applications.* New York: Wiley.

Raftery, Adrian E. (1986a). A note on Bayes factors for log-linear contingency table models with vague prior information. *Journal of the Royal Statistical Society, series B,* **48**, 249–250.

Raftery, Adrian E. (1986b). Choosing models for cross-classifications. *American Sociological Review,* **51**, 145–146.

Raftery, Adrian E. (1988). Approximate Bayes factors for generalized linear models. *Technical Report no. 121,* Department of Statistics, University of Washington.

Raftery, Adrian E. (1993a). Bayesian model selection in structural equation models. In *Testing Structural Equation Models* (K.A. Bollen and J.S. Long, eds.), Beverly Hills, Calif.: Sage, pp. 163–180.

Raftery, Adrian E. (1993b). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Technical Report 255,* Department of Statistics, University of Washington.

Raftery, Adrian E., Steven M. Lewis, Akbar Aghajanian and Michael J. Kahn (1993). Event history modeling of World Fertility Survey data. Working Paper no. 93-1, Center for Studies in Demography and Ecology, University of Washington.

Raftery, Adrian E., Steven M. Lewis and Akbar Aghajanian (1994). Demand or ideation? Evidence from the Iranian marital fertility decline. Working Paper no. 94-1, Center for Studies in Demography and Ecology, 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* (J.M. Bernardo *et al.*, eds.), Oxford University Press, to appear.

Raftery, Adrian E. and Sylvia Richardson (1995). Model selection for generalized linear models via GLIB, with application to epidemiology. In *Bayesian Biostatistics* (D.A. Berry and D.K. Stangl, eds.), New York: Dekker, to appear.

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

Schwarz, Gideon (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Stigler, George J. (1970). The optimum enforcement of laws. *Journal of Political Economy*, **78**, 526–536.

Tierney, Luke and Kadane, Joseph B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, 82–86.

Vandaele, Walter (1978). Participation in illegitimate activities: Ehrlich revisited. In *Deterrence and Incapacitation*, (eds. Blumstein, A., Cohen, J. and Nagin, D.). Washington, D.C.: National Academy of Sciences, pp. 270–335.

Weakliem, David L. (1992). Comparing non-nested models for contingency tables. In *Sociological Methodology 1992*, edited by Peter V. Marsden, 147–178, Cambridge, Mass.: Blackwell.

Weisberg, Sanford (1985). *Appied Linear Regression.* New York: Wiley.

Xie, Yu (1992). The log-multiplicative layer effect for comparing mobility tables. *American Sociological Review* 57: 380–395.

Xie, Yu (1994). Log-multiplicative models for discrete-time, discrete-covariate event-history data. In *Sociological Methodology 1994*, edited by Peter V. Marsden, 301–340, Cambridge, Mass.: Blackwell.

Appendix: Software

The BIC or BIC′ approximation can be readily calculated using the output from most standard statistical model fitting software. All that is needed is that they return either the deviance or the LRT statistic against a null model, along with the number of parameters or the degrees of freedom.

Finding the models in Occam's window and averaging across them to account for model uncertainty can also be done using only the output from standard software, but it is much more time-conuming. I will now describe three pieces of software that help to make it more automatic.

BICREG: Bayesian Model Selection for Linear Regression

BICREG is an S-PLUS function which can be obtained free of charge by sending the e-mail message "send bicreg from S" to the Internet address *statlib@stat.cmu.edu*. It implements the Occam's window algorithm for linear regression using the BIC′ approximation of equation (26).

For a given dependent variable and set of candidate independent variables, it finds the models

in Occam's window and their posterior probabilities, and for each independent variable it finds $\Pr[\beta_j \neq 0 | D]$ and the posterior mean and standard deviation. It was used to carry out the analysis in Table 10.

It uses the leaps and bounds algorithm of Furnival and Wilson (1974) to identify a reduced set of good models. When there are more than 30 variables, it first uses backwards elimination to reduce the initial set of variables to 30.

BIC.LOGIT: Bayesian Model Selection for Logistic Regression

BIC.LOGIT is another S-PLUS function which can be obtained free of charge by sending the e-mail message "send bic.logit from S" to *statlib@stat.cmu.edu.* It is an adaptation of BICREG to the logistic regression setting, and gives the same outputs.

It exploits the fact that at the MLE, logistic regression is approximately a weighted least squares problem with an adjusted dependent variable (McCullagh and Nelder, 1989). To reduce the set of models to a manageable number, it converts the logistic regression problem to the equivalent weighted least squares problem and applies a liberal version of BICREG. It then calculates BIC exactly for the remaining models, and finds those that lie in Occam's window

GLIB: Generalized LInear Bayesian Modeling

GLIB is another S-PLUS function which can be obtained free of charge by sending the message "send glib from S" to *statlib@stat.cmu.edu.* It does Bayesian model selection and accounting for model uncertainty for generalized linear models, notably logistic regression and log-linear models.

It differs from BICREG in two main respects, in addition to the class of models it deals with. It does not use the BIC approximation, but instead carries out a more exact Bayesian analysis using a reference set of prior distributions (Raftery, 1993b). Results are given for a range of priors. It does not yet implement Occam's window or any model search algorithm, but requires the user to specify all the models to be considered. An epidemiological application was reported in detail by Raftery and Richardson (1995).

Comment: Avoiding Model Selection in Bayesian Social Research[13]

Andrew Gelman, University of California, Berkeley

Donald B. Rubin, Harvard University

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 $3 \times 3 \times 16$ 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 ratio-

nale 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 obser-

vation 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[14]; 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.

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

---

[14]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.

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.

Comment: Better Rules for Better Decisions

Robert M. Hauser
University of Wisconsin

January 10, 1995

1   Introduction

About a decade ago, after David Grusky and I had suffered endlessly over the choices among alternative models in our comparative analyses of social mobility (Grusky and Hauser 1984), our satisfaction with the product of those analyses was temporarily shattered by the news that Adrian Raftery (1986) would publish a methodological comment on the work. What would he have to say? And could we defend our work? Would we take the standard defensive posture of sociologists whose work was under criticism?

In the event, Raftery's brief and elegant comment turned out not to require a defense at all. Rather, it outlined a superior way to think about the decisions that we had faced, namely, how to choose among alternative models in a sample so large that standard inferential methods would lead us to reject all but a saturated model. Raftery's proposal to use the Bayesian information criterion (BIC) relieved, rather than increased our discomfort at having ignored standard rules of statistical inference, and it even supported some — though not all — of the decisions that we had made.[15]

Pleased as I was that some parts of the Grusky-Hauser analysis survived Raftery's scrutiny, I am happier yet that our efforts prompted the introduction of a simple and defensible rule of thumb that could be used to improve decisions in discrete multivariate analysis and structural equation models (Raftery 1993). For the past several years, I have routinely used BIC as a guide in model selection (Hauser and Wong 1989; Wong and Hauser 1992; Hout and Hauser 1992; Hauser 1993; Hauser and Phang 1993; Kuo and Hauser 1995a, 1995b), sometimes without showing the details of inferential procedures in the text.

Raftery has now explained and elaborated the uses of Bayesian inference in a wider context. In this brief commentary, I should like to offer a few additional examples and suggestions about the use of Bayesian methods of model selection, as well as to encourage readers to read, evaluate, and use Raftery's contributions in their own teaching and research.

---

[15] At the time, as I recall, we even wanted to write a "reply" to the comment that would express our agreement and gratitude to Raftery, but my recollection is that the editor discouraged this unexciting, though original use of journal pages.

2   Further Reflections on Comparative Social Mobility

I should like to start by adding a footnote to Raftery's exposition of the right way to do the Grusky-Hauser analysis. In the early 1980s, the central idea in cross-national mobility analysis was the so-called Featherman-Jones-Hauser (FJH) hypothesis (Featherman, Jones, and Hauser 1975). It said that, once marginal distributions were controlled (in a log-linear model), the interactions (odds-ratios) in a mobility table were essentially invariant across societies. Grusky and Hauser wanted to show that mobility regimes in different countries were similar, but not identical, in order to motivate their effort to explain the cross- national variations in interactions without entirely denying the usefulness of the FJH hypothesis. Thus, among other models, they estimated a variant of the quasi-symmetry (quasi-perfect mobility) model, shown in line 3 of Raftery's Table 2, and a variant of the saturated model, shown in line 4 of Raftery's Table 2. Both models specified that the interaction effects were the same in all 17 countries. Each model yielded a nominally significant likelihood-ratio test statistic, relative to its less constrained counterpart (1350 with 45 df and 1329 with 60 df, respectively).

In the article, Grusky and Hauser (1984: 26) waffled about this, first noting that these contrasts were "highly significant statistically," then saying that "The results ... imply convergence ..." (both among industrialized countries and in the full sample of 17 countries), and, finally, going on to model cross-national differences. Their analysis eventually yielded the explanatory model whose satisfactory fit is reported in line 5 of Raftery's Table 2. Had Grusky and Hauser used BIC, they would have been able to accept the model of cross-nationally variable quasi-symmetry, but they would also have had to reject outright both the model of constant quasi-symmetry (BIC = 826) and that of constant association (BIC = 631). Grusky and Hauser did go on in the right direction, but not for quite the right reason. Had they been more forthright in this early part of the history of the FJH hypothesis, it is possible that others would not later have gone to such great lengths to sustain it (Erikson and Goldthorpe 1992).

What is the moral of this story? The first thing most researchers will notice about the decision rules that Raftery elaborates is that they are far more conservative, at least in survey-based samples, than application of the customary criteria of P = 0.05 or P = 0.01. In my experience, people are uncomfortable with this kind of conservatism – with or without the legitimation of the BIC approximation – even though its rewards are likely to include both scientific parsimony and external validity. I like this elaboration of Raftery's example because it shows how a conservative, but informal decision rule can also go wrong. In this case, BIC not only tells us which "significant" findings should be ignored, but which should have been pursued most seriously.

3   Assessing the Replicability of Findings on Occupational Scaling

Raftery's text several times notes the value of BIC in improving "out-of-sample predictions," that is, in yielding replicable findings. This phrase suggested another application of BIC, namely, to

an analysis in which the investigators could not determine the expected value of a new measure of correlation.

In the May 1992 American Journal of Sociology (pp. 1658 ff), there was an extensive symposium on the scaling of occupational status. Rytina (1992a) introduced a method for constrained (symmetric) canonical scoring of the rows and columns of an intergenerational social mobility table, the Symmetric Scaling of Intergenerational Continuity or SSIC. He estimated scores for a 308 by 308 occupational mobility table (from the General Social Survey, 1972 to 1986, $n = 7965$) and obtained a much higher correlation between father's and adult child's occupation (0.450) than was obtained with the Stevens-Featherman variant of the Duncan SEI (0.338). He concluded that the intergenerational correlation was truly higher and that the SEI failed to reflect a large share of the vertical dimension of occupational stratification.

Hauser and Logan (1992) responded by cross-validation and by example. First, in a fresh sample (the GSS from 1987 to 1990) Rytina's SSIC scores (from the 1972 to 1986 data) were less highly correlated across generations than was the SEI. Second, arguing by analogy, they showed that the canonical correlation (not that based on the SSIC), when corrected for loss of degrees of freedom using a formula of Lawley (1959), was of about the same magnitude as the observed correlation based on the SEI.[16] They were unable to come up with an analytic correction of the correlation based on the SSIC. Rytina (1992b) rejoined at length, refusing to accept the idea that sampling variability increased the SSIC correlation.

Raftery's paper suggests a simple analysis that should resolve the issue: Compute BIC' for the two competing models (from his equation 26). With $N = 7965$, $p_1 = 1$, and $R_1^2 = 0.338^2$, BIC' = -957. With the same sample size, but $p_2 = 307$ and $R_2^2 = 0.450^2$, BIC' = 955.[17] In other words, Rytina's model fares worse than no model at all. Rytina's model would be superior to that based on the SEI if it could produce a correlation as large as 0.450 with 92 or fewer parameters. That would yield BIC' = -976. Any larger number of parameters would yield a difference in BIC' (relative to the SEI equation) less than 10 (in absolute value). Obversely, using as many as 307 parameters, one would have to obtain a correlation of 0.62 or more to improve on what the SEI does with one parameter. In short, this simple analysis would lead one to expect the failure of cross-validation that was observed by Hauser and Logan.[18]

---

[16] The idea is the same as that of correcting a multiple correlation for its degrees of freedom.

[17] This illustrates a use of Raftery's equation 26 that he did not elaborate, namely, to evaluate differences in overall fit (R^2) between linear models that differ in more than one degree of freedom. The application to canonical scoring is a bit unusual; a more common application will be to tests of the effects of sets of categories, as in dummy-variable regression analysis.

[18] Hauser and Logan (1992: 1694) introduced a design effect of two-thirds for the GSS sample, and its use here would yield an even stronger rejection of Rytina's finding. In a sample two-thirds as large as the cumulative GSS ($n = 5310$), using 307 parameters, one would have to obtain an SSIC correlation of 0.68 to compete with that based on the SEI, or, obversely, one would have to obtain an SSIC correlation of 0.45 with only 64 parameters.

4 Selection of Variables in Single-Equation Models

When I first read an earlier draft of Raftery's paper, I was put off slightly by his emphasis on the problem of variable selection in single- equation models. As described above, my introduction to BIC took place in the context of alternative specifications of models for cross- classification tables. My research of late has focused on constraints on slopes, means, and variances within and between populations in structural equation models. Why should we worry so much about inference in single-equation models?

First, the message comes through loud and clear that Raftery's main ideas apply as well to these other modeling situations as to single- equation models, even if the machinery for identifying the sets in Occam's Window is not yet available. One most valuable suggestion of Raftery's paper in this respect is that standard software for structural equation models should extend displays of t-statistics and modification indices to approximate BIC (as well as reporting BIC along with other standard indices of model fit). Thus, as he notes, from estimation of a single model, we can easily approximate Bayes factors both for the deletion of every parameter in the model and for the release of every constrained parameter of the model.

Second, despite my interest in other aspects of structure, the more I think about it, the more I think sociologists can profit from better methods for the selection of variables in relatively simple models. In this context, Raftery's illustrative analysis of the problem of model uncertainty in the data on crime and punishment deserves close and repeated reading. It provides a powerful example of how we ought to proceed when our interest lies in the full set of variables that belong (or do not belong) in an equation.

For example, Jencks, Perman, and Rainwater (1988) introduced a new index of job desirability (IJD), which, they argued, might be preferable to occupational status or earnings in analyses of social stratification. To construct the index, they used a magnitude scaling item to measure the desirability of the job held by each adult in a national Survey of Job Characteristics (SJC, $N = 809$). Then, they regressed the desirability of the job on a vector of 48 job characteristics, eliminating all but the 14 that proved statistically significant.[19] Their proposal is that, rather than ascertaining occupation and mapping it into a status or prestige scale, future researchers should simply ascertain the significant occupational characteristics and weight them as they affected job desirability in the SJC.

Which and how many job characteristics would have appeared in the IJD if Jencks, Perman, and Rainwater had used Bayesian methods of model selection? A full analysis would require location of Occam's Window, but it is instructive to note that to meet Raftery's criterion of "very strong" evidence for inclusion (BIC = 10), a variable should have a t-statistic greater than 4.09 in a sample of 809. This criterion is met by only two of the variables in the published, 14-variable version of the

---

[19] Jencks, Perman, and Rainwater (1988: 1336-7) were not unmindful of the problem of replication. In fact, they estimated the reliability of the IJD using independent half-samples. However, they do not appear to have cross-validated their selection of regressors.

IJD equation, the earnings and the educational requirements of the job (p. 1336).[20] Only five of the 14 variables meet Raftery's less stringent criterion of "strong" evidence. Perhaps further analytic work could identify a small enough set of valid predictors of the IJD to warrant measurement of those components on a regular basis, even outside specialized studies of social stratification.

5   Introducing BIC in Research and Teaching

It is well that we now have a lucid, complete, and relatively non- technical guide to the use of BIC and BIC$'$. It is not easy to break old habits or to teach students to develop the right habits in the conduct of research. In the fall of 1994 I introduced BIC in my course in structural equation models, and it was often difficult and in some cases impossible to persuade students not to reject a null hypothesis, especially one pertaining to the overall fit of a model, merely because the deviance was nominally statistically significant with P = 0.05 or P = 0.01. We want so badly to reject null hypotheses that we often do so badly – and often, thanks to the miracles of modern software – without much regard for the theoretical ideas that we actually wish to test. Perhaps one inducement to good practice will be Raftery's finding that, just as we set nominal probability levels too high in large samples, we also set them too low in small samples.

While I hope that Raftery's paper will lead readers to appreciate the meaning of Bayes factors, in practice it may be difficult even to introduce the rules of thumb suggested by Raftery in his Tables 6, 7, 8, and 9. Perhaps it will help to photocopy those tables and post them visibly in statistical laboratories. They ought to be circulated by journal editors along with guidelines to authors. They ought to become as commonplace as tables of the distribution of t or of χ^2. The formulas on which the tables are based are so simple that I would encourage readers to validate the tables in a computer spreadsheet. Having done that, the spreadsheet will also estimate critical values of BIC for the sample at hand, and it can be used to calculate BIC from standard computer output.[21]

Another good way to introduce Bayesian selection methods will be to encourage students (and researchers) to compare standard model selection procedures with the use of BIC in simulated, null data or in repeated samples. Those who lack theoretical skills in statistics may well be able to carry out their own analyses of simulated data, just as described by Raftery. Another straightforward exercise, accessible to any analyst, will be to combine half-sample replication with the use of BIC and other model selection procedures. For example, split a sample into two random halves. Run a regression equation in one half-sample, and choose which variables should enter the equation using several decision rules, for example, all significant variables in the full equation, forward selection, backward selection, and BIC$'$. Then estimate the same equation in the second half-sample and

---

[20] It is striking that these two job-level variables are analogs of the components of occupational socioeconomic status.

[21] For many years, I have constructed statistical tables by importing standard output into a spreadsheet and using the spreadsheet both to format the table and to perform and document auxiliary calculations.

compare the pattern of significant coefficients with those determined in the first-half sample under varying decision rules. Such exercises ought to enter the statistical curriculum.

Over the past twenty years, the frontier of sociological modeling has advanced rapidly, but with few exceptions, there has been little progress in our practice of statistical inference. To Raftery's credit, he makes no grandiose claims or promises about the value of Bayesian methods, and he also emphasizes the importance of other theoretical and methodological aspects of the research process. This is good as well as modest advice, and I second Raftery's observation that there are no methodological panaceas. At the same time, I hope and expect that Raftery's exposition of Bayesian methods of model selection represents the beginning of their widespread use in sociological research as well as an invitation to further advances in inferential methods. Raftery's methodological work, like that of Leo Goodman, is truly useful to researchers, and in use it will repay close and repeated study.

References

Erikson, Robert, and John H. Goldthorpe. 1992. *The Constant Flux: A Study of Class Mobility in Industrial Societies.* Oxford: The Clarendon Press.

Featherman, David L., F. Lancaster Jones, and Robert M. Hauser. 1975. "Assumptions of Social Mobility Research in the U.S.: The Case of Occupational Status." *Social Science Research* 4 (December): 329-60.

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

Hauser, Robert M. 1993. "Trends in College Entry Among Blacks, Hispanics, and Whites." In *Studies of Supply and Demand in Higher Education.* Edited by Charles Clotfelter and Michael Rothschild. National Bureau of Economic Research. Chicago, Illinois: University of Chicago Press.

Hauser, Robert M., and John A. Logan. 1992. "How Not to Measure Intergenerational Occupational Persistence." *American Journal of Sociology* 97(6)(May): 1689-1711.

Hauser, Robert M., and Hanam Samuel Phang. 1993. "Trends in High School Dropout Among White, Black, and Hispanic Youth, 1973 to 1989." *IRP Discussion Paper*, vol. 1007-93. Madison, Wisconsin: Institute for Research on Poverty.

Hauser, Robert M., and Raymond Sin-Kwok Wong. 1989. "Sibling Resemblance and Inter-sibling Effects in Educational Attainment." *Sociology of Education* 62 (July): 149-71.

Hout, Michael, and Robert M. Hauser. 1992. "Symmetry and Hierarchy in Occupational Mobility: A Methodological Analysis of the CASMIN Model of Class Mobility." *European Sociological Review* 8 (December): 239-66.

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(6)(May): 1322-57.

Kuo, Hsiang-Hui Daphne, and Robert M. Hauser. 1995a. "Gender, Family Configuration, and the Effect of Family Background on Educational Attainment." *Social Biology* (forthcoming).

———- . 1995b. "Black-White Differentials in the Effect of Family Background on Educational Attainment of U.S. Men Born from 1907 to 1946: A Study of Sibling Resemblance." *Sociology of Education* (forthcoming).

Lawley, D.N. 1989. "Tests of Significance in Canonical Analysis." *Biometrika* 46: 59-66.

Raftery, Adrian E. 1986. "Choosing Models for Cross-Classifications (Comment on Grusky and Hauser)." *American Sociological Review* 51(1)(February): 145-6.

———- . 1993. "Bayesian Model Selection in Structural Equation Models." Pp 163-80 In *Testing Structural Equation Models*, edited by Kenneth A. Bollen and J. Scott Long. Newbury Park: SAGE Publications.

Rytina, Steve. 1992a. "Scaling the Intergenerational Continuity of Occupation: Is Occupational Inheritance Ascriptive After All?" it American Journal of Sociology 97(6)(May): 1658-88.

———- . 1992b. "Response to Hauser and Logan and Grusky and Van Rompaey." *American Journal of Sociology* 97(6)(May): 1729-48.

Wong, Raymond Sin-Kwok, and Robert M. Hauser. 1992. "Trends in Occupational Mobility in Hungary Under Socialism." *Social Science Research* 21(4)(December): 419-44.

Rejoinder: Model Selection is Unavoidable in Social Research[22]

Adrian E. Raftery
University of Washington

January 20, 1995

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

1 Introduction

I would like to thank Hauser and Gelman & 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 & 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 & 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.

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 (see Findley, 1991, and the discussion of it in Kass and Raftery, 1995, Section 8.3). 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% 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% 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 which 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 data sets. See Western (1994) for related discussion.

3 Response to Gelman & Rubin

3.1 Points of Agreement

Gelman & Rubin and I agree with one another (a) that classical frequentist methods fail and that my suggested methods based on BIC can point in better directions; and (b) 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 & 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.[23]

Gelman & 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 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.[24] 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 & 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 & 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 & 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.

This argument is made in detail by Berger and Sellke (1987), and pithily summarized in the

---

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

[24]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.

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

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

Gelman & 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 (a) a fairly general theoretical result; and (b) 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, equation (4)) gave a theoretical result showing that, on average over models and over data sets, 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 data sets 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%.

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, Section 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 & Rubin say that BIC corresponds to a uniform, improper prior. (This is a prior distribution which does not integrate to one, and 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 is especially accurate 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 & Rubin's last paragraph does not apply to the results in my chapter.

Gelman & 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 & 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 & 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.[25]

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 expected 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 (a) the researcher knows the purpose to which his or her results will be put; (b) the costs and benefits associated with plausible combinations of outcome and model selected can be assessed fairly accurately; and (c) 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 which 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 & 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 tranforming 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 which should be made based on the data.

Gelman & Rubin suggest that, rather than selecting a subset of the 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 sceptical 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 & Rubin's approach could also be

---

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

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 & Rubin dismiss the simulated examples of Sections 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 & 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 & 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 non-significant 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**, 317–352.

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**, 112–122.

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: 319–327.

Ehrlich, Isaac (1973). Participation in illegitimate activities: a theoretical and empirical investigation. *Journal of Political Economy*, **81**, 521–565.

Findley, David F. (1991). Counter examples to parsimony and BIC. *Annals of the Institute of Statistical Mathematics*, 43: 515–514.

Freedman, David A. (1983). A note on screening regression equations. *The American Statistician* 37: 152–155.

Gelman, Andrew, Xiao-Li Meng and Hal S. Stern (1995). Bayesian model checking using tail-area probabilities. Under revision for *Statistica Sinica*.

Good, Irving John (1952). Rational decisions. *Journal of the Royal Statistical Society, series B*, 14: 107–114.

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. Pp. 451–454 in *Bayesian Analysis in Econometrics and Statistics*, edited by Arnold Zellner, 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: 1322–1357.

Kadane, Joseph B. and Dickey, James M. (1980). Bayesian decision theory and the simplification of models. Pp. 245–268 in *Evaluation of Econometric Models* (J. Kmenta and J. Ramsey, eds.), New York: Academic Press.

Kass, Robert E. and Adrian E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association*, to appear.

Kass, Robert E. and Larry Wasserman (1995). A reference Bayesian test for nested hypotheses with large samples. *Journal of the American Statistical Association*, to appear.

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: 1535–1546.

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). To appear in *Bayesian Statistics 5* (J.M. Bernardo *et al.*, eds.), Oxford University Press.

Raftery, Adrian E. and Judith E. Zeh (1993). Estimation of bowhead whale, *Balaena mysticetus*, population size (with Discussion). Pp. 160–240 in *Bayesian Statistics in Science and Technology: Case Studies*, (C. Gatsonis, J.S. Hodges, R.E. Kass and N.D. Singpurwalla, eds.), 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: 1151–1172.

Rytina, Steve (1992). Scaling the intergenerational continuity of occupation: Is occupational inheritance ascriptive after all? *American Journal of Sociology* 97: 1658–1688.

Western, Bruce (1994). Vague theory in macrosociology. Paper presented to the Annual Meeting of the American Sociological Association, Los Angeles, August 1994.