Volume Title: Evaluation of Econometric Models

Volume Author/Editor: Jan Kmenta and James B. Ramsey, eds.

Volume Publisher: Academic Press

Volume ISBN: 978-0-12-416550-2

Volume URL: http://www.nber.org/books/kmen80-1

Publication Date: 1980

Chapter Title: Bayesian Decision Theory and the Simplification of Models

Chapter Author: Joseph B. Kadane, James M. Dickey

Chapter URL: http://www.nber.org/chapters/c11704

Chapter pages in book: (p. 245 - 268)

Bayesian Decision Theory and the Simplification of Models

JOSEPH B. KADANE

DEPARTMENT OF STATISTICS
CARNEGIE-MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA

and

JAMES M. DICKEY

DEPARTMENT OF MATHEMATICS AND STATISTICS
STATE UNIVERSITY OF NEW YORK AT ALBANY
ALBANY, NEW YORK

1. Introduction

The problem of evaluating econometric models is here viewed as a particular case of a general class of problems called decision problems. Since the authors are attracted to a particular approach to decision problems, the Bayesian approach, we have written this paper in order to see what light can be shed by Bayesian decision theory on the evaluation of econometric models.

Briefly, the decisions d that might be chosen lie in some decision space $\mathcal{D}$. There is some space Ω of possible alternative circumstances θ, and your opinion over this space is given by the probability density with element $p(\theta)\,d\theta$. If one's utility function is $U(d, \theta)$, representing the relative desirability of the decision d if the alternative θ were true, Bayesian decision theory recommends that you choose that decision d which maximizes the integral

$$\int_{\Omega} U(d, \theta)p(\theta)\,d\theta \tag{1.1}$$

(See F. Ramsey, 1931, Von Neumann & Morgenstern, 1947, and Savage,

ISBN 0-12-416550-8

1954, for the theoretical background of this choice, called the principle of maximum expected utility.)

The view that this simple paradigm might play a fundamental role in statistics by providing a theory analogous to price theory in microeconomics has been met with considerable controversy among statisticians. Some statisticians maintain that the inputs, particularly the opinion $p(\theta)\,d\theta$ and the utility function $U(d, \theta)$, are very difficult to obtain in applied contexts, while others resist the notion of any kind of organizing principle for statistics. The burden of proof is on Bayesians to show that the principle suggested by (1.1) is useful as a way of thinking about statistical problems and that the inputs are worth the trouble to discover. In this light, the problem of choice between econometric models is a fine test case.

While we will speak of econometric models in general, the specific model we will refer to and have in the back of our minds most often is the choice between linear regressors, especially the choice whether or not to add some specified variable to a model. However, because our viewpoint is very general (and this is a distinct advantage of Bayesian theory), much of what we say has much wider applicability.

An enormous literature already exists on these and related questions, some of which is reviewed in Gaver & Geisel (1973), Harvey & Collier (1977), Hocking (1976), and J. Ramsey (1974). We cannot hope to take up each of the methods reviewed in these papers, but rather will deal with strengths and weaknesses we find in classes of such methods.

Some methods, typified by the use of tests of hypotheses, ask whether or not the submodel is a "true" restriction of the larger model. The level of the test is the probability under the submodel that the statistic in question would be as large as observed or larger. However, practicing statisticians have long been aware that the question of the exact "truth" of a sharp statistical hypothesis is hardly ever appropriate (Berkson, 1938). Nearly every model used in econometrics is known *a priori* to be literally false. Because of this, although a typical statistical test with a small sample will rarely reject, for large enough samples it will nearly always reject. Consequently, whether or not a test rejects a null hypothesis appears to have more to do with the sample size than with how true the null hypothesis is. For these reasons we find methods based on classical tests of hypotheses to be most unsatisfactory. A purist may object, saying that the level of a test should be set in consideration of the power that can be achieved with that sample size, but after all these years there is no satisfactory theory of how to decide what level of test to set, and most practitioners use .05 or .01 in an automatic way, without justification.

The fundamental problem with the test of a hypothesis, we believe, lies not in the failure to jiggle the level satisfactorily, but rather with the underlying fact of having asked a poor question. No method geared to the question

Is the null hypothesis true? will be statisfactory in the most common situation where we know the null hypothesis to be false. Even if one attempts to wriggle out of this argument by claiming that what is really meant by $x \sim N(10.0, 1.0)$ is that the mean μ satisfies the inequality $9.950 \leq \mu \leq 10.049$ and the variance σ^2 satisfies the inequality $0.950 \leq \sigma^2 \leq 1.049$, and one modifies the test accordingly, the above argument still holds with "false" changed to "extremely unlikely". Of course, if the intervals are widened, their joint probability content grows, but then they become that much less a representation of $X \sim N(10.0, 1.0)$.

The important question in practice is not whether a true effect is nonzero, for it is already known not to do exactly zero, but rather, How large is the effect? But then this question is only relevant in terms of How large is how important? This question in turn depends on the use to which the inference will be put, namely, on the utility function of the concerned scientist. Approaches which attempt to explain model simplification from the viewpoint of the inappropriate question, Is it true that . . . ? have a common thread in that they all proceed without reference to the utility function of the scientist. And therefore, from the decision theory view, they all impose normative conditions on the utility function which are seldom explicit and often far from the case in practice.

We explore in this paper the consequences of asking the more relevant question Is it useful to assume that . . . ?, which requires us to be more explicit about the investigator's utility function. We find the utility function in Bayesian theory to be especially useful to explain the behavior of a scientist who, with the same data (and hence opinion) studying the same phenomenon, sometimes uses a simpler model and sometimes a more complex one, depending on his purpose. His view of the truth has not changed, but his purpose, and hence his utility function, has.

In response to the popular interest in questions of "truth" and the natural interpretation of the posterior distribution as a measure of knowledge on such questions, we begin by examining the role of the posterior odds for a statistical hypothesis within a general two-decision setting (Section 2). We identify the extent to which the posterior odds can be a useful summary of data for decision making independent of the detailed utility structure. Our finding is essentially negative, that the odds are typically not useful as a data summary in this sense.

We turn then to the specific problem of how to choose predictors in the normal multiple regression model (Section 3). Without nontrivial prior belief in any null hypothesis, we consider the question of how to decide whether or not to accept a restriction or "simplification" on the class of available predictors. In other words, attention is redirected from the simplification of sampling models to the simplification of predictors. Results are

summarized and compared for predictive utilities in the familiar squared-error-loss form (Lindley, 1968) and a new predictive utility in the form of a probability density function (Dickey & Kadane, 1977). Comparisons are made to various truth-test criteria and traditional predictor-selection criteria. A general montonicity is noted for the predictive expected utility as a function of the predictor class, whereby cost-free predictor variables tend to be included in the optimal predictor on a foregone basis. We discuss the consequent paradox of Bayesian overfitting (Section 4).

Alternative Bayesian discussions of selecting regression models are given by Leamer (1978a,b) and by Davis & DeGroot (1978).

2. Bayesian Response to Data; Odds

When one obtains new data D, Bayesian theory suggests how one should take this data into account in changing one's opinion about θ. Suppose that the probability of observing D, if θ were the true state of nature, is, under one's opinion, given by the function $l(D|\theta)$. Then one's new opinion should be represented by your conditional distribution of θ, given D, given by

$$p(\theta|D)\,d\theta = l(D|\theta)p(\theta)\,d\theta \Big/ \int_\Omega l(D|\theta)p(\theta)\,d\theta. \tag{2.1}$$

Technically, Eq. (2.1) has used a rather simple result from probability theory called Bayes' theorem, from which the name Bayesian statistics derives. Of course, the opinion used in the earlier Eq. (1.1) must be your own current opinion at the time of the decision, and hence must take into account all of the information available to you, that is, one's opinion must be changed by Eq. (2.1) every time new data appear.

We now undertake to model the simplest sort of decision that might be made. We suppose that there are two hypotheses, H and H^c. You are sure both cannot be true ($H \cap H^c = \varnothing$), but that one of them is true ($P(H \cup H^c) = 1$). We denote the *odds for* H, $O(H)$, as

$$O(H) = P(H)/P(H^c) = P(H)/(1 - P(H)) \tag{2.2}$$

When new data D are observed, the prior odds $O(H)$ change to posterior odds $O(H|D)$ in the following way, using Bayes Theorem:

$$O(H|D) = B_D(H)O(H), \tag{2.3}$$

where

$$B_D(H) = p(D|H)/p(D|H^c)$$

and

$$p(D|J) = \int_J l(D|\theta)p(\theta)\,d\theta \Big/ \int_J p(\theta)\,d\theta \qquad \text{for} \quad J = H, H^c.$$

The quantity $B_D(H)$ is called the Bayes factor and is constant among persons having the same likelihood and the same conditional priors, given H and H^c. The usefulness of the Bayes factor as an expression for how strongly your current beliefs support H has been explored in a series of papers by Jeffreys (1961), Edwards, Lindman, & Savage (1963), and Dickey (1971, 1973a, 1974, 1976, 1977, 1979); see also Lindley (1957, 1961).

The goal of this section is to show that Bayes factors have a very limited usefulness in the two-decision context. We note that Bayes factors do appear in other contexts (see, for example, Zellner & Vandaele (1975, pp. 640–641)) not discussed here.

2.1. Evidence for a Dichotomous Decision

Suppose that the relevant decision space D has only two desicions, d_1 and d_2, which we can think of as deciding in favor of or against H, respectively. With respect to the utility function $U(d, \theta)$, the optimal decision is to choose that decision d at which the maximum of $\{\bar{U}_1, \bar{U}_2\}$ occurs, where

$$\bar{U}_i = E[U(d_i, \tilde{\theta})] = \int_\Omega U(d_i, \theta)p(\theta)\,d\theta, \qquad i = 1, 2. \tag{2.4}$$

(Here the tilde over a symbol emphasizes that the symbol denotes a random variable.) The expectation is conditional on all information, including any sample data, as explained in Eq. (2.1).

We need a notation for the conditional expectation of a utility function under the different hypotheses. Thus we write

$$\bar{U}_i(J) = E[U(d_i, \tilde{\theta})|J] = \int_J U(d_i, \theta)p(\theta)\,d\theta \Big/ \int_J p(\theta)\,d\theta, \tag{2.5}$$

for $i = 1, 2$ and $J = H, H^c$. Using this notation, we can write

$$\bar{U}_i = \bar{U}_i(H)P(H) + \bar{U}_i(H^c)P(H^c), \qquad i = 1, 2. \tag{2.6}$$

Finally, we introduce the notation $V(\theta) = U(d_1, \theta) - U(d_2, \theta)$, and write

$$\bar{V}(J) = E(V(\tilde{\theta})|J) = \bar{U}_1(J) - \bar{U}_2(J), \qquad J = H, H^c. \tag{2.7}$$

Then

$$\begin{aligned} \bar{U}_1 - \bar{U}_2 &= \bar{U}_1(H)P(H) + \bar{U}_1(H^c)P(H^c) - \bar{U}_2(H)P(H) - \bar{U}_2(H^c)P(H^c) \\ &= \bar{V}(H)P(H) + \bar{V}(H^c)P(H^c). \end{aligned} \tag{2.8}$$

Without loss of generality, we assume $\bar{V}(H) > 0$, so that if we knew H were true, with our present distribution of probability over H, we would prefer to make decision d_1 (H is true), rather than decision d_2 (H^c is true). Then it is easy to see from (2.8) that $\bar{U}_1 \geq \bar{U}_2$ if and only if

$$O(H) = P(H)/P(H^c) \geq -\bar{V}(H^c)/\bar{V}(H). \tag{2.9}$$

It is tempting to consider (2.9) as a threshold criterion on the odds. To do so, we are particularly interested in the situation when only $O(H)$, and not the right-hand side of (2.9), depends on the distribution over Ω.

TABLE 1

		True	
		H	H^c
Accept	H	0	$-L_{12}$
	H^c	$-L_{21}$	0

An important special case of this model is where the utilities only depend on whether $\theta \in H$ or $\theta \in H^c$. Thus we might have the utility structure in Table 1 (see, for example, Edwards *et al.*, 1963; Zellner, 1971; and Gaver & Geisel, 1973), where L_{12} and L_{21} are both typically positive. The analysis above specializes to the criterion "accept" H (decide d_1) iff

$$O(H) \geq L_{12}/L_{21}. \tag{2.10}$$

After data D are observed, (2.10) can be expressed equivalently as

$$O(H|D) \geq L_{12}/L_{21}, \tag{2.11}$$

or, by using (2.3) with (2.11),

$$B_D(H) \geq L_{12}/[L_{21}O(H)]. \tag{2.12}$$

Thus the Bayes factor is the appropriate function of the data on which to have a threshold (and $L_{12}/[L_{21}O(H)]$ is the appropriate threshold), above which it is optimal to choose d_1, and below which it is optimal to choose d_2. This optimistic special case suggests that perhaps there is something special about the Bayes factor in general that makes it a canonical summary of the data. That this is not so is the burden of the next result.

For the Bayes factor to be a sufficient summary of the data, the right-hand side of (2.9) must not depend on the data. However, when data are available, the expectations on the right-hand side of (2.9) are *posterior* expectations and in general depend on the data. A sufficient condition for their independence

of the data is that $V(\theta)$ be constant within H and within H^c. In general, the condition is also necessary in a sense now to be defined. The result applies separately to the numerator and to the denominator of the right-hand side of (2.9), though the statement is made in terms of the denominator.

Theorem. *Let $\mathscr{P}_H$ be the class of all continuous-type distributions with densities $p_H(\theta)$ over an analytic segment H of a finite dimensional Euclidean space. Assume $V(\theta)$ is bounded and continuous. Then the expectation $E[V(\theta)|H] = \int V(\theta)p_H(\theta)\,d\theta$ is constant in p_H over $\mathscr{P}_H$ if and only if V is constant in θ over H.*

Proof. The sufficiency of constant V is immediate. To see that this condition is also necessary, assume that it does not hold. Suppose θ_1, θ_2 are two values in H for which $V(\theta_1) > V(\theta_2)$. Choose neighborhoods N_1 and N_2 of θ_1 and θ_2, respectively, for which the two image sets are disjoint, $V(N_1) \cap V(N_2) = \varnothing$. If two densities p_H^1 and p_H^2 have their support sets contained in N_1 and N_2, respectively, then $\int V(\theta)p_H^1(\theta)\,d\theta > \int V(\theta)p_H^2(\theta)\,d\theta$ (since the two integrands satisfy this same inequality), which then contradicts the constancy of the expection. ■

The proof extends easily to any subclass of $\mathscr{P}_H$ which contains distributions having arbitrarily small probability for the complements of arbitrary neighborhoods. This typically true for the set of conditional posterior distributions from all possible data in a given statistical decision problem. Thus we see that when utility takes more than two values within H and/or H^c, the threshold on the odds, and hence the preferred decision, will depend on the conditional uncertainties given H and given H^c.

Consequently, the adequacy of the Bayes factor as a summary of the data depends on the acceptability of a utility structure similar to Table 1. Such a structure cannot be close to correct for the most common problems because if I wrongly accept H (make decision d_1), it matters for almost every practical purpose whether, although wrong, H is close to correct or very far from correct. Thus in the simple versus composite case (say, testing whether or not a normal mean is 10.0), the Bayes factor cannot summarize the data. Of course, this holds as well for the yet more common composite versus composite cases.

For this reason it is our judgement that the Bayes factor theory, originally justified in the simple versus simple case as the Bayesian version of hypothesis testing, does not fill the need of econometricians for a useable, theoretically well-based way of choosing models. Note that we arrive at this conclusion even when positive probability is put on the truth of the null hypothesis, although, as explained in Section 1, we doubt that this is appropriate in general.

2.2. Hypothesis Defined by the Decision Problem

In Section 2.1 we started with a hypothesis H (for example, $\theta = 10.0$) and an alternative (for example, $\theta \neq 10.0$), and found (restrictive) conditions under which the Bayes factor is sufficient as a representation of the impact of new data. Our method was essentially to take the hypothesis H (and therefore also its complement H^c) as fixed and to try to modify the utility functions in ways that would lead to the Bayes factor. In this section, by contrast, we try the reverse, that is, we begin with a reasonable utility function for a two-decision problem and work backward to see what sort of hypothesis H a procedure based on the Bayes factor would imply. We propose to choose the hypothesis H by the requirement

$$\begin{aligned} &U(d_1,\theta) \geq U(d_2,\theta) \quad \text{for all} \quad \theta \in H, \\ &U(d_1,\theta) < U(d_2,\theta) \quad \text{for all} \quad \theta \in H^c. \end{aligned} \tag{2.13}$$

Then d_1 would be the preferred decision for any known θ in H, and d_2 would be preferred for any known θ in H^c. Consequently, we define a hypothesis H for inference relative to a given two-decision problem by

$$H = \{\theta : U(d_1,\theta) \geq U(d_2,\theta)\}. \tag{2.14}$$

To connect this theory with the kind of theory discussed in Section 2.1, consider, for example, what happens if our objective is to give an accurate estimate of θ. Let us suppose that decision d_1 simplifies the model by declaring that $\theta = \theta^*$, when θ^* is some particular value (like $\theta^* = 10$, for example). In this case, suppose that our reward is some function of θ, say $S_1(\theta)$. Alternatively, if θ is not declared to have the value 10, then it must be estimated, say by some $\hat{\theta}$. Then the reward that we receive is a function of both $\hat{\theta}$ and θ, say $S_2(\hat{\theta},\theta)$. Anscombe (1963) considers an estimation loss function which is constant if the parameter is freely estimated ($S_2(\hat{\theta},\theta) = U_2$) but proportional to the square of the parameter if the estimate is taken to be zero ($S_1(\theta) = U_1\theta^2$, where U_1 is negative).

Now of course, $\hat{\theta}$ must be chosen as best it can be if the second decision is made, that is, $\hat{\theta} = \hat{\theta}_D$, where $\hat{\theta}_D$ maximizes $E[S_2(\hat{\theta},\tilde{\theta})|D]$.

Of particular interest are utility pairs differing by a constant "reward for simplicity" $U^* > 0$, where

$$S_1(\theta) = S_2(\theta^*,\theta) + U^*. \tag{2.15}$$

In this case, we obtain by (2.14) the "hypothesis"

$$H = \{\theta : S_2(\hat{\theta}_D,\theta) - S_2(\theta^*,\theta) \leq U^*\}. \tag{2.16}$$

Consider, for example, the familiar negative squared error, $S_2(\hat{\theta},\theta) = -(\hat{\theta} - \theta)^2$. Then, if θ is to be freely estimated, the optimal estimate will be the

posterior mean $\hat{\theta}_D = E[\tilde{\theta}|D]$, and H takes the form

$$H = \{\theta\colon \theta \lesseqgtr \tfrac{1}{2}(\theta^* + \hat{\theta}_D) - \tfrac{1}{2}U^*/(\theta^* - \hat{\theta}_D)\}, \tag{2.17}$$

where the sense of the inequality is taken to match that of $\theta^* \lesseqgtr \hat{\theta}_D$. Thus H turns out to be an infinite half-line, and worse, one that depends on the data through $\hat{\theta}_D$.

Our attempts to find a link decision theory and the problem of choosing a probability model have failed, and we make no further such attempts here. In both cases we sought to retain some kind of formal decision theoretic meaning for classical hypothesis-testing ideas. A similar such failure is reported in Edwards *et al.*(1963).

The consequence of this failure is that something must go, either our Bayesian approach or classical hypothesis testing. Having declared our purpose in this paper to explore the consequences of Bayesian decision theory for choosing models, with particular reference to econometrics, it will come as no surprise to our readers that our choice is to keep our Bayesian decision theory and see what its consequences are. In the next section, then, we begin anew to see where Bayesian theory will lead us. We introduce the normal linear model and the briefly review the behavior of classical and Bayesian "truth-testing" procedures for later comparisons before turning to our preferred Bayesian methods. Our purpose is constructive and is designed to lead to theoretically based but useful procedures, rather than retrospective, that is, to see if we can find some explanation for old procedures derived from other perspectives.

3. Prediction

3.1. The Distributions

To obtain a context in which one can judge the suitability of a decision model, we specify the usual normal linear sampling model in which the response variable y has a normal distribution conditional on the concomitant r-dimensional vector variable $\mathbf{x}$ with linear mean $\mathbf{x}^T\boldsymbol{\beta}$ and variance σ^2. Given a sequence of values for $\mathbf{x}$, conditional independence is assumed under the sampling model for the corresponding values y. We shall consider the family of natural conjugate prior distributions for $\boldsymbol{\beta}$ and σ and make direct use of the posterior predictive distribution for future y given $\mathbf{x}$. This distribution is the mixture of sampling distributions for y given $\mathbf{x}$, mixed over the posterior distribution of $\boldsymbol{\beta}$ and σ. It models one's coherent postdata personal uncertainty about future y taking account of the fact that $\boldsymbol{\beta}$ and σ are not known. See Raiffa & Schlaifer (1961) for details.

For much of our development the distribution of future $\mathbf{x}$ is arbitrary. But special interest attaches to the case when the vectors $\mathbf{x}_i$ are "random," more precisely are independently sampled from a multivariate normal distribution, say with unknown mean and variance ξ, and Σ. Given that the prior uncertainty is independent between (ξ, Σ) and (β, σ), then by an obvious factorization of the likelihood function, so is the posterior uncertainty, and future $\mathbf{x}$ will be posterior independent of (β, σ). Geisser (1965), under "ignorance" priors on (ξ, Σ), and Ando and Kaufman (1965), under conjugate priors, obtain multivariate-t predictive distributions for $\mathbf{x}$ with moments approximately matching the empirical moments of observed $\mathbf{x}_1, \ldots, \mathbf{x}_n$

$$E(\mathbf{x}|D) = \bar{\mathbf{x}}, \qquad V(\mathbf{x}|D) = n^{-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T. \tag{3.1}$$

We shall use (3.1) as an approximation.

3.2. Truth-Test Criteria

A variety of rather arbitrary criteria for selecting a linear model appear in practice, such as the modified multiple correlation statistic $\bar{R}^2 = 1 - s^2/\{\sum(y_i - \bar{y})^2/(n-1)\}$, where vs^2 is the sample error sum of squares, n is the sample size, and $v = n - r$, where r is the number of concomitant variables included in the regression. This is typically preferred to the unmodified statistic, $R^2 = 1 - vs^2/\sum(y_i - \bar{y})^2$, merely because the latter is monotonically increasing for any nested sequence of models, whereas $\bar{R}^2$ may sometimes decrease when a new variable is added. On the other hand, the modified statistic need not lie in the unit interval $[0, 1]$. (See Gaver & Geisel, 1973, for a sense in which R^2 is a Bayesian criterion for choice when comparing two models of equal dimensionality.)

Some criteria are intended to answer the question, Is the model true? The traditional F test asks this question of the smaller model H in a nested pair of models $H \subset \Omega$. Denote by $r, \mathbf{x}, \beta$, and σ the parameters in the larger, unquestioned model Ω where r is the dimensionality of β. Then if H is defined by linear constraints on β, we can, without loss of generality, define the variables in such a way that H takes the nested discarded-variables form

$$H: \quad \beta_K = \mathbf{0}, \tag{3.2a}$$

where

$$\mathbf{x} = (\mathbf{x}_H^{\mathrm{T}}, \mathbf{x}_K^{\mathrm{T}})^{\mathrm{T}}, \qquad \beta = (\beta_H^{\mathrm{T}}, \beta_K^{\mathrm{T}})^{\mathrm{T}}. \tag{3.2b}$$

The parameters of H as a normal linear model will then be denoted by r_H, $\mathbf{x}_H, \beta_H, \sigma_H \equiv \sigma$.

Tests for inference on the hypothesis H are available based on a sample of n realizations of $(y, \mathbf{x})$. For example, the traditional F- test is based on the

error sums of squares (SSE) for the two models, SSE(H) and SSE(Ω), and the sum of squares between (SSB),

$$\text{SSB}(H,\Omega) = \text{SSE}(H) - \text{SSE}(\Omega). \tag{3.3}$$

The test rejects H if

$$\text{SSB}(H,\Omega) > (r - r_H)s^2 F_{r-r_H,\nu}(.95), \tag{3.4}$$

where $s^2 = \text{SSE}(\Omega)/\nu$, the usual unbiased estimate of σ^2. Note that this threshold on the between sum of squares is approximately constant in the sample size n for large n, as was noted by Lindley (1968) in the σ^2-known case.

Under this same nesting, the odds-ratio test for "truth" has been developed in Dickey (1971). Consider a mixed-type prior distribution with positive probabilities $P(H), P(H^c)$, and conditional prior densities $P(\boldsymbol{\beta}, \sigma | H^c)$ and $P(\boldsymbol{\beta}_H, \sigma | H)$ of the natural conjugate form.

Under the natural assumption of the *continuity condition*, that the conditional prior distributions $P(\boldsymbol{\beta}_H, \sigma | \boldsymbol{\beta}_K)$ are continuous-in-distribution in $\boldsymbol{\beta}_K$ at $\boldsymbol{\beta}_K = \mathbf{0}$, one obtains in the Bayes factor $B_D(H)$ a useful approximation to the Bayes factor for a more realistic hypothesis in the form of a neighborhood set closely surrounding H. The neighborhood Bayes factor would result from various integrable continuous-type distributions with a mound of high density over the neighbourhood set. Of course, the quality of the approximation depends on the data outcome D. (As always, the conditional distribution $P(\boldsymbol{\beta}_H, \sigma | H)$, and hence also the Bayes factor, is not unique to the conditioning event H, but depends on the choice of conditioning parameter $\boldsymbol{\beta}_K$ (Gunel & Dickey, 1974).)

The resulting approximate Bayes factor $B_D(H)$ is proportional to the *density ordinate* of the usual F statistic (as opposed to the traditional tail area). The decision threshold on the between sum of squares, following from a fixed threshold on the Bayes factor, is proportional to the logarithm of sample size in the known-variance case

$$\text{SSB}(H,\Omega) > C \log(n). \tag{3.5}$$

Contrast this with the threshold in (3.4). Even for small sample sizes, the tendency is for a Bayes factor to require more extreme data for rejection than the tail area test requires (Dickey, 1977).

For large sample sizes, either criterion, (3.4) or (3.5), is nearly certain to reject H in statistical practice, since H is nearly always known not to hold exactly, and the "power" of either test increases to unity in n. (This analysis for the Bayes factor assumes, of course, that the full model $\Omega = H \cup H^c$ holds exactly. However, we believe that the same phenomenon occurs in greater generality.) Of course, this phenomenon could be adjusted for by arbitrarily changing the level of the test (3.4) with sample size n or some similar *ad hoc*

response in the case of the Bayes factor (3.5). Rather than pursue such a line, we develop below a decision threshold on SSB(H, Ω) that is proportional to the sample size n—whereby it will no longer hold that one nearly always decides against simplicity when n is large.

3.3. Predictors and Utility

We now neglect the concept of a smaller sampling model H and the artificial question of the "truth" of H. As a subset of the parameter space, H will receive no special positive probability. The family of natural conjugate prior distributions will apply to β, σ over Ω.

We begin with a continuous action space and derive a discreteaction space in a natural way. Consider the problem of predicting a future value of y from a future concomitant value $\mathbf{x}$. That is, the decision maker must choose after sampling, based on the information he has gained, a *predictor*, or predictor function, $\hat{y}(\mathbf{x})$. He is assumed then to receive a utility W depending on his choice and on the future outcome. Typically, W is large for $\hat{y}(\mathbf{x})$ near y.

Since the decision maker must choose a whole function $\hat{y}(\cdot)$, it seems natural that the utility W should depend on the chosen function, as well as on the realized values $\hat{y}(\mathbf{x})$, y. W might, for example, impose penalties for choosing a complicated predictor function or a predictor depending on coordinate variables in $\mathbf{x}$ which are costly to observe. As an illustration, consider the utility

$$W\{\hat{y}(\mathbf{x}), y, \hat{y}(\cdot)\} = \begin{cases} c - [\hat{y} - y]^2 & \text{if } \hat{y}(\cdot) \text{ is constant in } \mathbf{x}, \\ -[\hat{y}(\mathbf{x}) - y]^2 & \text{otherwise.} \end{cases} \tag{3.6}$$

The utility function (3.6) offers a reward c if you are willing to have the same prediction $\hat{y}$ regardless of $\mathbf{x}$. This means that $\mathbf{x}$ need not be measured for the prediction. Thus agreeing in advance to this restriction save some effort, whose cost is c. Intuitively, if $\mathbf{x}$ does not vary much anyway, and/or if the connection between y and $\mathbf{x}$ is weak, and/or if c is large, the simpler predictor would be the more favored.

We shall assume in general that there are classes of predictor functions defined such that within each class I, W is constant in $\hat{y}(\cdot) \in I$ for given values $\hat{y}(\mathbf{x})$ and y. In the example shown in (3.6), there are two such classes: I_1 consisting of all constants $\hat{y}$ and I_2 of the general functions $\hat{y}(\mathbf{x})$. It will be natural in such predictor-choice problems to identify the derived problem of how to choose between classes I.

Thus we seek to compute

$$\max_I \max_{\hat{y}_I(\cdot)} \iint W(\hat{y}_I(\mathbf{x}), y, y_I(\cdot)) p(y|\mathbf{x})\, dy\, p(\mathbf{x})\, d\mathbf{x}. \tag{3.7}$$

This formula looks more formidable than it really is. The maximization with respect to I and $\hat{y}_I(\cdot)$ occurs after the integration with respect to y and $\mathbf{x}$ because they are chosen before y and $\mathbf{x}$ are known.

According to the principle of maximal expected utility, a single posterior-predictive distribution based on a single parametrized sampling model should be used to calculate the expected utilities of all the competing predictors. Strictly speaking, we are no longer concerned with deciding between probability models, but rather with choosing a class of predictors I.

Since the class of utilities of particular interest to us are those for which the only feature of $\hat{y}_I(\cdot)$ to affect W is its membership in the class I, we will write, with only slight abuse of notation, $W(\hat{y}_I(\mathbf{x}), y, I)$ instead of $W(\hat{y}(\mathbf{x}), y, \hat{y}(\cdot))$.

We now specialize our utility function to the form

$$W(\hat{y}, y, I) = U_I w(\hat{y}, y) + U_I^*, \tag{3.8}$$

where $w(\hat{y}, y)$ measures the utility of predicting $\hat{y}$ when y turns out to be the realization of the process and U_I and U_I^* are constants depending on I that reflect, for example, the cost of measuring and computing from the variables permitted by the class I.

To distinguish $w(\hat{y}, y)$ from $W(\hat{y}, y, I)$, we call the former a purely predictive utility component since it is, strictly speaking, not a utility function, but rather one aspect of the utility function (3.8). The illustration (3.6) is a special case of (3.8) in which

$$U_{I_1} = U_{I_2} = 1, \qquad U_{I_1}^* = c, \qquad U_{I_2}^* = 0, \qquad \text{and} \qquad w(\hat{y}, y) = -(\hat{y} - y)^2.$$

The form (3.8) is especially convenient because the associated expected utility preserves the linearity of W in w as

$$\bar{W}(I) = U_I \bar{w}(I) + U_I^*, \tag{3.9}$$

where

$$\bar{w}(I) = \max_{\hat{y}_I(\cdot)} \iint w(\hat{y}_I(\mathbf{x}), y) p(y|\mathbf{x})\, dy\, p(\mathbf{x})\, d\mathbf{x}$$

and

$$\bar{W}(I) = \max_{\hat{y}_I(\cdot)} \iint W(\hat{y}_I(\mathbf{x}), y, I) p(y|\mathbf{x})\, dy\, p(\mathbf{x})\, d\mathbf{x}.$$

Hence the optimal choice between two classes of predictors I_1 and I_2 will proceed by a comparison between the expected purely predictive performance of the two models ($\bar{w}(I_1)$ and $\bar{w}(I_2)$) to the multiplicative (U_{I_1} and U_{I_2}) and/or additive ($U_{I_1}^*$ and $U_{I_2}^*$) rewards for simplicity, as summarized by (3.9).

Two important special cases suggest themselves. If $U_{I_1} = U_{I_2} = U$, so that the rewards for simplicity are purely additive, maximum expected

utility reduces to the criterion: prefer I_1 over I_2 if and only if

$$\bar{w}(I_1) - \bar{w}(I_2) \geq (U^*_{I_1} - U^*_{I_2})/U. \tag{3.10}$$

Similarly, if $U^*_{I_1} = U^*_{I_2}$, so that the rewards for simplicity are purely multiplicative, maximum expected utility reduces to: prefer I_1 over I_2 if and only if

$$\bar{w}(I_1)/\bar{w}(I_2) \geq U_{I_2}/U_{I_1}. \tag{3.11}$$

The predictor classes considered here are of the discarded-variable type. The class of all predictor functions is restricted to dependence on only a specified subvector $\mathbf{x}_I$,

$$\hat{y}(\mathbf{x}) = \hat{y}_I(\mathbf{x}_I), \qquad \mathbf{x} = (\mathbf{x}_I^{\mathrm{T}}, \mathbf{x}_K^{\mathrm{T}})^{\mathrm{T}}, \tag{3.12}$$

where for ease of notation $\mathbf{x}_I$ consists of the first few coordinates of $\mathbf{x}$. The subscripts I and K are used analogously to H and K in (3.2). Note that if a specific independent variable, say x_q, is included in $\mathbf{x}_I$, then all functions of that variable are contained in the class I of available predictor functions. For example, higher powers of x_q cannot be excluded without excluding x_q itself. Another limitation of this form is that one cannot consider whether or not to have a zero intercept as a choice between classes I. An advantage of this form is that if one considers that the main cost of having x_q included in the set of variables on the basis of which a prediction $\hat{y}$ is to be made is the cost of measuring x_q, then it is quite reasonable to allow the most flexible use of x_q for prediction.

An alternative class of predictors not pursued here is the class of constrained linear predictors. This is the class of all linear forms satisfying a specified linear constraint on the coefficients, or without loss of generality, the class of all linear forms in a specified subvector $\mathbf{x}_I$,

$$\hat{y}(\mathbf{x}) = \mathbf{x}_I^{\mathrm{T}} \hat{\boldsymbol{\beta}}_I. \tag{3.13}$$

Constrained linear predictors are studied in papers of Goldstein (1975a,b, 1976).

Having stated that we will study the discarded-variables class of predictors under the general form of utility (3.8), it remains only to specify the purely predictive utility component in order to have a closed-form decision problem ready for analysis. In this paper we report results for two such choices: the usual negative squared error

$$w(\hat{y}, y) = -(\hat{y} - y)^2, \tag{3.14a}$$

which is taken up in Section 3.4, and a less familiar function is probability density form

$$w(\hat{y}, y) = f_\alpha(\hat{y} - y), \tag{3.14b}$$

where f_α is the probability density of a random variable α as developed in Section 3.5.

3.4. Negative Squared Error as Purely Predictive Utility Component

Having outlined a decision model for consideration, we now summarize the resulting decision criterion in the discarded-variable case for comparison to other Bayesian procedures and to classical procedures. The derivations are not dealt with at great length since they are given in a line of studies including Anscombe (1963), Dickey (1967), Lindley (1968), Harrison & Stevens (1976) and Dickey & Kadane (1977).

We use the negative squared error form for the purely predictive utility (3.8), (3.14a), and the optimization procedure outlined in (3.7). Lindley (1968) derived the corresponding criterion in the case of known σ^2. Starting with the expectations over y and $\mathbf{x}$, we have, given I and $\hat{y}_I(\cdot)$,

$$\begin{aligned} E[\{\tilde{y} - y_I(\tilde{\mathbf{x}})\}^2] &= E[E[\{\tilde{y} - E(y|\mathbf{x}_I) + E(y|x_I) - \hat{y}_I(\mathbf{x}_I)\}^2|\tilde{\mathbf{x}}_I]] \\ &= E[E[\{y - E(y|\mathbf{x}_I)\}^2|\mathbf{x}_I]] \\ &\quad + E[E[\{E(y|\mathbf{x}_I) - y_I(\mathbf{x}_I)\}^2|\mathbf{x}_I]]. \end{aligned} \tag{3.15a}$$

This expectation is minimized by the choice

$$\hat{y}_I(\mathbf{x}_I) = E(y|\mathbf{x}_I), \tag{3.16}$$

which makes the second summand zero. Using this choice and resuming the calculation in (3.15a), we have

$$\begin{aligned} -w(I) &= E[\mathrm{Var}(y|\mathbf{x}_I)] \\ &= \mathrm{Var}\ y - \mathrm{Var}[E(y|\mathbf{x}_I)] \\ &= E(\sigma^2) + \mathrm{Var}(\tilde{\mathbf{x}}^{\mathrm{T}}\tilde{\boldsymbol{\beta}}) - \mathrm{Var}[E(\boldsymbol{\beta})^{\mathrm{T}}E(\mathbf{x}|\mathbf{x}_I)] \\ &= E(\sigma^2) + E(\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathrm{T}}\tilde{\boldsymbol{\beta}}) - E(\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\tilde{\mathbf{x}})E(\tilde{\mathbf{x}}^{\mathrm{T}}\tilde{\boldsymbol{\beta}}) - \mathrm{Var}[E(\boldsymbol{\beta})^{\mathrm{T}}E(\mathbf{x}|\mathbf{x}_I)] \\ &= E(\sigma^2) + \mathrm{tr}[E(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathrm{T}})E(\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^{\mathrm{T}})] - E(\tilde{\boldsymbol{\beta}})^{\mathrm{T}}E(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathrm{T}})E(\boldsymbol{\beta}) \\ &\quad + E(\tilde{\boldsymbol{\beta}})^{\mathrm{T}}E(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathrm{T}})E(\boldsymbol{\beta}) - E(\tilde{\boldsymbol{\beta}})^{\mathrm{T}}E(\tilde{\mathbf{x}})E(\tilde{\mathbf{x}}^{\mathrm{T}})E(\tilde{\boldsymbol{\beta}}) - \mathrm{Var}(E(\boldsymbol{\beta})^{\mathrm{T}}E(\mathbf{x}|\mathbf{x}_I)] \\ &= E(\sigma^2) + \mathrm{tr}[E(\mathbf{x}\mathbf{x}^{\mathrm{T}})\ \mathrm{Var}(\boldsymbol{\beta})] + E(\boldsymbol{\beta})^{\mathrm{T}}\ \mathrm{Var}(\mathbf{x})E(\boldsymbol{\beta}) \\ &\quad - E(\boldsymbol{\beta})^{\mathrm{T}}\ \mathrm{Var}[E(\mathbf{x}|\mathbf{x}_I)]E(\boldsymbol{\beta}) \\ &= E(\sigma^2) + \mathrm{tr}[E(\mathbf{x}\mathbf{x}^{\mathrm{T}})\ \mathrm{Var}(\boldsymbol{\beta})] + E(\boldsymbol{\beta})^{\mathrm{T}}E\{\mathrm{Var}[\mathbf{x}|\mathbf{x}_I]\}E(\boldsymbol{\beta}), \end{aligned} \tag{3.15b}$$

which is equivalent to Lindley's (1968) Eq. (10).

To simplify the discussion of these findings, we specify the popular "ignorance" prior density within the natural conjugate family

$$p(\beta, \sigma) = c/\sigma. \tag{3.17}$$

The results for this case also hold as approximations in large samples for the more general case (Dickey, 1976).

For sample data $D = (\mathbf{y}, X)$, where $\mathbf{y} = (y_1, \ldots, y_n)$ and $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}}$, denote the usual sufficient statistics by

$$\begin{aligned} N &= X^{\mathrm{T}}X, \qquad & \mathbf{b} &= N^{-1}X^{\mathrm{T}}\mathbf{y}, \\ \nu &= n - r, \qquad & s^2 &= \nu^{-1}(\mathbf{y} - X\mathbf{b})^{\mathrm{T}}(\mathbf{y} - X\mathbf{b}) = \nu^{-1}\mathrm{SSE}(\Omega), \end{aligned} \tag{3.18}$$

where r is the dimension of $\mathbf{x}$, that is, the number of independent variables in the regression model Ω, and ν is the degrees of freedom. Then the resulting posterior distribution is given by

$$(\boldsymbol{\beta} \mid \sigma, D) \sim N^{(r)}(\mathbf{b}, \sigma^2 N^{-1}), \tag{3.19a}$$

an r-dimensional normal distribution with mean $\mathbf{b}$ and variance matrix $\sigma^2 N^{-1}$, and by

$$(\sigma^2 \mid D) \sim s^2/(\chi_\nu^2/\nu), \tag{3.19b}$$

s^2 divided by a chi-square random variable with ν degrees of freedom, divided by ν. The optimal predictor in class I (3.15) is then

$$\hat{y}_I(\mathbf{x}_I) = \mathbf{x}_I^{\mathrm{T}}\mathbf{c}_I, \tag{3.20}$$

where $\mathbf{c}_I$ denotes the usual subset-regression least-squares coefficient vector based on $D_I = (\mathbf{y}, X_I)$ where the matrix $X = (X_I, X_K)$ is partitioned according to $\mathbf{x}^{\mathrm{T}} = (\mathbf{x}_I^{\mathrm{T}}, \mathbf{x}_K^{\mathrm{T}})^{\mathrm{T}}$. The corresponding predictive utility (3.15b) becomes

$$-\bar{w}(I) = (1 + r/n)s^2\nu/(\nu - 2) + \mathrm{SSB}(I, \Omega)/n. \tag{3.21}$$

Note that in both (3.15b) and (3.21) one term alone depends on the class I. In the case of additive rewards ($U_I \equiv 1$), one prefers class I_2 over I_1 if and only if $n(U_{I_1}^* - U_{I_2}^*)$ is exceeded by the difference

$$\mathrm{SSB}(I_1, \Omega) - \mathrm{SSB}(I_2, \Omega). \tag{3.22}$$

This difference will itself be a between sum of squares $\mathrm{SSB}(I_1, I_2)$ in the nested case $I_1 \subset I_2$. Hence as stated previously, one obtains a threshold on SSB proportional to sample size, in distinction to the previous thresholds (3.4) and (3.5).

Note the lack of dependence of the criterion on any feature of the model Ω other than the property that the predictor variables of I_1 and I_2 are both subsets of those of some model to which a prior distribution in the "ignorance" form applies. In particular, this lack of dependence holds whether or not I_1 and I_2 are nested.

The first term of (3.21), $(1 + r/n)s^2\nu/(\nu - 2)$, is the Bayesian posterior predictive squared error of the optimal predictor $\hat{y}_\Omega(\mathbf{x})$, based on the full

set of r variables. This term refers literally to the context in which inference and prediction are *both* made from the full model Ω. Thus this term is the "cost of doing business," the cost we must pay for the prediction regardless of other costs. If we infer from the full model to obtain the posterior expected squared error for optimal prediction by a smaller class I, there is then the additional term $\mathrm{SSB}(I,\Omega)/n$. Hence a statistician obeying the principle of maximum expected utility and evaluating predictor classes from the standpoint of the full inference model would prefer the larger predictor class of any nested pair if no special rewards are given for simplicity (that is, if $U_I \equiv U$ and $U_I^* = U^*$). We shall return to this aspect in Section 4.

For the next few paragraphs we discuss expected squared errors calculated from a sampling theory viewpoint. Lest this change in point of view cause confusion, we pause here to remind the reader about the distinction. In Bayesian theory, the random variables are what is unknown, namely the values of parameters. The data, once observed, are no longer random (we know the outcome), and hence calculations like (2.1) are done conditional on the observed data outcomes. By contrast, in traditional sampling theory, the random variables are the "data that might have been," distributed conditionally on the values of the parameters. Each side in the discussion accepts the mathematics of the other side; the discussion is rather about which computations are relevant for inference.

Taking expectations now under the *sampling model*, holding the parameters fixed, we find

$$E(y - \hat{y}_I)^2 = (1 + r_I/n)\sigma_I^2, \tag{3.23}$$

where σ_I^2 is the sampling-model conditional variance of y given $\mathbf{x}_I$. This expectation is commonly estimated by the statistic,

$$(1 + r_I/n)\mathrm{SSE}(I)/v_I. \tag{3.24}$$

Traditionally, one compares values of (3.23) for various subsets I. Such a criterion is less than appropriate for several reasons, including that the expectation is not conditional on the observed data D. Instead, it is an average over values of the coefficient estimate $\mathbf{c}_I$ which did *not* occur, and therefore violates the likelihood principle, which is a consequence of the Bayesian axioms.

A statistician who uses the estimate (3.23) ignores the data X_K. The similarity in form of (3.24) to the first term of (3.21) yields an interpretation of the traditional statistic (3.24) as an approximate posterior predictive squared error of $\hat{y}_I$ for a Bayesian statistician whose prior uncertainty implies conditional independence between $(y, \mathbf{x}_I)$ and X_K given $\mathbf{y}$ and X_I. To take this position for various sets $\mathbf{x}_I$ would be to place very special

conditions on the joint distribution of future data $(y, \mathbf{x})$ and known data $D = (\mathbf{y}, X_I, X_K)$.

It can be shown that a Bayesian who believes $H: \boldsymbol{\beta}_K = \mathbf{0}$ but otherwise has prior "ignorance" would have the same predictor as someone who was "ignorant" within Ω but was forced to discard $\mathbf{x}_K$, namely,

$$\begin{aligned} \hat{y}(\mathbf{x}) &= E(\mathbf{y}|\mathbf{x}, \boldsymbol{\beta}_K = \mathbf{0}, D) = \mathbf{x}_I^{\mathrm{T}} E(\boldsymbol{\beta}_I | \boldsymbol{\beta}_K = \mathbf{0}, \mathrm{D}) \\ &= \mathbf{x}_I^{\mathrm{T}} \mathbf{c}_I, = \hat{y}_I(\mathbf{x}), \end{aligned} \tag{3.25}$$

using (3.20). The person believing H would interpret (3.24) as his optimal posterior predictive squared error (as opposed to (3.21) for the believer in Ω). To compare (3.24) for two difference choices I_1 and I_2 under this interpretation would be to first state a prior opinion that $\boldsymbol{\beta}_{J_1} = 0$ with "ignorance" on the remaining coefficients $\boldsymbol{\beta}_{I_1}$ and then an analogous prior with respect to I_2. To hold both opinions simultaneously would be a contradiction.

3.5. Probability Density Form of Purely Predictive Utility Component

We now consider the second kind of purely predictive utility component, that in the form of a probability density function (pdf) of f_α $(\hat{y} - y)$; see (3.14). It will help the reader's intuition to think of the density f_α of the random variable α as symmetric and unimodal with mode at zero. Thus we most prefer (have highest utility for) a zero predictive error $(\hat{y} - y)$. Our preference decreases as the error increases in absolute value, and we are equally displeased by an over-prediction ($\hat{y} - y$ positive) and an under-prediction ($\hat{y} - y$ negative) of equal magnitude. Such a utility function has the advantage of being bounded, and hence stable (Kadane & Chuang, 1978); by comparison, negative square error has neither property. Tiao and Afonja (1976) discuss the special case in which f is the normal pdf; some of these densities f_α are in the class of conjugate utilities proposed by Lindley (1976). Our analysis follows Dickey & Kadane (1977).

Suppose now that α has two moments, so that we may write

$$E(\alpha) = 0, \qquad \mathrm{Var}(\alpha) = a^2. \tag{3.26}$$

The variable a has an interesting interpretation. If "a miss is as good as a mile," that is, if any but the most accurate estimates are considered nearly equally worthless, then a would be taken to be quite small. On the other hand, if the idea is to be roughly correct, then a could be taken to be larger. Thus a measures an important aspect of an investigator's utility function, one that might change from one occasion to the next, even with the same

data and the same formal decision problem, depending on the investigator's purpose.

Conditional on I, the expected purely predictive utility component achieved is

$$\bar{w}(I) = \max_{\hat{y}_I(\cdot)} \iint f_\alpha(\hat{y}_I(\mathbf{x}) - y)p(y|\mathbf{x}_I)\,dy\,p(\mathbf{x}_I)\,dx_I, \tag{3.27}$$

where we have already used the fact that $\hat{y}_I(\mathbf{x})$ depends on $\mathbf{x}$ only through $\mathbf{x}_I$. We notice that the inner integral is a familiar form. In fact, it is the formula for the convolution of two random variables α and y, with the resulting density evaluated at $\hat{y}_I(\mathbf{x})$. Thus we may write

$$\int f_\alpha(\hat{y}_I(\mathbf{x}) - y)p(y|\mathbf{x}_I)\,dy = f_{\alpha+y|\mathbf{x}_I}(\hat{y}_I(\mathbf{x}_I)). \tag{3.28}$$

Substituting this expression in (3.27), we have

$$\bar{w}(I) = \max_{\hat{y}_I(\cdot)} \int f_{\alpha+y|\mathbf{x}_I}(\hat{y}_I(\mathbf{x}_I)p(\mathbf{x}_I)\,d\mathbf{x}_I. \tag{3.29}$$

Since the class of predictors $\hat{y}_I(\cdot)$ are all functions of $\mathbf{x}_I$, the maximization in (3.29) can be performed pointwise in $\mathbf{x}_I$ before integration. Hence

$$\bar{w}(I) = \int \left[\max_{\hat{y}_I(\mathbf{x}_I)} f_{\alpha+y|\mathbf{x}_I}(\hat{y}_I(\mathbf{x}_I))\right] p(\mathbf{x}_I)\,d\mathbf{x}_I. \tag{3.30}$$

Now α is unimodal and symmetric, with center of symmetry 0. Similarly, conditionally on $\mathbf{x}_I$, $y - E(y|\mathbf{x}_I)$ is assumed unimodal and symmetric with center of symmetry 0. Since these two random variables are independent, their convolution is also unimodal and symmetric with mode at 0 (Wintner, 1938, p. 30). Because the mode of this convolution is at 0, the mode of $f_{\alpha+y|\mathbf{x}_I}$ is at $E(y|\mathbf{x}_I)$, and hence the best choice of $\hat{y}_I(\mathbf{x}_I)$ is

$$\hat{y}_I(\mathbf{x}_I) = E(y|\mathbf{x}_I). \tag{3.31}$$

The resulting purely predictive utility component is

$$\bar{w}(I) = \int f^{(0)}_{\alpha+y|\mathbf{x}_I - E(y|\mathbf{x}_I)} p(\mathbf{x}_I)\,d\mathbf{x}_I. \tag{3.32}$$

This is in the form of a mixture of densities (here evaluated at zero). The corresponding mixture random variable has zero mean and variance V, where

$$V = a^2 + E[\mathrm{Var}(y|\mathbf{x}_I)]. \tag{3.33}$$

Suppose g is the density of some location-scale family (for specificity, one can think of g as φ, the standardized normal density). Then we assume the approximation

$$\int f^{(z)}_{\alpha+y|\mathbf{x}_I - E(y|\mathbf{x}_I)} p(\mathbf{x}_I)\,d\mathbf{x}_I \doteq g(z/V^{1/2})/V^{1/2}, \tag{3.34}$$

so that, in particular,

$$\int f^{(0)}_{a+y|\mathbf{x}_I - E(y|\mathbf{x}_I)} p(\mathbf{x}_I)\, d\mathbf{x}_I \doteq g(0)/V^{1/2}. \tag{3.35}$$

Hence we have the approximation

$$\bar{w}(I) \doteq g(0)/\{a^2 + E[\mathrm{Var}(y|x_I)]\}^{1/2} \tag{3.36}$$

For a normal approximation, $g(0) = (2\pi)^{-1/2}$. Note that $E[\mathrm{Var}(y|\mathbf{x}_I)]$ is evaluated in (3.15b). In the "ignorance" prior case, by (3.21),

$$V = (1 + r/n)s^2 v/(v - 2) + \mathbf{SSB}(I, \Omega)/n. \tag{3.37}$$

It is interesting to compare the decision criteria that result from (3.36) with those we have seen in (3.16) and (3.21) in the squared error loss case. In the case that all "rewards" are multiplicative ($U_I^* \equiv U$), the criterion for two nested predictor models, $I_1 \subset I_2$, prefers I_2 for

$$\mathbf{SSB}(I_1, I_2) > n[a^2 + V][(U_1/U_2)^2 - 1].$$

Compare this with the criterion (3.22) which also has the form of a threshold on the between sum of squares. Again, note the proportionality of the threshold to sample size.

4. Discussion

We have seen that the Bayesian odds ratio is not useful as a complete data summary for decision making. Indeed, the very concept of inference about the "truth" of a model has come under question. Also, in the specific decision problem of choice of predictor in a normal linear model, we have found that the predictive criterion behaved as a function of sample size in such a fashion that even if the F test tail area and the Bayesian odds for a submodel were very small (3.4), (3.5), the optimal predictor could still be the predictor based only on the submodel, (3.22), (3.36). In other words, one can have coherent preference for the use of a simple model even if it is known to be "false."

According to Occam's razor, one should prefer scientific models that are simple. In our decision criteria for prediction it has turned out that positive rewards for simplicity are necessary to imply preferences for simple nested predictor classes. For our utilities $W(\hat{y}, y, I)$ (3.8), the obvious monotonicity of the purely predictive expected utility component

$$I_1 \subset I_2 \quad \Rightarrow \quad \bar{w}(I_1) \leq \bar{w}(I_2),$$

implies that positive rewards, $(U_{I_1}^* - U_{I_2}^*)/U$ (3.10) or $\log(U_{I_1}/U_{I_2})$ (3.11), will

be needed. This result is quite general since to maximize a functional $\bar{w}[\hat{y}(\cdot)]$ over a larger class of functions $\hat{y}(\cdot)$ can never lead to a smaller maximum value $\bar{w}$. Hence, to the extent that real decision makers are coherent, the fact that simple predictor models are used could mean that positive rewards do exist.

The principle is clear that one should not deliberately exclude a cost-free variable from influence on one's predictor, provided that one maintains the freedom to use it in an optimal way. It could turn out, though, that the optimal way itself implies a predictor which is a function of the other variables only. In our treatment of the normal linear model, this happens with probability zero for a prior distribution of continuous type. However, a mixed-type prior distribution could lead to such an event with nontrivial probability as follows.

Using a probability density form of predictive utility w, one may consider linear predictors based on a parameter estimate $\hat{\boldsymbol{\beta}}$. If the probability is positive for a submodel of the form $H: \boldsymbol{\beta}_K = \mathbf{0}$, then the predictive expected utility $\bar{w}(\hat{\boldsymbol{\beta}}) = E(w|H)P(H) + E(w|H^c)P(H^c)$ will typically have two local maxima in $\hat{\boldsymbol{\beta}}$, one for which have $\hat{\boldsymbol{\beta}}_J = \mathbf{0}$ and the other with $\hat{\boldsymbol{\beta}}_J \neq \mathbf{0}$. A high value of $P(H)$ will tend to imply that the overall maximum is achieved at the first of these points. Hence, even without positive rewards for simplicity, $\bar{W}$ will be maximized by the simpler predictor. (For details, see Dickey & Kadane, 1977.) In our view, however, mixed-type priors that put positive probability on low dimensional subspaces are not usually realistic in the context of most social sciences. Consequently, positive rewards for simplicity seem to be required to explain why good economists work with simple models.

Common sense and statistical folklore warn against using too many variables in a regression relative to the sample size. Least squares estimates are notoriously prone to outliers and other problems and become even more suspect in high dimensions. For example, the traditionally estimated expected squared error of prediction (3.24) increases for large values of r_I/n. Yet the Bayesian methods developed in Section 3 require the use of all available cost-free variables. This phenomenon has been called by Lindley (1978) the paradox of Bayesian overfitting.

We offer two explanations of this difficulty. The first, also cited by Dickey (1973b) and by Lindley (1978) in his discussion of Young (1977), is that the model chosen for illustration in Section 3 takes as its prior the ignorance form (3.17), which is justified as an approximation to a proper subjective prior distribution by the theory of stable estimation (Edwards *et al.*, 1963). This approximation yields the least squares estimates. It will be valid when sample size n is large relative to the dimensionality of the regressors (r_I), so the approximation tends not to be valid precisely where the problem occurs (r_I large relative to n).

Our second explanation, a basic premise of this paper, is that previous efforts in this area have not taken explicit account of the utility function of the economist. Just as in the past many efforts have been made to avoid proper subjective prior distributions on the parameters, many economists and statisticians try to avoid the statement of an explicit utility function for the problem of model choice. In problems of estimation this is not such a severe difficulty because, if the posterior distribution is symmetric, every symmetric utility function having its maximum at zero will result in the same estimate, namely the posterior center of symmetry. Thus it is not necessary to think about whether your utility is negative squared error, negative absolute error, or normal pdf form since they all result in the same estimate. However, as we have seen, the choice of models problem is essentially different in this respect. For choosing models, there seems to be no alternative to an honest assessment of your utility function and, in particular, of your personal weighting of accuracy against parsimony.

ACKNOWLEDGMENTS

The authors are pleased to acknowledge the financial support of the National Science Foundation and the useful and helpful comments of Jan Kmenta, James B. Ramsey, and Michel Mouchart. Reproduction for any purpose of the U.S. Government is permitted.

REFERENCES

Ando, A., & Kaufman, G. W. Bayesian analysis of the independent multinormal process —neither mean nor precision known. *Journal of the American Statistical Association*, 1965, **60**, 347–358.

Anscombe, F. J. Bayesian inference concerning many parameters with reference to super-saturated designs. *Bulletin of the International Statistical Institute*, 1963, **40**, 721–733.

Berkson, J. Some difficulties of interpretation encountered in the application of the chi-squared test. *Journal of the American Statistical Association*, 1938, **33**, 526–542.

Davis, W. W., & DeGroot, M. H. A new look at selecting regression models. Unpublished technical report, Carnegie-Mellon University, 1978.

Dickey, J. M. A Bayesian hypothesis-testing procedure. *Annals of the Institute of Statistical Mathematics*, 1967, **19**, 367–369.

Dickey, J. M. The weighted likelihood ratio, linear hypotheses on normal location parameters. *Annals of Mathematical Statistics*, 1971, **42**, 204–223.

Dickey, J. M. Scientific reporting and personal probabilities: Student's hypothesis. *Journal of the Royal Statistical Society Series B*, 1973, **35**(2), 285–305.(a)

Dickey, J. M. Discussion to "Combining possibly related estimation problems" by B. Efron and C. Morris. *Journal of the Royal Statistical Society, Series B*, 1973, **35**(3), 379–421.(b)

Dickey, J. M. Bayesian alternatives to the F test and the least squares estimate in the normal linear model. In Stephen E. Fienberg & Arnold Zellner (Eds.), *Studies in Bayesian econometrics and statistics* (dedicated to Leonard J. Savage). Amsterdam: North-Holland Publ., 1974. Pp. 515–554.

Dickey, J. M. Approximate posterior distributions. *Journal of the American Statistical Association*, 1976, **71**(September), 680–689.

Dickey, J. M. Is the tail area useful as an approximate Bayes factor? *Journal of the American Statistical Association*, 1977, **72**(March), 138–142.

Dickey, J. M. Approximate coherence for regression model inference-with a new analysis of Fisher's broadback wheatfield data. In A. Zellner (Ed.) *Studies in Bayesian econometrics and statistics* (in honor of Sir Harold Jeffreys). Amsterdam: North-Holland Publ., 1979, in press.

Dickey, J. M. & Kadane, Joseph B. Simplification of predictors when the utility has probability density form. Unpublished technical report, Department of Statistics, Carnegie-Mellon University, 1977.

Edwards, W., Lindman, H., & Savage, L. J. Bayesian statistical inference for psychological research. *Psychological Review*, 1963, **70**, 193–242.

Gaver, K. M., & Geisel, M. S. Discriminating among alternative models: Bayesian and non-Bayesian methods. In Paul Zarembka (Ed.) *Frontiers of econometrics*. New York: Academic Press, 1973.

Geisser, S. Bayesian estimation in multivariate analysis. *Annals of Mathematical Statistics*, 1965, **36**, 150–159.

Goldstein, M. Approximate Bayes solutions to some non-parametric problems. *Annals of Statistics*, 1975, **3**, 512–517.(a)

Goldstein, M. A note on some Bayesian non-parametric estimates. *Annals of Statistics*, 1975, **3**, 736–740.(b)

Goldstein, M. Bayesian analysis of regression problems. *Biometrika*, 1976, **63**(1), 51–58.

Gunel, E. & Dickey, J. M. Bayes factors for independence in contingency tables. *Biometrika*, 1974, **61**, 545–571.

Harrison, P. J., & Stevens, C. F. Bayesian forecasting (with discussion). *Journal of the Royal Statistical Society, Series B*, 1976, **38**(3), 205–247.

Harvey, A. C., & Collier, P. Testing for functional misspecifications in regression analysis. *Journal of Econometrics*, 1977, **6**, 103–119.

Hocking, R. R. The analysis and selection of variables in linear regression. *Biometrics*, 1976, **32**(1), 1–49.

Jeffreys, H. *Theory of probability*. (3rd ed.) London and New York: Oxford Univ. Press (Clarendon), 1961.

Kadane, J. B., & Chuang, D. Stable decision problems. *Annals of Statistics*, 1978, **6**, 1095–1110.

Leamer, E. *Specification searches*. New York: Wiley, 1978.(a)

Leamer, E. Regression selection strategies and revealed priors. *Journal of the American Statistical Association*, 1978, **73**, 580–587.(b)

Lindley, D. V. A statistical paradox. *Biometrika*, 1957, 44, 187–192.

Lindley, D. V. The use of prior probability distributions in statistical inference and decision. *Proceedings of the 4th Berkeley symposium on statistics and probability*, 1961, **1**, 453–468.

Lindley, D. V. The choice of variables in multiple regression (with discussion). *Journal of the Royal Statistical Society, Series B*, 1968, **30**(1), 31–66.

Lindley, D. V. A class of utility functions. *Annals of Statistics*, 1976, 4, 1–10.

Lindley, D. V. The Bayesian approach (with discussion). *Scandinavian Journal of Statistics*, 1978, **5**, 1–26.

Raiffa, H., & Schlaifer, R. *Applied statistical decision theory*. Boston: Division of Research, Harvard Business School. Republished in paperback: Cambridge, Mass.: MIT Press, 1968.

Ramsey, F. P. *The foundations of mathematics and other essays*. London: Kegan, Paul, Trench, Trubner & Co., Ltd., 1931.

Ramsey, J. B. Classical model selection through specification error tests. In Paul Zarembka (Eds.), *Frontiers of econometrics.* New York: Academic Press, 1974. Chapter 1.

Savage, L. J. *The foundations of statistics.* New York: Wiley, 1954.

Tiao, G. C., & Afonja, B. Some Bayesian considerations of the choice of design for ranking, selection, and estimation. *Annals of the Institute of Statistical Mathematics,* 1976, **28**, 167–185.

Von Neumann, J., & Morgenstern, O. *Theory of games and economic behavior.* Princeton, N.J.: Princeton Univ. Press, 1947.

Wintner, A. *Asymptotic distributions and infinite convolutions.* Ann Arbor, Mich.: Edwards, 1938.

Young, A. S. A Bayesian approach to prediction using polynomials. *Biometrika,* 1977, **64**, 309–317.

Zellner, A. *An introduction to Bayesian inference in econometrics.* New York.: Wiley, 1971.

Zellner, A., & Vandaele, W. Bayes–Stein estimators for R-means, regression and simultaneous equation models. In S. E. Fienberg & A. Zellner (Eds.), Studies in Bayesian Econometrics and Statistics. Amsterdam: North-Holland Publ., 1974. Pp. 627–653.