Bayesian Variable Selection in Linear Regression

T. J. Mitchell; J. J. Beauchamp


Stable URL:
http://links.jstor.org/sici?sici=0162-1459%281988%2983%3A404%3C1023%3ABVSILR%3E2.0.CO%3B2-F


---

Your use of the JSTOR archive indicates your acceptance of JSTOR’s Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR’s Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/astata.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

---

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.
This article is concerned with the selection of subsets of predictor variables in a linear regression model for the prediction of a dependent variable. It is based on a Bayesian approach, intended to be as objective as possible. A probability distribution is first assigned to the dependent variable through the specification of a family of prior distributions for the unknown parameters in the regression model. The method is not fully Bayesian, however, because the ultimate choice of prior distribution from this family is affected by the data. It is assumed that the predictors represent distinct observables; the corresponding regression coefficients are assigned independent prior distributions. For each regression coefficient subject to deletion from the model, the prior distribution is a mixture of a point mass at 0 and a diffuse uniform distribution elsewhere, that is, a “spike and slab” distribution. The random error component is assigned a normal distribution with mean 0 and standard deviation σ, where ln(σ) has a locally uniform noninformative prior distribution. The appropriate posterior probabilities are derived for each submodel. If the regression coefficients have identical priors, the posterior distribution depends only on the data and the parameter γ, which is the height of the spike divided by the height of the slab for the common prior distribution. This parameter is not assigned a probability distribution; instead, it is considered a parameter that indexes the members of a class of Bayesian methods. Graphical methods are proposed as informal guides for choosing γ, assessing the complexity of the response function and the strength of the individual predictor variables, and assessing the degree of uncertainty about the best submodel. The following plots against γ are suggested: (a) posterior probability that a particular regression coefficient is 0; (b) posterior expected number of terms in the model; (c) posterior entropy of the submodel distribution; (d) posterior predictive error; and (e) posterior probability of goodness of fit. Plots (d) and (e) are suggested as ways to choose γ. The predictive error is determined using a Bayesian cross-validation approach that generates a predictive density for each observation, given all of the data except that observation, that is, a type of “leave one out” approach. The goodness-of-fit measure is the sum of the posterior probabilities of all submodels that pass a standard F test for goodness of fit relative to the full model, at a specified level of significance. The dependence of the results on the scale of the variables is discussed, and some ways to choose the scaling constants are suggested. Examples based on a large data set arising from an energy-conservation study are given to demonstrate the application of the methods.

KEY WORDS: Cross-validation; Linear models; Subset selection.

1. INTRODUCTION

We are concerned with predicting an unknown dependent variable, given known values of k predictor variables based on a statistical analysis of n cases in which all variables are measured. This is often done using linear regression methods, which are based on the statistical model

\[ y = \sum_{j=1}^{k} \beta_j x_j + \epsilon, \]  

(1.1)

where y corresponds to the dependent variable, x_j corresponds to the jth predictor, \( \beta_j \) is the jth regression coefficient, and \( \epsilon \) is a random error. The regression coefficients are model parameters whose values are constant over all cases, whereas the value of \( \epsilon \) varies randomly from case to case. It is usually desirable to include a constant term in the model; in this case define \( x_1 = 1 \).

At some point during the analysis, one may wish to delete some predictors from the model. We refer to predictors that are candidates for deletion as vulnerable predictors. The search for a best submodel (or set of submodels) is called variable selection or subset selection. Some reasons for undertaking this search are (a) to express the relationship between y and the predictors as simply as possible; (b) to reduce future cost of prediction; (c) to identify important and negligible predictors; or (d) to increase the precision of statistical estimates and predictions. There are numerous classical approaches to variable selection. [For some of the most popular, see Hocking (1976) or Draper and Smith (1981, chap. 6).] These are generally based on sequences of hypothesis tests (e.g., stepwise regression) or estimation of some type of mean squared error (MSE) (e.g., Allen 1971a,b; Mallows 1973).

Here we take a Bayesian approach. We assign a probability distribution (the prior distribution) to the \( \beta \)'s and \( \epsilon \)'s, and hence to the y's, through (1.1). Following Box and Meyer (1986), we require that the distribution of \( \beta_j \) associated with a vulnerable predictor include a discrete probability mass at the point \( \beta_j = 0 \). Posterior probabilities are computed as usual, using the Bayes theorem. We are concerned primarily with the posterior probabilities of the various submodels, where each submodel is defined by the event that a specific subset of the \( \beta \)'s has the value 0 and the remaining \( \beta \)'s do not. These posterior probabilities can be used to find a “best” submodel or collection of submodels, although we stop short of prescribing a rule for doing so. They can also be used to generate other probabilities of interest in variable selection; for example, \( \Pr(\beta_j = 0) \) for each j.

Other Bayesian approaches have been developed in the context of model discrimination; we discuss these in Section 5.
Frequentist and Bayesian approaches alike usually start with the assumption that the observations on y, given the value of the predictor vector x, are generated by a mechanism that produces stochastic data whose frequency distribution is based on the distribution of ε in (1.1). The elements of the parameter vector β are then regarded as "real" but unknown quantities. In many applications, however, it is more appropriate to regard the model solely as a predictive device. As such, its parameters are artificial components of that device, not properties of the real world that are to be estimated, so it is difficult to define what is meant by their "true" values. We think it is useful in this case to define probabilities about the model parameters as frequencies within a hypothetical population of predicting experts (Mitchell and Beauchamp 1986a). For example, the prior probability that \( \beta_i = 0 \) is the proportion of experts in this population who choose to omit \( x_i \) from the model they use to predict y. Posterior probabilities then describe the distribution of choices made by those experts who predicted all of the observed y's correctly (to within an arbitrarily small tolerance).

We assume that if a constant term is included in the full model (1.1) it is exempt from deletion. We also assume that all of the remaining predictors represent distinct observables. That is, none of them are defined as functions of others, nor does any subset of two or more of them represent a single observable (e.g., as in the representation of the effect of a single classification variable by a set of dummy variables). The reason for these restrictions is that our primary method is based on the specification of independent and identical prior distributions for the regression coefficients subject to deletion from the model. We think this is reasonable when each predictor is a suitably scaled distinct observable; however, it may not be reasonable otherwise (see Sec. 4).

In Section 2, we describe our family of Bayesian models, present the appropriate posterior distributions, and propose methods for evaluating the prior distributions within the family. Examples of the application of these methods are given in Section 3. Section 4 considers the scaling of the predictors, and Section 5 discusses the relationship of our approach to previous work on this subject. [For further discussion and some additional computational details, see Mitchell and Beauchamp (1986a).]

2. METHOD

2.1 The Data

The observed data on the predictors are contained in the \( n \times k \) matrix X, where the ith row of X contains the values of the predictors in the ith case. We assume throughout that the rank of X is k, and that \( k < n \). Here all probability statements are implicitly conditioned on X. The observed data on the dependent variable y are contained in the n vector y.

There are \( 2^k \) possible submodels of the model (1.1), where each submodel excludes a particular subset of the predictors and includes the rest. We denote by \( X_m \) the \( n \times k_m \) matrix consisting of those columns of X that correspond to the predictors included in the mth submodel \( A_m \) for \( m = 1, 2, \ldots, 2^k \). The least squares estimate of the corresponding parameter vector \( \hat{\beta}_m \) is given by

\[
\hat{\beta}_m = (X_m'X_m)^{-1}X_m'y.
\]

and the residual sum of squares for this fit is given by

\[
S_m^2 = (y - X_m\hat{\beta}_m)'(y - X_m\hat{\beta}_m).
\]

2.2 The Prior Distribution

For each case, we assign to \( \varepsilon \) in (1.1) a normal distribution with mean 0 and variance \( \sigma^2 \), and we take the \( \varepsilon_i \)'s in distinct cases to be independent. Given \( \beta \) and \( \sigma^2 \), the n vector y has a multivariate normal distribution with mean \( X\beta \) and covariance matrix \( \sigma^2 I \). In Bayesian regression, \( \beta \) and (usually) \( \sigma \) are also random variables; different Bayesian techniques are characterized by the choice of prior distributions for these parameters. Here we assign \( \sigma \) the standard noninformative prior, under which \( \ln(\sigma) \) is locally uniform. That is, \( \ln(\sigma) \) is uniformly distributed between \(-\ln(\sigma_0)\) and \(\ln(\sigma_0)\), where \( \sigma_0 \) is very large. We further assume that the prior distribution of \( \beta \) is independent of \( \sigma \) and that the individual \( \beta \)'s are mutually independent, each having a "spike and slab" distribution. That is, \( \beta_i \) is uniformly distributed between the two limits \(-f_j\) and \(f_j\), except for a bit of probability mass concentrated at 0 if \( x_i \) is vulnerable to deletion. We are interested in taking \( f_j \) to be very large for all \( j \), to specify prior uncertainty about the value of \( \beta_j \) in those models in which \( x_j \) appears. [This class of prior distributions is a special case of one considered by Lempers (1971, p. 35-37). Lempers’ approach to the limit differs from ours, however; see Remark 1, Sec. 2.3.]

Formally,

\[
\Pr(\beta_j = 0) = h_{0j},
\]

\[
\Pr(\beta_j < b, \beta_j \neq 0) = (b + f_j)h_{1j}, \quad -f_j < b < f_j,
\]

and

\[
\Pr(|\beta_j| > f_j) = 0,
\]

where \( h_{0j} > 0, h_{1j} > 0, \) and \( h_{0j} + 2h_{1j}f_j = 1 \). We take \( f_j \) and \( g_j \) as the parameters of this distribution, where

\[
g_j = h_{0j}/h_{1j} = 2h_{0j}f_j/(1 - h_{0j});
\]

that is, \( g_j \) is the height of the spike divided by the height of the slab. To exempt certain terms (e.g., the constant term) from deletion from the model, set the corresponding \( h_{0j} \) (and hence \( g_j \)) equal to 0.

With the aforementioned model specification, our prior distribution over the submodels is

\[
\Pr(A_m) = \prod_{j_{m}} h_{0j} \prod_{j_{m}} (2f_j h_{1j})
\]

\[
= \prod_{j_{m}} g_j \prod_{j_{m}} (2f_j) \prod_{j} (g_j + 2f_j)^{-1},
\]

where \( J_m \) is the set of subscripts corresponding to the terms
that are included in $A_m$, and $\bar{J}_m$ is the set of subscripts for the terms omitted from $A_m$.

### 2.3 The Posterior Distribution

Under the Bayesian model specified previously, the probability density function of $y$, given $A_m$, $\beta_m$, and $\sigma$, is

$$
p(y \mid A_m, \beta_m, \sigma) = (2\pi)^{-n/2} \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}(S_m^2 + (\beta_m - \hat{\beta}_m)')X_m'X_m(\beta_m - \hat{\beta}_m)\right].
$$

(2.8)

(We use $p$ generically to mean "the density function of ")

If we multiply (2.8) by $p(\beta_m \mid A_m, \sigma)$, which is $\prod_{f_j}(2\pi)^{-1}$ over the region of positive probability, and integrate over $\sigma$, we obtain

$$
p(y \mid A_m, \sigma) = \left(\prod_{f_j}(2\pi)^{-1}\right)
\times (2\pi)^{-(n-k_m)/2} |X_m'X_m|^{-1/2} \sigma^{-(n-k_m)} \exp[-S_m^2/(2\sigma^2)],
$$

(2.9)

where $k_m$ is the number of terms in submodel $A_m$. To obtain (2.9), we assume that each $f_m$ is large enough so that for all $\sigma < \sigma_0$, all of the integrals from $-\infty$ to $\infty$ can be replaced by integrals from $-\infty$ to $\infty$ with arbitrarily small error. Now multiply (2.9) by $p(\sigma \mid A_m)$, which is proportional to $\sigma^{-1}$, and integrate over $\sigma$ to obtain

$$
p(y \mid A_m) = (2\ln(\sigma_0))^{-1} \left(\prod_{f_j}(2\pi)^{-1}\right) \pi^{-(n-k_m)/2} \times \Gamma\left(\frac{n-k_m}{2}\right) |X_m'X_m|^{-1/2} \exp[-(S_m^2)/(n-k_m)].
$$

(2.10)

In deriving (2.10), we assume that $\sigma_0$ is large enough so that the integral from $\sigma_0^{-1}$ to $\sigma_0$ can be replaced by the integral from 0 to $\infty$ with arbitrarily small error. Now (2.10) may be multiplied by $Pr(A_m)$ from (2.7) to obtain $p(y \mid A_m)$, from which we obtain

$$
Pr(A_m \mid y) = g \left(\prod_{f_j} \gamma_j\right) \pi^{k_m/2} \times \Gamma\left(\frac{n-k_m}{2}\right) |X_m'X_m|^{-1/2} \exp[-(S_m^2)/(n-k_m)],
$$

(2.11)

where $g$ is a normalizing constant. (For the "empty" or "null" model, replace the determinant with 1.)

**Remark 1.** Equation (2.11) holds for any prior distribution in the class defined by (2.3)–(2.5), as long as each $f_j$ is sufficiently large in the sense indicated just after (2.9). Although this result does not require formally taking a limit, it is useful to consider the sense in which (2.11) is a limiting class of posterior distributions. Let $\{F_\alpha\}$ ($\alpha = 1, 2, \ldots$) be a nested increasing unbounded sequence of $k$-dimensional rectangles, each of which defines the region of positive probability in the prior distributions (2.3)–(2.5). Given $F_\alpha$ (a particular member of the sequence), the prior distribution is completely specified by choosing the $k$ vector $\beta_0$; let $G_\alpha$ be the class of posterior distributions that corresponds to all such choices. As $\alpha$ approaches $\infty$, $G_\alpha$ approaches the class given by (2.11), with each member of the class indexed by the $k$ vector $\beta_0$. Note that the limiting class cannot be obtained by considering the elementwise limit of $\{G_\alpha\}$, that is, by fixing $\beta_0$ while taking the limit as $\alpha$ goes to $\infty$. The elementwise limit is the distribution that places all posterior probability on the submodel that omits all vulnerable predictors, no matter what the fixed value of $\beta_0$ is, and no matter what the data are. This "pathological" case has received much attention in the literature. For example, see Bartlett's (1957) remarks about the one-dimensional example used by Lindley (1957), and the much more general presentation by Lempers (1971, pp. 35–37). Lempers noted that the limiting ratio of the posterior probabilities of two submodels is "indeterminate," which is true for our special case of Lempers's prior distribution unless one of them is nested within the other, in which case the ratio of the posterior probability of the more complicated model to that of the simpler model approaches 0. Various approaches to this kind of difficulty were suggested by Halpern (1973), Atkinson (1978), Smith and Spiegelhalter (1980), Spiegelhalter and Smith (1982), and Pericchi (1984).

**Remark 2.** The determinant in (2.11) can be written as

$$
|X_m'X_m| = |X'X| |V_m|(\sigma)^{-(k-k_m)},
$$

(2.12)

where $V_m$ is the variance–covariance matrix of the least squares estimates of the $\beta$'s omitted by submodel $m$. Thus, submodels for which the information about the omitted $\beta$'s is greatest (in the sense that the generalized variance $|V_m|$ is least) tend to be favored.

To use (2.11), one needs to specify all of the $\gamma_j$'s. This can be done using (2.6) if all of the parameters of the prior distributions (2.3)–(2.5) are prescribed. Here we propose a less direct approach requiring no prior input from the user regarding the relative effects of the predictors. A modification of this approach, which allows the user more control over the prior distribution, is suggested briefly in Section 5.

We suppose that the $\beta$'s corresponding to the vulnerable predictors have identical prior distributions. (Whether this is a sensible choice is bound to the question of how to scale the predictors; see Sec. 4. For now, assume that all vulnerable predictors have been suitably scaled.) For all submodels having positive prior probability, (2.11) becomes

$$
Pr(A_m \mid y) = g' \gamma^{-k_m} \pi^{k_m/2} \times \Gamma\left(\frac{n-k_m}{2}\right) |X_m'X_m|^{-1/2} \exp[-(S_m^2)/(n-k_m)],
$$

(2.13)

where $g'$ is another normalizing constant and $\gamma$ is the common value of all positive $\gamma_j$'s. We treat $\gamma$ as an adjustable parameter of the Bayes model; that is, we do not assign a distribution to it, nor do we intend it to be chosen a
priori. Methods for assessing values of \( \gamma \) are presented in Section 2.4 and illustrated in the examples of Section 3.

Equation (2.13) can also be written as

\[
\Pr(A_m \mid y) = \frac{w_m}{\sum_{u=1}^{2^s} w_u},
\]

(2.14)

where the logarithm of any positive "weight" \( w_m \) is given by

\[
\ln(w_m) = k_m(-\ln(\gamma) + \frac{1}{n}\ln(n)) + \ln(\Gamma(\frac{1}{2}(n - k_m)))
- \frac{1}{2} \ln|X_m^T X_m| - \frac{1}{2}(n - k_m) \ln(S_m^2).
\]

(2.15)

When there is interest in evaluating a particular coefficient, \( \beta_j \), then the following posterior probability is useful:

\[
\Pr(\beta_j = 0 \mid y) = \sum_m \Pr(A_m \mid y),
\]

(2.16)

where the summation is over all submodels that do not include the coefficient \( \beta_j \). The posterior probability given in (2.13) can also be used to calculate the posterior expected number of terms in the model,

\[
E(k_m \mid y) = \sum_m k_m \Pr(A_m \mid y),
\]

(2.17)

and the posterior entropy of the submodel distribution,

\[
H = -\sum_m \Pr(A_m \mid y) \ln(\Pr(A_m \mid y)).
\]

(2.18)

The entropy \( H \) is a measure of the degree of dispersion among the submodels. We find the analog of \( H \) easier to interpret, noting that if the posterior probability were distributed equally among \( s \) submodels the analog of \( H \) would be \( s \). Plots of (2.16)–(2.18) as functions of \( \gamma \) are useful in assessing the importance of individual predictors, the number of model terms required, and the extent of uncertainty about the choice of a best submodel.

Although we are primarily interested in the posterior distribution of the submodels, we note that the posterior density of \( \beta \) is given by

\[
p(\beta \mid A_m, y) = \sum_m \Pr(A_m \mid y) p(\beta \mid A_m, y),
\]

(2.19)

where \( p(\beta \mid A_m, y) \) is a multivariate \( t \) density centered at \( \hat{\beta}_{m,y} \) (Box and Tiao 1973, p. 117). In particular, the posterior distribution of \( \beta_j \) is a mixture of scaled and shifted \( t \) distributions. The \( m \)th such distribution has \( n - k_m \) df and is centered at \( \hat{\beta}_{m,j} \) with scale factor \( q_{m,j} \), where \( \hat{\beta}_{m,j} \) is the least squares estimate of \( \beta_j \) in \( A_m \), and \( q_{m,j} \) is its standard error.

The posterior distribution of the dependent variable \( y_p \) at a specified value \( x_p \) of the predictors is also a mixture of shifted and scaled \( t \) distributions. The \( m \)th of these has \( n - k_m \) df and is centered at \( x_p^T \hat{\beta}_{m,y} \) with scale factor \( s_m[1 + x_p^T (X_m^T X_m)^{-1} x_p]^2 \), where \( x_m \) is the subvector of predictors in \( x_p \) present in model \( A_m \) and \( S_m^2 = (n - k_m) \). These results can be obtained as special cases of results given by Geisser (1965). For fixed \( \gamma \), the posterior distribution of \( y_p \) is a Bayesian predictive distribution (Geisser 1971). In the next section, we suggest a way of using this to assess choices of \( \gamma \).

### 2.4 Evaluation of \( \gamma \)

The Bayesian development so far has led only to a class of posterior distributions (2.13), indexed by the parameter \( \gamma \). Because we could not think of a reasonable noninformative prior for \( \gamma \), we decided to consider it only as a parameter that indexes the members of a class of Bayesian methods, and its choice as that of choosing the method from the class that is best in some sense. We suggest two approaches, as follows.

#### 2.4.1 Bayesian Cross-Validation

Suppose that we use the Bayesian approach to generate a predictive density \( p_i \) of \( y_i \), the dependent variable in the \( i \)th case, given all of the data except \( Y_i \), the observed value of the dependent variable in the \( i \)th case. All submodels having positive posterior probability contribute to \( p_i \); that is, we do not require selection of a best submodel first. The goodness of the predictive distribution can be assessed by comparing it with \( Y_i \) using some loss function \( L(y, Y) \) that has a distribution generated by the predictive density \( p_i \). Often a single property of the loss distribution, such as the mean, can serve as a useful measure of deficiency in the predictive distribution. For squared error loss, the mean of the loss distribution is

\[
\text{MSE}_i = E(y_i - Y_i)^2,
\]

(2.20)

where \( Y_i \) is known and the expectation \( E \) is taken over \( p_i \). [If \( i = n + 1 \) so that \( Y_i \) is a future observation, then \( \text{MSE}_i \) is a utility function in the sense of San Martini and Spezzaferri (1984); however, these authors did not consider its use for cross-validation.]

In the spirit of cross-validation, \( \text{MSE}_i \) can be computed for each \( Y_i \) in the data set, where the predictive density \( p_i \) in each case is based on all data except \( Y_i \). We take as a measure of the deficiency of prediction the square root of the average of \( \text{MSE}_i \) over all cases. We refer to this as the predictive error,

\[
\text{PE} = \sqrt{(1/n) \sum_{i=1}^{n} \text{MSE}_i},
\]

(2.21)

and use it to assess \( \gamma \), knowing that large values of \( \text{PE} \) result from choices of \( \gamma \) that place too much posterior probability on submodels that are bad for prediction.

Most of the work in computing \( \text{PE} \) involves computing the \( n \) vectors \( h_m \) and \( e_m \) for each submodel, where \( h_m \) is the diagonal of the hat matrix \( X_m (X_m^T X_m)^{-1} X_m \), and \( e_m \) is the least squares residual vector. In Mitchell and Beauchamp (1986a) we showed that

\[
\text{MSE}_i = \sum_m \Pr_i(A_m)[f_{i,m}^2 + (S_m^2 - e_{i,m} - f_{i,m})(1 - h_{i,m})^{-1}(n - k_m - 3)^{-1}],
\]

(2.22)

where \( f_{i,m} = e_{i,m}/(1 - h_{i,m}) \), and \( \Pr_i(A_m) \) is the posterior probability of model \( m \) excluding the data for the \( i \)th case. The latter can be computed by using (2.13), where the residual sum of squares and the determinant of the \( X_m^T X_m \)
matrix that result when the $i$th case is omitted from the data set are given by

$$S_m(i) = S_m - e_{i,m} f_{i,m}$$

(2.23)

and

$$|X_m(i)X_m(i)|^{-1} = |X_mX_m|^{-1}/(1 - h_{i,m}).$$

(2.24)

A plot of PE as a function of $\gamma$ is useful, since it provides a visual assessment of the effect of $\gamma$ on the predictive ability of the posterior distribution. Note that PE is finite only if $n - k_m > 3$, so this approach to assessing $\gamma$ is not useful unless there are at least 4 df for error.

Remark 3. Our approach differs from the usual approach to cross-validation in that we define the deficiency of prediction at a given $x_i$ directly as a function of $Y_i$ and $p_i$. That is, we do not first reduce $p_i$ to scalar $Y_i$, which is then treated as the chosen predicted value. In this respect, our approach is similar to that of Geisser and Eddy (1979), who in effect defined the deficiency of prediction to be $-\log p_i(Y_i)$.

Remark 4. Although one could minimize PE as a formal way of choosing $\gamma$, we prefer to use the PE plot as an informal guide. Consideration of PE allows us to avoid choosing values of $\gamma$ that may lead to unacceptably large predictive errors. Note, however, that PE is a measure of predictive ability averaged over the cases in the data set at hand. Thus, it is useful when the cases at hand are representative of the cases for which one intends to make predictions, but not so useful otherwise.

2.4.2 Goodness-of-Fit Plot

Another useful way of evaluating $\gamma$ is to plot the posterior probability of goodness of fit, $G$, as a function of $\gamma$, where $G$ is the sum of the posterior probabilities of all submodels that pass a standard $F$ test for goodness of fit relative to the full $k$-variable model, at a specified level of significance.

Remark 5. In the present setting, Lindley’s paradox (Lindley 1957) occurs when the data are such that a particular submodel is given high posterior probability, although it is strongly rejected by a standard sampling theory lack-of-fit test. The implications of this have been the topic of much discussion in the literature [e.g., see Shafer (1982, including discussion); Smith and Spiegelhalter (1980)]. If $G$ is used to choose $\gamma$, this situation can be avoided, but only because we allow the data to help choose the prior distribution from the class we are considering. Similarly, both Smith and Spiegelhalter (1980) and Pericchi (1984) presented approaches that avoid the paradox, but in both cases the prior is design-dependent.

2.5 Computations

To calculate the posterior probabilities of the submodels, we compute the residual sum of squares $S_m$ and $|X_mX_m|$ for every submodel assigned a positive prior probability. Our initial program, which we used for the examples of Section 3, does the main computations for each submodel independently. We expect that this program could benefit from a more efficient approach, under which computations made for one submodel can be used in obtaining the results for another (e.g., see Furnival and Wilson 1974).

In anticipation of applications to data sets with many predictors, we also developed a branch-and-bound method that avoids the calculations for submodels that are known, by the results of earlier calculations, to have negligible posterior probability. We are still refining and testing this method; the approach was outlined in Mitchell and Beauchamp (1986b).

3. EXAMPLES

3.1 Background

The following examples are all based on data from an energy-conservation study (Hirst, White, Holub, and Goelitz 1985). These data consist of observations on the electricity savings (the dependent variable) for a sample of 401 households that participated in a residential weatherization program. The Appendix contains a list of the 10 predictor variables and dependent variable used in this analysis. The first example consists of an analysis of the actual data. In the second and third examples, we simulated some of the data to make the analysis more interesting.

3.2 Example 1: The Real Data

The first step in the analysis of this data set was to do the usual regression analyses, using a first-order model with 11 terms, including a constant term. [We depart slightly from the notation of (1.1) here, in that the subscripts run from 0 to 10, with the first term the constant $\beta_0$.] The results of this analysis included residual plots, calculations of $R^2$ and Mallows’s (1973) $C_p$ for all possible subset regressions, and calculations of regression diagnostics. To be consistent with the variables used in the Bayesian methodology, except for the constant, the original predictors (denoted by 1, $X_1$, $X_2$, . . . , $X_10$) were scaled by subtracting their means and dividing by their standard deviations. We denote these centered and scaled predictors by $x_1$, $x_2$, . . . , $x_{10}$. The least squares estimate of $y$ for these data was

$$\hat{y} = 5,068 - 116x_1 - 367x_2 + 353x_3 + 879x_4 - 370x_5$$

$$+ 618x_6 - 429x_7 + 2,812x_8 - 452x_9 + 433x_{10}.$$  

The residual plots did not show any obvious problems. The magnitude of the variance inflation factors (all <1.5) and the maximum condition index (2.04) did not imply any significant problems of collinearity among the predictor variables. The full model did not have a high $R^2$ (.465). All of the regression coefficients were significant ($P < .05$) according to Student’s $t$ test, except for $\beta_1$, for which $P$ was .497. The submodel that omitted $x_1$ was the only one for which $C_p$ was less than 1.1$k$. (Throughout this section, we somewhat arbitrarily regard all submodels that satisfy this criterion as acceptable with respect to $C_p$.)
We then analyzed these data using Bayesian variable selection, as described in Section 2, where all predictors other than the constant term were vulnerable to deletion. Again, except for the constant, the original predictors were centered by subtracting their means and scaled by dividing by their standard deviations.

Figure 1a shows the relative strengths of the individual predictors, with the weaker ones rising earlier and increasing to 1 sooner than the stronger ones, as $y$ increases. We see that the preprogram electricity use ($x_8$) is a far stronger predictor than the others. Figures 1d and 1e suggest that $\ln(y)$ should be less than 6 if we want to minimize our measure of PE and maintain a high posterior probability on the submodels that show no lack of fit. For values of $\ln(y)$ in this range, the posterior probability is concentrated on the full model, so we would not omit any predictors. Note in Figure 1c that the antilog of the entropy starts to rise sharply at $\ln(y) = 6$, showing increasing confusion about the best submodel. For values of $\ln(y)$ between 25 and 50, Figures 1a–c show that the posterior distribution is concentrated on a single submodel having only two terms, $\beta_0$ and $\beta_8 x_8$. Figure 1d shows that this greatly simplified model is acceptable if one is willing to tolerate a PE of about 5,000 instead of the minimum of 4,550. Neverthe-

less, Figure 1e reminds us that this model shows a significant lack of fit. For each value of $\ln(y)$, the time needed to make the necessary calculations for this data set was from 15 to 20 seconds on a Cray X-MP.

### 3.3 Example 2: Important and Unimportant Predictors

Here we used the same data set, but replaced the values of the dependent variable with values generated from a known set of coefficients. The predictors were the same as in Example 1.

The new observed values of the dependent variable were simulated with

$$y = f(x) + \varepsilon,$$

where

$$f(x) = 3,312 + 4,755 x_1 + 5,196 x_2 + 455 x_3$$

$$+ 532 x_4 + 538 x_5 + 104 x_6 + 91 x_7 \tag{3.2}$$

and $\varepsilon$ represents a normal random variable with mean 0 and standard deviation 1.000. We consider $\beta_1$ and $\beta_2$ large, $\beta_3$, $\beta_4$, and $\beta_5$ moderate, $\beta_6$ and $\beta_7$ small, and $\beta_8$, $\beta_9$, and $\beta_{10}$ null.

The least squares estimate of $f(x)$ based on the simu-
lated data was
\[
\hat{y} = 3.289 + 4.712x_1 + 5.148x_2 + 440x_3 \\
+ 444x_4 + 546x_5 + 51x_6 + 58x_7 \\
- 20x_8 + 85x_9 + 35x_{10}.
\] (3.3)
The value of \( R^2 \) for the full model was .980. The \( P \) values for Student's \( t \) tests of the significance of the small and null coefficients, \( \beta_i - \beta_{10} \), were .31, .27, .73, .13, and .48, respectively. There were 22 submodels for which \( C_p \) was less than 1.1\( k \). All of these included the terms in \( x_1-x_3 \); 11 of them included at least two of the three null terms. The lowest value of \( C_p \) was 5.598, for the model that omits the terms in \( x_6, x_7, x_9, \) and \( x_{10} \).

The results of the Bayesian procedure are illustrated in Figure 2. Figure 2a sorts out the predictors nicely with respect to strength, although it cannot distinguish between the small ones (\( x_1 \) and \( x_2 \)) and the null ones (\( x_6-x_{10} \)). Figures 2d and 2e suggest that \( \ln(\gamma) \) can be as great as about 32 without increasing PE or detracting from the goodness of fit. [The PE actually decreases slightly, from a value of 1,377 at \( \ln(\gamma) = -10 \) to a value of 1,372 at \( \ln(\gamma) = 25 \).] Figures 2b and 2c show that for \( \ln(\gamma) \) about 32, the posterior probability is concentrated on a single model having six terms; Figure 2a shows that this model contains the predictors \( x_1-x_3 \). This is clearly the model of choice, unless one is willing to tolerate an increase in PE to about 1,700 (Fig. 2d) to obtain an even simpler model. In this case, one would choose \( \ln(\gamma) \) greater than about 37, where the posterior probability is concentrated on the model containing predictors \( x_1 \) and \( x_3 \).

### 3.4 Example 3: Collinearity

In this example, strong collinearity was introduced among some of the predictor variables. This was done by replacing the predictor \( X_9 \) with
\[
X_{9,\text{New}} = \hat{X}_9 + .01(X_9 - \bar{X}_9),
\] (3.4)
where \( \hat{X}_9 \) is the least squares predicted value of \( X_9 \) when it is regressed on \( X_1, X_3, X_6, X_7, \) and \( X_8 \), with a constant included in the regression equation. A second collinear relation was introduced by replacing the predictor \( X_2 \) with
\[
X_{2,\text{New}} = \hat{X}_2 + .01(X_2 - \bar{X}_2),
\] (3.5)
where \( \hat{X}_2 \) is the least squares predicted value of \( X_2 \) when it is regressed on \( X_1, X_3, X_4, X_6, \) and \( X_7 \), with a constant included in the regression equation. Before the observed values of the dependent variable (\( y \)) were generated, the 10 predictor variables were centered and scaled with the same constants used in the previous examples. The generated values of the dependent variable were obtained with the same \( \beta \)'s and \( \epsilon \)'s used in Example 2. Thus the only difference between Examples 2 and 3 is in the data for predictors \( x_2 \) and \( x_9 \). The true response function, the vector of errors, and the prior distribution of the \( \beta \)'s are unchanged.

The least squares estimates of the regression coefficients, the \( P \) values for the associated \( t \) statistics, and the variance inflation factors are given in Table 1.

<table>
<thead>
<tr>
<th>( j )</th>
<th>( \beta )</th>
<th>( \hat{\beta} )</th>
<th>( P )</th>
<th>( \text{VIF} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.755</td>
<td>3.863</td>
<td>&lt;.01</td>
<td>123.0</td>
</tr>
<tr>
<td>2</td>
<td>5.196</td>
<td>338</td>
<td>.95</td>
<td>263.5</td>
</tr>
<tr>
<td>3</td>
<td>455</td>
<td>220</td>
<td>.47</td>
<td>40.6</td>
</tr>
<tr>
<td>4</td>
<td>332</td>
<td>824</td>
<td>.04</td>
<td>70.4</td>
</tr>
<tr>
<td>5</td>
<td>538</td>
<td>546</td>
<td>&lt;.01</td>
<td>1.2</td>
</tr>
<tr>
<td>6</td>
<td>604</td>
<td>-455</td>
<td>.37</td>
<td>112.1</td>
</tr>
<tr>
<td>7</td>
<td>91</td>
<td>-1,968</td>
<td>.16</td>
<td>832.2</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>-2,964</td>
<td>.13</td>
<td>1,679.8</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>8,507</td>
<td>.13</td>
<td>3,143.1</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>35</td>
<td>.48</td>
<td>1.1</td>
</tr>
</tbody>
</table>

The effects of the collinearity are evident. As one might expect, this leads to considerable confusion regarding the choice of a best model. The value of \( R^2 \) for the full model is .952; there are 89 submodels having values of \( R^2 \) greater than .950. There are 13 submodels having values of \( C_p \) less than 1.1\( k \); these all include the terms in \( x_1, x_4, x_5, \) and \( x_7 \). Six of the thirteen exclude \( x_2 \), and eight of them include at least two of the three null terms. The lowest \( C_p \) is 5.87, for the model that includes \( x_1, x_3, x_4, x_5, x_6, \) and \( x_7 \).

The results of the Bayesian analysis are illustrated in Figure 3. It appears that one can choose \( \ln(\gamma) \) as great as

![Figure 3](image-url)

Figure 3. Plots of (a) Posterior Probabilities That Each \( \beta_i = 0, \text{Pr}(\beta_i = 0 \mid y) \); (b) Posterior Expected Number of Terms in Model, \( E(k_n \mid y) \); (c) Antilog of the Posterior Entropy, \( \exp(H) \); (d) Predictive Error, \( \text{PE} \); and (e) Posterior Probability of Goodness of Fit, \( G \), as Functions of \( \ln(\gamma) \) for Weatherization Data With 10 Predictor Variables, Simulated Values of the Dependent Variable, and Collinear Predictor Variables (Sec. 3.4). The numbers on the curves in panel a indicate \( j \) for predictor variable \( X_j \) (see Appendix).
about 10 without degrading goodness of fit (Fig. 3c) or predictive capability (Fig. 3d). At this point, the posterior distribution is spread over several submodels (with the antilog of the entropy about 4 in Fig. 3c). Figure 3a suggests that at $\ln(\gamma) = 10$ the terms in $x_{r-x_{10}}$ can be omitted and the rest probably kept, though there is some doubt about $x_3$ and $x_4$. (The nonmonotone behavior of some of the curves in Fig. 3a is presumably caused by the collinearity, but we have not investigated the relationship.) An examination of the posterior probabilities of the submodels when $\ln(\gamma) = 10$ reveals that the two most favored submodels are $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$, which has probability .48, and $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$, which has probability .38. The next most favored submodel has probability .02. If one does not mind lack of fit, but is concerned mainly with predictive error, larger values of $\ln(\gamma)$ (up to about 40) are acceptable. At this point, all of the posterior probability is concentrated on the model that includes $x_1$, $x_4$, $x_5$, and $x_6$.

4. CENTERING AND SCALING

In general, the predictors $x_j$ used in the analysis are the result of centering and scaling the original predictors $X_j$; that is,

$$x_j = (X_j - c_j)/d_j,$$  \hspace{1cm} (4.1)

where $c_j$ is the centering constant and $d_j$ is the scaling constant for the $j$th predictor. We discuss the effect of these constants on the Bayesian analysis described in Section 2, and offer some suggestions for choosing them.

The choice of centering constants rarely causes any problems. Usually, there is a constant term in the model, exempt from deletion; then, any choice of centering constants will do, since the posterior probabilities in (2.13) are unaffected. If there is no constant term in the model, then all centering constants should be 0.

The posterior probabilities in (2.13) are also unaffected by the choice of scaling constant for any predictor exempt from deletion. It is obvious that the posterior probabilities are not invariant, for fixed $\gamma$, to scale changes in the dependent variable. Nevertheless, the class of posterior distributions given by (2.13) is invariant. It can easily be shown that scale changes in the dependent variable merely shift the $\ln(\gamma)$ axis, in plots such as those in Figures 1–3, by an amount equal to the log of the scaling constant. The same type of invariance holds for the vulnerable predictors, if all are scaled by the same constant. If the vulnerable predictors are scaled by different constants, however, the scaling matters. Since the method proposed in Section 2 rests largely on the specification of identical priors for the $\beta$'s associated with these predictors, our approach here is to devise scaling rules under which this seems reasonable.

Our intent in choosing identical priors was to reflect prior impartiality about the relative importance of the predictors. We propose that the importance of the $j$th predictor be defined in some reasonable way as a multiple of $|\theta_j|$, where $\theta_j$ is the coefficient of the original predictor $X_j$, and that the scaling constant $d_j$ be set equal to the multiplier for each $j$. The choice of identical priors for the $\beta$'s then implies that the prior distribution of importance is the same for all predictors.

We suggest the following types of scaling based on this approach:

1. For each raw predictor $X_j$, choose a range of prediction, an interval of width $r_j$ that covers the values of $X_j$ for which one expects to use the regression equation. Define the importance of $X_j$ as $|\theta_j| r_j$, the absolute change in $E(y)$ effected by moving $X_j$ across its range of prediction. Now, set the scaling constant $d_j$ equal to $r_j$; this makes the importance of the $j$th predictor equal to $|\beta_j|$.

2. Another definition of importance depends on the distribution of future predictor values. Suppose that the future distribution of values of $X_j$ at which one expects to use the regression equation has a standard deviation $\sigma_j$. Then, the standard deviation of the contribution $\theta_j X_j$ to the prediction of $y$ is $|\theta_j σ_j|$; define the importance of the $j$th predictor as equal to this. Now, choose the scaling constant $d_j$ equal to $\sigma_j$; this makes the importance of the $j$th predictor equal to $|\beta_j|$. If the values of $X_j$ in the data at hand are representative of the population for which one intends to make future predictions, then one can use the standard deviation of $X_j$ computed from the observed data to estimate $\sigma_j$. This justifies the standard scaling used in our examples.

Note, however, that the rationale for both types of scaling is much weaker if the values of the predictors at which one would like to make predictions are correlated. In this case, it is much more difficult to define the importance of a predictor, and hence to justify a particular choice of scaling constants. The same remark applies to models in which some of the predictors are known functions of others, such as a quadratic response surface model. More attention to this point is needed, we think, before our methods can be applied to such models.

5. DISCUSSION

Several related Bayesian variable selection methods have been considered by other authors, often from the viewpoint of model selection or model discrimination. In several of these (Atkinson 1978; Box and Meyer 1986; Halpern 1973; Pericchi 1984; Poirier 1983; Smith and Spiegelhalter 1980; Spiegelhalter and Smith 1982) a normal conjugate prior is used for $\beta_m$, given $σ^2$ and the submodel $A_m$. Within each submodel, the covariance matrix of $\beta_m$ is a known matrix $V_m × σ^2$; the elements of $V_m$ are often considered to be very large to simulate ignorance about $\beta_m$. When the prior information is in the form of a real or imagined prior experiment, the dependence of the uncertainty in $\beta_m$ on $σ^2$ seems reasonable. Other authors (e.g., Dickey 1975) prefer to specify prior independence between $β_m$ and $σ^2$. We think this is appropriate when attempting to formulate noninformative priors, so this is the course we have adopted. [For further discussion, see Tiao and Zellner (1964) and Lempers (1971, pp. 21–26).]

Many authors, particularly those who view the problem as one of model selection, do not relate the regression coefficients among the models in any way. For example,
the coefficient of $x_j$ is a different random variable in every model in which $x_j$ appears, and these random variables are independent (e.g., see Atkinson 1978; Halpern 1973; Pericchi 1984; Smith and Spiegelhalter 1980; Spiegelhalter and Smith 1982). This approach allows a great deal of flexibility, but it requires the specification of a great many prior parameters. Other authors (Box and Meyer 1986; Dickey 1975; Poirier 1985) approached the problem differently, by defining only one random variable for each regression term and specifying a joint distribution for the set of them. Under this approach, the distribution of the regression coefficients within a submodel is not arbitrary, but induced as a conditional distribution from the prior distribution of the coefficients in the full model. In addition to being intuitively sensible in many cases, this greatly reduces the number of prior parameters that need to be specified. This is our approach, with the additional restriction of independence among the coefficients.

A major concern of much of the literature is the behavior of the posterior distribution as the prior distribution becomes less and less informative. The usual approach is to define a sequence of priors, in which the distribution of the parameters of each submodel becomes progressively more diffuse, whereas the prior probabilities of the models remain fixed. Unfortunately, this leads to a posterior distribution that is indeterminate in the most general case (Lempers 1971, pp. 35–37), when the regression coefficients among the various models are unrelated. If the regression coefficients among the models are related to one another through the prior distribution on the full model (as in our approach), then a limiting posterior distribution exists; however, it places all probability on the empty model, that is, the one with no vulnerable predictors. Although the sequence of priors in this case may be decreasingly informative about the parameters within each model, it is evidently not so with respect to the choice of model. We have adopted a different approach to the limiting process—we consider a sequence of sets of prior distributions, where all possible configurations of model probabilities (subject to independence of the $\beta$'s) are represented within each set. This leads to a nontrivial limiting set of posterior distributions (as indicated in Sec. 2).

Several articles have dealt with specifying the parameters of the prior distribution, or the functions of these parameters that appear as arbitrary constants in the posterior distribution as one progresses to the limit. Various criteria have been discussed, such as equality (across models) of the prior generalized variance of the regression coefficients (Halpern 1973), consistency with the likelihood ratio (Atkinson 1978), avoidance of Lindley's paradox (Smith and Spiegelhalter 1980), consistency with inferences based on a thought experiment (Spiegelhalter and Smith 1982), equality (across models) of expected gain in information (Pericchi 1984), and prior experience (Box and Meyer 1986). These criteria are usually used to determine the prior precision of the model coefficients, but they can also be used to modify the prior probabilities of the models (Atkinson 1978; Pericchi 1984). Some of the prescriptions (Halpern 1973; Spiegelhalter and Smith 1982) depend only on the dimensionality of the models. Others (Atkinson 1978; Pericchi 1984; Smith and Spiegelhalter 1980) require knowledge of the experimental design. Although knowledge of the design is available, there may be philosophical difficulties with allowing one's prior uncertainty about the regression equation to depend on the design of the experiment.

Our approaches to determining the unspecified parameter $\gamma$ (Sec. 2.4) depend on both the design and the data. We choose, from a class of Bayesian methods indexed by $\gamma$, the one that is best in some sense. At this point, the interpretation of $\gamma$ in terms of prior uncertainty is no longer of interest to us. (The same attitude could also be used to defend, or at least ignore, the design dependency of the priors mentioned previously.) The two criteria we have suggested are predictive performance in a "leave one out" game (Sec. 2.4.1) and consistency with standard sampling theory lack-of-fit tests (Sec. 2.4.2). Taken as a whole, then, our method is really a hybrid rather than a fully Bayesian approach, although the final inferences are expressed in Bayesian terms as posterior probabilities.

Of course, there are numerous non-Bayesian techniques for variable selection (e.g., see Hocking 1976). Classical approaches have produced many useful statistics, such as Mallows's (1973) $C_p$, Allen's (1971b) PRESS (prediction sum of squares), $P$ values for specified lack-of-fit tests, and so forth, that should be considered in arriving at a reasoned opinion regarding the consequences of omitting predictors from the model. Most classical methods, however, are based on statements of the form, "If the data are generated by submodel $A_m$, then the probability of the event $E$ (or the expectation of the statistic $Z$) is $Q$." There are many potentially informative statements of this type, corresponding to different choices of $A_m$ and $E$ or $Z$. Synthesizing this information and resolving the inevitable conflicts is not easy, especially when data-guided strategies change the $Q$'s in unknown ways. For a good exposition of the difficulties, see Miller (1984).

In contrast, synthesizing the available information using a Bayesian approach is relatively straightforward, subject to computational limitations. All of the agony centers on the choice of prior distribution(s). We have reduced this agony to the choice of a single parameter $\gamma$ and the choice of scaling (which we think is unavoidable in variable selection), and we have suggested some ways to make these choices (see Secs. 2.4 and 4).

We have deliberately tried to choose a Bayesian approach in which the prior beliefs of the user are as unobtrusive as possible, not because of a philosophical commitment to "objective" Bayesian techniques, but because we think that such techniques should be available for routine use. The prior marginal distributions of the $\beta$'s are all the same, which under the suggested scaling conventions should make the procedure fair to all predictors. Our class of priors on the submodels was set up to be objective (even though individual members of the class are not), in that it contains a whole spectrum of priors, ranging from those that favor small submodels to those that favor large ones. Within each submodel, the prior is objective in the sense...
that the distribution of the regression coefficients is diffuse enough to cover the region of nonnegligible likelihood, and is uniform over that region. Finally, the user need not specify in advance the strength of belief that any particular predictor is null, since $\gamma$ is an adjustable constant that can be assessed using the data.

If desired, our approach can be modified to give the user more control over the prior distribution without having to specify all of its parameters explicitly. As before, we assume the predictors have been suitably scaled. Set all of the $f_j$'s in (2.3)–(2.5) equal to an unspecified constant; then, choose the odds for deletion $\phi_j = h_{0j}/(1 - h_{0j})$ for each predictor. By (2.6), $\gamma_i = \gamma \phi_j$, where $\gamma$ is an unspecified constant. Now, use (2.11) and evaluate the parameter $\gamma$ as described in Section 2.4. Note that it is only the relative odds for deletion that matter, since if all of the $\phi$'s are multiplied by a constant, the same family of posterior distribution arises.

We think that the major practical limitations of our method are the exclusion of functional dependencies among predictors, the exclusion of class variables having more than two levels, the computational effort required if there are many predictors, and the restriction of the error model to the normal distribution. We hope that these limitations will be removed as a result of further work in this area, and that meanwhile our method will be a useful adjunct to methods of variable selection currently in use.

**APPENDIX: VARIABLES FROM RESIDENTIAL WEATHERIZATION PROGRAM**

The dependent variable used in the analysis is $Y$: electricity savings (in kilowatt hours per year). The predictor variables used are as follows:

- $X_1$: presence of air-conditioning equipment (0 or 1)
- $X_2$: long-run heating degree days
- $X_3$: change in number of household members
- $X_5$: change in electricity price
- $X_6$: switched primary heating fuel from electricity (0 or 1)
- $X_7$: household income
- $X_8$: preprogram electricity use
- $X_9$: heated floor area
- $X_{10}$: audit prediction of saving.

*Received December 1986. Revised June 1987.*

**REFERENCES**

Allen, D. M. (1971a), "Mean Square Error of Prediction as a Criterion for Selecting Variables" (with discussion), Technometrics, 13, 469–481.


