Bayesian Model Averaging for Linear Regression Models

Adrian E. Raftery; David Madigan; Jennifer A. Hoeting

Journal of the American Statistical Association, Vol. 92, No. 437 (Mar., 1997), 179-191.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199703%2992%3A437%3C179%3ABMAFLR%3E2.0.CO%3B2-9

Bayesian Model Averaging for Linear Regression Models

Adrian E. RAFTERY, David MADIGAN, and Jennifer A. HOETING

We consider the problem of accounting for model uncertainty in linear regression models. Conditioning on a single selected model ignores model uncertainty, and thus leads to the underestimation of uncertainty when making inferences about quantities of interest. A Bayesian solution to this problem involves averaging over all possible models (i.e., combinations of predictors) when making inferences about quantities of interest. This approach is often not practical. In this article we offer two alternative approaches. First, we describe an ad hoc procedure, "Occam's window," which indicates a small set of models over which a model average can be computed. Second, we describe a Markov chain Monte Carlo approach that directly approximates the exact solution. In the presence of model uncertainty, both of these model averaging procedures provide better predictive performance than any single model that might reasonably have been selected. In the extreme case where there are many candidate predictors but no relationship between any of them and the response, standard variable selection procedures often choose some subset of variables that yields a high R^2 and a highly significant overall F value. In this situation, Occam's window usually indicates the null model (or a small number of models including the null model) as the only one (or ones) to be considered thus largely resolving the problem of selecting significant models when there is no signal in the data. Software to implement our methods is available from StatLib.

KEY WORDS: Bayes factor; Markov chain Monte Carlo model composition; Model uncertainty; Occam's window; Posterior model probability.

1. INTRODUCTION

Selecting subsets of predictor variables is a basic part of building a linear regression model. The objective of variable selection is typically stated as follows: Given a dependent variable $\mathbf{Y}$ and a set of a candidate predictors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k$, find the "best" model of the form

$$\mathbf{Y} = \beta_0 + \sum_{j=1}^{p} \beta_j \mathbf{X}_j + \varepsilon,$$

where $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p$ is a subset of $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k$. Here "best" may have any of several meanings; for example, the model providing the most accurate predictions for new cases exchangeable with those used to fit the model.

A typical approach to data analysis is to carry out a model selection exercise leading to a single "best" model and then to make inferences as if the selected model were the true model. However, this ignores a major component of uncertainty—namely, uncertainty about the model itself (Draper 1995; Hodges 1987; Leamer 1978; Moulton 1991; Raftery 1988, 1996). As a consequence, uncertainty about quantities of interest can be underestimated. (For striking examples of this see Draper 1995, Kass and Raftery 1995, Madigan and York 1995, Miller 1984, Raftery 1996, and Regal and Hook 1991.) A complete Bayesian solution to this problem involves averaging over *all* possible combinations of predictors when making inferences about quantities of interest. Indeed, this approach provides optimal predictive ability (Madigan and Raftery 1994). However, in many applications this averaging will not be a practical proposition. Here we present two alternative approaches.

First, we extend the Bayesian graphical model selection algorithm of Madigan and Raftery (1994) to linear regression models. We refer to this algorithm as "Occam's window." This approach involves averaging over a reduced set of models. Second, we directly approximate the complete solution by applying the Markov chain Monte Carlo model composition (MC^3) approach of Madigan and York (1995) to linear regression models. In this approach the posterior distribution of a quantity of interest is approximated by a Markov chain Monte Carlo method that generates a process that moves through model space. We show in an example that both of these model averaging approaches provide better predictive performance than any single model that might reasonably have been selected.

Freedman (1983) pointed out that when there are many predictors and there is no relationship between the predictors and the response, variable selection techniques can lead to a model with a high R^2 and a highly significant overall F value. By contrast, when a dataset is generated with no relationship between the predictors and the response, Occam's window typically indicates the null model as the "best" model or as one of a small set of "best" models, thus largely resolving the problem of selecting a significant model for a null relationship.

The background literature for our approach includes several areas of research: the selection of subsets of predictor variables in linear regression models (Breiman 1992, 1995; Breiman and Spector 1992; Draper and Smith 1981; Hocking 1976; Linhart and Zucchini 1986; Miller 1990; Shibata 1981), Bayesian approaches to the selection of subsets of predictor variables in linear regression models (George and McCulloch 1993; Laud and Ibrahim 1995; Mitchell and Beauchamp 1988; Schwarz 1978), and model uncertainty

Adrian E. Raftery is Professor of Statistics and Sociology, and David Madigan is Assistant Professor of Statistics, Department of Statistics, University of Washington, Seattle, WA 98195. Jennifer Hoeting is Assistant Professor of Statistics, Department of Statistics, Colorado State University, Fort Collins, CO 80523. The research of Raftery and Hoeting was partially supported by Office of Naval Research contract N-00014-91-J-1074. Madigan's research was partially supported by National Science Foundation grant DMS 92111627. The authors are grateful to Danika Lew for research assistance and the editor, the associate editor, two anonymous referees, and David Draper for very helpful comments that greatly improved the article.

Journal of the American Statistical Association
March 1997, Vol. 92, No. 437, Theory and Methods

(Freedman, Navidi, and Peters 1986; Leamer 1978; Madigan and Raftery 1994; Stewart 1987; Stewart and Davis 1986).

In the next section we outline the philosophy underlying our approach. In Section 3 we describe how we selected prior distributions and outline the two model averaging approaches in Section 4. In Section 5 we provide an example and describe our assessment of predictive performance. In Section 6 we compare the performance of Occam's window to that of standard variable selection methods when there is no relationship between the predictors and the response. Finally, in Section 7 we discuss related work and suggest future directions.

2. ACCOUNTING FOR MODEL UNCERTAINTY USING BMA

As described previously, basing inferences on a single "best" model as if the single selected model were true ignores model uncertainty, which can result in underestimating uncertainty about quantities of interest. There is a standard Bayesian solution to this problem, proposed by Leamer (1978). If $\mathcal{M} = \{M_1, \ldots, M_K\}$ denotes the set of all models being considered and if Δ is the quantity of interest such as a future observation or the utility of a course of action, then the posterior distribution of Δ given the data D is

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

This is an average of the posterior distributions under each model weighted by the corresponding posterior model probabilities. We call this Bayesian model averaging (BMA). In Equation (1) the posterior probability of model M_k is given by

$$\Pr(M_k|D) = \frac{\Pr(D|M_k)\Pr(M_k)}{\sum_{l=1}^{K} \Pr(D|M_l)\Pr(M_l)}, \tag{2}$$

where

$$\Pr(D|M_k) = \int \Pr(D|\boldsymbol{\theta}_k, M_k) \Pr(\boldsymbol{\theta}_k|M_k)\, d\boldsymbol{\theta}_k \tag{3}$$

is the marginal likelihood of model M_k, $\boldsymbol{\theta}_k$ is the vector of parameters of model M_k, $\Pr(\boldsymbol{\theta}_k|M_k)$ is the prior density of $\boldsymbol{\theta}_k$ under model M_k, $\Pr(D|\boldsymbol{\theta}_k, M_k)$ is the likelihood, and $\Pr(M_k)$ is the prior probability that M_k is the true model. All probabilities are implicitly conditional on $\mathcal{M}$, the set of all models being considered. In this article we consider $\mathcal{M}$ to be equal to the set of all possible combinations of predictors.

Averaging over *all* of the models in this fashion provides better predictive ability, as measured by a logarithmic scoring rule, than using any single model M_j:

$$-E\left[\log\left\{\sum_{k=1}^{K} \Pr(\Delta|M_k, D)\Pr(M_k|D)\right\}\right] \leq -E[\log\{\Pr(\Delta|M_j, D)\}] \qquad (j = 1, \ldots, K),$$

where Δ is the observable to be predicted and the expectation is with respect to $\sum_{k=1}^{K} \Pr(\Delta|M_k, D)\Pr(M_k|D)$. This follows from the nonnegativity of the Kullback–Leibler information divergence.

Implementation of Bayesian model averaging is difficult for two reasons. First, the integrals in (3) can be hard to compute. Second, the number of terms in (1) can be enormous. In this article we present solutions to both of these problems.

3. BAYESIAN FRAMEWORK

3.1 Modeling Framework

Each model that we consider is of the form

$$\mathbf{Y} = \beta_0 + \sum_{j=1}^{p} \beta_j \mathbf{X}_j + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{4}$$

where the observed data on p predictors are contained in the $n \times (p+1)$ matrix $\mathbf{X}$. The observed data on the dependent variable are contained in the n vector $\mathbf{Y}$. We assign to $\boldsymbol{\varepsilon}$ a normal distribution with mean zero and variance σ^2 and assume that the ε's in distinct cases are independent. We consider the $(p+1)$ individual parameters $\boldsymbol{\beta}$ and σ^2 to be unknown.

Where possible, informative prior distributions for $\boldsymbol{\beta}$ and σ^2 should be elicited and incorporated into the analysis (see Garthwaite and Dickey 1992 and Kadane, Dickey, Winkler, Smith, and Peters 1980). In the absence of expert opinion, we seek to choose prior distributions that reflect uncertainty about the parameters and also embody reasonable a priori constraints. We use prior distributions that are proper but reasonably flat over the range of parameter values that could plausibly arise. These represent the common situation where there is some prior information, but rather little of it, and put us in the "stable estimation" case where results are relatively insensitive to changes in the prior distribution (Edwards, Lindman, and Savage 1963). We use the standard normal gamma conjugate class of priors,

$$\boldsymbol{\beta} \sim N(\boldsymbol{\mu}, \sigma^2\mathbf{V})$$

and

$$\frac{\nu\lambda}{\sigma^2} \sim X^2_\nu.$$

Here ν, λ, the $(p+1) \times (p+1)$ matrix $\mathbf{V}$, and the $(p+1)$ vector $\boldsymbol{\mu}$ are hyperparameters to be chosen.

The marginal likelihood for $\mathbf{Y}$ under a model M_i based on the proper priors described earlier is given by

$$\begin{aligned} p(\mathbf{Y}|\mu_i, \mathbf{V}_i, \mathbf{X}_i, M_i) &= \frac{\Gamma\left(\frac{\nu+n}{2}\right)(\nu\lambda)^{\nu/2}}{\pi^{n/2}\Gamma\left(\frac{\nu}{2}\right)|I + \mathbf{X}_i\mathbf{V}_i\mathbf{X}_i^t|^{1/2}} \\ &\quad \times [\lambda\nu + (\mathbf{Y} - \mathbf{X}_i\mu_i)^t \\ &\qquad \times (I + \mathbf{X}_i\mathbf{V}_i\mathbf{X}_i^t)^{-1}(\mathbf{Y} - \mathbf{X}_i\mu_i)]^{-(\nu+n)/2}, \end{aligned} \tag{5}$$

where $\mathbf{X}_i$ is the design matrix and $\mathbf{V}_i$ is the covariance matrix for $\boldsymbol{\beta}$ corresponding to model M_i (Raiffa and Schlaifer 1961). The Bayes factor for M_0 versus M_1, the ratio of

Equation (5) for $i = 0$ and $i = 1$, is then given by

$$B_{01} = \left(\frac{|I + \mathbf{X}_1\mathbf{V}_1\mathbf{X}_1^t|}{|I + \mathbf{X}_0\mathbf{V}_0\mathbf{X}_0^t|}\right)^{1/2} \left[\frac{a_0}{a_1}\right]^{-(\nu+n)/2}, \quad (6)$$

where $a_i = \lambda\nu + (\mathbf{Y} - \mathbf{X}_i\mu_i)^t(I + \mathbf{X}_i\mathbf{V}_i\mathbf{X}_i^t)^{-1}(\mathbf{Y} - \mathbf{X}_i\mu_i)$, $i = 0, 1$.

3.2 Selection of Prior Distributions

The Bayesian framework described earlier gives the BMA user the flexibility to modify the prior setup as desired. In this section we describe the prior distribution setup that we adopt in our examples below.

For noncategorical predictor variables, we assume the individual β's to be independent a priori. We center the distribution of $\boldsymbol{\beta}$ on zero (apart from β_0) and choose $\boldsymbol{\mu} = (\hat{\beta}_0, 0, 0, \ldots, 0)$, where $\hat{\beta}_0$ is the ordinary least squares estimate of β_0. The covariance matrix $\mathbf{V}$ is equal to σ^2 multiplied by a diagonal matrix with entries $(s_Y^2, \phi^2 s_1^{-2}, \phi^2 s_2^{-2}, \ldots, \phi^2 s_p^{-2})$, where s_Y^2 denotes the sample variance of $\mathbf{Y}$, s_i^2 denotes the sample variance of $\mathbf{X}_i$ for $i = 1, \ldots, p$, and ϕ is a hyperparameter to be chosen. The prior variance of β_0 is chosen conservatively and represents an upper bound on the reasonable variance for this parameter. The variances of the remaining $\boldsymbol{\beta}$ parameters are chosen to reflect increasing precision about each β_i as the variance of the corresponding $\mathbf{X}_i$ increases and to be invariant to scale changes in both the predictor variables and the response variable.

For a categorical predictor variable $\mathbf{X}_i$ with $(c + 1)$ possible outcomes ($c \geq 2$), the Bayes factor should be invariant to the selection of the corresponding dummy variables $(X_{i1}, \ldots, X_{ic})$. To this end, we set the prior variance of $(\beta_{i1}, \ldots, \beta_{ic})$ equal to $\sigma^2\phi^2[(1/n)\mathbf{X}^{i^T}\mathbf{X}^i]^{-1}$, where $\mathbf{X}^i$ is the $n \times c$ design matrix for the dummy variables, where each dummy variable has been centered by subtracting its sample mean. This is related to the g prior of Zellner (1986). The complete prior covariance matrix for $\boldsymbol{\beta}$ is now given by

$$\mathbf{V}(\boldsymbol{\beta}) = \sigma^2 \begin{pmatrix} s_Y^2 & & & \\ & \phi^2 s_1^{-2} & & \\ & & \ddots & \\ & & & \phi^2 s_{i-1}^{-2} \\ & & & & \phi^2\left(\frac{1}{n}\mathbf{X}^{i^T}\mathbf{X}^i\right)^{-1} & & \\ & & & & & \phi^2 s_{i+1}^{-2} & \\ & & & & & & \ddots \\ & & & & & & & \phi^2 s_p^{-2} \end{pmatrix}.$$

To choose the remaining hyperparameters ν, λ, and ϕ, we define a number of reasonable desiderata and attempt to satisfy them. In what follows we assume that all the variables have been standardized to have mean zero and sample variance 1. We would like the following desiderata to hold

1. The prior density $\Pr(\beta_1, \ldots, \beta_p)$ is reasonably flat over the unit hypercube $[-1, 1]^p$.
2. $\Pr(\sigma^2)$ is reasonably flat over $(a, 1)$ for some small a.
3. $\Pr(\sigma^2 \leq 1)$ is large.

The order of importance of these desiderata is roughly the order in which they are listed. More formally, we maximize $\Pr(\sigma^2 \leq 1)$ subject to the following:

a. $\Pr(\beta_1 = 0, \ldots, \beta_p = 0)/\Pr(\beta_1 = 1, \ldots, \beta_p = 1) \leq K_1$. (Following Jeffreys (1961), we choose $K_1 = \sqrt{10}$.)
b. $\{\max_{a<\sigma^2<1} / \Pr(\sigma^2)\} \Pr(\sigma^2 = a) \leq K_2$.
c. $\{\max_{a<\sigma^2<1} / \Pr(\sigma^2)\} \Pr(\sigma^2 = 1) \leq K_2$.

Because desideratum 2 is less important than desideratum 1, we have chosen $K_2 = 10$.

For $a = .05$, this yields $\nu = 2.58, \lambda = .28$, and $\phi = 2.85$. For this set of hyperparameters, $\Pr(\sigma^2 \leq 1) = .81$. We use these settings of the hyperparameters in the examples that follow.

To compare our prior for $\beta_i, i = 1, \ldots, p$ for a noncategorical predictor with the actual distribution of coefficients from real data, we collected 13 datasets from several regression textbooks (see App. A). Figure 1 shows a histogram of the 100 coefficients from the standardized data plotted with the prior distribution resulting from the hyperparameters that we use. As desired, the prior density is relatively flat over the range of observed values.

4. TWO APPROACHES TO BAYESIAN MODEL AVERAGING

4.1 Occam's Window

Our first method for accounting for model uncertainty starting from Equation (1) involves applying the Occam's window algorithm of Madigan and Raftery (1994) to linear regression models. Two basic principles underly this ad hoc approach.

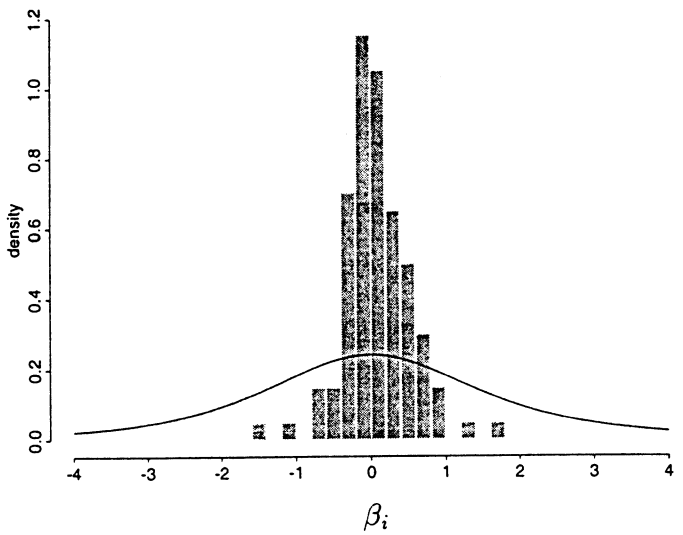

Figure 1. Histogram of 100 Coefficients from Standardized Data, from 13 Textbook Datasets. The solid line is the prior density for β_i, $i = 1, \ldots, p$.

First, if a model predicts the data far less well than the model that provides the best predictions, then it has effectively been discredited and should no longer be considered. Thus models not belonging to

$$\mathcal{A}' = \left\{ M_k: \frac{\max_l \{\Pr(M_l|D)\}}{\Pr(M_k|D)} \leq C \right\} \tag{7}$$

should be excluded from Equation (1), where C is chosen by the data analyst and $\max_l\{\Pr(M_l|D)\}$ denotes the model with the highest posterior model probability. In the examples that follow we use $C = 20$. The number of models in Occam's window increases as the value of C decreases.

Second, appealing to Occam's razor, we exclude models that receive less support from the data than any of their simpler submodels. More formally, we also exclude from (1) models belonging to

$$\mathcal{B} = \left\{ M_k: \exists M_l \in \mathcal{M}, M_l \subset M_k, \frac{\Pr(M_l|D)}{\Pr(M_k|D)} > 1 \right\}. \tag{8}$$

Equation (1) is then replaced by

$$\Pr(\Delta|D) = \frac{\sum_{M_k \in \mathcal{A}} \Pr(\Delta|M_k, D) \Pr(D|M_k) \Pr(M_k)}{\sum_{M_k \in \mathcal{A}} \Pr(D|M_k) \Pr(M_k)}, \tag{9}$$

where

$$\mathcal{A} = \mathcal{A}' \setminus \mathcal{B} \in \mathcal{M}. \tag{10}$$

This greatly reduces the number of models in the sum in Equation (1), and now all that is required is a search strategy

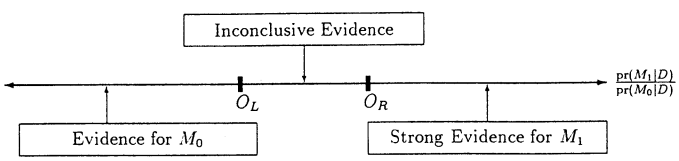

Figure 2. Occam's Window: Interpreting the Posterior Odds for Nested Models.

to identify the models in $\mathcal{A}$. Two further principles underly the search strategy. The first principle—Occam's window—concerns interpretating the ratio of posterior model probabilities $\Pr(M_1|D)/\Pr(M_0|D)$. Here M_0 is a model with one less predictor than M_1. The essential idea is shown in Figure 2. If there is evidence for M_0 then M_1 is rejected, but to reject M_0 we require strong evidence for the larger model, M_1. If the evidence is inconclusive (falling in Occam's window), then neither model is rejected. The second principle is that if M_0 is rejected, then so are all of the models nested within it.

These principles fully define the strategy. Typically, in our experience, the number of terms in (1) is reduced to fewer than 25, often to as few as 1 or 2. Madigan and Raftery (1994) provided a detailed description of the algorithm and showed how averaging over the selected models provides better predictive performance than basing inference on a single model in each of the examples that they considered.

4.2 Markov Chain Monte Carlo Model Composition

Our second approach is to approximate (1) using a Markov chain Monte Carlo (MCMC) approach (see, e.g., Smith and Roberts 1993). For our application, we adopt the MCMC model composition (MC^3) methodology of Madigan and York (1995), which generates a stochastic process that moves through model space. We can construct a Markov chain $\{M(t), t = 1, 2, \ldots\}$ with state space $\mathcal{M}$ and equilibrium distribution $\Pr(M_i|D)$. If we simulate this Markov chain for $t = 1, \ldots, N$, then under certain regularity conditions, for any function $g(M_i)$ defined on $\mathcal{M}$, the average

$$\hat{G} = \frac{1}{N} \sum_{t=1}^{N} g(M(t)) \tag{11}$$

converges almost surely to $E(g(M))$ as $N \to \infty$ (Smith and Roberts 1993). To compute (1) in this fashion, set $g(M) = \Pr(\Delta|M, D)$.

To construct the Markov chain, we define a neighborhood $\text{nbd}(M)$ for each $M \in \mathcal{M}$ that consists of the model M itself and the set of models with either one variable more or one variable fewer than M. Define a transition matrix $\mathbf{q}$ by setting $\mathbf{q}(M \to M') = 0$ for all $M' \notin \text{nbd}(M)$ and $\mathbf{q}(M \to M')$ constant for all $M' \in \text{nbd}(M)$. If the chain is currently in state M, then we proceed by drawing M' from $\mathbf{q}(M \to M')$. It is then accepted with probability

$$\min\left\{1, \frac{\Pr(M'|D)}{\Pr(M|D)}\right\}.$$

Otherwise, the state stays in state M. Madigan and York (1995) described MC^3 for discrete graphical models. Software for implementing the MC^3 algorithm is described in the Appendix.

5. MODEL UNCERTAINTY AND PREDICTION

5.1 Example: Crime and Punishment

5.1.1 Crime and Punishment: Overview. Up to the 1960s, criminal behavior was traditionally viewed as deviant and linked to the offender's presumed exceptional psychological, social, or family circumstances (Taft and England 1964). Becker (1968) and Stigler (1970) argued that on the contrary, the decision to engage in criminal activity is a rational choice determined by its costs and benefits relative to other (legitimate) opportunities.

In an influential article, Ehrlich (1973) developed this argument theoretically, specified it mathematically, and tested it empirically using aggregate data from 47 U.S. states in 1960. Errors in Ehrlich's empirical analysis were corrected by Vandaele (1978), who gave the corrected data, which we use here (see also Cox and Snell 1982). (Ehrlich's study has been much criticized (see, e.g., Brier and Fienberg 1980), and we cite it here for purely illustrative purposes. For economy of expression, we use causal language and speak of "effects," even though the validity of this language for these data is dubious. Because people, not states, commit crimes, these data may reflect aggregation bias.)

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

We thus have 15 candidate predictors of crime rate (Table 4), and so potentially $2^{15} = 32,768$ different models. As in the original analyses, all data were transformed logarithmically. Standard diagnostic checking (see, e.g., Draper and Smith 1981) did not reveal any gross violations of the assumptions underlying normal linear regression.

Ehrlich's analysis concentrated on the relationship between crime rate and predictors 14 and 15 (probability of imprisonment and average time served in state prisons). In his original analysis, Ehrlich (1973) focused on two regression models, consisting of the predictors (9, 12, 13, 14, 15) and (1, 6, 9, 10, 12, 13, 14, 15), which were chosen in advance based on theoretical grounds.

To compare Ehrlich's results with models that might be selected using standard techniques, we chose three popular variable selection techniques: Efroymson's stepwise method (Miller 1990), minimum Mallow's C_p, and maximum adjusted R^2 (Weisberg 1985). Efroymson's stepwise method is like forward selection except that when a new variable is added to the subset, partial correlations are considered to see whether any of the variables currently in the subset should be dropped. Similar hybrid methods are found in most standard statistical computer packages. Problems with stepwise regression, Mallow's C_p, and adjusted R^2 are well known (see, e.g., Weisberg 1985).

Table 1 displays the results from the full model with all 15 predictors, three models selected using standard variable selection techniques, and the two models chosen by Ehrlich on theoretical grounds. The three models chosen using variable selection techniques (models 2, 3, 4) share many of the same variables and have high values of R^2. Ehrlich's theoretically chosen models fit the data less well. There are striking differences—indeed, conflicts—between the results from the different models. Even the models chosen using statistical techniques lead to conflicting conclusions about the main questions of interest, despite the models' superficial similarity.

Consider first the predictor for probability of imprisonment, $\mathbf{X}_{14}$. This is a significant predictor in all six models, so interest focuses on estimating the size of its effect. To aid interpretation, recall that all variables have been transformed logarithmically, so that when all other predictors are held fixed, $\beta_{14} = -.30$ means roughly that a 10% increase in the probability of imprisonment produces a 3% reduction in the crime rate. The estimates of β_{14} fluctuate wildly between models. The stepwise regression model gives an estimate about one-third lower in absolute value than the full model, enough to be of policy importance; this difference is equal to about 1.7 standard errors. The Ehrlich models give estimates that are about one-half higher than the full model, and more than twice as big as those from stepwise regression (in absolute value). There is clearly considerable model uncertainty about this parameter.

Table 1. Models Selected for Crime Data

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

NOTE: P_{15} is the p value from a two-sided t test for testing $\beta_{15} = 0$. For the stepwise procedure, $F = 3.84$ was used for the F-to-enter and F-to-delete value. This corresponds approximately to the 5% level

Table 2. Crime Data: Occam's Window Posterior Model Probabilities

Model											Posterior model probability (%)
1		3	4			9	11		13	14	12.6
1		3	4				11		13	14	9.0
1		3	4			9			13	14	8.4
1		3		5		9	11		13	14	8.0
		3	4		8	9			13	14	7.6
1		3	4						13	14	6.3
1		3	4				11		13		5.8
1		3		5			11		13	14	5.7
1		3	4						13		4.9
1		3		5		9			13	14	4.8
		3		5	8	9			13	14	4.4
		3	4			9			13	14	4.1
		3		5		9			13	14	3.6
1		3		5					13	14	3.5
	2	3	4						13	14	2.0
1		3		5			11		13		1.9
		3	4						13	14	1.6
		3		5					13	14	1.6
		3	4						13		1.4
1		3		5					13		1.4
		3		5					13		.7
1			4					12	13		.7

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

Together these results paint a confused picture about β_{14} and β_{15}. Later we argue that the confusion can be resolved by taking explicit account of model uncertainty.

5.1.2 Crime and Punishment: Model Averaging. For the model averaging strategies, we assumed that all possible combinations of predictors were equally likely a priori. To implement Occam's window, we started from the null model and used the "up" algorithm only (see Madigan and Raftery 1994). The selected models and their posterior model probabilities are shown in Table 2. The models with posterior model probabilities of 1.2% or larger as indicated by MC^3 are shown in Table 3. In total, 1,772 different models were visited during 30,000 iterations of MC^3. Occam's window chose 22 models in this example, clearly indicating model uncertainty. Choosing any one model and making inferences as if it were the "true" model ignores model uncertainty. In the next section we further explore the consequences of basing inferences on a single model.

The top models indicated by the two methods (Tables 2 and 3) are quite similar. The posterior probabilities are normalized over all selected models for Occam's window and over all possible combinations of the 15 predictors for MC^3. So the posterior probabilities for the same models differ across the model averaging method, but this has little effect on the relationship between the models as measured by the Bayes factor.

Table 4 shows the posterior probability that the coefficient for each predictor does not equal 0—that is, $\Pr(\beta_i \neq 0|D)$—obtained by summing the posterior model probabilities across models for each predictor. The results from Occam's window and MC^3 are fairly close for most of the predictors. Predictors with high $\Pr(\beta_i \neq 0|D)$ include proportion of young males, mean years of schooling, police expenditure, income inequality, and probability of imprisonment.

Comparing the two models analyzed by Ehrlich (1973), consisting of the predictors (9, 12, 13, 14, 15) and (1, 6, 9, 10, 12, 13, 14, 15), with the results in Table 4, we see that several predictors included in Ehrlich's analysis receive little support from the data. The estimated $\Pr(\beta_i \neq 0|D)$ is quite small for predictors 6, 10, 12, and 15. Two predictors (3 and 4) have empirical support but were not included by Ehrlich. Indeed, Ehrlich's two selected models have very low posterior probabilities.

Ehrlich's work attracted attention primarily because of his conclusion that both the probability of imprisonment (predictor 14) and the average prison term (predictor 15) influenced the crime rate. The posterior distributions for the coefficients of these predictors, based on the model averaging results of MC^3, are shown in Figures 3 and 4. The MC^3 posterior distribution for β_{14} is indeed centered away from 0, with a small spike at 0 corresponding to $P(\beta_{14} = 0|D)$. The posterior distribution for β_{14} based on Occam's window is quite similar. The spike at 0 is an artifact of our approach, in which it is possible to consider models with a predictor fully removed from the model. This is in contrast to the practice of setting the predictor close to 0 with high probability (as in George and McCulloch 1993). In contrast to Figure 3, the MC^3 posterior distribution for the coefficient corresponding to average prison term is centered close to 0 and has a large spike at 0 (Fig. 4). Occam's window indicates a spike at 0 only, or no support for inclusion of this predictor. By averaging over all models, our results indicate support for a relationship between crime rate and predictor 14, but not predictor 15. Our model averaging results are consistent with those of Ehrlich for the probability of imprisonment, but not for the average prison term.

Table 3. Crime Data: MC^3, Models With Posterior Model Probabilities of 1.2% or Larger

Model										Posterior model probability (%)
1	3	4			9	11	13	14		2.6
1	3	4				11	13	14		1.8
1	3	4			9		13	14		1.7
1	3	4	5		9		13	14		1.6
1	3	4			9	11	13	14	15	1.6
1	3	4			9		13	14	15	1.6
	3	4		8	9		13	14		1.5
1	3	4					13	14		1.3
1	3	4				11	13			1.2
1	3		5			11	13	14		1.2

Table 4. Crime Data: $Pr(\beta_i \neq 0|D)$, Expressed as a Percentage

Predictor number	Predictor	Occam's window	MC^3	Ehrlich's models	
1	Percentage of males age 14–24	73	79		*
2	Indicator variable for southern state	2	17		
3	Mean years of schooling	99	98		
4	Police expenditure in 1960	64	72		
5	Police expenditure in 1959	36	50		
6	Labor force participation rate	0	6		*
7	Number of males per 1,000 females	0	7		
8	State population	12	23		
9	Number of nonwhites per 1,000 people	53	62	*	*
10	Unemployment rate of urban males age 14–24	0	11		*
11	Unemployment rate of urban males, age 35–39	43	45		
12	Wealth	1	30	*	*
13	Income inequality	100	100	*	*
14	Probability of imprisonment	83	83	*	*
15	Average time served in state prisons	0	22	*	*

NOTE: The last column indicates the predictors included in the two models considered by Ehrlich.
* Corresponds to Ehrlich model 1 and * corresponds to Ehrlich model 2.

Among the variables that measure the expected benefits from crime, Ehrlich concluded that both wealth and income inequality had an effect; we found this to be true for income inequality but not for wealth. For the predictors that represent the payoff from legitimate activities, Ehrlich found the effects of variables 1, 6, 10, and 11 to be unclear; he did not include mean schooling in his model. We found strong evidence for the effect of some of these variables, notably the percent of young males and mean schooling, but the effects of unemployment and labor force participation are either unproven or unlikely. Finally, the "control" variables that have no theoretical basis (2, 7, 8) turned out, satisfyingly, to have no empirical support either.

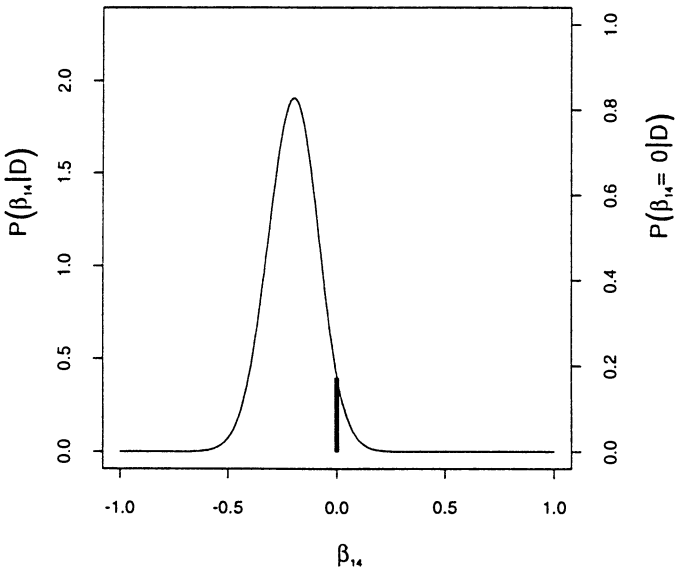

Figure 3. Posterior Distribution for β_{14}, the Coefficient for the Predictor "Probability of Imprisonment," Based on the MC^3 Model Average. The spike corresponds to $P(\beta_{14} = 0|D)$. The vertical axis on the left corresponds to the posterior distribution for β_{14}, and the vertical axis on the right corresponds to the posterior distribution for β_{14} equal to zero. The density is scaled so that the maximum of the density is equal to $P(\beta_{14} \neq 0|D)$ on the right axis.

The model averaging results for the predictors for police expenditures lead to an interesting interpretation. Police expenditure was measured in two successive years, and the measures are highly correlated ($r = .993$). The data show clearly that the 1960 crime rate is associated with police expenditures, and that only one of the two measures ($\mathbf{X}_4$ and $\mathbf{X}_5$) is needed, but they do not say for sure which measure should be used. Each model in Occam's window contains one predictor or the other, but not both. For both Occam's window and MC^3 $\Pr[(\beta_4 \neq 0) \cup (\beta_5 \neq 0)|D] = 1$, so the data provide very strong evidence for an association with police expenditures.

In summary, we found strong support for some of Ehrlich's conclusions but not for others. In particular, by averaging over all models, our results indicate support for a relationship between crime rate and probability of imprisonment, but not for average time served in state prisons.

5.1.3 Crime and Punishment: Assessment of Predictive Performance. We use the predictive ability of the selected models for future observations to measure the effectiveness of a model selection strategy. Our specific objective is to compare the quality of the predictions based on model averaging with the quality of predictions based on any single model that an analyst might reasonably have selected.

To measure performance, we randomly split the complete dataset into two subsets. Other percentage splits can be adopted. A 50-50 split was chosen here, so that each portion would contain enough data to be a representative sample. We ran Occam's window and MC^3 using half of the data. This set is called the *training set,* D^{T}. We evaluated performance using the *prediction set,* made up of the remaining half of the data, $D^{\mathrm{P}} = D \backslash D^{\mathrm{T}}$. Within this framework, we assessed predictive performance using numerical and graphical measures of performance.

Predictive coverage was measured using the proportion of observations in the performance set that fall in the corresponding 90% prediction interval. For both Occam's window and MC^3, 80% of the observations in the performance set fell in the 90% prediction intervals over the averaged

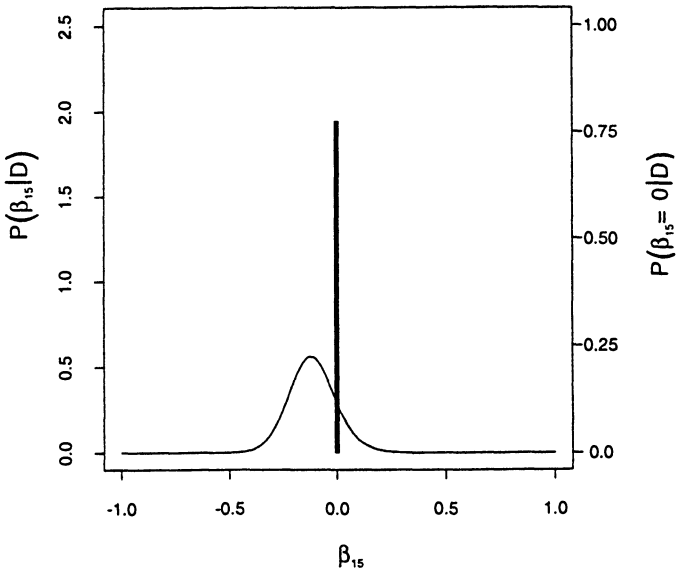

Figure 4. Posterior Distribution for β_{15}, the Coefficient for the Predictor "Average Time Served in State Prisons," Based on the Model Average Over a Large Set of Models From MC^3. See Figure 3.

models (Table 5). David Draper (personal communication) suggested that BMA falls somewhat short of nominal coverage here because aspects of model uncertainty other than model selection have not been assessed. In Hoeting, Raftery, and Madigan (1995, 1996), we extended BMA to account for uncertainty in the selection of transformations and in the identification of outliers.

For comparison with other standard variable selection techniques, we used the three popular variable selection procedures discussed earlier to select two or three "best" models. The models that we chose using these methods are given in Table 5. All of the individual models chosen using standard techniques performed considerably worse than the model averaging approaches, with prediction coverage ranging from 58% to 67%. Thus the model averaging strategies improved predictive coverage substantially as compared to any single model that might reasonably have been chosen.

A sensitivity analysis for priors chosen within the framework described in Section 3.2 indicates that the results for Occam's window and MC^3 are not highly sensitive to the choice of prior. The results for Occam's window and MC^3 using three different sets of priors were quite similar.

In an attempt to provide a graphical measure of predictive performance, we used a "calibration plot" to determine whether the predictions were well calibrated. A model is well calibrated if, for example, 70% of the observations in the test dataset are less than or equal to the 70th percentile of the posterior predictive distribution. The calibration plot shows the degree of calibration for different models, with the posterior predictive probability on the x-axis and the percentage of observed data less than or equal to the posterior predictive probability on the y-axis. In a calibration plot, perfect calibration is the 45-degree line; the closer a model's calibration line to the 45-degree line, the better calibrated the model. The calibration plot is similar to reliability diagrams used to assess probability forecasts (see, e.g., Murphy and Winkler 1977). The calibration plot for the model chosen by stepwise selection and for model averaging using Occam's window is shown in Figure 5. The shaded area in Figure 5 shows where the model averaging strategy produces predictions that are better calibrated than predictions from the model chosen by the stepwise model selection procedure. The calibration plot for MC^3 is similar.

These performance measures support our claim that conditioning on a single selected model ignores model uncertainty, which in turn leads to the underestimation of uncertainty when making inferences about quantities of interest. Model averaging leads to better-calibrated predictive distributions.

5.2 Simulated Examples: Predictive Performance

In the foregoing example, the true answer is unknown. To further demonstrate the usefulness of BMA, we use several simulated examples. In our examples, we follow the format of George and McCulloch (1993).

Example 5.2.1. In this example we investigate the impact of model averaging on predictive performance when there is little model uncertainty. For the training set, we simulated $p = 15$ predictors and $n = 50$ observations as independent standard normal vectors. We generated the re-

Table 5. Crime Data: Performance Comparison

Method	Model												Predictive coverage (%)
MC^3	Model averaging												80
Occam's window	Model averaging												80
Stepwise (5%)			3	4				9			13		67
Adjusted R^2 (2)	1	2	3	4	5		8		11	12	13	15	67
Adjusted R^2 (3)	1	2	3	4	5	6	8		11	12	13	15	67
Stepwise (15%)			3	4			8	9			13	15	63
C_p (2)	1	2	3	4					11		13		63
Adjusted R^2 (1)	1	2	3	4	5				11	12	13	15	58
C_p (1)	1	2	3	4					11		13	15	58
C_p (3)	1	2	3	4					11	12	13	15	58

NOTE: Predictive coverage is the percentage of observations in the performance set that fall in the 90% prediction interval. Method numbers correspond to the ith model chosen using the given model selection method. For example, C_p (1) is the first model chosen using the C_p method. The percentage values shown for the stepwise procedures correspond to the significance levels for the F-to-enter and F-to-delete values. For example, $F = 3.84$ corresponds approximately to the 5% level.

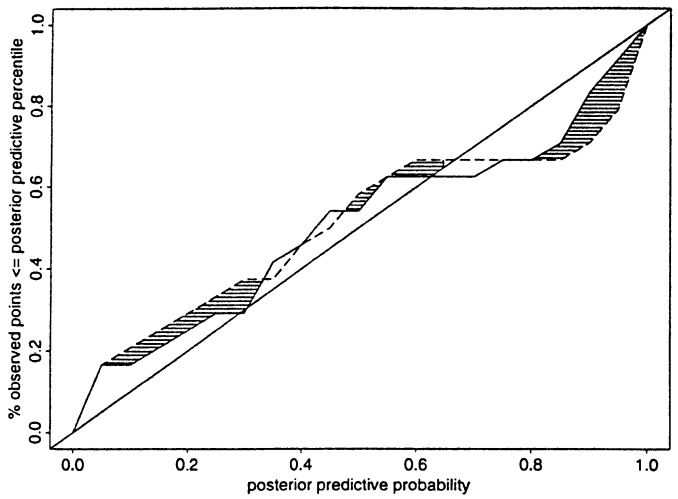

Figure 5. Crime Data: Calibration Plot. The solid line denotes model averaging (Occam's window); the dashed line, predictors 3, 4, 8, 9, 13, 15 (stepwise).

sponse using the model

$$\mathbf{Y} = \mathbf{X}_4 + \mathbf{X}_5 + \varepsilon, \tag{12}$$

where $\varepsilon \sim N_{50}(0, \sigma^2)$ with $\sigma = 2.5$. Least squares estimates for these data are given in Table 6. There is little model uncertainty in this example; only the p values for β_4 and β_5 were smaller than .1. We generated 50 additional observations in the same manner to create the prediction set.

In this example the true model, the model averaging techniques, and models selected using standard techniques all have poor predictive coverage (Table 7). It is slightly encouraging that BMA performs better than the true model, but the improvement is too small to be significant. This and other similar examples that we simulated show that when there is very little model uncertainty, predictive performance is not significantly improved by model averaging.

Example 5.2.2. This example demonstrates the performance of BMA when a subset of the predictors is correlated. For the training set, we simulated $p = 15$ predictors and $n = 50$ observations. We obtained predictors 1–10 as independent standard normal vectors, $\mathbf{X}_1, \ldots, \mathbf{X}_{10}$ iid $\sim N(0, 1)$, and generated predictors 11–15 using the framework

$$[\mathbf{X}_{11}, \ldots, \mathbf{X}_{15}] = [\mathbf{X}_1, \ldots, \mathbf{X}_5]([.3, .5, .7, .9, 1.1]^{\mathrm{T}}[11111]) + \varepsilon,$$

where $\varepsilon \sim N(0, 1)$. We generated the response using the model

$$\mathbf{Y} = \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3 + \mathbf{X}_4 + \mathbf{X}_5 + \varepsilon \tag{13}$$

where $\varepsilon \sim N_{50}(0, \sigma^2)$ with $\sigma = 2.5$. Least squares estimates for these data are given in Table 8. The correlation structure resulted in moderate pairwise correlation between predictors 1–5 and 11–15 ($\mathrm{corr}(\mathbf{X}_1, \mathbf{X}_{11}) = .39$, $\mathrm{corr}(\mathbf{X}_2, \mathbf{X}_{12}) = .41$, $\mathrm{corr}(\mathbf{X}_3, \mathbf{X}_{13}) = .56$, $\mathrm{corr}(\mathbf{X}_4, \mathbf{X}_{14}) = .71$, $\mathrm{corr}(\mathbf{X}_5, \mathbf{X}_{15}) = .69$) and small pairwise correlations elsewhere (median correlation equal to $-.02$). We generated 50 additional observations in the same manner to create the prediction set.

Table 9 shows that in this example, model averaging has better predictive performance than any single model that might have been selected. In this example, the poor performance of the true model and the other single models selected using standard techniques demonstrate that model uncertainty can strongly influence predictive performance.

6. SUCCESSFUL IDENTIFICATION OF THE NULL MODEL

Linear regression models are frequently used even when little is known about the relationship between the predictors and the response. When there is a weak relationship between the predictors and the response, the overall F statistic will be small and thus the null hypothesis that the null model is true fails to be rejected. However, many data analysts perform model selection regardless of the F statistic value for the overall model. Problems can then occur, as subsequent model selection techniques often choose a model that includes a subset of the predictors. Freedman (1983) has shown that in the extreme case where there is *no* relationship between the predictors and the response variable, omitting the predictors with the smallest t values (e.g., $p > .25$) can result in a model with a highly significant F statistic and high R^2. In contrast, if the response and predictors are independent, Occam's window typically indicates the null model only, or the null model as one of a small number of "best" models.

Following Freedman (1983), we generated 5,100 independent observations from a standard normal distribution to create a matrix with 100 rows and 51 columns. The first column was taken to be the dependent variable in a regression equation, and the other 50 columns were taken to be the predictors. Thus the predictors are independent of the response by construction. For the entire dataset, the multiple regression results were as follows:

- $R^2 = .55$ and $p = .29$.
- 18 coefficients out of 50 were significant at the .25 level.
- 4 coefficients out of 50 were significant at the .05 level.

We used three different variable selection procedures on the simulated data. The first of these was the method used

Table 6. Least Squares Estimates for Example 5.2.1 ($\hat{\sigma} = 2.9$)

	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}
β	0	0	0	0	1.00	1.00	0	0	0	0	0	0	0	0	0	0
$\hat{\beta}$	.42	.21	.40	.07	.95	1.72	.20	.34	−.32	.24	−.15	.6	−.45	−.08	.20	.18
$\hat{\sigma}_{\beta}$	.46	.55	.56	.36	.52	.47	.39	.58	.49	.45	.44	.55	.48	.52	.45	.47

Table 7. Performance Comparison for Example 5.2.1: Predictive Coverage for a 90% Prediction Interval

Method	Model	Predictive coverage (%)
BMA (estimated coverage)	Model averaging	72
Occam's window	Model averaging	70
Adjusted R^2 (3)	2 4 5 8 11	70
C_p (3)	4 5 11	70
True model and stepwise (5%)	4 5	68
Stepwise (15%) and C_p (2)	2 4 5	68
C_p (1)	4 5	68
Adjusted R^2 (2)	2 4 5 10 11	68
Adjusted R^2 (1)	2 4 5 11	66

NOTE: Predictive coverage for BMA (all models) is estimated using the 371 models with posterior model probabilities greater than .0001; see Table 5.

by Freedman (1983), in which all predictors with p values of .25 or lower were included in a second pass over the data. The results from this method were as follows:

- $R^2 = .40$ and $p = .0003$.
- 17 coefficients out of 18 were significant at the .25 level.
- 10 coefficients out of 18 were significant at the .05 level.

These results are highly misleading, as they indicate a definite relationship between the response and the predictors, whereas in fact the data are all noise.

The second model selection method used on the full dataset was Efroymson's stepwise method. This indicated a model with 15 predictors with the following results:

- $R^2 = .40$, and $p = .0001$.
- All 15 predictors were significant at the .25 level.
- 10 coefficients out of 15 were significant at the .05 level.

Again a model is chosen that misleadingly appears to have a great deal of explanatory power.

The third variable selection method that we used was Occam's window. The only model chosen by this method was the null model.

We repeated the foregoing procedure 10 times with similar results. In five simulations, Occam's window chose only the null model. For the remaining simulations, three models or fewer were chosen along with the null model. All the nonnull models chosen had R^2 values less than .15. For all of the simulations, the selection procedure used by Freedman (1983) and the stepwise method chose models with many predictors and highly significant R^2 values.

At best, Occam's window correctly indicates that the null model is the only model that should be chosen when there is no signal in the data. At worst, Occam's window chooses the null model along with several other models. The presence of the null model among those chosen by Occam's window should indicate to a researcher the possibility of evidence for a lack of signal in the data that he or she is analyzing.

To examine the possibility that our Bayesian approach favors parsimony to the extent that Occam's window finds no signal even when one exists, we did an additional simulation study. We generated 3,000 observations from a standard normal distribution to create a dataset with 100 observations and 30 candidate predictors. We allowed the response $\mathbf{Y}$ to depend only on $\mathbf{X}_1$, where $\mathbf{Y} = .5\mathbf{X}_1 + \varepsilon$ with $\varepsilon \sim N(0, .75)$. Thus $\mathbf{Y}$ still has unit variance, and the "true" R^2 for the model equals .20.

For this simulated data, Occam's window contained one model only—the correct model with $\mathbf{X}_1$. In contrast, the screening method used by Freedman produced a model with six predictors, including $\mathbf{X}_1$, with four of these significant at the .1 level. Stepwise regression indicated a model with two predictors, including $\mathbf{X}_1$, both of them significant at the .025 level. So the two standard variable selection methods indicated evidence for variables that in fact were not at all associated with the dependent variable, whereas Occam's window chose the correct model.

These examples provide evidence that Occam's window overcomes the problem of selection of the null model when there is no signal in the data.

Table 8. Least Squares Estimates for Example 5.2.2 ($\hat{\sigma}$ = 2.21)

	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}
β	0	1.00	1.00	1.00	1.00	1.00	0	0	0	0	0	0	0	0	0	0
$\hat{\beta}$	.12	.80	1.07	1.03	−.18	.55	−.67	.28	−.11	.31	.29	.11	−.09	−.39	.73	−.96
$\hat{\sigma}_\beta$	.38	.49	.41	.45	.53	.58	.37	.41	.49	.33	.34	.40	.32	.35	.37	.40

Table 9. Performance Comparison for Example 5.2.2: Predictive Coverage for a 90% Prediction Interval

Method	Model	Predictive coverage (%)
MC^3	Model averaging	92
Occam's window	Model averaging	86
Stepwise (5% and 15%)	1 2 3 4	80
True model	1 2 3 4 5	78
C_p (2)	1 2 3 6 13 14 15	72
Adjusted R^2 (1)		
C_p (3)	1 2 3 6 10 14 15	72
Adjusted R^2 (3)	1 2 3 6 7 13 14 15	72
C_p (1)	1 2 3 5 14 15	70
Adjusted R^2 (2)	1 2 3 6 10 13 14 15	70

NOTE: Predictive coverage for BMA (all models) is estimated using the 1,014 models with posterior model probabilities greater than .00005; see Table 5.

7. DISCUSSION

7.1 Related Work

Draper (1995) has also addressed the problem of assessing model uncertainty. Draper's approach is based on the idea of *model expansion;* that is, starting with a single reasonable model chosen by a data-analytic search, expanding model space to include those models suggested by context or other considerations, and then averaging over this model class. Draper did not directly address the problem of model uncertainty in variable selection. However, one could consider Occam's window to be a practical implementation of model expansion.

George and McCulloch (1993) have developed the stochastic search variable selection (SSVS) method, which is similar in spirit to MC^3. They defined a Markov chain that moves through model space and parameter space at the same time. Their method never actually removes a predictor from the full model, but only sets it close to zero with high probability. Our approach avoids this by integrating analytically over parameter space.

We have focused here on Bayesian solutions to the model uncertainty problem. Very little has been written about frequentist solutions to the problem. Perhaps the most obvious frequentist solution is to bootstrap the entire data analysis, including model selection. However, Freedman et al. (1986) have shown that this does not necessarily give a satisfactory solution to the problem.

7.2 Conclusions

The prior distribution of the covariance matrix for β described in Section 3.2 depends on the actual data, including both the dependent and the independent variables. A similar data-dependent approach to the assessment of the priors was used by Raftery (1996). Although at first this may appear to be contrary to the idea of a prior, our objective was to develop priors that lead to posteriors similar to those of a person with little prior information. Examples analyzed to date suggest that we achieved this objective. The priors for β lead to a reasonable prior variance and result in conclusions that are not highly sensitive to the choice of hyperparameters. Thus the data dependence does not appear to be a drawback.

In a strict sense, our data-dependent priors do not correspond to a Bayesian subjective prior. Our priors might be considered to be an approximation to a true Bayesian subjective prior and might be appropriate when little prior information is available. We have followed other authors, including George and McCullough (1993), Laud and Ibrahim (1995), and Zellner (1986), in referring to our approach as Bayesian.

The choice of which procedure to use—Occam's window or MC^3—will depend on the particular application. Occam's window will be most useful when one is interested in making inferences about the relationships between the variables. Occam's window also tends to be much faster computationally. MC^3 is the better procedure to choose if the goal is good predictions or if the posterior distribution of some quantity is of more interest than the nature of the "true" model and if computer time is not a critical consideration. However, each approach is flexible enough to be used successfully for both inference *and* prediction.

We have described two procedures that can be used to account for model uncertainty in variable selection for linear regression models. In addition to variable selection, uncertainty is also involved in the identification of outliers and in the choice of transformations in regression. To broaden the flexibility of our current procedures, and to improve our ability to account for model uncertainty, we have extended BMA to include transformation selection and outlier identification in work reported elsewhere (Hoeting et al. 1995, 1996).

APPENDIX A: DATA FOR FIGURE 1

The following data from selected textbooks were used to make Figure 1:

Dataset	Source	Page number	Number of observations	Number of predictors
Attitude survey	Chatterjee and Price (1991)	70	30	6
Equal education oppurtunity	Chatterjee and Price (1991)	176	70	3
Gasoline mileage	Chatterjee and Price (1991)	261	30	10
Nuclear power	Cox and Snell (1982)	81	32	10
Crime	Cox and Snell (1982)	170	47	13
Hald	Draper and Smith (1981)	630	13	4
Grades	Hamilton (1993)	83	118	3
Swiss fertility	Mosteller and Tukey (1977)	550	47	5
Surgical unit	Neter, Wasserman and Kutner (1990)	439, 468	108	4
Berkeley study	Weisberg (1985)			
Girls		56	32	10
Boys		57	26	10
Housing	Weisberg (1985)	241	27	9
Highway	Weisberg (1985)	206	39	13

APPENDIX B: SOFTWARE FOR IMPLEMENTING MC^3

BMA is a set of S-PLUS functions that can be obtained free of charge via the World Wide Web address http://lib.stat.cmu.edu/S/bma or by sending an e-mail message containing the text "send BMA from S" to the Internet address statlib@stat.cmu.edu.

The program MC3.REG performs MCMC model composition for linear regression. The set of programs fully implements the MC^3 algorithm described in Section 4.2.

[Received November 1993. Revised June 1996.]

REFERENCES

Becker, G. S. (1968), "Crime and Punishment: An Economic Approach," *Journal of Political Economy*, 76, 169–217.

Brier, S. S., and Fienberg, S. E. (1980), "Recent Econometric Modeling of Crime and Punishment: Support for the Deterrence Hypothesis?," *Evaluation Review*, 4, 147–191.

Breiman, L. (1968), *Probability*, Reading, MA: Addison-Wesley.

——— (1992), "The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: X-Fixed Prediction Error," *Journal of the American Statistical Association*, 87, 738–754.

——— (1995), "Better Subset Regression Using the Nonnegative Garrote," *Technometrics*, 37, 373–384.

Breiman, L., and Spector, P. (1992), "Submodel Selection and Evaluation in Regression," *International Statistical Review*, 60, 291–319.

Chatterjee, S., and Price, B. (1991), *Regression Analysis by Example* (2nd ed.), New York: Wiley.

Cox, D. R., and Snell, E. J. (1982), *Applied Statistics: Principles and Examples*, New York: Chapman and Hall.

Chung, K. L. (1967), *Markov Chains with Stationary Transition Probabilities* (2nd ed.), Berlin: Springer-Verlag.

Draper, D. (1995), "Assessment and Propagation of Model Uncertainty" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 57, 45–97.

Draper, N. R., and Smith, H. (1981), *Applied Regression Analysis* (2nd ed.), New York: Wiley.

Edwards, W., Lindman, H., and Savage, L. J. (1963), "Bayesian Statistical Inference for Psychological Research," *Psychological Review*, 70, 193–242.

Ehrlich, I. (1973), "Participation in Illegitimate Activities: A Theoretical and Empirical Investigation," *Journal of Political Economy*, 81, 521–565.

Freedman, D. A. (1983), "A Note on Screening Regression Equations," *The American Statistician*, 37, 152–155.

Freedman, D. A., Navidi, W. C., and Peters, S. C. (1986), "On the Impact of Variable Selection in Fitting Regression Equations," in *On Model Uncertainty and Its Statistical Implications*, ed. T. K. Dijkstra, Berlin: Springer-Verlag, pp. 1–16.

Garthwaite, P. H., and Dickey, J. M. (1992), "Elicitation of Prior Distributions for Variable Selection Problems in Regression," *The Annals of Statistics*, 20, 1697–1719.

Geisser, S. (1980), Discussion of "Sampling and Bayes' Inference in Scientific Modelling and Robustness" by G. E. P. Box, *Journal of the Royal Statistical Society*, Ser. A, 143, 416–417.

George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–890.

Good, I. J. (1952), "Rational Decisions," *Journal of the Royal Statistical Society*, Ser. B, 14, 107–114.

Hamilton, L. C. (1993), *Statistics With Stata 3*, Belmont, CA: Duxbury Press.

Hocking, R. R. (1976), "The Analysis and Selection of Variables in Linear Regression," *Biometrics*, 32, 1–51.

Hodges, J. S. (1987), "Uncertainty, Policy Analysis, and Statistics," *Statistical Science*, 2, 259–291.

Hoeting, J. A., Raftery, A. E., and Madigan, D. (1995), "Simultaneous Variable and Transformation Selection in Linear Regression," Technical Report 9506, Colorado State University, Dept. of Statistics.

——— (1996), "A Method for Simultaneous Variable Selection and Outlier Identification in Linear Regression," *Journal of Computational Statistics and Data Analysis*, 22, 251–270.

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

Kadane, J. B., Dickey, J. M., Winkler, R. L., Smith, W. S., and Peters, S. C. (1980), "Interactive Elicitation of Opinion for a Normal Linear Model," *Journal of the American Statistical Association*, 75, 845–854.

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Laud, P. W., and Ibrahim, J. G. (1995), "Predictive Model Selection," *Journal of the Royal Statistical Society*, Ser. B, 57, 247–262.

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

Linhart, H., and Zucchini, W. (1986), *Model Selection*, New York: Wiley.

Madigan, D., and Raftery, A. E. (1994), "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window," *Journal of the American Statistical Association*, 89, 1535–1546.

Madigan, D., and York, J. (1995), "Bayesian Graphical Models for Discrete Data," *International Statistical Review*, 63, 215–232.

Miller, A. J. (1984), "Selection of Subsets of Regression Variables" (with discussion), *Journal of the Royal Statistical Society*, Ser. A, 147, 389–425.

——— (1990), *Subset Selection in Regression*, New York: Chapman and Hall.

Mitchell, T. J., and Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression" (with discussion), *Journal of the American Statistical Association*, 83, 1023–1036.

Mosteller, F., and Tukey, J. W. (1977), *Data Analysis and Regression*, Reading, MA: Addison-Wesley.

Moulton, B. R. (1991), "A Bayesian Approach to Regression Selection

and Estimation With Application to a Price Index for Radio Services," *Journal of Econometrics*, 49, 169–193.

Murphy, A. H., and Winkler, R. L. (1977), "Reliability of Subjective Probability Forecasts of Precipitation and Temperature," *Applied Statistics*, 26, 41–47.

Neter, J., Wasserman, W., and Kutner, M. (1990), *Applied Linear Statistical Models*, Homewood, IL: Irwin.

Raftery, A. E. (1988), "Approximate Bayes Factors for Generalized Linear Models," Technical Report 121, University of Washington, Dept. of Statistics.

——— (1996), "Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models," *Biometrika*, 83, 251–266.

Raiffa, H., and Schlaifer, R. (1961), *Applied Statistical Decision Theory*, Cambridge, MA: MIT Press.

Regal, R., and Hook, E. B. (1991), "The Effects of Model Selection on Confidence Intervals for the Size of a Closed Population," *Statistics in Medicine*, 10, 717–721.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Shibata, R. (1981), "An Optimal Selection of Regression Variables," *Biometrika*, 68, 45–54.

Smith, A. F. M., and Roberts, G. O. (1993), "Bayesian Computation via Gibbs Sampler and Related Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society*, Ser. B, 55, 3–24.

Stewart, L. (1987), "Hierarchical Bayesian Analysis Using Monte Carlo Integration: Computing Posterior Distributions When There are Many Possible Models," *The Statistician*, 36, 211–219.

Stewart, L., and Davis, W. W. (1986), "Bayesian Posterior Distributions Over Sets of Possible Models With Inferences Computed by Monte Carlo Integration," *The Statistician*, 35, 175–182.

Stigler, G. J. (1970), "The Optimum Enforcement of Laws," *Journal of Political Economy*, 78, 526–536.

Taft, D. R., and England, R. W. (1964), *Criminology* (4th ed.), New York: Macmillan.

Vandaele, W. (1978), "Participation in Illegitimate Activities; Ehrlich Revisited," in *Deterrence and Incapacitation* (eds. A. Blumstein, J. Cohen, and D. Nagin), Washington, D.C.: National Academy of Sciences Press, pp. 270–335.

Weisberg, S. (1985), *Applied Linear Regression* (2nd ed.), New York: Wiley.

Zellner, A. (1986), "On Assessing Prior Distributions and Bayesian Regression Analysis With g Prior Distributions," in *Bayesian Inference and Decision Techniques—Essays in Honor of Bruno de Finetti*, eds. P. K. Goel and A. Zellner, Amsterdam: North-Holland, pp. 233–243.