

# Model Selection and Accounting for Model Uncertainty in Linear Regression Models

Adrian Raftery, David Madigan and Jennifer Hoeting  
University of Washington <sup>1</sup>

November 19, 1993

<sup>1</sup>Adrian E. Raftery is Professor of Statistics and Sociology, David Madigan is Assistant Professor of Statistics, and Jennifer Hoeting is a Ph.D. Candidate, all at the Department of Statistics, GN-22, University of Washington, Seattle, WA 98195. The research of Raftery and Hoeting was supported by ONR Contract N-00014-91-J-1074. Madigan's research was partially supported by NSF grant no. DMS 92111627. The authors are grateful to Danika Lew for research assistance.

## Abstract

We consider the problems of variable selection and 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. The complete Bayesian solution to this problem involves averaging over all possible models when making inferences about quantities of interest. This approach is often not practical. In this paper we offer two alternative approaches. First we describe a Bayesian model selection algorithm called “Occam’s Window” which involves averaging over a reduced set of models. Second, we describe a Markov chain Monte Carlo approach which directly approximates the exact solution. Both these model averaging procedures provide better predictive performance than any single model which might reasonably have been selected.

In the extreme case where there are many candidate predictors but there is 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. We refer to this unfortunate phenomenon as “Freedman’s Paradox” (Freedman, 1983). In this situation, Occam’s Window usually indicates the null model as the only one to be considered, or else a small number of models including the null model, thus largely resolving the paradox.

**Key Words:** Bayes factor; Freedman’s Paradox; Markov chain Monte Carlo model composition; Model uncertainty; Occam’s Window; Posterior model probability.

# Contents

- 1 Introduction** **1**
  
- 2 Methodology** **2**
  - 2.1 Accounting for Model Uncertainty . . . . . 2
  - 2.2 Bayesian Framework and Selection of Prior Distributions . . . . . 3
  - 2.3 Model Selection using Occam’s Window . . . . . 6
  - 2.4 Markov Chain Monte Carlo Model Composition . . . . . 7
  
- 3 Model uncertainty and prediction** **8**
  - 3.1 Example: Crime and Punishment . . . . . 8
  - 3.2 Assessment of Predictive Performance . . . . . 11
  
- 4 Freedman’s Paradox Resolved** **15**
  
- 5 Discussion** **18**
  - 5.1 Related Work . . . . . 18
  - 5.2 Conclusions . . . . . 18
  
- A Data for Figure 1** **20**
  
- B The Up-Down Algorithm** **20**
  
- References** **22**

# 1 Introduction

The selection of 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  $Y$  and a set of a candidate predictors  $X_1, X_2, \dots, X_k$ , find the “best” model of the form

$$Y = \beta_0 + \sum_{j=1}^p \beta_j X_j + \epsilon$$

where  $X_1, X_2, \dots, X_p$  is a subset of  $X_1, X_2, \dots, X_k$ .

In this work we embed this model selection problem in the larger framework of accounting for model uncertainty. We argue that conditioning on a single selected model ignores model uncertainty, and that this, in turn, leads to the underestimation of uncertainty when making inferences about quantities of interest. A complete Bayesian solution to this problem involves averaging over *all* possible models when making inferences about quantities of interest. Indeed, this approach provides optimal predictive ability (Madigan and Raftery, 1994). In many applications however, this averaging will not be a practical proposition and 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.” Appealing to scientific norms, this approach involves averaging over a reduced set of models and allows for effective communication of real model uncertainty to the analyst. Second, we directly approximate the complete solution by applying the Markov chain Monte Carlo approach of Madigan and York (1993) to linear regression models. In this approach the posterior distribution of a quantity of interest is approximated by a Markov chain Monte Carlo method which generates a process that moves through model space. We show in an example that both these model averaging approaches provide better predictive performance than any single model which 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 with a high  $R^2$  and a highly significant overall  $F$  value, a phenomenon we refer to as “Freedman’s Paradox”. By contrast, when a data set 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 paradox.

The background literature for our approach includes several areas of research, namely the selection of subsets of predictor variables in linear regression models (Hocking, 1976; Draper

and Smith, 1981; Linhart and Zucchini, 1986; Mitchell and Beauchamp, 1988; Miller, 1990; George and McCulloch, 1993) and model uncertainty (Raftery, 1993; Madigan and Raftery, 1994; Madigan and York, 1993; Kass and Raftery, 1993; Draper, 1994).

In the next section we outline the philosophy underlying our approach, describe how we selected prior distributions, and outline the two model averaging approaches. In Section 3 we provide an example and describe our assessment of predictive performance. In Section 4 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. In Section 5 we discuss related work and suggest future directions.

## 2 Methodology

### 2.1 Accounting for Model Uncertainty

A typical approach to data analysis is to carry out a model selection exercise leading to a single “best” model and to then make inference as if the selected model were the true model. However, this ignores a major component of uncertainty, namely uncertainty about the model itself (Leamer, 1978; Hodges, 1987; Raftery, 1988, 1993; Moulton, 1991; Draper, 1994). As a consequence, uncertainty about quantities of interest can be underestimated. For striking examples of this see Miller (1984), Regal and Hook (1991), Madigan and York (1993), Raftery (1993), and Kass and Raftery (1993).

There is a standard Bayesian solution to this problem. If  $\mathcal{M} = \{M_1, \dots, M_k\}$  denotes the set of all models being considered and if  $\Delta$  is the quantity of interest such as a future observation or the utility of a course of action, then the posterior distribution of  $\Delta$  given the data  $D$  is

$$\text{pr}(\Delta | D) = \sum_{k=1}^K \text{pr}(\Delta | M_k, D)\text{pr}(M_k | D). \quad (1)$$

This is an average of the posterior distributions under each model weighted by the corresponding posterior model probabilities. In equation (1), the posterior probability of model  $M_k$  is given by

$$\text{pr}(M_k | D) = \frac{\text{pr}(D | M_k)\text{pr}(M_k)}{\sum_{l=1}^K \text{pr}(D | M_l)\text{pr}(M_l)}, \quad (2)$$

where

$$\text{pr}(D | M_k) = \int \text{pr}(D | \theta_k, M_k)\text{pr}(\theta_k | M_k)d\theta_k, \quad (3)$$

is the marginal likelihood of model  $M_k$ ,  $\theta_k$  is the vector of parameters of model  $M_k$ ,  $\text{pr}(\theta_k | M_k)$  is the prior density of  $\theta_k$  under model  $M_k$ ,  $\text{pr}(D | \theta_k, M_k)$  is the likelihood, and  $\text{pr}(M_k)$

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.

Implementation of the above strategy 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 what follows, we present workable solutions to both of these problems.

## 2.2 Bayesian Framework and Selection of Prior Distributions

Each model we consider is of the form:

$$Y = \beta_0 + \sum_{j=1}^p \beta_j X_j + \epsilon = X\beta + \epsilon$$

where the observed data on the predictors are contained in the  $n \times (p+1)$  matrix  $X$  and the observed data on the dependent variable are contained in the  $n$ -vector  $Y$ . We assign to  $\epsilon$  a normal distribution with mean 0 and variance  $\sigma^2$  and assume that the  $\epsilon$ 's in distinct cases are independent. We consider the  $(p+1)$  parameter vector  $\beta$  and  $\sigma^2$  to be unknown.

Where possible, informative prior distributions for  $\beta$  and  $\sigma^2$  should be elicited and incorporated into the analysis—see Kadane *et al.* (1980) and Garthwaite and Dickey (1992). In the absence of expert opinion we seek to choose prior distributions which reflect uncertainty about the parameters and also embody reasonable *a priori* constraints. We use the standard normal-gamma conjugate class of priors,

$$\beta \sim N(\mu, \sigma^2 V),$$

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

Here  $\nu$ ,  $\lambda$ , the  $(p+1) \times (p+1)$  matrix  $V$  and the  $(p+1)$ -vector  $\mu$  are hyperparameters to be chosen.

For non-categorical predictor variables we assume the individual  $\beta$ 's to be independent *a priori*. We center the distribution of  $\beta$  on zero (apart from  $\beta_0$ ) and choose  $\mu = (\hat{\beta}_0, 0, 0, \dots, 0)$  where  $\hat{\beta}_0$  is the ordinary least squares estimate of  $\beta_0$ . The covariance matrix  $V$  is diagonal with entries  $(s_Y^2, \phi^2 s_1^{-2}, \phi^2 s_2^{-2}, \dots, \phi^2 s_p^{-2})$  where  $s_Y^2$  denotes the sample variance of  $Y$ ,  $s_i^2$  denotes the sample variance of  $X_i$  for  $i = 1, \dots, p$ , and  $\phi$  is a hyperparameter to be chosen. The prior variance of  $\beta_0$  is chosen conservatively and represents an upper bound on the reasonable variance for this parameter. The variances of the remaining  $\beta$ -parameters are chosen to reflect increasing precision about each  $\beta_i$  as the variance of the corresponding  $X_i$  increases and to be invariant to scale changes in both the predictor variables and the response variable.



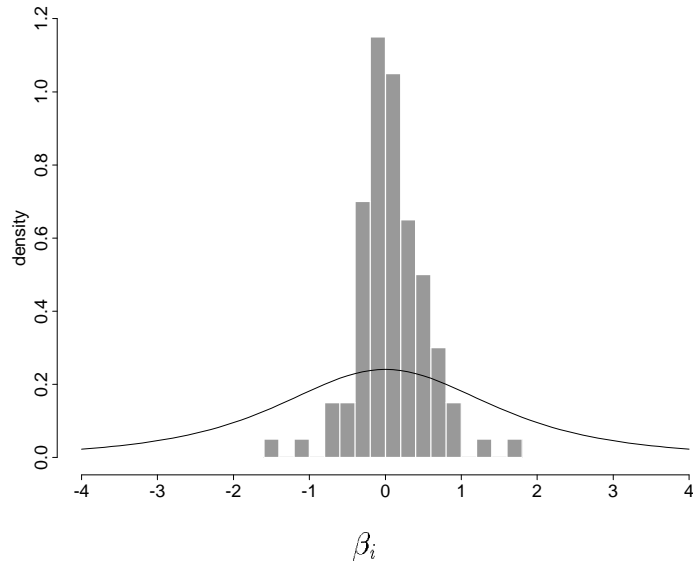


Figure 1: Histogram of 100 coefficients from standardized data, from 13 textbook data sets. The solid line is the prior density for  $\beta_i$ ,  $i = 1, \dots, p$ .

To compare our prior for  $\beta_i$ ,  $i = 1, \dots, p$  for a non-categorical predictor with the actual distribution of coefficients from real data, 13 data sets from several regression textbooks were collected (Appendix A). A histogram of the 100 coefficients from the standardized data plotted with the prior distribution resulting from the hyperparameters we use in this paper is shown in Figure 1. As desired, the prior density is relatively flat over the range of observed values.

The marginal likelihood for  $Y$  under a model  $M_i$  based on the proper priors described above is given by

$$p(Y|\mu_i, V_i, X_i, M_i) = \frac{\Gamma(\frac{\nu+n}{2})(\nu\lambda)^{\frac{\nu}{2}}}{\pi^{\frac{n}{2}}\Gamma(\frac{\nu}{2})|I + X_i V_i X_i^t|^{\frac{1}{2}}} \left[ \lambda\nu + (Y - X_i \mu_i)^t (I + X_i V_i X_i^t)^{-1} (Y - X_i \mu_i) \right]^{-\frac{(\nu+n)}{2}} \quad (4)$$

where  $X_i$  is the design matrix and  $V_i$  is the covariance matrix for  $\beta$  corresponding to model  $M_i$  (Raiffa and Schlaifer, 1961). The Bayes factor for  $M_0$  versus  $M_1$ , the ratio of equation (4) for  $i = 0$  and  $i = 1$ , is then given by

$$B_{01} = \left( \frac{|I + X_1 V_1 X_1^t|}{|I + X_0 V_0 X_0^t|} \right)^{\frac{1}{2}} \left[ \frac{\lambda\nu + (Y - X_0\mu_0)^t(I + X_0 V_0 X_0^t)^{-1}(Y - X_0\mu_0)}{\lambda\nu + (Y - X_1\mu_1)^t(I + X_1 V_1 X_1^t)^{-1}(Y - X_1\mu_1)} \right]^{-\frac{(\nu+n)}{2}}. \quad (5)$$

### 2.3 Model Selection using Occam's Window

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

$$\mathcal{A}' = \left\{ M_k : \frac{\max_l \{\text{pr}(M_l | D)\}}{\text{pr}(M_k | D)} \leq C \right\}, \quad (6)$$

should be excluded from equation (1) where  $C$  is chosen by that data analyst. In the examples we used  $C = 20$ . Second, appealing to Occam's razor, we exclude models which receive less support from the data than any of their simpler submodels. More formally we also exclude from (1) models belonging to:

$$\mathcal{B} = \left\{ M_k : \exists M_l \in \mathcal{A}, M_l \subset M_k, \frac{\text{pr}(M_l | D)}{\text{pr}(M_k | D)} > 1 \right\}. \quad (7)$$

Equation (1) is then replaced by

$$\text{pr}(\Delta | D) = \frac{\sum_{M_k \in \mathcal{A}} \text{pr}(\Delta | M_k, D) \text{pr}(D | M_k) \text{pr}(M_k)}{\sum_{M_k \in \mathcal{A}} \text{pr}(D | M_k) \text{pr}(M_k)} \quad (8)$$

where

$$\mathcal{A} = \mathcal{A}' \setminus \mathcal{B}. \quad (9)$$

This greatly reduces the number of models in the sum in equation (1) and now all that is required is a search strategy to identify the models in  $\mathcal{A}$ . Two further principles underly the search strategy. First, if a model is rejected then all its submodels are rejected. The second principle — “Occam's Window” — concerns the interpretation of the ratio of posterior model probabilities  $\text{pr}(M_1 | D) / \text{pr}(M_0 | D)$ . 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,

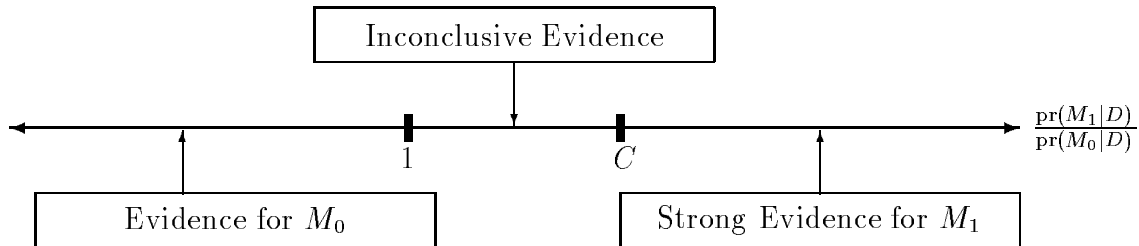


Figure 2: Occam’s Window: Interpreting the posterior odds for nested models.

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) neither model is rejected.

These principles fully define the strategy. Typically, in our experience, the number of terms in (1) is reduced to fewer than 25, and often to as few as one or two. Madigan and Raftery (1994) provide a detailed description of the algorithm and show how averaging over the selected models provides better predictive performance than basing inference on a single model in each of the examples they consider. The algorithm is reproduced in Appendix B.

## 2.4 Markov Chain Monte Carlo Model Composition

Our second approach is to approximate (1) using the Markov chain Monte Carlo model composition (MC<sup>3</sup>) approach of Madigan and York (1993). MC<sup>3</sup> generates a stochastic process which moves through model space. Specifically, let  $\mathcal{M}$  denote the space of models under consideration. We can construct a Markov chain  $\{M(t), t = 1, 2, \dots\}$  with state space  $\mathcal{M}$  and equilibrium distribution  $\text{pr}(M_i | D)$ . If we simulate this Markov chain for  $t = 1, \dots, N$ , then under certain regularity conditions, for any function  $g(M_i)$  defined on  $\mathcal{M}$ , the average:

$$\hat{G} = \frac{1}{N} \sum_{t=1}^N g(M(t)) \quad (10)$$

is a simulation-consistent estimate of  $E(g(M))$  (Smith and Roberts, 1993). To compute (1) in this fashion set  $g(M) = \text{pr}(\Delta | M, D)$ .

To construct the Markov chain we define a neighborhood  $\text{nbnd}(M)$  for each  $M \in \mathcal{M}$  which 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  $q$  by setting  $q(M \rightarrow M') = 0$  for all  $M' \notin \text{nbnd}(M)$  and  $q(M \rightarrow M')$  constant for all  $M' \in \text{nbnd}(M)$ . If the chain is currently in

state  $M$ , we proceed by drawing  $M'$  from  $q(M \rightarrow M')$ . It is then accepted with probability:

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

Otherwise the state stays in state  $M$ . MC<sup>3</sup> for discrete graphical models is described in Madigan and York (1993).

### 3 Model uncertainty and prediction

#### 3.1 Example: Crime and Punishment

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, on the contrary, that the decision to engage in criminal activity is a rational choice determined by its costs and benefits relative to other (legitimate) opportunities.

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 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. The benefits of crime are related to both the aggregate wealth and income inequality in the surrounding community. The expected net payoff from alternative legitimate activities is related to educational level and the availability of employment, the latter being measured by the unemployment and labor force participation rates. This payoff was expected to be lower (in 1960) for nonwhites and for young males than for others, so that states with high proportions of these were expected also to have higher crime rates. Ehrlich also raised the possibility of including the sex ratio and an indicator variable for southern states as explanatory variables, but the theoretical rationale for this is unclear.

We thus have 15 candidate predictors of crime rate (Table 3). As in the original analyses, all data were transformed logarithmically. Standard diagnostic checking (e.g. Draper and Smith, 1981) did not reveal any gross violations of the assumptions underlying normal linear regression. All possible models were assumed to be equally likely *a priori*.

To implement Occam’s Window, we started from the null model and used the “Up” algorithm only. The selected models and their posterior model probabilities are shown in

Table 1. The models with posterior model probabilities of 1.2% or larger as indicated by MC<sup>3</sup> are shown in Table 2. In total, 1772 different models were visited during 30,000 iterations of MC<sup>3</sup>. 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. The consequences of basing inferences on a single model will be explored in the next section.

Table 3 shows the posterior probability that the coefficient for each predictor does not equal 0, i.e.,  $\Pr(\beta_i \neq 0|D)$ , obtained by summing the posterior model probabilities across models for each predictor. The results from Occam’s Window and MC<sup>3</sup> are fairly close for most of the predictors. There are several predictors with high  $\Pr(\beta_i \neq 0|D)$  including the proportion of young males, mean years of schooling, police expenditure, income inequality, and probability of imprisonment.

In his original analysis of this data set, Ehrlich (1973) analyzed two regression models, consisting of the predictors (9, 12, 13, 14, 15) and (1, 6, 9, 10, 12, 13, 14, 15), respectively. Comparing these models with the results in Table 3, we see that there are several predictors included in Ehrlich’s analysis that receive little support from the data. The estimated  $\Pr(\beta_i \neq 0|D)$  is quite small for predictors 6, 10, 12, and 15. There are also variables for which there is empirical support but which Ehrlich did not include (3 and 4). 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 and the average prison term reduced the crime rate. Our results are consistent with this for the probability of imprisonment, but not for the average prison term. We also found evidence for an association between crime rate and police expenditures, net of other variables.

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.

In summary, we found strong support for some of Ehrlich’s predictions but not for others. It seems hard to reconcile the data with Ehrlich’s overall economic theory of crime.

Table 1: 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	0.7
1, 4, 12, 13	0.7

Table 2: Crime data: MC<sup>3</sup> Posterior Model Probabilities

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, 5, 9, 11, 13, 14	1.6
1, 3, 4, 9, 11, 13, 14, 15	1.6
1, 3, 4, 9, 13, 14	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

### 3.2 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 data set into two subsets. We ran Occam’s Window and MC<sup>3</sup> 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,  $D^P = D \setminus D^T$ . Within this framework, we assess predictive performance using two strategies.

The first measure of predictive ability is the logarithmic scoring rule of Good (1952) which is based on the conditional predictive ordinate (Geisser, 1980). Specifically, we measured the predictive ability of an individual model,  $M$ , with:

$$- \sum_{d \in D^P} \log \text{pr}(d \mid M, D^T).$$

We measured the predictive performance for model averaging with:

$$- \sum_{d \in D^P} \log \left\{ \sum_{M \in \mathcal{A}} \text{pr}(d \mid M, D^T) \text{pr}(M \mid D^T) \right\},$$

Table 3: Crime data:  $\Pr(\beta_i \neq 0|D)$ , expressed as a percentage

Predictor number	Predictor	Occam's Window	MC <sup>3</sup>
1	percent of males 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 1000 females	0	7
8	state population	12	23
9	number of nonwhites per 1000 people	53	62
10	unemployment rate of urban males 14-24	0	11
11	unemployment rate of urban males 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

Table 4: Crime data: Occam’s Window Predictive Performance. The models were selected and the posterior model probabilities calculated from half the data, and the log predictive score was calculated from the other half.

Model	Posterior model probability %	Log predictive score
1, 2, 3, 4, 11, 13	34.1	42.4
3, 4, 9, 13	30.3	42.6
1, 2, 3, 4, 10, 13	9.8	44.3
3, 5, 9, 13	8.8	38.5
1, 2, 3, 4, 13	7.9	39.2
2, 3, 4, 13	5.9	40.2
1, 2, 3, 5, 11, 13	3.3	37.0
model averaging		31.4

where for Occam’s Window  $\mathcal{A}$  is the set of selected models and for  $MC^3$   $\mathcal{A}$  is the set of visited models. We give results for Occam’s Window and  $MC^3$  in Tables 4 and 5. Table 4 shows the seven models selected by Occam’s Window. Table 5 shows models with posterior model probabilities of 0.9% and greater. However, all 2293 of the models visited during  $MC^3$  were included in the computation of the model averaging score. For both methods, model averaging had considerably better predictive ability than any individual model with high posterior probability. For example,  $MC^3$  outperformed the most likely model *a posteriori* by 13 points, or by 26 points on the scale of twice the log probability on which deviances are measured. The performance of Occam’s Window was almost as good.

For comparison with other standard variable selection techniques, three standard variable selection procedures were used to select a “best” model, Efroymsen’s stepwise method (Miller, 1990), minimum Mallows’ Cp and maximum adjusted  $R^2$ . Efroymsen’s stepwise method is like forward selection except that when a new variable is added to the subset, partial correlations are considered to see if any of the variables currently in the subset should be dropped. Similar hybrid methods are found in most standard statistical computer packages. The models chosen using these methods are given in Table 6. All three standard methods indicate models that have log predictive scores that are larger than the scores for both model averaging procedures. Thus the model averaging strategies predictively out-perform models

Table 5: Crime data: MC<sup>3</sup> Predictive Performance. (See Table 4).

Model	Posterior model probability %	Log predictive score
1, 2, 3, 4, 11, 13	3.1	42.4
3, 4, 9, 13	2.7	42.6
1, 2, 3, 4, 5, 11, 13	1.7	42.2
3, 4, 5, 9, 13	1.4	42.2
1, 2, 3, 4, 11, 13, 15	1.3	45.2
1, 2, 3, 4, 9, 11, 13	1.1	38.7
1, 2, 3, 4, 11, 12, 13	1.0	41.3
1, 3, 4, 11, 13	0.9	35.1
1, 2, 3, 4, 10, 13	0.9	44.3
2, 3, 4, 11, 13	0.9	42.8
3, 4, 7, 9, 13	0.9	46.6
model averaging		29.3

chosen using the standard techniques.

For other random splits of the data, different models were selected by Occam’s Window and MC<sup>3</sup>, but the log predictive scores for the model averaging strategies were very similar across the random splits. For these other random splits, MC<sup>3</sup> typically had better predictive performance than Occam’s Window as measured by the log predictive score.

A sensitivity analysis for priors chosen within the framework described in Section 2.2 indicates that the results for Occam’s Window and MC<sup>3</sup> are not highly sensitive to the choice of prior. The results for Occam’s Window and MC<sup>3</sup> using 3 different sets of priors were quite similar. In all 3 cases, the model averaging log predictive scores for Occam’s Window and MC<sup>3</sup> were smaller than the log predictive scores for the models chosen by the 3 standard variable selection procedures (the models shown in Table 6).

In an attempt to provide a more interpretable measure of performance, a graphical method was used to determine if the predictions were well calibrated. A model is well calibrated if, for example, 70% of the observations in the test data set 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 where perfect calibration is the 45° line. The calibration plot is similar to the reliability diagrams used to assess probability forecasts

Table 6: Crime data: Performance comparison

Method	Model	Log predictive score
MC <sup>3</sup>	model averaging	29.3
Occam’s Window	model averaging	31.4
Efroymsen	3, 4, 8, 9, 13, 15	39.1
Adjusted R <sup>2</sup>	1, 2, 3, 4, 5, 11, 12, 13, 15	44.9
C <sub>p</sub>	1, 2, 3, 4, 11, 13, 15	45.2

(see, for example, Murphy and Winkler, 1977). The calibration plot for the model chosen by Efroymsen and for model averaging using Occam’s Window is shown in Figure 3. The shaded area in Figure 3 shows where the model averaging strategy produces predictions that are better calibrated than predictions from the model chosen by Efroymsen’s model selection procedure. The calibration plot for MC<sup>3</sup> is similar. These performance measures support our claim that conditioning on a single selected model ignores model uncertainty which, in turn leads to the underestimating of uncertainty when making inferences about quantities of interest.

## 4 Freedman’s Paradox Resolved

Linear regression models are frequently used even when little is known about the relationship between the predictors and the response. 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.  $p > 0.25$ ) can result in a model with a highly significant F statistic and high R<sup>2</sup>. We will refer to this unfortunate phenomenon as “Freedman’s paradox.” In contrast, if the response and predictors are independent, Occam’s Window typically indicates the null model only, or as one of a small number of “best” models.

Like Freedman (1983), we generated 5100 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 data set, the multiple regression results were as follows:

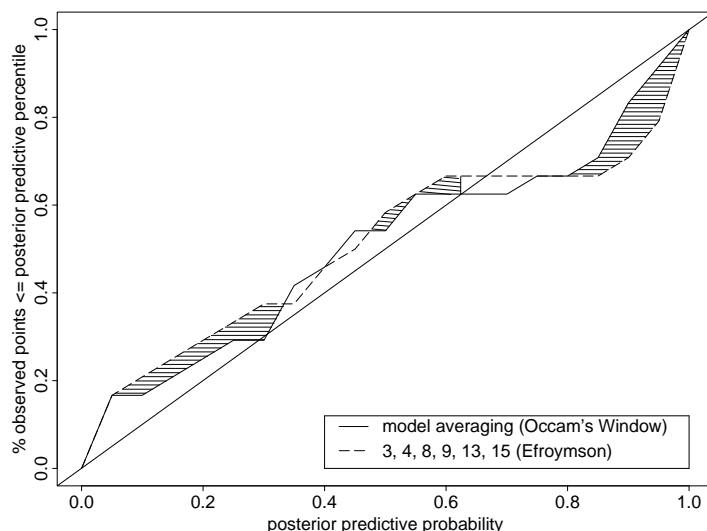


Figure 3: Crime data: Calibration

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

Three different variable selection procedures were used on the simulated data. The first of these was the method used by Freedman (1983) in which all predictors with  $p$ -values of 0.25 or lower were included in a second pass over the data. The results from this method were as follows:

- $R^2 = .40$ ,  $p = .0003$ ;
- 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 data set was Efroymsen's stepwise method. This indicated a model with 15 predictors with the following results:

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

Table 7: Log predictive scores for the Freedman simulated noise data.

Model	Method	Log predictive score
null model	Occam's Window	133
18 predictors	Freedman (1983)	174
15 predictors	Efroymsen	181

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

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

The procedure described above was repeated 10 times with similar results. In 5 simulations, Occam's Window chose only the null model. For the remaining simulations 3 models or fewer were chosen along with the null model. For the non-null models that were chosen, all models had  $R^2$  values less than 0.15. For all of the simulations the selection procedure used by Freedman (1983) and Efroymsen's stepwise method chose models with many predictors and highly significant  $R^2$  values.

To compare the predictive performance of the models chosen by the three methods, another data set with 100 rows and 51 columns was simulated and log predictive scores were calculated (Table 7). The log predictive score for the only model selected by Occam's Window (the null model) is considerably better than the log predictive score for the models chosen by the other two methods. In addition, the mean square predictive error (MSPR) was calculated. The MSPR for Freedman's (1983) method was 1.4 and the MSPR for the Efroymsen model was 1.5 while the MSPR for the null model was 0.9. Thus Occam's Window had considerably greater out-of-sample predictive power than the more standard variable selection methods considered.

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 that there may be evidence for a lack of signal in the data he is analyzing. Thus Occam's Window largely resolves "Freedman's Paradox".

## 5 Discussion

### 5.1 Related Work

Draper (1994) has also addressed the problem of assessing model uncertainty. Draper’s approach is based on the idea of *model expansion*, i.e., starting with a single reasonable model chosen by a data-analytic search, expanding model space to include those models which are suggested by context or other considerations, and then averaging over this model class. Draper does 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<sup>3</sup>. They define a Markov chain which moves through model space and parameter space at the same time. To make the chain irreducible, however, their method never actually removes a predictor from the full model, but only sets it close to zero with high probability. If this probability is very high, the algorithm has convergence difficulties, and if not the results can be hard to interpret. Our approach avoids this problem by integrating analytically over parameter space.

### 5.2 Conclusions

The prior distribution for the covariance matrix for  $\beta$  described in Section 2.2 depends on the actual data, including both the dependent and independent variables. A similar data-dependent approach to the assessment of the priors was used by Raftery (1993). While this may appear at first sight 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 this objective was achieved. The priors for  $\beta$  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.

The choice of which procedure to use — Occam’s Window or MC<sup>3</sup> — 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. In the example of Section 3, the CPU time for Occam’s Window was 1 hour on a SUN Sparc 2 workstation, compared with 16 hours for MC<sup>3</sup>. MC<sup>3</sup> 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.

In this paper 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, there is also uncertainty involved in the identification of outliers and in the choice of transformations in regression. To broaden the flexibility of our current procedures as well as to improve our ability to account for model uncertainty, we are currently working on extending our model averaging strategies to include transformation selection and outlier identification.

## A Data for Figure 1

Table 8: Data from selected textbooks used to make Figure 1.

Data set	Source	page number	number of obser- vations	number of predictors
Attitude Survey	Chatterjee and Price (1991)	70	30	6
Equal Education Opportunity	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

## B The Up-Down Algorithm

The search can proceed in two directions: “Up” from each starting model by adding variables, or “Down” from each starting model by dropping variables. When starting from a model made up of some subset of the variables, we first execute the “Down” algorithm. Then we execute the “Up” algorithm, using the models from the “Down” algorithm as a starting point. Experience to date suggests that the ordering of these operations has little impact on the final set of models. Let  $\mathcal{A}$  and  $\mathcal{C}$  be subsets of model space  $\mathcal{M}$ , where  $\mathcal{A}$  denotes the set of “acceptable” models and  $\mathcal{C}$  denotes the models under consideration. For both algorithms, we begin with  $\mathcal{A} = \emptyset$  and  $\mathcal{C} = \text{set of starting models}$ .

### Down Algorithm

1. Select a model  $M$  from  $\mathcal{C}$
2.  $\mathcal{C} \leftarrow \mathcal{C} - M$  and  $\mathcal{A} \leftarrow \mathcal{A} + M$
3. Select a submodel  $M_0$  of  $M$  by removing a variable from  $M$
4. Compute  $B = \log \frac{\text{pr}(M_0|D)}{\text{pr}(M|D)}$
5. If  $B > O_R$  then  $\mathcal{A} \leftarrow \mathcal{A} - M$  and if  $M_0 \notin \mathcal{C}, \mathcal{C} \leftarrow \mathcal{C} + M_0$
6. If  $O_L \leq B \leq O_R$  then if  $M_0 \notin \mathcal{C}, \mathcal{C} \leftarrow \mathcal{C} + M_0$
7. If there are more submodels of  $M$ , go to 3
8. If  $\mathcal{C} \neq \emptyset$ , go to 1

### Up Algorithm

1. Select a model  $M$  from  $\mathcal{C}$
2.  $\mathcal{C} \leftarrow \mathcal{C} - M$  and  $\mathcal{A} \leftarrow \mathcal{A} + M$
3. Select a supermodel  $M_1$  of  $M$  by adding a variable to  $M$
4. Compute  $B = \log \frac{\text{pr}(M|D)}{\text{pr}(M_1|D)}$
5. If  $B < O_L$  then  $\mathcal{A} \leftarrow \mathcal{A} - M$  and if  $M_1 \notin \mathcal{C}, \mathcal{C} \leftarrow \mathcal{C} + M_1$
6. If  $O_L \leq B \leq O_R$  then if  $M_1 \notin \mathcal{C}, \mathcal{C} \leftarrow \mathcal{C} + M_1$
7. If there are more supermodels of  $M$ , go to 3
8. If  $\mathcal{C} \neq \emptyset$ , go to 1

Upon termination,  $\mathcal{A}$  contains the set of potentially acceptable models. Finally, we remove all the models which satisfy equation (7), where 1 is replaced by  $\exp(O_R)$ , and those models  $M_k$  for which

$$\frac{\max_l \{\text{pr}(M_l | D)\}}{\text{pr}(M_k | D)} > C.$$

$\mathcal{A}$  now contains the acceptable models.

## References

- Becker, G.S. (1968), Crime and punishment: An economic approach. *Journal of Political Economy*, **76**, 169–217.
- Breiman, L. (1968), *Probability*, Addison-Wesley, Reading.
- Chatterjee, S. and Price, B. (1991), *Regression analysis by example*, 2nd edition, New York: Wiley.
- Cox, D. R. and Snell, E. J. (1982), *Applied statistics : principles and examples*, New York: Chapman and Hall.
- Chung, Kai Lai (1967), *Markov Chains with Stationary Transition Probabilities* (2nd ed), Berlin: Springer-Verlag.
- Draper, D. (1994), Assessment and Propagation of Model Uncertainty, *Journal of the Royal Statistical Society B*, to appear.
- Draper, N.R. and Smith, H. (1981), *Applied Regression Analysis*, (2nd. edition), New York: Wiley.
- 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**, No. 2, 152–155.
- Garthwaite, P.H. and Dickey, J.M. (1992), Elicitation of prior distributions for variable-selection problems in regression, *Annals of Statistics*, **20**, No. 4, 1697–1719.
- Geisser, S. (1980), Discussion on Sampling and Bayes' inference in scientific modelling and robustness (by G.E.P. Box), *Journal of the Royal Statistical Society A*, **143**, 416–417.
- George, E.I. and McCulloch, R.E. (1993), Variable selection via Gibbs sampling, *Journal of the American Statistical Society*, **88**, No. 423, 881–890.
- Good, I.J. (1952), Rational Decisions, *Journal of the Royal Statistical Society 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.
- Jeffreys, H. (1961), *Theory of Probability*, (3rd ed.), 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. (1993), Bayes factors and model uncertainty. *Technical Report no. 254*, Department of Statistics, University of Washington.
- 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*, to appear.
- Madigan, D. and York, J. (1993), Bayesian graphical models for discrete data, *Technical Report 259*, Department of Statistics, University of Washington.
- Miller, A.J. (1984), Selection of subsets of regression variables (with Discussion), *Journal of the Royal Statistical Society (Series A)*, **147**, 389–425.
- Miller, A.J. (1990), *Subset Selection in Regression*, New York: Chapman-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 no. 121*, Department of Statistics, University of Washington.
- Raftery, A.E. (1993), Approximate Bayes factors and accounting for model uncertainty in generalized linear models, *Technical Report 255*, Department of Statistics, University of Washington.
- Raiffa, H. and Schlaifer, R. (1961), *Applied Statistical Decision Theory*, Cambridge, MA: The 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.
- 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 Statistics Society B*, **55**, 3–24.
- 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. Blumstein, A., Cohen, J. and Nagin, D.). Washington, D.C.: National Academy of Sciences, 270–335.

Weisberg, S. (1985), *Applied Linear Regression*, 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), North-Holland: Amsterdam. 233–243.