COMMUN. STATIST.—SIMULA., 27(3), 711-734 (1998)

METHODS OF VARIABLE SELECTION IN REGRESSION MODELING

Paul A. Murtaugh

Department of Statistics
Oregon State University
Corvallis, Oregon 97331

Key Words: Bayes information criterion; Mallows' C_p statistic; multiple linear regression; regression tree; stepwise regression

ABSTRACT

Simulation was used to evaluate the performances of several methods of variable selection in regression modeling: stepwise regression based on partial F-tests, stepwise minimization of Mallows' C_p statistic and Schwarz's Bayes Information Criterion (BIC), and regression trees constructed with two kinds of pruning. Five to 25 covariates were generated in multivariate clusters, and responses were obtained from an ordinary linear regression model involving three of the covariates; each data set had 50 observations. The regression-tree approaches were markedly inferior to the other methods in discriminating between informative and noninformative covariates, and their predictions of responses in "new" data sets were much more variable and less accurate than those of the other methods. The F-test, C_p and BIC approaches were similar in their overall frequencies of "correct" decisions about inclusion or exclusion of covariates, with the C_p method leading to the largest models and the BIC

method to the smallest. The three methods were also comparable in their ability to predict "new" observations, with perhaps a tendency for the C_p approach to perform relatively more poorly for large covariate pools. The abilities of all methods to discriminate between informative and noninformative covariates and to predict "new" observations decreased with increasing size of the covariate pool.

1 Introduction

Regression modeling is a ubiquitous tool in science, with interest usually focusing on prediction of some response from measured values of one or more covariates, or on assessment of the strength of association between the response and the covariates. Methods of regression have been well developed for a variety of types of responses (e.g., see Weisberg 1985, McCullagh and Nelder 1989, Kalbfleisch and Prentice 1980).

A challenge in all types of regression is deciding which elements from a large pool of potential covariates are most suitable for inclusion in a predictive model of the response. A variety of methods is available for selecting such subsets. The methods differ in the algorithms used to identify subsets (e.g., forward, backward and stepwise regression, or enumeration of all possible subsets), as well as in the criteria used to judge the goodness-of-fit of candidate models (e.g., F-test, Mallows' C_p, or the Bayes Information Criterion). Inclusion of too few predictors will lead to bias, and inclusion of too many predictors will cause loss of precision in the estimation of regression coefficients and the prediction of new responses. These methods and issues are well discussed in many modern statistics texts (e.g., Draper and Smith 1981, Fisher and van Belle 1993, Ramsey and Schafer 1997), and the area of model selection is a central theme of statistical research (Linhart and Zucchini 1986, Miller 1990, Lehmann 1990, Cox 1990).

Here I use simulation to evaluate several methods of variable selection — three based on ordinary linear regression and two based on regression trees

(Breiman et al. 1984). I focus on the abilities of the different methods to discriminate between meaningful predictors and useless noise variables, and to identify the models that best predict future observations. Some data on the species richness of zooplankton in lakes (Dodson 1992) are used as an example of application of the different methods.

2 Methods

2.1 Simulation of data

In each simulated data set, values of covariates were randomly generated from multivariate distributions. The intent was to mimic the tendency in real data sets for predictors to occur as clusters of related variables. The following mutually-independent sets of predictors were generated:

1. a set of correlated, continuous predictors, $X_1, X_2, \ldots, X_k$, such that each $X_i \sim N(\mu, \tau^2)$ and Cor $(X_i, X_j) = \rho$ for $i \neq j$;
2. another set of predictors generated as above, labeled $W_1, W_2, \ldots, W_k$;
3. a set of correlated, binary predictors, $V_1, V_2, \ldots, V_k$, obtained by generating k correlated, normal random variables as above and transforming them to 0's or 1's, according to whether their values were less than or greater than μ;
4. a set of mutually-independent, continuous predictors, $Z_1, Z_2, \ldots, Z_k$, such that each $Z_i \sim N(\mu, \tau^2)$; and
5. a set of mutually-independent, binary predictors, $Z_{k+1}, Z_{k+2}, \ldots, Z_{2k}$, generated by transforming continuous predictors as in (4) to 0's or 1's, according to whether they were less than or greater than μ.

In all of the simulations, $\mu = 10$, $\tau^2 = 2$, and $\rho = 0.5$. Five sizes of covariate "pools" were simulated: $k = 1, \ldots, 5$, corresponding to 5, 10, ..., 25 covariates total.

Responses were generated from the above predictors according to the following model:

$$Y = \beta_0 + \beta_x X_1 + \beta_w W_1 + \beta_v V_1 + \epsilon, \quad (1)$$

where the β's are regression coefficients and $\epsilon \sim N(0, \sigma^2)$ represents random error. Default values were $\beta_0 = 0$, $\beta_x = 0.3$, $\beta_w = 0.2$, $\beta_v = 1$, and $\sigma^2 = 1$, and $n = 50$ responses were simulated in each data set. A single simulation consisted of applying the various analysis techniques to 500 simulated data sets.

Equation (1) implies that there are three clusters of potentially informative covariates (the X's, W's and V's), but, in each cluster, only one of the predictors has a true, functional relationship with the response. Nevertheless, by virtue of their correlation with the key predictor, the other members of the cluster have indirect associations with the response that could lead them to be selected in regression modeling.

The Z variables, which Equation (1) shows are unrelated to the response Y, are included to mimic the tendency for real data sets to contain noninformative covariates.

2.2 Methods of variable selection

The following methods of variable selection were evaluated, with shorthand labels in boldface:

- **F-test:** stepwise multiple linear regression with F-tests. Starting with a model having no predictors, we add the covariate achieving the most statistically significant ($P < 0.05$) reduction of the residual sum of squares. This procedure is repeated in subsequent steps, and, at each step, any predictor in the model whose association with the response becomes non-significant is dropped. Statistical significance is judged from the usual partial F statistic (e.g., see Draper and Smith 1981).

- **C_p:** stepwise selection based on Mallows' C_p. For a linear regression model having p regression coefficients based on n observations, Mallows' C_p statistic is defined as

$$C_p = p + (n - p) \cdot \frac{\hat{\sigma}^2 - \hat{\sigma}_F^2}{\hat{\sigma}_F^2}, \tag{2}$$

where $\hat{\sigma}^2$ is the mean squared error for the model and $\hat{\sigma}_F^2$ is the mean squared error for a "full" model containing all possible predictors. Starting with a model having no predictors, a stepwise procedure is used to add or delete predictors until a minimum value of C_p is obtained (e.g., see p. 175 of Venables and Ripley 1994). In actual practice, the analyst would be wise to consider the relationship between C_p and the number of regression coefficients in the model, p, since models for which $C_p > p$ are likely to produce biased predictions (e.g., see Fisher and van Belle 1993).

I also experimented with all-subset regression by leaps and bounds, to minimize the C_p statistic. This method uses a computational trick to identify the best few subsets of predictors in linear regression models, without having to perform most of the possible regressions (e.g., see Weisberg 1985). I used the "leaps" function in the S+ language (Becker et al. 1988, Statistical Sciences Inc. 1993), which is based on the algorithm of Furnival and Wilson (1974). Since the results of this method of model selection were quite similar to those of the stepwise procedure described above, I do not report the all-subset results in great detail.

- **BIC:** stepwise selection based on Schwarz's Bayes Information Criterion. A stepwise procedure is used as above to construct linear regression models, but the criterion that is minimized is

$$\text{BIC} = n \log(\hat{\sigma}^2) + p \log(n). \tag{3}$$

This form of the statistic is due to Schwarz (1978). Model selection

using the BIC is discussed by Ramsey and Schafer (1997), and a more technical treatment may be found in Kass and Raftery (1995).

- Regression trees are decision trees for predicting a response based on successive binary splits of independent variables (Breiman et al. 1984). Trees are "pruned" to avoid over-fitting the training data. Two methods of pruning are used here:

  AIC-tree: an approach designed to minimize an approximation to Akaike's information criterion (Venables and Ripley 1994). The α of the cost-complexity measure of Breiman et al. (1984) is set equal to $2\hat{\sigma}_F^2$, where $\hat{\sigma}_F^2$ is the mean squared error from a linear regression model including all possible predictors.

  CV-tree: an approach based on cross-validation, in which the data set is split into 10 parts. Nine parts are used to grow a tree, which is then tested on the tenth; this is done ten ways. The final tree is obtained by pruning the original tree back to the number of nodes at which the average of the cross-validated deviances is a minimum.

  The S+ language (version 3.4) was used for the growing and pruning of regression trees, as described by Venables and Ripley (1994).

Except for the second regression-tree method described above, I did not explore the use of resampling in conjunction with the subset selection procedures. Breiman and Spector (1992) found in simulation studies that cross-validation and bootstrapping can greatly reduce bias in subset selection.

2.3 Assessment of predictive ability

For each of the 500 models obtained with a particular variable-selection procedure in a simulation, I calculated the predicted response for a single hypothetical observation having mean covariate values, i.e., $X_1 = \cdots = X_k = W_1 = \cdots = W_k = Z_1 = \cdots = Z_k = 10$, and $V_1 = \cdots = V_k = Z_{k+1} = \cdots =$

$Z_{2k} = 0.5$. The distribution of those predictions about the expected value of the response yields information about the bias and variability of predictions made by that method.

The ability of regression and regression-tree models to predict "future" observations having *different* sets of covariate values was assessed by further simulations. For each size of covariate pool ($k = 1, \ldots, 5$), the following steps were taken:

1. Assemble the predictive models from the five variable-selection techniques applied to the 500 "training" data sets (i.e., sets of covariates and responses generated as described earlier).

2. Generate 200 "new" data sets consisting of covariates and responses generated by the same mechanisms used to obtain the training data.

3. For each new data set and variable-selection technique, use the 500 models based on the training data sets to predict responses for the new data set.

4. Summarize the agreement between the observed responses in the new data set and those predicted by each of the 500 models using the mean squared error of the predictions (MSE) and the sample correlation of predicted and observed responses (R):

$$\text{MSE} = \frac{\sum_{i=1}^{50}(Y_i - \hat{Y}_i)^2}{50};$$

$$\text{R} = \frac{\sum_{i=1}^{50}(Y_i - \bar{Y})(\hat{Y}_i - \bar{\hat{Y}})}{\sqrt{\sum_{i=1}^{50}(Y_i - \bar{Y})^2 \sum_{i=1}^{50}(\hat{Y}_i - \bar{\hat{Y}})^2}}, \tag{4}$$

where $\hat{Y}_i$ is the predicted and Y_i the observed response for subject i, $\bar{Y}$ is the mean of the 50 observations in the new data set, and $\bar{\hat{Y}}$ is the mean of the 50 predicted responses. (These means are identical for regression models, but not for the regression-tree models.)

5. Average the goodness-of-fit measures in Equations (4) over the 500 predictive models available for each variable-selection method, and summarize the predictive ability of each method as the distribution of those averages over the 200 new data sets.

3 Results of Simulations

3.1 Numbers and types of covariates selected

Figure 1 shows the average number of covariates selected by the various approaches; recall that there are three truly informative covariates in all cases. The C_p and AIC-tree methods consistently select more covariates than the F-test, BIC and CV-tree approaches. For all methods except the CV-tree, the number of covariates selected increases with the size of the pool of covariates being selected from.

The proportions of models that are "correct" (i.e., that include only the three informative covariates, X_1, W_1 and V_1) are shown in Figure 2. For the smallest covariate pools, the C_p and F-test methods do best at identifying the correct model, while, for larger pools, the F-test and BIC approaches are best — although the overall success rate is quite low.

Figure 3 shows probabilities of various predictors being included in models produced by the different methods. For the informative covariates (X_1, W_1 and V_1), the C_p method usually has the highest inclusion probabilities, and the CV-tree method has the lowest. For the two informative continuous predictors (X_1 and W_1), the AIC-tree method has fairly high inclusion probabilities, and the F-test and BIC methods have intermediate inclusion probabilities. Interestingly, the two regression-tree approaches select the informative binary predictor (V_1) markedly less often than do the other three methods. For all methods, the inclusion probabilities for informative predictors decrease as the size of the covariate pool increases — presumably because of an increase in the chance for correlated, surrogate covariates to be selected instead. This decrease seems especially pronounced for the regression-tree approaches.

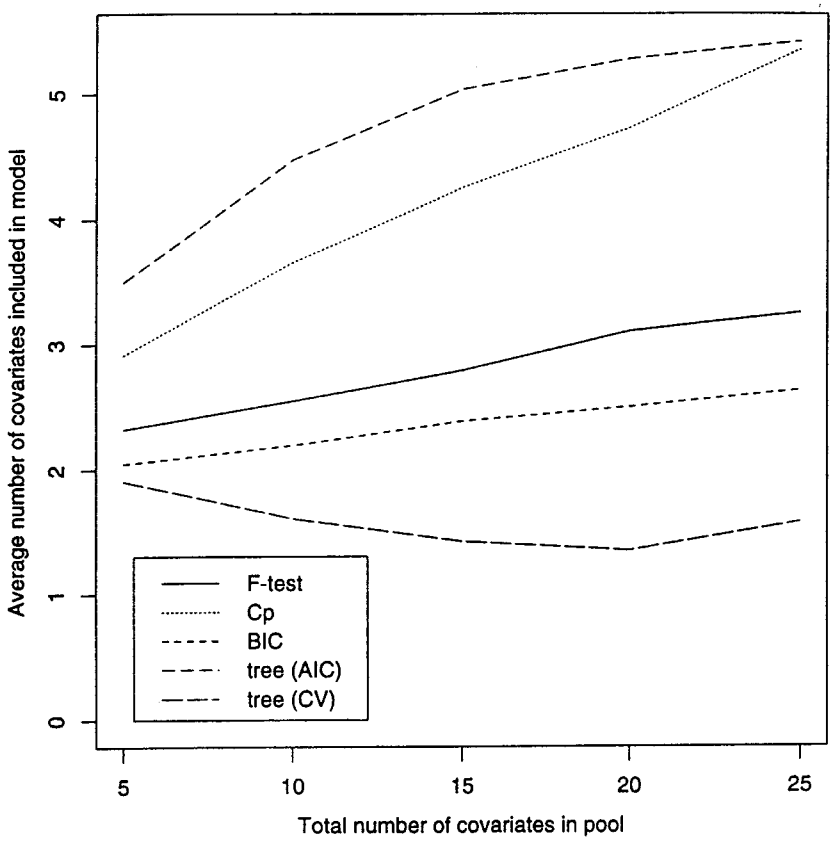

FIG. 1. Average number of covariates included in statistical models vs. size of the pool of available covariates (5, 10, 15, 20, or 25). Each point is an average from 500 simulated data sets.

For the noninformative covariates (Z's), the F-test and BIC methods generally have inclusion probabilities close to 0.05, while the C_p method yields substantially elevated inclusion probabilities (Figure 3). The AIC-tree approach includes noninformative continuous covariates with greater frequency than do the other methods, especially for small covariate pools, and its inclu-

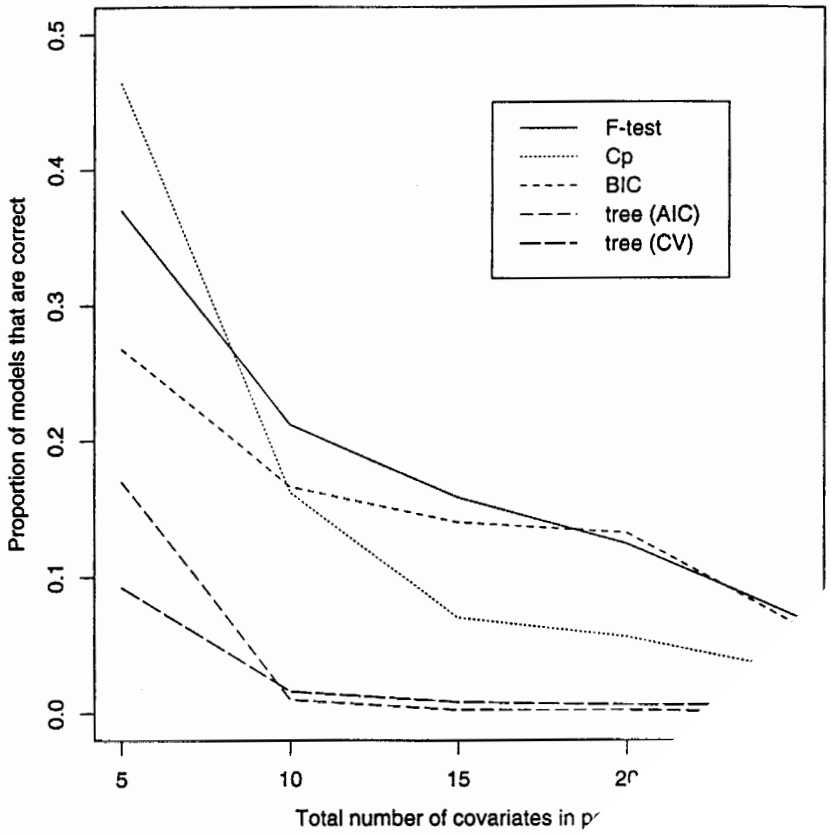

FIG. 2. Proportions of models fit to simulated
(i.e., contain only X_1, W_1 and V_1) vs. size of

sion of noninformative binary
methods except the C_p ap

Figure 4 is one po
informative covaria
across different
the covariat

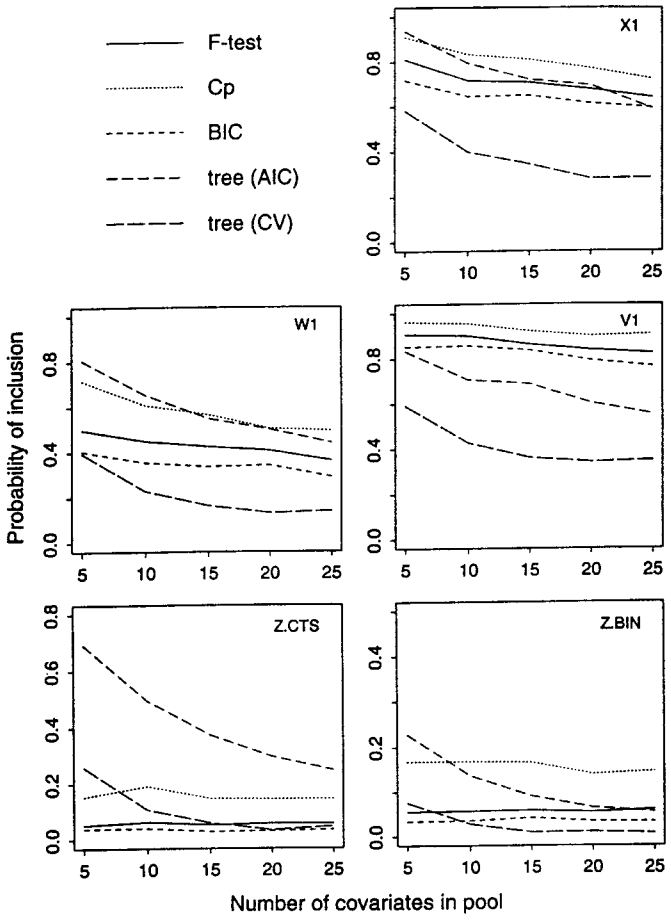

FIG. 3. Probabilities of including various covariates vs. size of the covariate pool. The plot labeled Z.CTS shows probabilities averaged over the k continuous, noninformative covariates, and Z.BIN represents averages over the k binary, noninformative covariates. Plots for the "semi-informative" covariates — $X_2, \ldots, X_k, W_2, \ldots, W_k, V_1, \ldots, V_k$ (not shown) — have patterns similar to those for the noninformative covariates, but with slightly higher inclusion probabilities overall.

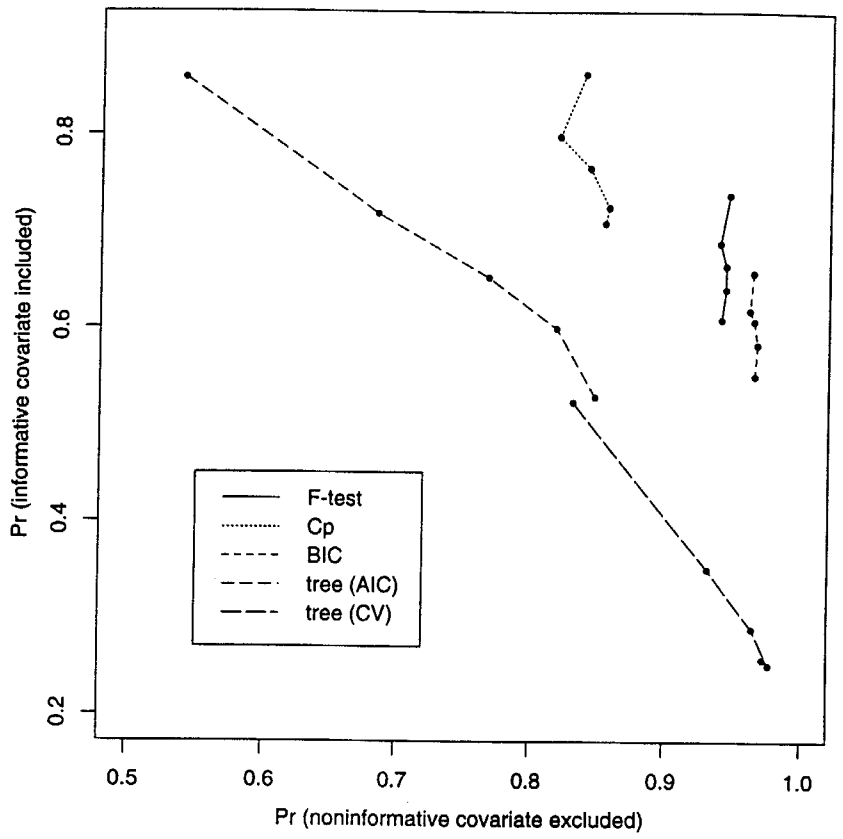

FIG. 4. Average probability of including an informative covariate (X_1, W_1, V_1) vs. average probability of excluding a noninformative covariate ($Z_1, \ldots Z_{2k}$). The points on each line correspond to the different-sized covariate pools, from five covariates (uppermost point) to 25 covariates (lowermost point). Each point is based on 500 simulated data sets.

tive covariate decreases, as seen earlier. The average probability of excluding a noninformative covariate is insensitive to the size of the covariate pool for the C_p, F-test and BIC methods, but increases dramatically with covariate-pool size for the two regression-tree approaches.

The C_p procedure has a relatively high probability of including an informative covariate, but a low probability of excluding a noninformative covariate (Figure 4). The F-test and BIC approaches "miss" informative covariates more frequently, but have a better chance of excluding noninformative covariates. The AIC-tree method includes informative covariates with reasonably high probability, but also has a high rate of inclusion of noninformative covariates. The CV-tree method is good at excluding noninformative covariates, but quite poor at including informative ones.

Which method is "best" depends on the relative costs of making the two possible kinds of errors (omitting informative covariates and including noninformative ones). If c_1 is the cost, or loss, accompanying the first kind of error, and c_2 the loss accompanying the second, we can write the expected loss as

$$\begin{aligned}\text{Expected loss} \;=\; & c_1 \cdot \Pr(\text{omitting an informative covariate}) + \\ & c_2 \cdot \Pr(\text{including a noninformative covariate}) . \qquad (5)\end{aligned}$$

Figure 5 shows the expected losses for the five methods, given that the two kinds of errors are equally costly ($c_1 = c_2 = 0.5$). In this case, the F-test and C_p approaches minimize the loss; the BIC method has a slightly higher expected loss; and the regression-tree approaches are clearly inferior to the other methods. Different values of c_1 and c_2, reflecting different costs of errors of omission and commission, would shift the balance among the different methods. It should also be realized that this loss analysis is specific to the definition of the errors implied by Figure 4. We might consider the trade-off between differently-defined errors, e.g., substitute Pr (*at least one* informative covariate excluded) and Pr (*at least one* noninformative covariate included) in Equation (5).

It is worth noting here that, in all of the simulations, the stepwise C_p procedure performed similarly to a leaps-and-bounds procedure to minimize

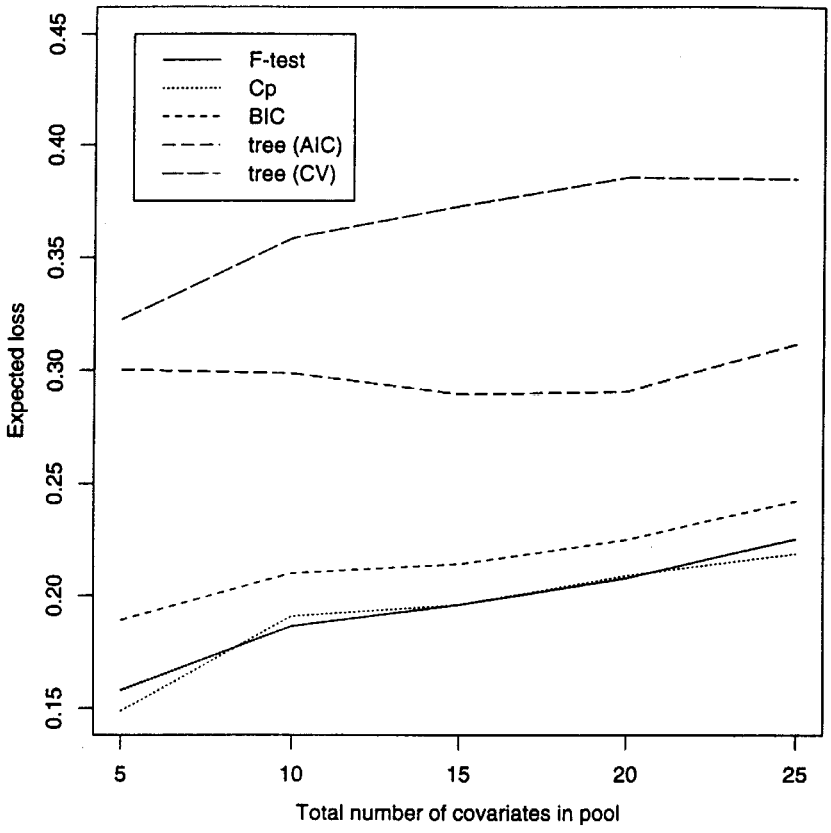

FIG. 5. Expected loss (with errors as implied by Figure 4, and $c_1 = c_2 = 0.5$ in Equation 5) for the five variable-selection methods, vs. size of covariate pool.

the C_p statistic. The leaps procedure yielded slightly higher inclusion probabilities for all covariates, but the two approaches chose identical models for 65% (25-covariate pool) to 99.6% (five-covariate pool) of the simulated data sets. The leaps-and-bounds algorithm effectively considers a larger universe of models than does the stepwise approach, but this does not result in a practically important difference in the nature of the models selected.

3.2 Prediction

Table I summarizes the results of using models based on simulated data sets to predict the response for a hypothetical observation having average covariate values. The three methods based on least-squares regression all appear to give unbiased estimates of the true response (5.5), but the regression-tree predictions are biased slightly upwards. More important, the average standard deviations of predictions from regression trees are 3.3 (CV-tree) to 4.5 (AIC-tree) times those for the other three methods.

Figure 6 shows the results of using the 500 models fit to simulated data sets to predict responses in 200 "new" data sets generated by the same mechanism. By any criterion, the regression-tree approaches do a poor job of predicting new observations, compared to the other three methods. There is a suggestion for large covariate pools that the mean squared error of prediction (see Equations 4) is higher for the C_p method than for the F-test and BIC approaches, although this difference is less evident in the sample correlations of predicted and actual responses. For all methods, the predictive ability of the models decreases as the size of the covariate pool increases.

4 An example with real data

I illustrate the different variable-selection techniques using data on the species richness of zooplankton communities in lakes (Dodson 1992). Eight covariates, some of which are strongly correlated, relate to lake size, location, and proximity to other lakes (see Table II). I modeled the species richness in 66

TABLE I

Prediction of response (true value = 5.5) for an observation having average covariate values. The overall mean is the average of the means from the five sizes of covariate pools, each of which involved 500 simulated data sets, and the average standard deviation is similarly averaged over covariate-pool sizes and simulated data sets.

Method	Overall mean	Average standard deviation
F-test	5.501	0.161
C_p	5.502	0.161
BIC	5.500	0.162
AIC-tree	5.836	$0.\bar{7}25$
CV-tree	5.740	0.536

North American lakes by applying different variable-selection techniques to the covariates listed in Table II, and then assessed the predictive ability of these models for a separate set of 37 lakes located in Europe and Asia (Dodson 1991, 1992).

Table III shows the results of the model building. At least four distinct models are obtained: three involving the covariates AREA, ELEV and NEAR, and one involving those three covariates plus LAT and L2. The AIC-based regression tree fit to the 66 North American lakes is shown in Figure 7. In this example, the regression-tree approaches lead to the highest correlations between species-richness values predicted and observed for the 37 Eurasian lakes, but confidence intervals for the correlation coefficients overlap substantially among methods.

Working with 38 North American lakes having complete data for a larger set of covariates than that shown in Table II, Dodson (1992) used a form of

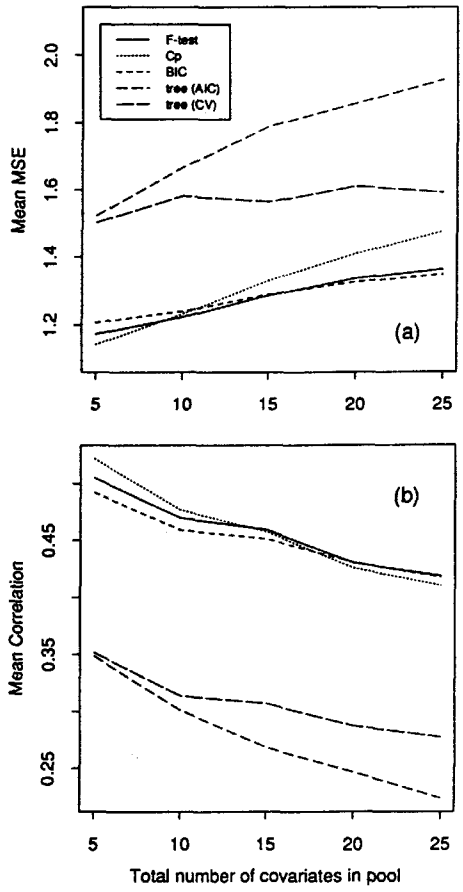

FIG. 6. Predictive ability of models fit to simulated data sets, as measured by (a) the mean squared error of predicted values about the true values, and (b) the sample correlation of predicted and true values (see Equations 4). For each of 200 "new" data sets, the measures were averaged over predictions from models fit to 500 "training" data sets; the plotted points are the means of these 200 averages. Averages (and ranges) of the widths of 95% confidence intervals about the 25 plotted points are 0.059 (0.054 to 0.070) for the MSE, and 0.020 (0.013 to 0.026) for the correlation coefficient.

TABLE II

Covariates used to predict zooplankton species richness in the lakes discussed by Dodson (1992). Logarithmic transformations were used to reduce skewness in the distributions of the response (SPP) and covariates.

Variable label	Meaning	Transformation used
SPP	Species richness	log(SPP)
AREA	Surface area (m^2)	log(AREA)
ZBAR	Mean depth (m)	log(ZBAR)
ZMAX	Maximum depth (m)	log(ZMAX)
ELEV	Elevation (m)	ELEV
LAT	Latitude	LAT
NEAR	Distance to nearest lake (km)	log(NEAR + 1)
L1	Number of lakes within 10 km	log(L1 + 1)
L2	Number of lakes between 10 and 20 km	log(L2 + 1)
LW20	Number of lakes within 20 km	log(LW20)

stepwise multiple linear regression to obtain a final model expressing species richness as an increasing function of area and number of lakes within 20 km (LW20), and a quadratic function of photosynthetic flux. I chose to exclude photosynthetic flux from the covariate pool, because it was not available for the Eurasian lakes on which I wanted to test the models. This example illustrates how the different methods of variable selection can lead to quite different models, with potentially different predictive abilities.

5 Discussion

All of the variable-selection techniques become increasingly challenged as the size of the covariate pool increases, leading to larger numbers of covariates

TABLE III

Models of zooplankton species richness obtained by the different variable-selection techniques, and results of predicting species richness for an independent set of 37 Eurasian lakes. The tranformations shown in Table II were used for both the response (SPP) and the covariates. The MSE and correlation coefficient measure the agreement between predicted and observed responses in the Eurasian lakes (see Equations 4). Numerical results for the CV-tree approach are average values from 10 runs of the cross-validation-based pruning method; seven of the final trees used all three listed covariates, while three trees did not use NEAR.

		Prediction for Eurasian lakes	
Method	Regression equation	MSE	Correlation
F-test, BIC	SPP = 1.55 + 0.080 AREA − 0.00015 ELEV − 0.348 NEAR	0.32	0.57
C_p	SPP = 2.07 + 0.073 AREA − 0.00018 ELEV − 0.0153 LAT − 0.346 NEAR + 0.072 L2	0.34	0.57
AIC-tree	SPP = AREA + ELEV + NEAR	0.28	0.65
CV-tree	SPP = AREA + ELEV (+ NEAR)	0.27	0.66

selected (Figure 1) and increasing rates of omission of informative covariates (Figure 3). The challenge of building parsimonious models when the number of potential parameters is large has been widely recognized (e.g., see Derksen and Keselman 1992, Anderson et al. 1994), and has led to rules-of-thumb such as Harrell's (1996) suggestion that the number of covariates in the pool should be no more than a tenth of the number of available observations.

The regression-tree approach can lead to quite different models, depending on the pruning method used (see Figures 1 and 3). The AIC-tree method

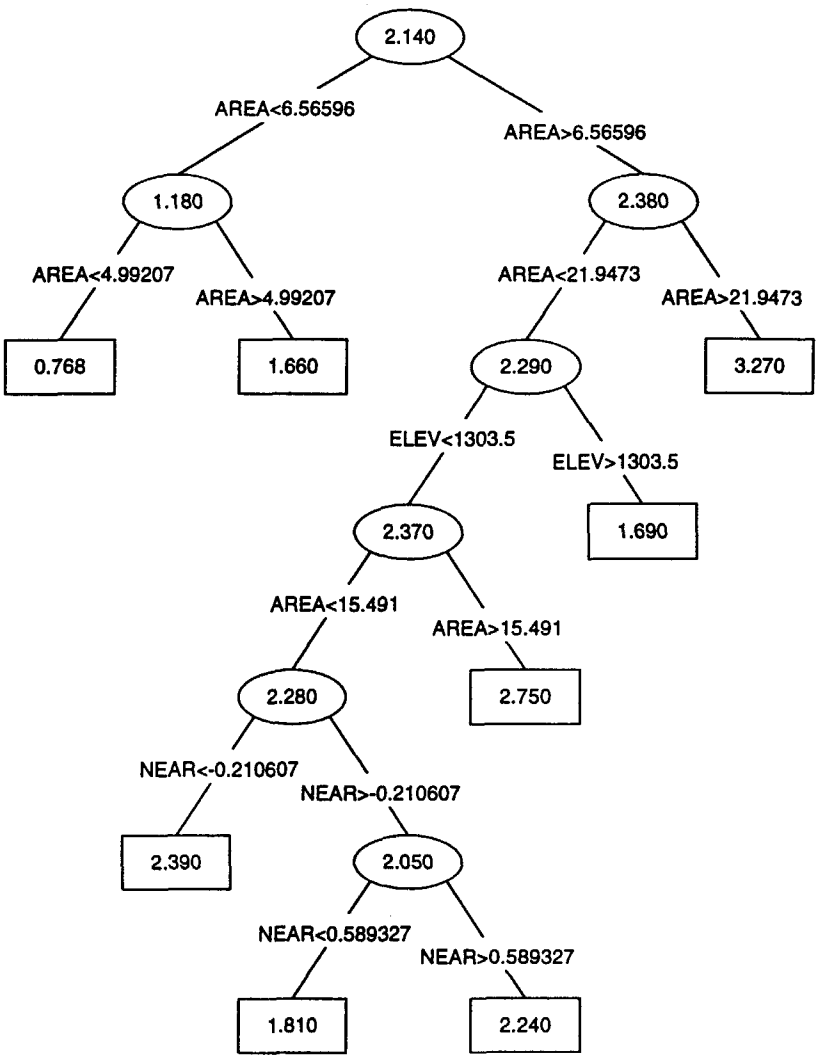

FIG. 7. Regression tree for the data on zooplankton species richness in 66 North American lakes. All numbers are on the tranformed scales shown in Table II. The numbers in the intermediate (oval) nodes and the terminal (rectangular) nodes, or "leaves", represent predicted values of log(SPP).

produces large models, including many noninformative covariates, while the CV-tree method produces smaller models, frequently omitting important predictors. Interestingly, the informative binary covariate, V_1, is included consistently *less* often by both of the regression-tree approaches than by the other methods (Figure 3) — even though regression trees are based on successive binary splits of independent variables.

The stepwise method to minimize Mallows' C_p statistic generally includes more covariates, informative and noninformative, than do the F-test and BIC approaches (Figures 1-4). Which approach is "best" depends on the relative costs of omitting important predictors and including noninformative ones. For the particular data structure modeled here, and for errors and costs as defined in the text, the regression-tree approaches are uniformly inferior to the other three approaches, and the F-test and C_p approaches enjoy a slight advantage over the BIC method (Figure 5).

For many applications of regression modeling, the "acid test" of the value of a model is how well it predicts new observations. For a constant covariate vector, the models fit to simulated data by the regression-tree approaches give somewhat biased "predictions" of the response, which are much more variable than those from the other three methods (Table 1). In spite of the different numbers of covariates ending up in the models, the other three methods do not differ substantially in the variability of their predictions for a constant covariate vector.

When we consider prediction for entirely new data sets, the regression-tree approaches are again substantially inferior to the other three approaches (Figure 6). The other three approaches are similar with respect to mean squared prediction error and the correlation between predicted and observed responses, although there is a suggestion that the F-test and BIC approaches may be preferable to the C_p method for large covariate pools (see Figure 6a).

The inferiority of the regression-tree approaches in this study is perhaps not surprising, given that the data were generated according to an ordinary linear regression model (Equation 1) — the "home turf" of the F-test, C_p

and BIC approaches. Regression trees might be expected to perform much better when covariate effects are nonlinear or discontinuous, or when there are important interactions among covariates. In addition, the sample size in the data sets simulated here (50) is quite small for the construction of regression trees, given the sequential partitioning that is the basis of the method (Miller 1994). Still, the high variability (Table 1) and low accuracy (Figure 6) of predictions of the regression-tree methods suggest that this tool may be best reserved for exploratory data analysis, at least until the relative merits of different kinds of pruning are better understood.

For the data structure used here, the F-test, C_p and BIC methods are fairly similar in their overall performance (Figures 5 and 6). One's choice of approach should be guided by the relative costs of the different kinds of errors in model building. If errors of omission of covariates are especially costly, the larger models produced by the C_p method might be preferred to the smaller models obtained by minimizing the BIC — and vice-versa if errors of commission are costlier. Whatever method of variable selection is used, the quality of models clearly decreases with increasing size of the covariate pool, pointing to the importance of the investigator deciding *a priori* on a manageable set of potentially informative covariates to use in model building.

ACKNOWLEDGEMENTS

I thank S. Dodson for making available the zooplankton data available, and J. Sifneos for helpful advice on implementing the regression-tree approaches.

BIBLIOGRAPHY

Anderson, D.R., Burnham, K.P. and White, G.C. (1994). "AIC model selection in overdispersed capture-recapture data," *Ecology*, 75, 1780-1793.

Becker, R.A., Chambers, J.M. and Wilks, A.R. (1988). *The New S Language: A Programming Environment for Data Analysis and Graphics.* Pacific Grove, California: Wadsworth & Brooks/Cole.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees.* Monterey: Wadsworth & Brooks/Cole.

Breiman, L. and Spector, P. (1992). "Submodel selection and evaluation in regression. The X-random case," *International Statistical Review*, 60, 291-319.

Cox, D.R. (1990). "Role of models in statistical analysis," *Statistical Science*, 5, 169-174.

Derkson, S. and Keselman, H.J. (1992). "Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables," *British Journal of Mathematical and Statistical Psychology*, 45, 265-282.

Dodson, S.I. (1991). "Species richness of crustacean zooplankton in European lakes of different sizes," *Internationale Vereinigung für Theoretische und Angewandte Limnologie, Verhandlungen*, 24, 1223-1229.

Dodson, S. (1992). "Predicting crustacean zooplankton species richness," *Limnology and Oceanography*, 37, 848-856.

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

Fisher, L.D. and van Belle, G. (1993). *Biostatistics: A Methodology for the Health Sciences.* New York: Wiley.

Furnival, G. and Wilson, R. (1974). "Regression by leaps and bounds," *Technometrics*, 16, 499-511.

Harrell, F.E., Jr., Lee, K.L. and Mark, D.B. (1996). "Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," *Statistics in Medicine*, 15, 361-387.

Kalbfleisch, J.D. and Prentice, R.L. (1980). *The statistical analysis of failure time data.* New York: Wiley.

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

Lehmann, E.L. (1990). "Model specification: The views of Fisher and Neyman, and later developments," *Statistical Science*, 5, 160-168.

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

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

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

Miller, T.W. (1994). "Model selection in tree-structured regression," *Proceedings of the Statistical Computing Section.* Alexandria, Virginia: American Statistical Association, 158-163.

Ramsey, F. and Schafer, D. (1997). *The Statistical Sleuth: A Course in Methods of Data Analysis.* Belmont, California: Duxbury Press.

Schwarz, G. (1978). "Estimating the dimension of a model," *The Annals of Statistics*, 6, 461-464.

Statistical Sciences Inc. (1993). *S-Plus Statistical Software, User's Manual,* Version 3.2. Seattle: Statistical Sciences Inc.

Venables, W.N. and Ripley, B.D. (1994). *Modern Applied Statistics with S-Plus.* New York: Springer-Verlag.

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

Received June, 1997; Revised February, 1998.