Critical Review and Comparison of Variable Selection Procedures for Linear Regression (Technical Report)

John Dziak, Runze Li, and Linda Collins
Methodology Center
Penn State University

December 7, 2005

Abstract

The derivation and behavior of various model selection and criteria (principally stepwise testing, pretesting, AIC, AIC_c, C_p, BIC, RIC) are reviewed and compared in the context of predictor selection for linear regression. Many such criteria are equivalent to penalized likelihood functions; of these, the prototypical examples are AIC and BIC, which differ only in the size of the penalty despite their disparate theoretical origins. The relationships among these procedures are rather complex and sometimes surprising, and no method dominates the others. However, different procedures give different weights to sensitivity versus specificity. Insight into these procedures is obtained from asymptotic results, orthogonal-case simplifications, and simulations.

Contents

1 Introduction **3**
1.1 Goals of Variable Selection 3

2 Testing-Based Methods **7**
2.1 Stepwise Testing 7
2.2 Pretesting 10

3 Prediction-Oriented Criteria **11**

4 Penalized Likelihood Criteria **16**
4.1 *AIC* 18
4.2 *BIC* 21
4.3 Comparing AIC and BIC 22
4.4 Similar Criteria 24
4.5 Finding the Model which Optimizes a Criterion 25
4.6 Handling Near-Best Subsets 27

5 Asymptotic Comparisons **29**

6 Orthogonal-Case Behavior **34**

7 Simulations **36**
7.1 Performance Measures 37
7.2 Data-Generating Models 39
7.3 Methods Compared 40
7.4 Results 41

8 Discussion **54**

1 Introduction

Model selection is an important but difficult part of data analysis, and much effort has been spent in developing powerful methods of data-driven model comparison and selection. This paper will describe some of the options available with their advantages and disadvantages, and provide practical suggestions for evaluating and using them in practice. Many variable selection techniques have been proposed, and the field is still rapidly expanding (see, e.g., Miller [2002]). We cannot review all techniques, but will discuss some of the more common ones, focusing especially on penalized-likelihood criteria like AIC and BIC.

We will focus mainly on variable selection (subset selection) in linear models. The concept of model selection is much broader than variable selection (e.g., choice of transformations, assumed response distributions, etc), and variable selection applies much more generally than to ordinary regression. However, the special case of selecting predictors for linear models allows simplified discussion of most of the important concepts and methods of interest, and extensions to more general settings are often straightforward. For example, AIC and BIC are applicable whenever a parametric likelihood exists and are very popular in categorical data analysis; they may even be extended to some semiparametric settings (see, e.g., Pan [2001]).

1.1 Goals of Variable Selection

Denote the model space (the set of all potential subsets being considered; in the current setting each model corresponds to a subset of the $\mathbf{x}$ variables)

as $\mathcal{M}$. We assume that $\mathcal{M}$ includes a full model with p observed predictors. The full model holds that the response variable y is related to covariates x_{ij} as follows:

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\boldsymbol{\beta}_j + \varepsilon_i,\ i = 1, ..., n \tag{1}$$

where the ε_i are independent errors. If we wish to consider interactions or transformations of the predictors, suppose that these derived terms have been included in the predictor vector, e.g., $x_3 = x_1x_2$, $x_4 = \log(x_1)$, etc.

Some of the p predictors in the full model may be unhelpful for predicting y, and can be deleted, i.e., some of the $\boldsymbol{\beta}_j$ are very close to zero and could be estimated as exactly zero to get a simpler model. The task of subset selection is to decide which set of variables should be deleted in order to get the best equation

$$y_i = \beta_0 + \sum_{j \in \mathcal{S} \subset \{1,...,p\}} \mathbf{x}_{ij}\boldsymbol{\beta}_j + \varepsilon_i \tag{2}$$

The confusion and controversy regarding variable selection in the literature are due in to the difficulty of defining "best" here. Thus before examining the procedures, it is important to consider possible goals. Among these are:

Sensitivity (Richness, Adequacy) We wish to include all important predictors in the model (roughly, those whose unknown true population coefficients β are not close to zero). In hypothesis testing this is Type II error and is usually considered less serious than Type I error, but in predictive contexts Type II error is sometimes more serious than

Type I. Excluding an important predictor adds to predictive error, and it may even creates a confounding variable and lead to misleading inference.

Specificity (Parsimony, Sparsity) We wish to exclude all unimportant predictors in the model (all predictors whose unknown true population coefficients β are zero, or all variables which do not improve prediction).

Future Predictive Ability Suppose we have observed data for a "training sample" of n members of a large population, and wish to predict the responses of other population members in the future. We wish to choose a subset of predictors which will give us low future prediction error. Several procedures, such as C_p and cross-validation, are especially tailored for this goal, but other procedures can also accomplish it. Note that in discussing this goal, we do not need to assume that any of the models in $\mathcal{M}$ are true descriptions of reality.

Selection Consistency Suppose we are willing to assume that one of the models in $\mathcal{M}$ is true, in the sense that it contains all possible important predictors and no unimportant predictors and that the relationships between the model predictors and the response are really linear. Our goal is then to find out which model this is. We want a procedure with a high probability of discovering that model. The best-known procedure tailored for this goal is Schwarz's BIC. (The term "consistency" here means that the chosen model size converges in probability to the true model size. It is not the same idea as consistency in estimating

the β parameters or the mean.)

Clearly, there is a trade-off between sensitivity and specificity, like the tension between Type I and II error risks in hypothesis testing. If we were only interested in sensitivity we could always use the full model, and if we were only interested in specificity we could use the mean-only model $Y_i = \beta_0 + \varepsilon_i$. By choosing somewhere between these extremes we are balancing the risk of making Type One or Type Two errors, in hopes of finding an adequate yet parsimonious model. The price we pay for seeking an intermediate-sized model is our uncertainty about just which such model to choose.

One might think that there is a trade-off between predictive ability and parsimony, but this is only partly true. Constraining the estimating equation in some way is important not only for interpretability, but even for prediction. Ordinary least squares and maximum-likelihood sometimes overestimate the amount of structure in the data ("fitting noise"), so some form of restriction (e.g., on the number of predictors) can actually improve parameter estimation and the fit of the model on future data. There is a need to balance the bias caused by too small a model with the variance (i.e., imprecision in estimation) caused by too large a model (Shibata [1989], Zucchini [2000] or Hastie et al. [2001]). Thus, the full model may not give the best predictions for future data, except when n (the amount of observed data available for fitting) is very large relative to p.

The relationship between future predictive ability and the goal of consistency is also complicated. Suppose that we believe that one of the models in $\mathcal{M}$ is really true. Then, provided that n is adequate to estimate all of the

parameters in the true model, the true model will also be the best model for future prediction. This is because the true model includes no bias and no unnecessary variance, and it is perfectly specific and sensitive. Thus we might conclude that, at least for large n, we should only use consistent procedures. Whether this conclusion is warranted is controversial.

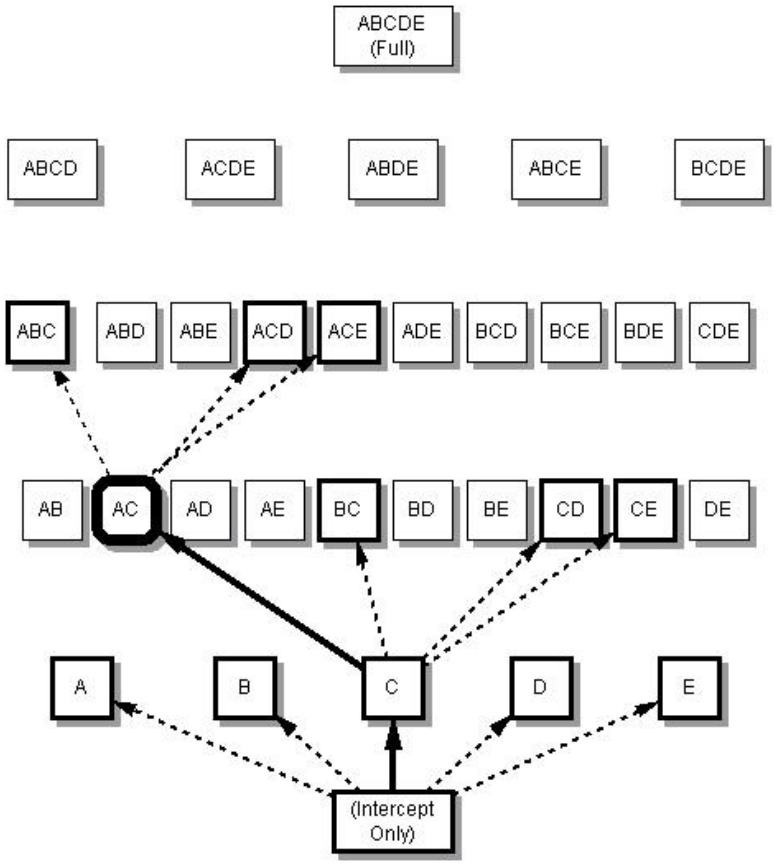
2 Testing-Based Methods

Some of the oldest approaches to model selection are based on sequential hypothesis tests. Although they provide a simple way to choose among predictors, they are not well-justified theoretically for use in model selection.

2.1 Stepwise Testing

Among the most widely used variable selection techniques are forward, backward, and stepwise (Efroymson [1960]) sequential testing, referred to collectively as stepwise methods. Forward testing starts with a null model and sequentially adds predictors if they are significant according to t or F tests. Backward testing starts with a full model and removes nonsignificant predictors one at a time. The stepwise algorithm alternates between adding and possibly removing. One problem with these methods is that they do not consider all possible models in $\mathcal{M}$ and may easily miss a good model. The problem is illustrated in Figure 1. Although model AB is judged as the best subset of size two, it is possible that ACD might be the best (or at least near best) subset of size three; but ACD will not be considered by forward or backward stepwise because it does not contain AB.

Figure 1: A Possible Stepwise Regression Path for Forward Stepwise Regression with p=5



Notes. Each box represents a model in $\mathcal{M}$. Dashed lines represent possible steps, while heavy lines represent steps taken (variables added). Letters in the boxes represent variables included in the models, out of an assumed pool of five available variables.

Now suppose $p = 15$. Then the ratio of the maximum number of models that would be considered by forward or backward stepwise, to the total number of possible models, is

$$\frac{15 + 14 + ... + 1}{2^{15} - 1} = \frac{120}{32767} \approx 0.4\%$$

The remaining 99.6% of models are hidden from consideration. Because not all models are considered, the final answer cannot be claimed to be the best available model.

The fact that not all models are considered was once beneficial. When stepwise methods were first introduced, they required much less computing time than would be required to fit all possible models. However, for reasonable p and modern computer hardware and software, the savings are essentially nil.

Furthermore, stepwise testing heuristics have no firm justification in statistical theory. Because the hypotheses of each test are chosen by the computer on the basis of previous tests, they do not adhere to the nominal α level (i.e., multiple testing increases the familywise type one error risk). Stepwise methods also have special problems with collinearity; Thompson [1995] suggested the analogy of selecting players (predictor variables) for a sports team. A stepwise approach would involve first choosing the best player, then choose the best second player *to support the first player*, and so on. This might not lead to the best possible performance, because players may complement or interfere with each other's performance, so that the best team need not include the best player. Neither does the resulting team nec-

essarily represent a true listing of players in order of ability, since all players after the first were selected based on the knowledge that the first had been added. The result would be a good team, but perhaps not the best possible.

2.2 Pretesting

Another common model selection heuristic is "screening" or "pretesting:" using initial t-tests for each predictor to decide which are significant, and removing those that are not. Usually the testing is done at $\alpha = .05$ (critical $t \approx 2$). Occasionally the one-standard-error rule ($t = 1$, i.e., $\alpha \approx .3$) is also proposed. Although these approaches are common practice, and are often preferable to using a full model, they involve potential problems (Freedman [1983], Freedman and Pee [1989], Babyak [2004]). They ignore the fact that when predictors are correlated, the significance of a predictor depends on which other predictors are in the model. Thus, although researchers often prefer pretesting over stepwise selection because the latter seems too *ad hoc* and exploratory, some of the criticisms that have been made about stepwise selection methods apply as well to pretesting.

Suppose the full model contains variables A, B, C, D, and E, and that B, D, and E are not significant at the .05 level. Then a naïve testing approach would be conclude that only A and C are significant predictors, and to leave B, D, and E out of the final model. However, this may be too parsimonious, as suggested by Figure 2. Of the 32 possible subsets, only six (ABCDE, ABCD, ABCE, ABDE, ACDE, BCDE) were fitted and evaluated before making a decision; and illogically, the one that was finally chosen (AC) was

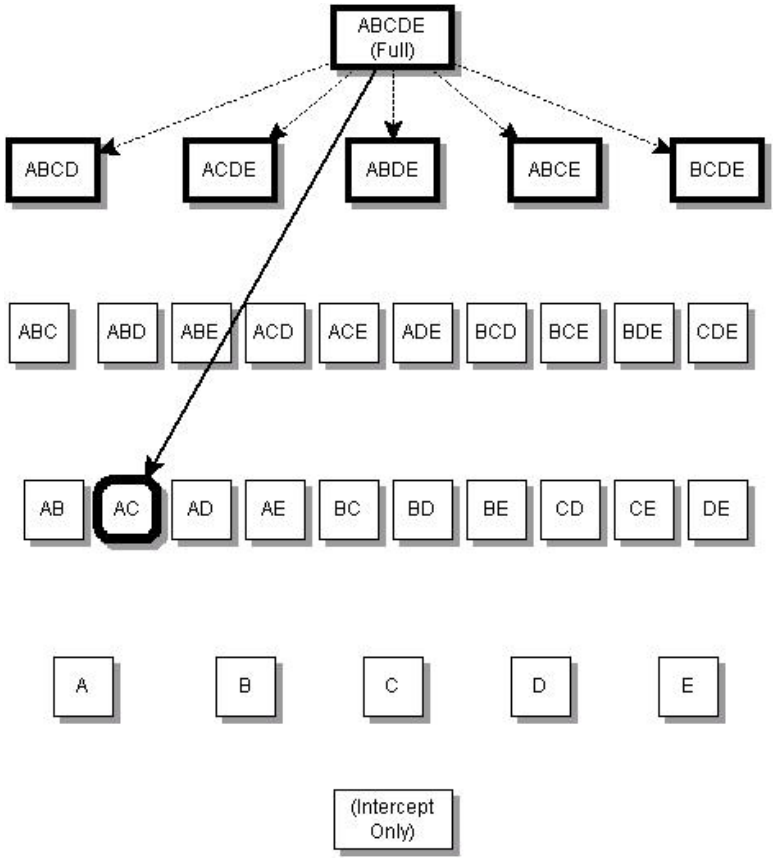
not one of those six. If the p-value for variable B is .09, can we safely exclude it? The answer would depend on whether we consider it more important to exclude inactive variables or to include active variables, i.e., whether we consider Type I or Type II error as worse. Also, suppose D and E are highly correlated with each other as well as with the response, and so their coefficients have very large standard errors. In such a situation, possibly neither variable would attain $p < .05$, and keeping both would lead to inflated variance and an unstable model; but dropping both would result in a loss of information and a biased model.

One reason for the shortcomings of the testing approaches is that they attempt to find the best model without first defining what a good model is. They simply base all decisions on significance tests; but classical testing and model selection address different questions, and do not always mix well (Burnham and Anderson [2002]; see also Christensen [2005]). In contrast, other methods first define a measure of model goodness balancing fit and parsimony, and then search for the model which optimizes this measure. This enables a richer theoretical treatment, and potentially more usefulness in practice.

3 Prediction-Oriented Criteria

Several popular methods are based on estimating the future predictive ability of each model. Suppose we wish to choose the model which minimizes

Figure 2: A Possible Screening-Based Variable Selection Choice for p=5



Notes. Notation is similar to that in Figure 1. Notice that model AC is chosen without being tested, after comparing ABCDE with the five four-variable models.

the long-run mean squared prediction error for future observations, i.e.,

$$EMSE(\mathrm{M}_j) = \mathrm{E}(\mathbf{y}_{\mathrm{new}} - \hat{\mathbf{y}}_{\mathrm{new}})^2$$

where $\mathbf{y}_{\mathrm{new}}$ is randomly selected new data from the population and where $\hat{\mathbf{y}}_{\mathrm{new}} = \mathbf{x}_{\mathrm{new}}\hat{\beta}(\mathrm{M}_j)$ is the prediction for $\mathbf{y}_{\mathrm{new}}$ generated by model M_j. If we could estimate $EMSE$ for each model available, we could choose the one with the smallest $EMSE$, or perhaps select a few with small $EMSE$ and choose among them on theoretical grounds. However, estimating $EMSE$ is difficult, since we do not have any data beyond the observed sample.

A naïve estimate for $EMSE$ would be $n^{-1}\mathrm{RSS}(\hat{\boldsymbol{\beta}})$, the average squared prediction error obtained by fitting the model of interest on the observed sample. Minimizing this criterion would be equivalent to maximizing R^2, i.e., it would always lead us to select the full model ($\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\mathrm{full}}$). Since the coefficients are being estimated from the same data on which the models are being tested, we get an inflated assessment of the reduction in prediction error available from each predictor. Thus the larger the number p_{in} of predictors used relative to n, the more $n^{-1}\mathrm{RSS}$ will be negatively biased as an estimate of $EMSE$, i.e., the more unduly "optimistic" (Hastie et al. [2001], pp. 200-3) we will be about how well our model works.

The desire for an unbiased estimate of $EMSE$ leads to Mallows' C_p criterion (see Hocking [1976], Olejnik et al. [2000]). Mallows [1973] showed that an appropriate bias correction to RSS is $2p_{\mathrm{in}}\sigma^2$, where p_{in} is the number of predictors in M_j and σ^2 is the variance of the unknown true model. Since σ^2 is of course unavailable, it is estimated in practice by the $\hat{\sigma}^2$ for the full

model. Thus we wish to minimize

$$n^{-1}\text{RSS} + 2n^{-1}p_{\text{in}}\hat{\sigma}^2 \tag{3}$$

Mallows' C_p itself is a linear function of (3), specifically $C_p = \text{RSS}/\hat{\sigma}^2 + 2p_{\text{in}} - n$. Under the classical linear regression assumptions, and if $\hat{\sigma}^2$ is a good approximation for the σ^2 in the full model, the second term in (3) serves asymptotically as the needed bias correction, so that a low C_p indicates good future predictive ability.

This is an interesting result, but there are reasons for uncertainty. First, it is not clear whether $\hat{\sigma}^2$ is really an appropriate substitute for σ^2. Second, the adjustment for optimism used in deriving C_p is itself optimistic. It assumes that future x will be the same as present x and that only y will be different. It also does not take into account the fact that the subset chosen was not set a priori but was selected by a data-driven method which introduces additional risk of overfitting (see Hastie et al. [2001, p. 204], Miller [2002]). Hence the bias correction is not quite large enough, so $EMSE$ is still somewhat underestimated, especially for larger p_{in}. Thus if we choose the model with the lowest C_p there will be a remaining tendency to overfit, as we will see in Sections 5, 6, and 7.

Another penalized statistic similar to C_p is the classic adjusted R^2 (Wherry [1931]; see Hocking [1972]). This attempts to correct the usual R^2 directly for optimism:

$$R^2_{adj} = \frac{(1 - (n-1)(1-R^2))}{n-p}$$

Like C_p, adjusted R^2 has poor specificity (see Section 6) and a remaining

tendency to overestimate fit.

Cross-validation provides a more direct, although more computationally intensive, approach to estimating $EMSE$. By setting aside part of the original sample to serve as new test data, we may reduce the problem of optimistic overfitting. Several approaches to cross-validation are available. Another possibility is to use a holdout sample: divide the dataset into fitting and testing subsamples. This is generally considered an inefficient use of the data, and generally leads to poorer performance unless n is very large.[1] A more efficient approach analyzes the same sample many times, each time partitioning it randomly into *fitting* and *testing* subsamples. We estimate the coefficients for each model using the fitting subsample and then evaluate each average performance (in terms of mean squared error) in predicting the data in the testing subsamples. There are several variations on this:

1. *Leave-one-out (Predictive Sum Squares).* Here n analyses are done for each model, with each observation serving once as the testing subsample. Computationally we do n regressions, with sample size $n-1$ each, for each possible model (of which there may be 2^p). This would be extremely time-consuming, except that an easier computational formula exists for linear models.

2. *k-fold* (k is usually five or ten). Here we divide the data into k parts and use each part as a testing subsample. The whole process of partitioning and fitting could also be repeated several times.

3. *Leave-d-out where* $d/n \to 1$. Here we divide the data into many parts and use each part as the *fitting* subsample. This scheme will perform

poorly for small n, since the resulting fitting subsamples become very small.

Leave-one-out cross-validation is asymptotically equivalent to C_p and AIC, under some conditions (see Stone [1977], Shibata [1981], Shao [1997]). Like them it tends to overfit. On the other hand, when n is large, leave-d-out cross-validation has better specificity and acts like BIC. k-fold has intermediate performance (Shao [1997]; see also Breiman [1995]). Apparently, the size $n - d$ of the fitting subsample relative to the testing subsample determines the power of the procedure to detect structure in the data (including spurious structure if the size is too large).[2]

C_p as described here is only applicable in linear models. Cross-validation can be computationally difficult. The criteria described in the next part are both very general and very easy to use.

4 Penalized Likelihood Criteria

If the correct subset were known, we would estimate its parameters by maximizing the fit to the observed data, i.e., maximizing the log-likelihood (in Gaussian linear models this is equivalent to minimizing the residual sum squares). However, if the true form of the model is not known, we might consider parsimony along with fit, and maximize a *penalized log-likelihood*:

$$\ell(\mathbf{X}, \mathbf{y}, \hat{\boldsymbol{\beta}}) - \mathrm{c}(\hat{\boldsymbol{\beta}}) \tag{4}$$

where ℓ is the likelihood, $\boldsymbol{\beta}$ is the vector of regression coefficients, and $\mathrm{c}(\hat{\boldsymbol{\beta}})$ is a measurement of model complexity: specifically, some kind of norm measuring how "big" $\hat{\boldsymbol{\beta}}$ is. Thus the task is not only to make ℓ large but to keep $\hat{\boldsymbol{\beta}}$ small. A simple way (although not the only way; see, e.g., Tibshirani [1996], Fu [1998]) to operationalize the size of $\hat{\boldsymbol{\beta}}$ is in terms of the number of free nonzero coefficients in the model. Thus, $\mathrm{c}(\hat{\boldsymbol{\beta}}) = \lambda p_{\text{in}}$ where p_{in} is the number of predictors included in the model and λ is some constant which controls the relative importance of parsimony versus good sample fit. [3]

The Akaike Information Criterion (AIC; Akaike [1973]), Schwarz's Bayesian Information Criterion (BIC; Schwarz [1978]), and Foster and George's newer Risk Inflation Criterion (*RIC*; Foster and George [1994]), although each is motivated differently, are equivalent to different choices of λ, i.e., to maximizing

$$\ell(\mathbf{X}, \mathbf{y}, \boldsymbol{\beta}) - \lambda p_{\text{in}} \tag{5}$$

In the notation of (5), AIC uses $\lambda = 1$, BIC uses $\lambda = \frac{1}{2}\ln(n)$, and RIC uses $\lambda = \ln(p)$ (Foster and George [1994]). Thus, BIC penalizes complex models heavily and prefers simpler models, while AIC is less strict. Ordinary maximum likelihood ($\lambda = 0$) does not consider parsimony. There are many other possible similar criteria such as the Hannan and Quinn [1979] information criterion with $\lambda \propto \ln(\ln n)$, and the "generalized information criterion" with any $\lambda \propto \ln n$ (see, e.g., Bozdogan [1987], Rao and Wu [1989], Shao [1997]), which which we do not review here. There is also some interest in choosing λ more adaptively; see, e.g., George and Foster [2000].

Notice that maximizing the penalized likelihood (5) is equivalent to minimizing $n \log(\hat{\sigma}^2) + 2\lambda p_{\text{in}}$ if we use $\hat{\sigma}^2 = \text{RSS}/n$ and assume a normal distribution for the errors. (Thus, the penalty constants for AIC and BIC are usually written as 2 and $\ln(n)$.) It is also equivalent to minimizing the penalized least-squares criterion

$$\text{RSS} + 2\sigma^2 \lambda p_{\text{in}} \tag{6}$$

if σ^2 is known or consistently estimated and if the distribution is normal. This suggests the asymptotic equivalence of C_p to AIC for linear modeling (Shao [1997]).

The seemingly arbitrary choices of λ used by AIC, BIC, and RIC are each optimal under a different theory of model selection. They are explained in the following sections.

4.1 *AIC*

The AIC of Akaike [1973] is motivated by the Kullback-Leibler discrepancy, which (loosely) measures how far a model is from the truth. We assume that the population or process from which the data was sampled is governed by an unknown, perhaps nonparametric, true likelihood function $f(\mathbf{X}, \mathbf{y})$, and we want to approximate the unknown f by a model-specific parametric likelihood $g(\mathbf{X}, \mathbf{y}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of parameters (for linear models, this is $(\boldsymbol{\beta}, \sigma^2)$). The "discrepancy" between f and g, or "information lost"

by representing f by g, is defined as

$$\begin{aligned} KL(f,g) &= \int \ln\left(\frac{f(x)}{g(\mathbf{X},\mathbf{y}|\boldsymbol{\theta})}\right) f(x)dx \\ &= \mathrm{E}\left(\ln f(x)\right) - \mathrm{E}\left(\ln g(\mathbf{X},\mathbf{y}|\boldsymbol{\theta})\right) \end{aligned}$$

Generally we have a number of possible g's (one is defined by each subset of available predictors) and we want to find the one that minimizes $KL(f,g)$, i.e., maximizes the expected log-likelihood $\mathrm{E}\left(\ln g(\mathbf{X},\mathbf{y}|\boldsymbol{\theta})\right)$. The unknown term $\mathrm{E}\left(\ln f(x)\right)$ is not important for choosing a model, since it is the same for each model being compared. The second term must be adjusted for the fact that $\boldsymbol{\theta}$ is not prespecified but estimated (usually by maximum likelihood) from $\mathbf{X}$ and $\mathbf{y}$. Thus $\mathrm{E}\left(\ln g(\mathbf{X},\mathbf{y}|\boldsymbol{\theta})\right)$ is replaced by the expected fitted log-likelihood

$$\mathrm{E}_y\,\mathrm{E}_x\left(\ln g(\mathbf{X},\mathbf{y}|\hat{\boldsymbol{\theta}}(\mathbf{X},\mathbf{y}))\right) \tag{7}$$

We cannot simply estimate (7) by using the observed fitted log-likelihood

$$\ell(\mathbf{X},\mathbf{y},\hat{\boldsymbol{\theta}}) = \ln g(\mathbf{X},\mathbf{y},\hat{\boldsymbol{\theta}}) \tag{8}$$

since this would be a biased estimate of (7), especially when p_{in} is large. Maximum likelihood tends to overfit; the fitted models seem better than they really are since the parameters are selected using the same data on which their fit is tested. Akaike [1973] showed that a roughly unbiased estimator of (7) is a constant plus

$$\ell(\mathbf{X},\mathbf{y},\hat{\boldsymbol{\theta}}) - p_{\text{in}}$$

Thus, a good guess at the model in $\mathcal{M}$ which best approximates the true likelihood function would be the one which minimizes

$$\text{AIC} = -2\ell(\mathbf{X}, \mathbf{y}, \boldsymbol{\beta}) + 2p_{\text{in}} \tag{9}$$

More information about AIC can be found in Burnham and Anderson [2002], or in Shibata [1989], Hastie et al. [2001], Burnham and Anderson [2002, 2004] or Kuha [2004].

Notice that the motivation of AIC is similar to that of C_p. Both assume that we wish to choose the model that optimizes some theoretical measure of model performance and estimate this measure by adjusting a sample estimate to reduce overfitting due to optimism. Neither considers parsimony to be a value in itself; the complexity penalty is only present to improve estimation and prediction by reducing overfitting.

The asymptotic approximation on which the AIC is based is rather poor when n is small (say, $n < 40p$; Burnham and Anderson [2004]). Therefore, Sugiura [1978] and Hurvich and Tsai [1989] proposed a small-sample correction, leading to the AIC_c statistic

$$\begin{aligned} AIC_c &= AIC + \frac{2p_{\text{in}}(p_{\text{in}}+1)}{n - p_{\text{in}} - 1} \\ &= -2\ell(\mathbf{X}, \mathbf{y}, \boldsymbol{\beta}) + \left(2 + \frac{2(p_{\text{in}}+1)}{n - p_{\text{in}} - 1}\right) p_{\text{in}} \\ &= -2\ell(\mathbf{X}, \mathbf{y}, \boldsymbol{\beta}) + 2\left(\frac{n}{n - p_{\text{in}} - 1}\right) p_{\text{in}} \end{aligned}$$

By making the complexity penalty larger when p_{in} is large relative to n, AIC_c can perform better than AIC. Thus Burnham and Anderson [2002]

recommend that AIC_c used in place of AIC for modest n.

4.2 BIC

The Bayesian Information Criterion of Schwarz [1978] is derived from Bayesian theory. Many methods of Bayesian variable selection are available, but the BIC is by far the simplest (and to its critics, most simplistic) to use. Typically, in Bayesian model selection we set a prior probability for each possible model M_j, perhaps setting all prior probabilities equal to express ignorance. Equivalently, we could give each coefficient β_j an independent prior probability of being nonzero; the prior probability for a given subset is then a function of these. We also set prior distributions for the nonzero coefficients in each model, e.g., we might say that a given coefficient, if the model does not set it to zero, will instead be normally distributed with some high variance. If we then assume that one and only one model in $\mathcal{M}$, along with its associated priors, is appropriate, we can then use Bayes' Theorem to find the posterior probability of each model given the data:

$$\Pr(\mathcal{M}_i|\mathbf{x},\mathbf{y}) = \Pr(\mathcal{M}_i)\Pr(\mathbf{y}|\mathbf{x},\mathcal{M}_i)$$

Thus, assuming the prior probabilities are equal, $\Pr(\mathcal{M}_i|\mathbf{x},\mathbf{y}) \propto \Pr(\mathbf{y}|\mathbf{x},\mathcal{M}_i)$. Schwarz [1978] and Kass and Wasserman [1995] showed that if we assume equal prior probabilities for all models and a certain "unit information prior" for the coefficients, then $\Pr(\mathbf{y}|\mathbf{x},\mathcal{M}_i)$ can be well approximated (see Raftery

[1995a]) by $\exp(\frac{1}{2}BIC)$, where

$$BIC = -2\ell(\mathbf{y}|\mathbf{x}, \hat{\beta}_{\mathcal{M}_i}) - \ln(n)(p_{\text{in}} + 2)$$

$(p_{\text{in}} + 2)$ here is the number of parameters in the model, including the intercept and variance. Thus if we want to choose the model with the highest posterior probability, we need only choose the one with lowest BIC. Classically, n here is the number of subjects and is thought to represent the amount of information in the sample; however, there are nonstandard situations (e.g., clustered data, categorical predictors with some categories very small) in which it is unclear what n is (see Weakliem [1999] and Raftery [1995a]).

Raftery [1995a] discusses potential uses of BIC in social science research, focusing mainly on its uses for model comparison (i.e., Bayesian hypothesis testing) and multimodel inference (model averaging) rather than exploratory model selection as in this paper. Other positive reviews of BIC are given in Hauser [1995] and Rust et al. [1995].

4.3 Comparing AIC and BIC

AIC and C_p tend to agree with each other, but BIC with its stricter penalty sometimes chooses a smaller model. When AIC and BIC indicate the same subset as best, we may feel rather confident in that model; but when they disagree, which one should we believe? BIC is generally favored because it is a "consistent" model selection technique, unlike AIC which overfits. That is, assuming that there is a fixed finite number of models in $\mathcal{M}$ and that one

of them is the true model, then as the sample size n gets large enough, the the lowest-BIC model will be the true model, with probability approaching 100% . This would seem to establish BIC as the best criterion. However, in practice sample sizes are limited, and the BIC may underfit, i.e., prefer overly simple models (,Hastie et al. [2001, pp. 206-208]; see also Lin and Dayton [1997] and Yang [2004]). Critics of BIC (e.g., Shibata [1989], Burnham and Anderson [2002], Leeb and Pötscher [2005]) say not only that it underfits but that its asymptotic properties are based on unrealistic assumptions, such as the availability of a fixed true model and a sample size which approaches infinity. In a way, the debate centers on the relative importance of sensitivity versus specificity. Prediction-oriented and consistency-oriented procedures differ in the weight they give to parsimony when n is large. When n is small, parsimony is clearly important because there is not enough information to estimate many parameters well, and poorly estimated parameters may do more harm than good to the fit; hence the recommendation to use AIC_c instead of AIC. But if n is large and our goal is prediction, then sensitivity becomes much more important than specificity, although they are equally important if our goal is to choose a true subset. For large n, so that a truly zero coefficient will be estimated as near zero with high probability; thus, although it complicates interpretation, it does relatively slight harm to prediction. Meanwhile, setting a nonzero coefficient to zero adds bias and harms prediction. In other words, when n is large variance is small, so we mainly need to control bias by allowing a rich model. Hence a prediction-oriented procedure like C_p or AIC will tend to play it safe by deleting rather few predictors. In contrast, the BIC does not wish to include a predictor

unless there is significant evidence that it is useful.

Ironically, while BIC is convenient and has good asymptotic properties, it has been severely criticized from a Bayesian perspective (Weakliem [1999]; see also Gelman and Rubin [1995] and the rejoinder in Raftery [1995b]). The uniform prior on the models and the relatively noninformative prior on the coefficients may not reflect the analyst's real prior beliefs or information. Thus, BIC is more appealing to frequentists than to theoretical Bayesians.

It is sometimes said that the difference between AIC and BIC is found in the interpretation of the target model, i.e., the goal of each procedure. The motivation of the BIC imagines that one of the models in $\mathcal{M}$ is *true* and that we want to find this true model. The motivation of the AIC (minimizing loss of information, or minimizing distance from the truth) is subtler, asking only for *best* model available (Burnham and Anderson [2002],Kuha [2004]). However, this philosophical argument implies nothing about the actual behavior of the two criteria and can even be misleading. Use of BIC in practice does not really require that one believe that one of the models in $\mathcal{M}$ is true (see Burnham and Anderson [2004], Kuha [2004]). One could also consider the BIC as trying to find a practically true or "quasi-true"rather than true model (perhaps something like the model with the least bias; see Shao [1997], Burnham and Anderson [2004], Kuha [2004]).[4]

4.4 Similar Criteria

We mention here one other interesting criterion, the Risk Inflation Criterion (RIC) of Foster and George [1994]. This is (5) with $\lambda = \ln(p)$, where p is the number of available predictors in the full model; note the dependence on

p rather than n. Intuitively, this penalty considers that the more predictors are available, the more strict we should be in deciding which predictors are worthy to enter the final model, in order to get a model of reasonable size. The RIC is motivated as an attempt to minimize a modified risk criterion that measures the increase in $EMSE$ caused by choosing the wrong size of model. It can also be motivated as a cutoff based on the highest t-statistic that would be expected by chance alone from a set of p inactive predictors (George and Foster [2000]).

4.5 Finding the Model which Optimizes a Criterion

So far we have only discussed the information criteria themselves, not how to find the model which optimizes them. The most general means of doing this is an exhaustive search through $\mathcal{M}$ to find the best subset (i.e., an "all-possible-regressions" or "all-subsets" algorithm). With reasonable p (say, less than 40) and modern computer software and hardware, for linear models this is easy and almost instantaneous.

Within a given value of the model size p_{in} (e.g., for all four-variable subsets), the best model in terms of AIC, AIC_c, BIC, RIC, or adjusted R^2 will be the model with the best R^2 (lowest RSS and hence highest likelihood; of course, the penalty term in (5) is the same for all models of a given size). These methods differ only in their way of determining the best subset size, not choosing the best subset within this size. Thus all that is needed is software to find the best subset for each p_{in} (e.g., the RSQUARE option in SAS PROC REG). In the case of normal linear models with reasonable p, the best in each class can be found using a specialized approach such as

the quick and efficient "leaps-and-bounds" algorithm of Furnival and Wilson [1974] (see Miller [2002]). These approaches, used by most popular software today, can determine the subset (or group of k subsets) having the highest R^2 for each size p_{in}, without having to fit each one separately. Evaluating each of these candidates on AIC, etc., is then extremely quick.

If p is very large so that evaluating all models is impossible even using special algorithms (e.g., if $p = 40$ so that there are $2^{40} \approx 10^{12}$ possible models), or in nonclassical situations (e.g., some generalized linear models) for which special algorithms are unavailable and computation is burdensome, it becomes important to impose some additional structure on the search. Thus there has been much work on efficient ways searching for promising models in $\mathcal{M}$. These include stochastic searches such as "simulated annealing;" these are not new selection criteria but ways of searching an otherwise intractably large $\mathcal{M}$. The most commonly used replacements for an exhaustive search are simple stepwise searches, similar to the stepwise testing of Section 2.1, but with a decision rule based on increases in AIC or BIC rather than on F- or t- tests. This stepwise approach is not needed for linear models unless p is huge – depending on the software used it may even take longer than all-subsets – and it may easily miss the best-AIC or best-BIC model. Stepwise searches do not necessarily pick the best model even for the sample, much less the population (see Section 2.1). Also, a stepwise search will only give one subset for each model size; it is much better, when possible, to ask for at least the three or four best subsets of each size. One argument *in favor* of a stepwise approach is that, since it considers fewer models (e.g., forward stepwise only considers subsets which contain the subsets chosen

in previous steps), there may be less capitalizing on chance (i.e., less Type One error risk inflation) and less bias in coefficient estimates and measures of fit (see Miller [2002]). There may also be situations in which the heuristic restriction implied by stepwise selection – that a good model should be one which contains other good models – is reasonable in itself.

4.6 Handling Near-Best Subsets

An important advantage of an all-subsets search with a fit criterion such as *AIC* or *BIC*, over a sequential testing approach, is that with the former we may ask for the ten best subsets, the five best for each model size, etc. This allows the analyst to think about different possibilities, and reduces the temptation to grab one seemingly best model and pretend it is the whole truth. However, it raises a difficult question: what to do in the (very common) case of a near-tie.

Suppose we are considering variables A, B, C, D, and E; suppose also that the subset with the lowest BIC is {A,B,C} with a BIC of 34.2, while the second-best is {B,C,D} has a BIC of 34.3. A naïve approach would be to conclude that A is an important predictor and D is not, and then conduct all later estimates and analyses using only the subset {A,B,C}. If we had gathered an even slightly different sample, though, we might be just as likely to make the opposite conclusion. What should we do? Some researchers might just report one model as being the correct one and ignore the other. Many methodologists hold that it is unscientific or even "unconscionable" (Burnham and Anderson [2002, p. 271]) to do this, since it seriously understates the true degree of uncertainty present.

A rather safe approach in this simple example would be to use {A,B,C,D} – this might in fact be the model AIC would recommend in such a situation – but this solution might be less relevant if several models were tied. A sophisticated method might handle this situation automatically by including both A and D, but each with attenuated coefficients. That is, both A and D might be partly *in* and partly *out*, and the resulting model size might be between 3 and 4. This is reminiscent of Bayesian model averaging (Hoeting et al. [1999], Wasserman [2000], Burnham and Anderson [2004]), and of combined shrinkage and selection methods like ridge-with-selection (see Hocking [1976]), garotte (Breiman [1995]), LASSO (Tibshirani [1996], Efron et al. [2004]), and SCAD (Fan and Li [2001]).

Another possibility is to fit each model and report the results of each fit. The theoretical meaning of each model could be discussed. It is also important to consider outliers, possible transformations, and diagnostic plots, as well as expert judgment about the practical usefulness and theoretical plausibility of each model (Henderson and Vellemen [1981]). It has even been suggested (e.g., by Kadane and Lazar [2004]) that model selection criteria should be viewed as allowing a researcher to identify a few good models by filtering out ("deselecting") obviously inferior models, so that the adequate models could then be considered in greater depth. Thus, near-ties should be considered the usual case, not a pathology. In the absence of a clear answer from the data, model choice can also be done subjectively, or by combining prior knowledge and goals with the consideration of one or more model fit measures, and this may be the best approach in many situations (Henderson and Vellemen [1981]), although statistical model selection criteria

remain valuable (Raftery [1995b]).

Some researchers have proposed benchmarks for judging the size of a difference in AIC or BIC between two models, to determine if it is in some sense significant. For example, an AIC difference between two models of less than 2 provides little evidence for choosing one over the other; an AIC difference of 10 or more is strong evidence. Note that the raw difference between two models' AIC values is much more informative than the relative difference. This is because AIC is estimating Kullback-Leibler information plus a somewhat arbitrary data-dependent additive constant. Then, two models with AIC's of 5 and 2 are in some sense about equally different as two models with AIC's of 10005 and 10002 (see Burnham and Anderson [2002]). Benchmarks for BIC differences are very similar, and can conveniently be interpreted as functions of Bayes factors or posterior probabilities (see Raftery [1995a]). Unfortunately, as with comparing likelihoods, comparison of AIC or BIC values requires the assumption that both models have been fit to the same data; this may be problematic in contexts with much missing data.

5 Asymptotic Comparisons

Much effort has gone into determining asymptotic properties of model selection criteria. In a beautiful synthesis of much of the literature, Shao [1997] divided classical model selection methods into three classes based on their asymptotic properties.

Class I includes AIC, C_p, GCV, and leave-one-out cross-validation. These procedures are asymptotically equivalent under certain assumptions. They

can be seen as maximizing (5) with $\lambda \approx 1$. They are generally not consistent and tend to overfit in many situations.

Class II includes BIC, as well as leave-d-out cross-validation where $d/n \to 1$ (i.e., breaking the sample into many very small groups). Class II methods maximize (5) with some λ such that $\lambda \to \infty$ (i.e., λ grows as n grows; e.g., with $\frac{1}{2}\log(n)$ in the case of BIC). The "consistent AIC" (CAIC, not to be confused with the AIC_c described earlier in this review) proposed by Bozdogan [1987], and the Generalized Information Criterion of Rao and Wu [1989], are also related to this class. These procedures are asymptotically consistent and do not tend to overfit, but in practice they may underfit.

Class III includes estimators which try to compromise between Classes I and II. They might include five- or tenfold cross-validation, as well as (5) with λ chosen (either subjectively or in a data-driven way) between 1 and $\frac{1}{2}\log(n)$.

Shao argues that Class I methods are the best available if no fixed-dimensional correct model is possible, and Class II methods are the best available if a fixed-dimensional correct model does exist. Class III might do better than either in practical situations, but its theoretical properties are always poorer than one or the other. To explore this assertion, consider the following two asymptotic scenarios:

Scenario A ($\mathcal{M}$ Fixed, f Contained in $\mathcal{M}$, $n \ggg p$) There is a fixed number p of predictors available and hence a fixed number of models in $\mathcal{M}$. One of the models in $\mathcal{M}$ is true. [5] Also, n is large (i.e., approaching infinity). The goal is to choose the minimal correct subset. Here BIC

is optimal.

Scenario B ($\mathcal{M}$ Growing, f Approximated by $\mathcal{M}$, $n \gg p$) The number of available predictors p grows as n grows (e.g., perhaps $p = \sqrt{n}$); hence $\mathcal{M}_n$ expands. [6] The true model is not in any of the $\mathcal{M}_n$, but as p grows, the list of predictors becomes more and more adequate to approximate the true model. The number of terms that should be included in the final model is limited only by the need to have an adequate sample size to estimate each coefficient well. Thus the goal is to choose the best subset in some practical sense which depends on n, rather than the literally correct subset.

In Scenario A, BIC and similar criteria are consistent model selectors (see Schwarz [1978], Shao [1997]) while AIC and similar estimators overfit and are not consistent in general. Thus asymptotically, AIC will keep all of the good predictors, but might also keep some null predictors, while BIC will keep the good predictors and discard the null predictors. In Scenario A then, BIC and similar procedures have better asymptotic behavior than AIC; by selecting the true model with high probability, they also provide good prediction and estimation.

In Scenario B, choosing an optimal procedure requires a more subtle condition than consistency. Here AIC and similar estimators have a kind of optimality (see Stone [1979], Shibata [1981], Li [1987], Shao [1997],Yang [2003]): they are asymptotically "risk efficient" in the sense that they reduce prediction risk about as much as any selection method can, at least as long as the number of models in $\mathcal{M}$ is not growing too fast, and so they may do

better than BIC.

Ultimately, these asymptotic arguments are very difficult to apply in real research. Neither of the two abstract scenarios apply fully to the actual situation of a researcher. Both are asymptotic arguments and hence assume n is very large, and both, especially the former, assume that p is much smaller than n. Furthermore, it is difficult to decide whether we should suppose that a finite-dimensional true model exists or not. Clearly in most research no simple, fixed-dimensional model could tell us *everything* we could want to know about y, but perhaps such a model could tell us everything we could reasonably expect from our dataset. Usually we do not know whether all the causes or correlates of y have even been measured, so at least in this sense we cannot have a completely true model. However, it is not clear whether an excluded *unobserved* confounder should cause a model to be considered invalid or not. Certainly an extraneous variable changes the coefficient estimates, unless it is orthogonal to all of the observed variables, so that in some sense a model is valid only if it has no confounders (Wold [1993]; see also Cohen et al. [2003] for discussion of confounding variables). However, whether a confounding variable makes the estimates *wrong* depends on our interpretation of the meaning of true value of β. If we consider the β as true constants representing causal effects, then any confounding variable makes a model completely invalid. However, if we simply treat $\mathbf{X}^T\beta$ as an approximation for modeling $E(y|\mathbf{x})$, with $\mathbf{x}$ representing only the observed variables in (1), then nonlinearity or unmodeled interactions are problematic, but excluded covariates are not.

We here deemphasize the question of the true model, as it leads into

a mire of confusing distinctions without generating much insight. In fact, the only reason AIC has better properties than BIC in Scenario B is because its tolerance of many small effects may allow it to approximate the nonparametric truth with less bias as n increases. If certain aspects of the truth are simply unobserved (e.g., missing predictor variables), then AIC has no necessary advantage and tends to be worse than BIC again (see Shao [1997]). AIC does not get any advantage in practice simply from its alleged epistemological sophistication.

Shao, like many researchers, assumes fixed x for his proofs. In the fixed-x case, as in a designed experiment, we only want to get a good prediction of future observations at the same $\mathbf{x}_i$ as our sample. In the random-x case, as in an observational study, we want to get a good prediction for any $\mathbf{x}_i$ from the distribution from which our $\mathbf{x}_i$ came (roughly, to interpolate). However, asymptotic results that hold for fixed-x generally hold for random-x as well, provided that the new x are from the same distribution as the old. Thus the fixed-x assumption is not problematic. On the other hand, if we want to get good predictions for data $\mathbf{x}$ at points far from the sample points (to extrapolate), then the usual asymptotic results might not hold. Unfortunately, almost no attention has been given to extrapolation in model selection research (an exception is Forster [2000]); the problem appears very difficult.

6 Orthogonal-Case Behavior

Much insight into the various model selection approaches can be gained by studying their behavior in a simple special case, that of orthogonal predictors. Imagine that the predictor variables each have mean 0 and sample variance 1, and that the responses have mean 0 (so that we do not have to fit an intercept term) and known homoskedastic conditional variance $E(y|\mathbf{x}) = \sigma^2$. Furthermore, assume that the predictors are orthogonal, i.e., their sample correlation coefficients with each other are exactly 0. Then the design matrix is $\mathbf{X}^T\mathbf{X} = n\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Thus the least-squares coefficient estimates are $\hat{\beta}_j = \left[\mathbf{X}^T\mathbf{y}\right]_j = n^{-1}\sum_{i=1}^{n} x_{ij}y_j$; the standard error for each coefficient is exactly $\sigma/n^{-1/2}$; and the correlation between X_j and Y is $\left(\sigma^2 + \sum_{\ell=1}^{d}\beta_\ell^2\right)^{-1/2}\beta_j$.

Because there is no collinearity, we can consider each coefficient separately, and stepwise testing becomes equivalent to pretesting. If we use the usual t-test of predictor significance then we will include predictor j if and only if it has $|t_j| = \left|\hat{\beta}_j/SE(\hat{\beta}_j)\right| > \sqrt{F^*}$ where F^* is the "F-to-enter" or critical value specified by the user (i.e., about 4 for $\alpha = .05$). Thus, we are applying a thresholding rule: keep $\hat{\beta}_j$ unchanged if $\hat{\beta}_j > \sigma\sqrt{F^*/n}$, or delete x_j (set $\hat{\beta}_j$ to 0) otherwise.

It can also be shown that in this simple case many of the criterion-based model selection methods also become equivalent to testing, with $F^* = 2\lambda$ where λ is as given in (5), e.g., 1 for AIC and $\frac{1}{2}\ln(n)$ for BIC. For example, using AIC is like using an F-to-enter of 2 (Foster and George [1994, p. 1947]). Thus the acting significance levels α for the different methods can

Table 1: Orthogonal-Case Thresholding Behavior of Various Methods

Method	Approx. F-to-Enter	Approx. t-to-Enter		Effective α Level		Approx $\lvert\hat{\beta}\rvert/\sigma$ to Enter	
		n=100	1000	n=100	1000	100	1000
Full model	0	0	0	1	1	0	0
Adjusted R^2	1	1	1	.32	.32	.10	.03
Best AIC, GCV or C_p	2	1.41	1.41	.16	.16	.14	.04
Testing with $\alpha = .05$	4	1.96	1.96	.05	.05	.20	.06
Best BIC	$\ln(n)$	2.15	2.63	.03	.009	.21	.08
Best RIC	$2\ln(p)$	(depends on p)					

be found as $P(|t_j| > \sqrt{F^*}) \approx 2(1-\Phi(\sqrt{F^*}))$ where Φ is the standard normal cumulative distribution function. The acting α levels are shown in Table 1 for the methods of most interest and for two representative sample sizes ($n = 100$ and $n = 1000$).

Table 1 gives an indication of the relative preference for sensitivity or specificity of the various methods. BIC is the most specific but least sensitive of the methods available. AIC is surprisingly liberal if viewed as a test, in keeping with its reputation for overfitting. Best adjusted-R^2 is even less specific than AIC and is similar to a one-standard-error rule.

Considerations of orthogonal-case behavior can also provide some intuitive insight into asymptotic results, especially the ambiguity over whether AIC or BIC provides the better selective and predictive performance. Suppose that some of the true β_j are zero and some are not, and assume that the number of predictors available, as well as the values of their true coefficients, are held constant while n is allowed to increase. Since for BIC the α

level is decreasing with n, BIC will have a progressively smaller Type I error rate, i.e., include fewer noise variables. However, BIC will not exclude many useful variables either, since power increases with n at a rate at least comparable to the decrease in the α level (note that the minimum $\left|\hat{\beta}\right|$ required for entry is shrinking for all methods as n grows). Hence, if we imagine that true coefficients stay the same but n grows, both the Type I and Type II error rates of BIC will approach zero as $n \to \infty$, hence the consistency property of BIC. AIC will do more poorly, since although its Type II error rate goes to zero as $n \to \infty$, its Type I error rate will not. Specifically, AIC will include the useful variables as well as about 16% of the available noise variables. In this scenario, BIC is vastly better than AIC.

However, if n is small and β contains many small coefficients, then using AIC, or even using the full model, would be better than using BIC. Remaining in the orthogonal case, suppose that $\sigma = 1$, $\beta = (.15, .15, .15, .15, 0, 0, 0, 0)$, and $\hat{\beta} \approx \beta$. AIC will include the nonzero predictors while BIC will consider them nonsignificant and delete them, discarding useful information. Thus the relative performance of AIC and BIC unfortunately depends on the unknowable true values of the coefficients.[7] Note also that in the orthogonal case, testing (with a usual α such as .05, .10, or .15) is intermediate in behavior between AIC ($\alpha \approx .16$) and BIC ($\alpha \to 0$).

7 Simulations

Another possible approach to studying the performance of various methods is to test them on simulated data for which the correct answer is known.

For example, if we set the true coefficients to $\boldsymbol{\beta} = [0, 1, 2, 3, 0]$, we will want to see whether a procedure can choose the correct subset (i.e., keep variables B, C, D, remove A, E) and whether its estimates for the remaining coefficients are close to their true values.[8]

We can address several questions using simulation here: How do the various selection methods perform on various goals such as sensitivity, specificity, estimation accuracy, prediction accuracy, and identification of the correct subset? How does the true regression function affect the relative performance of the methods? How does sample size affect the relative performance of the methods? To explore these issues, an extensive series of simulations were performed using `R` (R Development Core Team [2004]).

7.1 Performance Measures

The simulation setting, in which the true answer is known, makes it easy to operationalize several measures of performance.

1. *Sensitivity* is one minus the false deletion rate. For example, suppose that the true coefficients are $[0, 1/2, 1/2, 1/2, 0]$, so that the best model would include predictors 2, 3, and 4, but that our model includes only predictors 3 and 4. Then the false deletion rate is 33% ($\frac{1}{3}$ of the good predictors were lost) so sensitivity is 67% ($\frac{2}{3}$ were kept).

2. *Specificity* is one minus the false inclusion rate. For example, suppose that the true model would include predictors 2, 3, and 4 (and exclude predictors 1 and 5), but our estimate contains 1, 2, 3, and 4. Then the false inclusion rate and the specificity rate would both be 50%, since

of the two inactive predictors, one was included and one was not. We calculate the sensitivity and specificity attained by each method in each simulation and then average them.

3. *Correct model rate* is the proportion of simulations in which exactly the correct subset was identified (i.e., in which sensitivity and specificity were both 100%). If a model is not correctly identified, then it is overfit (has too many predictors), underfit (too few), or misfit (the wrong ones).

4. *Model error* is a measure of how far the model's estimate of $\hat{\mu} = \mathbf{X}\hat{\beta}$ differs from the true $\mu = E(y|\mathbf{x})$ and is defined as

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T E(\mathbf{X}\mathbf{X}^T)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

Since accuracy in estimating $\mathbf{X}\beta$ depends on accuracy in estimating β, model error depends on accurate estimation of the parameters in the full model (1), which includes appropriate decisions on which such parameters are 0. Model error is also closely related to predictive accuracy, since for linear models $\hat{y} = \hat{\mu}$. Specifically,

$$\begin{aligned} E(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^2 &= E(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta})^2 + E(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^2 \\ &= (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T E(\mathbf{X}\mathbf{X}^T)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \sigma^2 \end{aligned} \tag{10}$$

In the simulation context, we can thus use model error as a measure of error both in parameter estimation and in future prediction.

7.2 Data-Generating Models

It is important not to overgeneralize the results of any single simulation; e.g., a procedure that works well at estimating $[0, 1, 2, 3, 0]$ might not work well at estimating $[0, .1, .2, .3, 0]$ or $[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]$. Thus we need to test the methods using several different scenarios. We consider five such scenarios, each having the form $20 + \sum_{j=1}^{12} \beta_j x_j + \varepsilon$ with ε $N(0, 1)$. Consider Scenario One:

$$S_I : \boldsymbol{\beta} = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0].$$

Here there are twelve candidate predictors, all of which are useless, so specificity is vital and sensitivity is unimportant. On the other hand, consider Scenario Two:

$$S_{II} : \boldsymbol{\beta} = [.1, .1, -.1, .1, .1, -.1, .1, .1, -.1, .1, .1, -.1].$$

where $\sigma = 1$. Here all of the predictors are useful, and there is no good basis for excluding any of them, but all of them are quite small. Now specificity is less important, so BIC may do poorly here.

Most simulations in the model selection literature use a somewhat sparse true model where predictors are useless and some are useful, as in

$$S_{III} : \boldsymbol{\beta} = [\frac{1}{2}, 0, \frac{1}{4}, , \ \frac{1}{2}, 0, -\frac{1}{4}, \frac{1}{2}, 0, -\frac{1}{4}, \ \frac{1}{2}, 0, -\frac{1}{4}]$$

or

$$S_{IV} : \boldsymbol{\beta} = [\frac{3}{4}, 0, 0, \ -\frac{3}{4}, 0, 0, \ \frac{3}{4}, 0, 0, \ -\frac{3}{4}, 0, 0].$$

Some have argued (e.g., see Burnham and Anderson [2002]) that it is unrealistic to imagine any β_j as exactly zero (recall the classic controversy over the relevance of point null hypothesis testing in psychology), and more realistic to imagine "tapering" effects having various degrees of practical significance. To address this concern we also include

$$S_V : \boldsymbol{\beta} = [\frac{1}{2}, \frac{1}{4}, \frac{1}{6}, \frac{1}{8}..., \frac{1}{24}].$$

Here no predictor has zero effect, but some have practically negligible effect. This is the sort of situation for which AIC seems well-suited, since we are trying to find a practical model for a given sample size, rather than a supposedly true subset.

7.3 Methods Compared

On each simulation, we tried various methods of variable selection to compare their ability to approximate the true model. These methods included:

- Full model (no selection). We include all 12 predictors.
- Pretesting with $\alpha = .05$ (critical $t \approx 2$), $\alpha = .15$ ($t \approx 1.4$), or $\alpha = .05/12$ (an attempt at a Bonferroni correction; $t \approx 2.86$). After the subset was determined by testing, the parameters were reestimated using the reduced model, because simply using the parameter estimates from the full model could lead to very poor performance in the presence of collinearity.
- Stepwise regression (Efroymson [1960], Miller [1996]), either with $\alpha =$

.05 to enter and $\alpha = .10$ to delete, or $\alpha = .15$ to enter and $\alpha = .25$ to delete.

- Best-subset with adjusted R^2, AIC, AIC_c, C_p, or BIC. This was done using the fast `leaps` package (Lumley [2004]).

- Intercept only (no predictors). This was expected to perform poorly, but it occasionally performed relatively well because of its simplicity and stability; loosely, if we do nothing there is nothing to go wrong.

We also recorded the predictive performance of an "oracle" model (a model which knows in advance exactly what subset is nonzero; this is impossible in real life but easy in a simulation). Lastly, we considered the performance of ridge regression (Hoerl and Kennard [1970]; see Sen and Srivastava [1990] or Hastie et al. [2001]) with λ chosen using generalized cross-validation.

7.4 Results

For each combination of the five true model scenarios, and for three sample sizes (n =26, 260, and 780, representing 2, 20 and 60 times the number of coefficients to be estimated), we generated data 6000 times and used the methods considered in this paper to try to select and estimating the coefficients of the true model. First we randomly generated an $\mathbf{X}$ matrix with moderate correlation among the predictors. Specifically, in each simulation we generated the $\mathbf{X}$ matrix as an $n \times p$ matrix of random normals, with $p = 12$, such that $\text{E}(x_{ij}) = 0$, $\text{Var}(x_{ij}) = 1$, and $\text{Corr}(x_{ij}, x_{ik}) = 0.5^{|j-k|}$.

We then generated the y_i as $\mathbf{x}_i^T\boldsymbol{\beta}$ plus an intercept term $\beta_0 = 20$ and an independent N(0,1) error term.

Specificity and sensitivity rates are shown in Tables 2 and 3 respectively. In each table, we have omitted results from scenarios for which the performance measure is difficult to define; for example, Scenario II is not shown in Table 2 because in that scenario there are no truly zero coefficients.

Specificity depends on n and also on the degree of conservatism of the method. For large n, specificity values tend to be very close to what we would expect from the orthogonal-case results in Section 6, i.e., about 84% for AIC-like methods and much higher for BIC-like methods. Testing methods tend to have a specificity of about one minus their nominal alpha, as we would hope. However, when n is small relative to p, overfitting occurs much more often than we would expect from the nominal α. Notice that AIC_c has better specificity, though somewhat poorer sensitivity, then classic AIC. All methods were reasonably sensitive and specific when n was high, except in the more difficult scenarios such as Scenario II. The least specific methods were the one-standard-error rule and adjusted R^2. The most specific were BIC, RIC, and testing with small α. The most sensitive methods were adjusted R^2 and the one-standard-error rule. The least sensitive method was Bonferroni-corrected testing.

Table 2: Specificity Rates for Selection Methods under Various Scenarios

Scenario	I (All Zero)			III (Sparse)			IV (Sparse)		
$n=$	26	260	780	26	260	780	26	260	780
Full Model	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Test t=1	0.66	0.68	0.68	0.67	0.68	0.68	0.66	0.68	0.68
Test t=2	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
Test t=2.86	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Stepwise α=.05	0.96	0.96	0.97	0.95	0.95	0.95	0.95	0.95	0.95
Stepwise α=.15	0.87	0.88	0.88	0.84	0.85	0.85	0.84	0.85	0.85
Best Adjusted R^2	0.59	0.67	0.68	0.60	0.68	0.68	0.60	0.67	0.68
Best AIC	0.71	0.83	0.84	0.69	0.83	0.84	0.69	0.83	0.84
Best AIC_c	0.85	0.84	0.84	0.85	0.85	0.84	0.86	0.85	0.84
Best BIC	0.86	0.98	0.99	0.82	0.97	0.99	0.82	0.98	0.99
Best RIC	0.96	0.98	0.98	0.93	0.97	0.97	0.92	0.98	0.98
Oracle	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 3: Sensitivity Rates for Selection Methods under Various Scenarios

True Model	II (All Tiny)			III (Sparse))			IV (Sparse)		
$n=$	26	260	780	26	260	780	26	260	780
Full Model	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Test t=1	0.35	0.58	0.85	0.55	0.98	1.00	0.85	1.00	1.00
Test t=2	0.06	0.21	0.53	0.18	0.91	1.00	0.49	1.00	1.00
Test t=2.86	0.00	0.04	0.20	0.03	0.75	0.99	0.15	1.00	1.00
Stepwise α=.05	0.07	0.22	0.50	0.20	0.92	1.00	0.41	1.00	1.00
Stepwise α=.15	0.17	0.39	0.71	0.38	0.96	1.00	0.73	1.00	1.00
Best Adjusted R^2	0.43	0.59	0.85	0.62	0.99	1.00	0.89	1.00	1.00
Best AIC	0.31	0.43	0.73	0.54	0.97	1.00	0.86	1.00	1.00
Best AIC_c	0.18	0.41	0.72	0.37	0.96	1.00	0.76	1.00	1.00
Best BIC	0.17	0.16	0.31	0.39	0.88	1.00	0.77	1.00	1.00
Best RIC	0.08	0.18	0.42	0.24	0.89	1.00	0.60	1.00	1.00
Oracle	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 4: Correct Model Identification Rates for Selection Methods under Various Scenarios

True Model	I (All Zero)			II (All Tiny)			III (Sparse)			IV (Sparse)		
$n=$	26	260	780	26	260	780	26	260	780	26	260	780
Full Model	0.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Test t=1	0.04	0.01	0.01	0.00	0.00	0.14	0.00	0.19	0.22	0.05	0.05	0.05
Test t=2	0.66	0.54	0.55	0.00	0.00	0.00	0.00	0.35	0.80	0.06	0.68	0.67
Test t=2.86	0.96	0.95	0.95	0.00	0.00	0.00	0.00	0.05	0.93	0.00	0.97	0.97
Stepwise α=.05	0.67	0.67	0.69	0.00	0.00	0.00	0.00	0.38	0.81	0.09	0.68	0.68
Stepwise α=.15	0.31	0.30	0.32	0.00	0.00	0.02	0.00	0.40	0.52	0.14	0.29	0.29
Best Adjusted R^2	0.01	0.01	0.01	0.00	0.00	0.14	0.00	0.19	0.22	0.03	0.05	0.05
Best AIC	0.08	0.15	0.16	0.00	0.00	0.03	0.00	0.38	0.50	0.08	0.24	0.26
Best AIC_c	0.21	0.16	0.16	0.00	0.00	0.02	0.00	0.40	0.51	0.20	0.27	0.27
Best BIC	0.33	0.82	0.91	0.00	0.00	0.00	0.00	0.23	0.94	0.16	0.86	0.92
Best RIC	0.70	0.78	0.79	0.00	0.00	0.00	0.00	0.27	0.89	0.16	0.83	0.83

Table 5: Proportions of Correct and Incorrect Choices by Selection Methods under Model III (Sparse)

	$n = 26$				$n = 260$				$n = 780$			
	True	Over	Under	Mis	True	Over	Under	Mis	True	Over	Under	Mis
Full Model	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00
Test t=1	0.00	0.01	0.24	0.74	0.19	0.69	0.02	0.10	0.22	0.78	0.00	0.00
Test t=2	0.00	0.00	0.84	0.16	0.35	0.08	0.46	0.10	0.80	0.19	0.00	0.00
Test t=2.86	0.00	0.00	0.98	0.02	0.05	0.00	0.93	0.01	0.93	0.02	0.05	0.00
Stepwise α=.05	0.00	0.00	0.81	0.19	0.38	0.08	0.43	0.12	0.81	0.19	0.00	0.00
Stepwise α=.15	0.00	0.00	0.51	0.49	0.40	0.33	0.13	0.14	0.52	0.48	0.00	0.00
Best Adjusted R^2	0.00	0.02	0.14	0.84	0.19	0.69	0.02	0.09	0.22	0.78	0.00	0.00
Best AIC	0.00	0.01	0.26	0.73	0.38	0.38	0.11	0.14	0.50	0.50	0.00	0.00
Best AIC_c	0.00	0.00	0.53	0.47	0.40	0.33	0.13	0.14	0.51	0.49	0.00	0.00
Best BIC	0.00	0.00	0.49	0.50	0.23	0.02	0.67	0.08	0.94	0.04	0.02	0.00
Best RIC	0.00	0.00	0.75	0.24	0.27	0.03	0.61	0.09	0.89	0.10	0.01	0.00

Table 6: Proportions of Correct and Incorrect Choices by Selection Methods under Model IV (Sparse)

	$n = 26$				$n = 260$				$n = 780$			
	True	Over	Under	Mis	True	Over	Under	Mis	True	Over	Under	Mis
Full Model	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00
Test t=1	0.05	0.49	0.04	0.42	0.05	0.95	0.00	0.00	0.05	0.95	0.00	0.00
Test t=2	0.06	0.04	0.68	0.23	0.68	0.32	0.00	0.00	0.67	0.33	0.00	0.00
Test t=2.86	0.00	0.00	0.96	0.03	0.97	0.03	0.00	0.00	0.97	0.03	0.00	0.00
Stepwise α=.05	0.09	0.04	0.61	0.26	0.68	0.32	0.00	0.00	0.68	0.32	0.00	0.00
Stepwise α=.15	0.14	0.29	0.15	0.42	0.29	0.71	0.00	0.00	0.29	0.71	0.00	0.00
Best Adjusted R^2	0.03	0.63	0.00	0.34	0.05	0.95	0.00	0.00	0.05	0.95	0.00	0.00
Best AIC	0.08	0.51	0.03	0.39	0.24	0.76	0.00	0.00	0.26	0.74	0.00	0.00
Best AIC_c	0.20	0.21	0.14	0.45	0.27	0.73	0.00	0.00	0.27	0.73	0.00	0.00
Best BIC	0.16	0.29	0.13	0.42	0.86	0.14	0.00	0.00	0.92	0.08	0.00	0.00
Best RIC	0.16	0.09	0.42	0.32	0.83	0.17	0.00	0.00	0.83	0.17	0.00	0.00

Sensitivity, which is analogous to power in the hypothesis testing framework, is strongly related to n; it also depends upon the degree of conservatism of the method and upon the effect size (here, the size of the smallest active coefficient). AIC-like methods (AIC, AIC_c, C_p, GCV) are very similar to each other, and are more sensitive than BIC-like methods when n is modest. The stepwise and testing methods are not very sensitive, which is not surprising in light of the discussion in Section 2.1 and from Figures 1 and 2.

Under most scenarios, BIC acts as we would hope from its asymptotic consistency property, i.e., with large samples it has excellent specificity and sensitivity. The exception was Scenario II, which had been chosen specifically to illustrate the weakness of BIC, its tendency to overlook small but real effects. The idea that BIC is good for identifying models with a few large effects but sometimes not as good as AIC for identifying models with many small effects is consistent with the literature (Lin and Dayton [1997], Shao [1997], Yang [2004], Burnham and Anderson [2002]). Ridge regression and similar stabilization techniques may do well for such diffuse models; see Hocking [1976], Tibshirani [1996], Fu [1998], and Fan and Li [2001] for the relationship of ridge regression to model selection and for combined approaches.

Correct model identification rates are shown in Table 4. They are generally low, since correct decisions must be made for all 12 predictors in order to get the correct final subset. They are also harder to interpret because they are highly dependent on factors which could not be confidently determined in a real-world setting – basically, how well-suited the method is to

the unknowable true model. They tend to be quite low unless the sample size is huge; i.e., with a modest n relative to p there is no reasonable hope of finding the true model regardless of which criterion is used. When n is large, BIC is about the best method among those considered, except in Scenario II. These findings are not surprising in light of the discussion in Section 5. Scenario II, in which all of the effects were real but tiny, was a special case, apparently fooling all of the selection procedures almost all of the time. We would have expected BIC to have trouble with Scenario II, but here apparently the other methods had just as much trouble.

For Scenarios III and IV, a more detailed analysis is given in Tables 7.4 and 7.4 respectively. These tables show the proportions of simulations in which underfitting (false deletions), overfitting (false inclusions) and misfitting (both false deletions and false inclusions) occurred. For small n, methods often underfit or misfit. Underfitting was more common in Scenario III than Scenario IV, since the smallest active true coefficient is much smaller in Scenario IV ($\frac{1}{4}$ vs. $\frac{3}{4}$), i.e., it was harder to distinguish between zero and nonzero coefficients. In both cases, underfitting and misfitting become rare when n is very large, but overfitting remains a problem, except with BIC and RIC.

We now consider mean-squared model estimation error as defined in 4. These error rates, averaged across simulations for each method under each scenario, are shown in Table 7.4. Since the error rates have a skewed distribution, we also show median error rates in Table 7.4.

Table 7: **Mean** Model Errors for Selection Methods under Various Scenarios

True Model	I (All Zero)			II (All Tiny)			III (Sparse)			IV (Sparse)			V (Tapering)		
$n=$	26	260	780	26	260	780	26	260	780	26	260	780	26	260	780
Full Model	0.98	0.05	0.02	0.96	0.05	0.02	0.96	0.05	0.02	0.96	0.05	0.02	0.98	0.05	0.02
Test t=1	0.62	0.04	0.01	0.66	0.06	0.02	0.90	0.05	0.02	0.86	0.05	0.01	0.83	0.06	0.02
Test t=2	0.13	0.01	0.00	0.25	0.09	0.04	0.90	0.06	0.01	0.91	0.03	0.01	1.07	0.11	0.04
Test t=2.86	0.05	0.01	0.00	0.18	0.13	0.08	1.09	0.12	0.01	1.11	0.02	0.01	1.36	0.23	0.08
Stepwise α=.05	0.14	0.01	0.00	0.26	0.08	0.04	0.74	0.06	0.01	0.92	0.03	0.01	0.53	0.07	0.03
Stepwise α=.15	0.30	0.03	0.01	0.40	0.07	0.03	0.78	0.05	0.01	0.76	0.04	0.01	0.56	0.06	0.02
Best Adjusted R^2	0.75	0.04	0.01	0.77	0.06	0.02	0.92	0.05	0.02	0.87	0.05	0.01	0.83	0.06	0.02
Best AIC	0.60	0.03	0.01	0.64	0.07	0.03	0.90	0.05	0.01	0.83	0.04	0.01	0.75	0.06	0.02
Best AIC_c	0.38	0.03	0.01	0.45	0.07	0.03	0.84	0.05	0.01	0.75	0.04	0.01	0.60	0.06	0.02
Best BIC	0.36	0.01	0.00	0.45	0.08	0.05	0.85	0.07	0.01	0.78	0.02	0.01	0.63	0.08	0.03
Best RIC	0.16	0.01	0.00	0.29	0.08	0.05	0.80	0.06	0.01	0.82	0.02	0.01	0.57	0.08	0.03
Intercept Only	0.04	0.00	0.00	0.18	0.14	0.14	1.16	1.13	1.13	1.16	1.13	1.13	1.44	1.40	1.40
Ridge with GCV	0.11	0.01	0.00	0.19	0.04	0.01	0.46	0.05	0.02	0.58	0.05	0.02	0.27	0.04	0.02
Oracle	0.04	0.00	0.00	0.96	0.05	0.02	0.50	0.04	0.01	0.22	0.02	0.01	0.98	0.05	0.02

Table 8: Median Model Errors for Selection Methods under Various Scenarios

True Model	I (All Zero)			II (All Tiny)			III (Sparse)			IV (Sparse)			V (Tapering)		
$n=$	26	260	780	26	260	780	26	260	780	26	260	780	26	260	780
Full Model	0.84	0.05	0.02	0.82	0.05	0.02	0.83	0.05	0.02	0.82	0.05	0.02	0.83	0.05	0.02
Test t=1	0.48	0.04	0.01	0.52	0.06	0.02	0.77	0.05	0.02	0.73	0.04	0.01	0.70	0.06	0.02
Test t=2	0.04	0.01	0.00	0.18	0.09	0.04	0.87	0.06	0.01	0.99	0.02	0.01	1.40	0.10	0.04
Test t=2.86	0.02	0.00	0.00	0.16	0.14	0.07	1.14	0.11	0.01	1.13	0.02	0.01	1.41	0.19	0.07
Stepwise α=.05	0.05	0.00	0.00	0.20	0.08	0.04	0.66	0.05	0.01	1.01	0.02	0.01	0.49	0.07	0.03
Stepwise α=.15	0.22	0.02	0.01	0.31	0.07	0.03	0.69	0.05	0.01	0.67	0.03	0.01	0.49	0.06	0.02
Best Adjusted R^2	0.62	0.04	0.01	0.64	0.06	0.02	0.79	0.05	0.02	0.73	0.04	0.01	0.68	0.05	0.02
Best AIC	0.47	0.03	0.01	0.52	0.06	0.03	0.77	0.05	0.01	0.69	0.04	0.01	0.61	0.06	0.02
Best AIC_c	0.29	0.03	0.01	0.36	0.07	0.03	0.73	0.05	0.01	0.64	0.03	0.01	0.52	0.06	0.02
Best BIC	0.24	0.00	0.00	0.33	0.08	0.05	0.73	0.06	0.01	0.66	0.02	0.01	0.53	0.08	0.03
Best RIC	0.05	0.00	0.00	0.20	0.08	0.05	0.68	0.06	0.01	0.74	0.02	0.01	0.52	0.08	0.03
Intercept Only	0.02	0.00	0.00	0.16	0.14	0.14	1.14	1.13	1.13	1.14	1.13	1.13	1.42	1.40	1.40
Ridge with GCV	0.05	0.00	0.00	0.14	0.03	0.01	0.43	0.05	0.02	0.53	0.05	0.02	0.21	0.04	0.01
Oracle	0.02	0.00	0.00	0.82	0.05	0.02	0.43	0.03	0.01	0.19	0.02	0.01	0.83	0.05	0.02

Table 9: Median **Relative** Model Errors for Selection Methods under Various Scenarios

True Model	I (All Zero)			II (All Tiny)			III (Sparse)			IV (Sparse)			V (Tapering)		
$n=$	26	260	780	26	260	780	26	260	780	26	260	780	26	260	780
Full Model	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Test t=1	0.59	0.79	0.80	0.67	1.17	1.22	0.96	0.95	0.94	0.87	0.86	0.87	0.85	1.11	1.16
Test t=2	0.06	0.15	0.15	0.26	1.74	2.46	0.99	1.05	0.78	1.01	0.49	0.50	1.14	2.04	2.20
Test t=2.86	0.02	0.04	0.05	0.21	2.63	4.47	1.33	2.13	0.72	1.34	0.37	0.38	1.68	4.08	4.42
Stepwise α=.05	0.06	0.11	0.12	0.28	1.54	2.45	0.86	0.99	0.78	1.01	0.50	0.51	0.60	1.42	1.67
Stepwise α=.15	0.28	0.50	0.51	0.42	1.36	1.65	0.89	0.90	0.86	0.78	0.70	0.71	0.63	1.18	1.34
Best Adjusted R^2	0.79	0.82	0.81	0.83	1.16	1.22	0.98	0.95	0.94	0.90	0.88	0.88	0.87	1.06	1.13
Best AIC	0.62	0.59	0.59	0.70	1.33	1.59	0.97	0.90	0.87	0.86	0.73	0.73	0.80	1.16	1.32
Best AIC_c	0.37	0.58	0.59	0.49	1.35	1.61	0.94	0.90	0.87	0.75	0.71	0.72	0.65	1.18	1.33
Best BIC	0.32	0.06	0.05	0.47	1.63	3.23	0.95	1.27	0.73	0.79	0.42	0.40	0.68	1.62	2.14
Best RIC	0.06	0.07	0.08	0.30	1.60	2.82	0.92	1.21	0.75	0.87	0.44	0.44	0.64	1.57	1.88
Intercept Only	0.02	0.04	0.04	0.21	2.91	8.78	1.41	>20	>70	1.43	>20	>70	1.74	>20	>80
Ridge with GCV	0.06	0.09	0.09	0.19	0.63	0.83	0.51	0.96	1.17	0.62	1.06	1.47	0.28	0.76	0.91
Oracle	0.02	0.04	0.04	1.00	1.00	1.00	0.57	0.71	0.71	0.23	0.36	0.36	1.00	1.00	1.00

Table 10: Median Relative *Prediction* Errors for Selection Methods under Various Scenarios

True Model	I (All Zero)			II (All Tiny)			III (Sparse)			IV (Sparse)			V (Tapering)		
$n=$	26	260	780	26	260	780	26	260	780	26	260	780	26	260	780
Full Model	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Test t=1	0.83	0.99	1.00	0.86	1.01	1.00	0.98	1.00	1.00	0.94	0.99	1.00	0.93	1.01	1.00
Test t=2	0.60	0.96	0.99	0.68	1.03	1.02	1.00	1.00	1.00	1.00	0.98	0.99	1.07	1.05	1.02
Test t=2.86	0.57	0.96	0.99	0.65	1.08	1.05	1.15	1.06	1.00	1.15	0.97	0.99	1.31	1.13	1.05
Stepwise α=.05	0.61	0.97	0.99	0.69	1.03	1.02	0.94	1.00	1.00	1.01	0.98	0.99	0.82	1.02	1.01
Stepwise α=.15	0.70	0.98	0.99	0.76	1.02	1.01	0.95	1.00	1.00	0.91	0.99	1.00	0.84	1.01	1.01
Best Adjusted R^2	0.91	0.99	1.00	0.93	1.01	1.00	0.99	1.00	1.00	0.96	0.99	1.00	0.94	1.00	1.00
Best AIC	0.84	0.98	0.99	0.87	1.02	1.01	0.99	1.00	1.00	0.94	0.99	1.00	0.91	1.01	1.00
Best AIC_c	0.74	0.98	0.99	0.79	1.02	1.01	0.97	1.00	1.00	0.90	0.99	1.00	0.85	1.01	1.00
Best GCV	0.80	0.98	0.99	0.84	1.02	1.01	0.98	1.00	1.00	0.91	0.99	1.00	0.89	1.01	1.00
Best BIC	0.72	0.96	0.99	0.78	1.03	1.04	0.98	1.01	1.00	0.91	0.97	0.99	0.87	1.03	1.02
Best RIC	0.62	0.96	0.99	0.70	1.03	1.03	0.96	1.01	1.00	0.94	0.98	0.99	0.85	1.03	1.01
Intercept Only	0.56	0.96	0.99	0.64	1.09	1.12	1.19	2.03	2.09	1.19	2.03	2.09	1.33	2.29	2.36
Ridge with GCV	0.60	0.96	0.99	0.65	0.98	1.00	0.78	1.00	1.00	0.83	1.00	1.01	0.68	0.99	1.00
Oracle	0.56	0.96	0.99	1.00	1.00	1.00	0.82	0.99	1.00	0.67	0.97	0.99	1.00	1.00	1.00

Clearly, sample size strongly affects model error: as n grows, estimation error falls quickly. Model selection is of secondary importance; mainly, it is important that active terms not be deleted, and when n is large it is not very important to delete all the inactive terms. To separate the effect of model selection from the effect of increasing n, in Table 7.4 we show the medians of the ratios of estimation errors between a given method and the full model. Thus an entry of, say, 0.56 here means that in a certain situation, the method had 56% as much error as the full model, so it was somehow $1/.56 = 179\%$ as effective as the full model in estimating μ.

Lastly, in Table 7.4 we show the median relative predictive errors, where relative predictive error is here defined as the model error plus σ^2 (here $\sigma = 1$, divided by the model error for the full model plus σ^2. This estimates the success we would have in actually predicting new observations from the population using our model. For instance, a 0.96 in this table would mean that the method would be expected to have mean squared error of about 96% relative to that of the full model in predicting future y given their x values. These ratios tend to be close to 1 when n is large, since when the model is estimated well the prediction error is dominated by unavoidable random error in the response.

The most interesting columns in Table 7.4 are those corresponding to Scenarios II and V, in which none of the coefficients are truly zero. In these scenarios, deleting a coefficient always adds bias to the model, since we are estimating a nonzero quantity at zero. Thus for a large enough sample size, the best performer in terms of estimation error will be the full model, or equivalently any model selection procedure too sensitive to delete any of the

predictors. Deleting some of the predictors can still reduce error, though, by reducing variance when the sample size is very small.

For a small sample size, simplicity and stability become as important as accuracy. Strangely, for the smallest sample size in Scenario II, the intercept-only model – which was in some sense *furthest* from the truth – gave the *best* estimation, much better than the full model, which was correct but not sufficiently well identified ($n \approx 2p$) to be useful. Apparently, the intercept-only model was simple enough so that, while it could never be very good, it could not be grossly bad either; this is related to the concept of instability discussed in Breiman [1996a].

It might be surprising that estimation error can ever higher for the full model than for a subset, especially in a situation like Model V in which the full model is correct. It is well-known that the least-squares estimate is the best unbiased estimator in the normal model. The explanation is that by adding bias (setting unimportant coefficients to zero) we reduce variance. This shows that the conventional wisdom stating that model selection methods perform poorly when n is not large relative to p is incomplete. Although very true, it fails to mention that ordinary regression on the full model performs at least as poorly and sometimes worse.

8 Discussion

The theoretical literature on model selection is vast. Zucchini [2000] gives an excellent theoretical introduction. Recent papers on basic issues include Zhang [1992], Wasserman [2000], and Kadane and Lazar [2004]. The oldest

methods of variable selection are reviewed in Hocking [1976], and some of the newer ones are reviewed in Shao [1997]; a nice overview is also included in Faraway [2002]. The books by Hastie et al. [2001], Burnham and Anderson [2002], McQuarrie and Tsai [1998], and especially Miller [2002] provide thorough but accessible discussions of model selection. Stine [2004] reviews the concept of "minimum description length," an alternative perspective on model selection based on coding theory.

Clearly, one's choice of model selection method should depend on one's preferences in a model; e.g., if we most value sensitivity, we should choose AIC or even the full model, whereas if we highly value specificity we should choose BIC or RIC. Also important are our prior beliefs about the unobserved true regression coefficients, especially whether they express a few large effects (in which case we should use BIC or RIC) or many small ones (in which case we should use AIC, ridge regression, or perhaps just the full model). Either AIC-like or BIC-like methods might give better prediction than the other class, depending on unknown characteristics of the true population. AIC-like procedures often overfit; and BIC-like procedures sometimes underfit. It is difficult to tell from asymptotic arguments which approach is better, since much depends on what conditions we are willing to assume about the true model, but simulations usually favor BIC.

All forms of data-driven variable selection, whether subjective or computerized, lead to certain theoretical problems which have not yet been resolved, generally based on the problem of model uncertainty. Theoretical discussions of model uncertainty are given in Zhang [1992], Draper [1995], Miller [2002], Chatfield [1995], and Burnham and Anderson [2002]. In gen-

eral, it is not reasonable to claim that a subset selected is definitively true. There may not be a single best subset (Hocking [1976]); the most useful subset may depend upon the use for which the model is intended (see Hocking [1976], Henderson and Vellemen [1981], Gelman and Rubin [1995]; but see also Raftery [1995b]); there may be problems with outliers or a need for transformations (Henderson and Vellemen [1981]); and even if there is a generally best subset we are likely to fail to find it due to sampling error (see Table 4).

Therefore, much attention has recently been given to the problems of spurious findings (Derksen and Keselman [1992], Thompson [1995]), instability (Breiman [1996a]), bias (Miller [2002]), and uncertainty (Chatfield [1995]) in model selection. Subset selection typically proceeds by evaluating some or all of the members of $\mathcal{M}$ on the basis of some criterion of good fit, and then selecting the one which is somehow best. The model is then treated as a true model and used as a basis for further inference. As a result, classical inferences may be invalid. Specifically, since the same data is used choose the model, estimate the coefficients, and test the significance of the coefficients (i.e., we capitalize on chance by using the same data to generate hypotheses and to test them), standard error estimates are biased down. Thus tests done after model selection do not validly control the α level, and confidence intervals do not have the correct coverage. Some researchers have taken this as a severe flaw in all data-driven subset selection and have claimed that such methods should not be used, but this advice is too extreme. Pretesting and refitting as in Section 2.2, as well as subjective and interactive model choice, are also forms of data-driven model selection

and may also invalidate the usual tests. Thus, it would be better to urge careful interpretation than to reject all model selection criteria.

Collinearity interacts with model uncertainty in interesting ways. Collinearity is one reason why predictors should sometimes be deleted: coefficients may become uninterpretable if the corresponding predictors are collinear with other predictors in the model. Suppose that we have two predictors, A and B, such that $B \approx 2A$ and $r_{AB} = .95$. Then, for purposes of fitting the sample data, the equations $Y = 10A$, $Y = 5B$, $Y = 30A - 10B$, $Y = -2A+6B$, etc., are roughly equivalent. The final choice of equation will be somewhat arbitrary and highly dependent on sampling error. Attempts to interpret the significance, magnitude, or even sign of the coefficients for A or B would then be extremely misleading. The same problem can occur in a way that is more difficult to detect if, say, $A + B - C \approx D$, etc; such instability is probably common in real datasets. However, collinearity also degrades the performance of selection methods; in the $B \approx 2A$ example, a data-driven choice of whether A or B is better would be arbitrary. Biases introduced by data-driven selection in the presence of multiple good models (both biases to standard errors, and biases to coefficient estimates themselves) are explored further by Miller [2002]. Model uncertainty in the presence of collinearity can also make interpretation difficult, since now the true value, estimated value, and interpretation of a predictor's coefficient, as well as its effect on fit, all depend on which other predictors are included. Consider the classic "correlation vs. causation" examples sometimes given in introductory classes on regression. For instance, what is the sign of the true regression coefficient for "number of firefighters present" in predicting

"damage done by a fire"? The true answer – not just the estimate – depends on whether "initial size of fire" is in the model. It becomes conceptually confusing to try to estimate "true parameters" without knowing the model to which they belong. Does β_j represent the effect of y when x_j is increased and all other possible predictors are held constant, or only the other predictors in the full model, or only the other predictors in the final selected model, or when all predictors are allowed to covary naturally. The third interpretation is the most accurate mathematically, but it implies that not only the value but also the meaning of the coefficient is data-driven and subject to sampling error! Perhaps the best answer here is that unless we are explicitly and carefully doing causal modeling based on preexisting theory, or analyzing data from a well-designed experiment, regression coefficients should not be thought of as effects at all, but simply functions of the partial correlations of the x_j with y. Thus a careful analyst considers the selected model to be only predictive, or at best descriptive in an exploratory way, not as a final answer.

How can the effects of model uncertainty be reduced? As the simulations suggested, it is much easier to distinguish between good and poor models when n is much larger than p. Thus having a large sample can be very helpful. Another way to deal with model uncertainty is to incorporate prior knowledge in addition to statistical procedures, perhaps by reducing the number of models in $\mathcal{M}$. The size of $\mathcal{M}$ (i.e., 2^p) is reduced by half for each variable that is forced in or out, thus reducing both the computational and theoretical problems caused by model uncertainty. The simplest way to customize a model search is to incorporate expert judgment and force certain

predictors into or out of the model. Most modern statistical software allows this easily. Predictors which are of theoretical interest for their own sake, or which are believed in advance to be strongly related to the response, should be forced in, since their absence would make the final model less meaningful. Incorporating expert opinion in this way is subjective, but certainly not wrong; even before the data analysis begins, an unlimited number of potential predictor variables (e.g., zodiac sign, shoe size) have already been excluded from the study because of prior beliefs about their irrelevance.

Many commentators such as Burnham and Anderson [2002] suggest that if possible we should use a model space consisting of only a few models, each with its own theoretical interpretation and justification given *a priori*, rather than the full list of 2^p models. They claim that the former is more reasonable and scientific, and that the latter, if used naïvely, may degenerate into a sloppy "let the computer sort it out"(Burnham and Anderson [2002, p. 244]) approach, leading to poor performance. An advantage of a very restricted approach is that, by only considering meaningful models, we will not be at risk of choosing a nonsensical model. A serious disadvantage is that one of the models we choose to ignore might possibly be a very good model, and by not considering enough predictors we may introduce substantial modeling bias. Furthermore, it may be very difficult to come up with a short list of good models before doing the analysis. Thus, limiting $\mathcal{M}$ may be helpful but is tricky in practice.

Bayesian methods (other than BIC) are beyond the scope of this paper, but it is important to note that Bayesian theory offers a powerfully sim-

ple approach to the idea of model uncertainty, at least if we are willing to assume that one of the models in $\mathcal{M}$ is true. Then the index of the true model is just one more quantity on which to set a prior and calculate a posterior distribution. Bayesian model selection ideas in general are reviewed in Mitchell and Beauchamp [1988], Weiss [1995], Wasserman [2000], Chipman et al. [2001], and Kadane and Lazar [2004]. An example of their use is given in Volinsky et al. [1997]. Generally, the approach is as follows: For each model, set a prior probability of that model being the true or best one, and set prior distributions for the parameters under each model. Then, find the posterior distribution of β given these priors. This gives not only predictions and parameter estimates but also the posterior probability that a given coefficient is zero or that a given model is the best, potentially a very helpful heuristic result.

Of course, it may be that the model with the best posterior probability may still have low posterior probability (e.g., maybe it has 10% when others have only 5%). This suggests model averaging. Improved predictive performance can often be obtained by averaging the predictions generated by many models (Breiman [1996b]), possibly weighting by the posterior probability or the BIC of each model. More information on model averaging can be found in the review by Clyde and George [2004] or in the excellent and accessible introductions by Wasserman [2000] and Hoeting et al. [1999]; it is also recommended by Burnham and Anderson [2002]. One weakness of model averaging is that the resulting prediction equation is no longer parsimonious and may be harder to interpret. Proponents of model averaging approaches believe that prediction, rather than model simplicity, is of

primary importance (Breiman [2001]).

Among the most important recent developments in frequentist model selection have been new penalty functions which combine shrinkage and selection, considering both the number of nozero coefficients in a model, and their size. Their goal is to combine the stability of ridge regression with the parsimony of subset selection. The most famous such proposal is the LASSO (Tibshirani [1996]) penalty. The SCAD penalty, similar to the LASSO but with less bias, is proposed in Fan and Li [2001]. More research into these new methods is needed.

Notes

[1]When n is quite large, however, we may wish to use *two* holdout samples (Hastie et al. [2001, p. 196]). Randomly divide the data into three sets: Use about 50% for fitting each model and 25% for testing each model. After selecting a model, use the remaining (and so far untouched) 25% to do hypothesis testing and error estimation. While this obviously requires a huge sample, it addresses a problem which almost all other approaches to model selection have difficulty with: how to estimate standard errors and do valid tests when the model is being selected and tested on the same data.

[2]As a caveat, it should be noted that the idea of cross-validation as predicting future model fit relies heavily on the assumption of an independent random sample from a defined population. In social science practice it cannot really be assumed, say, that a sample of n persons divided into two halves is really equivalent to to independent random samples of $n/2$ persons. Both subsamples are part of the same study and are similarly affected by any methodological biases and flaws. Thus cross-validation, even with a holdout sample, does not provide the same level of confirmatory validity as a separately performed confirmatory study. However, the same objection could be made to any other conceivable statistical criterion or adjustment, and cross-validation is better than nothing.

[3]We define p_{in} here somewhat loosely for convenience. In general p_{in} should be the number of free parameters, not the number of predictors; i.e., we also should count β_0 and perhaps σ^2. In most respects this does not matter since β_0 and σ^2 are present in all models.

[4]In fact, just as there is a sense in which BIC does not assume a true model, there is a sense in which AIC does. The arguments used to motivate AIC (i.e., show that using $\lambda = 1$ in (5) is the appropriate magnitude of penalty for estimating relative Kullback-Leibler information), make sense primarily in a context in which all models are assumed to be valid, which is obviously unrealistic (except if all models are nested and the smallest is adequate). In defense of AIC, Burnham and Anderson [2002] argue that the AIC approximation is adequate for relatively good models, and unimportant for poor models (since for poor models $\ell(\mathbf{X}, \mathbf{y}, \hat{\boldsymbol{\beta}})$ is already small, so that (5) will be small even if the penalty is wrong). They use the fact that the AIC is actually a good approximation of a more difficult-to-estimate function called Takeuchi's Information Criterion, which in turn is a good estimator of relative Kullback-Leibler discrepancy even without the assumption of a true model in $\mathcal{M}$ (Takeuchi [1976]; see Shibata [1989]).

[5]A subset M_k can be called *correct* if it is true in the population that

$E(\mathbf{y}|\mathbf{x}_i) = \mathbf{x}_i^T\boldsymbol{\beta}$ for some β having zeroes in at least all of the places corresponding to the predictors excluded from M_k, so that no information is lost by using only M_k. Notice that if any subset is a correct model, then the full model is a correct model (e.g., if $E(y|x_1, x_2, x_3) = 3x_1 + 2x_2$, then $E(y|x_1, x_2, x_3) = 3x_1 + 2x_2 + 0x_3$). It is possible that none of the subsets is a correct model, e.g., if the true regression structure is nonlinear or if it includes interaction terms not being considered. The *true model*, or *smallest correct model* would be the subset containing all of the predictors with nonzero true coefficients and none of the predictors with zero true coefficients. Asymptotically, the true model will also be the lowest-risk model for predicting future data.

[6]Since the number of models in $\mathcal{M}_n$ is 2^p and p is growing, some authors make additional assumptions, e.g., that all models are nested, in order to keep $\mathcal{M}_n$ from growing too fast, for an effective search. It is not known whether this might be giving an unfair advantage to AIC-like methods in Scenario B, since it limits the temptation to overfit.

[7]This is related to the idea that AIC is better than BIC in a "minimax" or "worst-case" sense discussed more technically in Yang [2004] and Leeb and Pötscher [2005]: roughly, their probability of getting much worse predictive performance than the full model, by wrongly deleting valuable predictors, is smaller than that of BIC at any given n. This latter advantage for *AIC* is not too impressive, since the *best* worst-case performance in this sense is trivially available by always using the full model (see Kempthorne [1984]), but it reflects the tension between sensitivity and specificity. There has been some recent research, with mixed results, on the question of whether the strengths of AIC and BIC can be combined ; see Yang [2003], de Luna and Skouras [2003] but theoretically this tradeoff is unavoidable (Shao [1997], Yang [2003], Leeb and Pötscher [2005]).

[8]The concept of a true value of a parameter when the true model is unknown is very slippery. We could either consider it to mean the true value in the full model, or in the chosen model. By the true value in the model we mean the value of that coefficient in the projection of the mean structure onto the span of the data, i.e., the estimates we would obtain if we fit the model to the entire population. Issues of model uncertainty are addressed further in the Discussion.

References

H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281. Akademai Kiado, Budapest, 1973.

M. A. Babyak. What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, 66:411–421, 2004.

H. Bozdogan. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52:345–370, 1987.

L. Breiman. Better subset regression using the nonnegative garotte. *Technometrics*, 37:373–384, 1995.

L. Breiman. Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24:2350–2383, 1996a.

L. Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16: 199–215, 2001.

L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996b.

K. P. Burnham and D. R. Anderson. Multimodel inference: Understanding AIC and BIC in model selection. *Sociological methods and research*, 33: 261–304, 2004.

K. P. Burnham and D. R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer-Verlag, New York, 2nd edition, 2002.

C. Chatfield. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series A*, 158:419–466, 1995.

H. Chipman, E. I. George, and R. E. McCulloch. The practical implementation of Bayesian model selection. In P. Lahiri, editor, *Model Selection*, pages 65–134. Institute for Mathematical Statistics, 2001.

R. Christensen. Testing Fisher, Neyman, Pearson and Bayes. *The American Statistician*, 59:121–126, 2005.

M. Clyde and E. I. George. Model uncertainty. *Statistical Science*, 19:81–94, 2004.

J. Cohen, S. G. West, L. Aiken, and P. Cohen. *Applied multiple regression/correlation analysis for the behavioral sciences*. L. Erlbaum Associates, 3rd edition, 2003.

X. de Luna and K. Skouras. Choosing a model selection strategy. *Scandinavian Journal of Statistics*, 30:113–128, 2003.

S. Derksen and H. J. Keselman. 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, 1992.

D. Draper. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society, Series B*, 57:45–97, 1995.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

M. A. Efroymson. Multiple regression analysis. In *Mathematical Methods for Digital Computers*. Wiley, New York, 1960.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96: 1348–1360, 2001.

J. J. Faraway. *Practical Regression and ANOVA using R*. 2002. URL `http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf`.

M. Forster. Key concepts in model selection: Performance and generalizability. *Journal of Mathematical Psychology*, 44:205–231, 2000.

D. P. Foster and E. I. George. The Risk Inflation Criterion for multiple regression. *Annals of Statistics*, 22:1947–1975, 1994.

D. A. Freedman. A note on screening regression equations. *American Statistician*, 37:152–155, 1983.

L. S. Freedman and D. Pee. Return to a note on screening regression equations. *American Statistician*, 43:279–82, 1989.

W. J. Fu. Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7:397–416, 1998.

G. M. Furnival and R. W. Wilson. Regression by leaps and bounds. *Technometrics*, 16:499–512, 1974.

A. Gelman and D. Rubin. Avoiding model selection in Bayesian social research (in discussion on Raftery). *Sociological Methodology*, 25:165–173, 1995.

E. I. George and D. Foster. Calibration and empirical Bayes variable selection. *Biometrika*, 87:731–747, 2000.

E. J. Hannan and B. G. Quinn. The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B*, 41:190–5, 1979.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: Data mining, inference and prediction*. Springer, New York, 2001.

R. M. Hauser. Better rules for better decisions. *Sociological Methodology*, 25:175–183, 1995.

H. V. Henderson and P. F. Vellemen. Building multiple regression models interactively. *Biometrics*, 37:391–411, 1981.

R. R. Hocking. Criteria for selection of a subset regression: Which one should be used? *Technometrics*, 14:967–970, 1972.

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

A. E. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14:382–417, 1999. URL `http://www.research.att.com/ volinsky/bma.html`.

C. M. Hurvich and C. Tsai. Regression and time series model selection in small samples. *Biometrika*, 76:297–307, 1989.

J. B. Kadane and N. A. Lazar. Methods and criteria for model selection. *Journal of the American Statistical Association*, 99:279–290, 2004.

R. E. Kass and L. Wasserman. A reference Bayesian test for nested hypotheses and its relationship to the Schwartz criterion. *Journal of the American Statistical Association*, 90:928–34, 1995.

P. J. Kempthorne. Admissible variable-selection procedures when fitting regression models by least squares for prediction. *Biometrika*, 71:593–597, 1984.

J. Kuha. AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods and Research*, 33:188–229, 2004.

H. Leeb and B. M. Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, 21:21–59, 2005.

K. Li. Asymptotic optimality for C_p, C_l cross-validation and generalized cross-validation: Discrete index set. *Annals of Statistics*, 15:958–975, 1987.

T. H. Lin and C. M. Dayton. Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics*, 22:249–264, 1997.

T. Lumley. *leaps: regression subset selection*, 2004. R package version 2.7. Uses Fortran code by Alan Miller.

C. L. Mallows. Some comments on C_p. *Technometrics*, 15:661–676, 1973.

A. D. R. McQuarrie and C. Tsai. *Regression and Time Series Model Selection*. World Scientific, Singapore, 1998.

A. J. Miller. The convergence of Efroymson's stepwise regression algorithm. *American Statistician*, 50:180–181, 1996.

A. J. Miller. *Subset selection in regression*. Chapman and Hall, New York, 2nd edition, 2002.

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83:1023–1032, 1988.

S. Olejnik, J. Mills, and H. Keselman. Using Wherry's adjusted R^2 and Mallows' C_p for model selection from all possible regressions. *Journal of Experimental Education*, 68:365–380, 2000.

W. Pan. Akaike's information criterion in generalized estimating equations. *Biometrics*, 57:120–125, 2001.

R Development Core Team. *R: A language and environment for technical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. URL `http://www.R-project.org`. ISBN 3-900051-00-3.

A. E. Raftery. Bayesian model selection in social research. *Sociological Methodology*, 25:111–163, 1995a.

A. E. Raftery. Rejoinder: Model selection is unavoidable in social research. *Sociological Methodology*, 25:185–195, 1995b.

C. R. Rao and Y. Wu. A strongly consistent procedure for model selection in a regression problem. *Biometrika*, 76:369–374, 1989.

R. T. Rust, D. Simester, R. J. Brodie, and V. Nilikant. Model selection criteria: An investigation of relative accuracy, posterior probabilities, and combinations of criteria. *Management Science*, 41:322–333, 1995.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6: 461–464, 1978.

A. Sen and M. Srivastava. *Regression analysis*. Springer-Verlag, New York, 1990.

J. Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–264, 1997.

R. Shibata. An optimal selection of regression variables. *Biometrika*, 68: 45–54, 1981.

R. Shibata. Statistical aspects of model selection. In J. C. Willems, editor, *From Data to Model*, chapter 5. Springer-Verlag, New York, 1989.

R. A. Stine. Model selection using information theory and the *mdl* principle. *Sociological Methods and Research*, 33:230–260, 2004.

M. Stone. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B*, 39:44–47, 1977.

M. Stone. Comments on model selection criteria of Akaike and Schwarz. *Journal of the Royal Statistical Society, Series B*, 41:276–278, 1979.

N. Sugiura. Further analysis of the data by Akaike's Information Criterion and the finite corrections. *Communications in Statistics, Theory, and Methods*, A7:13–26, 1978.

K. Takeuchi. Distribution of information statistics and a criterion of model fitting (in Japanese). *Suri-Kagaku (Mathematical Sciences)*, 153:12–18, 1976.

B. Thompson. Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement*, 55:525–534, 1995.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

C. T. Volinsky, D. Madigan, A. E. Raftery, and R. A. Kronmal. Bayesian model averaging in proportional hazard models: Assessing the risk of a stroke. *Applied Statistics*, 46:433–448, 1997.

L. Wasserman. Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44:92–107, 2000.

D. L. Weakliem. A critique of the Bayesian Information Criterion for model selection. *Sociological Methods and Research*, 27:359–397, 1999.

R. E. Weiss. The influence of variable selection: a Bayesian diagnostic perspective. *Journal of the American Statistical Association*, 90:619–625, 1995.

R. J. Wherry. A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Annals of Mathematical Statistics*, 2:440–457, 1931.

S. Wold. Discussion: PLS in chemical practice. *Technometrics*, 35(2):136–139, 1993.

Y. Yang. Can the strengths of AIC and BIC be shared? Technical report, Department of Statistics Iowa State University, Ames, IA, 2003.

Y. Yang. Prediction/estimation with simple linear models: Is it really that simple? Preprint, Institute for Mathematics and its Applications, University of Minnesota, 2004. URL `citeseer.ist.psu.edu/yang04predictionestimation.html`.

P. Zhang. Inference after variable selection in linear regression models. *Biometrika*, 79:741–746, 1992.

W. Zucchini. An introduction to model selection. *Journal of Mathematical Psychology*, 44:41–61, 2000.