Bootstrap Model Selection

Jun SHAO

In a regression problem, typically there are p explanatory variables possibly related to a response variable, and we wish to select a subset of the p explanatory variables to fit a model between these variables and the response. A bootstrap variable/model selection procedure is to select the subset of variables by minimizing bootstrap estimates of the prediction error, where the bootstrap estimates are constructed based on a data set of size n. Although the bootstrap estimates have good properties, this bootstrap selection procedure is inconsistent in the sense that the probability of selecting the optimal subset of variables does not converge to 1 as $n \to \infty$. This inconsistency can be rectified by modifying the sampling method used in drawing bootstrap observations. For bootstrapping pairs (response, explanatory variable), it is found that instead of drawing n bootstrap observations (a customary bootstrap sampling plan), much less bootstrap observations should be sampled: The bootstrap selection procedure becomes consistent if we draw m bootstrap observations with $m \to \infty$ and $m/n \to 0$. For bootstrapping residuals, we modify the bootstrap sampling procedure by increasing the variability among the bootstrap observations. The consistency of the modified bootstrap selection procedures is established in various situations, including linear models, nonlinear models, generalized linear models, and autoregressive time series. The choice of the bootstrap sample size m and some computational issues are also discussed. Some empirical results are presented.

KEY WORDS: Autoregressive time series; Bootstrap sample size; Generalized linear model; Nonlinear regression; Prediction error.

1. INTRODUCTION

In a regression problem, typically there is a vector $\mathbf{x}$ of p explanatory variables to be used to fit a model between $\mathbf{x}$ and a response variable y. Because some of the components of $\mathbf{x}$ may not be related to y, using all p components of $\mathbf{x}$ does not necessarily produce a better model than using part of the components of $\mathbf{x}$. Because the relative performance of each model (corresponding to a set of components of $\mathbf{x}$) is usually unknown, we have to select a set of explanatory variables (components of $\mathbf{x}$) based on a data set $\{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$, where y_i is the response at $\mathbf{x} = \mathbf{x}_i$. This variable selection problem is equivalent to a model selection problem in which each model corresponds to a particular set of the p components of $\mathbf{x}$.

There exist many variable/model selection procedures in the case where the relationship between $\mathbf{x}$ and y is linear; for example, the Akaike information criterion (AIC) (Akaike 1970); the C_p method (Mallows 1973); the Bayes information criterion (BIC) (Hannan and Quinn 1979; Schwartz 1978); the final prediction error (FPE_λ) method (Shibata 1984); the generalized information criterion (Rao and Wu 1989) and its analogs (Pötscher 1989); the delete-one cross-validation (Allen 1974; Stone 1974); the generalized cross-validation (Craven and Wahba 1979); and the delete-d cross-validation (Burman 1989; Geisser 1975; Shao 1993; Zhang 1993a). This article introduces some selection methods based on the bootstrap.

Besides the theoretical and empirical properties of the bootstrap selection procedures established in this article, there are at least two other reasons to use a bootstrap model selection procedure:

1. In the linear regression context, the bootstrap method provides inference procedures (e.g., confidence sets) that are asymptotically more accurate than those produced by the other methods (Adkins and Hill 1990; Hall 1989). It may be preferable to use the same method both in model selection and in the subsequent inference based on the selected model. In addition, if we use the bootstrap for both model selection and the subsequent inference, then the bootstrap observations generated for model selection can also be used in the subsequent inference; that is, in terms of generating bootstrap observations, there is no extra cost for using a bootstrap model selection procedure when the bootstrap is also used for inference. If a cross-validation method is used for model selection and the bootstrap is used for the subsequent inference, then the extra computations in generating resamples for cross-validating cannot be avoided.

2. The bootstrap selection procedure developed in the linear regression case can be extended, without any theoretical derivation, to more complicated problems such as the nonlinear regression models, generalized linear models, and autoregression models. The cross-validation method, which is also a data-resampling method, can also be easily extended to nonlinear regression and generalized linear models, but not to autoregression models.

In Section 2 we focus on the case where the relationship between $\mathbf{x}$ and y is linear. We consider two different ways of generating bootstrap observations: bootstrapping residuals and bootstrapping pairs $(\mathbf{x}, y)$. The main theoretical study of a bootstrap selection procedure is its consistency; that is, whether the probability of selecting a nonoptimal model vanishes as the sample size n increases to infinity. Finite-sample performances of some bootstrap selection procedures are studied by simulation. We consider more complicated cases in Section 3 and establish some results similar to those in Section 2 in nonlinear regression, generalized linear, and autoregression models.

Jun Shao is Associate Professor, Department of Statistics, University of Wisconsin, Madison, WI 53706. The author would like to thank R. Tibshirani for conversations that led to this study and the referees for helpful comments. The research was supported by NSF Grant DMS-9504425.

Journal of the American Statistical Association
June 1996, Vol. 91, No. 434, Theory and Methods

Our main discovery is that a straightforward application of the bootstrap does not yield a consistent model selection procedure—although some simple modifications can be used to rectify this inconsistency. Consider, for example, the method of bootstrapping pairs. One usually generates n independent and identically distributed (iid) bootstrap observations from $\hat{F}$, the empirical distribution putting mass n^{-1} on each pair $(\mathbf{x}_i, y_i), i = 1, \ldots, n$ (Efron 1982, 1983; Freedman 1981). But our results in Sections 2 and 3 show that this leads to an inconsistent bootstrap selection procedure. A simple modification that results in a consistent bootstrap selection procedure is to generate fewer bootstrap observations from $\hat{F}$. More precisely, if m (instead of n) iid bootstrap observations are generated from $\hat{F}$, then the bootstrap selection procedure is consistent if and only if $m/n \to 0$ and $m \to \infty$. Changing the bootstrap sample size to rectify the inconsistency of the bootstrap has been shown to be successful in various other problems (Arcones and Gine 1989; Deheuvels, Mason, and Shorack 1993; Hall 1990; Huang, Sen, and Shao 1996; Shao 1994; Swanepoel 1986).

2. LINEAR MODELS

Let $\{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ be the available data set, where $\mathbf{x}_i$ is the ith value of a p vector of explanatory variables and y_i is the response at $\mathbf{x}_i$. We confine our study to the case where p is fixed; that is, p does not increase as n increases. The explanatory variable $\mathbf{x}$ is either random or deterministic. In the former case we assume that $(\mathbf{x}_i, y_i), i = 1, \ldots, n$, are iid. In the latter case, we assume that $y_i, i = 1, \ldots, n$, are independent. In both cases, we assume that $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ is of full rank and

$$\mu_i = E(y_i|\mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\beta}, \qquad \operatorname{var}(y_i|\mathbf{x}_i) = \sigma^2, \qquad i = 1, \ldots, n, \quad (1)$$

where $\boldsymbol{\beta}$ is a p vector of unknown parameters.

2.1 The Optimal Model

Let α be a subset of $\{1, \ldots, p\}$ of size p_α and let $\mathbf{x}_{i\alpha}$ (or $\boldsymbol{\beta}_\alpha$) be the subvector of $\mathbf{x}_i$ (or $\boldsymbol{\beta}$) containing the components of $\mathbf{x}_i$ (or $\boldsymbol{\beta}$) indexed by the integers in α. Then a model corresponding to α, called model α for simplicity, is

$$\mu_{i\alpha} = E(y_i|\mathbf{x}_i) = \mathbf{x}_{i\alpha}'\boldsymbol{\beta}_\alpha, \qquad \operatorname{var}(y_i|\mathbf{x}_i) = \sigma^2, \qquad i = 1, \ldots, n. \quad (2)$$

For a given α, model α is not necessarily a correct model in the sense that $E(y_i|\mathbf{x}_i)$ is *actually not always equal to* $\mathbf{x}_{i\alpha}'\boldsymbol{\beta}_\alpha$. If $\boldsymbol{\beta}_\alpha$ contains all nonzero components of $\boldsymbol{\beta}$, then $\mathbf{x}_i'\boldsymbol{\beta} = \mathbf{x}_{i\alpha}'\boldsymbol{\beta}_\alpha$ for any $\mathbf{x}_i$ and model (2) is called a correct model. There may be more than one correct model.

Suppose that under each α, the model is fit using the least squares method; that is, $\boldsymbol{\beta}_\alpha$ is estimated by the least squares estimator (LSE),

$$\hat{\boldsymbol{\beta}}_\alpha = (\mathbf{X}_\alpha'\mathbf{X}_\alpha)^{-1}\mathbf{X}_\alpha'\mathbf{y},$$

where $\mathbf{X}_\alpha = (\mathbf{x}_{1\alpha}, \ldots, \mathbf{x}_{n\alpha})'$ and $\mathbf{y} = (y_1, \ldots, y_n)'$. Then the efficiency of model α can be measured by the average loss,

$$L_n(\alpha) = \frac{1}{n}\sum_{i=1}^{n}(\mu_i - \mathbf{x}_{i\alpha}'\hat{\boldsymbol{\beta}}_\alpha)^2 = \frac{\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_\alpha\|^2}{n},$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)', \hat{\boldsymbol{\mu}}_\alpha = \mathbf{X}_\alpha\hat{\boldsymbol{\beta}}_\alpha$, and $\|\mathbf{a}\| = \sqrt{\mathbf{a}'\mathbf{a}}$ for any vector $\mathbf{a}$. After observing the data, our concern is to select a model $\alpha \in \mathcal{A}$ so that $L_n(\alpha)$ may be as small as possible, where $\mathcal{A}$ is a collection of some subsets of $\{1, \ldots, p\}$. The largest possible $\mathcal{A}$ is the one containing all nonempty subsets of $\{1, \ldots, p\}$. But in a practical problem, we may consider a smaller collection of subsets.

Let z_i be a future response at $\mathbf{x}_i, i = 1, \ldots, n$, and assume that the z_i are independent of the y_i. Then the average conditional expected loss in prediction is

$$\Gamma_n(\alpha) = E\left[\frac{1}{n}\sum_{i=1}^{n}(z_i - \mathbf{x}_{i\alpha}'\hat{\boldsymbol{\beta}}_\alpha)^2 \,\middle|\, \mathbf{y}, \mathbf{X}\right] = \sigma^2 + L_n(\alpha). \quad (3)$$

Thus selecting a model with the smallest $L_n(\alpha)$ over all $\alpha \in \mathcal{A}$ is equivalent to selecting a model with the best prediction ability over all $\alpha \in \mathcal{A}$.

Let $\boldsymbol{\varepsilon} = \mathbf{y} - \boldsymbol{\mu}, \mathbf{H}_\alpha = \mathbf{X}_\alpha(\mathbf{X}_\alpha'\mathbf{X}_\alpha)^{-1}\mathbf{X}_\alpha$, and

$$\Delta_n(\alpha) = \frac{\|\boldsymbol{\mu} - \mathbf{H}_\alpha\boldsymbol{\mu}\|^2}{n}. \quad (4)$$

Then

$$L_n(\alpha) = \Delta_n(\alpha) - \frac{2(\boldsymbol{\mu} - \mathbf{H}_\alpha\boldsymbol{\mu})'\boldsymbol{\varepsilon}}{n} + \frac{\|\mathbf{H}_\alpha\boldsymbol{\varepsilon}\|^2}{n}. \quad (5)$$

When model α is correct, $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} = \mathbf{X}_\alpha\boldsymbol{\beta}_\alpha = \mathbf{H}_\alpha\boldsymbol{\mu}$, $\Delta_n(\alpha) = 0$, and

$$L_n(\alpha) = \frac{\|\mathbf{H}_\alpha\boldsymbol{\varepsilon}\|^2}{n}. \quad (6)$$

Let α_0 be the subset corresponding to the correct model with the smallest size; that is, $\boldsymbol{\beta}_{\alpha 0}$ contains exactly all nonzero components of $\boldsymbol{\beta}$. Then, under (1) and

$$\liminf_{n\to\infty} \Delta_n(\alpha) > 0 \quad \text{for any incorrect model } \alpha, \quad (7)$$

model α_0 is optimal in the sense that it minimizes $L_n(\alpha)$ over $\alpha \in \mathcal{A}$ for sufficiently large n; that is,

$$\lim_{n\to\infty} P\{L_n(\alpha_0) = \min_{\alpha\in\mathcal{A}} L_n(\alpha)\} = 1. \quad (8)$$

Because $L_n(\alpha)$ involves the unknown parameter $\boldsymbol{\beta}$, the optimal α_0 must be estimated. Selecting a model is the same as finding an estimate of α_0. Let $\hat{\alpha}$ be the estimate of α based on a model selection procedure. Then the model selection procedure is said to be consistent if

$$\lim_{n\to\infty} P\{\hat{\alpha} = \alpha_0\} = 1. \quad (9)$$

2.2 Bootstrap Selection Procedures

We now introduce bootstrap model selection procedures (bootstrap estimators of α_0). Under linear model (1), there are different ways of generating bootstrap observations:

1. *Bootstrapping residuals* (Efron 1979). Let $r_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}$ be the ith residual, where $\hat{\boldsymbol{\beta}}$ is the LSE under model (1) (or model $\alpha = \{1, \ldots, p\}$). Generate iid $\varepsilon_1^*, \ldots, \varepsilon_n^*$ from the empirical distribution that puts mass n^{-1} on $(r_i - \bar{r})/\sqrt{1 - p/n}, i = 1, \ldots, n$, where $\bar{r}$ is the average of the r_i. The bootstrap observations under model α are $\{(\mathbf{x}_{i\alpha}, y_{i\alpha}^*), i = 1, \ldots, n\}$, where $y_{i\alpha}^* = \mathbf{x}_{i\alpha}'\hat{\boldsymbol{\beta}}_\alpha + \varepsilon_i^*$, $i = 1, \ldots, n$. The bootstrap analog of $\hat{\boldsymbol{\beta}}_\alpha$ is

$$\hat{\boldsymbol{\beta}}_\alpha^* = (\mathbf{X}_\alpha'\mathbf{X}_\alpha)^{-1}\mathbf{X}_\alpha'\mathbf{y}_\alpha^*,$$

where $\mathbf{y}_\alpha^* = (y_{1\alpha}^*, \ldots, y_{n\alpha}^*)'$.

2. *Bootstrapping pairs* (Efron 1982). Let $\hat{F}$ be the empirical distribution putting mass n^{-1} on each pair $(\mathbf{x}_i, y_i), i = 1, \ldots, n$. Generate iid bootstrap data $\{(\mathbf{x}_i^*, y_i^*), i = 1, \ldots, n\}$ from $\hat{F}$. The bootstrap analog of $\hat{\boldsymbol{\beta}}_\alpha$ is

$$\tilde{\boldsymbol{\beta}}_\alpha^* = (\mathbf{X}_\alpha^{*\prime}\mathbf{X}_\alpha^*)^{-1}\mathbf{X}_\alpha^{*\prime}\mathbf{y}^*, \tag{10}$$

where $\mathbf{X}^* = (\mathbf{x}_1^*, \ldots, \mathbf{x}_n^*)'$ and $\mathbf{y}^* = (y_1^*, \ldots, y_n^*)'$. Under the weak condition that $\mathbf{X}'\mathbf{X} \to \infty, \mathbf{X}^{*\prime}\mathbf{X}^* \to \infty$ almost surely. Hence $(\mathbf{X}_\alpha^{*\prime}\mathbf{X}_\alpha^*)^{-1}$ exists asymptotically. In applications, $\tilde{\boldsymbol{\beta}}_\alpha^*$ can be replaced by $\hat{\boldsymbol{\beta}}_\alpha$ in the event that $(\mathbf{X}_\alpha^{*\prime}\mathbf{X}_\alpha^*)^{-1}$ does not exist.

Bootstrapping residuals is more suitable for the case of deterministic $\mathbf{x}$, whereas bootstrapping pairs is more appropriate for the case of random $\mathbf{x}$. But bootstrapping pairs can also be used for deterministic $\mathbf{x}$ (Efron 1982).

Efron (1982, 1983) derived the following bootstrap estimate of the mean of the prediction error $\Gamma_n(\alpha)$ in (3). First, define the expected excess error under model α by

$$e_n(\alpha) = E\left[\Gamma_n(\alpha) - \frac{\|\mathbf{y} - \mathbf{X}_\alpha\hat{\boldsymbol{\beta}}_\alpha\|^2}{n}\right].$$

Then

$$E[\Gamma_n(\alpha)] = E\,\frac{\|\mathbf{y} - \mathbf{X}_\alpha\hat{\boldsymbol{\beta}}_\alpha\|^2}{n} + e_n(\alpha).$$

A bootstrap estimate of $e_n(\alpha)$ is

$$\hat{e}_n(\alpha) = E_*\left[\frac{\|\mathbf{y} - \mathbf{X}_\alpha\boldsymbol{\beta}_\alpha^*\|^2}{n} - \frac{\|\mathbf{y}^* - \mathbf{X}_\alpha^*\boldsymbol{\beta}_\alpha^*\|^2}{n}\right],$$

where $\boldsymbol{\beta}_\alpha^* = \hat{\boldsymbol{\beta}}_\alpha^*$ or $\tilde{\boldsymbol{\beta}}_\alpha^*$, and E_* is the expectation with respect to the bootstrap sampling described under bootstrapping residuals or bootstrapping pairs. (Note that $\mathbf{X}_\alpha^* = \mathbf{X}_\alpha$ and $\mathbf{y}^* = \mathbf{y}_\alpha^*$ for bootstrapping residuals.) A bootstrap estimate of $E[\Gamma_n(\alpha)]$ is

$$\hat{\Gamma}_n(\alpha) = \frac{\|\mathbf{y} - \mathbf{X}_\alpha\hat{\boldsymbol{\beta}}_\alpha\|^2}{n} + \hat{e}_n(\alpha). \tag{11}$$

This estimator is almost unbiased. Some similar estimates were provided by Bunke and Droge (1984).

It seems natural to define the bootstrap estimate of the optimal model α_0 as the model $\hat{\alpha}_n \in \mathcal{A}$ that minimizes $\hat{\Gamma}_n(\alpha)$. But this procedure is inconsistent:

$$\lim_{n\to\infty} P\{\hat{\alpha}_n = \alpha_0\} < 1, \tag{12}$$

unless $\alpha_0 = \{1, \ldots, p\}$; that is, model (1) is the only correct model. The empirical result in Section 2.5 shows that this inconsistency can be quite serious: The probability in (12) can be very low.

It is interesting to note that the bootstrap selection procedures described here are asymptotically equivalent to the selection procedures using the information criterion, C_p, and delete-one cross-validation methods (see (14) and Shao 1993, form. (3.6)).

2.3 The Reason for Inconsistency

Let $\alpha_p = \{1, \ldots, p\}$ be the largest subset. For any $\alpha \in \mathcal{A}$, define

$$\begin{aligned} D_n(\alpha) &= E[\Gamma_n(\alpha) - \Gamma_n(\alpha_p)] = E[L_n(\alpha) - L_n(\alpha_p)] \\ &= \frac{(p_a - p)\sigma^2}{n} + \Delta_n(\alpha). \end{aligned} \tag{13}$$

Minimizing $E[\Gamma_n(\alpha)]$ or $E[L_n(\alpha)]$ is then the same as minimizing $D_n(\alpha)$. Although the bootstrap estimator $\hat{\Gamma}_n(\alpha)$ in (11) is a reasonably good estimator of $E[\Gamma_n(\alpha)]$, the difference $\hat{\Gamma}_n(\alpha) - \hat{\Gamma}_n(\alpha_p)$ is not a consistent estimator of $D_n(\alpha)$ when $\alpha \neq \alpha_p$ and α is also correct. More precisely, it is shown in the Appendix that when α is correct but $\alpha \neq \alpha_p$,

$$\begin{aligned} \hat{\Gamma}_n(\alpha) - \hat{\Gamma}_n(\alpha_p) = \frac{2(p_\alpha - p)\sigma^2}{n} \\ - \frac{\boldsymbol{\varepsilon}'(\mathbf{H}_\alpha - \mathbf{H}_{\alpha_p})\boldsymbol{\varepsilon}}{n} + o_p\left(\frac{1}{n}\right). \end{aligned} \tag{14}$$

Then, by (13)–(14) and the fact that $\Delta_n(\alpha) = 0$,

$$\frac{\hat{\Gamma}_n(\alpha) - \hat{\Gamma}_n(\alpha_p)}{D_n(\alpha)} \not\to 1 \quad \text{in probability.}$$

This leads to the inconsistency of the bootstrap selection procedures described in Section 2.2.

2.4 Modified Bootstrap Selection Procedures

It is clear that if we can find a consistent estimator $\hat{D}_n(\alpha)$ of $D_n(\alpha)$ in the sense that

$$\frac{\hat{D}_n(\alpha)}{D_n(\alpha)} \to 1 \quad \text{in probability,} \qquad \alpha \in \mathcal{A},$$

then we can drive a consistent model selection procedure. Unfortunately, a consistent estimator of $D_n(\alpha)$ is not available, unless α_p is the only correct model.

Let $\{m_n\}$ be a sequence of integers such that $\lim_{n\to\infty} m_n = \infty$ and $\lim_{n\to\infty} m_n/n = 0$. Then with n pairs of data we can find a consistent estimator of $D_{m_n}(\alpha)$. Under condition (7),

$$\lim_{n\to\infty} P\{L_{m_n}(\alpha_0) = \min_{\alpha\in\mathcal{A}} L_{m_n}(\alpha)\} = 1;$$

that is, $D_n(\alpha)$ and $D_{m_n}(\alpha)$ share the same minimizer α_0 for sufficiently large n. This leads us to obtain consistent model selection procedures by minimizing consistent estimators of $D_{m_n}(\alpha)$ or $E[\Gamma_{m_n}(\alpha)]$.

First, consider bootstrapping pairs. For $m < n$, a simple bootstrap estimator of $E[\Gamma_m(\alpha)]$ is

$$\hat{\Gamma}_{n,m}(\alpha) = E_* \frac{\|\mathbf{y} - \mathbf{X}_\alpha \tilde{\boldsymbol{\beta}}^*_{\alpha,m}\|^2}{n}, \tag{15}$$

where $\tilde{\boldsymbol{\beta}}^*_{\alpha,m}$ is the bootstrap analog of $\hat{\boldsymbol{\beta}}_\alpha$ based on m iid pairs $(\mathbf{x}^*_i, y^*_i)$ generated from the empirical distribution putting mass n^{-1} on $(\mathbf{x}_i, y_i), i = 1, \ldots, n$; that is,

$$\tilde{\boldsymbol{\beta}}^*_{\alpha,m} = \left(\sum_{i=1}^m \mathbf{x}^*_{i\alpha}\mathbf{x}^{*\prime}_{i\alpha}\right)^{-1} \sum_{i=1}^m \mathbf{x}^*_{i\alpha} y^*_i. \tag{16}$$

A modified bootstrap model selection procedure is to select a model $\hat{\alpha}_{n,m} \in \mathcal{A}$ that minimizes $\hat{\Gamma}_{n,m}(\alpha)$.

Next, consider bootstrapping residuals. Unless there is a special structure in the $\mathbf{x}_i$ (see, e.g., Hall 1990 for the case where $\mathbf{x}_i = i/n$), it may not be easy to find a way to bootstrap residuals with a bootstrap sample size $m < n$. In view of the fact that only the first two moments of the bootstrap distribution are involved in $\hat{\Gamma}_{n,m}(\alpha)$ and the fact that

$$E_*\tilde{\boldsymbol{\beta}}^*_{\alpha,m} \approx E_*\tilde{\boldsymbol{\beta}}^*_\alpha, \qquad \text{var}_*\tilde{\boldsymbol{\beta}}^*_{\alpha,m} \approx \frac{n}{m}\,\text{var}_*\tilde{\boldsymbol{\beta}}^*_\alpha,$$

where $\tilde{\boldsymbol{\beta}}^*_\alpha$ and $\tilde{\boldsymbol{\beta}}^*_{\alpha,m}$ are defined in (10) and (16), we can modify the procedure in bootstrapping residuals by multiplying a factor $\sqrt{n/m}$ to the values from which the bootstrap data are generated. That is, let $\varepsilon^*_i, i = 1, \ldots, n$, be iid from the distribution that puts mass n^{-1} on each $\sqrt{n/m}(r_i - \bar{r})/\sqrt{1 - p/n}, i = 1, \ldots, n$, and let $y^*_{i\alpha} = \mathbf{x}'_{i\alpha}\hat{\boldsymbol{\beta}}_\alpha + \varepsilon^*_i$ and $\hat{\boldsymbol{\beta}}^*_{\alpha,m} = (\mathbf{X}'_\alpha\mathbf{X}_\alpha)^{-1}\sum_{i=1}^n \mathbf{x}_{i\alpha}y^*_{i\alpha}$. Then estimate $E[\Gamma_m(\alpha)]$ by

$$\hat{\Gamma}_{n,m}(\alpha) = E_* \frac{\|\mathbf{y} - \mathbf{X}_\alpha \hat{\boldsymbol{\beta}}^*_{\alpha,m}\|^2}{n}.$$

The model selected by this modified bootstrap procedure is still denoted by $\hat{\alpha}_{n,m}$, which minimizes $\hat{\Gamma}_{n,m}(\alpha)$ over $\alpha \in \mathcal{A}$.

For bootstrapping residuals, it can be shown that under the linear model (1),

$$\hat{\Gamma}_{n,m}(\alpha) = \frac{\|\mathbf{y} - \mathbf{X}_\alpha\hat{\boldsymbol{\beta}}_\alpha\|^2}{n} + \frac{p_\alpha}{m(n-p)}\sum_{i=1}^n (r_i - \bar{r})^2.$$

Hence the method of bootstrapping residuals is the same as the generalized information criterion (Rao and Wu 1989); however, this is not true in nonlinear models (see Sec. 3).

As a special case of the general result in Section 3, both modified bootstrap selection procedures are consistent; that is,

$$\lim_{n\to\infty} P\{\hat{\alpha}_{n,m} = \alpha_0\} = 1,$$

provided that m satisfies $m/n \to 0$ and $m \to \infty$.

2.5 A Simulation Study

A simulation study was carried out to examine the finite-sample performance of the selection procedures based on bootstrapping pairs with different m. Model (1) was considered, with $p = 5, n = 40$, and iid standard normal errors $\varepsilon_i, i = 1, \ldots, n$. The first component of each $\mathbf{x}_i$ is 1 and the values of other components of $\mathbf{x}_i$ were taken from the solid waste data example of Gunst and Mason (1980) (see also Shao 1993, table 1). The ratio of a component of $\boldsymbol{\beta}$ over σ was chosen to be ≥ 2. If this ratio is too small, then one must increase the sample size n to show a good performance of any model selection procedure.

The bootstrap estimators $\hat{\Gamma}_{n,m}(\alpha)$ were computed by Monte Carlo with size $B = 100$. The computation was done on an IBM 3090 at University of Ottawa. IMSL subroutines DRNNOA and RNUND were used for random number generation.

Table 1 reports the empirical probabilities (based on 1,000 simulations) of selecting each model using the modified bootstrap with various m. When $m = 40$, the bootstrap procedure is the unmodified bootstrap; that is, the model is selected by minimizing $\hat{\Gamma}_n$ in (11). For comparison, empirical selection probabilities using the C_p and the BIC are included.

The following is a summary of the simulation results in Table 1:

1. The empirical results clearly support the asymptotic result previously stated. First, the unmodified bootstrap selection procedure ($m = 40$) performs poorly unless the optimal model is the full model (the largest model). Second, the modified bootstrap selection procedure with an m smaller than 40 clearly improves the unmodified bootstrap selection procedure unless the optimal model is the full model.
2. The C_p and the unmodified bootstrap selection procedures perform almost the same.
3. The modified bootstrap selection procedure can be substantially better than the BIC.
4. The optimal choice of m depends on the parameter $\boldsymbol{\beta}$.

2.6 Discussions

2.6.1 The Choice of the Bootstrap Sample Size m. The previous discussion indicates that for the consistency of the bootstrap selection procedure, m should satisfy $m \to \infty$ and $m/n \to 0$. For practical uses, m needs to be specified for a fixed n. One restriction on m is that p/m should be reasonably small; we should choose an m so that the least squares fitting of a regression model with p regressors does not have too high a variability.

Zhang (1993b) derived the convergence rates for the C_p and BIC procedures. It would be nice if we could choose $m = m_n$ so that the probability $P\{\hat{\alpha}_{n,m_n} = \alpha_0\}$ converges to 1 in the fastest speed. But such an optimal m_n may depend on model parameters and thus may be very difficult or impossible to determine. For example, the results in Table 1 indicate that if model {1, 2, 3, 4, 5} is not the optimal model, then $m = 15$ is the best choice among all the bootstrap sample sizes considered in the simulation study; otherwise, $m = n = 40$ is the best choice.

Table 1. Selection Probabilities Based on 1000 Simulations

True β'	Model	Method for model selection: Bootstrap $m = 15$	Bootstrap $m = 20$	Bootstrap $m = 25$	Bootstrap $m = 30$	Bootstrap $m = 40$	C_p	BIC
(2, 0, 0, 4, 0)	1, 4*	.951 (.007)	.868 (.011)	.764 (.013)	.651 (.015)	.582 (.016)	.594 (.016)	.804 (.013)
	1, 2, 4	.025 (.005)	.045 (.007)	.067 (.008)	.097 (.009)	.115 (.010)	.110 (.010)	.049 (.007)
	1, 3, 4	.017 (.004)	.061 (.006)	.100 (.009)	.138 (.011)	.137 (.011)	.113 (.010)	.065 (.008)
	1, 4, 5	.006 (.002)	.019 (.004)	.047 (.007)	.060 (.008)	.077 (.008)	.095 (.009)	.057 (.007)
	1, 2, 3, 4	.001 (.001)	.004 (.002)	.013 (.004)	.029 (.005)	.043 (.007)	.028 (.005)	.009 (.003)
	1, 2, 4, 5	.000 (.000)	.002 (.001)	.005 (.002)	.013 (.004)	.018 (.004)	.027 (.005)	.007 (.003)
	1, 3, 4, 5	.000 (.000)	.001 (.001)	.004 (.002)	.011 (.003)	.024 (.005)	.026 (.005)	.008 (.003)
	1, 2, 3, 4, 5	.000 (.000)	.000 (.000)	.000 (.000)	.001 (.001)	.004 (.002)	.007 (.003)	.001 (.001)
(2, 0, 0, 4, 8)	1, 4, 5*	.953 (.007)	.899 (.010)	.820 (.012)	.737 (.014)	.662 (.015)	.690 (.015)	.881 (.010)
	1, 2, 4, 5	.026 (.005)	.051 (.007)	.075 (.008)	.111 (.010)	.134 (.011)	.129 (.011)	.045 (.007)
	1, 3, 4, 5	.021 (.005)	.046 (.007)	.094 (.009)	.130 (.011)	.166 (.012)	.142 (.011)	.067 (.008)
	1, 2, 3, 4, 5	.000 (.000)	.004 (.002)	.011 (.003)	.022 (.005)	.038 (.005)	.039 (.006)	.007 (.003)
(2, 9, 0, 4, 8)	1, 4, 5	.012 (.003)	.006 (.002)	.002 (.002)	.000 (.000)	.002 (.002)	.000 (.000)	.000 (.000)
	1, 2, 4, 5*	.974 (.005)	.956 (.006)	.916 (.009)	.859 (.011)	.815 (.012)	.817 (.012)	.939 (.008)
	1, 3, 4, 5	.002 (.002)	.002 (.002)	.000 (.000)	.000 (.000)	.000 (.000)	.000 (.000)	.000 (.000)
	1, 2, 3, 4, 5	.012 (.003)	.036 (.006)	.082 (.009)	.141 (.011)	.183 (.012)	.183 (.012)	.061 (.008)
(2, 9, 6, 4, 8)	1, 2, 3, 5	.008 (.003)	.000 (.000)	.000 (.000)	.000 (.000)	.000 (.000)	.000 (.000)	.000 (.000)
	1, 2, 4, 5	.012 (.003)	.001 (.001)	.000 (.000)	.000 (.000)	.000 (.000)	.000 (.000)	.000 (.000)
	1, 3, 4, 5	.054 (.007)	.012 (.003)	.005 (.002)	.001 (.001)	.002 (.002)	.000 (.000)	.000 (.000)
	1, 2, 3, 4, 5*	.926 (.008)	.987 (.004)	.995 (.003)	.999 (.001)	.998 (.002)	1.00 (.000)	1.00 (.000)

* The optimal model.
NOTE: The numbers in parentheses are empirical standard deviations. When $m = 40$, the model is selected by minimizing $\hat{\Gamma}_n$ in (11).

In a practical problem, statistical inference usually is required after model selection. The bootstrap sample size m for model selection can then be determined by minimizing an accuracy measure of the inference procedure after model selection. We illustrate this idea by considering the case where a confidence interval for $\mathbf{c}'\boldsymbol{\beta}$ with a fixed vector $\mathbf{c}$ is required after model selection. Under model α, a bootstrap-t confidence interval for $\mathbf{c}_\alpha'\boldsymbol{\beta}_\alpha$ with approximate level $1 - 2a$ $(0 < a < 1/2)$ is

$$[\mathbf{c}_\alpha'\hat{\boldsymbol{\beta}}_\alpha - \hat{\sigma}_\alpha\hat{G}_\alpha^{-1}(1-\alpha), \mathbf{c}_\alpha'\hat{\boldsymbol{\beta}}_\alpha - \hat{\sigma}_\alpha\hat{G}_\alpha^{-1}(a)], \quad (17)$$

where $\mathbf{c}_\alpha$ is the subvector of $\mathbf{c}$ containing the components of $\mathbf{c}$ indexed by the integers in α,

$$\hat{\sigma}_\alpha^2 = \frac{1}{n - p_\alpha}\sum_{i=1}^{n}(y_i - \mathbf{x}_{i\alpha}'\hat{\boldsymbol{\beta}}_\alpha)^2,$$

$\hat{G}_\alpha^{-1}(a)$ is the quantile function of the bootstrap distribution

$$\hat{G}_\alpha(t) = P_*\{\mathbf{c}_\alpha'(\boldsymbol{\beta}_\alpha^* - \hat{\boldsymbol{\beta}}_\alpha)/\sigma_\alpha^* \le t\}, \quad (18)$$

$\boldsymbol{\beta}_\alpha^*$ equals $\hat{\boldsymbol{\beta}}_\alpha^*$ for bootstrapping residuals and $\tilde{\boldsymbol{\beta}}_\alpha^*$ for bootstrapping pairs, and $\hat{\sigma}_\alpha^*$ is the bootstrap analog of $\hat{\sigma}_\alpha$. An important accuracy measure for the confidence interval in (17) is its length,

$$l(\alpha) = \hat{\sigma}_\alpha[\hat{G}_\alpha^{-1}(1-a) - \hat{G}_\alpha^{-1}(a)].$$

Then we can choose an $\hat{m}_n$ and select a model, $\hat{\alpha}_{n,\hat{m}_n}$, by solving

$$l(\hat{\alpha}_{n,\hat{m}_n}) = \min_{\alpha_n \le m \le b_n} l(\hat{\alpha}_{n,m}), \quad (19)$$

where $\hat{\alpha}_{n,m}$ is the model selected using one of the modified bootstrap methods described in Section 2.4, and $\{a_n\}$ and $\{b_n\}$ are two sequences satisfying $a_n \to \infty$ and $b_n/n \to 0$ (e.g., $a_n = \log\log n$ and $b_n = n/\log\log n$). This choice of a_n and b_n ensures that $\hat{m}_n \to \infty$ and $\hat{m}_n/n \to 0$, and hence the selected model $\hat{\alpha}_{n,\hat{m}_n}$ is consistent; that is, (9) holds with $\hat{\alpha} = \hat{\alpha}_{n,\hat{m}_n}$.

A similar result can be obtained if simultaneous confidence intervals for $\mathbf{c}'\boldsymbol{\beta}, \mathbf{c} \in \mathcal{C}$ are required. Hall and Pittelkow (1990) derived the following bootstrap simultaneous confidence intervals:

$$[\mathbf{c}_\alpha'\hat{\boldsymbol{\beta}}_\alpha - \hat{\sigma}_\alpha u_\alpha^-, \mathbf{c}_\alpha'\hat{\boldsymbol{\beta}}_\alpha - \hat{\sigma}_\alpha u_\alpha^+], \qquad \mathbf{c} \in \mathcal{C},$$

where u_α^- and u_α^+ satisfy

$$P_*\{u_\alpha^+ \le \mathbf{c}_\alpha'(\boldsymbol{\beta}_\alpha^* - \hat{\boldsymbol{\beta}}_\alpha)/\hat{\sigma}_\alpha^* \le u_\alpha^- \ \forall\, \mathbf{c} \in \mathcal{C}\} = 1 - 2a.$$

Then $\hat{m}_n$ and $\hat{\alpha}_{n,\hat{m}_n}$ can be obtained by solving (19) with $l(\alpha) = \hat{\sigma}_\alpha(u_\alpha^- - u_\alpha^+)$.

Rao and Tibshirani (1993) developed a method of determining the parameter λ in the FPE_λ method (Shibata 1984), which can also be used here to determine m.

2.6.2 Bootstrap Monte Carlo Approximations. Computation of $\hat{\Gamma}_{n,m}(\alpha)$ in (15) may require a Monte Carlo approximation. For bootstrapping pairs, $\hat{\Gamma}_{n,m}(\alpha)$ can be approximated by

$$\hat{\Gamma}_{n,m}^{(B)}(\alpha) = \frac{1}{B}\sum_{b=1}^{B}\frac{\|\mathbf{y} - \mathbf{X}_\alpha\tilde{\boldsymbol{\beta}}_{\alpha,m}^{*b}\|^2}{n},$$

where B is the Monte Carlo sample size, $\tilde{\beta}^{*b}_{\alpha,m}$ is computed according to (16) with $(\mathbf{x}^*_i, y^*_i)$ replaced by $(\mathbf{x}^{*b}_i, y^{*b}_i)$, and

$$(\mathbf{x}^{*b}_i, y^{*b}_i), \qquad i = 1, \ldots, m, \qquad b = 1, \ldots, B \quad (20)$$

are mB independent bootstrap data generated from the empirical distribution putting mass n^{-1} on $(\mathbf{x}_i, y_i)$. Note that these bootstrap data are used for computing $\hat{\Gamma}^{(B)}_{n,m}(\alpha)$ for all $\alpha \in \mathcal{A}$.

The bootstrap data in (20) can still be used in inference after model selection. In bootstrap inference, such as setting bootstrap confidence interval (17), we need to compute Monte Carlo approximations to $\hat{G}_\alpha$ in (18) and its quantiles. With the bootstrap data in (20), we need only generate $(n - m)B$ additional pairs of bootstrap data,

$$(\mathbf{x}^{*b}_i, y^{*b}_i), \qquad i = m + 1, \ldots, n, \qquad b = 1, \ldots, B.$$

This means that the total number of bootstrap data generated for model selection and the subsequent inference is nB, which is the same as that required in bootstrap inference without performing model selection.

3. GENERAL RESULTS

We now consider more complicated situations where the relationship between the mean response and the explanatory variables can be nonlinear.

3.1 Nonlinear Regression

The following model is an extension of the linear model (1):

$$\mu_i = E(y_i|\mathbf{x}_i) = f(\mathbf{x}_i, \boldsymbol{\beta}), \qquad \operatorname{var}(y_i|\mathbf{x}_i) = \sigma^2, \qquad i = 1, \ldots, n,$$

where f is a known function defined on $\mathcal{X} \times \mathcal{B}$, and $\mathcal{X}$ and $\mathcal{B}$ are admissible sets for $\mathbf{x}_i$ and $\boldsymbol{\beta}$. Let $f_i(\boldsymbol{\beta}) = f(\mathbf{x}_i, \boldsymbol{\beta})$, $\mathcal{A}$ be a collection of some subsets of $\{1, \ldots, p\}$, and let $f_{i\alpha}(\boldsymbol{\beta}_\alpha) = f_\alpha(\mathbf{x}_{i\alpha}, \boldsymbol{\beta}_\alpha)$, where $\alpha \in \mathcal{A}$ and f_α is the restriction of the function f to the admissible set of $(\mathbf{x}_{i\alpha}, \boldsymbol{\beta}_\alpha)$. Let

$$\mathcal{A}_c = \{\alpha \in \mathcal{A}: f_\alpha(\mathbf{x}_\alpha, \boldsymbol{\beta}_\alpha) = f(\mathbf{x}, \boldsymbol{\beta}) \; \forall \, \mathbf{x} \in \mathcal{X}\}$$

be the collection of correct models, and assume that $\mathcal{A}_c$ is nonempty.

The simplest example is $f(\mathbf{x}, \boldsymbol{\beta}) = \phi(\mathbf{x}'\boldsymbol{\beta})$ with a function ϕ on $\mathcal{R}$ (the real line). Then $f_\alpha(\mathbf{x}_\alpha, \boldsymbol{\beta}_\alpha) = \phi(\mathbf{x}'_\alpha \boldsymbol{\beta}_\alpha)$, and the correctness of a model is defined the same as that in Section 2. Another example is

$$\boldsymbol{\beta} = (a, b)', \qquad a \in \mathcal{R}, \qquad b \in [0, \infty),$$

$$\mathbf{x} = (1, z)', \qquad z \in (0, \infty),$$

$$f(\mathbf{x}, \boldsymbol{\beta}) = a + e^{-bz}.$$

In this case, $\mathcal{A} = \{\alpha_i, i = 1, 2, 3\}$,

$$f_{\alpha_1}(\mathbf{x}_{\alpha_1}, \boldsymbol{\beta}_{\alpha_1}) = a + 1, \qquad (b = 0)$$

$$f_{\alpha_2}(\mathbf{x}_{\alpha_2}, \boldsymbol{\beta}_{\alpha_2}) = e^{-bz}, \qquad (a = 0)$$

$$f_{\alpha_3}(\mathbf{x}_{\alpha_3}, \boldsymbol{\beta}_{\alpha_3}) = f(\mathbf{x}, \boldsymbol{\beta}) = a + e^{-bz}.$$

If $a = 0$, then $\mathcal{A}_c = \{\alpha_2, \alpha_3\}$ and $\alpha_0 = \alpha_2$ (the correct model with the smallest size). If $b = 0$, then $\mathcal{A}_c = \{\alpha_1, \alpha_3\}$ and $\alpha_0 = \alpha_1$. If $a \neq 0$ and $b \neq 0$, then $\mathcal{A}_c = \{\alpha_3\}$ and $\alpha_0 = \alpha_3$.

The model selection problem in nonlinear regression is similar to that in linear regression. The model corresponding to $\alpha \in \mathcal{A}$ is

$$\mu_i = E(y_i|\mathbf{x}_i) = f_{i\alpha}(\boldsymbol{\beta}_\alpha), \qquad \operatorname{var}(y_i|\mathbf{x}_i) = \sigma^2, \qquad i = 1, \ldots, n.$$

We wish to select a model that minimizes the loss

$$L_n(\alpha) = \frac{\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_\alpha\|^2}{n}$$

over $\alpha \in \mathcal{A}$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$, $\hat{\boldsymbol{\mu}}_\alpha = (f_{1\alpha}(\hat{\boldsymbol{\beta}}_\alpha), \ldots, f_{n\alpha}(\hat{\boldsymbol{\beta}}_\alpha))'$, and $\hat{\boldsymbol{\beta}}_\alpha$ is the LSE of $\boldsymbol{\beta}_\alpha$. For any function $g(\gamma)$, let $\dot{g}(\gamma) = \partial g(\gamma)/\partial\gamma$ and $\ddot{g}(\gamma) = \partial^2 g(\gamma)/\partial\gamma\partial\gamma'$. Then $\hat{\boldsymbol{\beta}}_\alpha$ is a solution of

$$s_\alpha(\gamma) = \sum_{i=1}^n [y_i - f_{i\alpha}(\gamma)]\dot{f}_{i\alpha}(\gamma) = 0, \qquad \gamma \in \mathcal{B}.$$

Result (8) still holds in this case; that is, the correct model with the smallest size is the optimal model.

The two modified bootstrap model selection procedures in Section 2.4 can be extended to this case. First, consider bootstrapping pairs. Let $(\mathbf{x}^*_i, y^*_i), i = 1, \ldots, m$, be iid from the empirical distribution putting mass n^{-1} to each $(\mathbf{x}_i, y_i), i = 1, \ldots, n$, and let $f^*_{i\alpha}(\gamma) = f_\alpha(\mathbf{x}^*_{i\alpha}, \gamma)$ and

$$s^*_{\alpha,m}(\gamma) = \sum_{i=1}^m [y^*_i - f^*_{i\alpha}(\gamma)]\dot{f}^*_{i\alpha}(\gamma).$$

Define

$$\tilde{\boldsymbol{\beta}}^*_{\alpha,m} = \hat{\boldsymbol{\beta}}_\alpha - [\dot{s}^*_{\alpha,m}(\hat{\boldsymbol{\beta}}_\alpha)]^{-1} s^*_{\alpha,m}(\hat{\boldsymbol{\beta}}_\alpha). \quad (21)$$

Note that the exact bootstrap estimator of $\boldsymbol{\beta}_\alpha$ is the solution of $s^*_{\alpha,m}(\gamma) = 0$ and $\tilde{\boldsymbol{\beta}}^*_{\alpha,m}$ is the result from the first-step iteration in solving the exact bootstrap estimator using Newton's method. Hence $\tilde{\boldsymbol{\beta}}^*_{\alpha,m}$ in (21) is an approximation to the exact bootstrap estimator and is much easier to compute. Define

$$\hat{\Gamma}_{n,m}(\alpha) = E_* \sum_{i=1}^n \frac{[y_i - f_{i\alpha}(\tilde{\boldsymbol{\beta}}^*_{\alpha,m})]^2}{n}.$$

The model selected by this bootstrap procedure is $\hat{\alpha}_{n,m}$, which minimizes $\hat{\Gamma}_{n,m}(\alpha)$ over $\alpha \in \mathcal{A}$.

Next, consider bootstrapping residuals. Let $\varepsilon^*_i, i = 1, \ldots, n$, be iid from the empirical distribution putting mass n^{-1} on each $\sqrt{n/m}(r_i - \bar{r})/\sqrt{1 - p/n}, i = 1, \ldots, n$, where $r_i = y_i - f_i(\hat{\boldsymbol{\beta}}), \hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_\alpha$ with $\alpha = \{1, \ldots, p\}$, and $\bar{r} = \sum_{i=1}^n r_i/n$. Let

$$\hat{\boldsymbol{\beta}}^*_{\alpha,m} = \hat{\boldsymbol{\beta}}_\alpha + [M_\alpha(\hat{\boldsymbol{\beta}}_\alpha)]^{-1} \sum_{i=1}^n \varepsilon^*_i \dot{f}_{i\alpha}(\hat{\boldsymbol{\beta}}_\alpha),$$

where

$$M_\alpha(\gamma) = \sum_{i=1}^n \dot f_{i\alpha}(\gamma)\dot f_{i\alpha}(\gamma)',$$

and let

$$\hat\Gamma_{n,m}(\alpha) = E_* \sum_{i=1}^n \frac{[y_i - f_{i\alpha}(\hat\beta^*_{\alpha,m})]^2}{n}.$$

The model selected by this bootstrap procedure is $\hat\alpha_{n,m}$, which minimizes $\hat\Gamma_{n,m}(\alpha)$ over $\alpha \in \mathcal{A}$.

The following regularity conditions are required in studying the consistency of bootstrap procedures:

C1. For each $\alpha \in \mathcal{A}$, $f_\alpha(\cdot,\cdot)$, $\dot f_\alpha(\cdot,\cdot)$, and $\ddot f_\alpha(\cdot,\cdot)$ are continuous functions on $\mathcal{X}\times\mathcal{B}$.

C2. For each $\alpha \in \mathcal{A}_c$, $\hat\beta_\alpha \to \beta_\alpha$ a.s.

C3. a. For deterministic $\mathbf{x}_i$, $\sup_i \|\mathbf{x}_i\| < \infty$ and $\liminf_n \lambda_{\alpha,n} > 0$, where $\lambda_{\alpha,n}$ is the smallest eigenvalue of $M_\alpha(\beta_\alpha)/n$ and $\alpha \in \mathcal{A}_c$.
b. For random iid $\mathbf{x}_i$, there is a function $h_\alpha(\mathbf{x}_\alpha)$ such that $E[h_\alpha(\mathbf{x}_\alpha)] < \infty$, $\|\dot f(\mathbf{x}_\alpha,\gamma_\alpha)\|^4 \le h_\alpha(\mathbf{x}_\alpha)$ and $\|\ddot f(\mathbf{x}_\alpha,\gamma_\alpha)\|^2 \le h_\alpha(\mathbf{x}_\alpha)$ for $\gamma_\alpha \in \{\gamma_\alpha\colon \|\gamma_\alpha - \beta_\alpha\| \le \varepsilon_0\}$, where $\varepsilon_0 > 0$ is fixed and $\alpha \in \mathcal{A}_c$. Also, $E\dot f(\mathbf{x}_\alpha,\beta_\alpha)\dot f(\mathbf{x}_\alpha,\beta_\alpha)' > 0$.

C4. For any incorrect model α,

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^n [y_i - f_{i\alpha}(\hat\beta_\alpha)]^2 > \sigma^2 \quad \text{a.s.}$$

Regularity conditions C1–C3 are types of conditions in establishing asymptotic normality of $\hat\beta_\alpha$ and its bootstrap analog. Condition C4 is reasonable because

$$\frac{1}{n}\sum_{i=1}^n [y_i - f_{i\alpha}(\hat\beta_\alpha)]^2 \to \sigma^2 \quad \text{a.s.}$$

for any correct model α. Under the linear model (1), C1–C2 are clearly satisfied; C4 is the same as (7); and C3 can be replaced by $\lim_{n\to\infty}\max_{i\le n}\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i = 0$.

The following result shows the consistency of the two bootstrap model selection procedures.

Theorem 1. Assume that conditions C1–C4 hold and that $m\to\infty$ and $m/n \to 0$ as $n\to\infty$. Then (9) holds with $\hat\alpha = \hat\alpha_{n,m}$.

3.2 Generalized Linear Models

A generalized linear model is characterized by the following structure: the responses $y_1,\dots,y_n$ are independent and

$$\mu_i = E(y_i|\mathbf{x}_i) = \mu(\eta_i), \qquad \sigma_i^2 = \operatorname{var}(y_i|\mathbf{x}_i) = \phi\dot\mu(\eta_i), \qquad i = 1,\dots,n, \quad (22)$$

where $\phi > 0$ is an unknown scale parameter; $\mu(\eta)$ is a known differentiable function with derivative $\dot\mu(\eta) > 0$; the η_i are related to $\mathbf{x}_i$, the values of explanatory variables, by a known injective and third-order continuously differentiable link function f,

$$f(\mu(\eta_i)) = \mathbf{x}_i'\beta; \quad (23)$$

and β is a p vector of unknown parameters. Examples of generalized linear models, including logit models, log-linear models, gamma-distributed data models, and survival data models, have been provided by McCullagh and Nelder (1989). The linear model (1) is clearly a special case of model (22)–(23).

Let $\mathcal{A}$ be a collection of subsets of $\{1,\dots,p\}$ and let

$$\mu_i = \mu(\eta_{i\alpha}), \qquad \sigma_i^2 = \phi\dot\mu(\eta_{i\alpha}),$$

$$\eta_{i\alpha} = (f\circ\mu)^{-1}(\mathbf{x}_{i\alpha}'\beta_\alpha), \qquad i = 1,\dots,n,$$

be the model corresponding to α, where the $\mathbf{x}_{i\alpha}$ and β_α are defined the same as before. The correctness of a model is defined the same as that in Section 2, and the optimal model is still the correct model with the smallest size.

Note that in model (22)–(23) the distribution of y_i is not specified. Hence we may not be able to obtain the maximum likelihood estimator of β_α. We consider the general estimation equation approach. That is, under model α, β_α is estimated by $\hat\beta_\alpha$, a solution of

$$\sum_{i=1}^n \mathbf{x}_{i\alpha}\psi(\mathbf{x}_{i\alpha}'\gamma)[y_i - f^{-1}(\mathbf{x}_{i\alpha}'\gamma)] = 0,$$

where ψ is the first-order derivative of $(f\circ\mu)^{-1}$. $\hat\beta_\alpha$ can be called a weighted least squares estimator of β_α.

The modified bootstrap model selection procedures in Section 2.4 can be used here for selecting a model from $\mathcal{A}$; that is, we select a model that minimizes

$$\hat\Gamma_{n,m}(\alpha) = E_*\sum_{i=1}^n \frac{[y_i - \mu(\hat\eta^*_{i\alpha})]^2}{nv_{i\alpha}}$$

over $\alpha\in\mathcal{A}$, where $v_{i\alpha} \equiv \dot\mu(\hat\eta_{i\alpha})$, $\hat\eta_{i\alpha} = (f\circ\mu)^{-1}(\mathbf{x}_{i\alpha}'\hat\beta_\alpha)$, $\hat\eta^*_{i\alpha} = (f\circ\mu)^{-1}(\mathbf{x}_{i\alpha}'\beta^*_\alpha)$, and β^*_α is a bootstrap analog of $\hat\beta_\alpha$ obtained by either bootstrapping residuals or bootstrapping pairs. For bootstrapping residuals, we generate iid $\varepsilon_1^*,\dots,\varepsilon_n^*$ from the distribution putting mass n^{-1} to $\sqrt{n/m}(r_i - \bar r)$, where $r_i = [y_i - f^{-1}(\mathbf{x}_i'\hat\beta)]/\sqrt{v_i}$, and $\hat\beta$ and v_i are $\hat\beta_\alpha$ and $v_{i\alpha}$ with $\alpha = \{1,\dots,p\}$. Then we can define β^*_α to be the linear bootstrap estimator

$$\hat\beta^*_\alpha = \hat\beta_\alpha - \hat M_\alpha^{-1}\sum_{i=1}^n \mathbf{x}_{i\alpha}\psi(\mathbf{x}_{i\alpha}'\hat\beta_\alpha)\sqrt{v_{i\alpha}}\varepsilon_i^*,$$

where $\hat M_\alpha = \sum_{i=1}^n \psi^2(\mathbf{x}_{i\alpha}'\hat\beta_\alpha)v_{i\alpha}\mathbf{x}_{i\alpha}\mathbf{x}_{i\alpha}'$. For bootstrapping pairs, we generate iid pairs $(\mathbf{x}_1^*,y_1^*),\dots,(\mathbf{x}_m^*,y_m^*)$ from the distribution putting mass n^{-1} to each $(\mathbf{x}_i,y_i)$ and define β^*_α to be the linear bootstrap estimator

$$\tilde\beta^*_\alpha = \hat\beta_\alpha - \hat M_\alpha^{-1}\sum_{i=1}^n \mathbf{x}^*_{i\alpha}\psi(\mathbf{x}^{*\prime}_{i\alpha}\hat\beta_\alpha)[y_i^* - f^{-1}(\mathbf{x}^{*\prime}_{i\alpha}\hat\beta_\alpha)].$$

Theorem 2. Assume that conditions C2–C3 hold, with $\dot{f}_{i\alpha}(\boldsymbol{\beta}_\alpha)$ replaced by $\psi(\mathbf{x}'_{i\alpha}\boldsymbol{\beta}_\alpha)\sqrt{\ddot{\mu}(\eta_{i\alpha})}\mathbf{x}_{i\alpha}$, and that

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \frac{[y_i - \mu(\hat{\eta}_{i\alpha})]^2}{v_{i\alpha}} > \phi \quad \text{a.s.}$$

for any incorrect α. If $m \to \infty$ and $m/n \to 0$, then (9) holds, with $\hat{\alpha}$ the model selected by the modified bootstrap.

The proof of Theorem 2 is very similar to the proof of Theorem 1 given in the Appendix and thus is omitted.

3.3 Autoregressive Time Series

A series $\{y_t, t = 0, \pm1, \pm2, \ldots\}$ is called an autoregressive time series of order p if

$$y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + \cdots + \theta_p y_{t-p} + \varepsilon_t, \qquad t = 0, \pm1, \pm2, \ldots, \quad (24)$$

where p is a fixed positive integer, $\theta_i, i = 1, \ldots, p$, are unknown parameters, and the ε_t are iid random variables with mean zero and variance σ^2. The observed data are $y_{1-p}, \ldots, y_0, y_1, \ldots, y_n$.

In many practical problems the order of an autoregressive series is unknown and must be estimated using the data. Estimating the order can be formulated as a model selection problem in which we select a model α from $\mathcal{A} = \{1, \ldots, p\}$ and each α corresponds to the autoregressive model of order α:

$$y_t = \theta_1 y_{t-1} + \cdots + \theta_\alpha y_{t-\alpha} + \varepsilon_t, \qquad t = 0, \pm1, \pm2, \ldots \quad (25)$$

Under model $\alpha, \boldsymbol{\beta}_\alpha = (\theta_1, \ldots, \theta_\alpha)'$ is estimated by the LSE

$$\hat{\boldsymbol{\beta}}_\alpha = \mathbf{S}_\alpha^{-1}\left(\sum_{t=1}^{n} y_{t-1}y_t, \ldots, \sum_{t=1}^{n} y_{t-\alpha}y_t\right)', \quad (26)$$

where

$$\mathbf{S}_\alpha = \sum_{t=1}^{n} \mathbf{z}_{t\alpha}\mathbf{z}'_{t\alpha} \quad \text{and} \quad \mathbf{z}_{t\alpha} = (y_{t-1}, \ldots, y_{t-\alpha})'.$$

We assume that $\alpha = p$ is the largest possible model. The optimal model is

$$\alpha_0 = \max\{j : 1 \le j \le p, \theta_j \ne 0\}.$$

The modified bootstrap model selection procedure can be extend to this problem as follows. Let $\varepsilon_t^*, t = 0, \pm1, \pm2, \ldots$, be iid from the distribution putting mass n^{-1} to $r_i - \bar{r}, i = 1, \ldots, n$, where $r_i = y_i - \mathbf{z}'_{ip}\hat{\boldsymbol{\beta}}_p$ is the ith residual under the largest model $\alpha = p$. The bootstrap analog $\hat{\boldsymbol{\beta}}^*_{\alpha,m}$ of $\hat{\boldsymbol{\beta}}_\alpha$ is defined by (26) with n replaced by m and with y_t replaced by

$$y_t^* = \mathbf{z}'_{t\alpha}\hat{\boldsymbol{\beta}}_\alpha + \varepsilon_t^*, \qquad t = 1 - \alpha, \ldots, 0, 1, \ldots, m,$$

where $\{y^*_{1-2p}, \ldots, y^*_{-p}\} = \{y_{1-p}, \ldots, y_0\}$. The model selected by the bootstrap, denoted by $\hat{\alpha}_{n,m}$, is then the minimizer of

$$\hat{\Gamma}_{n,m}(\alpha) = E_* \sum_{t=1}^{n} \frac{(y_t - \mathbf{z}_{t\alpha}\hat{\boldsymbol{\beta}}^*_{\alpha,m})^2}{n}$$

over $\alpha = 1, \ldots, p$.

Theorem 3. Assume that the roots of $1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_p z^p = 0$ are outside of the unit circle, $E|\varepsilon_1|^{2(s+1)} < \infty$ for some $s \ge 3$, and that $(\varepsilon_1, \varepsilon_1^2)$ satisfies Cramér's condition; that is, for every $\mathbf{c} > 0$, there exists $\delta_c > 0$ such that $\sup_{\|u\|\ge c} |E\exp[i(\varepsilon_1, \varepsilon_1^2)u]| \le e^{-\delta_c}$. If $m \to \infty$ and $m/n \to 0$, then the bootstrap model selection procedure is consistent; that is, (9) holds with $\hat{\alpha} = \hat{\alpha}_{n,m}$.

The result in Theorem 3 can be easily extended to the case where a constant term μ is added to models given by (24) and (25).

4. CONCLUSIONS

We have studied bootstrap model selection procedures in linear regression, nonlinear regression, generalized linear models, and autoregressive time series models. We have shown that the procedure that selects a model by minimizing Efron's (1982, 1983) estimators of prediction error is inconsistent as the sample size tends to infinity. We have proposed two consistent modified bootstrap selection procedures. For bootstrapping pairs, we suggest generating m pairs of bootstrap data; for bootstrapping residuals, we suggest multiplying the residuals by a factor $\sqrt{n/m}$, where m satisfies $m/n \to 0$ and $m \to \infty$.

APPENDIX: PROOFS

Throughout this article, E_* and var_* should be understood as the asymptotic expectation and variance (see Akahira and Takeuchi 1991), conditioned on $y_1, \ldots, y_n$ (and $\mathbf{x}_1, \ldots, \mathbf{x}_n$ if they are random). Thus if ξ_n^* is a function of the bootstrap sample, then

$$E_*\{\xi_n^*[1 + o_p^*(1)]\} = E_*(\xi_n^*)[1 + o_p(1)],$$

where $o_p^*(1)$ denotes a quantity ζ_n^* satisfying $P_*\{|\zeta_n^*| > \varepsilon\} = o_p(1)$ for any $\varepsilon > 0$.

Proof of (14)

We provide a proof only for bootstrapping pairs; the proof for bootstrapping residuals is similar. From the definition of $\tilde{\boldsymbol{\beta}}^*_\alpha$ and $\hat{\boldsymbol{\beta}}_\alpha$,

$$\tilde{\boldsymbol{\beta}}^*_\alpha - \hat{\boldsymbol{\beta}}_\alpha = (\mathbf{X}^{*\prime}_\alpha\mathbf{X}^*_\alpha)^{-1}\sum_{i=1}^{n} \mathbf{x}^*_{i\alpha}(y_i^* - \mathbf{x}^{*\prime}_{i\alpha}\hat{\boldsymbol{\beta}}_\alpha).$$

Because $(\mathbf{X}'_\alpha\mathbf{X}_\alpha)(\mathbf{X}^{*\prime}_\alpha\mathbf{X}^*_\alpha)^{-1} \to 1$ a.s.,

$$\begin{aligned}
\text{var}_*\tilde{\boldsymbol{\beta}}^*_\alpha &= (\mathbf{X}'_\alpha\mathbf{X}_\alpha)^{-1} \\
&\quad \times \text{var}_*\left[\sum_{i=1}^{n} \mathbf{x}^*_{i\alpha}(y_i^* - \mathbf{x}^{*\prime}_{i\alpha}\hat{\boldsymbol{\beta}}_\alpha)\right](\mathbf{X}'_\alpha\mathbf{X}_\alpha)^{-1}[1 + o_p(1)] \\
&= (\mathbf{X}'_\alpha\mathbf{X}_\alpha)^{-1}\sum_{i=1}^{n} \mathbf{x}_{i\alpha}\mathbf{x}'_{i\alpha}(y_i - \mathbf{x}'_{i\alpha}\hat{\boldsymbol{\beta}}_\alpha)^2 \\
&\quad \times (\mathbf{X}'_\alpha\mathbf{X}_\alpha)^{-1}[1 + o_p(1)] \\
&= \sigma^2(\mathbf{X}'_\alpha\mathbf{X}_\alpha)^{-1}[1 + o_p(1)], \qquad (A.1)
\end{aligned}$$

where the last equality holds when $\alpha \in \mathcal{A}_c$ and $\sigma_i^2 = \text{var}(y_i|\mathbf{x}_i) = \sigma^2$. Because $E_*(P_i^0 - P_i^*) = 0$,

$$\begin{aligned}
\hat{e}_n(\alpha) &= E_* \sum_{i=1}^n (P_i^0 - P_i^*)(y_i - \mathbf{x}_{i\alpha}'\tilde{\boldsymbol{\beta}}_\alpha^*)^2 \\
&= 2E_* \sum_{i=1}^n (P_i^0 - P_i^*)(y_i - \mathbf{x}_{i\alpha}'\hat{\boldsymbol{\beta}}_\alpha)\mathbf{x}_{i\alpha}'(\hat{\boldsymbol{\beta}}_\alpha - \tilde{\boldsymbol{\beta}}_\alpha^*) \\
&\quad + E_* \sum_{i=1}^n (P_i^0 - P_i^*)[\mathbf{x}_{i\alpha}'(\hat{\boldsymbol{\beta}}_\alpha - \tilde{\boldsymbol{\beta}}_\alpha^*)]^2 \\
&= 2E_* \sum_{i=1}^n P_i^*(y_i - \mathbf{x}_{i\alpha}'\hat{\boldsymbol{\beta}}_\alpha)\mathbf{x}_{i\alpha}'(\hat{\boldsymbol{\beta}}_\alpha^* - \hat{\boldsymbol{\beta}}_\alpha)[1 + o_p(1)] \\
&\quad + E_*(\hat{\boldsymbol{\beta}}_\alpha - \tilde{\boldsymbol{\beta}}_\alpha^*)' \sum_{i=1}^n (P_i^0 - P_i^*)\mathbf{x}_{i\alpha}\mathbf{x}_{i\alpha}'(\hat{\boldsymbol{\beta}}_\alpha - \tilde{\boldsymbol{\beta}}_\alpha^*) \\
&= \frac{2}{n} E_*(\tilde{\boldsymbol{\beta}}_\alpha^* - \hat{\boldsymbol{\beta}}_\alpha)'(\mathbf{X}_\alpha^{*\prime}\mathbf{X}_\alpha^*)(\tilde{\boldsymbol{\beta}}_\alpha^* - \hat{\boldsymbol{\beta}}_\alpha)[1 + o_p(1)] \\
&\quad + \frac{1}{n} E_*(\tilde{\boldsymbol{\beta}}_\alpha^* - \hat{\boldsymbol{\beta}}_\alpha)'[(\mathbf{X}_\alpha'\mathbf{X}_\alpha) - (\mathbf{X}_\alpha^{*\prime}\mathbf{X}_\alpha^*)](\tilde{\boldsymbol{\beta}}_\alpha^* - \hat{\boldsymbol{\beta}}_\alpha) \\
&= \frac{2}{n} E_*(\tilde{\boldsymbol{\beta}}_\alpha^* - \hat{\boldsymbol{\beta}}_\alpha)'(\mathbf{X}_\alpha'\mathbf{X}_\alpha)(\tilde{\boldsymbol{\beta}}_\alpha^* - \hat{\boldsymbol{\beta}}_\alpha)[1 + o_p(1)] \\
&= \frac{2}{n} \text{tr}[(\mathbf{X}_\alpha'\mathbf{X}_\alpha)\text{var}_*\tilde{\boldsymbol{\beta}}_\alpha^*][1 + o_p(1)].
\end{aligned}$$

For a correct model α, $\|\mathbf{y} - \mathbf{X}_\alpha\hat{\boldsymbol{\beta}}_\alpha\|^2 = \|\boldsymbol{\varepsilon}\|^2 - \boldsymbol{\varepsilon}'\mathbf{H}_\alpha\boldsymbol{\varepsilon}$, and thus

$$\begin{aligned}
\hat{\Gamma}_n(\alpha) &= \frac{\|\mathbf{y} - \mathbf{X}_\alpha\hat{\boldsymbol{\beta}}_\alpha\|^2}{n} + \frac{2\sigma^2}{n} \text{tr}[(\mathbf{X}_\alpha'\mathbf{X}_\alpha)(\mathbf{X}_\alpha'\mathbf{X}_\alpha)^{-1}] \\
&\quad \times [1 + o_p(1)] = \frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{n} - \frac{\boldsymbol{\varepsilon}'\mathbf{H}_\alpha\boldsymbol{\varepsilon}}{n} + \frac{2p_\alpha\sigma^2}{n} + o_p\left(\frac{1}{n}\right),
\end{aligned}$$

assuming that $\sigma_i^2 = \sigma^2$. This proves (14).

Proof of Theorem 1

First, consider bootstrapping pairs. From Conditions C1–C3, (21) gives that for $\alpha \in \mathcal{A}_c$,

$$\begin{aligned}
\text{var}_*(\tilde{\boldsymbol{\beta}}_{\alpha,m}^*) &= \left(\frac{n}{m}\right)^2 [\dot{s}_\alpha(\hat{\boldsymbol{\beta}}_\alpha)]^{-1}\text{var}_*[s_{\alpha,m}^*(\hat{\boldsymbol{\beta}}_\alpha)] \\
&\quad \times [\dot{s}_\alpha(\hat{\boldsymbol{\beta}}_\alpha)]^{-1}[1 + o_p(1)] \\
&= \frac{n}{m} M_\alpha^{-1}(\hat{\boldsymbol{\beta}}_\alpha) \sum_{i=1}^n [y_i - f_{i\alpha}(\hat{\boldsymbol{\beta}}_\alpha)]^2 \dot{f}_{i\alpha}(\hat{\boldsymbol{\beta}}_\alpha) \\
&\quad \times \dot{f}_{i\alpha}(\hat{\boldsymbol{\beta}}_\alpha)' M_\alpha^{-1}(\hat{\boldsymbol{\beta}}_\alpha) + o_p\left(\frac{1}{m}\right) \\
&= \frac{n\sigma^2}{m} M_\alpha^{-1}(\hat{\boldsymbol{\beta}}_\alpha) + o_p\left(\frac{1}{m}\right)
\end{aligned}$$

and

$$\begin{aligned}
&E_* \sum_{i=1}^n \frac{[f_{i\alpha}(\hat{\boldsymbol{\beta}}_\alpha) - f_{i\alpha}(\tilde{\boldsymbol{\beta}}_{\alpha,m}^*)]^2}{n} \\
&= E_* \sum_{i=1}^n \frac{[\dot{f}_{i\alpha}(\hat{\boldsymbol{\beta}}_\alpha)'(\hat{\boldsymbol{\beta}}_\alpha - \tilde{\boldsymbol{\beta}}_{\alpha,m}^*)]^2}{n} [1 + o_p(1)] \\
&= \frac{1}{n} \sum_{i=1}^n \dot{f}_{i\alpha}(\hat{\boldsymbol{\beta}}_\alpha)'\text{var}_*(\tilde{\boldsymbol{\beta}}_{\alpha,m}^*)\dot{f}_{i\alpha}(\hat{\boldsymbol{\beta}}_\alpha) + o_p\left(\frac{1}{m}\right) \\
&= \frac{p_\alpha\sigma^2}{m} + o_p\left(\frac{1}{m}\right).
\end{aligned}$$

Because

$$\sum_{i=1}^n [y_i - f_{i\alpha}(\hat{\boldsymbol{\beta}}_\alpha)]\dot{f}_{i\alpha}(\hat{\boldsymbol{\beta}}_\alpha) = 0,$$

we can similarly show that

$$\begin{aligned}
&E_* \sum_{i=1}^n \frac{[y_i - f_{i\alpha}(\hat{\boldsymbol{\beta}}_\alpha)][f_{i\alpha}(\hat{\boldsymbol{\beta}}_\alpha) - f_{i\alpha}(\tilde{\boldsymbol{\beta}}_{\alpha,m}^*)]}{n} \\
&= E_* \sum_{i=1}^n \frac{[y_i - f_{i\alpha}(\hat{\boldsymbol{\beta}}_\alpha)]\dot{f}_{i\alpha}(\hat{\boldsymbol{\beta}}_\alpha)'(\tilde{\boldsymbol{\beta}}_{\alpha,m}^* - \hat{\boldsymbol{\beta}}_\alpha)}{n} \\
&\quad + o_p\left(\frac{1}{m}\right) = o_p\left(\frac{1}{m}\right).
\end{aligned}$$

Then, for $\alpha \in \mathcal{A}_c$,

$$\hat{\Gamma}_{n,m}(\alpha) = \frac{1}{n} \sum_{i=1}^n [y_i - f_{i\alpha}(\hat{\boldsymbol{\beta}}_\alpha)]^2 + \frac{p_\alpha\sigma^2}{m} + o_p\left(\frac{1}{m}\right).$$

The same result also holds for bootstrapping residuals. Let $\varepsilon_i = y_i - f_i(\boldsymbol{\beta})$. For $\alpha \in \mathcal{A}_c$,

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n [y_i - f_{i\alpha}(\hat{\boldsymbol{\beta}}_\alpha)]^2 &= \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 + \frac{1}{n} \sum_{i=1}^n [f_{i\alpha}(\hat{\boldsymbol{\beta}}_\alpha) - f_{i\alpha}(\boldsymbol{\beta}_\alpha)]^2 \\
&\quad + \frac{2}{n} \sum_{i=1}^n \varepsilon_i[f_{i\alpha}(\hat{\boldsymbol{\beta}}_\alpha) - f_{i\alpha}(\boldsymbol{\beta}_\alpha)],
\end{aligned}$$

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^n [f_{i\alpha}(\hat{\boldsymbol{\beta}}_\alpha) - f_{i\alpha}(\boldsymbol{\beta}_\alpha)]^2 \\
&= \frac{1}{n} \sum_{i=1}^n [\dot{f}_{i\alpha}(\boldsymbol{\beta}_\alpha)'(\hat{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta}_\alpha)]^2 + o_p\left(\frac{1}{n}\right) = O_p\left(\frac{1}{n}\right),
\end{aligned}$$

and

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \varepsilon_i[f_{i\alpha}(\hat{\boldsymbol{\beta}}_\alpha) - f_{i\alpha}(\boldsymbol{\beta}_\alpha)] &= \frac{1}{n} \sum_{i=1}^n \varepsilon_i\dot{f}_{i\alpha}(\boldsymbol{\beta}_\alpha)'(\hat{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta}_\alpha) \\
&\quad + o_p\left(\frac{1}{n}\right) = O_p\left(\frac{1}{n}\right).
\end{aligned}$$

Hence for $\alpha \in \mathcal{A}_c$,

$$\hat{\Gamma}_{n,m}(\alpha) = \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 + \frac{p_\alpha\sigma^2}{m} + o_p\left(\frac{1}{m}\right). \tag{A.2}$$

Because $1/n\sum_{i=1}^n \varepsilon_i^2 \to \sigma^2$ a.s., it follows from C4 that

$$P\left\{\frac{1}{n} \sum_{i=1}^n [y_i - f_{i\alpha}(\hat{\boldsymbol{\beta}}_\alpha)]^2 > \hat{\Gamma}_{n,m}(\alpha_0)\right\} \to 1$$

for $\alpha \notin \mathcal{A}_c$. From the definition of $\hat{\boldsymbol{\beta}}_\alpha$,

$$\hat{\Gamma}_{n,m}(\alpha) \geq \frac{1}{n} \sum_{i=1}^n [y_i - f_{i\alpha}(\hat{\boldsymbol{\beta}}_\alpha)]^2.$$

Hence for $\alpha \notin \mathcal{A}_c$,

$$P\{\hat{\Gamma}_{n,m}(\alpha) > \hat{\Gamma}_{n,m}(\alpha_0)\} \to 1.$$

By (A.2), for $\alpha \in \mathcal{A}_c$ and $\alpha \neq \alpha_0$,

$$P\{\hat{\Gamma}_{n,m}(\alpha) > \hat{\Gamma}_{n,m}(\alpha_0)\} = P\{(p_\alpha - p_{\alpha_0})\sigma^2 + o_p(1) > 0\} \to 1.$$

This proves that (9) holds.

The result for bootstrapping residuals can be shown similarly.

Proof of Theorem 3

Let Σ_α be the $\alpha \times \alpha$ matrix whose (i,j)th element is $\text{cov}(y_i, y_j)/\sigma^2$ and let $\hat{\Sigma}_\alpha$ be the $\alpha \times \alpha$ matrix whose (i,j)th element is $\text{cov}_*(y_i^*, y_j^*)/\text{var}_*(y_i^*)$. Bose (1988) showed that

$$\sup_x |\hat{H}_{\alpha,m}(x) - H_\alpha(x)| \to 0 \quad \text{a.s.,} \tag{A.3}$$

where $H_\alpha(x)$ is the distribution of $\sqrt{n}\Sigma_\alpha^{-1/2}(\hat{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta}_\alpha)$ and $\hat{H}_{\alpha,m}$ is the bootstrap distribution of $\sqrt{m}\hat{\Sigma}_\alpha^{-1/2}(\hat{\boldsymbol{\beta}}^*_{\alpha,m} - \hat{\boldsymbol{\beta}}_\alpha)$.

When $\alpha \geq \alpha_0$, using result (A.3), we obtain that

$$\begin{aligned}
\hat{\Gamma}_{n,m}(\alpha) &= \frac{1}{n}\sum_{t=1}^{n}(y_t - \mathbf{z}'_{t\alpha}\hat{\boldsymbol{\beta}}_\alpha)^2 \\
&\quad + \frac{1}{n}\sum_{t=1}^{n}\mathbf{z}'_{t\alpha}\text{var}_*(\hat{\boldsymbol{\beta}}^*_{\alpha,m})\mathbf{z}_{t\alpha} + o_p\left(\frac{1}{m}\right) \\
&= \frac{1}{n}\sum_{t=1}^{n}(y_t - \mathbf{z}'_{t\alpha}\hat{\boldsymbol{\beta}}_\alpha)^2 + \frac{\text{tr}(\hat{\Sigma}_\alpha^{-1}\mathbf{S}_\alpha)}{mn} + o_p\left(\frac{1}{m}\right) \\
&= \frac{1}{n}\sum_{t=1}^{n}\varepsilon_t^2 + \frac{\text{tr}(\hat{\Sigma}_\alpha^{-1}\mathbf{S}_\alpha)}{mn} + o_p\left(\frac{1}{m}\right),
\end{aligned}$$

where the last equality follows from the fact that when $\alpha \geq \alpha_0$,

$$\mathbf{z}'_{t\alpha}\boldsymbol{\beta}_\alpha = \theta_1 y_{t-1} + \cdots + \theta_p y_{t-p} = y_t - \varepsilon_t$$

and

$$\hat{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta}_\alpha = O_p(n^{-1/2}).$$

Because both $\hat{\Sigma}_\alpha$ and $n^{-1}\mathbf{S}_\alpha/\sigma^2$ converge to Σ_α (Bose 1988), we have

$$\hat{\Gamma}_{n,m}(\alpha) = \frac{1}{n}\sum_{t=1}^{n}\varepsilon_t^2 + \frac{\sigma^2\alpha}{m} + o_p\left(\frac{1}{m}\right). \tag{A.4}$$

When $\alpha < \alpha_0$,

$$\hat{\Gamma}_{n,m}(\alpha) \geq \frac{1}{n}\sum_{t=1}^{n}(y_t - \mathbf{z}'_{t\alpha}\hat{\boldsymbol{\beta}}_\alpha)^2 \tag{A.5}$$

and

$$\liminf_{n\to\infty}\left[\frac{1}{n}\sum_{t=1}^{n}(y_t - \mathbf{z}'_{t\alpha}\hat{\boldsymbol{\beta}}_\alpha)^2 - \frac{1}{n}\sum_{t=1}^{n}(y_t - \mathbf{z}'_{t\alpha_0}\hat{\boldsymbol{\beta}}_{\alpha_0})^2\right] > 0 \quad \text{a.s.} \tag{A.6}$$

(Wei 1992). It follows from (A.4)–(A.6) that

$$P\{\hat{\alpha}_{n,m} = \alpha_0\} \to 1.$$

[Received October 1993. Revised July 1995.]

REFERENCES

Adkins, L. C., and Hill, R. C. (1990), "An Improved Confidence Ellipsoid for the Linear Regression Models," *Journal of Statistical Computation and Simulations*, 36, 9–18.

Akahira, M., and Takeuchi, K. (1991), "On the Definition of Asymptotic Expectation," in *Asymptotic Theory of Statistical Estimation*, ed. M. Akahira, Institute of Mathematics, Univ. of Tsukuba, Japan.

Akaike, H. (1970), "Statistical Predictor Identification," *Annals of the Institute of Statistical Mathematics*, 22, 203–217.

Allen, D. M. (1974), "The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction," *Technometrics*, 16, 125–127.

Arcones, M. A., and Gine, E. (1989), "The Bootstrap of the Mean With Arbitrary Bootstrap Sample Size," *Annals of the Institute of Henri Poincaré*, 25, 457–481.

Bickel, P. J., and Freedman, D. A. (1981), "Some Asymptotic Theory for the Bootstrap," *The Annals of Statistics*, 9, 1196–1217.

Bose, A. (1988), "Edgeworth Correction by Bootstrap in Autoregressions," *The Annals of Statistics*, 16, 1709–1722.

Bunke, O., and Droge, B. (1984), "Bootstrap and Cross-Validation Estimates of the Prediction Error for Linear Regression Models," *The Annals of Statistics*, 12, 1400–1424.

Burman, P. (1989), "A Comparative Study of Ordinary Cross-Validation, v-Hold Cross-Validation and Repeated Learning-Testing Methods," *Biometrika*, 76, 503–514.

Craven, P., and Wahba, G. (1979), "Smoothing Noisy Data With Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation," *Numerical Mathematics*, 31, 377–403.

Deheuvels, P., Mason, D. M., and Shorack, G. R. (1993), "Some Results on the Influence of Extremes on the Bootstrap," *Annals of the Institute Henri Poincaré*, 29, 83–103.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1–26.

——— (1982), *The Jackknife, the Bootstrap, and Other Resampling Plans*, Philadelphia: SIAM.

——— (1983), "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation," *Journal of the American Statistical Association*, 78, 316–331.

Freedman, D. A. (1981), "Bootstrapping Regression Models," *The Annals of Statistics*, 9, 1218–1228.

Geisser, S. (1975), "The Predictive Sample Reuse Method With Applications," *Journal of the American Statistical Association*, 70, 320–328.

Gunst, G. F., and Mason, R. L. (1980), *Regression Analysis and Its Applications*, New York: Marcel Dekker.

Hall, P. (1989), "Unusual Properties of Bootstrap Confidence Intervals in Regression Problem," *Probability Theory and Related Fields*, 81, 247–273.

——— (1990), "Using the Bootstrap to Estimate Mean Squared Error and Select Smoothing Parameters in Nonparametric Problems," *Journal of Multivariate Analysis,* 32, 177–203.

Hall, P., and Pittelkow, Y. E. (1990), "Simultaneous Bootstrap Confidence Bands in Regression," *Journal of Statistical Computation and Simulation*, 37, 99–113.

Hannan, E. J., and Quinn, B. G. (1979), "The Determination of the Order of an Autoregression," *Journal of the Royal Statistical Society*, Ser. B, 41, 190–195.

Huang, J. S., Sen, P. K., and Shao, J. (1996), "Bootstrapping a Sample Quantile When the Density Has a Jump," *Statistica Sinica*, 6, 299–309.

Mallows, C. L. (1973), "Some Comments on C_p," *Technometrics*, 15, 661–675.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, 2nd ed., London: Chapman and Hall.

Pötscher, B. M. (1989), "Model Selection Under Nonstationary: Autoregressive Models and Stochastic Linear Regression Models," *The Annals of Statistics*, 17, 1257–1274.

Rao, C. R., and Wu, Y. (1989), "A Strongly Consistent Procedure for Model Selection in a Regression Problem," *Biometrika*, 76, 369–374.

Rao, J. S., and Tibshirani, R. (1993), "Bootstrap Model Selection via the Cost Complexity Parameter in Regression," technical report, University of Toronto.

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

Shao, J. (1993), "Linear Model Selection by Cross-Validation," *Journal of*

the American Statistical Association, 88, 486–494.
——— (1994), "Bootstrap Sample Size in Non-Regular Cases," *Proceedings of American Mathematical Society*, 122, 1251–1262.
Shibata, R. (1984), "Approximate Efficiency of a Selection Procedure for the Number of Regression Variables," *Biometrika*, 71, 43–49.
Stone, M. (1974), "Cross-Validation Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society*, Ser. B, 36, 111–147.
Swanepoel, J. W. H. (1986), "A Note on Proving That the (Modified) Bootstrap Works, *Communications in Statistics*, *Part A—Theory and Methods*, 15, 3193–3203.
Wei, C. Z. (1992), "On Predictive Least Squares Principles," *The Annals of Statistics*, 20, 1–42.
Zhang, P. (1993a), "Model Selection Via Multifold Cross-Validation," *The Annals of Statistics*, 21, 299–313.
——— (1993b), "On the Convergence Rate of Model Selection Criteria," *Communications in Statistics*, *Part A—Theory and Methods*, 22, 2765–2775.