Noname manuscript No.
(will be inserted by the editor)

Bootstrapping Multiple Linear Regression After Variable Selection

Lasanthi C.R. Pelawa Watagoda · David J. Olive

Received: date / Accepted: date

Abstract This paper suggests a method for bootstrapping the multiple linear regression model $Y = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e$ after variable selection. We develop asymptotic theory for some common least squares variable selection estimators such as forward selection with C_p. Then hypothesis testing is done using three confidence regions, one of which is new. Theory suggests that the three confidence regions tend to have coverage at least as high as the nominal coverage if the sample size is large enough.

Keywords Bagging · Confidence Region · Forward Selection

1 Introduction

In this section we review the variable selection model and some results on bootstrap confidence regions. Section 2 will give large sample theory for some ordinary least squares (OLS) variable selection estimators. Section 3 will give theory for bootstrap confidence regions. Section 4 will show how to bootstrap some variable selection estimators. We assume the number of predictors, p, is fixed.

Suppose that the response variable Y_i and at least one predictor variable $x_{i,j}$ are quantitative with $x_{i,1} \equiv 1$. Let $\boldsymbol{x}_i = (x_{i,1}, ..., x_{i,p})^T$ and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ where β_1 corresponds to the intercept. Then the multiple linear

Lasanthi C.R. Pelawa Watagoda
Department of Mathematical Sciences, Appalachian State University, Boone, NC 28608-2092, USA.
E-mail: lasanthi@appstate.edu

David J. Olive
Department of Mathematics, Southern Illinois University, Mailcode 4408, Carbondale, IL 62901-4408, USA.
E-mail: dolive@siu.edu

regression model is

$$Y_i = \beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$$

for $i = 1, ..., n$. This model is also called the full model. Here n is the sample size, and assume that the random variables e_i are independent and identically distributed (iid) with variance $V(e_i) = \sigma^2$. In matrix notation, these n equations become

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of response variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors. The ith fitted value $\hat{Y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ and the ith residual $r_i = Y_i - \hat{Y}_i$ where $\hat{\boldsymbol{\beta}}$ is an estimator of $\boldsymbol{\beta}$. Ordinary least squares is often used for inference if n/p is large.

Next, we describe variable selection, and then develop theory in Section 2. Variable selection is the search for a subset of predictor variables that can be deleted with little loss of information if n/p is large. Following Olive and Hawkins (2005), a *model for variable selection* can be described by

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S \quad (1)$$

where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$, $\boldsymbol{x}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p - a_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$ and E denotes the subset of terms that can be eliminated given that the subset S is in the model. Let $\boldsymbol{x}_I$ be the vector of a terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). Suppose that S is a subset of I and that model (1) holds. Then

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_{I/S}^T\boldsymbol{\beta}_{(I/S)} + \boldsymbol{x}_O^T\boldsymbol{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I$$

where $\boldsymbol{x}_{I/S}$ denotes the predictors in I that are not in S. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \boldsymbol{0}$ if $S \subseteq I$.

Forward selection forms a sequence of submodels $I_1, ..., I_p$ where I_j uses j predictors including the constant. Let I_1 use $x_1^* = x_1 \equiv 1$: the model has a constant but no nontrivial predictors. To form I_2, consider all models I with two predictors including x_1^*. Compute $Q_2(I) = SSE(I) = RSS(I) = \boldsymbol{r}^T(I)\boldsymbol{r}(I) = \sum_{i=1}^n r_i^2(I) = \sum_{i=1}^n (Y_i - \hat{Y}_i(I))^2$. Let I_2 minimize $Q_2(I)$ for the $p-1$ models I that contain x_1^* and one other predictor. Denote the predictors in I_2 by x_1^*, x_2^*. In general, to form I_j consider all models I with j predictors including variables $x_1^*, ..., x_{j-1}^*$. Compute $Q_j(I) = \boldsymbol{r}^T(I)\boldsymbol{r}(I) = \sum_{i=1}^n r_i^2(I) = \sum_{i=1}^n (Y_i - \hat{Y}_i(I))^2$. Let I_j minimize $Q_j(I)$ for the $p - j + 1$ models I that contain $x_1^*, ..., x_{j-1}^*$ and one other predictor not already selected. Denote the predictors in I_j by $x_1^*, ..., x_j^*$. Continue in this manner for $j = 2, ..., M = p$.

When there is a sequence of M submodels, the final submodel I_d needs to be selected. Let the candidate model I contain a terms, including a constant, and let $\boldsymbol{x}_I$ and $\hat{\boldsymbol{\beta}}_I$ be $a \times 1$ vectors. Then there are many criteria used to select the final submodel I_d. For a given data set, the quantities p, n, and $\hat{\sigma}^2$ act as

constants, and a criterion below may add a constant or be divided by a positive constant without changing the subset I_{min} that minimizes the criterion.

Let criteria $C_S(I)$ have the form

$$C_S(I) = SSE(I) + aK_n\hat{\sigma}^2.$$

These criteria need a good estimator of σ^2 and n/p large. The criterion $C_p(I) = AIC_S(I)$ uses $K_n = 2$ while the $BIC_S(I)$ criterion uses $K_n = \log(n)$. See Jones (1946) and Mallows (1973) for C_p. Typically $\hat{\sigma}^2$ is the OLS full model

$$MSE = \sum_{i=1}^{n} \frac{r_i^2}{n-p}$$

when n/p is large. Then $\hat{\sigma}^2 = MSE$ is a $\sqrt{n}$ consistent estimator of σ^2 under mild conditions by Su and Cook (2012).

The following criteria also need n/p large. AIC is due to Akaike (1973) and BIC to Schwarz (1978).

$$AIC(I) = n\log\left(\frac{SSE(I)}{n}\right) + 2a, \quad \text{and}$$

$$BIC(I) = n\log\left(\frac{SSE(I)}{n}\right) + a\log(n).$$

Let p be fixed and let I_{min} be the submodel that minimizes the criterion using variable selection with OLS. Following Nishii (1984) and Shao (1993), $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$ if C_p or AIC is used for forward selection, backward elimination, or all subsets. Seber and Lee (2003, p. 448) and Claeskens and Hjort (2008) summarize related results. Also see Li (1987).

Inference will consider bootstrap hypothesis testing. Applying certain prediction intervals or prediction regions to the bootstrap sample will result in confidence intervals or confidence regions. See Olive (2018). The prediction intervals and regions are based on samples of size n, while the bootstrap sample size is $B = B_n$. To help motivate this idea, let $Z_{(1)}, ..., Z_{(n)}$ be the order statistics of n iid random variables $Z_1, ..., Z_n$. Let a future random variable Z_f be such that $Z_1, ..., Z_n, Z_f$ are iid. Let $k_1 = \lceil n\delta/2 \rceil$ and $k_2 = \lceil n(1-\delta/2) \rceil$ where $\lceil x \rceil$ is the smallest integer $\geq x$. For example, $\lceil 7.7 \rceil = 8$. Then a common nonparametric large sample $100(1-\delta)\%$ prediction interval (PI) for Z_f is $[Z_{(k_1)}, Z_{(k_2)}]$ where $0 < \delta < 1$. See Frey (2013) for references. Let T_n be an estimator of a parameter θ such as $T_n = \overline{Z} = \sum_{i=1}^{n} Z_i/n$ with $\theta = E(Z_1)$. Let $T_1^*, ..., T_B^*$ be a bootstrap sample for T_n. Then a bootstrap percentile method large sample $100(1-\delta)\%$ confidence interval for θ is an interval $[T^*_{(k_L)}, T^*_{(K_U)}]$ containing $\approx \lceil B(1-\delta) \rceil$ of the T_i^*. A common choice is $[T^*_{(k_1)}, T^*_{(k_2)}]$ where the k_i are as above with B used instead of n. See Efron (1982, p. 78). Note that $[T^*_{(k_1)}, T^*_{(k_2)}]$ is a large sample confidence interval for θ and a large sample prediction interval for a future value of T_f^*.

The shorth(c) estimator is useful for making prediction intervals. Then the shorth estimator can be applied to a bootstrap sample $\hat{\beta}^*_{i1}, ..., \hat{\beta}^*_{iB}$ to get a

confidence interval for β_i. Here $T_n = \hat{\beta}_i$ and $\theta = \beta_i$. With the Z_i and $Z_{(i)}$ as in the above paragraph, let the shortest closed interval containing at least c of the Z_i be

$$\text{shorth(c)} = [\text{Z}_{(\text{s})}, \text{Z}_{(\text{s+c}-1)}]. \tag{2}$$

Let

$$k_n = \lceil n(1-\delta) \rceil. \tag{3}$$

Frey (2013) showed that for large $n\delta$ and iid data, the shorth(k_n) prediction interval has maximum undercoverage $\approx 1.12\sqrt{\delta/n}$, and used the shorth($c$) estimator as the large sample $100(1-\delta)\%$ PI where

$$c = \min(n, \lceil n[1 - \delta + 1.12\sqrt{\delta/n}\ ]\ \rceil). \tag{4}$$

Applied to a bootstrap sample, the Frey shorth interval can be regarded as the shortest percentile method confidence interval, asymptotically. Hence the shorth confidence interval is a practical implementation of the Hall (1988) shortest bootstrap interval based on all possible bootstrap samples. Some theory for the bootstrap shorth confidence interval is given in the last paragraph of Section 3.

Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_0$ is a known $g \times 1$ vector. Then a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is a set $\mathcal{A}_n$ such that $P(\boldsymbol{\theta} \in \mathcal{A}_n) \rightarrow 1-\delta$ as $n \rightarrow \infty$. Then reject H_0 if $\boldsymbol{\theta}_0$ is not in the confidence region $\mathcal{A}_n$. A prediction region will be applied to iid random vectors $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$. Then confidence region (11) will apply the prediction region to the bootstrap sample $T_1^*, ..., T_B^*$. Context will be used to determine whether $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$ are iid random vectors or the observed sample (the training data).

For a confidence region, let the $g \times 1$ vector T_n be an estimator of the $g \times 1$ parameter vector $\boldsymbol{\theta}$. Let $T_1^*, ..., T_B^*$ be the bootstrap sample for T_n. Let $\boldsymbol{A}$ be a full rank $g \times p$ constant matrix. For variable selection, consider testing $H_0 : \boldsymbol{A\beta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{A\beta} \neq \boldsymbol{\theta}_0$ with $\boldsymbol{\theta} = \boldsymbol{A\beta}$ where often $\boldsymbol{\theta}_0 = \boldsymbol{0}$. Then let $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$ and let $T_i^* = \boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{I_{min},0,i}$ for $i = 1, ..., B$. The statistic $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is the variable selection estimator padded with zeroes. See the second paragraph of Section 2.

To bootstrap a confidence region, Mahalanobis distances will be useful. Let the $g \times 1$ column vector $T = T_n$ be a multivariate location estimator, and let the $g \times g$ symmetric positive definite matrix $\boldsymbol{C}$ be a dispersion estimator. Then the ith *squared sample Mahalanobis distance* is the scalar

$$D_i^2 = D_i^2(T, \boldsymbol{C}) = D_{\boldsymbol{z}_i}^2(T, \boldsymbol{C}) = (\boldsymbol{z}_i - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z}_i - T) \tag{5}$$

for each observation $\boldsymbol{z}_i$. Notice that the Euclidean distance of $\boldsymbol{z}_i$ from the estimate of center T is $D_i(T, \boldsymbol{I}_g)$ where $\boldsymbol{I}_g$ is the $g \times g$ identity matrix. The classical Mahalanobis distance D_i uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{z}}, \boldsymbol{S})$, the sample mean and sample covariance matrix where

$$\overline{\boldsymbol{z}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i \ \text{ and } \ \boldsymbol{S} = \frac{1}{\text{n}-1}\sum_{\text{i}=1}^{\text{n}} (\boldsymbol{z}_\text{i} - \overline{\boldsymbol{z}})(\boldsymbol{z}_\text{i} - \overline{\boldsymbol{z}})^\text{T}. \tag{6}$$

Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + g/n)$ for $\delta > 0.1$ and

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta g/n), \quad \text{otherwise.} \tag{7}$$

If $1 - \delta < 0.999$ and $q_n < 1 - \delta + 0.001$, set $q_n = 1 - \delta$. Let

$$c = \lceil nq_n \rceil. \tag{8}$$

Let $(T, \boldsymbol{C}) = (\overline{\boldsymbol{z}}, \boldsymbol{S})$, and let $D_{(U_n)}$ be the $100q_n$th sample quantile of the D_i. Then the Olive (2013, 2017b) large sample $100(1 - \delta)\%$ nonparametric prediction region for a future value $\boldsymbol{z}_f$ given iid data $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$ is

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{z}}, \boldsymbol{S}) \leq D^2_{(U_n)}\}, \tag{9}$$

while the classical large sample $100(1 - \delta)\%$ prediction region is

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{z}}, \boldsymbol{S}) \leq \chi^2_{g,1-\delta}\}. \tag{10}$$

The Olive (2017ab, 2018) prediction region method obtains a confidence region for $\boldsymbol{\theta}$ by applying the nonparametric prediction region (9) to the bootstrap sample $T^*_1, ..., T^*_B$. Let $\overline{T}^*$ and $\boldsymbol{S}^*_T$ be the sample mean and sample covariance matrix of the bootstrap sample. Assume $n\boldsymbol{S}^*_T \xrightarrow{P} \boldsymbol{\Sigma}_A$. See Machado and Parente (2005) for regularity conditions for this assumption.

Following Bickel and Ren (2001), let the vector of parameters $\boldsymbol{\theta} = T(F)$, the statistic $T_n = T(F_n)$, and $T^* = T(F^*_n)$ where F is the cdf of iid $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, F_n is the empirical cdf, and F^*_n is the empirical cdf of $\boldsymbol{x}^*_1, ..., \boldsymbol{x}^*_n$, a sample from F_n using the nonparametric bootstrap. If $\sqrt{n}(F_n - F) \xrightarrow{D} \boldsymbol{z}_F$, a Gaussian random process, and if T is sufficiently smooth (has a Hadamard derivative $\dot{T}(F)$), then $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ and $\sqrt{n}(T^*_i - T_n) \xrightarrow{D} \boldsymbol{u}$ with $\boldsymbol{u} = \dot{T}(F)\boldsymbol{z}_F$. Olive (2017b, 2018) used these results to show that if $\boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$, then $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{D} \boldsymbol{0}$, $\sqrt{n}(T^*_i - \overline{T}^*) \xrightarrow{D} \boldsymbol{u}$, $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, and that the prediction region method large sample $100(1 - \delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - \overline{T}^*)^T [\boldsymbol{S}^*_T]^{-1} (\boldsymbol{w} - \overline{T}^*) \leq D^2_{(U_B)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{S}^*_T) \leq D^2_{(U_B)}\} \tag{11}$$

where $D^2_{(U_B)}$ is computed from $D^2_i = (T^*_i - \overline{T}^*)^T [\boldsymbol{S}^*_T]^{-1} (T^*_i - \overline{T}^*)$ for $i = 1, ..., B$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(\overline{T}^* - \boldsymbol{\theta}_0)^T [\boldsymbol{S}^*_T]^{-1} (\overline{T}^* - \boldsymbol{\theta}_0) > D^2_{(U_B)}$. A simpler proof than the Olive (2017b, 2018) proof is given in Section 3.

The modified Bickel and Ren (2001) large sample $100(1 - \delta)\%$ confidence region is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}^*_T]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B,T)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}^*_T) \leq D^2_{(U_B,T)}\} \tag{12}$$

where the cutoff $D^2_{(U_B,T)}$ is the $100q_B$th sample quantile of the $D^2_i = (T^*_i - T_n)^T [\boldsymbol{S}^*_T]^{-1} (T^*_i - T_n)$. Note that q_B is found from (7) and (8) by replacing

n by B. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(T_n - \boldsymbol{\theta}_0)^T [\boldsymbol{S}_T^*]^{-1} (T_n - \boldsymbol{\theta}_0) > D^2_{(U_B,T)}$.

Shift region (11) to have center T_n, or equivalently, change the cutoff of region (12) to $D^2_{(U_B)}$ to get the new hybrid large sample $100(1-\delta)\%$ confidence region: $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{(U_B)}\}. \tag{13}$$

Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(T_n - \boldsymbol{\theta}_0)^T [\boldsymbol{S}_T^*]^{-1} (T_n - \boldsymbol{\theta}_0) > D^2_{(U_B)}$.

Hyperellipsoids (11) and (13) have the same volume since they are the same region shifted to have a different center. The ratio of the volumes of regions (11) and (12) is

$$\frac{|\boldsymbol{S}_T^*|^{1/2}}{|\boldsymbol{S}_T^*|^{1/2}} \left(\frac{D_{(U_B)}}{D_{(U_B,T)}} \right)^g = \left(\frac{D_{(U_B)}}{D_{(U_B,T)}} \right)^g. \tag{14}$$

For $g = 1$, the percentile method uses an interval that contains $U_B \approx k_B = \lceil B(1-\delta) \rceil$ of the T_i^* from a bootstrap sample $T_1^*, ..., T_B^*$ where the statistic T_n is an estimator of θ based on a sample of size n. Note that the squared Mahalanobis distance $D_\theta^2 = (\theta - \overline{T}^*)^2 / S_T^{*2} \leq D^2_{(U_B)}$ is equivalent to $\theta \in [\overline{T}^* - S_T^* D_{(U_B)}, \overline{T}^* + S_T^* D_{(U_B)}]$, which is an interval centered at $\overline{T}^*$ just long enough to cover U_B of the T_i^*. Hence the prediction region method is a special case of the percentile method if $g = 1$. Efron (2014) used a similar large sample $100(1-\delta)\%$ confidence interval assuming that $\overline{T}^*$ is asymptotically normal. The Frey (2013) shorth(c) interval (2) (with c given by (4)) applied to the T_i^* is recommended since the shorth confidence interval can be much shorter than the Efron (2014) or prediction region method confidence intervals if $g = 1$.

The bootstrap confidence region (11) is centered at $\overline{T}^*$, which is closely related to a model averaging estimator. Wang and Zhou (2013) show that the Hjort and Claeskens (2003) confidence intervals based on frequentist model averaging are asymptotically equivalent to those obtained from the full model. See Buckland et al. (1997) and Schomaker and Heumann (2014) for standard errors when using the bootstrap or model averaging for linear model confidence intervals. Additional references are in Section 6.

Sections 2 and 3 give large sample theory for some OLS variable selection estimators and for the confidence regions. Section 4 considers using the confidence regions after variable selection, and Section 5 gives a simulation.

2 Large Sample Theory for Some OLS Variable Selection Estimators

Large sample theory is often tractable if the optimization problem is convex. The optimization problem for variable selection is not convex, so new tools are needed. Tibshirani et al. (2018) and Leeb and Pötscher (2006, 2008) note

that we can not find the limiting distribution of $\boldsymbol{Z}_n = \sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\beta}}_{I_{min}} - \boldsymbol{\beta}_I)$ after variable selection. One reason is that with positive probability, $\hat{\boldsymbol{\beta}}_{I_{min}}$ does not have the same dimension as $\boldsymbol{\beta}_I$ if AIC or C_p is used. Hence $\boldsymbol{Z}_n$ is not defined with positive probability.

We will show that large sample theory becomes simple by using zero padding. If $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$, form the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I,0}$ from $\hat{\boldsymbol{\beta}}_I$ by adding 0s corresponding to the omitted variables. For example, if $p = 4$ and $\hat{\boldsymbol{\beta}}_{I_{min}} = (\hat{\beta}_1, \hat{\beta}_3)^T$, then $\hat{\boldsymbol{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$. Since fewer than 2^p regression models I contain the true model S, and each such model gives a $\sqrt{n}$ consistent estimator $\hat{\boldsymbol{\beta}}_{I,0}$ of $\boldsymbol{\beta}$, the probability that I_{min} picks one of these models goes to one as $n \to \infty$ by Nishii (1984). Hence $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ under model (1) if AIC or C_p is used with forward selection, backward elimination, or all subsets. Olive (2017a: p. 123, 2017b: p. 176) showed that $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is a consistent estimator. This section will use mixture distributions to find the limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_{min},0} - \boldsymbol{\beta})$.

Mixture distributions are useful for variable selection since $\hat{\boldsymbol{\beta}}_{I_{min},0}$ has a mixture distribution of the $\hat{\boldsymbol{\beta}}_{I_j,0}$. A random vector $\boldsymbol{u}$ has a mixture distribution of random vectors $\boldsymbol{u}_j$ with probabilities π_j if $\boldsymbol{u}$ equals random vector $\boldsymbol{u}_j$ with probability π_j for $j = 1, ..., J$. Let $\boldsymbol{u}$ and $\boldsymbol{u}_j$ be $p \times 1$ random vectors. Then the cumulative distribution function (cdf) of $\boldsymbol{u}$ is

$$F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})$$

where the probabilities π_j satisfy $0 \leq \pi_j \leq 1$ and $\sum_{j=1}^{J} \pi_j = 1$, $J \geq 2$, and $F_{\boldsymbol{u}_j}(\boldsymbol{t})$ is the cdf of $\boldsymbol{u}_j$.

Suppose $E(h(\boldsymbol{u}))$ and the $E(h(\boldsymbol{u}_j))$ exist. Then

$$E(h(\boldsymbol{u})) = \sum_{j=1}^{J} \pi_j E[h(\boldsymbol{u}_j)] \text{ and } \mathrm{E}(\boldsymbol{u}) = \sum_{\mathrm{j}=1}^{\mathrm{J}} \pi_{\mathrm{j}} \mathrm{E}[\boldsymbol{u}_{\mathrm{j}}].$$

Hence $\mathrm{Cov}(\boldsymbol{u}) = E(\boldsymbol{u}\boldsymbol{u}^T) - E(\boldsymbol{u})E(\boldsymbol{u}^T) = E(\boldsymbol{u}\boldsymbol{u}^T) - E(\boldsymbol{u})[E(\boldsymbol{u})]^T =$
$\sum_{j=1}^{J} \pi_j E[\boldsymbol{u}_j \boldsymbol{u}_j^T] - E(\boldsymbol{u})[E(\boldsymbol{u})]^T =$

$$\sum_{j=1}^{J} \pi_j \mathrm{Cov}(\boldsymbol{u}_j) + \sum_{j=1}^{J} \pi_j E(\boldsymbol{u}_j)[E(\boldsymbol{u}_j)]^T - E(\boldsymbol{u})[E(\boldsymbol{u})]^T.$$

If $E(\boldsymbol{u}_j) = \boldsymbol{\theta}$ for $j = 1, ..., J$, then $E(\boldsymbol{u}) = \boldsymbol{\theta}$ and

$$\mathrm{Cov}(\boldsymbol{u}) = \sum_{j=1}^{J} \pi_j \mathrm{Cov}(\boldsymbol{u}_j).$$

Now suppose that T_n is equal to the estimator T_{jn} with probability π_{jn} for $j = 1, ..., J$ where $\sum_j \pi_{jn} = 1$, $\pi_{jn} \to \pi_j$ as $n \to \infty$, and $\boldsymbol{u}_{jn} = \sqrt{n}(T_{jn} - \boldsymbol{\theta}) \xrightarrow{D}$

$\boldsymbol{u}_j$ with $E(\boldsymbol{u}_j) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{u}_j) = \boldsymbol{\Sigma}_j$. Then T_n has a mixture distribution of the T_{jn} with probabilities π_{jn}, and the cdf of T_n is $F_{T_n}(\boldsymbol{z}) = \sum_j \pi_{jn} F_{T_{jn}}(\boldsymbol{z})$ where $F_{T_{jn}}(\boldsymbol{z})$ is the cdf of T_{jn}. Hence $\sqrt{n}(T_n - \boldsymbol{\theta})$ has a mixture distribution of the $\sqrt{n}(T_{jn} - \boldsymbol{\theta})$, and

$$\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u} \tag{15}$$

where the cdf of $\boldsymbol{u}$ is $F_{\boldsymbol{u}}(\boldsymbol{z}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{z})$ and $F_{\boldsymbol{u}_j}(\boldsymbol{z})$ is the cdf of $\boldsymbol{u}_j$. Thus, $\boldsymbol{u}$ is a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j, $E(\boldsymbol{u}) = \boldsymbol{0}$, and $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \sum_j \pi_j \boldsymbol{\Sigma}_j$.

Applying the above results with large sample theory for OLS makes large sample theory for OLS variable selection simple. Assume the maximum leverage $\max_{i=1,\ldots,n} \boldsymbol{x}_{iI}^T (\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1} \boldsymbol{x}_{iI} \to 0$ in probability as $n \to \infty$ for each I with $S \subseteq I$. For the full OLS model, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2 \boldsymbol{V})$ where $(\boldsymbol{X}^T \boldsymbol{X})/n \xrightarrow{P} \boldsymbol{V}^{-1}$. See, for example, Olive (2017a, p. 39) and Sen and Singer (1993, p. 280). For OLS variable selection with C_p, let $\hat{\boldsymbol{\beta}}_{I_j} = (\boldsymbol{X}_{I_j}^T \boldsymbol{X}_{I_j})^{-1} \boldsymbol{X}_{I_j}^T \boldsymbol{Y} = \boldsymbol{D}_j \boldsymbol{Y}$, $T_n = \hat{\boldsymbol{\beta}}_{I_{min},0}$ and $T_{jn} = \hat{\boldsymbol{\beta}}_{I_j,0} = \boldsymbol{D}_{j,0} \boldsymbol{Y}$ where $\boldsymbol{D}_{j,0}$ adds rows of zeroes to $\boldsymbol{D}_j$ corresponding to the x_i not in I_j. Let $T_n = T_{kn} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities π_{kn} where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the π_k with $S \subseteq I_k$ by π_j. The other $\pi_k = 0$ by Nishii (1984). Then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \sigma^2 \boldsymbol{V}_j)$ and $\boldsymbol{u}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}_j \sim N_p(\boldsymbol{0}, \sigma^2 \boldsymbol{V}_{j,0})$ where $n(\boldsymbol{X}_{I_j}^T \boldsymbol{X}_{I_j})^{-1} \xrightarrow{P} \boldsymbol{V}_j$ and $\boldsymbol{V}_{j,0}$ adds columns and rows of zeroes corresponding to the x_i not in I_j. Hence $\boldsymbol{\Sigma}_j = \sigma^2 \boldsymbol{V}_{j,0}$ is singular unless I_j corresponds to the full model.

Then Equation (15) holds:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_{min},0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u} \tag{16}$$

where the cdf of $\boldsymbol{u}$ is $F_{\boldsymbol{u}}(\boldsymbol{z}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{z})$. Thus $\boldsymbol{u}$ is a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j, $E(\boldsymbol{u}) = \boldsymbol{0}$, and $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \sum_j \pi_j \sigma^2 \boldsymbol{V}_{j,0}$. The values of π_j depend on the OLS variable selection method with C_p, such as backward elimination, forward selection, and all subsets. The results also hold if $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Hence the results hold if BIC or AIC is used instead of C_p, and, under regularity conditions, for the relaxed lasso estimator that fits OLS to the predictors than had nonzero lasso coefficients. See Efron et al. (2004), Meinshausen (2007), and Tibshirani (1996). Let $\boldsymbol{A}$ be a $g \times p$ full rank matrix with $1 \leq g \leq p$. Then

$$\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} = \boldsymbol{v} \tag{17}$$

where $\boldsymbol{A}\boldsymbol{u}$ has a mixture distribution of the $\boldsymbol{A}\boldsymbol{u}_j \sim N_g(\boldsymbol{0}, \sigma^2 \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T)$ with probabilities π_j.

Two special cases are interesting. First, suppose $\pi_d = 1$ so $\boldsymbol{u} \sim \boldsymbol{u}_d \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_d)$. This special case occurs for C_p if $a_S = p$ so S is the full model, and for all subsets variable selection with methods like BIC that choose I_S with probability going to one.

The second special case occurs if for each $\pi_j > 0$, $\boldsymbol{A}\boldsymbol{u}_j \sim N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\Sigma}_j\boldsymbol{A}^T) = N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$. Then $\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$. This special case occurs for $\hat{\boldsymbol{\beta}}_S$ if the nontrivial predictors are orthogonal or uncorrelated with zero mean so $\boldsymbol{X}^T\boldsymbol{X}/n \to diag(d_1, ..., d_p)$ as $n \to \infty$ where each $d_i > 0$. Then $\hat{\boldsymbol{\beta}}_S$ has the same multivariate normal limiting distribution for I_{min} and for the OLS full model.

3 Theory for the Confidence Regions

Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}$ is $g \times 1$. This section gives some theory for bootstrap confidence regions and for the bagging estimator $\overline{T}^*$, also called the smoothed bootstrap estimator. Empirically, bootstrapping with the bagging estimator often outperforms bootstrapping with T_n. See Breiman (1996), Yang (2003), and Efron (2014). See Büchlmann and Yu (2002) and Friedman and Hall (2007) for theory and references for the bagging estimator. Since (12) is a large sample confidence region by Bickel and Ren (2001), (11) and (13) are too, provided $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \boldsymbol{0}$.

If i) $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, then under regularity conditions, ii) $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$, iii) $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, iv) $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{u}$, and v) $n\boldsymbol{S}_T^* \xrightarrow{P} \mathrm{Cov}(\boldsymbol{u})$.

Suppose i) and ii) hold with $E(\boldsymbol{u}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}}$. With respect to the bootstrap sample, T_n is a constant and the $\sqrt{n}(T_i^* - T_n)$ are iid for $i = 1, ..., B$. Let $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{v}_i \sim \boldsymbol{u}$ where the $\boldsymbol{v}_i$ are iid with the same distribution as $\boldsymbol{u}$. Fix B. Then the average of the $\sqrt{n}(T_i^* - T_n)$ is

$$\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{D} \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{v}_i \sim AN_g\left(\boldsymbol{0}, \frac{\boldsymbol{\Sigma}_{\boldsymbol{u}}}{B}\right)$$

where $\boldsymbol{z} \sim AN_g(\boldsymbol{0}, \boldsymbol{\Sigma})$ is an asymptotic multivariate normal approximation. Hence as $B \to \infty$, $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \boldsymbol{0}$, and iii) and iv) hold. If B is fixed and $\boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{u}})$, then

$$\frac{1}{B}\sum_{i=1}^{B} \boldsymbol{v}_i \sim N_g\left(\boldsymbol{0}, \frac{\boldsymbol{\Sigma}_{\boldsymbol{u}}}{B}\right) \text{ and } \sqrt{\mathrm{B}}\sqrt{\mathrm{n}}(\overline{\mathrm{T}}^* - \mathrm{T_n}) \xrightarrow{\mathrm{D}} \mathrm{N_g}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{u}}).$$

Hence the prediction region method gives a large sample confidence region for $\boldsymbol{\theta}$ provided that the sample percentile $\hat{D}^2_{1-\delta}$ of the $D^2_{T_i^*}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(T_i^* - \overline{T}^*)^T(n\boldsymbol{S}_T^*)^{-1}\sqrt{n}(T_i^* - \overline{T}^*)$ is a consistent estimator of the percentile $D^2_{n,1-\delta}$ of the random variable $D^2_{\boldsymbol{\theta}}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(\boldsymbol{\theta} - \overline{T}^*)^T(n\boldsymbol{S}_T^*)^{-1}\sqrt{n}(\boldsymbol{\theta} - \overline{T}^*)$ in that $\hat{D}^2_{1-\delta} - D^2_{n,1-\delta} \xrightarrow{P} 0$. Since iii) and iv) hold, the sample percentile will be consistent under much weaker conditions than v) if $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ is nonsingular. Olive (2017b: § 5.3.3, 2018) proved that the prediction region method gives a large sample confidence region under the much stronger conditions of v) and $\boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{u}})$, but the above proof is simpler.

A geometric argument is useful. Assume $T_1, ..., T_B$ are iid with nonsingular covariance matrix $\boldsymbol{\Sigma}_{T_n}$. Then the large sample $100(1-\delta)\%$ prediction region $R_p = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}, \boldsymbol{S}_T) \leq D^2_{(U_B)}\}$ centered at $\overline{T}$ contains a future value of the statistic T_f with probability $1 - \delta_B \rightarrow 1 - \delta$ as $B \rightarrow \infty$. Hence the region $R_c = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T) \leq D^2_{(U_B)}\}$ centered at a randomly selected T_n contains $\overline{T}$ with probability $1 - \delta_B$. If $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ with $E(\boldsymbol{u}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}}$, then for fixed B with $\boldsymbol{v}_i \sim \boldsymbol{u}$,

$$\sqrt{n}(\overline{T} - \boldsymbol{\theta}) \xrightarrow{D} \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{v}_i \sim AN_g\left(\boldsymbol{0}, \frac{\boldsymbol{\Sigma}_{\boldsymbol{u}}}{B}\right).$$

Hence $(\overline{T} - \boldsymbol{\theta}) = O_P((nB)^{-1/2})$, and $\overline{T}$ gets arbitrarily close to $\boldsymbol{\theta}$ compared to T_n as $B \rightarrow \infty$. Hence R_c is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ as $n, B \rightarrow \infty$. We also need $(n\boldsymbol{S}_T)^{-1}$ to be fairly well behaved (not too ill conditioned) for each $n \geq 20g$, say. This condition is weaker than $(n\boldsymbol{S}_T)^{-1} \xrightarrow{P} \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}$.

If $\sqrt{n}(T_n - \boldsymbol{\theta})$ and $\sqrt{n}(T_i^* - T_n)$ both converge in distribution to $\boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$, say, then the bootstrap sample data cloud of $T_1^*, ..., T_B^*$ is like the data cloud of iid $T_1, ..., T_B$ shifted to be centered at T_n. Then the hybrid region (13) is a confidence region by the geometric argument, and (11) is a confidence region if $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \boldsymbol{0}$.

Note that if $\sqrt{n}(T_n - \theta) \xrightarrow{D} U$ and $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} U$ where U has a unimodal probability density function symmetric about zero, then the confidence intervals from the three confidence regions, the shorth confidence interval, and the usual percentile method confidence interval are asymptotically equivalent (use the central proportion of the bootstrap sample, asymptotically).

4 Bootstrapping Variable Selection Estimators

Olive (2017a: p. 128, 2017b: p. 181, 2018) showed that the prediction region method can simulate well for the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I_{min},0}$. This section will explain why the bootstrap confidence regions (11), (12), and (13) give useful results. Much of the theory in Section 3 does not apply to the variable selection estimator $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$ with $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta}$, because T_n is not smooth since T_n is equal to the estimator T_{jn} with probability π_{jn} for $j = 1, ..., J$. Here $\boldsymbol{A}$ is a known full rank $g \times p$ matrix with $1 \leq g \leq p$. We have $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{v}$ by (17) where $E(\boldsymbol{v}) = \boldsymbol{0}$, and $\boldsymbol{\Sigma}_{\boldsymbol{v}} = \sum_j \sigma^2 \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T$. Hence the geometric argument of Section 3 holds: applying the prediction region (9) to an iid sample $T_1, ..., T_B$ and then centering the region at T_n gives a large sample confidence region for $\boldsymbol{\theta}$. For variable selection, this section will show that the bootstrap sample data cloud $T_1^*, ..., T_B^*$ tends to be slightly more variable than the data cloud of iid $T_1, ..., T_B$ for large n.

Assume p is fixed, $n \geq 20p$, and that the error distribution is unimodal and not highly skewed. The response plot and residual plot are plots with

$\hat{Y} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ on the horizontal axis and Y or r on the vertical axis, respectively. Then the plotted points in these plots should scatter in roughly even bands about the identity line (with unit slope and zero intercept) and the $r = 0$ line, respectively. If the error distribution is skewed or multimodal, then much larger sample sizes may be needed.

For the bootstrap, suppose that T_i^* is equal to T_{ij}^* with probability ρ_{jn} for $j = 1, ..., J$ where $\sum_j \rho_{jn} = 1$, and $\rho_{jn} \to \pi_j$ as $n \to \infty$. Let B_{jn} count the number of times $T_i^* = T_{ij}^*$ in the bootstrap sample. Then the bootstrap sample $T_1^*, ..., T_B^*$ can be written as

$$T_{1,1}^*, ..., T_{B_{1n},1}^*, ..., T_{1,J}^*, ..., T_{B_{Jn},J}^*$$

where the B_{jn} follow a multinomial distribution and $B_{jn}/B \xrightarrow{P} \rho_{jn}$ as $B \to \infty$. Denote $T_{1j}^*, ..., T_{B_{jn},j}^*$ as the jth bootstrap component of the bootstrap sample with sample mean $\overline{T}_j^*$ and sample covariance matrix $\boldsymbol{S}_{T,j}^*$. Then

$$\overline{T}^* = \frac{1}{B}\sum_{i=1}^{B} T_i^* = \sum_j \frac{B_{jn}}{B}\frac{1}{B_{jn}}\sum_{i=1}^{B_{jn}} T_{ij}^* = \sum_j \hat{\rho}_{jn}\overline{T}_j^*.$$

Similarly, we can define the jth component of the iid sample $T_1, ..., T_B$ to have sample mean $\overline{T}_j$ and sample covariance matrix $\boldsymbol{S}_{T,j}$.

For the residual bootstrap, we use the fitted values and residuals from the OLS full model to obtain $\boldsymbol{Y}^*$, but fit $\hat{\boldsymbol{\beta}}$ for a method such as forward selection, lasso, et cetera. Consider forward selection where each component uses a $\hat{\boldsymbol{\beta}}_{I_j}$. Let $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{Y}}_{OLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{H}\boldsymbol{Y}$ be the fitted values from the OLS full model where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Let $\boldsymbol{r}^W$ denote an $n \times 1$ random vector of elements selected with replacement from the OLS full model residuals. Following Freedman (1981) and Efron (1982, p. 36), $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} + \boldsymbol{r}^W$ follows a standard linear model where the elements r_i^W of $\boldsymbol{r}^W$ are iid from the empirical distribution of the OLS full model residuals r_i. Hence

$$E(r_i^W) = \frac{1}{n}\sum_{i=1}^{n} r_i = 0, \quad V(r_i^W) = \sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} r_i^2 = \frac{n-p}{n}MSE,$$

$$E(\boldsymbol{r}^W) = \boldsymbol{0}, \text{ and } \text{Cov}(\boldsymbol{Y}^*) = \text{Cov}(\boldsymbol{r}^{\text{W}}) = \sigma_{\text{n}}^2\boldsymbol{I}_{\text{n}}.$$

Then $\hat{\boldsymbol{\beta}}_{I_j}^* = (\boldsymbol{X}_{I_j}^T\boldsymbol{X}_{I_j})^{-1}\boldsymbol{X}_{I_j}^T\boldsymbol{Y}^* = \boldsymbol{D}_j\boldsymbol{Y}^*$ with $\text{Cov}(\hat{\boldsymbol{\beta}}_{I_j}^*) = \sigma_n^2(\boldsymbol{X}_{I_j}^T\boldsymbol{X}_{I_j})^{-1}$ and $E(\hat{\boldsymbol{\beta}}_{I_j}^*) = (\boldsymbol{X}_{I_j}^T\boldsymbol{X}_{I_j})^{-1}\boldsymbol{X}_{I_j}^T E(\boldsymbol{Y}^*) = (\boldsymbol{X}_{I_j}^T\boldsymbol{X}_{I_j})^{-1}\boldsymbol{X}_{I_j}^T\boldsymbol{H}\boldsymbol{Y} = \hat{\boldsymbol{\beta}}_{I_j}$ since $\boldsymbol{H}\boldsymbol{X}_{I_j} = \boldsymbol{X}_{I_j}$. The expectations are with respect to the bootstrap distribution where $\hat{\boldsymbol{Y}}$ acts as a constant.

For the above residual bootstrap with C_p, let $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$ and $T_{jn} = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0} = \boldsymbol{A}\boldsymbol{D}_{j,0}\boldsymbol{Y}$ where $\boldsymbol{D}_{j,0}$ adds rows of zeroes to $\boldsymbol{D}_j$ corresponding to the x_i not in I_j. If $S \subseteq I_j$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \sigma^2\boldsymbol{V}_j)$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}_j \sim N_p(\boldsymbol{0}, \sigma^2\boldsymbol{V}_{j,0})$ where $\boldsymbol{V}_{j,0}$ adds columns and rows

of zeroes corresponding to the x_i not in I_j. Using results from Section 2, $E(T^*) = \sum_j \rho_{jn} T_{jn} = \sum_j \rho_{jn} \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0}$ and $\boldsymbol{S}_T^*$ is a consistent estimator of

$$\mathrm{Cov}(T^*) = \sum_j \rho_{jn} \mathrm{Cov}(T_{jn}^*) + \sum_j \rho_{jn} \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0} \hat{\boldsymbol{\beta}}_{I_j,0}^T \boldsymbol{A}^T - E(T^*)[E(T^*)]^T$$

where asymptotically the sum is over $j : S \subseteq I_j$. If $\boldsymbol{\theta}_0 = \boldsymbol{0}$, then $n\boldsymbol{S}_T^* = \boldsymbol{\Sigma}_A + O_P(1)$ where

$$n\mathrm{Cov}(T_n) \xrightarrow{P} \boldsymbol{\Sigma}_A = \sum_j \sigma^2 \pi_j \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T.$$

Then $(n\boldsymbol{S}_T^*)^{-1}$ tends to be "well behaved" if $\boldsymbol{\Sigma}_A$ is nonsingular.

For the residual bootstrap with forward selection $n\mathrm{Cov}(T_{jn})$ and $n\mathrm{Cov}(T_{jn}^*)$ both converge in probability to $\sigma^2 \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T$, and are close for $n \geq 20p$ since $\mathrm{Cov}(T_{jn}^*) \approx (n-p)\mathrm{Cov}(T_{jn})/n$. Hence the jth component of an iid sample $T_1, ..., T_B$ and the jth component of the bootstrap sample $T_1^*, ..., T_B^*$ have the same variability asymptotically. Since $E(T_{jn}) = \boldsymbol{\theta}$, each component of the iid sample is centered at $\boldsymbol{\theta}$. Since $E(T_{jn}^*) = T_{jn} = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0}$, the bootstrap components are centered at T_{jn}. Geometrically, separating the component clouds so that they are no longer centered at one value makes the overall data cloud larger. Thus the variability of T_n^* is larger than that of T_n for variable selection, asymptotically. Hence the prediction region applied to the bootstrap sample is slightly larger than the prediction region applied to the iid sample, asymptotically (we want $n \geq 20p$). Hence cutoff $\hat{D}_{1,1-\delta}^2 = D_{(U_B)}^2$ gives coverage close to or higher than the nominal coverage for confidence regions (11) and (13), using the geometric argument. The deviation $T_i^* - T_n$ tends to be larger in magnitude than the deviations $\overline{T}^* - \boldsymbol{\theta}$, $T_n - \boldsymbol{\theta}$, and $T_i^* - \overline{T}^*$. Hence the cutoff $\hat{D}_{2,1-\delta}^2 = D_{(U_B,T)}^2$ tends to be larger than $D_{(U_B)}^2$, and region (12) tends to have higher coverage than region (13) for a mixture distribution. The bootstrap sample data cloud is centered at $\overline{T}^* \approx \sum_j \rho_{jn} T_{jn}$. The T_{jn} are computed from the same data set and hence correlated. In simulations for $n \geq 20p$ and (11) and (13), the coverage tends to get close to $1-\delta$ for $B \geq \max(400, 50p)$ so that $\boldsymbol{S}_T^*$ is a good estimator of $\mathrm{Cov}(T^*)$.

In the simulations where S is not the full model, inference with forward selection with I_{min} using C_p was often more precise than inference with the OLS full model if $n \geq 20p$ and $B \geq 50p$. It is possible that $\boldsymbol{S}_T^*$ is singular if a column of the bootstrap sample is equal to $\boldsymbol{0}$.

Undercoverage can occur if the bootstrap sample data cloud is less variable than the iid data cloud, e.g., if $(n-p)/n$ is not close to one. Coverage can be higher than the nominal coverage for two reasons: i) the bootstrap data cloud is more variable than the iid data cloud of $T_1, ..., T_B$, and ii) zero padding.

To see the effect of zero padding, consider $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{\beta}_O = \boldsymbol{0}$ where $\boldsymbol{\beta}_O = (\beta_{i_1}, ..., \beta_{i_g})^T$ and $O \subseteq E$ in (1) so that H_0 is true. Suppose a nominal 95% confidence region is used and $U_B = 0.96$. Hence the confidence region (11) or (12) covers at least 96% of the bootstrap sample. If $\hat{\boldsymbol{\beta}}_{O,j}^* = \boldsymbol{0}$ for more than

4% of the $\hat{\boldsymbol{\beta}}^*_{O,1}, ..., \hat{\boldsymbol{\beta}}^*_{O,B}$, then $\mathbf{0}$ is in the confidence region and the bootstrap test fails to reject H_0. If this occurs for each run in the simulation, then the observed coverage will be 100%.

Now suppose $\hat{\boldsymbol{\beta}}^*_{O,j} = \mathbf{0}$ for $j = 1, ..., B$. Then $\boldsymbol{S}^*_T$ is singular, but the singleton set $\{\mathbf{0}\}$ is the large sample $100(1-\delta)\%$ confidence region (11), (12), or (13) for $\boldsymbol{\beta}_O$ and $\delta \in (0, 1)$, and the pvalue for $H_0 : \boldsymbol{\beta}_O = \mathbf{0}$ is one. (This result holds since $\{\mathbf{0}\}$ contains 100% of the $\hat{\boldsymbol{\beta}}^*_{O,j}$ in the bootstrap sample.) For large sample theory tests, the pvalue estimates the population pvalue. Let I denote the other predictors in the model so $\boldsymbol{\beta} = (\boldsymbol{\beta}_I^T, \boldsymbol{\beta}_O^T)^T$. For the I_{min} model from forward selection, there may be strong evidence that $\boldsymbol{x}_O$ is not needed in the model given $\boldsymbol{x}_I$ is in the model if the "100%" confidence region is $\{\mathbf{0}\}$, $n \geq 20p$, $B \geq 50p$, and the error distribution is unimodal and not highly skewed. (Since the pvalue is one, this technique may be useful for data snooping: applying OLS theory to submodel I may have negligible selection bias.)

Note that there are several important variable selection models, including the model given by Equation (1). Another model is $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_{S_i}^T\boldsymbol{\beta}_{S_i}$ for $i = 1, ..., K$. Then there are $K \geq 2$ competing "true" nonnested submodels where $\boldsymbol{\beta}_{S_i}$ is $a_{S_i} \times 1$. See Ferrari and Yang (2015). For example, suppose the $K = 2$ models have predictors x_1, x_2, x_3 for S_1 and x_1, x_2, x_4 for S_2. Then x_3 and x_4 are likely to be selected and omitted often by forward selection for the B bootstrap samples. Hence omitting all predictors x_i that have a $\beta^*_{ij} = 0$ for at least one of the bootstrap samples $j = 1, ..., B$ could result in underfitting, e.g. using just x_1 and x_2 in the above $K = 2$ example. If n and B are large enough, the singleton set $\{\mathbf{0}\}$ could still be the "100%" confidence region for a vector $\boldsymbol{\beta}_O$.

Suppose the predictors x_i have been standardized. Then another important regression model has the β_i taper off rapidly, but no coefficients are equal to zero. For example, $\beta_i = e^{-i}$ for $i = 1, ..., p$.

5 Example and Simulations

Example. Cook and Weisberg (1999, pp. 351, 433, 447) gives a data set on 82 mussels sampled off the coast of New Zealand. The data set can be found at the URL listed for Olive (2019, Example 2.8) which has *R* code to reproduce the example. Let the response variable be the logarithm $\log(M)$ of the *muscle mass*, and the predictors are the *length* L and *height* H of the shell in mm, the logarithm $\log(W)$ of the *shell width* W, the logarithm $\log(S)$ of the *shell mass* S and a constant. Inference for the full model is shown along with the shorth(c) nominal 95% confidence intervals for β_i computed using the nonparametric and residual bootstraps. As expected, the residual bootstrap intervals are close to the classical least squares confidence intervals $\approx \hat{\beta}_i \pm 2SE(\hat{\beta}_i)$.

The minimum C_p model from forward selection used a constant, H, and $\log(S)$. The shorth(c) nominal 95% confidence intervals for β_i using the residual bootstrap are shown. Note that the intervals for W and H are right skewed

and contain 0 when closed intervals are used instead of open intervals. The least squares output is also shown, but should only be used for inference if the model was selected before looking at the data.

It was expected that $\log(S)$ may be the only predictor needed, along with a constant, since $\log(S)$ and $\log(M)$ are both log(mass) measurements and likely highly correlated. Hence we want to test $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ with the I_{min} model selected by forward selection. (Of course this test would be easy to do with the full model using least squares theory.) Then $H_0 : \boldsymbol{A\beta} = (\beta_2, \beta_3, \beta_4)^T = \boldsymbol{0}$. Using the prediction region method with the full model had $[0, D_{(U_B)}] = [0, 2.908]$ with $D_{\boldsymbol{0}} = 1.577$. Note that $\sqrt{\chi^2_{3,0.95}} = 2.795$. So fail to reject H_0. Using the prediction region method with the I_{min} forward selection model had $[0, D_{(U_B)}] = [0, 3.258]$ while $D_{\boldsymbol{0}} = 1.245$. So fail to reject H_0. The ratio of the volumes of the bootstrap confidence regions for this test was 0.392. (Use (14) with $\boldsymbol{S}^*_T$ and D from forward selection for the numerator, and from the full model for the denominator.) Hence the forward selection bootstrap test was more precise than the full model bootstrap test.

```
      large sample full model inference
      Est.    SE  t   Pr(>|t|)   nparboot      resboot
int -1.249 0.838 -1.49 0.14 [-2.93,-0.093][-3.045,0.473]
L   -0.001 0.002 -0.28 0.78 [-0.005,0.003][-0.005,0.004]
logW 0.130 0.374  0.35 0.73 [-0.457,0.829][-0.703,0.890]
H    0.008 0.005  1.50 0.14 [-0.002,0.018][-0.003,0.016]
logS 0.640 0.169  3.80 0.00 [ 0.244,1.040][ 0.336,1.012]
output and shorth intervals for the min Cp submodel
      Est.     SE      t      Pr(>|t|)  95% shorth CI
int   -0.9573  0.1519 -6.3018 0.0000  [-3.294, 0.495]
L      0                              [-0.005, 0.004]
logW   0                              [ 0.000, 1.024]
H      0.0072  0.0047  1.5490 0.1254  [ 0.000, 0.016]
logS   0.6530  0.1160  5.6297 0.0000  [ 0.322, 0.901]
```

Next, we describe a small simulation study that was done using $B = \max(1000, n/25, 50p)$ and 5000 runs. The simulation used $p = 4, 6, 7, 8$, and 10; $n = 25p$ and $50p$; $\psi = 0, 1/\sqrt{p}$, and 0.9; and $k = 1$ and $p-2$ where k and ψ are defined in the following paragraph. Larger simulation studies are in Imhoff (2018), Murphy (2018), and Pelawa Watagoda (2017). In the simulations, we use $\theta = \boldsymbol{A\beta} = \beta_i$, $\boldsymbol{\theta} = \boldsymbol{A\beta} = \boldsymbol{\beta}_S = \boldsymbol{1}$ and $\boldsymbol{\theta} = \boldsymbol{A\beta} = \boldsymbol{\beta}_E = \boldsymbol{0}$.

Let $\boldsymbol{x} = (1 \ \boldsymbol{u}^T)^T$ where $\boldsymbol{u}$ is the $(p-1) \times 1$ vector of nontrivial predictors. In the simulations, for $i = 1, ..., n$, we generated $\boldsymbol{w}_i \sim N_{p-1}(\boldsymbol{0}, \boldsymbol{I})$ where the $m = p - 1$ elements of the vector $\boldsymbol{w}_i$ are iid N(0,1). Let the $m \times m$ matrix $\boldsymbol{A} = (a_{ij})$ with $a_{ii} = 1$ and $a_{ij} = \psi$ where $0 \leq \psi < 1$ for $i \neq j$. Then the vector $\boldsymbol{u}_i = \boldsymbol{A}\boldsymbol{w}_i$ so that $Cov(\boldsymbol{u}_i) = \boldsymbol{\Sigma_u} = \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ where the diagonal entries $\sigma_{ii} = [1+(m-1)\psi^2]$ and the off diagonal entries $\sigma_{ij} = [2\psi+(m-2)\psi^2]$. Hence the correlations are $cor(x_i, x_j) = \rho = (2\psi+(m-2)\psi^2)/(1+(m-1)\psi^2)$ for $i \neq j$ where x_i and x_j are nontrivial predictors. If $\psi = 1/\sqrt{cp}$, then $\rho \to 1/(c+1)$ as $p \to \infty$ where $c > 0$. As ψ gets close to 1, the predictor vectors cluster about

the line in the direction of $(1, ..., 1)^T$. Let $Y_i = 1 + 1x_{i,2} + \cdots + 1x_{i,k+1} + e_i$ for $i = 1, ..., n$. Hence $\boldsymbol{\beta} = (1, .., 1, 0, ..., 0)^T$ with $k + 1$ ones and $p - k - 1$ zeros. The zero mean errors e_i were iid from five distributions: i) N(0,1), ii) t_3, iii) EXP(1) - 1, iv) uniform$(-1, 1)$, and v) 0.9 N(0,1) + 0.1 N(0,100). Only distribution iii) is not symmetric.

When $\psi = 0$, the full model least squares confidence intervals for β_i should have length near $2t_{96,0.975}\sigma/\sqrt{n} \approx 2(1.96)\sigma/10 = 0.392\sigma$ when $n = 100$ and the iid zero mean errors have variance σ^2. The simulation computed the Frey shorth(c) interval for each β_i and used bootstrap confidence regions to test $H_0 : \boldsymbol{\beta}_S = \mathbf{1}$ (whether first $k + 1$ $\beta_i = 1$) and $H_0 : \boldsymbol{\beta}_E = \mathbf{0}$ (whether the last $p - k - 1$ $\beta_i = 0$). The nominal coverage was 0.95 with $\delta = 0.05$. Observed coverage between 0.94 and 0.96 suggests coverage is close to the nominal value.

The regression models used the residual bootstrap on the forward selection estimator $\hat{\boldsymbol{\beta}}_{I_{min},0}$. Table 1 gives results for when the iid errors $e_i \sim N(0, 1)$ with $n = 100$, $p = 4$, and $k = 1$. Table 1 shows two rows for each model giving the observed confidence interval coverages and average lengths of the confidence intervals. The term "reg" is for the full model regression, and the term "vs" is for forward selection. The last six columns give results for the tests. The terms pr, hyb, and br are for the prediction region method (11), hybrid region (13), and Bickel and Ren region (12). The 0 indicates the test was $H_0 : \boldsymbol{\beta}_E = \mathbf{0}$, while the 1 indicates that the test was $H_0 : \boldsymbol{\beta}_S = \mathbf{1}$. The length and coverage = P(fail to reject H_0) for the interval $[0, D_{(U_B)}]$ or $[0, D_{(U_B,T)}]$ where $D_{(U_B)}$ or $D_{(U_B,T)}$ is the cutoff for the confidence region. The cutoff will often be near $\sqrt{\chi^2_{g,0.95}}$ if the statistic T is asymptotically normal. Note that $\sqrt{\chi^2_{2,0.95}} = 2.448$ is close to 2.45 for the full model regression bootstrap tests.

Volume ratios of the three confidence regions can be compared using (14), but there is not enough information in Table 1 to compare the volume of the confidence region for the full model regression versus that for the forward selection regression since the two methods have different determinants $|\boldsymbol{S}^*_T|$.

Table 1 Bootstrapping OLS Forward Selection with C_p, $e_i \sim N(0, 1)$

ψ	β_1	β_2	β_{p-1}	β_p	pr0	hyb0	br0	pr1	hyb1	br1
reg,0	0.946	0.950	0.947	0.948	0.940	0.941	0.941	0.937	0.936	0.937
len	0.396	0.399	0.399	0.398	2.451	2.451	2.452	2.450	2.450	2.451
vs,0	0.948	0.950	0.997	0.996	0.991	0.979	0.991	0.938	0.939	0.940
len	0.395	0.398	0.323	0.323	2.699	2.699	3.002	2.450	2.450	2.457
reg,0.5	0.946	0.944	0.946	0.945	0.938	0.938	0.938	0.934	0.936	0.936
len	0.396	0.661	0.661	0.661	2.451	2.451	2.452	2.451	2.451	2.452
vs,0.5	0.947	0.968	0.997	0.998	0.993	0.984	0.993	0.955	0.955	0.963
len	0.395	0.658	0.537	0.539	2.703	2.703	2.994	2.461	2.461	2.577
reg,0.9	0.946	0.941	0.944	0.950	0.940	0.940	0.940	0.935	0.935	0.935
len	0.396	3.257	3.253	3.259	2.451	2.451	2.452	2.451	2.451	2.452
vs,0.9	0.947	0.968	0.994	0.996	0.992	0.981	0.992	0.962	0.959	0.970
len	0.395	2.751	2.725	2.735	2.716	2.716	2.971	2.497	2.497	2.599

The inference for forward selection was often as precise or more precise than the inference for the full model. The coverages were near 0.95 for the regression bootstrap on the full model, although there was slight undercoverage for the tests since $(n-p)/n = 0.96$ when $n = 25p$. Suppose $\psi = 0$. Then from Section 2, $\hat{\boldsymbol{\beta}}_S$ has the same limiting distribution for I_{min} and the full model. Note that the average lengths and coverages were similar for the full model and forward selection I_{min} for β_1, β_2, and $\boldsymbol{\beta}_S = (\beta_1, \beta_2)^T$. Forward selection inference was more precise for $\boldsymbol{\beta}_E = (\beta_3, \beta_4)^T$. The Bickel and Ren (12) cutoffs and coverages were at least as high as those of the hybrid region (13).

For $\psi > 0$ and I_{min}, the coverages for the β_i corresponding to $\boldsymbol{\beta}_S$ were near 0.95, but the average length could be shorter since I_{min} tends to have less multicorrelation than the full model. For $\psi \geq 0$, the I_{min} coverages were higher than 0.95 for β_3 and β_4 and for testing $H_0 : \boldsymbol{\beta}_E = \mathbf{0}$ since zeros often occurred for $\hat{\beta}_j^*$ for $j = 3, 4$. The average CI lengths were shorter for I_{min} than for the OLS full model for β_3 and β_4. Note that for I_{min}, the coverage for testing $H_0 : \boldsymbol{\beta}_S = \mathbf{1}$ was higher than that for the OLS full model.

Table 2 Bootstrap CIs with C_p, $p = 10, k = 8, \psi = 0.9$, error type v)

n	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
250	0.945	0.824	0.822	0.827	0.827	0.824	0.826	0.817	0.827	0.999
shlen	0.825	6.490	6.490	6.482	6.485	6.479	6.512	6.496	6.493	6.445
250	0.946	0.979	0.980	0.985	0.981	0.983	0.983	0.977	0.983	0.998
prlen	0.807	7.836	7.850	7.842	7.830	7.830	7.851	7.840	7.839	7.802
250	0.947	0.976	0.978	0.984	0.978	0.978	0.979	0.973	0.980	0.996
brlen	0.811	8.723	8.760	8.765	8.736	8.764	8.745	8.747	8.753	8.756
2500	0.951	0.947	0.948	0.948	0.948	0.947	0.949	0.944	0.951	0.999
shlen	0.263	2.268	2.271	2.271	2.273	2.262	2.632	2.277	2.272	2.047
2500	0.945	0.961	0.959	0.955	0.960	0.960	0.961	0.958	0.961	0.998
prlen	0.258	2.630	2.639	2.640	2.632	2.632	2.641	2.638	2.642	2.517
2500	0.946	0.958	0.954	0.960	0.956	0.960	0.962	0.955	0.961	0.997
brlen	0.258	2.865	2.875	2.882	2.866	2.871	2.887	2.868	2.875	2.830
25000	0.952	0.940	0.939	0.935	0.940	0.942	0.938	0.937	0.942	1.000
shlen	0.083	0.809	0.808	0.806	0.805	0.807	0.808	0.808	0.809	0.224
25000	0.948	0.964	0.968	0.962	0.964	0.966	0.964	0.964	0.967	0.991
prlen	0.082	0.806	0.805	0.801	0.800	0.805	0.805	0.803	0.806	0.340
25000	0.949	0.969	0.972	0.968	0.967	0.971	0.969	0.969	0.973	0.999
brlen	0.082	0.810	0.810	0.805	0.804	0.809	0.810	0.808	0.810	0.317

Results for other values of n, p, k, and distributions of e_i were similar. For forward selection with $\psi = 0.9$ and C_p, the hybrid region (13) and shorth confidence intervals occasionally had coverage less than 0.93. It was also rare for the bootstrap to have one or more columns of zeroes so $\boldsymbol{S}_T^*$ was singular. For error distributions i)-iv) and $\psi = 0.9$, sometimes the shorth CIs needed $n \geq 100p$ for all p CIs to have good coverage. For error distribution v) and $\psi = 0.9$, even larger values of n were needed. Confidence intervals based on (11) and (12) worked for much smaller n, but tended to be longer than the shorth CIs.

See Table 2 for one of the worst scenarios for the shorth, where shlen, prlen, and brlen are for the average CI lengths based on the shorth, (11), and (12), respectively. In Table 2, $k = 8$ and the two nonzero π_j correspond to

the full model $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{S,0}$. Hence $\beta_i = 1$ for $i = 1, ..., 9$ and $\beta_{10} = 0$. Hence confidence intervals for β_{10} had the highest coverage and usually the shortest average length (for $i \neq 1$) due to zero padding. Theory in this paper showed that the CI lengths are proportional to $1/\sqrt{n}$. When $n = 25000$, the shorth CI uses the 95.16th percentile while CI (11) uses the 95.00th percentile, allowing the average CI length of (11) to be shorter than that of the shorth CI, but the distribution for $\hat{\beta}_i^*$ is likely approximately symmetric for $i \neq 10$ since the average lengths of the three confidence intervals were about the same for each $i \neq 10$.

When BIC was used, undercoverage was a bit more common and severe, and undercoverage occasionally occurred with regions (11) and (12). BIC also occasionally had 100% coverage since BIC produces more zeroes than C_p.

Limited simulations for the Tibshirani (1996) lasso estimator were similar, but the confidence intervals were longer than those for forward selection. Ridge regression only simulated well for $\psi = 0$.

6 Conclusion

Another way to look at the bootstrap confidence region for OLS variable selection estimators is to consider the estimator $T_{2,n}$ that chooses I_j with probability equal to the observed bootstrap proportion $\hat{\rho}_{jn}$. The bootstrap sample $T_1^*, ..., T_B^*$ tends to be slightly more variable than an iid sample $T_{2,1}, ..., T_{2,B}$, and the geometric argument suggests that the large sample coverage of the nominal $100(1-\delta)\%$ confidence region will be at least as large as the nominal coverage $100(1-\delta)\%$.

The hybrid confidence region was motivated by the geometric argument. The modified Bickel and Ren confidence region can be motivated by the hybrid region with a larger cutoff. The prediction region method works since the bagging estimator $\overline{T}^*$ tends estimate $\boldsymbol{\theta}$ at least as well as T_n. See Breiman (1996) and Yang (2003). Note that $\overline{T}^* \approx E(T^*) = \sum_j \rho_{jn} T_{jn}$, a version of model averaging.

In simulations for forward selection with C_p, bootstrap confidence regions (11) and (12) performed well. BIC seems to need larger sample size n than C_p to perform well. For some data sets, $\boldsymbol{S}_T^*$ may be singular due to one or more columns of zeroes in the bootstrap sample for $\beta_1, ..., \beta_p$. The variables corresponding to these columns are likely not needed in the model given that the other predictors are in the model. Confidence intervals can be computed without $\boldsymbol{S}_T^*$ for (11), (12), and (13).

Under regularity conditions, applying the prediction region (9) to a bootstrap sample results in a confidence region. Knight and Fu (2000) have some results on the residual bootstrap that use residuals from one estimator, such as full model OLS, but fit another estimator, such as lasso. Schomaker (2012) suggests bootstrap estimates of the standard error of $\hat{\beta}_i$ for shrinkage estimators. Firinguetti and Bobadilla (2011) suggest confidence intervals for β_i for ridge regression.

There is a massive literature on variable selection and a fairly large literature for inference after variable selection. See, for example, Leeb and Pötscher (2006, 2008), Leeb et al. (2015), Tibshirani et al. (2016), and Tibshirani et al. (2018). Recent papers on large sample theory for multiple linear regression estimators include Cook and Forzani (2018, 2019), Knight and Fu (2000), and Zhang (2018).

Results in Claeskens and Hjort (2008, pp. 101, 102, 232) suggest that the probability that AIC underfits goes to zero for many models. Hence with AIC variable selection,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_{min},0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}$$

for many time series models, generalized linear models, and survival regression models. Confidence regions (11) and (12) may be useful for the estimator $\hat{\boldsymbol{\beta}}_{I_{min},0}$ selected by AIC.

In addition to large sample theory, shrinkage and variable selection estimators can be compared with asymptotically optimal prediction intervals, even if n/p is not large. See Pelawa Watagoda and Olive (2019).

Response plots of the fitted values $\hat{Y}$ versus the response Y are useful for checking linearity of the model and for detecting outliers. Residual plots should also be made.

The simulations were done in *R*. See R Core Team (2016). We used several *R* functions including forward selection as computed with the `regsubsets` function from the `leaps` library. The collection of Olive (2019) *R* functions *slpack*, available from (http://lagrange.math.siu.edu/Olive/slpack.txt), has some useful functions for the inference. Table 1 was made using `regbootsim3` for the OLS full model and `vsbootsim4` for forward selection. The functions `bicboot` and `bicbootsim` are useful if BIC is used instead of C_p. The function `lassobootsim3` uses the prediction region method for lasso and ridge regression. The function `lassobootsim4` can be useful if $\boldsymbol{S}_T^*$ is singular for lasso. For forward selection with C_p, the function `vscisim` was used to make Table 2, and can be used to compare the shorth, prediction region method, and Bickel and Ren CIs for β_i. The function `fselboot` can be used to bootstrap the forward selection model.

Acknowledgements The authors thank the Editor and two referees for their work.

References

1. Akaike H (1973) Information theory as an extension of the maximum likelihood principle. In Proceedings, 2nd international symposium on information theory, eds. Petrov BN, Csakim F, Akademiai Kiado, Budapest, 267-281.
2. Bickel PJ, Ren JJ (2001) The bootstrap in hypothesis testing. In State of the art in probability and statistics: festschrift for William R. van Zwet, eds. de Gunst M, Klaassen C, van der Vaart A, The Institute of Mathematical Statistics, Hayward, CA, 91-112.
3. Breiman L (1996) Bagging predictors. Machine Learning 24:123-140.
4. Büchlmann P, Yu B (2002) Analyzing bagging. Ann Stat 30:927-961.

5. Buckland ST, Burnham KP, Augustin NH (1997) Model selection: an integral part of inference. Biometrics 53:603-618.
6. Claeskens G, Hjort NL (2008) Model Selection and Model Averaging, Cambridge University Press, New York, NY.
7. Cook RD, Forzani L (2018) Big data and partial least squares prediction. Can J Stat 46:62-78.
8. Cook RD, Forzani L (2019) Partial least squares prediction in high-dimensional regression. Ann Stat 47:884-908.
9. Cook RD, Weisberg S (1999), Applied Regression Including Computing and Graphics, Wiley, New York, NY.
10. Efron B (1982) The Jackknife, the Bootstrap and Other Resampling Plans, SIAM, Philadelphia, PA.
11. Efron B (2014) Estimation and accuracy after model selection (with discussion). J Am Stat Assoc 109:991-1007.
12. Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression (with discussion). Ann Stat 32:407-451.
13. Ferrari D, Yang Y (2015) Confidence sets for model selection by F–testing. Stat Sinica 25:1637-1658.
14. Firinguetti L, Bobadilla G (2011) Asymptotic confidence intervals in ridge regression based on the Edgeworth expansion. Stat Pap 52:287-307.
15. Freedman DA (1981) Bootstrapping regression models. Ann Stat 9:1218-1228.
16. Frey J (2013) Data-driven nonparametric prediction intervals. J Stat Plan Inference 143:1039-1048.
17. Friedman JH, Hall P (2007) On bagging and nonlinear estimation. J Stat Plan Inference 137:669-683.
18. Hall P (1988) Theoretical comparisons of bootstrap confidence intervals (with discussion). Ann Stat 16:927-985.
19. Hjort G, Claeskens NL (2003) The focused information criterion. J Am Stat Assoc 98:900-945.
20. Imhoff DC (2018) Bootstrapping forward selection with C_p. Master's Research Paper, Southern Illinois University.
21. Jones HL (1946) Linear regression functions with neglected variables. J Am Stat Assoc 41:356-369.
22. Knight K, Fu WJ (2000) Asymptotics for lasso-type estimators. Ann Stat 28:1356–1378.
23. Leeb H, Pötscher BM (2006) Can one estimate the conditional distribution of post-model-selection estimators? Ann Stat 34:2554-2591.
24. Leeb H, Pötscher BM (2008) Can one estimate the unconditional distribution of post-model-selection estimators? Econometrics Theory 24:338-376.
25. Leeb H, Pötscher BM, Ewald K (2015) On various confidence intervals post-model-selection. Stat Sci 30:216-227.
26. Li K–C (1987) Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. Ann Stat 15:958-975.
27. Machado JAF, Parente P (2005) Bootstrap estimation of covariance matrices via the percentile method. Econometrics J 8:70-78.
28. Mallows C (1973) Some comments on C_p. Technom 15:661-676.
29. Meinshausen N (2007) Relaxed lasso. Comput Stat Data Anal 52:374-393.
30. Murphy C (2018). Bootstrapping forward selection with BIC. Master's Research Paper, Southern Illinois University.
31. Nishii R (1984) Asymptotic properties of criteria for selection of variables in multiple regression. Ann Stat 12:758-765.
32. Olive DJ (2013) Asymptotically optimal regression prediction intervals and prediction regions for multivariate data. Internat J Stat Probab 2:90-100.
33. Olive DJ (2017a) Linear Regression, Springer, New York, NY.
34. Olive DJ (2017b) Robust Multivariate Analysis, Springer, New York, NY.
35. Olive DJ (2018) Applications of hyperellipsoidal prediction regions. Stat Pap 59:913-931.
36. Olive DJ (2019) Prediction and statistical learning, online course notes, see (http://lagrange.math.siu.edu/Olive/slearnbk.htm).
37. Olive DJ, Hawkins DM (2005) Variable selection for 1D regression models. Technom 47:43-50.

38. Pelawa Watagoda LCR (2017) Inference after variable selection, PhD Thesis, Southern Illinois University. See (http://lagrange.math.siu.edu/Olive/slasanthiphd.pdf).
39. Pelawa Watagoda LCR, Olive DJ (2019) Comparing shrinkage estimators with asymptotically optimal prediction intervals. Unpublished manuscript at (http://lagrange.math.siu.edu/Olive/pppicomp.pdf).
40. R Core Team (2016) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
41. Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461-464.
42. Shao J (1993) Linear model selection by cross-validation. J Am Stat Assoc 88:486-494.
43. Schomaker M (2012) Shrinkage averaging estimation. Stat Pap 53:1015-1034.
44. Schomaker M, Heumann C (2014) Model selection and model averaging after multiple imputation. Computat Stat Data Anal 71:758-770.
45. Seber GAF, Lee AJ (2003) Linear Regression Analysis, 2nd ed., Wiley, New York, NY.
46. Sen PK, Singer JM (1993) Large Sample Methods in Statistics: an Introduction with Applications, Chapman & Hall, New York, NY.
47. Su Z, Cook RD (2012) Inner envelopes: efficient estimation in multivariate linear regression. Biometrika 99:687-702.
48. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Royal Stat Soc B 58:267-288.
49. Tibshirani RJ, Rinaldo A, Tibshirani R, Wasserman L (2018) Uniform asymptotic inference and the bootstrap after model selection. Ann Stat 46:1255-1287.
50. Tibshirani RJ, Taylor J, Lockhart R, Tibshirani R (2016) Exact post-selection inference for sequential regression procedures. J Am Stat Assoc 111:600-620.
51. Wang H, Zhou SZF (2013) Interval estimation by frequentist model averaging. Commun Stat Theory Meth 42:4342-4356.
52. Yang Y (2003) Regression with multiple candidate models: selecting or mixing? Stat Sinica 13:783-809.
53. Zhang J (2018) Consistency of MLE, LSE and M-estimation under mild conditions. Stat Pap to appear.