Bootstrapping Some GLM and Survival Regression Variable Selection Estimators

Rasanji C. Rathnayake and David J. Olive
School of Mathematical & Statistical Sciences
Southern Illinois University
Carbondale, Illinois 62901-4408
rasanji@siu.edu dolive@siu.edu

Keywords Backward elimination; bagging; confidence region; forward selection.

Mathematics Subject Classification Primary 62F40; Secondary 62J12.

Abstract

Inference after variable selection is a very important problem. This paper derives the asymptotic distribution of many variable selection estimators, such as forward selection and backward elimination, when the number of predictors is fixed. Under strong regularity conditions, the variable selection estimators are asymptotically normal, but generally the asymptotic distribution is a nonnormal mixture distribution. The theory shows that the lasso variable selection and elastic net variable selection estimators are $\sqrt{n}$ consistent estimators of $\boldsymbol{\beta}$ when lasso and elastic net are consistent estimators of $\boldsymbol{\beta}$. A bootstrap technique to eliminate selection bias is to fit the variable selection estimator $\hat{\boldsymbol{\beta}}^*_{VS}$ to a bootstrap sample to find a submodel, then draw another bootstrap sample and fit the same submodel to get the bootstrap estimator $\hat{\boldsymbol{\beta}}^*_{MIX}$. Bootstrap confidence regions were used for hypothesis testing.

1. Introduction

This section reviews regression models, variable selection, and some results on bootstrap confidence regions. Consider regression models where the response variable Y is independent of the $p \times 1$ vector of predictors $\boldsymbol{x}$ given $\boldsymbol{x}^T\boldsymbol{\beta}$, written $Y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta}$. Many important regression models satisfy this condition, including multiple linear regression, the Nelder and

Wedderburn (1972) generalized linear models (GLMs), and the Cox (1972) proportional hazards regression model. Forward selection or backward elimination with the Akaike (1973) AIC criterion or Schwarz (1978) BIC criterion are often used for variable selection.

Some shrinkage methods do variable selection: the regression method, such as a GLM, uses the predictors that had nonzero shrinkage estimator coefficients. These methods include least angle regression, lasso, relaxed lasso, and elastic net. Least angle regression variable selection is the LARS-OLS hybrid estimator of Efron et al. (2004, p. 421). Lasso variable selection is called relaxed lasso by Hastie, Tibshirani, and Wainwright (2015, p. 12), and the relaxed lasso estimator with $\phi = 0$ by Meinshausen (2007, p. 376). Also see Fan and Li (2001), Friedman, Hastie, and Tibshirani (2010), Simon et al. (2011), Tibshirani (1996), and Zou and Hastie (2005). The Meinshausen (2007) relaxed lasso estimator fits lasso with penalty λ_n to get a subset of variables with nonzero coefficients, and then fits lasso with a smaller penalty ϕ_n to this subset of variables where n is the sample size.

Two important quantities for a regression model are the sufficient predictor $SP = \boldsymbol{x}^T\boldsymbol{\beta}$, and the estimated sufficient predictor $ESP = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$. For the regression models, the conditioning and subscripts, such as i, will often be suppressed. The multiple linear regression model is $Y|\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{\beta} + \boldsymbol{e}$ or $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ for $i = 1, ..., n$. Consider a parametric regression model $Y|\boldsymbol{x} \sim D(\boldsymbol{x}^T\boldsymbol{\beta}, \boldsymbol{\gamma})$ where D is a parametric distribution that depends on the $p \times 1$ vector of predictors $\boldsymbol{x}$ only through $\boldsymbol{x}^T\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ is a $q \times 1$ vector of parameters. Three examples used in the simulations follow. The *binomial logistic regression model* is $Y_i \sim \text{binomial}\left(\mathrm{m_i}, \rho(\mathrm{SP}) = \dfrac{\mathrm{e^{SP}}}{1 + \mathrm{e^{SP}}}\right)$. The binary logistic regression model has $m_i \equiv 1$ for $i = 1, ..., n$. A useful *Poisson regression model* is $Y \sim \text{Poisson}\left(\mathrm{e^{SP}}\right)$. The *Weibull proportional hazards regression model* is

$$Y|SP \sim W(\gamma, \lambda_0 \exp(SP))$$

where Y has a Weibull $W(\gamma, \lambda)$ distribution if the probability density function of Y is

$$f(y) = \lambda\gamma y^{\gamma-1} \exp[-\lambda y^\gamma] \ \text{ for } \ \mathrm{y} > 0.$$

Following Olive and Hawkins (2005), a *model for variable selection* can be described by

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S \tag{1}$$

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 from 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

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_I^T\boldsymbol{\beta}_I + \boldsymbol{x}_O^T\boldsymbol{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I.$$

Thus $\boldsymbol{\beta}_O = \boldsymbol{0}$ if $S \subseteq I$. The model using $\boldsymbol{x}^T\boldsymbol{\beta}$ is the full model.

To clarify notation, suppose $p = 4$, a constant $x_1 = 1$ corresponding to β_1 is always in the model, and $\boldsymbol{\beta} = (\beta_1, \beta_2, 0, 0)^T$. Then there are $J = 2^{p-1} = 8$ possible subsets of $\{1, 2, ..., p\}$ that contain 1, including $I_1 = \{1\}$ and $S = I_2 = \{1, 2\}$. There are $2^{p-a_S} = 4$ subsets such that $S \subseteq I_j$. Let $\hat{\boldsymbol{\beta}}_{I_2} = (\hat{\beta}_1, \hat{\beta}_2)^T$ and $\boldsymbol{x}_{I_2} = (x_1, x_2)^T$.

Let I_{min} correspond to the set of predictors selected by a variable selection method such as forward selection or lasso variable selection. If $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$, use zero padding to 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 the observed variable selection estimator $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$. As a statistic, $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities $\pi_{kn} = P(I_{min} = I_k)$ for $k = 1, ..., J$ where there are J subsets, e.g. $J = 2^p - 1$.

The large sample theory for $\hat{\boldsymbol{\beta}}_{MIX}$, defined below, is useful for explaining the large sample theory of $\hat{\boldsymbol{\beta}}_{VS}$. Let $\hat{\boldsymbol{\beta}}_{MIX}$ be a random vector with a mixture distribution of the $\hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities equal to π_{kn}. Hence $\hat{\boldsymbol{\beta}}_{MIX} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with the same probabilities π_{kn} of the variable selection estimator $\hat{\boldsymbol{\beta}}_{VS}$, but the I_k are randomly selected. A random vector $\boldsymbol{u}$ has a mixture distribution of random vectors $\boldsymbol{u}_j$ with probabilities π_j if $\boldsymbol{u}$ equals the randomly selected 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

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

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

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

Inference will consider bootstrap hypothesis testing with confidence intervals (CIs) and regions. 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. A large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is a set $\mathcal{A}_n$ such that $P(\boldsymbol{\theta} \in \mathcal{A}_n)$ is eventually bounded below by $1-\delta$ as the sample size $n \to \infty$. Then reject H_0 if $\boldsymbol{\theta}_0$ is not in the confidence region. Let the $g \times 1$ vector T_n be an estimator of $\boldsymbol{\theta}$. 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, test $H_0 : \boldsymbol{A\beta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{A\beta} \neq \boldsymbol{\theta}_0$ with $\boldsymbol{\theta} = \boldsymbol{A\beta}$. Then let $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{SEL}$ and let $T_i^* = \boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{SEL}$ for $i = 1, ..., B$ and SEL is VS or MIX. Let $\lceil x \rceil$ be the smallest integer $\geq x$. For $g = 1$, let the shortest closed interval containing at least c of the T_i^* be the shorth(c) estimator. See Frey (2013). Then the large sample $100(1-\delta)\%$ shorth(c) CI for θ is

$$[T^*_{(s)}, T^*_{(s+c-1)}] \text{ with } \mathrm{c = min(B, \lceil B[1 - \delta + 1.12\sqrt{\delta/n}\ ]\ \rceil)}. \tag{2}$$

The shorth confidence interval is a practical implementation of the Hall (1988) shortest bootstrap interval based on all possible bootstrap samples.

The confidence regions use Mahalanobis distances D_i and a correction factor to get better coverage when $B \geq 50g$. This result is useful because the bootstrap confidence regions can be slow to simulate. Let

$$q_B = \min(1 - \delta + 0.05, 1 - \delta + g/B) \text{ for } \delta > 0.1 \text{ and}$$

$$q_B = \min(1 - \delta/2, 1 - \delta + 10\delta g/B), \quad \text{otherwise.} \tag{3}$$

If $1 - \delta < 0.999$ and $q_B < 1 - \delta + 0.001$, set $q_B = 1 - \delta$. Let $D_{(U_B)}$ be the $100q_B$th sample percentile of the D_i. Let T be $g \times 1$ and let $\boldsymbol{C}$ be a $g \times g$ symmetric positive definite matrix. Then the ith *squared sample Mahalanobis distance* is the scalar

$$D_i^2 = D_i^2(T, \boldsymbol{C}) = D^2_{\boldsymbol{z}_i}(T, \boldsymbol{C}) = (\boldsymbol{z}_i - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z}_i - T)$$

for each observation $\boldsymbol{z}_i$. Let $\overline{T}^*$ and $\boldsymbol{S}_T^*$ be the sample mean and sample covariance matrix of the bootstrap sample.

The Olive (2017ab, 2018) prediction region method (4), modified Bickel and Ren (2001) (5), and Pelawa Watagoda and Olive (2021) hybrid (6) large sample $100(1-\delta)\%$ confidence regions for $\boldsymbol{\theta}$ are $\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{S}^*_T) \le D^2_{(U_B)}\} =$

$$\{\boldsymbol{w} : (\boldsymbol{w} - \overline{T}^*)^T [\boldsymbol{S}^*_T]^{-1} (\boldsymbol{w} - \overline{T}^*) \le D^2_{(U_B)}\} \tag{4}$$

where $D^2_{(U_B)}$ is computed from $D_i^2 = (T_i^* - \overline{T}^*)^T [\boldsymbol{S}^*_T]^{-1} (T_i^* - \overline{T}^*)$ for $i = 1, ..., B$ (if $g = 1$, (4) is a closed interval centered at $\overline{T}^*$ just long enough to cover U_B of the T_i^*), $\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}^*_T) \le D^2_{(U_B,T)}\} =$

$$\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}^*_T]^{-1} (\boldsymbol{w} - T_n) \le D^2_{(U_B,T)}\} \tag{5}$$

where the cutoff $D^2_{(U_B,T)}$ is the $100q_B$th sample percentile of the $D_i^2 =$
$(T_i^* - T_n)^T [\boldsymbol{S}^*_T]^{-1} (T_i^* - T_n)$, and $\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}^*_T) \le D^2_{(U_B)}\} =$

$$\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}^*_T]^{-1} (\boldsymbol{w} - T_n) \le D^2_{(U_B)}\}. \tag{6}$$

Under regularity conditions, Olive (2017b, 2018) proved that (4) is a large sample confidence region. See Bickel and Ren (2001) for (5), while Pelawa Watagoda and Olive (2021) gave simpler proofs and proved that (2) is a large sample CI. Assume $\boldsymbol{u}_n \xrightarrow{D} \boldsymbol{u}$ where

$\boldsymbol{u}_n = \sqrt{n}(T_i^* - T_n)$, $\sqrt{n}(T_i^* - \overline{T}^*)$, $\sqrt{n}(T_n - \boldsymbol{\theta})$, or $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta})$, and $n\boldsymbol{S}_T^* \xrightarrow{P} \boldsymbol{C}$ where $\boldsymbol{C}$ is nonsingular. Let

$$D_1^2 = D_{T_i^*}^2(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(T_i^* - \overline{T}^*)^T (n\boldsymbol{S}_T^*)^{-1} \sqrt{n}(T_i^* - \overline{T}^*),$$

$$D_2^2 = D_{\boldsymbol{\theta}}^2(T_n, \boldsymbol{S}_T^*) = \sqrt{n}(T_n - \boldsymbol{\theta})^T (n\boldsymbol{S}_T^*)^{-1} \sqrt{n}(T_n - \boldsymbol{\theta}),$$

$$D_3^2 = D_{\boldsymbol{\theta}}^2(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(\overline{T}^* - \boldsymbol{\theta})^T (n\boldsymbol{S}_T^*)^{-1} \sqrt{n}(\overline{T}^* - \boldsymbol{\theta}), \quad \text{and}$$

$$D_4^2 = D_{T_i^*}^2(T_n, \boldsymbol{S}_T^*) = \sqrt{n}(T_i^* - T_n)^T (n\boldsymbol{S}_T^*)^{-1} \sqrt{n}(T_i^* - T_n).$$

Then $D_j^2 \approx \boldsymbol{u}^T (n\boldsymbol{S}_T^*)^{-1} \boldsymbol{u} \approx \boldsymbol{u}^T \boldsymbol{C}^{-1} \boldsymbol{u}$, and the percentiles of D_1^2 and D_4^2 can be used as cutoffs. Confidence regions (4) and (6) have the same volume.

The ratio of the volumes of regions (4) and (5) is

$$\frac{|\boldsymbol{S}_T^*|^{1/2}}{|\boldsymbol{S}_T^*|^{1/2}} \left(\frac{D_{(U_B)}}{D_{(U_B,T)}}\right)^g = \left(\frac{D_{(U_B)}}{D_{(U_B,T)}}\right)^g. \tag{7}$$

The volume of confidence region (5) tends to be greater than that of (4) since the T_i^* are closer to $\overline{T}^*$ than T_n on average.

Section 2 gives large sample theory for $\hat{\boldsymbol{\beta}}_{MIX}$ and $\hat{\boldsymbol{\beta}}_{VS}$. Section 3 shows how to bootstrap these two estimators, and Section 4 gives a simulation.

2. Large Sample Theory For Variable Selection Estimators

The new Theorems 1 and 3 in this section generalize the Pelawa Watagoda and Olive (2020, 2021) theory for multiple linear regression to many other models. Theorem 2, due to Pelawa Watagoda and Olive (2021), is added for reference with an improved proof. The theory assumes that there is a "true model" S and that at least one subset I is considered such that $S \subseteq I$. For example, with forward selection and backward elimination, the theory assumes that the full model contains S. The theory does not hold if the true model S is not a subset of any of the considered models. For example, S could contain some interactions that were not included in the "full" model. Checking that the full model is good is important.

Assume p is fixed. Suppose model (1) holds, and that if $S \subseteq I_j$ where the dimension of I_j is a_j, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \boldsymbol{V}_j)$ where $\boldsymbol{V}_j$ is the covariance matrix of the asymptotic multivariate normal distribution. Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}_{j,0}) \tag{8}$$

where $\boldsymbol{V}_{j,0}$ adds columns and rows of zeros corresponding to the x_i not in I_j, and $\boldsymbol{V}_{j,0}$ is singular unless I_j corresponds to the full model. This large sample theory holds for many models, including multiple linear regression fit by least squares (OLS), GLMs fit by maximum likelihood, and Cox regression fit by maximum partial likelihood. See, for example, Sen and Singer (1993, pp. 280, 309).

The first assumption in Theorem 1 is $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Then the variable selection estimator corresponding to I_{min} underfits with probability going to zero, and the assumption holds under regularity conditions if BIC or AIC is used for many parametric regression models such as GLMs. See Charkhi and Claeskens (2018) and Claeskens and Hjort (2008, pp. 70, 101, 102, 114, 232). This assumption is a necessary condition for a variable selection estimator to be a consistent estimator. See Zhao and Yu (2006). Thus if a shrinkage estimator that does variable selection is a consistent estimator of $\boldsymbol{\beta}$, then $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Hence Theorem 1c) proves that the lasso variable selection and elastic net variable selection estimators are $\sqrt{n}$ consistent estimators of $\boldsymbol{\beta}$ if lasso and elastic net are consistent. Also see Theorem 3. The assumption on $\boldsymbol{u}_{jn}$ in Theorem 1 is reasonable by (8) since $S \subseteq I_j$ for each π_j, and since $\hat{\boldsymbol{\beta}}_{MIX}$ uses random selection.

Consider the assumption $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$ for multiple linear regression. Charkhi and Claeskens (2018) proved the assumption holds for AIC for a wide variety of error distributions. Shao (1993) gave similar results for AIC, BIC, and C_p. The assumption holds for lasso variable selection and elastic net variable selection provided that $\hat{\lambda}_n/n \to 0$ as $n \to \infty$ so lasso and elastic net are consistent estimators. Here $\hat{\lambda}_n$ is the shrinkage penalty parameter selected after k-fold cross validation. See Pelawa Watogoda and Olive (2020) and Knight and Fu (2000). Next we give a new argument for the Mallows (1973) C_p criterion when each submodel contains a constant. Let submodel I have $k \leq p$ predictors including a constant. Then

$$C_p(I) = \frac{SSE(I)}{MSE} + 2k - n$$

where MSE is for the full model, and $C_p(I) \geq -p$. Assume the full model is one of the submodels considered with $C_p(full) = p$, e.g. forward selection, backward elimination, stepwise selection, and all subsets selection. Then $-p \leq C_p(I_{min}) \leq p$. Let $\boldsymbol{r}$ be the residual vector for the full model and $\boldsymbol{r}_I$ that for the submodel. Then the correlation

$$corr(r, r_I) = \sqrt{\frac{n-p}{C_p(I) + n - 2k}}$$

by Olive and Hawkins (2005). Thus $corr(r, r_{I_{min}}) \to 1$ as $n \to \infty$. Suppose S is not a subset of I. Under the model $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$, $corr(r, r_I)$ will not converge to 1 as $n \to \infty$, and for large enough n, $[corr(r, r_I)]^2 \leq \gamma < 1$. Thus $C_p(I) \to \infty$ as $n \to \infty$. Hence $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$ if the zero mean iid errors have constant variance σ^2.

Theorem 1 a) proves that $\boldsymbol{u}$ is a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j, $E(\boldsymbol{u}) = \boldsymbol{0}$, and $\text{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma_u} = \sum_j \pi_j \boldsymbol{V}_{j,0}$. Some of the submodels I_k will have $\pi_k = 0$. For example, since the probability of underfitting goes to zero, every submodel I_k that underfits has $\pi_k = 0$. Hence $S \subseteq I_j$ corresponding to the $\pi_j > 0$. If $\pi_d = 1$, then submodel I_d is picked with probability going to 1 as $n \to \infty$, and I_d is the only submodel with a positive π_k. Often $\pi_d = \pi_S$ in the literature.

Theorem 1 *Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and let $\hat{\boldsymbol{\beta}}_{MIX} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities π_{kn} where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the positive π_k by π_j. Assume $\boldsymbol{u}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}_j \sim N_p(\boldsymbol{0}, \boldsymbol{V}_{j,0})$. a) Then*

$$\boldsymbol{u}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u} \tag{9}$$

where the cdf of $\boldsymbol{u}$ is $F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})$.

b) Let $\boldsymbol{A}$ be a $g \times p$ full rank matrix with $1 \leq g \leq p$. Then

$$\boldsymbol{v}_n = \boldsymbol{A}\boldsymbol{u}_n = \sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} = \boldsymbol{v} \tag{10}$$

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

c) The estimator $\hat{\boldsymbol{\beta}}_{VS}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$: $\sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) = O_P(1)$.

d) If $\pi_d = 1$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{SEL} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u} \sim N_p(\boldsymbol{0}, \boldsymbol{V}_{d,0})$ where SEL is VS or MIX.

Proof. a) Since $\boldsymbol{u}_n$ has a mixture distribution of the $\boldsymbol{u}_{kn}$ with probabilities π_{kn}, the cdf of $\boldsymbol{u}_n$ is $F_{\boldsymbol{u}_n}(\boldsymbol{t}) = \sum_k \pi_{kn} F_{\boldsymbol{u}_{kn}}(\boldsymbol{t}) \to F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})$ at continuity points of the $F_{\boldsymbol{u}_j}(\boldsymbol{t})$ as $n \to \infty$.
b) Since $\boldsymbol{u}_n \xrightarrow{D} \boldsymbol{u}$, then $\boldsymbol{A}\boldsymbol{u}_n \xrightarrow{D} \boldsymbol{A}\boldsymbol{u}$.
c) The result follows since selecting from a finite number J of $\sqrt{n}$ consistent estimators (even on a set that goes to one in probability) results in a $\sqrt{n}$ consistent estimator by Pratt (1959).
d) If $\pi_d = 1$, there is no selection bias, asymptotically. The result also follows by Pötscher (1991, Lemma 1). □

The following subscript notation is useful. Subscripts before the MIX are used for subsets of $\hat{\boldsymbol{\beta}}_{MIX} = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$. Let $\hat{\boldsymbol{\beta}}_{i,MIX} = \hat{\beta}_i$. Similarly, if $I = \{i_1, ..., i_a\}$, then $\hat{\boldsymbol{\beta}}_{I,MIX} = (\hat{\beta}_{i_1}, ..., \hat{\beta}_{i_a})^T$. Subscripts after MIX denote the ith vector from a sample $\hat{\boldsymbol{\beta}}_{MIX,1}, ..., \hat{\boldsymbol{\beta}}_{MIX,B}$. Similar notation is used for other estimators such as $\hat{\boldsymbol{\beta}}_{VS}$. The subscript 0 is still used for zero padding. We may use $FULL$ to denote the full model $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{FULL}$.

Typically the mixture distribution is not asymptotically normal unless a $\pi_d = 1$ (e.g. if S is the full model), or if for each π_j, $\boldsymbol{A}\boldsymbol{u}_j \sim N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T) = N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$. Then $\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$. This special case occurs for $\hat{\boldsymbol{\beta}}_{S,MIX}$ if $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V})$ where the asymptotic covariance matrix $\boldsymbol{V}$ is diagonal and nonsingular. Then $\hat{\boldsymbol{\beta}}_{S,MIX}$ and $\hat{\boldsymbol{\beta}}_{S,FULL}$ have the same multivariate normal limiting distribution. For several criteria, this result should hold for $\hat{\boldsymbol{\beta}}_{VS}$ since asymptotically, $\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{A}\boldsymbol{\beta})$ is selecting from the $\boldsymbol{A}\boldsymbol{u}_j$ which have the same distribution. In the simulations when $\boldsymbol{V}$ is diagonal, the confidence regions applied to $\boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{SEL} = \boldsymbol{B}\hat{\boldsymbol{\beta}}^*_{S,SEL}$ had similar volume and cutoffs where SEL is MIX, VS, or $FULL$.

Theorem 1 can be used to justify prediction intervals after variable selection. See Olive, Rathnayake, and Haile (2021). Theorem 1d) is useful for *variable selection consistency* and the *oracle property* where $\pi_d = \pi_S = 1$ if $P(I_{min} = S) \to 1$ as $n \to \infty$. See Claeskens and Hjort (2008, pp. 101-114) and Fan and Li (2001) for references. A necessary condition for $P(I_{min} = S) \to 1$ is that S is one of the models considered with probability going to one. This condition holds under strong regularity conditions for fast methods. See Wieczorek

(2018) for forward selection and Hastie, Tibshirani, and Wainwright (2015, pp. 295-302) for lasso, where the predictors need a "near orthogonality" condition.

The following Pelawa Watagoda and Olive (2021) theorem is useful for bootstrapping variable selection estimators. Let $(\overline{T}, \boldsymbol{S}_T)$ be the sample mean and sample covariance matrix computed from $T_1, ..., T_B$ which have the same distribution as T_n where $T_i = T_{in}$. Let $D^2_{(U_B)}$ be the cutoff computed from the $D^2_i(\overline{T}, \boldsymbol{S}_T)$ for $i = 1, ..., B$. The hyperellipsoids corresponding to $D^2(T_n, \boldsymbol{C})$ and $D^2(\overline{T}, \boldsymbol{C})$ are centered at T_n and $\overline{T}$, respectively. Note that $D^2_{\overline{T}}(T_n, \boldsymbol{C}) = D^2_{T_n}(\overline{T}, \boldsymbol{C})$. Thus $D^2_{\overline{T}}(T_n, \boldsymbol{C}) \leq D^2_{(U_B)}$ iff $D^2_{T_n}(\overline{T}, \boldsymbol{C}) \leq D^2_{(U_B)}$. In Theorem 2, since R_p contains T_f with probability $1 - \delta_B$, the region R_c contains $\overline{T}$ with probability $1 - \delta_B$. Since T_n depends on the sample size n, we need $(n\boldsymbol{S}_T)^{-1}$ to be fairly well behaved, e.g. $(n\boldsymbol{S}_T)^{-1} \xrightarrow{P} \boldsymbol{\Sigma}_A^{-1}$.

Theorem 2: Geometric Argument. *Suppose* $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ *with* $E(\boldsymbol{u}) = \boldsymbol{0}$ *and* $Cov(\boldsymbol{u}) = \boldsymbol{\Sigma_u} \neq \boldsymbol{0}$. *Assume* $T_1, ..., T_B$ *are iid with nonsingular covariance matrix* $\boldsymbol{\Sigma}_{T_n}$ *where* $(n\boldsymbol{S}_T)^{-1} \xrightarrow{P} \boldsymbol{\Sigma}_A^{-1}$. *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* $1 - \delta_B$ *which is eventually bounded below by* $1 - \delta$ *as* $B \to \infty$. *Hence the region* $R_c = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T) \leq D^2_{(U_B)}\}$ *is a large sample* $100(1-\delta)\%$ *confidence region for* $\boldsymbol{\theta}$ *where* T_n *is a randomly selected* T_i.

Proof. The region R_c centered at a randomly selected T_n contains $\overline{T}$ with probability $1 - \delta_B$ which is eventually bounded below by $1 - \delta$ as $B \to \infty$. Since the $\sqrt{n}(T_i - \boldsymbol{\theta})$ are iid,

$$\begin{bmatrix} \sqrt{n}(T_1 - \boldsymbol{\theta}) \\ \vdots \\ \sqrt{n}(T_B - \boldsymbol{\theta}) \end{bmatrix} \xrightarrow{D} \begin{bmatrix} \boldsymbol{v}_1 \\ \vdots \\ \boldsymbol{v}_B \end{bmatrix}$$

where the $\boldsymbol{v}_i$ are iid with the same distribution as $\boldsymbol{u}$. For fixed B, the average of these random vectors is

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

where AN_g denotes an approximate multivariate normal distribution. Hence $(\overline{T} - \boldsymbol{\theta}) =$

$O_P((nB)^{-1/2})$, and $\overline{T}$ gets arbitrarily close to $\boldsymbol{\theta}$ compared to T_n as $B \to \infty$. Thus R_c is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ as $n, B \to \infty$. $\square$

Examining the iid data cloud $T_1, ..., T_B$ and the bootstrap sample data cloud $T_1^*, ..., T_B^*$ is often useful for understanding the bootstrap. If $\sqrt{n}(T_n - \boldsymbol{\theta})$ and $\sqrt{n}(T_i^* - T_n)$ both converge in distribution to $\boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{\Sigma})$, 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 (6) is a confidence region by the geometric argument (as is region (5) which tends to use a larger cutoff), and (4) is a confidence region if $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \boldsymbol{0}$.

For $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX}$ with $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta}$, we have $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{v}$ by (10) where $E(\boldsymbol{v}) = \boldsymbol{0}$, and $\boldsymbol{\Sigma}_{\boldsymbol{v}} = \sum_j \pi_j \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T$. By Theorem 2, if we had iid data $T_1, ..., T_B$, then R_c would be a large sample confidence region for $\boldsymbol{\theta}$. If $\sqrt{n}(T_n^* - T_n) \xrightarrow{D} \boldsymbol{v}$, then we could use the bootstrap sample and confidence regions (4) to (6). This condition holds only under strong regularity conditions such as $\pi_d = 1$ or $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{B}\boldsymbol{\beta}_S$ if $\boldsymbol{V}$ was diagonal. Section 3 will explain why the bootstrap confidence regions may still be useful.

Pötscher (1991) used the conditional distribution of $\hat{\boldsymbol{\beta}}_{VS}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})$ to find the distribution of $\boldsymbol{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta})$. Define $P(A|B_k)P(B_k) = 0$ if $P(B_k) = 0$. Let $\hat{\boldsymbol{\beta}}^C_{I_k,0}$ be a random vector from the conditional distribution $\hat{\boldsymbol{\beta}}_{I_k,0}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})$. Let $\boldsymbol{w}_{kn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta})|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}) \sim \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_k,0} - \boldsymbol{\beta})$. Denote $F_{\boldsymbol{z}}(\boldsymbol{t}) = P(z_1 \leq t_1, ..., z_p \leq t_p)$ by $P(\boldsymbol{z} \leq \boldsymbol{t})$. Then Pötscher (1991) and Pelawa Watagoda and Olive (2020) show

$$F_{\boldsymbol{w}_n}(\boldsymbol{t}) = P[n^{1/2}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \leq \boldsymbol{t}] = \sum_{k=1}^{J} F_{\boldsymbol{w}_{kn}}(\boldsymbol{t})\pi_{kn}.$$

Hence $\hat{\boldsymbol{\beta}}_{VS}$ has a mixture distribution of the $\hat{\boldsymbol{\beta}}^C_{I_k,0}$ with probabilities π_{kn}, and $\boldsymbol{w}_n$ has a mixture distribution of the $\boldsymbol{w}_{kn}$ with probabilities π_{kn}.

Charkhi and Claeskens (2018) showed that $\boldsymbol{w}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w}_j$ if $S \subseteq I_j$ for the maximum likelihood estimator (MLE) with AIC, and gave a forward selection example. Here $\boldsymbol{w}_j$ is a multivariate truncated normal distribution (where no truncation is possible) that is symmetric about $\boldsymbol{0}$. Note that both $\sqrt{n}(\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{\beta})$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta})$ are selecting from the $\boldsymbol{u}_{kn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta})$ and asymptotically from the $\boldsymbol{u}_j$. The random selection for $\hat{\boldsymbol{\beta}}_{MIX}$

does not change the distribution of $\boldsymbol{u}_{jn}$, but selection bias does change the distribution of the selected $\boldsymbol{u}_{jn}$ and $\boldsymbol{u}_j$ to that of $\boldsymbol{w}_{jn}$ and $\boldsymbol{w}_j$. The assumption that $\boldsymbol{w}_{jn} \xrightarrow{D} \boldsymbol{w}_j$ may not be mild. The proof for Equation (11) is the same as that for (9). Theorem 3 proves that $\boldsymbol{w}$ is a mixture distribution of the $\boldsymbol{w}_j$ with probabilities π_j.

Theorem 3. *Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and let $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities π_{kn} where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the positive π_k by π_j. Assume $\boldsymbol{w}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w}_j$. Then*

$$\boldsymbol{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w} \tag{11}$$

where the cdf of $\boldsymbol{w}$ is $F_{\boldsymbol{w}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{w}_j}(\boldsymbol{t})$.

3. Bootstrapping Variable Selection Estimators

Obtaining the bootstrap samples for $\hat{\boldsymbol{\beta}}_{VS}$ and $\hat{\boldsymbol{\beta}}_{MIX}$ is simple. Generate $\boldsymbol{Y}^*$ and $\boldsymbol{X}^*$ that would be used to produce $\hat{\boldsymbol{\beta}}^*$ if the full model estimator $\hat{\boldsymbol{\beta}}$ was being bootstrapped. Instead of computing $\hat{\boldsymbol{\beta}}^*$, compute the variable selection estimator $\hat{\boldsymbol{\beta}}^*_{VS,1} = \hat{\boldsymbol{\beta}}^{*C}_{I_{k_1},0}$. Then generate another $\boldsymbol{Y}^*$ and $\boldsymbol{X}^*$ and compute $\hat{\boldsymbol{\beta}}^*_{MIX,1} = \hat{\boldsymbol{\beta}}^*_{I_{k_1},0}$ (using the same subset I_{k_1}). This process is repeated B times to get the two bootstrap samples for $i = 1, ..., B$. Let the selection probabilities for the bootstrap variable selection estimator be ρ_{kn}. Then this bootstrap procedure bootstraps both $\hat{\boldsymbol{\beta}}_{VS}$ and $\hat{\boldsymbol{\beta}}_{MIX}$ with $\pi_{kn} = \rho_{kn}$. Then apply the confidence regions (4), (5), and (6) on the bootstrap sample $T_1^*, ..., T_B^*$ where $T_i^* = \boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{SEL,i}$ where SEL is VS or MIX.

By Section 2, we expect the confidence regions to simulate well (have coverage close to or higher than the nominal level so that the type I error is close to or less than the nominal level) if $\pi_d = 1$ or if the asymptotic covariance matrix for the full model is nonsingular and diagonal, but these conditions are very strong. In simulations for $\hat{\boldsymbol{\beta}}_{VS}$ with $n \geq 20p$, if the confidence regions (4) and (5) simulated well for the full model bootstrap, then (4) and (5) also simulated well for $\hat{\boldsymbol{\beta}}_{VS}$. The hybrid confidence region (6) had poorer performance, and confidence regions for $\hat{\boldsymbol{\beta}}_{VS}$ tended to have less undercoverage than confidence regions for $\hat{\boldsymbol{\beta}}^*_{MIX}$.

Undercoverage can occur if the bootstrap data cloud is less variable than the iid data cloud, e.g., if $n < 20p$. Heuristically, if $n \geq 20p$, then coverage can be higher than the nominal coverage for two reasons: i) the bootstrap data cloud $T_1^*, ..., T_B^*$ is more variable than the iid data cloud of $T_1, ..., T_B$, and ii) zero padding. In the simulations for $H_0 : \boldsymbol{A\beta} = \boldsymbol{B\beta}_S = \boldsymbol{\theta}$, the simulated coverage for confidence intervals and confidence regions (4) and (5) was roughly 2% less than to 2% higher than the nominal 95% coverage due to i). In the simulations for $H_0 : \boldsymbol{A\beta} = \boldsymbol{B\beta}_E = \boldsymbol{0}$, the simulated coverage for confidence intervals and confidence regions (4) and (5) tended to be close to 99% when the nominal coverage was 95%, but the nominal 95% confidence intervals tended to be shorter than those for the full model, and the confidence region volumes were often much smaller than those for the full model. See Pelawa Watagoda and Olive (2021) for more on why zero padding tends to increase the coverage while decreasing the volume of the confidence regions and confidence intervals. The simulations also used $B \geq \max(200, 50p)$ so that $\boldsymbol{S}_T^*$ is a good estimator of $\text{Cov}(T^*)$.

The matrix $\boldsymbol{S}_T^*$ can be singular due to one or more columns of zeros 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. A simple remedy is to add d bootstrap samples of the full model estimator $\hat{\boldsymbol{\beta}}^* = \hat{\boldsymbol{\beta}}^*_{FULL}$ to the bootstrap sample. For example, take $d = \lceil cB \rceil$ with $c = 0.01$. A confidence interval $[L_n, U_n]$ can be computed without $\boldsymbol{S}_T^*$ for (4), (5), and (6). Using the confidence interval $[\max(L_n, T^*_{(1)}), \min(U_n, T^*_{(B)})]$ can give a shorter covering region.

Next we examine why the bootstrap data cloud tends to be more variable than the iid data cloud. 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

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

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 $\boldsymbol{S}^*_{T,j}$. 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 $\boldsymbol{S}_{T,j}$.

Let $T_n = \hat{\boldsymbol{\beta}}_{MIX}$. If $S \subseteq I_j$, assume $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \boldsymbol{V}_j)$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}^*_{I_j} - \hat{\boldsymbol{\beta}}_{I_j}) \xrightarrow{D}$

$N_{a_j}(\mathbf{0}, \boldsymbol{V}_j)$. Then by Equation (8),

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}_{j,0}) \text{ and } \sqrt{\mathrm{n}}(\hat{\boldsymbol{\beta}}^*_{\mathrm{I_j,0}} - \hat{\boldsymbol{\beta}}_{\mathrm{I_j,0}}) \xrightarrow{\mathrm{D}} \mathrm{N_p}(\mathbf{0}, \boldsymbol{V}_{\mathrm{j,0}}). \tag{12}$$

If Equation (12) holds, then the component clouds have the same variability asymptotically, and the confidence regions will shrink to a point at $\boldsymbol{\beta}$ as $n \to \infty$, giving good test power, asymptotically. The iid data component clouds are all centered at $\boldsymbol{\beta}$. If the bootstrap data component clouds were all centered at the same value $\tilde{\boldsymbol{\beta}}$, then the bootstrap cloud would be like an iid data cloud shifted to be centered at $\tilde{\boldsymbol{\beta}}$, and (5) and (6) would be confidence regions for $\boldsymbol{\theta} = \boldsymbol{\beta}$ by Theorem 2. Instead, the bootstrap data component clouds are shifted slightly from a common center, and are each centered at a $\hat{\boldsymbol{\beta}}_{I_j,0}$. Geometrically, the shifting of the bootstrap component data clouds makes the bootstrap data cloud more variable than the iid data cloud, asymptotically (we want $n \geq 20p$). The shifting also makes the T_i^* further from $\overline{T}^*$ than if there is no shifting. A similar argument can be given for $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX}$ and $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta}$. Region (4) has the same volume as region (6), but tends to have higher coverage since empirically, the bagging estimator $\overline{T}^*$ tends to estimate $\boldsymbol{\theta}$ at least as well as T_n for a mixture distribution. See Breiman (1996) and Yang (2003).

The above argument is heuristic since we have not been able to prove that the coverage is $\geq 1 - \delta$, asymptotically, except under strong regularity conditions. Then the type I error $\leq \delta$, asymptotically. Confidence region (5) rejects H_0 if $(T_n - \boldsymbol{\theta}_0)^T[\boldsymbol{S}^*_T]^{-1}(T_n - \boldsymbol{\theta}_0) > D^2_{(U_B,T)}$. If an iid data cloud was available, the cutoff $D^2_{(U_B)}(T_n, \boldsymbol{S}^*_T)$ could be computed from $D^2_i = (T_i - \boldsymbol{\theta}_0)^T[\boldsymbol{S}^*_T]^{-1}(T_i - \boldsymbol{\theta}_0)$ for $i = 1, ..., B$. Hence the type I error is controlled if $D^2_{(U_B,T)}$ tends to be larger than $D^2_{(U_B)}(T_n, \boldsymbol{S}^*_T)$.

The bootstrap component clouds for $\hat{\boldsymbol{\beta}}^*_{VS}$ are again separated compared to the iid clouds for $\hat{\boldsymbol{\beta}}_{VS}$, which are centered about $\boldsymbol{\beta}$. Heuristically, most of the selection bias is due to predictors in E, not to the predictors in S. Hence $\hat{\boldsymbol{\beta}}^*_{S,VS}$ is roughly similar to $\hat{\boldsymbol{\beta}}^*_{S,MIX}$. Typically the distributions of $\hat{\boldsymbol{\beta}}^*_{E,VS}$ and $\hat{\boldsymbol{\beta}}^*_{E,MIX}$ are not similar, but use the same zero padding.

Next we will examine when Equation (12) holds. If $S \subseteq I_j$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\mathbf{0}, \boldsymbol{V}_j)$ by the large sample theory (8) for the estimator. Bootstrap theory should show that $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V})$, but showing $\sqrt{n}(\hat{\boldsymbol{\beta}}^*_{I_j} - \hat{\boldsymbol{\beta}}_{I_j}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \boldsymbol{V}_j)$ is often difficult.

The nonparametric bootstrap (also called the empirical bootstrap, naive bootstrap, and the pairs bootstrap) draws a sample of n cases $(Y_i^*, \boldsymbol{x}_i^*)$ with replacement from the n cases $(Y_i, \boldsymbol{x}_i)$, and regresses the Y_i^* on the $\boldsymbol{x}_i^*$ to get $\hat{\boldsymbol{\beta}}^*_{VS,1}$, and then draws another sample to get $\hat{\boldsymbol{\beta}}^*_{MIX,1}$. This process is repeated B times to get the two bootstrap samples for $i = 1, ..., B$. If $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V})$ for the full model, then $\sqrt{n}(\hat{\boldsymbol{\beta}}^*_{I_j} - \hat{\boldsymbol{\beta}}_{I_j}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \boldsymbol{V}_j)$ when $S \subseteq I_j$: just use I_j as the new full model. Thus Equation (12) should hold when the full model bootstrap works. The method is used for multiple linear regression, Cox proportional hazards regression with right censored Y_i, and GLMs. See, for example, Burr (1994), Efron and Tibshirani (1986), Freedman (1981), and Shao and Tu (1995, pp. 335-349).

For the parametric regression model $Y_i|\boldsymbol{x}_i \sim D(\boldsymbol{x}_i^T\boldsymbol{\beta}, \boldsymbol{\gamma})$, assume $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}(\boldsymbol{\beta}))$, and that $\boldsymbol{V}(\hat{\boldsymbol{\beta}}) \xrightarrow{P} \boldsymbol{V}(\boldsymbol{\beta})$ as $n \to \infty$. These assumptions tend to be mild for a parametric regression model where the MLE $\hat{\boldsymbol{\beta}}$ is used. Then $\boldsymbol{V}(\boldsymbol{\beta}) = \boldsymbol{I}^{-1}(\boldsymbol{\beta})$, the inverse Fisher information matrix. For GLMs, see, for example, Sen and Singer (1993, p. 309). For the parametric regression model, we regress $\boldsymbol{Y}$ on $\boldsymbol{X}$ to obtain $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ where the $n \times 1$ vector $\boldsymbol{Y} = (Y_i)$ and the ith row of the $n \times p$ design matrix $\boldsymbol{X}$ is $\boldsymbol{x}_i^T$.

The parametric bootstrap uses $\boldsymbol{Y}_j^* = (Y_i^*)$ where $Y_i^*|\boldsymbol{x}_i \sim D(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ for $i = 1, ..., n$. Regress $\boldsymbol{Y}_j^*$ on $\boldsymbol{X}$ to get $\hat{\boldsymbol{\beta}}_j^*$ for $j = 1, ..., B$. The large sample theory for $\hat{\boldsymbol{\beta}}^*$ is simple. Note that if $Y_i^*|\boldsymbol{x}_i \sim D(\boldsymbol{x}_i^T\boldsymbol{b}, \hat{\boldsymbol{\gamma}})$ where $\boldsymbol{b}$ does not depend on n, then $(\boldsymbol{Y}^*, \boldsymbol{X})$ follows the parametric regression model with parameters $(\boldsymbol{b}, \hat{\boldsymbol{\gamma}})$. Hence $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \boldsymbol{b}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}(\boldsymbol{b}))$. Now fix large integer n_0, and let $\boldsymbol{b} = \hat{\boldsymbol{\beta}}_{n_o}$. Then $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}_{n_o}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}(\hat{\boldsymbol{\beta}}_{n_o}))$. Since $N_p(\boldsymbol{0}, \boldsymbol{V}(\hat{\boldsymbol{\beta}})) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}(\boldsymbol{\beta}))$, we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}(\boldsymbol{\beta})) \tag{13}$$

as $n \to \infty$.

Now suppose $S \subseteq I$. Without loss of generality, let $\boldsymbol{\beta} = (\boldsymbol{\beta}_I^T, \boldsymbol{\beta}_O^T)^T$ and $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}(I)^T, \hat{\boldsymbol{\beta}}(O)^T)^T$. Then $(\boldsymbol{Y}, \boldsymbol{X}_I)$ follows the parametric regression model with parameters $(\boldsymbol{\beta}_I, \boldsymbol{\gamma})$. Hence $\sqrt{n}(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \boldsymbol{V}(\boldsymbol{\beta}_I))$. Now $(\boldsymbol{Y}^*, \boldsymbol{X}_I)$ only follows the parametric regression model asymptotically, since $\hat{\boldsymbol{\beta}}(O) \neq \boldsymbol{0}$. Then showing $\sqrt{n}(\hat{\boldsymbol{\beta}}^*_{I_j} - \hat{\boldsymbol{\beta}}_{I_j}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \boldsymbol{V}_j)$ is often difficult.

For the multiple linear regression model, $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$, assume a constant x_1 is in the model, and the zero mean e_i are iid with variance $V(e_i) = \sigma^2$. Let $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. For each I with $S \subseteq I$, assume the maximum leverage $\max_{i=1,...,n} \boldsymbol{x}_{iI}^T(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{x}_{iI} \to 0$ in probability as $n \to \infty$. For least squares with $S \subseteq I$, $\sqrt{n}(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \boldsymbol{V}_I)$ where $(\boldsymbol{X}_I^T\boldsymbol{X}_I)/(n\sigma^2) \xrightarrow{P} \boldsymbol{V}_I^{-1}$. See, for example, Sen and Singer (1993, p. 280).

Consider the parametric bootstrap for the above model with $\boldsymbol{Y}^* \sim N_n(\boldsymbol{X}\hat{\boldsymbol{\beta}}, \hat{\sigma}_n^2\boldsymbol{I}) \sim N_n(\boldsymbol{HY}, \hat{\sigma}_n^2\boldsymbol{I})$ where **we are not assuming** that the $e_i \sim N(0, \sigma^2)$, and

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

where the residuals are from the full OLS model. Then MSE is a $\sqrt{n}$ consistent estimator of σ^2 under mild conditions by Su and Cook (2012). Thus $\hat{\boldsymbol{\beta}}_I^* = (\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{Y}^* \sim N_{a_I}(\hat{\boldsymbol{\beta}}_I, \hat{\sigma}_n^2(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1})$ since $E(\hat{\boldsymbol{\beta}}_I^*) = (\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{HY} = \hat{\boldsymbol{\beta}}_I$ because $\boldsymbol{HX}_I = \boldsymbol{X}_I$, and $\text{Cov}(\hat{\boldsymbol{\beta}}_I^*) = \hat{\sigma}_n^2(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}$. Hence

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_I^* - \hat{\boldsymbol{\beta}}_I) \sim N_{a_I}(\boldsymbol{0}, n\hat{\sigma}_n^2(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \boldsymbol{V}_I)$$

as $n, B \to \infty$ if $S \subseteq I$. Hence Equation (12) holds under mild conditions.

When $\boldsymbol{V}$ is diagonal, $\sqrt{n}(\hat{\boldsymbol{\beta}}_{S,full} - \boldsymbol{\beta}_S) \xrightarrow{D} N_{a_S}(\boldsymbol{0}, \boldsymbol{V}_S)$ where $\boldsymbol{V}_S$ is a diagonal matrix using the relevant diagonal elements of $\boldsymbol{V}$. For multiple linear regression with the parametric bootstrap, the full model $\hat{\boldsymbol{\beta}}^* \sim N_p(\hat{\boldsymbol{\beta}}, \hat{\sigma}_n^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}) \approx N_p(\hat{\boldsymbol{\beta}}, \boldsymbol{V}/n)$. If the columns of $\boldsymbol{X}$ are orthogonal and $S \subseteq I$, then $\hat{\boldsymbol{\beta}}_{S,I}^* = \hat{\boldsymbol{\beta}}_{S,full}^*$ and $\hat{\boldsymbol{\beta}}_{S,I} = \hat{\boldsymbol{\beta}}_{S,full}$. Hence $\sqrt{n}(\hat{\boldsymbol{\beta}}_{S,MIX}^* - \hat{\boldsymbol{\beta}}_{S,full}) \xrightarrow{D} N_{a_S}(\boldsymbol{0}, \boldsymbol{V}_S)$. When $\boldsymbol{V}$ is diagonal, the columns of $\boldsymbol{X}$ are asymptotically orthogonal. Hence if $S \subseteq I$, $\hat{\boldsymbol{\beta}}_{S,I} \approx \hat{\boldsymbol{\beta}}_{S,full} \approx \overline{T}^*$, and the bootstrap component clouds have the same asymptotic variability as the iid data clouds. Hence we expect the bootstrap cutoffs for $\boldsymbol{A}\hat{\boldsymbol{\beta}}_{S,MIX}^*$ to be near $\chi^2_{g,1-\delta}$. Results in Pelawa Watagoda and Olive (2021) show that the residual bootstrap behaves similarly to the parametric bootstrap, with $\hat{\sigma}_n^2 = MSE$ replaced by $\tilde{\sigma}_n^2 = (n-p)MSE/n$.

The weighted least squares formulation of the GLM maximum likelihood estimator, given for example by Hillis and Davis (1994) and Sen and Singer (1993, p. 307), suggests that similar results hold for the GLM when $\boldsymbol{V}$ is diagonal.

4. Example and Simulations

Example. Lindenmayer et al. (1991) and Cook and Weisberg (1999, p. 533) gave a data set with 151 cases where Y is the number of possum species found in a tract of land in Australia. The predictors are *acacia*=basal area of acacia + 1, *bark*=bark index, *habitat*=habitat score, *shrubs*=number of shrubs + 1, *stags*= number of hollow trees + 1, *stumps*=indicator for presence of stumps, and a constant. For the full Poisson regression model, the bootstrap shorth CIs were close to the large sample GLM confidence intervals $\approx \hat{\beta}_i \pm 2SE(\hat{\beta}_i)$ (not shown). The data set is available from the Cook and Weisberg (1999) *Arc* software (https://www.stat.umn.edu/arc/).

The minimum AIC model from backward elimination used a constant, *bark*, *habitat*, and *stags*. The 95% shorth(c) confidence intervals for β_i using the parametric bootstrap are shown in Table 1. Note that most of the CIs contain 0 when closed intervals are used instead of open intervals.

Table 1: Shorth CIs for the example

variable	$\hat{\beta}_i$	95% shorth CI: VS	95% shorth CI: MIX
intercept	-0.8994	$[-1.5662, -0.5169]$	$[-1.3680, -0.3553]$
acacia	0	[0, 0.0.0384]	$[-0.0004, 0.0397]$
bark	0.0336	[0, 0.05928]	[0, 0.0563]
habitat	0.1069	[0, 0.1524]	[0, 0.1584]
shrubs	0	[0, 0.05582]	$[-0.01560, 0.04532]$
stags	0.0302	[0, 0.0540]	[0, 0.0540]
stumps	0	$[-0.9326,\ 0.0000]$	$[-0.8402, 0.1515]$

We tested $H_0 : \beta_2 = \beta_5 = \beta_7 = 0$ with the I_{min} model selected by backward elimination. (Of course this test would be easy to do with the full model using GLM theory.) Then $H_0 : \boldsymbol{A\beta} = (\beta_2, \beta_5, \beta_7)^T = \boldsymbol{0}$. Using the prediction region method with the full model had $[0, D_{(U_B)}] = [0, 2.773]$ with $D_{\boldsymbol{0}} = 2.067$. 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} backward elimination model had $[0, D_{(U_B)}] = [0, 2.702]$ while $D_{\mathbf{0}} = 1.327$. So fail to reject H_0. The ratio of the volumes of the bootstrap confidence regions for this test was 0.322. (Use (7) with $\boldsymbol{S}_T^*$ and D from backward elimination for the numerator, and from the full model for the denominator.) Hence the backward elimination bootstrap test was more precise than the full model bootstrap test. The test with $\hat{\boldsymbol{\beta}}_{MIX}$ had $[0, D_{(U_B)}] = [0, 3.157]$ while $D_{\mathbf{0}} = 1.066$. So fail to reject H_0. The ratio of the volumes of the bootstrap confidence regions for this test (MIX vs. FULL) was 0.117.

Now we describe simulations for multiple linear regression, binomial regression, Cox regression, and Poisson regression. There is a massive literature on variable selection, but most of the methods for confidence intervals and hypothesis testing are conditional on the subset I_{min} of predictors selected by the variable selection method. Then inference for the predictors that were not selected is difficult. Data splitting, the Charkhi and Claeskens (2018) method, and most high dimensional variable selection methods are conditional on I_{min}. Also, most of the methods are for multiple linear regression. Previous inference methods for forward selection, backward elimination, and lasso variable selection did not use the large sample theory given by this paper and Pelawa Watagoda and Olive (2020). For variable selection estimators of $\boldsymbol{\beta}$ with n/p large, the most important competitor is inference from the full model.

Hence our simulations compare bootstrap inference for the full model with that for the variable selection estimator of $\boldsymbol{\beta}$. We used 5000 runs, $\theta = \boldsymbol{A\beta} = \beta_i$, $\boldsymbol{\theta} = \boldsymbol{A\beta} = \boldsymbol{\beta}_S = (\beta_1, 1, ..., 1)^T$ and $\boldsymbol{\theta} = \boldsymbol{A\beta} = \boldsymbol{\beta}_E = \mathbf{0}$. The simulations often used $n = 25p$, $n = 50p$; $\psi = 0, 1/\sqrt{p}$, and 0.9; and $k = 1$ and 2 where k and ψ are defined in the following paragraph. We often used $p = 4$ since the simulations with 5000 runs take a long time.

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 $\boldsymbol{w}_i \sim N_{p-1}(\mathbf{0}, \boldsymbol{I})$ where the $q = p - 1$ elements of the vector $\boldsymbol{w}_i$ are iid N(0,1). Let the $q \times q$ matrix $\boldsymbol{A} = (a_{ij})$ with $a_{ii} = 1$ and $a_{ij} = \psi$ where $0 \le \psi < 1$ for $i \ne j$. Then the vector $\boldsymbol{z}_i = \boldsymbol{A}\boldsymbol{w}_i$ so that $\text{Cov}(\boldsymbol{z}_i) = \boldsymbol{\Sigma}_{\boldsymbol{z}} = \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ where the diagonal entries $\sigma_{ii} = [1 + (q-1)\psi^2]$ and the off diagonal entries $\sigma_{ij} = [2\psi + (q-2)\psi^2]$. Hence the correlations are $cor(z_i, z_j) = \rho = (2\psi + (q-2)\psi^2)/(1 + (q-1)\psi^2)$ for $i \neq j$. Then $\sum_{j=1}^k z_j \sim N(0, k\sigma_{ii} + k(k-1)\sigma_{ij}) = N(0, v^2)$. Let $\boldsymbol{u} = a\boldsymbol{z}/v$. Then $cor(x_i, x_j) = \rho$ 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 $\boldsymbol{u}_i$ cluster about the line in the direction of $(1, ..., 1)^T$. Let $SP = \boldsymbol{x}^T\boldsymbol{\beta} = \beta_1 + 1x_{i,2} + \cdots + 1x_{i,k+1} \sim N(\beta_1, a^2)$ for $i = 1, ..., n$. Hence $\boldsymbol{\beta} = (\beta_1, 1, ..., 1, 0, ..., 0)^T$ with β_1, k ones, and $p - k - 1$ zeros. Binomial regression used $\beta_1 = 0, a = 5/3$, and $m_i = m$ with $m = 1$ or 20. Poisson regression used $\beta_1 = 1 = a$ and $\beta_1 = 5$ with $a = 2$. The simulation for multiple linear regression was similar, but $\beta_1 = 1$ and $\boldsymbol{z}$ was used instead of $\boldsymbol{u}$. The Cox regression simulation changes are described above Table 5. In the tables, $\psi = 0$ means the correlation $\rho = 0$. If $\psi = 0.9$, then $\rho = 0.996$ if $p = 4$ and $\rho = 0.999$ if $p = 10$. In Table 5, if $\psi = 0.5$, then $\rho = 0.857$.

The simulation computed the shorth(c) CI for each β_i and used bootstrap confidence regions to test $H_0 : \boldsymbol{\beta}_S = (\beta_1, 1, ..., 1)^T$ where $\beta_2 = \cdots = \beta_{k+1} = 1$, and $H_0 : \boldsymbol{\beta}_E = \boldsymbol{0}$ (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 would suggest coverage is close to the nominal value. The parametric bootstrap was used with AIC for the GLMs, multiple linear regression used the residual bootstrap with Mallows (1973) C_p, and Cox regression used the nonparametric bootstrap with lasso variable selection.

In Tables 2-5, there are two rows for each model giving the observed confidence interval coverages and average lengths of the confidence intervals. The term "reg,0" is for the full model regression with $\psi = 0$, the term "vs,0.9" is for variable selection with $\psi = 0.9$, and "mix,0" for random selection with $\psi = 0$. The last six columns give results for the tests. The terms pr, hyb, and br are for the prediction region method (4), hybrid region (6), and Bickel and Ren region (5). The 0 indicates the test was $H_0 : \boldsymbol{\beta}_E = \boldsymbol{0}$, while the 1 indicates that the test was $H_0 : \boldsymbol{\beta}_S = (\beta_1, 1..., 1)^T$. 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 for $\boldsymbol{\beta}_S$ if $k = 1$ (if $k = 2$ for Cox regression).

Volume ratios of the three confidence regions can be compared using (7), but there is not enough information in the tables to compare the volume of the confidence region for the full model versus that for the variable selection or random selection since the three methods have different determinants $|\boldsymbol{S}^*_T|$. For random selection, the random vector $\hat{\boldsymbol{\beta}}_{MIX}$ is not observed. Hence for the hybrid region and Bickel and Ren region $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{VS}$ was used, and the coverage for the hybrid region for $\hat{\boldsymbol{\beta}}_{MIX}$ was often 5% too low in the hyb0 and hyb1 columns with $\psi = 0.9$.

The inference for variable selection was often as precise or more precise than the inference for the full model. The coverages tended to be near 0.95 for the bootstrap for the full model. Variable selection coverage tended to be near 0.95 unless the $\hat{\beta}_i$ could equal 0 or if the hybrid region was used with $\hat{\boldsymbol{\beta}}_{MIX}$. An exception was binary logistic regression with $m = 1$ where variable selection and the full model often had higher coverage than the nominal 0.95 for the hypothesis tests, especially for $n = 25p$. For binary regression, the bootstrap confidence regions using smaller a and larger n resulted in coverages closer to 0.95 for the full model, and convergence problems caused the programs to fail for $a > 4$. (The MLE tends to converge if $\max(|\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}|) \leq 7$ and if the Y values of 0 and 1 are not nearly perfectly classified by the rule $\hat{Y} = 1$ if $\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} > 0$ and $\hat{Y} = 0$, otherwise.) The Bickel and Ren (5) average cutoffs were rarely lower than those of the hybrid region (6). For Poisson regression for $\hat{\boldsymbol{\beta}}_{MIX}$ with $p = 10$ and $\psi = 0.9$, the coverages for $H_0 : \boldsymbol{\beta}_S = \mathbf{1}$ were about 4% too low in Table 4. One of the ten shorth confidence intervals also had coverage about 2% too low for this case.

If β_i was a component of β_E, then the variable selection confidence intervals had higher coverage but were shorter than those of the full model due to zero padding. The zeros in $\hat{\boldsymbol{\beta}}_E$ tend to result in higher than nominal coverage for the variable selection estimator, but can greatly decrease the volume of the confidence region compared to that of the full model.

For the simulated data, when $\psi = 0$, the asymptotic covariance matrix, e.g. $\boldsymbol{I}^{-1}(\boldsymbol{\beta})$, is diagonal. Hence $\hat{\boldsymbol{\beta}}_S$ has the same multivariate normal limiting distribution for $\hat{\boldsymbol{\beta}}_{MIX}$ and the full model $\hat{\boldsymbol{\beta}}$, and possibly for $\hat{\boldsymbol{\beta}}_{VS}$, by Section 2. For Tables 2-5, $\boldsymbol{\beta}_S = (\beta_1, \beta_2)^T$, and β_{p-1} and β_p are components of $\boldsymbol{\beta}_E$. For the n in the tables and $\psi = 0$, the coverages and "lengths" did tend to be close for the β_i that are components of $\boldsymbol{\beta}_S$, and for pr1, hyb1, and br1.

Table 2 was for multiple linear regression with forward selection, the residual bootstrap, $n = 100, p = 4, k = 1$, and $B = 1000$. There was slight undercoverage for $\psi = 0$ since n is small for the skewed error distribution. For the full model, and for $\psi = 0$ with $S = \{1, 2\}$, the CI length should be close to $2(1.96)\sigma/10 = 0.392$ when $n = 100$. A larger simulation study, with p as large as 10 and without the MIX rows, is in Pelawa Watagoda and Olive (2021).

Table 2: Bootstrapping OLS Forward Selection with C_p, $e_i \sim EXP(1) - 1$

ψ	β_1	β_2	β_{p-1}	β_p	pr0	hyb0	br0	pr1	hyb1	br1
reg,0	0.939	0.949	0.949	0.944	0.941	0.942	0.942	0.937	0.937	0.938
len	0.393	0.400	0.399	0.400	2.474	2.474	2.475	2.453	2.453	2.455
vs,0	0.938	0.944	0.999	0.997	0.994	0.984	0.995	0.932	0.933	0.934
len	0.392	0.398	0.323	0.323	2.709	2.709	3.014	2.453	2.453	2.460
mix,0	0.937	0.943	0.999	0.998	0.998	0.987	0.995	0.930	0.931	0.931
len	0.391	0.397	0.274	0.278	3.072	3.072	3.288	2.455	2.455	2.459
reg,0.9	0.940	0.948	0.953	0.953	0.946	0.948	0.948	0.933	0.933	0.932
len	0.392	3.249	3.249	3.249	2.475	2.475	2.476	2.454	2.454	2.455
vs,0.9	0.941	0.966	0.996	0.997	0.992	0.983	0.994	0.957	0.953	0.962
len	0.393	2.754	2.721	2.713	2.712	2.712	2.950	2.492	2.492	2.597
mix,0.9	0.941	0.972	0.997	0.998	0.995	0.873	0.996	0.938	0.892	0.935
len	0.391	2.105	1.999	2.000	2.547	2.547	2.828	2.448	2.448	2.610

Tables 3 and 4 are for binary logistic regression and Poisson regression with backward elimination. In Table 3, the coverages for $H_0 : \boldsymbol{\beta}_S = (0, 1)^T$ were a bit high. In Table 4, the coverages for $H_0 : \boldsymbol{\beta}_S = (1, 1)^T$ were low for $\hat{\boldsymbol{\beta}}_{MIX}$ and $\psi = 0.9$.

For Cox proportional hazards regression, the cases were $(Z_i, \delta_i, \boldsymbol{x}_i)$ where $Z_i = Y_i$ is

Table 3: Bootstrapping Binomial Logistic Regression, Backward Elimination with AIC, $B = 200$, $n = 200$, $p = 4$, $k = 1$, and $m = 1$

ψ	β_1	β_2	β_{p-1}	β_p	pr0	hyb0	br0	pr1	hyb1	br1
reg,0	0.950	0.944	0.955	0.954	0.958	0.966	0.967	0.958	0.966	0.973
len	0.754	0.677	0.458	0.459	2.488	2.488	2.499	2.485	2.485	2.575
vs,0	0.954	0.948	0.998	0.998	0.996	0.993	0.997	0.962	0.968	0.976
len	0.750	0.675	0.393	0.390	2.725	2.725	3.031	2.482	2.482	2.575
mix,0	0.956	0.947	0.999	0.998	0.998	0.994	0.998	0.962	0.965	0.970
len	0.740	0.663	0.321	0.322	3.129	3.129	3.341	2.482	2.482	2.548
reg,0.9	0.946	0.954	0.952	0.950	0.955	0.964	0.966	0.950	0.961	0.963
len	0.755	6.084	6.069	6.080	2.489	2.489	2.499	2.486	2.486	2.497
vs,0.9	0.954	0.949	0.996	0.997	0.993	0.991	0.996	0.976	0.980	0.984
len	0.750	5.320	5.385	5.388	2.788	2.788	3.039	2.588	2.588	2.723
mix,0.9	0.955	0.966	0.997	0.997	0.996	0.989	0.996	0.977	0.968	0.974
len	0.741	3.938	3.859	3.865	2.869	2.869	3.046	2.604	2.604	2.707

uncensored if $\delta_i = 1$, and Z_i is right censored if $\delta_i = 0$. We used the nonparametric bootstrap on the cases with lasso variable selection: fit the Cox model on the predictors with nonzero lasso coefficients. *R* code similar to that of Zhou (2001) was used to generate data from the Weibull proportional hazards regression model. The correlations for the predictors were similar to those for the Poisson and binomial regression, but no constant was used so replace q by p. Then $SP = \boldsymbol{x}_i^T\boldsymbol{\beta} = 1x_{i,1} + \cdots + 1x_{i,k} \sim N(0, a^2)$ for $i = 1, ..., n$. The simulations use $a = 1$ where $\boldsymbol{\beta} = (1, ..., 1, 0, ..., 0)^T$ with k ones and $p - k$ zeros. We used $\psi = 0.5$ since $\psi = 0.9$ gave convergence problems. See Table 5.

6. Conclusions

Pelawa Watagoda and Olive (2020) showed that $\hat{\boldsymbol{\beta}}_{VS}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ for several important variable selection estimators for multiple linear regression. This paper extended the theory for several important variable selection estimators for many other regression estimators. The random vector $\hat{\boldsymbol{\beta}}_{MIX}$ has simple large sample theory, and is

Table 4: Bootstrapping Poisson Regression, Backward Elimination with AIC, $B = 500$, $n = 250$, $p = 10$, $k = 1$, $a = 1$, $\beta_1 = 1$

ψ	β_1	β_2	β_{p-1}	β_p	pr0	hyb0	br0	pr1	hyb1	br1
reg,0	0.948	0.953	0.953	0.952	0.951	0.951	0.952	0.943	0.945	0.947
len	0.175	0.133	0.128	0.128	3.986	3.986	3.990	2.453	2.453	2.474
vs,0	0.947	0.952	0.998	0.999	0.998	0.997	0.997	0.950	0.954	0.958
len	0.175	0.132	0.105	0.104	4.303	4.303	4.740	2.452	2.452	2.499
mix,0	0.948	0.953	0.999	1−	1.000	1−	1−	0.951	0.949	0.951
len	0.174	0.129	0.088	0.088	5.122	5.122	5.396	2.453	2.453	2.475
reg,0.9	0.948	0.953	0.957	0.955	0.946	0.948	0.948	0.945	0.945	0.948
len	0.175	3.287	3.286	3.291	3.983	3.983	3.987	2.454	2.454	2.469
vs,0.9	0.948	0.943	0.998	0.999	0.998	0.997	0.999	0.968	0.967	0.973
len	0.175	2.862	2.810	2.820	4.261	4.261	4.685	2.480	2.480	2.638
mix,0.9	0.951	0.921	0.998	0.999	1−	0.999	1−	0.899	0.874	0.898
len	0.174	2.527	2.231	2.228	4.971	4.971	5.287	2.569	2.569	2.711

useful for understanding the more complicated large sample theory of $\hat{\boldsymbol{\beta}}_{VS}$. Theory for the C_p criterion for multiple linear regression with an unknown error distribution was also given. The hybrid confidence region is useful for explaining the three bootstrap confidence regions with Theorem 2, but often has lower coverage than the other two confidence regions.

More theory is needed for the bootstrap confidence regions for variable selection. The method works well when $\pi_d = 1$, which is a very strong assumption, but weaker than the assumption $\pi_S = 1$ that is often made for variable selection consistency and the oracle property. The confidence regions (4) and (5) simulate well for many variable selection estimators, especially when $\boldsymbol{V}$ is diagonal. Augmenting the bootstrap sample for the variable selection estimator with T^* from the full model makes $\boldsymbol{S}_T^*$ nonsingular.

Heuristically, the iid component data clouds are centered at $\boldsymbol{\beta}$ while the bootstrap component data clouds are centered at $\hat{\boldsymbol{\beta}}_{I_j,0}$. Hence the T_i^* tend to be further from $\overline{T}^*$ than the T_i are from $\overline{T}$. Then the bootstrap cutoffs tend to result in conservative tests provided

Table 5: Bootstrapping Cox Regression, Lasso Variable Selection, $B = 200$, $n = 100$, $p = 4$, $k = 2$

ψ	β_1	β_2	β_{p-1}	β_p	pr0	hyb0	br0	pr1	hyb1	br1
reg,0	0.936	0.934	0.952	0.956	0.951	0.962	0.968	0.945	0.965	0.974
len	0.850	0.852	0.744	0.744	2.525	2.525	2.552	2.514	2.514	2.640
vs,0	0.937	0.942	0.989	0.988	0.970	0.974	0.977	0.947	0.966	0.975
len	0.852	0.853	0.728	0.726	2.544	2.544	2.647	2.515	2.515	2.640
mix,0	0.940	0.942	0.992	0.992	0.979	0.978	0.981	0.946	0.966	0.976
len	0.842	0.841	0.667	0.666	2.691	2.691	2.758	2.515	2.515	2.623
reg,0.5	0.945	0.953	0.954	0.951	0.951	0.963	0.965	0.944	0.962	0.968
len	2.372	2.373	2.333	2.338	2.529	2.529	2.556	2.522	2.522	2.571
vs,0.5	0.968	0.960	0.992	0.990	0.980	0.979	0.984	0.976	0.965	0.975
len	2.064	2.064	2.046	2.045	2.784	2.784	2.931	2.558	2.558	2.688
mix,0.5	0.944	0.940	0.993	0.992	0.984	0.984	0.990	0.933	0.944	0.953
len	2.220	2.212	1.971	1.970	2.784	2.784	2.929	2.546	2.546	2.667

$(n - p)/n$ is near 1. The theory, heuristics, and good simulation results suggest that (4) and (5) are useful for exploratory purposes.

There is a massive literature on variable selection and a fairly large literature for inference after variable selection. See, for example, Guan and Tibshirani (2020), Lee and Wu (2018), Leeb and Pötscher (2003), Leeb, Pötscher, and Ewald (2015), Lu et al. (2017), Ning and Liu (2017), Pötscher (1991), and Tibshirani et al. (2018). High dimensional testing has n/p small, and often assumes that n/a_S is large. Ewald and Schneider (2018) note several methods basically use the OLS full model when n/p is large. Rinaldo, Wasserman, and G'Sell (2019) show data splitting is useful and discuss problems with inference after variable selection. Su (2018) shows that fast variable selection methods tend to select spurious variables quickly if $k = a_S$ is not small. Recent papers on large sample theory for multiple linear regression estimators include Cook and Forzani (2018, 2019), Pelawa Watagoda and Olive (2020), and Zhang (2020). Also see Knight and Fu (2000).

See Efron (1979, 1982) for more on the bootstrap. The bagging estimator, $\overline{T}^*$, 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), Schomaker (2012), and Schomaker and Heumann (2014) for standard errors when using the bootstrap or model averaging for linear model confidence intervals.

The simulations were done in *R*. See R Core Team (2016). We used several *R* functions including backward elimination computed with the `step` function from the Venables and Ripley (2010) `MASS` library, forward selection computed with the Lumley (2009) `leaps` function, and lasso computed from the Friedman et al. (2015) `glmnet` library. The collection of Olive (2021) *R* functions *slpack*, available from (http://parker.ad.siu.edu/Olive/slpack.txt), has some useful functions for the inference. The functions `regbootsim3` and `vsbootsim5` were to bootstrap the full model and forward selection for multiple linear regression. The functions `binregbootsim` and `pregbootsim` are useful for the full binomial regression and full Poisson regression models. The functions `vsbrbootsim2` and `vsprbootsim2` were used to bootstrap backward elimination for binomial and Poisson regression. The functions `LRboot` and `vsLRboot2` bootstrap the logistic regression full model and backward elimination. The functions `PRboot` and `vsPRboot2` bootstrap the Poisson regression full model and backward elimination. The function `PHboot` bootstraps the full Cox PH model. The function `PHbootsim` is used to simulate the bootstrap for the full Cox PH model. The function `RLPHboot2` bootstraps a Cox PH model with lasso variable selection. The function `RLPHbootsim2` is used to simulate the bootstrap for lasso variable selection with Cox regression. Sample *R* code is available from (http://parker.ad.siu.edu/Olive/ppRcodebootglm.pdf).

Acknowledgments

The authors thank the referees and Editors for their work.

References

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle.

In *Proceedings, 2nd international symposium on information theory*, ed. B. N. Petrov and F. Csakim, 267-281. Budapest: Akademiai Kiado.

Bickel, P. J., and J. J. Ren. 2001. The bootstrap in hypothesis testing. In *State of the art in probability and statistics: festschrift for William R. van Zwet*, ed. M. de Gunst, C. Klaassen, and A. van der Vaart, 91-112. Hayward, CA: The Institute of Mathematical Statistics.

Breiman, L. 1996. Bagging predictors. *Machine Learning* 24:123-140. doi:10.1023/A:1018054314350.

Buckland S. T., K. P. Burnham, and N. H. Augustin. 1997. Model selection: An integral part of inference. *Biometrics* 53 (2):603-618. doi:10.2307/2533961.

Burr, D. 1994. A comparison of certain bootstrap confidence intervals in the Cox model. *Journal of the American Statistical Association* 89 (42):1290-1302. doi:10.2307/2290992.

Charkhi, A., and G. Claeskens. 2018. Asymptotic post-selection inference for the Akaike information criterion. *Biometrika* 105 (3):645-664. doi:10.1093/biomet/asy018.

Claeskens, G., and N. L. Hjort. 2008. *Model selection and model averaging.* New York, NY: Cambridge University Press.

Cook, R. D., and L. Forzani. 2018. Big data and partial least squares prediction. *The Canadian Journal of Statistics* 46 (1):62-78. doi:10.1002/cjs.11316.

Cook R. D., and L. Forzani. 2019. Partial least squares prediction in high-dimensional regression. *The Annals of Statistics* 47 (2):884-908. doi:10.1214/18-AOS1681.

Cook, R. D., and S. Weisberg. 1999. *Applied regression including computing and graphics.* New York, NY: Wiley.

Cox, D. R. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 34 (2):187-220. doi:10.1111/j.2517-6161.1972.tb00899.x.

Efron, B. 1979. Bootstrap methods, another look at the jackknife. *The Annals of Statistics* 7 (1):1-26. doi:10.1214/aos/1176344552.

Efron, B. 1982. *The jackknife, the bootstrap and other resampling plans.* Philadelphia, PA: SIAM.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani. 2004. Least angle regression (with discussion). *The Annals of Statistics* 32 (2): 407-451. doi:10.1214/009053604000000067.

Efron, B., and R. Tibshirani. 1986. Bootstrap methods for standard errors, confidence intervals, and other methods of statistical accuracy (with discussion). *Statistical Science* 1 (1):54-77. doi:10.1214/ss/1177013815.

Ewald, K., and U. Schneider. 2018. Uniformly valid confidence sets based on the lasso. *Electronic Journal of Statistics* 12 (1):1358-1387. doi:10.1214/18-EJS1425.

Fan, J., and R. Li. 2001. Variable selection via noncave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96 (456):1348-1360. doi:10.1198/016214501753382273.

Freedman, D. A. 1981. Bootstrapping regression models. *The Annals of Statistics* 9 (6):1218-1228. doi:10.1214/aos/1176345638.

Frey, J. 2013. Data-driven nonparametric prediction intervals. *Journal of Statistical Planning and Inference* 143 (6):1039-1048. doi:10.1016/j.jspi.2013.01.004.

Friedman, J., T. Hastie, N. Simon, and R. Tibshirani. 2015. *glmnet*: Lasso and elastic-net regularized generalized linear models. *R* package version 2.0. http://cran.r-project.org/ package=glmnet.

Friedman, J., T. Hastie, and R. Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33 (1):1-22.

Guan, L., and R. Tibshirani. 2020. Post model-fitting exploration via a “next-door” analysis. *The Canadian Journal of Statistics* 48 (3):447-470. doi:10.1002/cjs.

Hall, P. 1988. Theoretical comparisons of bootstrap confidence intervals (with discussion). *The Annals of Statistics* 16 (3):927-985. doi:10.1214/aos/1176350933.

Hastie, T., R. Tibshirani, and M. Wainwright. 2015. *Statistical learning with sparsity: The lasso and generalizations*. Boca Raton, FL: CRC Press Taylor & Francis.

Hillis, S. L., and C. S. Davis. 1994. A simple justification of the iterative fitting procedure for generalized linear models. *The American Statistician* 48 (4):288-289. doi:10.1080/00031305.1994.10476082.

Hjort, G., and N. L. Claeskens. 2003. The focused information criterion. *Journal of the*

American Statistical Association 98 (464):900-945. doi:10.1198/016214503000000819.

Knight, K., and W. J. Fu. 2000. Asymptotics for lasso-type estimators. *The Annals of Statistics* 28 (5):1356–1378. doi:10.1214/aos/1015957397.

Lee, S. M. S., and Y. Wu. 2018. A bootstrap recipe for post-model-selection inference under linear regression models. *Biometrika* 105 (4):873-890. doi:10.1093/biomet/asy046.

Leeb, H., and B. M. Pötscher. 2003. The finite-sample distribution of post-model selection estimators and uniform versus nonuniform approximations. *Econometric Theory* 19 (1):100-142. doi:10.1017/S0266466603191050.

Leeb, H., B. M. Pötscher, and K. Ewald. 2015. On various confidence intervals post-model-selection. *Statistical Science* 30 (2):216-227. doi:10.1214/14-STS507.

Lindenmayer. D. B., R. Cunningham, M. T. Tanton, H. A. Nix, and A. P. Smith. 1991. The conservation of arboreal marsupials in the montane ash forests of central highlands of Victoria, South-East Australia: III. The habitat requirement's of Leadbeater's possum Gymnobelideus Leadbeateri and models of the diversity and abundance of arboreal marsupials. *Biological Conservation* 56 (3):295-315. doi:10.1016/0006-3207(91)90063-F.

Lu, S., Y. Liu, L. Yin, and K. Zhang. 2017. Confidence intervals and regions for the lasso by using stochastic variational inequality techniques in optimization. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79 (2):589-611. doi:10.1111/rssb.12184.

Lumley, T. 2009. *leaps*: Regression subset selection. *R* package version 2.9. https://cran.r-project.org/package=leaps.

Mallows, C. 1973. Some comments on C_p. *Technometrics* 15 (4):661-676. doi:10.2307/1267380.

Meinshausen, N. 2007. Relaxed lasso. *Computational Statistics & Data Analysis* 52 (1):374-393. doi:10.1016/j.csda.2006.12.019.

Nelder, J. A., and R. W. M. Wedderburn. 1972. Generalized linear models. *Journal of the Royal Statistical Society, A* 135 (3):370-380. doi:10.2307/2344614.

Ning, Y., and H. Liu. 2017. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics* 45 (1):158-195. doi:10.1214/16-AOS1448.

Olive, D. J. 2017a. *Linear regression.* New York, NY: Springer.

Olive, D. J. 2017b. *Robust multivariate analysis.* New York, NY: Springer.

Olive, D. J. 2018. Applications of hyperellipsoidal prediction regions. *Statistical Papers* 59 (3):913-931. doi:10.1007/s00362-016-0796-1.

Olive, D. J. 2021. *Prediction and statistical learning.* Online course notes. http://parker.ad.siu.edu/Olive/slearnbk.htm.

Olive, D. J., and D. M. Hawkins. 2005. Variable selection for 1D regression models. *Technometrics* 47 (1):43-50. doi:10.1198/004017004000000590.

Olive, D. J., R. C. Rathnayake, and M. G. Haile. 2021. Prediction intervals for GLMs, GAMs, and some survival regression models. *Communications in Statistics: Theory and Methods* to appear. doi:10.1080/03610926.2021.1887238.

Pelawa Watagoda, L. C. R., and D. J. Olive. 2020. Comparing six shrinkage estimators with large sample theory and asymptotically optimal prediction intervals. *Statistical Papers* to appear. doi:10.1007/s00362-020-01193-1.

Pelawa Watagoda, L. C. R., and D. J. Olive. 2021. Bootstrapping multiple linear regression after variable selection. *Statistical Papers* 62 (2):681-700. doi:10.1007/s00362-019-01108-9.

Pötscher, B. 1991. Effects of model selection on inference. *Econometric Theory* 7 (2):163-185. doi:10.1017/S0266466600004382.

Pratt, J. W. 1959. On a general concept of "in Probability". *The Annals of Mathematical Statistics* 30 (2):549-558. doi:10.1214/aoms/1177706267.

R Core Team. 2018. *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing. www.R-project.org.

Rinaldo, A., L. Wasserman, and M. G'Sell. 2019. Bootstrapping and sample splitting for high-dimensional, assumption-free inference. *The Annals of Statistics* 47 (6):3438-3469. doi:10.1214/18-AOS1784.

Schomaker, M. 2012. Shrinkage averaging estimation. *Statistical Papers* 53:1015-1034. doi:10.1007/s00362-011-0405-2.

Schomaker, M., and C. Heumann. 2014. Model selection and model averaging after multiple imputation. *Computational Statistics & Data Analysis* 71:758-770. doi:10.1016/j.csda.2013.02.017.

Schwarz, G. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6 (2):461-464. doi:10.1214/aos/1176344136.

Sen, P. K., and J. M. Singer. 1993. *Large sample methods in statistics: An introduction with applications*. New York, NY: Chapman & Hall.

Shao, J. 1993. Linear model selection by cross-validation. *Journal of the American Statistical Association* 88 (422):486-494. doi:10.1080/01621459.1993.10476299.

Shao, J., and D. S. Tu. 1995. The jackknife and the bootstrap. New York, NY: Springer.

Simon, N., J. Friedman, T. Hastie, and R. Tibshirani. 2011. Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software* 39 (5):1-13. doi:10.18637/jss.v039.i05.

Su, W. J. 2018. When is the first spurious variable selected by sequential regression procedures? *Biometrika* 105 (3):517-527. doi:10.1093/biomet/asy032.

Su, Z., and R. D. Cook. 2012. Inner envelopes: efficient estimation in multivariate linear regression. *Biometrika* 99 (3):687-702. doi:10.1093/biomet/ass024.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 58 (1):267-288. doi:10.1111/j.2517-6161.1996.tb02080.x.

Tibshirani, R. J., A. Rinaldo, R. Tibshirani, and L. Wasserman. 2018. Uniform asymptotic inference and the bootstrap after model selection. *The Annals of Statistics* 46 (3):1255-1287. doi:10.1214/17-AOS1584.

Venables, W. N., and B. D. Ripley. 2010. *Modern applied statistics with S*. 4th ed. New York, NY: Springer.

Wang. H., and S. Z. F. Zhou. 2013. Interval estimation by frequentist model averaging. *Communications in Statistics: Theory and Methods* 42: (23):4342-4356. doi:10.1080/03610926.2011.647218.

Wieczorek, J. A. 2018. Model selection and stopping rules for high-dimensional forward selection. Ph.D. Thesis, Carnegie Mellon University.

Yang, Y. 2003. Regression with multiple candidate models: selecting or mixing? *Statistica Sinica* 13 (3):783-809.

Zhang, J. 2020. Consistency of MLE, LSE and M-estimation under mild conditions. *Statistical Papers* 61:189-199. doi:10.1007/s00362-017-0928-2.

Zhao, P., and B. Yu. 2006. On model selection consistency of lasso. *Journal of Machine Learning Research* 7:2541-2563.

Zhou, M. 2001. Understanding the Cox regression models with time–change covariates. *The American Statistician* 55 (2):153-155. doi:10.1198/000313001750358491.

Zou, H., and T. Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2):301-320. doi:10.1111/j.1467-9868.2005.00503.x.