BOOTSTRAPPING SOME GLM AND SURVIVAL REGRESSION VARIABLE SELECTION ESTIMATORS

BY RASANJI C. RATHNAYAKE AND DAVID J. OLIVE

Southern Illinois University

Consider a regression model where the response variable $Y$ depends on the $p \times 1$ vector of predictors $x$ only through $x^T \beta$. We show that some common variable selection estimators $\hat{\beta}_{VS}$, such as forward selection with AIC, are $\sqrt{n}$ consistent, and also develop a related random variable $\hat{\beta}_{MIX}$ that has simple large sample theory. Then a method is suggested for bootstrapping many of these variable selection estimators, and hypothesis testing is done using three confidence regions. Theory for $\hat{\beta}_{MIX}$ suggests that two of the three confidence regions tend to have coverage at least as high as the nominal coverage if the sample size is large enough.

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 $x$ given $x^T \beta$, written $Y \perp \!\!\!\!\perp x | x^T \beta$. Many important regression models satisfy this condition, including multiple linear regression, generalized linear models (GLMs), negative binomial regression, beta–binomial regression, the Cox (1972) proportional hazards regression model, and some other survival regression models such as Weibull regression and accelerated failure time models. Also see Nelder and Wedderburn (1972), Olive (2017a), and Yee (2015). Forward selection or backward elimination with the Akaike (1973) AIC criterion or Schwarz (1978) BIC criterion are often used for variable selection for regression models.

Some shrinkage methods do variable selection: the regression method, such as a GLM or Cox (1972) proportional hazards regression, uses the predictors that had nonzero coefficients. These methods include least angle regression, lasso, relaxed lasso, and elastic net. See, for example, Fan and Li (2001), Friedman et al. (2007), Friedman, Hastie, and Tibshirani (2010), Hastie, Tibshirani, and Wainwright (2015, ch. 5), Simon et al. (2011), Sun and Zhang (2012), Tibshirani (1996), and Zou and Hastie (2005). The Meinshausen (2007) relaxed lasso estimator for multiple linear regression fits lasso
with penalty $\lambda_n$ to get a subset of variables with nonzero coefficients, and then fits lasso with a smaller penalty $\phi_n$ to this subset of variables where $n$ is the sample size. This two stage procedure could be used for other estimators such as Poisson regression, binomial regression, and Cox proportional hazards regression. A three stage procedure uses this relaxed lasso estimator for variable selection.

These variable selection estimators have had several names in the literature. For multiple linear regression with ordinary least squares (OLS), 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).

Two important quantities for a regression model are the sufficient predictor $SP = x^T\beta$, and the estimated sufficient predictor $ESP = \hat{x}^T\hat{\beta}$. For the regression models, the conditioning and subscripts, such as $i$, will often be suppressed. The multiple linear regression model is $Y|x = x^T\beta + e$ or $Y_i = x_i^T\beta + e_i$ for $i = 1, ..., n$. Consider a parametric regression model $Y|x \sim D(x^T\beta, \gamma)$ where $D$ is a parametric distribution that depends on the $p \times 1$ vector of predictors $x$ only through $x^T\beta$, and $\gamma$ is a $q \times 1$ vector of parameters. Three examples used in the simulations follow. The binomial logistic regression model is $Y_i \sim \text{binomial}(m_i, \rho(SP) = \exp(SP))$. Then $E(Y_i|SP) = m_i\rho(SP)$, $V(Y_i|SP) = m_i\rho(SP)(1 - \rho(SP))$, and $\hat{E}(Y_i|SP) = m_i\hat{\rho} = m_i\rho(ESP)$ is the estimated mean function. The binary logistic regression model has $m_i \equiv 1$ for $i = 1, ..., n$. A useful Poisson regression model $Y \sim \text{Poisson}(\exp(SP))$ has $E(Y|SP) = V(Y|SP) = \exp(SP)$. The estimated mean and variance functions are $\hat{E}(Y|SP) = \exp(ESP) = \hat{V}(Y|SP)$. The Weibull proportional hazards regression model is $Y|SP \sim W(\gamma = 1/\sigma, \lambda_0 \exp(SP))$

where $\lambda_0 = \exp(-\alpha/\sigma)$, and $Y$ has a Weibull $W(\gamma, \lambda)$ distribution if the probability density function of $Y$ is $f(y) = \lambda\gamma y^{\gamma-1} \exp[-\lambda y^\gamma]$ for $y > 0$.

This model is useful for simulating the Cox proportional hazards regression model.

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 where the number
of predictors $p$ is fixed. Following Olive and Hawkins (2005), a model for variable selection can be described by

\begin{equation}
\mathbf{x}^T \beta = \mathbf{x}_S^T \beta_S + \mathbf{x}_E^T \beta_E = \mathbf{x}_S^T \beta_S
\end{equation}

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

\[ \mathbf{x}_I^T \beta = \mathbf{x}_S^T \beta_S + \mathbf{x}_I^T \beta_I = \mathbf{x}_S^T \beta_S \]

where $\mathbf{x}_{I/S}$ denotes the predictors in $I$ that are not in $S$. Since this is true regardless of the values of the predictors, $\beta_O = \mathbf{0}$ if $S \subseteq I$. The model using $\mathbf{x}_I^T \beta$ is the full model.

To clarify notation, suppose $p = 4$, a constant $x_1 = 1$ corresponding to $\beta_1$ is always in the model, and $\beta = (\beta_1, \beta_2, 0, 0)^T$. Then the $J = 2^{p-1} = 8$ possible subsets of $\{1, 2, ..., p\}$ that always contain 1 are $I_1 = \{1\}$, $S = I_2 = \{1, 2\}$, $I_3 = \{1, 3\}$, $I_4 = \{1, 4\}$, $I_5 = \{1, 2, 3\}$, $I_6 = \{1, 2, 4\}$, $I_7 = \{1, 3, 4\}$, and $I_8 = \{1, 2, 3, 4\}$. There are $2^{p-a_S} = 4$ subsets $I_2, I_5, I_6$, and $I_8$ such that $S \subseteq I_2$. Also, $\beta_I = (\hat{\beta}_1, \hat{\beta}_3, \hat{\beta}_7)^T$ is obtained by regressing $Y$ on $\mathbf{x}_I$, $\mathbf{x}_I = (x_1, x_3, x_4)^T$.

Let $I_{\min}$ correspond to the set of predictors selected by a variable selection method such as forward selection or lasso variable selection. If $\hat{\beta}_I$ is a $1 \times 1$, form the $p \times 1$ vector $\hat{\beta}_{I,0}$ from $\hat{\beta}_I$ by adding 0s corresponding to the omitted variables. Also use zero padding for the model $I_{\min}$. For example, if $p = 4$ and $\hat{\beta}_{I_{\min}} = (\hat{\beta}_1, \hat{\beta}_3)^T$, then the observed $\beta_{VS} = \hat{\beta}_{I_{\min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$. As a statistic, $\hat{\beta}_{VS} = \hat{\beta}_{I_k,0}$ with probabilities $\pi_{kn} = P(I_{\min} = I_k)$ for $k = 1, ..., J$ where there are $J$ subsets. For example, if each subset contains at least one variable, then there are $J = 2^p - 1$ subsets.

Let $\hat{\beta}_{MIX}$ be a random vector with a mixture distribution of the $\hat{\beta}_{I_k,0}$ with probabilities equal to $\pi_{kn}$. Hence $\hat{\beta}_{MIX} = \hat{\beta}_{I_k,0}$ with same probabilities $\pi_{kn}$ of the variable selection estimator $\hat{\beta}_{VS}$, but the $I_k$ are randomly selected. A random vector $\mathbf{u}$ has a mixture distribution of random vectors $\mathbf{u}_j$ with probabilities $\pi_j$ if $\mathbf{u}$ equals the randomly selected random vector $\mathbf{u}_j$ with probability $\pi_j$ for $j = 1, ..., J$. Let $\mathbf{u}$ and $\mathbf{u}_j$ be $p \times 1$ random vectors. Then the cumulative distribution function (cdf) of $\mathbf{u}$ is

\[ F_{\mathbf{u}}(t) = \sum_{j=1}^{J} \pi_j F_{\mathbf{u}_j}(t) \]
where the probabilities $\pi_j$ satisfy $0 \leq \pi_j \leq 1$ and $\sum_{j=1}^{J} \pi_j = 1$, $J \geq 2$, and $F_{u_j}(t)$ is the cdf of $u_j$. Suppose $E(h(u))$ and the $E(h(u_j))$ exist. Then

$$E(h(u)) = \sum_{j=1}^{J} \pi_j E[h(u_j)]$$

and

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

If $E(u_j) = \theta$ for $j = 1, \ldots, J$, then $E(u) = \theta$ and

$$\text{Cov}(u) = \sum_{j=1}^{J} \pi_j \text{Cov}(u_j).$$

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. 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 $Z_{(1)}, \ldots, Z_{(n)}$ be the order statistics of $n$ independent and identically distributed (iid) random variables $Z_1, \ldots, Z_n$. Let a future random variable $Z_f$ be such that $Z_1, \ldots, Z_n, Z_f$ 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 $\theta$ such as $T_n = \bar{Z} = \frac{1}{n} \sum_{i=1}^{n} Z_i$ with $\theta = E(Z_1)$. Let $T_1^*, \ldots, T_B^*$ be a bootstrap sample for $T_n$. The bootstrap percentile method large sample 100(1 - $\delta$)% confidence interval $[T_{(k_1)}^*, T_{(k_2)}^*]$ for $\theta$ contains $\approx [B(1 - \delta)]$ of the $T_i^*$. A common choice is to apply the above PI on the bootstrap sample to get $[T_{(k_1)}^*, T_{(k_2)}^*]$ where the $k_i$ are as above with $B$ used instead of $n$. See Efron (1982, p. 78).

Let the shortest closed interval containing at least $c$ of the $Z_i$ be the shorth($c$) estimator. Frey (2013) showed that for large $n\delta$ and iid data, the shorth($kn$ = $n(1 - \delta)$) prediction interval has maximum undercoverage $\approx 1.12\sqrt{\delta/n}$, and used the large sample 100(1 - $\delta$)% PI shorth($c$) =

$$[Z_{(s)}, Z_{(s+c-1)}] \quad \text{with} \quad c = \min(n, \lceil n[1 - \delta + 1.12\sqrt{\delta/n}] \rceil).$$

Apply the shorth estimator to a bootstrap sample $\hat{\beta}_{i1}^*, \ldots, \hat{\beta}_{iB}^*$ to get a confidence interval for $\theta = \beta_i$ where $T_n = \hat{\beta}_i$. Then (2) can be regarded as the
shortest percentile method confidence interval, asymptotically. Hence the shortest confidence interval is a practical implementation of the Hall (1988) shortest bootstrap interval based on all possible bootstrap samples.

Consider testing $H_0: \theta = \theta_0$ versus $H_1: \theta \neq \theta_0$ where $\theta_0$ is a known $g \times 1$ vector. We will form a large sample confidence region and reject $H_0$ if $\theta_0$ is not in the confidence region. A large sample $100(1 - \delta)\%$ confidence region for $\theta$ is a set $A_n$ such that $P(\theta \in A_n)$ is eventually bounded below by $1 - \delta$ as the sample size $n \to \infty$.

For a confidence region, let the $g \times 1$ vector $T_n$ be an estimator of the $g \times 1$ parameter vector $\theta$. Let $T_1^*, ..., T_B^*$ be the bootstrap sample for $T_n$. Let $A$ be a full rank $g \times p$ constant matrix. For variable selection, consider testing $H_0: A\beta = \theta_0$ versus $H_1: A\beta \neq \theta_0$ with $\theta = A\beta$ where often $\theta_0 = 0$. Then let $T_n = A\beta_{SEL}$ and let $T_i^* = A\beta_{SEL}^*$ for $i = 1, ..., B$ and $SEL$ is $VS$ or $MIX$.

The bootstrap confidence regions, given below, use Mahalanobis distances. Let the $g \times 1$ column vector $T$ be a multivariate location estimator, and let the $g \times g$ symmetric positive definite matrix $C$ be a dispersion estimator. Then the $i$th squared sample Mahalanobis distance is the scalar

$$D_i^2 = D_i^2(T, C) = D_i^2(T, C) = (z_i - T)^T C^{-1} (z_i - T)$$

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

$$\bar{z} = \frac{1}{B} \sum_{i=1}^{B} z_i \quad \text{and} \quad S = \frac{1}{B - 1} \sum_{i=1}^{B} (z_i - \bar{z})(z_i - \bar{z})^T.$$ 

The first confidence region applies a prediction region to the bootstrap sample $T_1^*, ..., T_B^*$ where the statistic of center $T$ is an estimator of $\theta$. For iid data from a distribution with a $g \times g$ nonsingular covariance matrix, it was found that the simulated maximum undercoverage was about 0.05 when $n = 20g$. Hence a correction factor was used to give better coverage for small $n$. When applied to a bootstrap sample of size $B$, the correction factor gives better coverage when $B \geq 50g$. This result is useful because the bootstrap confidence regions can be slow to simulate. Hence we want to use small values of $B \geq 50g$. Let $q_B = \min(1 - \delta + 0.05, 1 - \delta + g/B)$ for $\delta > 0.1$ and

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

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 quantile of the $D_i$. If $B$ is large enough, $D(U_B)$ is the
100(1 - \delta)th quantile. Let \( \mathbf{T}' \) and \( S_T^* \) be the sample mean and sample covariance matrix of the bootstrap sample.

The Olive (2017ab, 2018) prediction region method large sample 100(1 - \delta)\% confidence region for \( \theta \) is \( \{ \mathbf{w} : (\mathbf{w} - \mathbf{T}')^T [S_T^*]^{-1} (\mathbf{w} - \mathbf{T}') \leq D_{(U_B)}^2 \} \)

\[ \text{(4)} \]
\[ = \{ \mathbf{w} : D_{\mathbf{w}}^2 (\mathbf{T}', S_T^*) \leq D_{(U_B)}^2 \} \]

where \( D_{(U_B)}^2 \) is computed from \( D_{i}^2 = (T_{i}^* - \mathbf{T}')^T[S_T^*]^{-1}(T_{i}^* - \mathbf{T}') \) for \( i = 1, \ldots, B \). Note that the corresponding test for \( H_0 : \theta = \theta_0 \) rejects \( H_0 \) if \( (\mathbf{T}' - \theta_0)^T[S_T^*]^{-1}(\mathbf{T}' - \theta_0) > D_{(U_B)}^2 \). Note that \( q_B \) is found from (3) with \( z_i = T_{i}^* \). If \( g = 1 \), (4) is a closed interval centered at \( \mathbf{T}' \) just long enough to cover \( U_B \) of the \( T_i^* \). This confidence interval is similar to that of Efron (2014).

The modified Bickel and Ren (2001) large sample 100(1 - \delta)\% confidence region is \( \{ \mathbf{w} : (\mathbf{w} - T_n)^T[S_T^*]^{-1}(\mathbf{w} - T_n) \leq D_{(U_B,T)}^2 \} \)

\[ \text{(5)} \]
\[ = \{ \mathbf{w} : D_{\mathbf{w}}^2 (T_n, S_T^*) \leq D_{(U_B,T)}^2 \} \]

where the cutoff \( D_{(U_B,T)}^2 \) is the 100\% sample quantile of the \( D_{i}^2 = (T_{i}^* - T_n)^T[S_T^*]^{-1}(T_{i}^* - T_n) \). Shift region (4) to have center \( T_n \), or equivalently, change the cutoff of region (5) to \( D_{(U_B,T)}^2 \) to get the Pelawa Watagoda and Olive (2019a) hybrid large sample 100(1 - \delta)\% confidence region: \( \{ \mathbf{w} : (\mathbf{w} - T_n)^T[S_T^*]^{-1}(\mathbf{w} - T_n) \leq D_{(U_B)}^2 \} \)

\[ \text{(6)} \]
\[ = \{ \mathbf{w} : D_{\mathbf{w}}^2 (T_n, S_T^*) \leq D_{(U_B)}^2 \} \].

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 (2019a) gave simpler proofs, and proved that the shorth(c) interval applied to a bootstrap sample of a random variable gives a large sample confidence interval. If \( g = 1 \), if \( \sqrt{n}(T_n - \theta) \overset{D}{\to} U \), and if \( \sqrt{n}(T_{i}^* - T_n) \overset{D}{\to} U \) where \( U \) has a unimodal probability density function symmetric about zero, then the confidence intervals from the three confidence regions (4)-(6), the shorth confidence interval (2), and the “usual” percentile method confidence interval are asymptotically equivalent (use the central proportion of the bootstrap sample, asymptotically).

Note that if (5) is a large sample confidence regions, then so are (4) and (6) is \( \sqrt{n}(\mathbf{T}' - T_n) \overset{D}{\to} \mathbf{0} \) as \( n \to \infty \). Pelawa Watagoda and Olive (2019a) showed that this condition holds if \( \sqrt{n}(T_n - \theta) \overset{D}{\to} \mathbf{u} \) and \( \sqrt{n}(T_{i}^* - T_n) \overset{D}{\to} \mathbf{u} \) where \( E(\mathbf{u}) = \mathbf{0} \) and \( \text{Cov}(\mathbf{u}) = \Sigma_{\mathbf{u}} \neq \mathbf{0} \). Thus \( \sqrt{n}(\mathbf{T}' - \theta) \overset{D}{\to} \mathbf{u} \) and
\sqrt{n}(T_i^* - \mathbf{T}^*) \overset{D}{\rightarrow} \mathbf{u}. In addition, assume \( nS_T^* \overset{P}{\rightarrow} \mathbf{C} \) where \( \mathbf{C} \) is nonsingular. Let
\[
D_1^2 = D_{1,T_i^*}^2(\mathbf{T}^*, S_T^*) = \sqrt{n}(T_i^* - \mathbf{T}^*)^T (nS_T^*)^{-1}\sqrt{n}(T_i^* - \mathbf{T}^*),
\]
\[
D_2^2 = D_{2,T}^2(T_n, S_T^*) = \sqrt{n}(T_n - \theta)^T (nS_T^*)^{-1}\sqrt{n}(T_n - \theta),
\]
\[
D_3^2 = D_{3,\theta}^2(\mathbf{T}^*, S_T^*) = \sqrt{n}(\mathbf{T}^* - \theta)^T (nS_T^*)^{-1}\sqrt{n}(\mathbf{T}^* - \theta), \quad \text{and}
\]
\[
D_4^2 = D_{4,T}^2(T_n, S_T^*) = \sqrt{n}(T_i^* - T_n)^T (nS_T^*)^{-1}\sqrt{n}(T_i^* - T_n).
\]
Then \( D_j^2 \approx \mathbf{u}^T (nS_T^*)^{-1} \mathbf{u} \approx \mathbf{u}^T \mathbf{C}^{-1} \mathbf{u} \), and the percentiles of \( D_1^2 \) and \( D_2^2 \) can be used as cutoffs. If \( n \) and \( B \) are large enough and \((nS_T^*)^{-1}\) is “not too ill conditioned,” then the confidence regions (4), (5), and (6) should still have coverage near 1 − \( \delta \). The regularity conditions for (4)–(6) are weaker when \( g = 1 \), since \( S_T^* \) does not need to be computed.

The ratio of the volumes of regions (4) and (5) is
\[
(7) \quad \frac{|S_T^*|^{1/2}}{|S_T^*|^{1/2}} \left( \frac{D_{(U_B)}}{D_{(U_B,T)}} \right)^g = \left( \frac{D_{(U_B)}}{D_{(U_B,T)}} \right)^g.
\]

The volume of confidence region (5) tends to be greater than that of (4) since the \( T_i^* \) are closer to \( \mathbf{T}^* \) than \( T_n \) on average. Regions (4) and (6) have the same volume.

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

2. Large sample theory for variable selection estimators. The theory for \( \hat{\beta}_{MIX} \) in this section applies to many regression models including many GLMs, some time series models, some survival regression models such as the Cox (1972) proportional hazards survival regression model, and the multiple linear regression model where the error distribution is unknown. Suppose the regression model satisfies \( Y \overset{\text{i.i.d.}}{\sim} \mathbf{x}^T \mathbf{\beta} \), that model (1) holds, and that if \( S \subseteq I_j \) where the dimension of \( I_j \) is \( a_j \), then \( \sqrt{n}(\hat{\beta}_{I_j} - \beta_{I_j}) \overset{D}{\rightarrow} N_{a_j}(\mathbf{0}, \mathbf{V}_j) \) where \( \mathbf{V}_j \) is the covariance matrix of the asymptotic multivariate normal distribution. Then
\[
(8) \quad \sqrt{n}(\hat{\beta}_{I_j,0} - \beta) \overset{D}{\rightarrow} N_{a_j}(\mathbf{0}, \mathbf{V}_{j,0})
\]
where \( \mathbf{V}_{j,0} \) adds columns and rows of zeros corresponding to the \( x_i \) not in \( I_j \), and \( \mathbf{V}_{j,0} \) is singular unless \( I_j \) corresponds to the full model.

The following theorem generalizes the Pelawa Watagoda and Olive (2019a) theorem for multiple linear regression. The first assumption in Theorem 1
is \( P(S \subseteq I_{\min}) \to 1 \) as \( n \to \infty \). This assumption means that 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. See Charkhi and Claeskens (2018) and Claeskens and Hjort (2008, pp. 70, 101, 102, 114, 232). For multiple linear regression with Mallows (1973) \( C_p \) or AIC, see Li (1987), Nishii (1984), and Shao (1993). Let \( \hat{\beta}_{I_{\min}} \) be the regression estimator, such as a GLM, applied to a constant and the variables with nonzero shrinkage estimator coefficients. If the shrinkage estimator is a consistent estimator of \( \beta \), then \( P(S \subseteq I_{\min}) \to 1 \) as \( n \to \infty \). The assumption on \( u_{jn} \) in Theorem 1 is reasonable by (8) since \( S \subseteq I_j \) for each \( \pi_j \), and since \( \hat{\beta}_{\text{MIX}} \) uses random selection.

**Theorem 1.** Assume \( P(S \subseteq I_{\min}) \to 1 \) as \( n \to \infty \), and let \( \hat{\beta}_{\text{MIX}} = \hat{\beta}_{I_{\min},0} \) with probabilities \( \pi_{kn} \) where \( \pi_{kn} \to \pi_k \) as \( n \to \infty \). Denote the positive \( \pi_k \) by \( \pi_j \). Assume \( u_{jn} = \sqrt{n}(\hat{\beta}_{j,0} - \beta) \overset{D}{\to} u_j \sim N_p(0, V_j,0) \). a) Then

\[
(9) \quad u_n = \sqrt{n}(\hat{\beta}_{\text{MIX}} - \beta) \overset{D}{\to} u
\]

where the cdf of \( u \) is \( F_u(t) = \sum_j \pi_j F_{u,j}(t) \). Thus \( u \) is a mixture distribution of the \( u_j \) with probabilities \( \pi_j \), \( E(u) = 0 \), and \( \text{Cov}(u) = \Sigma_u = \sum_j \pi_j V_j,0 \).

b) Let \( A \) be a \( g \times p \) full rank matrix with \( 1 \leq g \leq p \). Then

\[
(10) \quad v_n = Au_n = \sqrt{n}(A\hat{\beta}_{\text{MIX}} - A\beta) \overset{D}{\to} Au = v
\]

where \( v \) has a mixture distribution of the \( v_j = Au_j \sim N_g(0, AV_j,0A^T) \) with probabilities \( \pi_j \).

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

d) If \( \pi_d = 1 \), then \( \sqrt{n}(\hat{\beta}_{\text{SEL}} - \beta) \overset{D}{\to} u \sim N_p(0, \Sigma_{d,0}) \) where \( \text{SEL} \) is \( \text{VS} \) or \( \text{MIX} \).

**Proof.** a) Since \( u_n \) has a mixture distribution of the \( u_{kn} \) with probabilities \( \pi_{kn} \), the cdf of \( u_n \) is \( F_{u_n}(t) = \sum_k \pi_{kn} F_{u_{kn}}(t) \to F_u(t) = \sum_j \pi_j F_{u,j}(z) \) at continuity points of the \( F_{u,j}(t) \) as \( n \to \infty \).

b) Since \( u_n \overset{D}{\to} u \), then \( Au_n \overset{D}{\to} Au \).

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

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 $\pi_j$, $A\mathbf{u}_j \sim N_g(0, A\Sigma_j A^T) = N_g(0, A\Sigma A^T)$. Then this result also holds for $\hat{\beta}_{VS}$ since asymptotically, $\sqrt{n}(A\hat{\beta}_{VS} - A\beta) \overset{D}{\to} N_g(0, A\Sigma A^T)$. Then this result also holds for $\hat{\beta}_{SEL}$ since asymptotically, $\sqrt{n}(A\hat{\beta}_{SEL} - A\beta)$ is selecting from the $A\mathbf{u}_j$ which have the same distribution. Also see Theorem 3. This special case occurs for $\hat{\beta}_{S, MIX}$ if $\sqrt{n}(\hat{\beta} - \beta) \overset{D}{\to} N_p(0, V)$ where the asymptotic covariance matrix $V$ is diagonal and nonsingular. Then $\hat{\beta}_{S, MIX}$, $\hat{\beta}_{S, VS}$, and $\hat{\beta}_{S, FULL}$ have the same multivariate normal limiting distribution. Hence the prediction regions applied to $A\hat{\beta}^*_SEL = B\hat{\beta}^*_S, SEL$ will have similar volume and cutoffs where $SEL$ is MIX, VS, or FULL.

Let the multiple linear regression model in matrix form be $Y = X\beta + \epsilon$. Then the large sample theory for lasso and elastic net given by Knight and Fu (2000) and Slawski, zu Castell, and Tutz (2010) is summarized by Pelawa Watagoda and Olive (2019b). If $\lambda_n/\sqrt{n} \to 0$ as $n \to \infty$, then the two estimators are asymptotically equivalent to the full OLS model, but the full OLS model large sample theory is better than that of lasso and elastic net otherwise if $X^TX$ is well behaved. Lasso and elastic net do variable selection better if $\lambda_n/\sqrt{n} \to \infty$ as $n \to \infty$ than if $\lambda_n/\sqrt{n} \to 0$, as noted by Fan and Li (2001), and are consistent if $\lambda_n/n \to 0$ as $n \to \infty$. The large sample theory uses $\sqrt{n}(\hat{\beta}_Q - \beta)$ where $Q$ stands for lasso or elastic net. Hence the coefficients estimated to be 0 are not discarded. We can modify lasso or elastic net that used $\lambda_n$ to use $\hat{\lambda}_n = \min(\lambda_n, 100n/(\log(n)))$ to insure consistency. Then Theorem 1 applies to $\hat{\beta}_{MIX}$ using the $\pi_{kn}$ from lasso variable selection or elastic net variable selection. By Theorem 1(c), the lasso variable selection and elastic net variable selection estimators are $\sqrt{n}$ consistent if lasso and elastic net are consistent.

Theorem 1 has several other applications. First, the theory gives the asymptotic distribution of $\hat{\beta}_{MIX}$ corresponding to many variable selection estimators. Second, the theory is useful for explaining why $\hat{\beta}_{I_{min}}$ should not be used, but $\hat{\beta}_{I_{min, 0}}$ is a good estimator. For a random quantity to be a $k \times 1$ random vector, the dimension of the random quantity needs to be $k$ (with probability one). Since the dimension of $\beta_{I_{min}}$ is a random variable, the random quantity $\hat{\beta}_{I_{min}}$ is neither a random vector nor a statistic. Then $\sqrt{n}(\hat{\beta}_{I_{min}} - \beta_{I_{j}})$
can not be used as an asymptotic pivot even if \( I_{min} = I_j \) is observed. Compare Leeb and Pötscher (2006). A third application is bootstrap inference for hypothesis testing, as discussed in Section 3 and by Pelawa Watagoda and Olive (2019a) for multiple linear regression. Fourth, the theory can be used to justify prediction intervals after variable selection. See Pelawa Watagoda and Olive (2019b) and Olive, Rathnayake, and Haile (2020).

Fifth, there is a large literature on 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) 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 only under very strong regularity conditions for fast methods such as forward selection, backward elimination, and lasso. 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. If \( \pi_S = 1 \), then

\[
\sqrt{n}(\hat{\beta}_{I_{min}, 0} - \beta) \xrightarrow{D} N_p(0, V_{S, 0}),
\]

a singular quantity distribution unless \( S \) is the full model. As discussed above, the random quantity \( \hat{\beta}_{I_{min}} \) is not a statistic, and \( \sqrt{n}(\hat{\beta}_{I_{min}} - \beta_S) \) is not well defined and does not converge in distribution. If \( S \) is known before looking at the data, there would be no need to do variable selection, but

\[
\sqrt{n}(\hat{\beta}_{S, I_{min}, 0} - \beta_S) \xrightarrow{D} N_{as}(0, V_S) \quad \text{and} \quad \sqrt{n}(\hat{\beta}_S - \beta_S) \xrightarrow{D} N_{as}(0, V_S).
\]

The following Pelawa Watagoda and Olive (2019a) theorem is useful for bootstrapping variable selection estimators. Let \((\hat{T}, S_T)\) be the sample mean and sample covariance matrix computed from \(T_1, \ldots, T_B\) which have the same distribution as \(T_n\). Let \(D^2_{(UB)}\) be the cutoff computed from the \(D^2_T(\hat{T}, S_T)\) for \(i = 1, \ldots, B\). Note that \(T_i = T_n\). The hyperellipsoid corresponding to the squared Mahalanobis distance \(D^2(T_n, C)\) is centered at \(T_n\), while the hyperellipsoid corresponding to \(D^2(\hat{T}, C)\) is centered at \(\hat{T}\). Note that \(D^2_T(T_n, C) = (\hat{T} - T_n)^T C^{-1} (\hat{T} - T_n) = (T_n - T)^T C^{-1} (T_n - T) = D^2_{T_n}(\hat{T}, C)\). Thus \(D^2_T(T_n, C) \leq D^2_{(UB)}\) iff \(D^2_{(UB)}(\hat{T}, C) \leq D^2_{(UB)}\). In Theorem 2, since \(R_p\) contains \(T_i\) with probability \(1 - \delta_B\), the region \(R_c\) contains \(\hat{T}\) with probability \(1 - \delta_B\). Since \(T_n\) depends on the sample size \(n\), we need \((nS_T)^{-1}\) to be fairly well behaved (“not too ill conditioned”) for each \(n \geq 20g\), say. This condition is weaker than the stated assumption \((nS_T)^{-1} \xrightarrow{P} \Sigma_A^{-1}\) where \(\Sigma_A\) is some nonsingular matrix. Often \(\Sigma_A = \lim_{n \to \infty} n\Sigma_{T_n}\).

**Theorem 2: Geometric Argument.** Suppose \(\sqrt{n}(T_n - \theta) \xrightarrow{D} U\) with \(E(U) = 0\) and \(\text{Cov}(U) = \Sigma U \neq 0\). Assume \(T_1, \ldots, T_B\) are iid with nonsin-
gular covariance matrix $\Sigma_T$, where $(nS_T)^{-1} \xrightarrow{P} \Sigma^{-1}$. Then the large sample 100$(1 - \delta)$% prediction region $R_p = \{ w : D^2 w(T, S_T) \leq D^2(U_B) \}$ centered at $T$ contains a future value of the statistic $T_f$ with probability $1 - \delta_B \rightarrow 1 - \delta$ as $B \rightarrow \infty$. Hence the region $R_c = \{ w : D^2 w(T_n, S_T) \leq D^2(U_B) \}$ is a large sample 100$(1 - \delta)$% confidence region for $\theta$ where $T_n$ is a randomly selected $T_i$.

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

$$\begin{bmatrix} \sqrt{n}(T_1 - \theta) \\ \vdots \\ \sqrt{n}(T_B - \theta) \end{bmatrix} \xrightarrow{D} \begin{bmatrix} v_1 \\ \vdots \\ v_B \end{bmatrix}$$

where the $v_i$ are iid with the same distribution as $u$. For fixed $B$, the average of these random vectors is

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

where $AN_g$ denotes an approximate multivariate normal distribution. Hence $(T - \theta) = O_P((nB)^{-1/2})$, and $T$ gets arbitrarily close to $\theta$ compared to $T_n$ as $B \rightarrow \infty$. Thus $R_c$ is a large sample 100$(1 - \delta)$% confidence region for $\theta$ as $n, B \rightarrow \infty$. (If $B$ is finite, the coverage is less than $1 - \delta_B$, and using (3) increases the coverage.)

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 $\sqrt{n}(T_n - \theta)$ and $\sqrt{n}(T^*_i - T_n)$ both converge in distribution to $u \sim N_g(0, \Sigma_A)$, 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 (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 $\sqrt{n}(T^* - T_n) \xrightarrow{D} 0$.

Let the variable selection estimator $T_n = A\beta_{MLX}$ with $\theta = A\beta$. Here $A$ is a known full rank $g \times p$ matrix with $1 \leq g \leq p$. We have $\sqrt{n}(T_n - \theta) \xrightarrow{D} v$ by (10) where $E(v) = 0$, and $\Sigma_v = \sum_{j} \pi_j A V_{j,0} A^T$. Hence the above geometric argument holds: if we had iid data $T_1, ..., T_B$, then $R_c$ would be a large sample confidence region for $\theta$. If $\sqrt{n}(T^*_n - T_n) \xrightarrow{D} v$, 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 $\theta = A\beta = B\beta_S$ if $V$ was.
diagonal. Section 3 will explain why the bootstrap confidence regions are still useful.

Pötscher (1991) used the conditional distribution of \( \hat{\beta}_{VS}(\hat{\beta}_{VS} = \beta_{I_k,0}) \) to find the distribution of \( w_n = \sqrt{n}(\beta_{VS} - \beta) \). Let \( W = W_{VS} = k \) if \( \beta_{VS} = \beta_{I_k,0} \) where \( P(W_{VS} = k) = \pi_{kn} \) for \( k = 1, \ldots, J \). Then \( (\beta_{VS:n}, W_{VS:n}) = (\beta_{VS}, W_{VS}) \) has a joint distribution where the sample size \( n \) is usually suppressed. Note that \( \hat{\beta}_{VS} = \beta_{I_{W},0} \). Define \( P(A|B_k)P(B_k) = 0 \) if \( P(B_k) = 0 \).

Let \( \hat{\beta}_{I_k,0} \) be a random vector from the conditional distribution \( \hat{\beta}_{I_k,0}(W_{VS} = k) \). Let \( w_{kn} = \sqrt{n}(\beta_{I_k,0} - \beta)(W_{VS} = k) = \sqrt{n}(\beta_{I_k,0} - \beta) \). Denote \( F_{w}(t) = P(z_1 \leq t_1, \ldots, z_p \leq t_p) \) by \( P(z \leq t) \). Then

\[
F_{w}(t) = P[n^{1/2}(\hat{\beta}_{VS} - \beta) \leq t] = \sum_{k=1}^{J} P[n^{1/2}(\hat{\beta}_{VS} - \beta) \leq t|(\hat{\beta}_{VS} = \hat{\beta}_{I_k,0})]P(\hat{\beta}_{VS} = \hat{\beta}_{I_k,0}) = \sum_{k=1}^{J} P[n^{1/2}(\hat{\beta}_{I_k,0} - \beta) \leq t|(\hat{\beta}_{VS} = \hat{\beta}_{I_k,0})]P_{kn} = \sum_{k=1}^{J} F_{w_{kn}}(t)\pi_{kn}.
\]

Hence \( \hat{\beta}_{VS} \) has a mixture distribution of the \( \hat{\beta}_{I_k,0} \) with probabilities \( \pi_{kn} \), and \( w_n \) has a mixture distribution of the \( w_{kn} \) with probabilities \( \pi_{kn} \).

Charkhi and Claeskens (2018) showed that \( \omega_{kn} = \sqrt{n}(\beta_{I_k,0} - \beta) \overset{D}{\to} w_j \) if \( S \subseteq I_j \) for the MLE with AIC. Here \( w_j \) is a multivariate truncated normal distribution (where no truncation is possible) that is symmetric about \( 0 \). Hence \( E(w_j) = 0 \), and Cov\( (w_j) = \Sigma_j \) exits. Note that both \( \sqrt{n}(\hat{\beta}_{MIX} - \beta) \) and \( \sqrt{n}(\hat{\beta}_{VS} - \beta) \) are selecting from the \( u_{kn} = \sqrt{n}(\beta_{I_k,0} - \beta) \) and asymptotically from the \( u_j \). The random selection for \( \hat{\beta}_{MIX} \) does not change the distribution of \( u_{jn} \), but selection bias does change the distribution of the selected \( u_{jn} \) to that of \( w_{jn} \). Similarly, selection bias does change the distribution of the selected \( u_j \) to that of \( w_j \). Let \( W = W_{VS:n} \) where \( P(W = k) = \pi_k \). In Theorem 1, the assumption \( \omega_{jn} = \sqrt{n}(\beta_{I_j,0} - \beta) \overset{D}{\to} u_j \) is a mild assumption that is made for large sample tests for whether a reduced model \( I_j \) is good. The reasonable Theorem 3 assumption that \( \omega_{jn} \overset{D}{\to} w_j \) may not be mild. Regularity conditions for the \( w_j \) to have \( E(w_j) = 0 \) may be strong. The proof for Equation (11) is the same as that for (9). Pelawa Watagod and Olive (2019b) have a similar result for multiple linear regression.
BOOTSTRAPPING VARIABLE SELECTION ESTIMATORS

Theorem 3. Assume \( P(S \subseteq I_{\min}) \to 1 \) as \( n \to \infty \), and let \( \hat{\beta}_{VS} = \hat{\beta}_{I_k,0} \) with probabilities \( \pi_{kn} \) where \( \pi_{kn} \to \pi_k \) as \( n \to \infty \). Denote the positive \( \pi_k \) by \( \pi_j \). Assume \( w_{jn} = \sqrt{n}(\beta_{I_j,0} - \beta) \xrightarrow{D} w_j \). Then

\[
(11) \quad w_n = \sqrt{n}(\beta_{VS} - \beta) \xrightarrow{D} w
\]

where the cdf of \( w \) is \( F_w(t) = \sum_j \pi_j F_{w_j}(t) \). Thus \( w \) is a mixture distribution of the \( w_j \) with probabilities \( \pi_j \).

The estimator \( \hat{\beta}_{MIX} \) is the random vector with selection probabilities \( P(W_{MIX} = k) = \pi_{kn} = P(W_{VS} = k) \) where \( W_{MIX} \) is independent of the \( \hat{\beta}_{I_k,0} \). Simulating \( \hat{\beta}_{MIX} \) and \( \hat{\beta}_{VS} \) is informative. We consider \( X = X_n \) fixed or condition on \( X_n \). The probabilities \( \pi_{kn} \) depend on \( X_n, n, p \), the variable selection estimator, and the population model that generates \( Y \). Consider the experiment of generating \( Y \) from the model. (For example, i) for a parametric regression model, generate \( Y_i \sim D(x_i^T \beta, \gamma) \) for \( i = 1, \ldots, n \) to form \( Y \), and ii) for multiple linear regression, generates \( e \) from the population of \( e \), and form \( Y = X_n \beta + e \). Then regress \( Y \) on \( X_n \) with variable selection to generate \( (\hat{\beta}_{VS}, W_{VS}) \). Generate another \( Y \) from the model, and generate \( (\hat{\beta}_{MIX}, W_{VS}) \). Then \( \hat{\beta}_{MIX} = \hat{\beta}_{I_k,0} \) whenever \( \hat{\beta}_{VS} = \beta_{I_k,0} \), but \( I_k \) was chosen for \( \hat{\beta}_{MIX} \) before generating the new \( Y \), so there is no selection bias. Repeat to get the sample \((\hat{\beta}_{I_{k1},0}, \hat{\beta}_{I_{k1},0}), \ldots, (\hat{\beta}_{I_{kB},0}, \hat{\beta}_{I_{kB},0})\).

3. Bootstrapping variable selection estimators. Obtaining the bootstrap samples for \( \hat{\beta}_{VS} \) and \( \hat{\beta}_{MIX} \) is simple. Generate \( Y^* \) and \( X^* \) that would be used to produce \( \hat{\beta}^* \) if the full model estimator \( \hat{\beta} \) was being bootstrapped. (The nonparametric bootstrap, parametric bootstrap, and residual bootstrap using the residuals from the full OLS model are discussed below.) Often \( X^* = X_n \). Instead of generating \( \hat{\beta}^* \), compute the variable selection estimator \( \hat{\beta}_{VS}^* = \hat{\beta}_{I_{k1},0}^* \). Then generate another \( Y^* \) and \( X^* \) and compute \( \hat{\beta}_{MIX}^* = \hat{\beta}_{I_{k1},0}^* \) (using the same subset \( I_{k1} \)). Note that the \( \epsilon_{i} = \hat{\beta}_{VS,i}^* - \hat{\beta}_{MIX,i}^* \) are iid with respect to the bootstrap distribution for \( i = 1, \ldots, B \). Let \( T = T_n = \hat{\beta} \). Then \( \sqrt{n}(T_{VS}^* - T_{MIX}^*) = \sqrt{n} \sum_{i=1}^B \epsilon_i^*/B \), and the two bagging estimators may be asymptotically equivalent. Let the selection probabilities for the bootstrap variable selection estimator be \( \rho_{kn} \). Then this bootstrap procedure bootstraps both \( \hat{\beta}_{VS} \) and \( \hat{\beta}_{MIX} \) with \( \pi_{kn} = \rho_{kn} \).

The key idea is to show that the bootstrap data cloud is slightly more variable than the iid data cloud, so confidence region (6) applied to the bootstrap data cloud has coverage bounded below by \( (1 - \delta) \) for large enough \( n \) and \( B \). 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

$$T^*_{1,1}, ..., T^*_{B_{1n}}, ..., T^*_{1,j}, ..., T^*_{B_{jn}, j}.$$ 

Denote $T^*_1, ..., T^*_B$ as the $j$th bootstrap component of the bootstrap sample with sample mean $\overline{T}_j^*$ and sample covariance matrix $S^*_T$. Similarly, we can define the $j$th component of the iid sample $T_1, ..., T_B$ to have sample mean $T^*_j$ and sample covariance matrix $S^*_T$.

Let $T_n = \hat{\beta}_{MIX}$ and $T_{ij} = \hat{\beta}_{I_j,0}$. If $S \subseteq I_j$, assume $\sqrt{n} (\hat{\beta}_{I_j} - \beta_{I_j}) \xrightarrow{D} N_{a_j}(0, V_j)$ and $\sqrt{n} (\hat{\beta}^*_I - \hat{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(0, V_j)$. Then by Equation (8),

(12) $\sqrt{n} (\hat{\beta}_{I_j,0} - \beta) \xrightarrow{D} N_p(0, V_{j,0})$ and $\sqrt{n} (\hat{\beta}^*_{I_j,0} - \hat{\beta}_{I_j,0}) \xrightarrow{D} N_p(0, V_{j,0})$.

This result means that the component clouds have the same variability asymptotically. The iid data component clouds are all centered at $\beta$. If the bootstrap data component clouds were all centered at the same value $\hat{\beta}$, then the bootstrap cloud would be like an iid data cloud shifted to be centered at $\hat{\beta}$, and (6) would be a confidence region for $\theta = \beta$. Instead, the bootstrap data component clouds are shifted slightly from a common center, and are each centered at a $\hat{\beta}_{I_j,0}$. Geometrically, the shifting of the bootstrap component data clouds makes the bootstrap data cloud similar but more variable than the iid data cloud asymptotically (we want $n \geq 20p$), and centering the bootstrap data cloud at $T_n$ results in the confidence region (6) having slightly higher asymptotic coverage than applying (6) to the iid data cloud. Then (5) tends to have higher coverage than (6) since the cutoff for (5) tends to be larger than the cutoff for (6). 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 $\theta$ at least as well as $T_n$ for a mixture distribution. See Breiman (1996) and Yang (2003). A similar argument holds if $T_n = A\hat{\beta}_{MIX}$, $T_{ij} = A\hat{\beta}_{I_j,0}$, and $\theta = A\beta$.

In the simulations of Section 4 for $H_0 : A\beta = B\beta_S = \theta_0$ with $n \geq 20p$, the coverage tended to get close to $1 - \delta$ for $B \geq \max(200, 50p)$ so that $S^*_T$ is a good estimator of $\text{Cov}(T^*)$. In the simulations where $S$ is not the full model, inference with backward elimination with $I_{\min}$ using $AIC$ was often more precise than inference with the full model if $n \geq 20p$ and $B \geq 50p$. Pelawa Watagoda and Olive (2019a) had similar results for multiple linear regression using forward selection for with $C_p$.

It is possible that $S^*_T$ is singular if a column of the bootstrap sample is equal to $0$. If the regression model has a $q \times 1$ vector of parameters $\gamma$, we may need to replace $p$ by $p + q$. 


Undercoverage can occur if 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 $H_0: A\beta = \beta_0 = 0$ where $\beta_0 = (\beta_1, ..., \beta_g)^T$ and $O \subseteq E$ in (1) so that $H_0$ is true. Suppose a nominal 95% confidence region is used and $U_B$ is the 96th percentile. Hence the confidence region (4) or (5) covers at least 96% of the bootstrap sample. If $\hat{\beta}_{O,j} = 0$ for more than 4% of the $\hat{\beta}_{O,1}, ..., \hat{\beta}_{O,B}$, then 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 $\hat{\beta}_{O,j} = 0$ for $j = 1, ..., B$. Then $S_I$ is singular, but the singleton set $\{0\}$ is the large sample 100(1 - $\delta$)% confidence region (4), (6), or (8) for $\beta_0$ and $\delta \in (0, 1)$, and the pvalue for $H_0: \beta_0 = 0$ is one. (This result holds since $\{0\}$ contains 100% of the $\hat{\beta}_{O,j}$ 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 $\beta = (\hat{\beta}_I^T, \hat{\beta}_O^T)^T$. For the $I_{min}$ model from variable selection, there may be strong evidence that $x_O$ is not needed in the model given $x_I$ is in the model if the “100%” confidence region is $\{0\}$, $n \geq 20p$, and $B \geq 50p$. (Since the pvalue is one, this technique may be useful for data snooping: applying regression model theory to submodel $I$ may have negligible selection bias.)

Assume $P(S \subseteq I_{min}) \rightarrow 1$ as $n \rightarrow \infty$, and that $S \subseteq I_j$. We want to examine when Equation (12) holds or when

$$\text{Cov}(\hat{\beta}_I^*) - \text{Cov}(\hat{\beta}_I) \rightarrow 0$$

as $n, B \rightarrow \infty$. Then the component clouds of the iid data and bootstrap data clouds have the same asymptotic variability, and the bootstrap confidence regions may give good results.

For multiple linear regression with the residual bootstrap that uses residuals from the full OLS model, Pelawa Watagoda and Olive (2019a) showed that Equation (13) holds for OLS variable selection. The nonparametric bootstrap (also called the empirical bootstrap, naive bootstrap, and the pairs bootstrap) draws a sample $n$ cases ($Y_i^*, x_i^*$) with replacement from the $n$ cases ($Y_i, x_i$), and regresses the $Y_i^*$ on the $x_i^*$ to get $\hat{\beta}_{I_{min},0,1}$. This process is repeated $B$ times to get the bootstrap sample $\hat{\beta}_{I_{min},0,i}^*$ for $i = 1, ..., n$. Under regularity conditions, Equation (12) holds. 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). Assumptions for the nonparametric bootstrap tend to be rather strong: often one assumption is that the \( n \) cases \((Y_i, x_i^T)^T\) are iid from some population.

Next, consider the parametric regression model \( Y_i|\mathbf{x}_i \sim D(x_i^T \beta, \gamma) \), and the parametric bootstrap. Suppose \( \sqrt{n}(\hat{\beta} - \beta) \xrightarrow{D} \mathcal{N}_p(0, V(\beta)) \), and that \( V(\hat{\beta}) \xrightarrow{p} V(\beta) \) as \( n \to \infty \). These assumptions tend to be mild for a parametric regression model where the maximum likelihood estimator (MLE) \( \hat{\beta} \) is used. Then \( V(\beta) = I^{-1}(\beta) \), the inverse Fisher information matrix. If \( \mathbf{I}_n(\beta) \) is the Fisher information matrix based on a sample of size \( n \), then \( \mathbf{I}_n(\beta)/n \xrightarrow{p} \mathbf{I}(\beta) \). For GLMs, see, for example, Sen and Singer (1993, p. 309).

For the parametric regression model, we regress \( Y \) on \( X \) to obtain \((\beta, \gamma)\) where the \( n \times 1 \) vector \( Y = (Y_i) \) and the \( i \)th row of the \( n \times p \) design matrix \( X \) is \( x_i^T \).

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

\[
\sqrt{n}(\hat{\beta}^* - \hat{\beta}) \xrightarrow{D} \mathcal{N}_p(0, V(\beta))
\]
as \( n \to \infty \).

Now suppose \( S \subseteq I \). Without loss of generality, let \( \beta = (\beta_T, \beta_O)^T \) and \( \hat{\beta} = (\hat{\beta}_T, \hat{\beta}_O)^T \). Then \((Y, X_I)\) follows the parametric regression model with parameters \((\beta_I, \gamma)\). Hence \( \sqrt{n}(\hat{\beta}_I - \beta_I) \xrightarrow{D} \mathcal{N}_{d_I}(0, V(\beta_I)) \). Now \((Y^*, X_I)\) only follows the parametric regression model asymptotically, since \( \hat{\beta}(O) \neq 0 \). However, under regularity conditions, \( \mathcal{E}(\hat{\beta}_I^*) \approx \beta_I \) and \( \text{Cov}(\hat{\beta}_I^*) \approx \text{Cov}(\beta_I) \) as \( n, B \to \infty \). See the following example.

Consider the multiple linear regression model \( Y_i = \beta_1 + x_{i12}\beta_2 + \cdots + x_{ip}\beta_p + e_i = x_i^T \beta + e_i \) for \( i = 1, \ldots, n \) where the random variables \( e_i \) are iid with variance \( V(e_i) = \sigma^2 \). In matrix notation, these \( n \) equations become \( Y = X\beta + e \). Let \( H = X(X^T X)^{-1}X^T \). Assume the maximum leverage \( \max_{i=1, \ldots, n} x_{i1}^T (X_I^T X_I)^{-1} x_{i1} \to 0 \) in probability as \( n \to \infty \) for each \( I \) with \( S \subseteq I \). For the OLS model with \( S \subseteq I \), \( \sqrt{n}(\hat{\beta}_I - \beta_I) \xrightarrow{D} \mathcal{N}_{d_I}(0, V_I) \) where \( (X_I^T X_I)/(n\sigma^2) \xrightarrow{p} V_I^{-1} \). See, for example, Sen and Singer (1993, p. 280).
Consider the parametric bootstrap for the above model with \( Y^* \sim N_n(X\hat{\beta}, \hat{\sigma}^2_n I) \sim N_n(HY, \hat{\sigma}^2_n I) \) where we are not assuming that the \( e_i \sim N(0, \sigma^2) \), and

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

where the residuals are from the full OLS model. Then \( MSE \) is a \( \sqrt{n} \) consistent estimator of \( \sigma^2 \) under mild conditions by Su and Cook (2012). Thus \( \hat{\beta}_I^* \) is a standard normal distribution since \( E(\hat{\beta}_I^*) = \hat{\beta}_I \) because \( HX_I = X_I \), and \( \text{Cov}(\hat{\beta}_I^*) = \hat{\sigma}^2_n(X_I^T X_I)^{-1} \).

Hence

\[
\sqrt{n}(\hat{\beta}_I^* - \hat{\beta}_I) \sim N_{a_1}(0, n\hat{\sigma}^2_n(X_I^T X_I)^{-1}) \overset{D}{\rightarrow} N_{a_1}(0, V_I)
\]

as \( n, B \to \infty \) if \( S \subseteq I \). For the residual bootstrap using residuals from the full OLS model, Pelawa Watagoda and Olive (2019a) showed that \( E(\hat{\beta}_I^*) = \hat{\beta}_I \) and \( \text{Cov}(\hat{\beta}_I^*) = [(n-p)/n]\hat{\sigma}^2_n(X_I^T X_I)^{-1} \overset{D}{\rightarrow} V_I \). Note that both the residual bootstrap and above parametric bootstrap for OLS are robust to the unknown error distribution of the iid \( e_i \).

4. **Examples and simulations.** The full model should be checked before doing variable selection inference. Assume \( p \) is fixed and \( n \geq 20p \). The response plot has \( ESP = x^T\hat{\beta} \) on the horizontal axis and \( Y \) on the vertical axis, respectively. Often the model mean function and a scatterplot smoother are added to the plots as visual aids to check the model. See Olive (2013, 2017a, ch. 13). Pelawa Watagoda and Olive (2019a) have an example and simulations for multiple linear regression using the residual bootstrap. We will also use Poisson regression, binomial regression, and Cox proportional hazards regression.

**Example 1.** Lindenmayer et al. (1991) and Cook and Weisberg (1999, p. 533) give 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 \textit{acacia}=basal area of acacia + 1, \textit{bark}=bark index, \textit{habitat}=habitat score, \textit{shrubs}=number of shrubs + 1, \textit{stags}=number of hollow trees + 1, \textit{stumps}=indicator for presence of stumps, and a constant. Inference for the full Poisson regression model is shown along with the shorth(c) nominal 95% confidence intervals for \( \beta_i \) computed using the parametric bootstrap with \( B = 1000 \). As expected, the bootstrap intervals are close to the large sample GLM confidence intervals \( \approx \hat{\beta}_i \pm 2SE(\hat{\beta}_i) \).

The minimum AIC model from backward elimination used a constant, \textit{bark}, \textit{habitat}, and \textit{stags}. The shorth(c) nominal 95% confidence intervals for \( \beta_i \) using the parametric bootstrap are shown. Note that most of the
confidence intervals contain 0 when closed intervals are used instead of open intervals. The Poisson regression output is also shown, but should only be used for inference if the model was selected before looking at the data.

### Large sample full model inference

|        | Est. | SE   | z    | Pr(>|z|) | 95% shorth CI          |
|--------|------|------|------|----------|------------------------|
| int    | -1.0428 | 0.2480 | -4.205 | 0.0000 | [-1.5619,-0.5381]     |
| acacia | 0.0166 | 0.0103 | 1.612  | 0.1070 | [-0.0042, 0.0345]     |
| bark   | 0.0361 | 0.0140 | 2.579  | 0.0099 | [0.00736, 0.0650]     |
| habitat| 0.0762 | 0.0375 | 2.032  | 0.0422 | [-0.0032, 0.1436]     |
| shrubs | 0.0145 | 0.0205 | 0.707  | 0.4798 | [-0.0279, 0.0557]     |
| stags  | 0.0325 | 0.0103 | 3.161  | 0.0016 | [0.0127, 0.0538]      |
| stumps | -0.3907 | 0.2866 | -1.364 | 0.1727 | [-1.0103, 0.1713]     |

### Output and shorth intervals for the min AIC submodel

|         | Est.  | SE    | z    | Pr(>|z|) | 95% shorth CI          |
|---------|-------|-------|------|----------|------------------------|
| int     | -0.8994 | 0.2135 | -4.212 | 0.0000 | [-1.4382,-0.4277]     |
| acacia  | 0     |       |      | 0.0000  | [0.0000, 0.0368]      |
| bark    | 0.0336 | 0.0121 | 2.773 | 0.0056  | [0.0000, 0.0601]      |
| habitat | 0.1069 | 0.0297 | 3.603 | 0.0003  | [0.0000, 0.1559]      |
| shrubs  | 0     |       |      | 0.0000  | [0.0000, 0.0602]      |
| stags   | 0.0302 | 0.0094 | 3.210 | 0.0013  | [0.0000, 0.0542]      |
| stumps  | 0     |       |      | 0.0000  | [-0.9702, 0.0000]     |

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 $H_0 : A\beta = (\beta_2, \beta_5, \beta_7)^T = 0$. Using the prediction region method with the full model had $[0, D_{(UB)}] = [0, 2.836]$ with $D_0 = 2.135$. 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 $[0, D_{(UB)}] = [0, 2.804]$ while $D_0 = 1.269$. So fail to reject $H_0$. The ratio of the volumes of the bootstrap confidence regions for this test was 0.322. (Use (12) with $S^*_T$ 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.

### Large sample full model inference

|        | Est. | SE   | z    | Pr(>|z|) | 95% shorth CI          |
|--------|------|------|------|----------|------------------------|
| int    | -475.5814 | 404.9133 | -1.175 | 0.2402 | [-83274.992,1939.714]  |
| length | 0.3753 | 1.4175 | 0.265 | 0.7912  | [-98.9019,137.5894]   |
EXAMPLE 2. For binary logistic regression, the MLE tends to converge if $\max(|x_i^T\hat{\beta}|) \leq 7$ and if the $Y$ values of 0 and 1 are not nearly perfectly classified by the rule $\hat{Y} = 1$ if $x_i^T\hat{\beta} > 0.5$ and $\hat{Y} = 0$, otherwise. If there is perfect classification, the MLE does not exist. Let $\hat{\rho}(x) = \hat{P}(Y = 1|x)$ under the binary logistic regression. If $|x_i^T\hat{\beta}| \geq 10$, some of the $\hat{\rho}(x_i)$ tend to be estimated to be exactly equal to 0 or 1, which causes problems for the MLE. The Flury and Riedwyl (1988, pp. 5-6) banknote data consists of 100 counterfeit and 100 genuine Swiss banknote. The response variable is an indicator for whether the banknote is counterfeit. The six predictors are measurements on the banknote: bottom, diagonal, left, length, right, and top. When the logistic regression model is fit with these predictors, there is almost perfect classification and backward elimination had problems. We deleted diagonal, which is likely an important predictor, so backward elimination would run. For this full model, classification is very good, but the $x_i^T\hat{\beta}$ run from $-20$ to 20. In a plot of $x_i^T\hat{\beta}$ versus $Y$ on the vertical axis (not shown), the logistic regression mean function is tracked closely by the lowess scatterplot smoother. The full model and backward elimination output is above. Inference using the logistic regression normal approximation appears to greatly underestimate the variability of $\hat{\beta}$ compared to the parametric full model bootstrap variability. We tested $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ with the $I_{min}$ model selected by backward elimination. Using the prediction region method with the full model had $[0, D_{(UB)}] = [0, 1.763]$ with $D_0 = 0.2046$. 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 $[0, D_{(UB)}] = [0, 1.511]$ while $D_0 = 0.2297$. So fail to reject $H_0$. The ratio of the volumes of the bootstrap confidence regions for this test was 16.2747. Hence the full model bootstrap inference was much more precise. Backward
elimination produced many zeros, but also produced many estimates that were very large in magnitude.

Binary regression data sets like the one in Example 2 are common: the response plot of \( x_i^T \hat{\beta} \) versus \( Y \) suggests that the logistic regression mean function is good, but the range of \( x_i^T \hat{\beta} \) is such that the GLM normal approximation to the MLE \( \hat{\beta} \) is likely invalid. Since the parametric bootstrap produces datasets very similar to the actual dataset, the bootstrap distribution of the logistic regression MLE may be superior to the GLM normal approximation. For Example 2, the GLM and bootstrap inference for the full model both suggest that bottom and top are important predictors.

In the simulations, we used 5000 runs, \( \theta = A \beta = \beta_c, \theta = A \beta = \beta_S = (\beta_1, 1, \ldots, 1)^T \) and \( \theta = A \beta = \beta_E = 0 \). The simulations often used \( n = 25p, n = 50p; \psi = 0, 1/\sqrt{p}, \) and 0.9; and \( k = 1 \) and 2 where \( k \) and \( \psi \) are defined in the following paragraph. We tend to use \( p = 4 \) since the simulations with 5000 runs take a long time. A larger simulation study for \( \hat{\beta}_Y \) for Poisson and binomial regression with \( p \) as large as 10 is in Rathnayake (2019).

Let \( x = (1 \ u^T)^T \) where \( u \) is the \((p - 1) \times 1\) vector of nontrivial predictors. In the simulations, for \( i = 1, \ldots, n \), we generated \( w_i \sim N_{p-1}(0, I) \) where the \( q = p - 1\) elements of the vector \( w_i \) are iid \( N(0,1) \). Let the \( q \times q \) matrix \( A = (a_{ij}) \) with \( a_{ii} = 1 \) and \( a_{ij} = \psi \) where \( 0 \leq \psi < 1 \) for \( i \neq j \). Then the vector \( z_i = Aw_i \) so that \( \text{Cov}(z_i) = \Sigma_z = AA^T = (\sigma_{ij}) \) where the diagonal entries \( \sigma_{ii} = [1+(q-1)^2] \) and the off diagonal entries \( \sigma_{ij} = [2\psi+(q-2)^2] \). Hence the correlations are \( \text{cor}(z_i, z_j) = \rho = (2\psi+(q-2)^2)/(1+(q-1)^2) \) 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 \( u = az/v \). Then \( \text{cor}(x_i, x_j) = \rho \) for \( i \neq j \) where \( x_i \) and \( x_j \) are nontrivial predictors. If \( \psi = 1/\sqrt{cp} \), then \( p \rightarrow 1/(c+1) \) as \( p \rightarrow \infty \) where \( c > 0 \). As \( \psi \) gets close to 1, the predictor vectors \( u_i \) cluster about the line in the direction of \((1, \ldots, 1)^T\). Let \( S \beta = x^T \beta = \beta_1 + 1x_{i,2} + \cdots + 1x_{i,k+1} \sim N(\beta_1, a^2) \) for \( i = 1, \ldots, n \). Hence \( \beta = (\beta_1, 1, \ldots, 1, 0, \ldots, 0)^T \) with \( \beta_1 \), \( k \) 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 computed the Frey short(\( c \)) interval for each \( \beta_i \) and used bootstrap confidence regions to test \( H_0 : \beta_S = (\beta_1, 1, \ldots, 1)^T \) where \( \beta_2 = \cdots = \beta_{k+1} = 1 \), and \( H_0 : \beta_E = 0 \) (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.

In the tables, there are 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 variable selection, and “mix” for random selection. 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 $H_0 : \beta_E = 0$, while the 1 indicates that the test was $H_0 : \beta_S = (\beta_1, 1, ..., 1)^T$. 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 $\beta_S$ if $k = 1$.

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 regression versus that for the variable selection regression or random selection since the three methods have different determinants $|S^*_T|$. For random selection, there is no $\hat{\beta}_{MIX}$. Hence for the hybrid region and Bickel and Ren region $T_n = A\hat{\beta}_{VS}$ was used, and the coverage for the hybrid region was often 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 parametric bootstrap on the full model. Variable selection coverage tended to be near 0.95 unless the $\hat{\beta}_i$ could equal 0. 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 Bickel and Ren (5) average cutoffs were at least as high as those of the hybrid region (6).

If $\beta_i$ was a component of $\beta_E$, 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 $\hat{\beta}_E$ 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 $I^{-1}(\beta)$ is diagonal. Hence $\hat{\beta}_S$ has the same multivariate normal limiting distribution for $I_{min}$ and the full model by Section 2. For Tables 1-4, $\beta_S = (\beta_1, \beta_2)^T$, and $\beta_{p-1}$ and $\beta_p$ are components of $\beta_E$. For Table 5,
\[ \beta_s = (\beta_1, ..., \beta_9)^T. \] Hence \( \beta_1, \beta_2, \) and \( \beta_{p-1} \) are components of \( \beta_s \), while \( \beta_E = \beta_{10} \). For the \( n \) in the tables and \( \psi = 0 \), the coverages and “lengths” did tend to be close for the \( \beta_i \) that are components of \( \beta_s \), and for \( \text{pr1, hyb1, and br1} \).

Table 1 was for multiple linear regression with forward selection, the residual bootstrap, \( n = 100, p = 4, k = 1 \) and \( B = 1000 \). There was slight undercoverage 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, without the MIX rows, is in Pelawa Watagoda (2017).

<table>
<thead>
<tr>
<th>( \psi )</th>
<th>( \beta_1 )</th>
<th>( \beta_2 )</th>
<th>( \beta_{p-1} )</th>
<th>( \beta_p )</th>
<th>pr0</th>
<th>hyb0</th>
<th>br0</th>
<th>pr1</th>
<th>hyb1</th>
<th>br1</th>
</tr>
</thead>
<tbody>
<tr>
<td>reg,0</td>
<td>0.939</td>
<td>0.949</td>
<td>0.949</td>
<td>0.944</td>
<td>0.941</td>
<td>0.942</td>
<td>0.942</td>
<td>0.937</td>
<td>0.937</td>
<td>0.938</td>
</tr>
<tr>
<td>len</td>
<td>0.393</td>
<td>0.400</td>
<td>0.399</td>
<td>0.400</td>
<td>2.4474</td>
<td>2.474</td>
<td>2.475</td>
<td>2.453</td>
<td>2.453</td>
<td>2.455</td>
</tr>
<tr>
<td>vs,0</td>
<td>0.938</td>
<td>0.944</td>
<td>0.999</td>
<td>0.997</td>
<td>0.994</td>
<td>0.984</td>
<td>0.995</td>
<td>0.932</td>
<td>0.933</td>
<td>0.934</td>
</tr>
<tr>
<td>len</td>
<td>0.392</td>
<td>0.398</td>
<td>0.323</td>
<td>0.323</td>
<td>2.709</td>
<td>2.709</td>
<td>3.014</td>
<td>2.453</td>
<td>2.453</td>
<td>2.490</td>
</tr>
<tr>
<td>mix,0</td>
<td>0.937</td>
<td>0.943</td>
<td>0.999</td>
<td>0.998</td>
<td>0.998</td>
<td>0.987</td>
<td>0.995</td>
<td>0.930</td>
<td>0.931</td>
<td>0.931</td>
</tr>
<tr>
<td>len</td>
<td>0.391</td>
<td>0.397</td>
<td>0.274</td>
<td>0.278</td>
<td>3.072</td>
<td>3.072</td>
<td>3.288</td>
<td>2.455</td>
<td>2.455</td>
<td>2.459</td>
</tr>
<tr>
<td>reg,0.9</td>
<td>0.940</td>
<td>0.948</td>
<td>0.953</td>
<td>0.953</td>
<td>0.946</td>
<td>0.948</td>
<td>0.948</td>
<td>0.933</td>
<td>0.933</td>
<td>0.932</td>
</tr>
<tr>
<td>len</td>
<td>0.392</td>
<td>3.249</td>
<td>3.249</td>
<td>3.249</td>
<td>2.475</td>
<td>2.475</td>
<td>2.476</td>
<td>2.454</td>
<td>2.454</td>
<td>2.455</td>
</tr>
<tr>
<td>vs,0.9</td>
<td>0.941</td>
<td>0.966</td>
<td>0.966</td>
<td>0.997</td>
<td>0.992</td>
<td>0.989</td>
<td>0.994</td>
<td>0.975</td>
<td>0.975</td>
<td>0.962</td>
</tr>
<tr>
<td>len</td>
<td>0.393</td>
<td>2.754</td>
<td>2.721</td>
<td>2.713</td>
<td>2.712</td>
<td>2.712</td>
<td>2.950</td>
<td>2.492</td>
<td>2.492</td>
<td>2.507</td>
</tr>
<tr>
<td>mix,0.9</td>
<td>0.941</td>
<td>0.972</td>
<td>0.977</td>
<td>0.998</td>
<td>0.995</td>
<td>0.873</td>
<td>0.996</td>
<td>0.938</td>
<td>0.892</td>
<td>0.935</td>
</tr>
<tr>
<td>len</td>
<td>0.391</td>
<td>2.105</td>
<td>1.999</td>
<td>2.000</td>
<td>2.547</td>
<td>2.547</td>
<td>2.828</td>
<td>2.448</td>
<td>2.448</td>
<td>2.610</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( \psi )</th>
<th>( \beta_1 )</th>
<th>( \beta_2 )</th>
<th>( \beta_{p-1} )</th>
<th>( \beta_p )</th>
<th>pr0</th>
<th>hyb0</th>
<th>br0</th>
<th>pr1</th>
<th>hyb1</th>
<th>br1</th>
</tr>
</thead>
<tbody>
<tr>
<td>reg,0.9</td>
<td>0.946</td>
<td>0.954</td>
<td>0.952</td>
<td>0.950</td>
<td>0.955</td>
<td>0.955</td>
<td>0.964</td>
<td>0.966</td>
<td>0.966</td>
<td>0.966</td>
</tr>
<tr>
<td>len</td>
<td>0.755</td>
<td>0.684</td>
<td>0.669</td>
<td>0.680</td>
<td>2.489</td>
<td>2.489</td>
<td>2.499</td>
<td>2.486</td>
<td>2.486</td>
<td>2.497</td>
</tr>
<tr>
<td>vs,0.9</td>
<td>0.954</td>
<td>0.949</td>
<td>0.996</td>
<td>0.997</td>
<td>0.993</td>
<td>0.991</td>
<td>0.996</td>
<td>0.976</td>
<td>0.980</td>
<td>0.984</td>
</tr>
<tr>
<td>len</td>
<td>0.750</td>
<td>5.320</td>
<td>5.385</td>
<td>5.388</td>
<td>2.788</td>
<td>2.788</td>
<td>3.039</td>
<td>2.588</td>
<td>2.588</td>
<td>2.723</td>
</tr>
<tr>
<td>mix,0.9</td>
<td>0.955</td>
<td>0.966</td>
<td>0.997</td>
<td>0.997</td>
<td>0.996</td>
<td>0.989</td>
<td>0.996</td>
<td>0.977</td>
<td>0.968</td>
<td>0.974</td>
</tr>
<tr>
<td>len</td>
<td>0.741</td>
<td>3.938</td>
<td>3.859</td>
<td>3.865</td>
<td>2.869</td>
<td>2.869</td>
<td>3.046</td>
<td>2.604</td>
<td>2.604</td>
<td>2.707</td>
</tr>
</tbody>
</table>

For Cox proportional hazards regression, the cases were \( (Z_i, \delta_i, x_i) \) where
Table 3

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

<table>
<thead>
<tr>
<th>$\psi$</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\beta_{p-1}$</th>
<th>$\beta_p$</th>
<th>pr0</th>
<th>hyb0</th>
<th>br0</th>
<th>pr1</th>
<th>hyb1</th>
<th>br1</th>
</tr>
</thead>
<tbody>
<tr>
<td>reg,0</td>
<td>0.948</td>
<td>0.953</td>
<td>0.953</td>
<td>0.952</td>
<td>0.951</td>
<td>0.951</td>
<td>0.952</td>
<td>0.943</td>
<td>0.945</td>
<td>0.947</td>
</tr>
<tr>
<td>len,0</td>
<td>0.175</td>
<td>0.133</td>
<td>0.128</td>
<td>0.128</td>
<td>3.986</td>
<td>3.986</td>
<td>3.990</td>
<td>2.453</td>
<td>2.453</td>
<td>2.474</td>
</tr>
<tr>
<td>vs,0</td>
<td>0.947</td>
<td>0.952</td>
<td>0.998</td>
<td>0.999</td>
<td>0.998</td>
<td>0.997</td>
<td>0.997</td>
<td>0.950</td>
<td>0.954</td>
<td>0.958</td>
</tr>
<tr>
<td>len,0</td>
<td>0.175</td>
<td>0.132</td>
<td>0.105</td>
<td>0.104</td>
<td>4.303</td>
<td>4.303</td>
<td>4.740</td>
<td>2.452</td>
<td>2.452</td>
<td>2.499</td>
</tr>
<tr>
<td>mix,0</td>
<td>0.948</td>
<td>0.953</td>
<td>0.999</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.951</td>
<td>0.949</td>
<td>0.951</td>
</tr>
<tr>
<td>len,0</td>
<td>0.174</td>
<td>0.129</td>
<td>0.088</td>
<td>0.088</td>
<td>5.122</td>
<td>5.122</td>
<td>5.396</td>
<td>2.453</td>
<td>2.453</td>
<td>2.475</td>
</tr>
<tr>
<td>reg,0.9</td>
<td>0.948</td>
<td>0.953</td>
<td>0.957</td>
<td>0.955</td>
<td>0.946</td>
<td>0.948</td>
<td>0.948</td>
<td>0.945</td>
<td>0.945</td>
<td>0.948</td>
</tr>
<tr>
<td>len,0</td>
<td>0.175</td>
<td>3.287</td>
<td>3.286</td>
<td>3.291</td>
<td>3.983</td>
<td>3.983</td>
<td>3.987</td>
<td>2.454</td>
<td>2.454</td>
<td>2.469</td>
</tr>
<tr>
<td>vs,0.9</td>
<td>0.948</td>
<td>0.943</td>
<td>0.998</td>
<td>0.999</td>
<td>0.998</td>
<td>0.997</td>
<td>0.999</td>
<td>0.968</td>
<td>0.967</td>
<td>0.973</td>
</tr>
<tr>
<td>len,0</td>
<td>0.175</td>
<td>2.862</td>
<td>2.810</td>
<td>2.820</td>
<td>4.261</td>
<td>4.261</td>
<td>4.685</td>
<td>2.480</td>
<td>2.480</td>
<td>2.638</td>
</tr>
<tr>
<td>mix,0.9</td>
<td>0.951</td>
<td>0.921</td>
<td>0.998</td>
<td>0.999</td>
<td>1−</td>
<td>0.999</td>
<td>1−</td>
<td>0.899</td>
<td>0.874</td>
<td>0.898</td>
</tr>
<tr>
<td>len,0</td>
<td>0.174</td>
<td>2.527</td>
<td>2.231</td>
<td>2.228</td>
<td>4.971</td>
<td>4.971</td>
<td>5.287</td>
<td>2.569</td>
<td>2.569</td>
<td>2.711</td>
</tr>
</tbody>
</table>

Table 4

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

<table>
<thead>
<tr>
<th>$\psi$</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\beta_{p-1}$</th>
<th>$\beta_p$</th>
<th>pr0</th>
<th>hyb0</th>
<th>br0</th>
<th>pr1</th>
<th>hyb1</th>
<th>br1</th>
</tr>
</thead>
<tbody>
<tr>
<td>reg,0</td>
<td>0.936</td>
<td>0.934</td>
<td>0.952</td>
<td>0.956</td>
<td>0.951</td>
<td>0.962</td>
<td>0.968</td>
<td>0.945</td>
<td>0.965</td>
<td>0.974</td>
</tr>
<tr>
<td>len,0</td>
<td>0.850</td>
<td>0.852</td>
<td>0.744</td>
<td>0.744</td>
<td>2.525</td>
<td>2.525</td>
<td>2.552</td>
<td>2.514</td>
<td>2.514</td>
<td>2.640</td>
</tr>
<tr>
<td>vs,0</td>
<td>0.937</td>
<td>0.942</td>
<td>0.989</td>
<td>0.988</td>
<td>0.970</td>
<td>0.974</td>
<td>0.977</td>
<td>0.947</td>
<td>0.966</td>
<td>0.975</td>
</tr>
<tr>
<td>len,0</td>
<td>0.852</td>
<td>0.853</td>
<td>0.728</td>
<td>0.726</td>
<td>2.544</td>
<td>2.544</td>
<td>2.647</td>
<td>2.515</td>
<td>2.515</td>
<td>2.640</td>
</tr>
<tr>
<td>mix,0</td>
<td>0.940</td>
<td>0.942</td>
<td>0.992</td>
<td>0.992</td>
<td>0.979</td>
<td>0.978</td>
<td>0.981</td>
<td>0.946</td>
<td>0.966</td>
<td>0.976</td>
</tr>
<tr>
<td>len,0</td>
<td>0.842</td>
<td>0.841</td>
<td>0.667</td>
<td>0.666</td>
<td>2.691</td>
<td>2.691</td>
<td>2.758</td>
<td>2.515</td>
<td>2.515</td>
<td>2.623</td>
</tr>
<tr>
<td>reg,0.5</td>
<td>0.945</td>
<td>0.953</td>
<td>0.954</td>
<td>0.951</td>
<td>0.951</td>
<td>0.963</td>
<td>0.965</td>
<td>0.944</td>
<td>0.962</td>
<td>0.968</td>
</tr>
<tr>
<td>len,0</td>
<td>2.372</td>
<td>2.373</td>
<td>2.333</td>
<td>2.338</td>
<td>2.529</td>
<td>2.529</td>
<td>2.556</td>
<td>2.522</td>
<td>2.522</td>
<td>2.571</td>
</tr>
<tr>
<td>vs,0.5</td>
<td>0.968</td>
<td>0.960</td>
<td>0.992</td>
<td>0.990</td>
<td>0.980</td>
<td>0.979</td>
<td>0.984</td>
<td>0.976</td>
<td>0.965</td>
<td>0.975</td>
</tr>
<tr>
<td>len,0</td>
<td>2.064</td>
<td>2.064</td>
<td>2.046</td>
<td>2.045</td>
<td>2.784</td>
<td>2.784</td>
<td>2.931</td>
<td>2.558</td>
<td>2.558</td>
<td>2.688</td>
</tr>
<tr>
<td>mix,0.5</td>
<td>0.944</td>
<td>0.940</td>
<td>0.993</td>
<td>0.992</td>
<td>0.984</td>
<td>0.984</td>
<td>0.990</td>
<td>0.933</td>
<td>0.944</td>
<td>0.953</td>
</tr>
<tr>
<td>len,0</td>
<td>2.220</td>
<td>2.212</td>
<td>1.971</td>
<td>1.970</td>
<td>2.784</td>
<td>2.784</td>
<td>2.929</td>
<td>2.546</td>
<td>2.546</td>
<td>2.667</td>
</tr>
</tbody>
</table>
$Z_i = Y_i$ is 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 $SP = x_i^T \beta = 1x_{i,1} + \cdots + 1x_{i,k} \sim N(0, a^2)$ for $i = 1, \ldots, n$. The simulations use $a = 1$ where $\beta = (1, \ldots, 1, 0, \ldots, 0)^T$ with $k$ ones and $p - k$ zeros. We used $\psi = 0.5$ since $\psi = 0.9$ gave convergence problems.

5. Discussion. One problem with the bootstrap methods is that $S_T^*$ can be singular due to one or more columns of zeros in the bootstrap sample for $\beta_1, \ldots, \beta_p$. 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 $\hat{\beta}^* = \hat{\beta}_{FULL}$ to the bootstrap sample. For example, take $d = \lceil cB \rceil$ with $c = 0.01$. Let $S_A^*$ be the covariance matrix from the augmented bootstrap sample. Then apply the confidence regions to the augmented bootstrap sample or plug in $S_A^*$ for $S_T^*$ for the confidence regions from the unaugmented bootstrap sample. Augmentation changes the probabilities $\rho_{kn}$ to $\rho_{kn}/1.01$ except the full model selection probability changes from $\rho_{fn}$ to $0.01 + \rho_{fn}/1.01$. A confidence interval $[L_n, U_n]$ can be computed without $S_T^*$ for $S_T^*$ for the confidence regions from the unaugmented bootstrap sample. Using the confidence interval $[\max(L_n, T_{(1)}^B), \min(U_n, T_{(B)}^B)]$ can give a shorter covering region.

For binomial regression full model, inference from the three bootstrap confidence regions may be more reliable than inference from MLE theory if there is perfect separation or if the range of the ESP is so large that some estimated probabilities are exactly 0 or 1.

Note that there are several important variable selection models, including the model given by Equation (1). Another model is $x_i^T \beta = x_S^i \beta_S_i$ for $i = 1, \ldots, K$. Then there are $K \geq 2$ competing “true” nonnested submodels where $\beta_S_i$ is $a_S_i \times 1$. 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 $\hat{\beta}_{ij}^* = 0$ for at least one of the bootstrap samples $j = 1, \ldots, B$ 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 $\{0\}$ could still be the “100%” confidence region for a vector $\beta_O$. Theorem 1 holds if “$P(S \subseteq I_{min}) \to 1$” is replaced by “$P(S_i \subseteq I_{min}$ for some $i) \to 1$,”
and the bootstrap sample is still more variable than the iid sample.

There is a massive literature on variable selection and a fairly large literature for inference after variable selection. See, for example, Kabaila and Leeb (2006), Leeb and Pötscher (2006), Leeb, Pötscher, and Ewald (2015), Lu et al. (2017), Ning and Liu (2017), Pötscher (1991), Tibshirani et al. (2016, 2018), and Zhang (1992). Su (2018) shows that fast variable selection methods tend to select spurious variables quickly if \( k \) is not small. Recent papers on large sample theory for multiple linear regression estimators include Cook and Forzani (2019), Knight and Fu (2000), and Pelawa Watagoda and Olive (2019ab). Also see Cook (2018).

One of the best methods for inference after variable or model selection is “data splitting,” which can be used even if \( p >> n \). Data splitting uses a training set to find a model, e.g. \( I_{min} = I_j \). Then \( I_j \) is used as the full model for the validation set, avoiding selection bias so valid inference can be done. See, for example, Hurvich and Tsai (1990, p. 216) and Rinaldo et al. (2019).

For post-selection inference, the methods in the literature are often for multiple linear regression assuming normality, or are asymptotically equivalent to using the full model, or find a quantity to test that is not \( A\beta \). See Ewald and Schneider (2018). The bootstrap confidence region (9) is centered at \( T^* \), 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.

The simulations were done in R. See R Core Team (2016). We used several R functions including backward elimination computed with the \texttt{step} function from the Venables and Ripley (1997) \texttt{MASS} library. The collection of Olive (2020) \( R \) functions \texttt{slpack}, available from (http://lagrange.math.siu.edu/Olive/slpack.txt), has some useful functions for the inference. The functions \texttt{regbootsim3} and \texttt{vsbootsim5} were to bootstrap the full model and forward selection for multiple linear regression. The functions \texttt{binregbootsim} and \texttt{pregbootsim} are useful for the full binomial regression and full Poisson regression models. The functions \texttt{vsbrbootsim} and \texttt{vsprbootsim} were used to bootstrap backward elimination for binomial and Poisson regression. The functions \texttt{LRboot} and \texttt{vsLRboot} bootstrap the logistic regression full model and backward elimination. The functions \texttt{PRboot} and \texttt{vsPRboot} bootstrap the Poisson regression full model and backward elimination. The function \texttt{PHboot} bootstraps the full Cox PH model. The function \texttt{PHbootsim} is used to simulate the bootstrap for the full Cox PH model. The functions \texttt{LPHboot}
and RLPHboot bootstrap a Cox PH model with lasso and lasso variable selection. The function RLPHbootsim is used to simulate the bootstrap for lasso variable selection with Cox regression.

REFERENCES


Department of Mathematics, Carbondale, IL
E-mail: rathnaya@sisu.edu, dolive@sisu.edu