Bootstrapping Some GLM and Survival Regression Variable Selection Estimators

Rasanji C. Rathnayake
Department of Mathematics, Southern Illinois University, Carbondale, USA.
E-mail: rasanji@siu.edu

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

Summary. 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. Hence bootstrap confidence regions that can handle this complicated distribution were used for hypothesis testing. A bootstrap technique to eliminate selection bias is to fit the variable selection estimator $\hat{\beta}_{VS}^*$ to a bootstrap sample to find a submodel, then draw another bootstrap sample and fit the same submodel to get the bootstrap estimator $\hat{\beta}_{MIX}^*$. The theory shows that the lasso and elastic net variable selection estimators are $\sqrt{n}$ consistent when lasso and elastic net are consistent.

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 $Y \perp x | x^T \beta$. Many important regression models satisfy this condition, including 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). Lasso variable selection is called relaxed lasso by Hastie et al. (2015), and the relaxed lasso estimator with $\phi = 0$ by Meinshausen (2007). Also see Fan and Li (2001), Friedman et al. (2010), Simon et al. (2011), Tibshirani (1996), and Zou and Hastie (2005). The Meinshausen (2007) relaxed lasso estimator 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.
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 \( i = 1, \ldots, n \). 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

\[
Y_i \sim \text{binomial} \left( m_i, \rho(SP) = \frac{e^{SP}}{1 + e^{SP}} \right).
\]

The binary logistic regression model has \( Y \sim \text{Poisson} \left( e^{SP} \right) \). The Weibull proportional hazards regression model is

\[
Y|SP \sim W(\gamma = 1/\sigma, \lambda_0 \exp(SP))
\]

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

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

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

\[
x^T \beta = x^T_S \beta_S + x^T_E \beta_E = x^T_S \beta_S
\]

(1)

where \( x = (x^T_S, x^T_E)^T \), \( x_S \) is an \( a_S \times 1 \) vector, and \( x_E \) is a \((p - a_S) \times 1 \) vector. Given that \( x_S \) is in the model, \( \beta_E = 0 \) and \( E \) denotes the subset of terms that can be eliminated given that the subset \( S \) is in the model. Let \( x_I \) be the vector of \( a \) terms from a candidate subset indexed by \( I \), and let \( x_O \) 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

\[
x^T \beta = x^T_S \beta_S = x^T_I \beta_I + x^T_O 0 = x^T_I \beta_I.
\]

Thus \( \beta_O = 0 \) if \( S \subseteq I \). The model using \( x^T \beta \) 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 \( \{1, 2, \ldots, p\} \) 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 \( S \subseteq I_j \). Let \( \hat{\beta}_{I_k} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_4)^T \) and \( x_{I_k} = (x_1, x_3, x_4)^T \).

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

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

\[
F_u(t) = \sum_{j=1}^{J} \pi_j F_{u_j}(t)
\]

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

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

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

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 with confidence intervals (CIs) and regions. Consider testing \( H_0 : \theta = \theta_0 \) versus \( H_1 : \theta \neq \theta_0 \) where \( \theta_0 \) is a known \( g \times 1 \) vector. A large sample \( 100(1 - \delta)\% \) confidence region for \( \theta \) is a set \( \mathcal{A}_n \) such that \( P(\theta \in \mathcal{A}_n) \) is eventually bounded below by \( 1 - \delta \) as the sample size \( n \to \infty \). Then reject \( H_0 \) if \( \theta_0 \) is not in the confidence region. Let the \( g \times 1 \) vector \( T_n \) be an estimator of \( \theta \). Let \( T_1^*, \ldots, T_B^* \) be the bootstrap sample for \( T_n \). Let \( \mathbf{A} \) be a full rank \( g \times p \) constant matrix. For variable selection, test \( H_0 : \mathbf{A} \beta = \theta_0 \) versus \( H_1 : \mathbf{A} \beta \neq \theta_0 \) with \( \theta = \mathbf{A} \beta \). Then let \( T_n = \mathbf{A} \hat{\beta}_{SEL} \) and let \( T_i^* = \mathbf{A} \hat{\beta}_{SEL}^* \) for \( i = 1, \ldots, B \) 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 \( \theta \) is

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

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

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

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

\[
D_i^2 = D_i^2(T, \mathbf{C}) = D_{z_i}^2(T, \mathbf{C}) = (z_i - T)^T \mathbf{C}^{-1}(z_i - T)
\]
for each observation $z_i$. Let $\overline{T}$ and $S_{T}^*$ be the sample mean and sample covariance matrix of the bootstrap sample.

The Olive (2017a, b, 2018) prediction region method (4), modified Bickel and Ren (2001) (5), and Pelawa Watagoda and Olive (2019) hybrid (6) large sample 100(1−δ)% confidence regions for $\theta$ are

$$
\{ w : (w - \overline{T})^T [S_{T}^*]^{-1} (w - \overline{T}) \leq D_{2(U_B)}^2 \} = \{ w : D_w^2(\overline{T}, S_{T}^*) \leq D_{2(U_B)}^2 \}
$$

(4)

where $D_{2(U_B)}^2$ is computed from $D_{i}^2 = (T_{i}^* - \overline{T})^T [S_{T}^*]^{-1} (T_{i}^* - \overline{T})$ 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 $T_{i}^*$).

$$
\{ w : (w - T_n)^T [S_{T}^*]^{-1} (w - T_n) \leq D_{2(U_B,T)}^2 \} = \{ w : D_w^2(T_n, S_{T}^*) \leq D_{2(U_B,T)}^2 \}
$$

(5)

where the cutoff $D_{2(U_B,T)}^2$ is the 100$q_{0.025}$th sample percentile of the $D_{i}^2 = (T_{i}^* - T_n)^T [S_{T}^*]^{-1} (T_{i}^* - T_n)$, and

$$
\{ w : (w - T_n)^T [S_{T}^*]^{-1} (w - T_n) \leq D_{2(U_B)}^2 \} = \{ w : D_w^2(T_n, S_{T}^*) \leq D_{2(U_B)}^2 \}.
$$

(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 (2019) gave simpler proofs and proved that (2) is a large sample CI. Assume $u_n \overset{D}{\rightarrow} u$ where $u_n = \sqrt{n}(T_{i}^* - T_n), \sqrt{n}(T_{i}^* - \overline{T}), \sqrt{n}(T_n - \theta)$, or $\sqrt{n}(\overline{T} - \theta)$, and $nS_{T}^* \overset{P}{\rightarrow} C$ where $C$ is nonsingular. Let

$$
D_{2}^2 = D_{2(T)}^2(\overline{T}, S_{T}^*) = \sqrt{n}(T_{i}^* - \overline{T})^T (nS_{T}^*)^{-1} \sqrt{n}(T_{i}^* - \overline{T}),
$$

$$
D_{2}^2 = D_{2(\theta)}^2(T_n, S_{T}^*) = \sqrt{n}(T_n - \theta)^T (nS_{T}^*)^{-1} \sqrt{n}(T_n - \theta),
$$

$$
D_{3}^2 = D_{3(\theta)}^2(\overline{T}, S_{T}^*) = \sqrt{n}(\overline{T} - \theta)^T (nS_{T}^*)^{-1} \sqrt{n}(\overline{T} - \theta),
$$

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 $D_{2}^2 \approx u^T (nS_{T}^*)^{-1} u \approx u^T C^{-1} u$, and the percentiles of $D_{2}^2$ 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

$$
\frac{|S_{T}^*|^{1/2}}{|S_{T}^*|^{1/2}} \left( \frac{D_{2(U_B)}}{D_{2(U_B,T)}} \right)^g = \left( \frac{D_{2(U_B)}}{D_{2(U_B,T)}} \right)^g.
$$

(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 $\hat{\beta}_{MIX}$ 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

Theorems 1 and 3 in this section generalize the Pelawa Watagoda and Olive (2019, 2020) theory for multiple linear regression to many other models. 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 \( \sqrt{n}(\beta_{I_j} - \beta) \overset{D}{\rightarrow} N_{a_j}(0, V_j) \) where \( V_j \) is the covariance matrix of the asymptotic multivariate normal distribution. Then

\[
\sqrt{n}(\beta_{I_j,0} - \beta) \overset{D}{\rightarrow} N_p(0, V_{j,0})
\]

where \( V_{j,0} \) adds columns and rows of zeros corresponding to the \( x_i \) not in \( I_j \), and \( V_{j,0} \) is singular unless \( I_j \) corresponds to the full model.

The first assumption in Theorem 1 is \( P(S \subseteq I_{\text{min}}) \rightarrow 1 \) as \( n \rightarrow \infty \). Then the variable selection estimator corresponding to \( I_{\text{min}} \) 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). If a shrinkage estimator that does variable selection is a consistent estimator of \( \beta \), then \( P(S \subseteq I_{\text{min}}) \rightarrow 1 \) as \( n \rightarrow \infty \). Hence Theorem 1c) proves that the lasso variable selection and elastic net variable selection estimator corresponding to \( u \) is a consistent estimator by (8) since \( S \subseteq I_j \) for each \( \pi_j \), and since \( \beta_{\text{MIX}} \) uses random selection.

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

a) Then

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

where the cdf of \( u \) is \( F_u(t) = \sum_j \pi_j F_{u_j}(t) \). Thus \( u \) has 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

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

where \( v \) has a mixture distribution of the \( v_j = A\hat{\beta}_{I_{\text{min}}} \sim N_g(0, AV_{j,0}A^T) \) with probabilities \( \pi_j \).

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

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

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

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

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 Lemma 1 of \( \text{P"otscher} \) (1991).

The following subscript notation is useful. Subscripts before the \( MIX \) are used for subsets of \( \hat{\beta}_{MIX} = (\hat{\beta}_1, ..., \hat{\beta}_p)^T \). Let \( \hat{\beta}_{i,MIX} = \hat{\beta}_i \). Similarly, if \( I = \{i_1, ..., i_a\} \), then \( \hat{\beta}_{I,MIX} = (\hat{\beta}_{i_1}, ..., \hat{\beta}_{i_a})^T \). Subscripts after \( MIX \) denote the \( i \)th vector from a sample \( \hat{\beta}_{MIX,i,1}, ..., \hat{\beta}_{MIX,i,B} \). 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 \( \beta = \beta_{FULL} \).

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 \( \pi_j \), \( Au_j \sim \mathcal{N}(0, \mathbf{A}\Sigma\mathbf{A}^T) \). Then \( \sqrt{n}(\mathbf{A}^\hat{\beta}_{MIX} - \mathbf{A}\beta) \overset{D}{\rightarrow} \mathbf{Au} \sim \mathcal{N}(0, \mathbf{A}\Sigma\mathbf{A}^T) \). This special case occurs for \( \hat{\beta}_{S,MIX} \) if \( \sqrt{n}(\hat{\beta} - \beta) \overset{D}{\rightarrow} \mathcal{N}(0, \mathbf{V}) \) where the asymptotic covariance matrix \( \mathbf{V} \) is diagonal and nonsingular. Then \( \hat{\beta}_{S,MIX} \) and \( \hat{\beta}_{S,\text{FULL}} \) have the same multivariate normal limiting distribution. For several criteria, this result should hold for \( \hat{\beta}_{VS} \) since asymptotically, \( \sqrt{n}(\mathbf{A}^\hat{\beta}_{VS} - \mathbf{A}\beta) \) is selecting from the \( \mathbf{Au} \) which have the same distribution. Then the confidence regions applied to \( \mathbf{A}^\star_{SEL} = \mathbf{B}\hat{\beta}_{S,SEL}^\star \) should have similar volume and cutoffs where \( \text{SEL} \) is \( MIX, VS \), or \( \text{FULL} \).

Theorem 1 can be used to justify prediction intervals after variable selection. See \( \text{Olive et al.} \) (2020). Theorem 1d) is useful for \textit{variable selection consistency} and the \textit{oracle property} where \( \pi_d = \pi_S = 1 \) if \( P(I_{\text{min}} = S) \rightarrow 1 \) as \( n \rightarrow \infty \). See \( \text{Claeskens and Hjort} \) (2008) and \( \text{Fan and Li} \) (2001) for references. A necessary condition for \( P(I_{\text{min}} = S) \rightarrow 1 \) 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 \( \text{Wieczorek} \) (2018) for forward selection and \( \text{Hastie et al.} \) (2015) for lasso, where the predictors need a “near orthogonality” condition.

The following \( \text{Pelawa Watagoda and Olive} \) (2019) theorem is useful for bootstrapping variable selection estimators. Let \( (\overline{T}, S_T) \) 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 \( D^2_{(U|u)} \) be the cutoff computed from the \( D^2_{i}(\overline{T}, S_T) \) for \( i = 1, ..., B \). The hyperellipsoids corresponding to \( D^2(\overline{T}_n, C) \) and \( D^2(\overline{T}, C) \) are centered at \( T_n \) and \( \overline{T} \), respectively. Note that \( D^2_{\overline{T}}(T_n, C) = D^2_{\overline{T}_n}(T, C) \). Thus \( D^2_{\overline{T}}(T_n, C) \leq D^2_{\overline{T}_n}(T, C) \), iff \( D^2_{\overline{T}_n}(T, C) \leq D^2_{\overline{T}}(T_n, C) \).

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 \( (nS_T)^{-1} \) to be fairly well behaved, e.g. \( (nS_T)^{-1} \overset{P}{\rightarrow} \Sigma_A^{-1} \).

**Theorem 2: Geometric Argument.** Suppose \( \sqrt{n}(T_n - \theta) \overset{D}{\rightarrow} u \) with \( E(u) = 0 \) and \( \text{Cov}(u) = \Sigma_u \neq 0 \). Assume \( T_1, ..., T_B \) are iid with nonsingular covariance matrix \( \Sigma_{T_n} \), where \( (nS_T)^{-1} \overset{P}{\rightarrow} \Sigma_A^{-1} \). Then the large sample \( 100(1 - \delta)\% \) prediction region \( R_P = \{u : D^2_u(\overline{T}, S_T) \leq D^2_{(U|u)}\} \) 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 \( B \rightarrow \infty \). Hence the region \( R_c = \{u : D^2_u(T_n, S_T) \leq D^2_{(U|u)}\} \) is a large sample \( 100(1 - \delta)\% \) confidence region for \( \theta \) where \( T_n \) is a randomly selected \( T_i \).
Hence the mixture distribution of the \( \hat{\mathbf{w}} \) is possible) that is symmetric about \( \mathbf{w} \). Here for the maximum likelihood estimator (MLE) with AIC, and gave a forward selection example. Here \( \mathbf{w}_j \) is a multivariate truncated normal distribution (where no truncation is possible) that is symmetric about \( \mathbf{0} \). Note that both \( \sqrt{n}(\hat{\beta}_{MLX} - \beta) \) and \( \sqrt{n}(\hat{\beta}_{VS} - \beta) \) are selecting from the \( u_{kn} = \sqrt{n}(\hat{\beta}_{I_k,0} - \beta) \) and asymptotically from the \( u_j \). The random

\[
\sqrt{n}(T - \theta) \overset{D}{\rightarrow} \frac{1}{B} \sum_{i=1}^{B} \mathbf{v}_i \sim AN_g \left( \mathbf{0}, \frac{\Sigma_u}{B} \right)
\]

where \( AN_g \) denotes an approximate multivariate normal distribution. Hence \( (\mathbf{T} - \theta) = O_P((nB)^{-1/2}) \), and \( \mathbf{T} \) gets arbitrarily close to \( \theta \) compared to \( T_n \) as \( B \to \infty \). Thus \( R_c \) is a large sample 100(1 – \( \delta \))% confidence region for \( \theta \) as \( n, B \to \infty \). □

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 \( \sqrt{n}(T^*_n - T_n) \) both converge in distribution to \( \mathbf{u} \sim N_g(\mathbf{0}, \Sigma_A) \), 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 \( T_0 \). 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 \( \sqrt{n}(\mathbf{T}^* - T_n) \overset{P}{\rightarrow} \mathbf{0} \).

For \( T_n = A\hat{\beta}_{MLX} \) with \( \theta = A\beta \), we have \( \sqrt{n}(T_n - \theta) \overset{D}{\rightarrow} \mathbf{v} \) by (10) where \( E(\mathbf{v}) = \mathbf{0} \), and \( \Sigma_\mathbf{v} = \sum_j \pi_j A\Sigma_{\beta_j,0}A^T \). By Theorem 2, if we had iid data \( T_1, ..., T_B \), then \( R_c \) would be a large sample confidence region for \( \theta \). If \( \sqrt{n}(T_n^* - T_n) \overset{D}{\rightarrow} \mathbf{v} \), then we could use the bootstrap sample and confidence regions (4) to (6). This condition holds only under strong regularity conditions such as \( \pi_\delta = 1 \) 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 \( \hat{\beta}_{VS}(\hat{\beta}_{VS} = \hat{\beta}_{I_k,0}) \) to find the distribution of \( \mathbf{w}_n = \sqrt{n}(\hat{\beta}_{VS} - \beta) \). 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 \( \hat{\beta}_{I_k,0}(\hat{\beta}_{VS} = \hat{\beta}_{I_k,0}) \). Let \( \mathbf{w}_{kn} = \sqrt{n}(\hat{\beta}_{I_k,0} - \beta) \sim \sqrt{n}(\hat{\beta}_{I_k,0} - \beta) \). Denote \( F_\mathbf{w}(t) = P(z_1 \leq t_1, ..., z_p \leq t_p) \) by \( P(z \leq t) \). Then Pötscher (1991) and Pelawa Watagoda and Olive (2020) show

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

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

Charkhi and Claeskens (2018) showed that \( \mathbf{w}_j = \sqrt{n}(\hat{\beta}_{I_j,0} - \beta) \overset{D}{\rightarrow} \mathbf{w}_j \) if \( S \subseteq I_j \) for the maximum likelihood estimator (MLE) with AIC, and gave a forward selection example. Here \( \mathbf{w}_j \) is a multivariate truncated normal distribution (where no truncation is possible) that is symmetric about \( \mathbf{0} \). Note that both \( \sqrt{n}(\hat{\beta}_{MLX} - \beta) \) and \( \sqrt{n}(\hat{\beta}_{VS} - \beta) \) are selecting from the \( u_{kn} = \sqrt{n}(\hat{\beta}_{I_k,0} - \beta) \) 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 $u_{jn}$ and $u_j$ to that of $w_{jn}$ and $w_j$. The assumption that $w_{jn}$ and $w_j$ may not be mild. The proof for Equation (11) is the same as that for (9).

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 $\pi_k$ where $\pi_k \to \pi_k$ as $n \to \infty$. Denote the positive $\pi_k$ by $\pi_j$. Assume $w_{jn} = \sqrt{n}(\beta_{I_{k,0}} - \beta) \xrightarrow{D} w_j$. Then

$$w_n = \sqrt{n}(\beta_{VS} - \beta) \xrightarrow{D} w$$

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

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. Instead of computing $\hat{\beta}^*$, compute the variable selection estimator $\hat{\beta}_{VS,1}^* = \hat{\beta}_{I_{k,0}}^*$. Then generate another $Y^*$ and $X^*$ and compute $\hat{\beta}_{MIX,1}^* = \hat{\beta}_{I_{k,0}}^*$ (using the same subset $I_{k,1}$). This process is repeated $B$ times to get the two bootstrap samples for $i = 1, ..., B$. Let $B_{jn}$ 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

$$T_1^*, ..., T_B^*, T_{1,j}, ..., \hat{T}_{B_{jn},j}.$$ 

Denote $T_{1,j}, ..., T_{B_{jn},j}$ as the $j$th bootstrap component of the bootstrap sample with sample mean $\overline{T}_j$ and sample covariance matrix $S_{T,j}$. Similarly, we can define the $j$th component of the iid sample $T_1, ..., T_B$ to have sample mean $\overline{T}_j$ and sample covariance matrix $S_{T,j}$.

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

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

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 $\beta$, then the bootstrap cloud would be like an iid data cloud shifted to be centered at $\beta$, and (5) 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 (5) having slightly higher asymptotic coverage than applying (5) to the iid data cloud. Also, (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 $\hat{\theta}$ 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 $T_n = A\hat{\beta}_{ML}, T_{ij} = A\hat{\beta}_{I,i,0}$, 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 $B \geq 50p$. If the regression model has a $q \times 1$ vector of parameters $\gamma$, we may need to replace $p$ by $p + q$.

The matrix $S_T^*$ 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 $\hat{\beta}^* = \hat{\beta}^*_\text{FULL}$ 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 $S_T^*$ 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.

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 $T_1, \ldots, T_B$, and ii) zero padding.

The nonparametric bootstrap (also called the empirical bootstrap, naive bootstrap, and the pairs bootstrap) draws a sample of $n$ 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 $\hat{\beta}_{Y,V,1}^*$, and then draws another sample to get $\hat{\beta}_{MIX,1}^*$. This process is repeated $B$ times to get the two bootstrap samples for $i = 1, \ldots, B$. 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).

For the parametric regression model $Y_i|x_i \sim D(x_i^T \beta, \gamma)$, assume $\sqrt{n}(\hat{\beta} - \beta) \sim N_p(0, V(\beta))$, and that $V(\hat{\beta}) \sim V(\beta)$ as $n \to \infty$. These assumptions tend to be mild for a parametric regression model where the MLE $\hat{\beta}$ is used. Then $V(\beta) = I^{-1}(\beta)$, the inverse Fisher information matrix. For GLMs, see, for example, Sen and Singer (1993). For the parametric regression model, we regress $Y$ on $X$ to obtain $(\hat{\beta}, \hat{\gamma})$ 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 $Y_j^* = (Y_j^*)$ where $Y_i^*|x_i \sim D(x_i^T \hat{\beta}, \hat{\gamma})$ for $i = 1, \ldots, n$. Regress $Y_j^*$ on $X$ to get $\hat{\beta}_j^*$ for $j = 1, \ldots, B$. The large sample theory for $\hat{\beta}_j^*$ is simple. Note that if $Y_i^*|x_i \sim D(x_i^T \beta, \gamma)$ where $\beta$ does not depend on $n$, then $(Y^*, X)$ follows the parametric regression model with parameters $(\beta, \gamma)$. Hence $\sqrt{n}(\hat{\beta}^* - \beta) \sim N_p(0, V(\beta))$. 
Now fix large integer \( n_0 \), and let \( b = \hat{\beta}_{n_0} \). Then \( \sqrt{n}(\hat{\beta}^* - \hat{\beta}_{n_0}) \overset{D}{\to} N_p(0, V(\hat{\beta}_{n_0})) \). Since \( N_p(0, V(\beta)) \overset{D}{\to} N_p(0, V(\beta)) \), we have

\[
\sqrt{n}(\hat{\beta}^* - \beta) \overset{D}{\to} N_p(0, V(\beta)) \tag{13}
\]
as \( n \to \infty \).

Now suppose \( S \subseteq I \). Without loss of generality, let \( \beta = (\beta_1^T, \beta_0^T)^T \), where \( \beta_1, \gamma \). Hence \( \sqrt{n}(\bar{\beta}_I - \beta) \overset{D}{\to} N_{p_1}(0, V(\beta_I)) \). Now \( (Y^*, X_I) \) only follows the parametric regression model asymptotically, since \( \gamma \neq 0 \). However, under regularity conditions, \( E(\hat{\beta}_I) \approx \beta_I \) and \( \text{Cov}(\hat{\beta}_I) \approx \text{Cov}(\beta_I) \to 0 \) as \( n, B \to \infty \). See the following example.

For the multiple linear regression model, \( Y = X\beta + e \), 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 \( H = X(X^TX)^{-1}X^T \).

For each \( I \) with \( S \subseteq I \), assume the maximum leverage \( \max_{i=1,\ldots,n} x_{it}^T(X_I^T X_I)^{-1}x_{it} \to 0 \) in probability as \( n \to \infty \). For least squares (OLS) with \( S \subseteq I \), \( \sqrt{n}(\bar{\beta}_I - \beta_I) \overset{D}{\to} N_{p_1}(0, V_I) \) where \( (X_I^T X_I)/(n\sigma^2) \overset{P}{\to} V_I^{-1} \). See, for example, Sen and Singer (1993).

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

\[
\hat{\sigma}^2_n = MSE = \frac{1}{n-p} \sum_{i=1}^{n} r_i^2
\]
where the residuals are from the full OLS model. Then \( MSE \) is a \( \sqrt{n} \) consistent estimator of \( \sigma^2 \) under mild conditions by Su and Cook (2012). Thus \( \hat{\beta}_I^* = (X_I^T X_I)^{-1}X_I^T Y^* \overset{D}{\sim} N_{p_1}(\bar{\beta}_I, \hat{\sigma}^2_n (X_I^T X_I)^{-1}) \) since \( E(\beta^*_I) = (X_I^T X_I)^{-1}X_I^T HY = \beta_I \) because \( HX_I = X_I \), and \( \text{Cov}(\hat{\beta}_I) = \hat{\sigma}^2_n (X_I^T X_I)^{-1} \). Hence

\[
\sqrt{n}(\hat{\beta}_I^* - \beta_I) \overset{D}{\sim} N_{p_1}(0, (X_I^T X_I)^{-1}) \overset{D}{\to} N_{p_1}(0, V_I)
\]
as \( n, B \to \infty \) if \( S \subseteq I \).

The bootstrap component clouds for \( \hat{\beta}_{VS}^* \) are again separated compared to the iid clouds for \( \hat{\beta}_{VS} \), which are centered about \( \beta \). Heuristically, most of the selection bias is due to predictors in \( E \), not to the predictors in \( S \). Hence \( \hat{\beta}_{VS}^* \) is roughly similar to \( \hat{\beta}_{MIX}^* \). Typically the distributions of \( \hat{\beta}_{E,VS}^* \) and \( \hat{\beta}_{E,MIX}^* \) are not similar, but use the same zero padding. In simulations, confidence regions for \( \hat{\beta}_{VS}^* \) tended to have less undercoverage than confidence regions for \( \hat{\beta}_{MIX}^* \).

### 4. Example and simulations

**Example.** Lindenmayer et al. (1991) and Cook and Weisberg (1999) 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,
Table 1. Shorth CIs for the example.

<table>
<thead>
<tr>
<th>variable</th>
<th>$\beta_i$</th>
<th>95% shorth CI: VS</th>
<th>95% shorth CI: MIX</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>-0.8994</td>
<td>[-1.5662, -0.5169]</td>
<td>[-1.3680, -0.3553]</td>
</tr>
<tr>
<td>acacia</td>
<td>0</td>
<td>[0, 0.0.0384]</td>
<td>[-0.0004, 0.0397]</td>
</tr>
<tr>
<td>bark</td>
<td>0.0336</td>
<td>[0, 0.05928]</td>
<td>[0, 0.0563]</td>
</tr>
<tr>
<td>habitat</td>
<td>0.1069</td>
<td>[0, 0.1524]</td>
<td>[0, 0.1584]</td>
</tr>
<tr>
<td>shrubs</td>
<td>0</td>
<td>[0, 0.05582]</td>
<td>[-0.01560, 0.04532]</td>
</tr>
<tr>
<td>stags</td>
<td>0.0302</td>
<td>[0, 0.0540]</td>
<td>[0, 0.0540]</td>
</tr>
<tr>
<td>stumps</td>
<td>0</td>
<td>[-0.9326, 0.0000]</td>
<td>[-0.8402, 0.1515]</td>
</tr>
</tbody>
</table>

$\psi$=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 $\approx \beta_i \pm 2SE(\hat{\beta}_i)$ (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 $\beta_i$ 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.

We tested $H_0: \beta_2 = \beta_3 = \beta_4 = 0$ 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 $H_0: A\beta = (\beta_2, \beta_3, \beta_4)^T = 0$. Using the prediction region method with the full model had $[0, D_{(U\beta)}] = [0, 2.773]$ with $D_0 = 2.067$. 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 $[0, D_{(U\beta)}] = [0, 2.702]$ while $D_0 = 1.327$. 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 $S^*$ 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 $\hat{\beta}_MIX$ had $[0, D_{(U\beta)}] = [0, 3.157]$ while $D_0 = 1.066$. 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. We used 5000 runs, $\theta = A\beta = \beta_i, \theta = A\beta = \beta_S = (\beta_1, 1, ..., 1)^T$ 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 often used $p = 4$ since the simulations with 5000 runs take a long time. A larger simulation study for $\hat{\beta}_{VS}$ 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 $i = 1, ..., n$, 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 $\sigma_{ii} = [1 + (q - 1)\psi^2]$ and the off diagonal entries $\sigma_{ij} = [2\psi + (q - 2)\psi^2]$. Hence the correlations are $\text{cov}(z_i, z_j) = \rho = (2\psi + (q -
the programs to fail for
resulted in coverages closer to 0.95 for the full model, and convergence problems caused
The cutoff will often be near $H(5)$. The 0 indicates the test was
are for the prediction region method (4), hybrid region (6), and Bickel and Ren region
selection. The last six columns give results for the tests. The terms pr, hyb, and br
the full model regression, the term “vs” is for variable selection, and “mix” for random
interval coverages and average lengths of the confidence intervals. The term “reg” is for
the nonparametric bootstrap with lasso variable selection.

For binary regression, the bootstrap confidence regions using smaller
$g$, $h$, and $i$ which 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 $\max(|x_i^T \beta|) \leq 7$ and if the
Table 2. Bootstrapping OLS Forward Selection with $C_p, c_i \sim EXP(1) - 1$

<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>0.939</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>
<td></td>
</tr>
<tr>
<td>len</td>
<td>0.393</td>
<td>0.400</td>
<td>0.399</td>
<td>0.400</td>
<td>2.474</td>
<td>2.474</td>
<td>2.475</td>
<td>2.453</td>
<td>2.453</td>
<td></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></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></td>
</tr>
<tr>
<td>mix,0</td>
<td>0.937</td>
<td>0.943</td>
<td>0.999</td>
<td>0.998</td>
<td>0.987</td>
<td>0.995</td>
<td>0.930</td>
<td>0.931</td>
<td>0.931</td>
<td></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></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></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></td>
</tr>
<tr>
<td>vs,0.9</td>
<td>0.941</td>
<td>0.966</td>
<td>0.996</td>
<td>0.997</td>
<td>0.992</td>
<td>0.983</td>
<td>0.994</td>
<td>0.957</td>
<td>0.953</td>
<td></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></td>
</tr>
<tr>
<td>mix,0.9</td>
<td>0.941</td>
<td>0.972</td>
<td>0.997</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></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></td>
</tr>
</tbody>
</table>

$Y$ values of 0 and 1 are not nearly perfectly classified by the rule $\hat{Y} = 1$ if $x_i^T \hat{\beta} > 0$ 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 $\hat{\beta}_{MIX}$ with $p = 10$ and $\psi = 0.9$, the coverages for $H_0: \beta_S = 1$ were about 4% too low in Table 4. One of the ten short confidence intervals also had coverage about 2% too low for this case.

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, e.g. $\mathbf{I}^{-1}(\beta)$, is diagonal. Hence $\hat{\beta}_S$ has the same multivariate normal limiting distribution for $\hat{\beta}_{MIX}$ and the full model $\hat{\beta}$, and possibly for $\hat{\beta}_{VS}$, by Section 2. For Tables 2-5, $\beta_S = (\beta_1, \beta_2)^T$, and $\beta_{p-1}$ and $\beta_p$ are components of $\beta_E$. 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 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 $n = 100$. Larger simulations, with $p$ as large as 10 and without the MIX rows, are in Pelawa Watagoda and Olive (2019).

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

For Cox proportional hazards regression, the cases were $(Z_i, \delta_i, x_i)$ where $Z_i = Y_i$ 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
Table 3. Bootstrapping Binomial Logistic Regression, Backward Elimination with AIC, $B = 200$, $n = 200$, $p = 4$, $k = 1$, and $m = 1$

<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>0.950</td>
<td>0.944</td>
<td>0.955</td>
<td>0.954</td>
<td>0.958</td>
<td>0.966</td>
<td>0.958</td>
<td>0.966</td>
<td>0.973</td>
<td></td>
</tr>
<tr>
<td>len</td>
<td>0.754</td>
<td>0.677</td>
<td>0.458</td>
<td>0.459</td>
<td>2.488</td>
<td>2.488</td>
<td>2.499</td>
<td>2.485</td>
<td>2.575</td>
<td></td>
</tr>
<tr>
<td>vs, 0</td>
<td>0.954</td>
<td>0.948</td>
<td>0.998</td>
<td>0.998</td>
<td>0.996</td>
<td>0.993</td>
<td>0.997</td>
<td>0.962</td>
<td>0.976</td>
<td></td>
</tr>
<tr>
<td>len</td>
<td>0.750</td>
<td>0.675</td>
<td>0.393</td>
<td>0.390</td>
<td>2.725</td>
<td>2.725</td>
<td>3.031</td>
<td>2.482</td>
<td>2.575</td>
<td></td>
</tr>
<tr>
<td>mix, 0</td>
<td>0.956</td>
<td>0.947</td>
<td>0.999</td>
<td>0.998</td>
<td>0.998</td>
<td>0.994</td>
<td>0.998</td>
<td>0.962</td>
<td>0.976</td>
<td></td>
</tr>
<tr>
<td>len</td>
<td>0.740</td>
<td>0.663</td>
<td>0.321</td>
<td>0.322</td>
<td>3.129</td>
<td>3.129</td>
<td>3.341</td>
<td>2.482</td>
<td>2.548</td>
<td></td>
</tr>
<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>0.964</td>
<td>0.966</td>
<td>0.950</td>
<td>0.963</td>
<td></td>
</tr>
<tr>
<td>len</td>
<td>0.755</td>
<td>0.684</td>
<td>0.606</td>
<td>0.608</td>
<td>2.489</td>
<td>2.489</td>
<td>2.499</td>
<td>2.486</td>
<td>2.497</td>
<td></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></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>2.723</td>
<td></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>0.978</td>
<td></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>2.707</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. 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</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>2.474</td>
<td></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>0.958</td>
<td></td>
</tr>
<tr>
<td>len</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>2.499</td>
<td></td>
</tr>
<tr>
<td>mix, 0</td>
<td>0.948</td>
<td>0.953</td>
<td>0.999</td>
<td>1−</td>
<td>1.000</td>
<td>1−</td>
<td>1−</td>
<td>0.951</td>
<td>0.949</td>
<td>0.951</td>
</tr>
<tr>
<td>len</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>2.475</td>
<td></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>0.946</td>
<td>0.948</td>
<td>0.945</td>
<td>0.945</td>
<td></td>
</tr>
<tr>
<td>len</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>2.469</td>
<td></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>0.973</td>
<td></td>
</tr>
<tr>
<td>len</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>2.638</td>
<td></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></td>
</tr>
<tr>
<td>len</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>2.711</td>
<td></td>
</tr>
</tbody>
</table>
Table 5. 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</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</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</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</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</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</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>

for the predictors were similar to those for the Poisson and binomial regression, but no constant was used so replace $q$ by $p$. Then $SP = x_i^{T}\beta = x_{i,1} + \cdots + x_{i,k} \sim N(0, a^2)$ 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. See Table 5.

5. Conclusion

The random vector $\hat{\beta}_{\text{MIX}}$ has simple large sample theory, and is useful for understanding the more complicated large sample theory of $\hat{\beta}_{\text{VS}}$. 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 would be useful for why confidence regions (4) and (5) simulate well for so many variable selection estimators.

There is a massive literature on variable selection and a fairly large literature for inference after variable selection. Data splitting is useful. See, for example, Leeb et al. (2015), Lu et al. (2017), Ning and Liu (2017), Pötscher (1991), Rinaldo et al. (2019), and Tibshirani et al. (2018). Su (2018) shows that fast variable selection methods tend to select spurious variables quickly if $k$ is not small.

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 (1997) 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 (2020) $R$ 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 \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 function \texttt{RLPHboot2} bootstraps a Cox PH model with lasso variable selection. The function \texttt{RLPHbootsim2} is used to simulate the bootstrap for lasso variable selection with Cox regression.

\section*{References}


