Inference For Multiple Linear Regression After Model or Variable Selection

Lasanthi C. R. Pelawa Watagoda and David J. Olive*
Southern Illinois University

August 16, 2017

Abstract

This paper presents inference for the multiple linear regression model $Y = \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$ after model or variable selection, including testing hypotheses with the bootstrap. If $n$ is the sample size, prediction intervals are developed that may be useful even if $p > n$.

KEY WORDS: Bootstrap; Forward Selection; Lasso; Partial Least Squares; Prediction Interval; Principal Components Regression; Relaxed Lasso; Ridge Regression.

*David J. Olive is Professor, Lasanthi C. R. Pelawa Watagoda is Ph.D. student, Department of Mathematics, Southern Illinois University, Carbondale, IL 62901, USA.
1 INTRODUCTION

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

$$Y_i = \beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \mathbf{x}_i^T\beta + e_i$$  \hfill (1.1)

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

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e},$$  \hfill (1.2)

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

It is often convenient to use the centered response $\mathbf{Z} = \mathbf{Y} - \bar{Y}$ where $\bar{Y} = \bar{Y}_1$, and the $n \times (p - 1)$ matrix of standardized nontrivial predictors $\mathbf{W} = (W_{ij})$. For $j = 1, \ldots, p - 1$, let $W_{ij}$ denote the $(j + 1)$th variable standardized so that $\sum_{i=1}^n W_{ij} = 0$ and $\sum_{i=1}^n W_{ij}^2 = n$. Note that the sample correlation matrix of the nontrivial predictors $u_i$ is $\mathbf{R}_u = \mathbf{W}^T\mathbf{W}/n$. Then regression through the origin is used for the model

$$\mathbf{Z} = \mathbf{W}\eta + \mathbf{e}$$  \hfill (1.3)

where the vector of fitted values $\hat{\mathbf{Y}} = \bar{Y} + \hat{\mathbf{Z}}$.

There are many methods for estimating $\beta$, including forward selection with OLS, principal component regression (PCR), partial least squares (PLS) due to Wold (1975), lasso due to Tibshirani (1996), and ridge regression (RR): see Hoerl and Kennard (1970). We also used the variant of relaxed lasso that applies OLS to a constant and the predictors that had nonzero lasso coefficients, which is the LARS-OLS hybrid estimator of Efron, Hastie, Johnstone, and Tibshirani (2004), also called the relaxed lasso ($\phi = 0$) estimator by Meinshausen (2007).

These six methods produce $M$ models and use a criterion to select the final model (e.g. $C_p$ or 10-fold cross validation (CV)). The number of models $M$ depends on the method. Lasso and ridge regression have a parameter $\lambda$. When $\lambda = 0$, the OLS full model is used. These methods also use a maximum value $\lambda_M$ of $\lambda$ and a grid of $M$ $\lambda$ values $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_{M-1} < \lambda_M$ where often $\lambda_1 = 0$. For lasso, $\lambda_M$ is the smallest value of $\lambda$ such that $\hat{\eta}_{\lambda_M} = 0$. Hence $\hat{\eta}_{\lambda_i} \neq 0$ for $i < M$. For forward selection, PCR, and PLS, $M \leq p$. See James, Witten, Hastie, and Tibshirani (2013, ch. 6).

Consider choosing $\hat{\eta}$ to minimize the criterion

$$Q(\eta) = \frac{1}{a}(\mathbf{Z} - \mathbf{W}\eta)^T(\mathbf{Z} - \mathbf{W}\eta) + \frac{\lambda_{1,n}}{a} \sum_{i=1}^{p-1} |\eta_i|^j$$  \hfill (1.4)

where $\lambda_{1,n} \geq 0$, $a > 0$, and $j > 0$ are known constants. Then $j = 2$ corresponds to ridge regression, $j = 1$ corresponds to lasso, and $a = 1, 2, n$, and $2n$ are common. The residual
sum of squares $RSS(\eta) = (Z - W\eta)^T(Z - W\eta)$, and $\lambda_{1,n} = 0$ corresponds to the OLS estimator $\hat{\eta}_{OLS} = (W^TW)^{-1}W^TZ$.

In the following three paragraphs, assume $p$ is fixed. Knight and Fu (2000) prove that i) $\hat{\eta}$ is a consistent estimator of $\eta$ if $\lambda_{1,n} = o(n)$ so $\lambda_{1,n}/n \to 0$ as $n \to \infty$, ii) $\hat{\eta}_{OLS}$ and $\hat{\eta}$ are asymptotically equivalent if $\lambda_{1,n} \to \infty$ too slowly as $n \to \infty$, iii) $\hat{\eta}$ is a $\sqrt{n}$ consistent estimator of $\eta$ if $\lambda_{1,n} = O(\sqrt{n})$ (so $\lambda_{1,n}/\sqrt{n}$ is bounded), and iv) if $\lambda_{1,n}/\sqrt{n} \to \tau \geq 0$, then for ridge regression

$$\sqrt{n}(\hat{\eta}_{RR} - \eta) \overset{D}{\to} N_{p-1}(-\tau V\eta, \sigma^2 V)$$

where

$$Ru = \frac{W^TW}{n} \overset{P}{\to} V^{-1}$$

as $n \to \infty$. If $\tau = 0$, then OLS and ridge regression have the same limiting distribution. Note that $V^{-1} = \rho_u$ if the $u_i$ are a random sample from a population with a nonsingular population correlation matrix $\rho_u$. Under (1.5), if $\lambda_{1,n}/n \to 0$ then

$$\frac{W^TW + \lambda_{1,n}I_{p-1}}{n} \overset{P}{\to} V^{-1}, \text{ and } n(W^TW + \lambda_{1,n}I_{p-1})^{-1} \overset{P}{\to} V.$$

The following identity from Gunst and Mason (1980, p. 342) is useful for ridge regression inference:

$$\hat{\eta}_{RR} = (W^TW + \lambda_{1,n}I_{p-1})^{-1}W^TZ = (W^TW + \lambda_{1,n}I_{p-1})^{-1}W^TW\hat{\eta}_{OLS} = A_n\hat{\eta}_{OLS}$$

$$= [I_{p-1} - \lambda_{1,n}(W^TW + \lambda_{1,n}I_{p-1})^{-1}]\hat{\eta}_{OLS} = B_n\hat{\eta}_{OLS}$$

since $A_n - B_n = 0$. If $\lambda_{1,n}/\sqrt{n} \to \tau \geq 0$, then

$$\sqrt{n}(\hat{\eta}_{RR} - \eta) = \sqrt{n}(\hat{\eta}_{RR} - \hat{\eta}_{OLS} + \hat{\eta}_{OLS} - \eta) =$$

$$\sqrt{n}(\hat{\eta}_{OLS} - \eta) - \sqrt{n}\frac{\lambda_{1,n}}{n}(W^TW + \lambda_{1,n}I_{p-1})^{-1}\hat{\eta}_{OLS}$$

$$\overset{D}{\to} N_{p-1}(0, \sigma^2 V) - \tau V\eta \sim N_{p-1}(-\tau V\eta, \sigma^2 V).$$

The following identity from Efron and Hastie (2016, p. 308), for example, is useful for inference for the lasso estimator $\hat{\eta}_L$:

$$-\frac{1}{n}W^T(Z - W\hat{\eta}_L) + \frac{\lambda_{1,n}}{2n}s_n = 0 \text{ or } -\frac{1}{n}W^T(Z - W\hat{\eta}_L) + \frac{\lambda_{1,n}}{2n}s_n = 0$$

where $s_m \in [-1, 1]$ and $s_m = \text{sign}(\hat{\eta}_{i,L})$ if $\hat{\eta}_{i,L} \neq 0$. Here $\text{sign}(\eta_i) = 1$ if $\eta_i > 1$ and $\text{sign}(\eta_i) = -1$ if $\eta_i < 1$. Note that $s_n = s_n\hat{\eta}_L$ depends on $\hat{\eta}_L$. Thus

$$\hat{\eta}_L = (W^TW)^{-1}W^TZ - n(W^TW)^{-1}\frac{\lambda_{1,n}}{2n}s_n.$$
\[
\sqrt{n}(\hat{\eta}_{\text{OLS}} - \eta) - \sqrt{n} \frac{\lambda_{1,n}}{2n} n(W^T W)^{-1} s_n \xrightarrow{P} N_{p-1}(0, \sigma^2 V) - \frac{\tau}{2} V s \sim N_{p-1} \left( \frac{-\tau}{2} V s, \sigma^2 V \right)
\]

If none of the elements of \( \eta \) are zero, and if \( \hat{\eta}_I \) is a consistent estimator of \( \eta \), then \( s_n \xrightarrow{P} s = s \eta \). If \( \lambda_{1,n}/\sqrt{n} \to 0 \), then OLS and lasso are asymptotically equivalent even if \( s_n \) does not converge to a vector \( s \) as \( n \to \infty \) since \( s_n \) is bounded.

The results in the above three paragraphs hold after model selection if \( \lambda_{1,n} \) is replaced by \( \hat{\lambda}_{1,n} \) and \( o \) and \( O \) are replaced by \( o_P \) and \( O_P \), e.g. \( \hat{\lambda}_{1,n} = o_P(\sqrt{n}) \) makes lasso or ridge regression asymptotically equivalent to OLS. For model selection, the \( M \) values of \( \lambda \) are denoted by \( \lambda_1, \lambda_2, ..., \lambda_M \) where \( \lambda_j = \lambda_{1,n,j} \) depends on \( n \) for \( j = 1, ..., M \). If \( \lambda_s \) corresponds to the model selected, then \( \hat{\lambda}_{1,n} = \lambda_s \).

Variable selection is the search for a subset of predictor variables that can be deleted without important loss of information. Following Olive and Hawkins (2005), a model for variable selection can be described by

\[
x^T \beta = x_S^T \beta_S + x_E^T \beta_E = x_S^T \beta_S
\]

where \( x = (x_S^T, x_E^T)^T \), \( x_S \) is a \( a_S \times 1 \) 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 \( x_j \) 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 (1.6) holds. Then

\[
x^T \beta = x_S^T \beta_S = x_S^T \beta_S + x_{I/S}^T \beta_{(I/S)} + x_O^T 0 = x_I^T \beta_I
\]

where \( x_{I/S} \) denotes the predictors in \( I \) that are not in \( S \). Since this is true regardless of the values of the predictors, \( \beta_O = 0 \) if \( S \subseteq I \).

Forward selection forms a sequence of submodels \( I_1, ..., I_M \) where \( I_j \) uses \( j \) predictors including the constant. Let \( I_1 \) use \( x_1^* = x_1 \equiv 1 \): the model has a constant but no nontrivial predictors. To form \( I_2 \), consider all models \( I \) with two predictors including \( x_1^* \). Compute \( Q_2(I) = \text{SSE}(I) = \text{RSS}(I) = r^T(I)r(I) = \sum_{i=1}^n r_i^2(I) = \sum_{i=1}^n(Y_i - \hat{Y}_i(I))^2 \). Let \( I_2 \) minimize \( Q_2(I) \) for the \( p - 1 \) models \( I \) that contain \( x_1^* \) and one other predictor. Denote the predictors in \( I_2 \) by \( x_1^*, x_2^* \). In general, to form \( I_j \) consider all models \( I \) with \( j \) predictors including variables \( x_1^*, ..., x_{j-1}^* \). Compute \( Q_j(I) = r^T(I)r(I) = \sum_{i=1}^n r_i^2(I) = \sum_{i=1}^n(Y_i - \hat{Y}_i(I))^2 \). Let \( I_j \) minimize \( Q_j(I) \) for the \( p - j + 1 \) models \( I \) that contain \( x_1^*, ..., x_{j-1}^* \) and one other predictor not already selected. Denote the predictors in \( I_j \) by \( x_1^*, ..., x_j^* \). Continue in this manner for \( j = 2, ..., M \). Often \( M = \min(\lfloor n/J \rfloor, p) \) for some integer \( J \) such as \( J = 5, 10, \) or \( 20 \). Here \( \lfloor x \rfloor \) is the smallest integer \( \geq x \), e.g., \( \lfloor 7.7 \rfloor = 8 \).

Consider the six methods forward selection with OLS, PCR, PLS, lasso, relaxed lasso, and ridge regression. When there is a sequence of \( M \) submodels, the final submodel \( I_d \) needs to be selected. Let the candidate model \( I \) contains \( a \) terms, including a constant. For example, let \( x_I \) and \( \hat{\beta}_I \) be \( a \times 1 \) vectors for the methods excluding PCR and PLS. Then there are many criteria used to select the final submodel \( I_d \). A simple method is to take the model that uses \( d = \min(\lfloor n/J \rfloor, p) \) variables. If \( p \) is fixed, the method will use the OLS full model once \( n/J \geq p \). For a given data set, \( p, n \) and \( \sigma^2 \) act as constants, and a criterion below may add a constant or be divided by a positive constant without changing the subset \( I_{\text{min}} \) that minimizes the criterion.
Let criteria $C_S(I)$ have the form

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

These criteria need a good estimator of $\sigma^2$. The criterion $C_p(I) = AIC_S(I)$ uses $K_n = 2$ while the $BIC_S(I)$ criterion uses $K_n = \log(n)$. Typically $\hat{\sigma}^2$ is the OLS full model

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

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

It is hard to get a good estimator of $\sigma^2$ when $n/p$ is not large. The following criterion are described in Burnham and Anderson (2004), but still need $n/p$ large. $AIC$ is due to Akaike (1973) and $BIC$ to Schwarz (1978).

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

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

Let $I_{\min}$ be the submodel that minimizes the criterion using variable selection with OLS. Following Seber and Lee (2003, p. 448) and Nishi (1984), the probability that model $I_{\min}$ from $C_p$ or $AIC$ underfits goes to zero as $n \to \infty$ with $p$ fixed. If $\hat{\beta}_I$ is $a \times 1$, form the $p \times 1$ vector $\hat{\beta}_{I,0}$ from $\hat{\beta}_I$ by adding 0s corresponding to the omitted variables. Since fewer than $2^p$ regression models $I$ contain the true model, and each such model gives a $\sqrt{n}$ consistent estimator $\hat{\beta}_{I,0}$ of $\beta$, the probability that $I_{\min}$ picks one of these models goes to one as $n \to \infty$. Hence $\hat{\beta}_{I_{\min},0}$ is a $\sqrt{n}$ consistent estimator of $\beta$ under model (1.6). Olive (2017a § 3.4.1, 2017c: § 5.3.4) showed that $\hat{\beta}_{I_{\min},0}$ is a consistent estimator.

The EBIC criterion given in Luo and Chen (2013) may work when $n/p$ is not large. Let $0 \leq \gamma \leq 1$ and $|I| = a \leq \min(n, p)$ if $\hat{\beta}_I$ is $a \times 1$. We may use $a \leq \min(n/5, p)$. Then

$$EBIC(I) = n \log \left( \frac{SSE(I)}{n} \right) + a \log(n) + 2\gamma \log \left( \frac{p}{a} \right) = BIC(I) + 2\gamma \log \left( \frac{p}{a} \right).$$

This criterion can give good results if $p = p_n = O(n^k)$ and $\gamma > 1 - 1/(2k)$. Hence we will use $\gamma = 1$.

The above criteria can be applied to forward selection and relaxed lasso. The $C_p$ criterion can also be applied to lasso. See Efron and Hastie (2016, pp. 221, 231).

Inference will consider prediction intervals and bootstrap hypothesis testing. Consider predicting a future test response variable $Y_f$ given a $p \times 1$ vector of predictors $x_f$ and training data $(x_1, Y_1), \ldots, (x_n, Y_n)$. A large sample $100(1 - \delta)\%$ prediction interval (PI) for $Y_f$ has the form $[\hat{L}_n, \hat{U}_n]$ where $P(\hat{L}_n \leq Y_f \leq \hat{U}_n) \to 1 - \delta$ as the sample size $n \to \infty$. 

5
The shorth($c$) estimator is useful for making prediction intervals. Let $Z_{(1)}, \ldots, Z_{(n)}$ be the order statistics of $Z_1, \ldots, Z_n$. Then let the shortest closed interval containing at least $c$ of the $Z_i$ be
\[
\text{shorth}(c) = [Z_{(s)}, Z_{(s+c-1)}].
\] (1.8)

Let
\[
k_n = \lfloor n(1 - \delta) \rfloor.
\] (1.9)

Frey (2013) showed that for large $n\delta$ and iid data, the shorth($k_n$) PI has maximum undercoverage $\approx 1.12\sqrt{\delta/n}$, and used the shorth($c$) estimator as the large sample $100(1 - \delta)%$ PI where
\[
c = \min(n, \lfloor n[1 - \delta + 1.12\sqrt{\delta/n}] \rfloor).
\] (1.10)

Section 2 will develop two prediction intervals that are useful after model selection. We also want to use bootstrap tests. Consider testing $H_0 : \mu = c$ versus $H_1 : \mu \neq c$ where $c$ is a known full rank $r \times 1$ vector. Given training data $z_1, \ldots, z_n$, a large sample $100(1 - \delta)%$ confidence region for $\mu$ is a set $A_n$ such that $P(\mu \in A_n) \to 1 - \delta$ as $n \to \infty$. Then reject $H_0$ if $c$ is not in the confidence region $A_n$. For model (1.1), let $\mu = A\beta$ where $A$ is a known full rank $r \times p$ matrix with $1 \leq r \leq p$.

To bootstrap a confidence region, Mahalanobis distances and prediction regions will be useful. Consider predicting a future test value $z_f$, given past training data $z_1, \ldots, z_n$, where the $z_i$ are $r \times 1$ random vectors. A large sample $100(1 - \delta)%$ prediction region is a set $A_n$ such that $P(z_f \in A_n) \to 1 - \delta$ as $n \to \infty$. Let the $r \times 1$ column vector $T$ be a multivariate location estimator, and let the $r \times r$ symmetric positive definite matrix $C$ be a dispersion estimator. Then the $i$th squared sample Mahalanobis distance is the scalar
\[
D_i^2 = D_i^2(T, C) = D_{z_i,T}(T, C) = (z_i - T)^TC^{-1}(z_i - T)
\] (1.11)
for each observation $z_i$. Notice that the Euclidean distance of $z_i$ from the estimate of center $T$ is $D_i(T, I_r)$ where $I_r$ is the $r \times r$ identity matrix. The classical Mahalanobis distance $D_i$ uses $(T, C) = (\overline{z}, S)$, the sample mean and sample covariance matrix where
\[
\overline{z} = \frac{1}{n}\sum_{i=1}^{n} z_i \quad \text{and} \quad S = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \overline{z})(z_i - \overline{z})^T.
\] (1.12)

Let $q_n$ and $c$ be given by (2.1) and (2.2) with $d$ replaced by $r$. Let $(T, C) = (\overline{z}, S)$, and let $D_{(U_n)}$ be the $100q_n$th sample quantile of the $D_i$. Then the Olive (2013) large sample $100(1 - \delta)%$ nonparametric prediction region for a future value $z_f$ given iid data $z_1, \ldots, z_n$ is
\[
\{z : D^2_{z_i}(\overline{z}, S) \leq D^2_{(U_n)}\},
\] (1.13)
while the classical large sample $100(1 - \delta)%$ prediction region is
\[
\{z : D^2_{z_i}(\overline{z}, S) \leq \chi^2_{r,1-\delta}\}.
\] (1.14)

The Olive (2017abce) prediction region method obtains a confidence region for $\mu$ by applying the nonparametric prediction region (1.13) to the bootstrap sample $T^*_1, \ldots, T^*_B$, and the theory for the method is sketched below. Let $\overline{T^*}$ and $S^*_T$ be the sample mean and
sample covariance matrix of the bootstrap sample. Assume \( \text{n} \mathbf{S}_T^* \overset{P}{\rightarrow} \mathbf{\Sigma}_A \). See Machado and Parente (2005) for regularity conditions for this assumption.

Following Bickel and Ren (2001), let the vector of parameters \( \mathbf{\mu} = T(F) \), the statistic \( T_n = T(F_n) \), and \( T^* = T(F^*_n) \) where \( F \) is the cdf of iid \( x_1, \ldots, x_n \), \( F_n \) is the empirical cdf, and \( F^*_n \) is the empirical cdf of \( x_1^*, \ldots, x_n^* \), a sample from \( F_n \) using the nonparametric bootstrap. If \( \sqrt{n}(F_n - F) \overset{D}{\rightarrow} z_F \), a Gaussian random process, and if \( T \) is sufficiently smooth (has a Hadamard derivative \( \dot{T}(F) \)), then \( \sqrt{n}(T_n - \mathbf{\mu}) \overset{D}{\rightarrow} \mathbf{U} \) and \( \sqrt{n}(T^*_n - T_n) \overset{D}{\rightarrow} \mathbf{U} \) with \( \mathbf{U} = T(F)z_F \). Olive (2017ce) used these results to show that if \( \mathbf{U} \sim \mathcal{N}_r(0, \mathbf{\Sigma}_A) \), then \( \sqrt{n}(\mathbf{T}^* - T_n) \overset{D}{\rightarrow} \mathbf{0} \), \( \sqrt{n}(\mathbf{T}_i^* - T^*_n) \overset{D}{\rightarrow} \mathbf{U_i} \), \( \sqrt{n}(\mathbf{T}^* - \mathbf{\mu}) \overset{D}{\rightarrow} \mathbf{U} \), and that the prediction region method large sample 100(1 − \( \delta \))% confidence region for \( \mathbf{\mu} \) is

\[
\{ \mathbf{w} : (\mathbf{w} - \mathbf{T}^*)^T \mathbf{S}_T^* \mathbf{S}_T^* (\mathbf{w} - \mathbf{T}^*) \leq D^2_w(\mathbf{U}_B)\} = \{ \mathbf{w} : D^2_w(\mathbf{T}^*_i, \mathbf{S}_T^*) \leq D^2_w(\mathbf{U}_B)\} \tag{1.15} \]

where \( D^2_w(\mathbf{U}_B) \) is computed from \( D^2_i = (T^*_i - \mathbf{T}^*)^T \mathbf{S}_T^* \mathbf{S}_T^* (T^*_i - \mathbf{T}^*) \) for \( i = 1, \ldots, B \). Note that the corresponding test for \( H_0 : \mathbf{\mu} = \mathbf{\mu}_0 \) rejects \( H_0 \) if \( (\mathbf{T}^* - \mathbf{\mu}_0)^T \mathbf{S}_T^* (\mathbf{T}^* - \mathbf{\mu}_0) > D^2_w(\mathbf{U}_B) \).

The prediction region method for testing \( H_0 : \mathbf{\mu} = \mathbf{c} \) versus \( H_1 : \mathbf{\mu} \neq \mathbf{c} \) is simple. Let \( \hat{\mathbf{\mu}} \) be a consistent estimator of \( \mathbf{\mu} \) and make a bootstrap sample \( \mathbf{w}_i = \hat{\mathbf{\mu}}_i - \mathbf{c} \) for \( i = 1, \ldots, B \). Make the nonparametric prediction region (1.15) for the \( \mathbf{w}_i \) and fail to reject \( H_0 \) if \( \mathbf{0} \) is in the prediction region (if \( D_0 \leq D(\mathbf{U}_B) \), reject \( H_0 \) otherwise.

The Bickel and Ren (2001) hypothesis testing method is equivalent to using confidence region (1.15) with \( \mathbf{T}^* \) replaced by \( T_n \) and \( \mathbf{U}_B \) replaced by \( k_B = \lfloor B(1 - \delta) \rfloor \). If region (1.15) or the Bickel and Ren (2001) region is a large sample 100(1 − \( \delta \))% confidence region, then so is the other region if \( \sqrt{n}(\mathbf{T}^* - T_n) \overset{D}{\rightarrow} \mathbf{0} \). Hadamard differentiability and asymptotic normality are two of the sufficient conditions for both regions to be large sample confidence regions if \( \text{n} \mathbf{S}_T^* \overset{P}{\rightarrow} \mathbf{\Sigma}_A \), but Bickel and Ren (2001) showed that their method can work when Hadamard differentiability fails.

For \( r = 1 \), the percentile method uses an interval that contains \( \mathbf{U}_B \approx k_B = \lfloor B(1 - \delta) \rfloor \) of the \( \mathbf{T}^*_n \) from a bootstrap sample \( \mathbf{T}^*_1, \ldots, \mathbf{T}^*_B \) where the statistic \( \mathbf{T}_n \) is an estimator of \( \mathbf{\mu} \) based on a sample of size \( n \). Note that the squared Mahalanobis distance \( D^2_\mu = (\mathbf{\mu} - \mathbf{T}^*)^2 / \mathbf{S}_T^* \leq D^2_w(\mathbf{U}_B) \) is equivalent to \( \mathbf{\mu} \in [\mathbf{T}^* - \mathbf{S}_T^* D(\mathbf{U}_B), \mathbf{T}^* + \mathbf{S}_T^* D(\mathbf{U}_B)] \), which is an interval centered at \( \mathbf{T}^* \) just long enough to cover \( \mathbf{U}_B \) of the \( \mathbf{T}^*_n \). Hence the prediction region method is a special case of the percentile method if \( r = 1 \). Efron (2014) used a similar large sample 100(1 − \( \delta \))% confidence interval assuming that \( \mathbf{T}_n \) is asymptotically normal. The Frey (2013) shorth(c) interval (1.8) (with \( c \) given by (1.10)) applied to the \( \mathbf{T}^*_n \) is recommended since the shorth confidence interval can be much shorter than the Efron (2014) or prediction region method confidence intervals if \( r = 1 \). The shorth confidence interval is a practical implementation of the Hall (1988) shortest bootstrap interval based on all possible bootstrap samples.

Section 2 considers prediction intervals after model selection where \( n/p \) need not be large. Section 3 discusses hypothesis testing when \( n/p \) is large, and Section 4 gives a simulation.
2 PREDICTION INTERVALS AFTER MODEL SELECTION

Results from Hastie, Tibshirani, and Wainwright (2015, pp. 20, 296, ch. 6, ch. 11) suggest that lasso can perform well for sparse models: the subset $S$ in (1.6) contains $a_S$ predictors where $a_S/n \to 0$ as $n \to \infty$. Let $d$ be a crude estimate of the model degrees of freedom. With the exception of ridge regression, $d$ is the number of “variables” used by the method. Forward selection, lasso, and relaxed lasso use variables $x_1^*, \ldots, x_d^*$ while PCR and PLS use variables that are linear combinations of the predictors $V_j = \gamma_{f}^j x$ for $j = 1, \ldots, d$. See Efron and Hastie (2016, pp. 221, 222, 231) and Tibshirani (2015) for lasso degrees of freedom.

For $n/p$ large, Olive (2013) developed prediction intervals for models of the form $Y_i = m(x_i) + e_i$, and variable selection models for (1.1) have this form, as noted by Olive (2017b). The first new PI, that can be useful even if $n/p$ is not large, is defined below. The PI is similar to the Olive (2013) PI with $p$ replaced by $d$, if $d$ is not too large.

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

\[ q_n = \min(1 - \delta/2, 1 - \delta + 10\delta d/n), \quad \text{otherwise.} \] (2.1)

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

\[ c = \lceil nq_n \rceil, \] (2.2)

and let

\[ b_n = \left(1 + \frac{15}{n}\right)^{1/2} \sqrt{\frac{n + 2d}{n - d}}, \] (2.3)

if $d \leq 8n/9$, and

\[ b_n = 5 \left(1 + \frac{15}{n}\right), \] otherwise. Compute the shorth$(c)$ of the residuals $[r(s), r(s+c-1)] = [\xi_{\Delta 1}, \xi_{\Delta 2}]$. Then the first new $100(1 - \delta)$% large sample PI for $Y_f$ is

\[ [\hat{m}(x_f) + b_n \xi_{\Delta 1}, \hat{m}(x_f) + b_n \xi_{\Delta 2}]. \] (2.4)

The second new PI randomly divides the data into two half sets $H$ and $V$ where $H$ has $n_H = \lceil n/2 \rceil$ of the cases and $V$ has the remaining $n_V = n - n_H$ cases $i_1, \ldots, i_{n_V}$. The estimator $\hat{m}_H(x)$ is computed using the training data set $H$. Then the validation residuals $v_j = Y_i - \hat{m}_H(x_i)$ are computed for the $j = 1, \ldots, n_V$ cases in the validation set $V$. Find the Frey PI $[v(s), v(s+c-1)]$ of the validation residuals (replacing $n$ in (1.10) by $n_V = n - n_H$). Then second new $100(1 - \delta)$% large sample PI for $Y_f$ is

\[ [\hat{m}_H(x_f) + v(s), \hat{m}_H(x_f) + v(s+c-1)]. \] (2.5)

The PIs (2.4) and (2.5) are asymptotically equivalent if $p$ is fixed and $n \to \infty$, but $\hat{m}_H$ has about half the efficiency of $\hat{m}$. When PI (2.4) has severe undercoverage because
\( \hat{m} \) is a poor estimator of \( m \), it is expected that PI (2.5) may have coverage closer to the nominal coverage. For example, if \( \hat{m} \) interpolates the data and \( \hat{m}_H \) interpolates the training data from \( H \), then the validation residuals will be huge. Hence PI (2.5) will be long compared to PI (2.4).

We can also motivate PI (2.5) by modifying the justification for the Lei, G’Sell, Rinaldo, Tibshirani, and Wasserman (2017) split conformal prediction interval \([\hat{m}_H(x_f) - a_q, \hat{m}_H(x_f) + a_q]\) where \( a_q \) is an appropriate quantile of the absolute validation residuals. Suppose \((Y_i, x_i)\) are iid for \( i = 1, ..., n, n + 1 \) where \((Y_f, x_f) = (Y_{n+1}, x_{n+1})\). Compute \( \hat{m}_H(x) \) from the cases in \( H \). For example, get \( \hat{\beta}_H \) from the cases in \( H \). Consider the validation residuals \( v_i \) for \( i = 1, ..., n_V \) and the validation residual \( v_{n_V + 1} \) for case \((Y_f, x_f)\). Since these \( n_V + 1 \) cases are iid, the probability that \( v_i \) has rank \( j \) for \( j = 1, ..., n_V + 1 \) is \( 1/(n_V + 1) \) for each \( t \), i.e., the ranks follow the discrete uniform distribution. Let \( t = n_V + 1 \) and let the \( v(j) \) be the ordered residuals using \( j = 1, ..., n_V \). That is, get the order statistics without using the unknown validation residual \( v_{n_V + 1} \). Then \( v(i) \) has rank \( i \) if \( v(i) < v_{n_V + 1} \) but rank \( i + 1 \) if \( v(i) > v_{n_V + 1} \). Thus

\[
P(Y_f \in [\hat{m}_H(x_f) + v(k), \hat{m}_H(x_f) + v(k+b-1)]) = P(v(k) \leq v_{n_V + 1} \leq v(k+b-1)) \geq \]

\[
P(v_{n_V + 1}) \text{ has rank between } k + 1 \text{ and } k + b - 1 \text{ and there are no tied ranks} \geq \frac{(b - 1)}{(n_V + 1)} \approx 1 - \delta \text{ if } b = [(n_V + 1)(1 - \delta)] + 1 \text{ and } k + b - 1 \leq n_V. \]

This probability statement holds for a fixed \( k \) such as \( k = \lceil n_V \delta/2 \rceil \). The statement is not true when the shorth \((b)\) estimator is used since the shortest interval using \( k = s \) can have \( s \) change with the data set. That is, \( s \) is not fixed. Hence if PI’s were made from \( J \) independent data sets, the PI’s with fixed \( k \) would contain \( Y_f \) about \( J(1 - \delta) \) times, but this value would be smaller for the shorth \((b)\) prediction intervals where \( s \) can change with the data set.

The PIs (2.4) and (2.5) can be used with \( \hat{m}(x) = \hat{Y}_f = x_I^T \hat{\beta}_I \) where \( I_d \) denotes the index of predictors selected from the model or variable selection method. The split conformal PI can have good coverage, but PI (2.5) does not need the error distribution to be symmetric to be asymptotically optimal.

Note that correction factors \( b_n \to 1 \) are used in large sample confidence intervals and tests if the limiting distribution is \( N(0,1) \) or \( \chi^2_p \), but a \( t_{d_n} \) or \( pF_{p,d_n} \) cutoff is used: \( t_{d_n,1-\delta}/z_{1-\delta} \to 1 \) and \( pF_{p,d_n,1-\delta}/\chi^2_{p,1-\delta} \to 1 \) if \( d_n \to \infty \) as \( n \to 1 \). Using correction factors for prediction intervals and bootstrap confidence regions improves the performance for moderate sample size \( n \).

### 3 HYPOTHESIS TESTING AFTER MODEL SELECTION, \( n/p \) LARGE

Assume \( n \geq 20p \) and that the error distribution is unimodal and not highly skewed. The response plot and residual plot are plots with \( \hat{Y} = x^T \hat{\beta} \) on the horizontal axis and \( Y \) or \( r \) on the vertical axis, respectively. Then the plotted points in these plots should scatter in roughly even bands about the identity line with unit slope and zero intercept and the \( r = 0 \) line, respectively. See Figure 1. If the plots for the OLS full model suggest that the error distribution is skewed or multimodal, then much larger sample sizes may be needed.
If the error distribution is unknown, then large sample theory tests are straightforward if the estimator is asymptotically equivalent to the OLS full model, e.g. \( \hat{\lambda}_{1,n} = o_P(\sqrt{n}) \), or choose the OLS full model if \( n \geq 50p \). The latter technique may be reasonable if the large sample theory of the method is not better than that of the OLS full model (lasso and ridge regression), or if it is not known how to do inference unless the model is asymptotically equivalent to the OLS full model, e.g., if \( P(d \to p) \to 1 \) (PCR, PLS, forward selection).

The residual bootstrap with the residuals from the OLS full model can provide a lot of information. Olive (2017a: p. 128, 2017b) showed that the prediction region method can simulate well for the \( p \times 1 \) vector \( \hat{\beta}_{I_{min},0} \) obtained by adding zeros to \( \hat{\beta}_{I_{min}} \) where \( I_{min} \) is the model that minimizes \( C_p \) for forward selection. Asymptotically, \( \hat{\beta}_{I_{min},0} \) is a mixture \( \sum_{j} \pi_j \hat{\beta}_{I_j,0} \) where \( 0 \leq \pi_j \leq 1 \) and \( \sum_{j} \pi_j = 1 \) where the sum is over all \( 2^{p-a_S} \) submodels \( I_j \) that contain \( S \). Results from Knight and Fu (2000) show that this residual bootstrap (with residuals from the OLS full model) works for each component \( \hat{\beta}_{I_j} \) of the mixture distribution for forward selection, but we may need at least \( 50p \) bootstrap samples per component with nonnegligible \( \pi_j \). The number of nonnegligible \( \pi_j \) can be small if \( p - a_S \) is small or if a criterion that picks \( S \) with high probability, such as BIC, is used. Here \( Y = x^T \beta + e = x_S^T \beta_S + e \) where \( \beta_S = a_s \times 1 \).

Examining \( \hat{\beta}_S \) and \( \hat{\beta}_E \) is informative for \( I_{min} \). First assume that the nontrivial predictors are orthogonal or uncorrelated with zero mean mean so \( X^T X / n \to \text{diag}(d_1, \ldots, d_p) \) as \( n \to \infty \) where each \( d_i > 0 \). Then \( \hat{\beta}_S \) has the same limiting distribution for \( I_{min} \) and for the OLS full model. The bootstrap distribution for \( \beta_E \) is a mixture of zeros and a distribution that would produce a confidence region for \( A \beta_E = 0 \) that has asymptotic coverage of \( 0 \) equal to \( 100(1 - \delta)\% \). Hence the asymptotic coverage is greater than the nominal coverage provided that \( S^* \) in nonsingular with probability going to one (e.g., \( p - a_S \) is small), where \( T = A \beta_{E,I_{min},0} \). With uncorrelated predictors, the number of bootstrap samples \( B \geq 50p \) may work well for the short confidence intervals and for testing \( A \beta_S = 0 \).

We do not yet have a proof that the prediction region method works when the estimator is not asymptotically multivariate normal, but in the simulations for forward selection, coverages were similar regardless of the correlation of the predictors. Let \( \beta_O \) be a vector component of \( \beta_E \), and consider testing \( H_0 : A \beta_O = 0 \). If \( A \hat{\beta}_{O,i} = 0 \) for greater than \( B \delta \) of the bootstrap samples \( i = 1, \ldots, B \), then the \( 100(1 - \delta)\% \) prediction region method confidence region will contain \( 0 \), and the test will fail to reject \( H_0 \).

Suppose we want to bootstrap \( T = \hat{\beta}_O \), where \( \beta = (\beta_I^T, \beta_O^T)^T \), and all \( \hat{\beta}_{O,i} = 0 \) for \( i = 1, \ldots, B \). Then \( S^*_T \) is singular, but the singleton set \( \{0\} \) is the large sample prediction region method \( 100(1 - \delta)\% \) confidence region for \( \beta_O \) and \( \delta \in (0, 1) \), and the pvalue for \( H_0 : \beta_O = 0 \) is one. For large sample theory tests, the pvalue estimates the population pvalue. For the \( I_{min} \) model from forward 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 \( \{0\} \), \( n \geq 20p \), \( B \geq 50p \), and the error distribution is unimodal and not highly skewed. (Since the pvalue is one, this technique may be useful for data snooping: applying OLS theory to submodel \( I \) may have negligible selection bias.)
4 EXAMPLE AND SIMULATIONS

Example. The Hebbler (1847) data was collected from \( n = 26 \) districts in Prussia in 1843. We will study the relationship between \( Y = \text{the number of women married to civilians} \) in the district with the predictors \( x_1 = \text{constant} \), \( x_2 = \text{pop} = \text{the population of the district in 1843} \), \( x_3 = \text{mmen} = \text{the number of married civilian men} \) in the district, \( x_4 = \text{mmilmen} = \text{number of married men in the military} \) in the district, and \( x_5 = \text{milwmn} = \text{the number of women married to husbands in the military} \) in the district. Sometimes the person conducting the survey would not count a spouse if the spouse was not at home. Hence \( Y \) and \( X_3 \) are highly correlated but not equal. Similarly, \( x_4 \) and \( x_5 \) are highly correlated but not equal. We expect that \( Y = x_3 + e \) is a good model.

Consider PI (2.4). Forward selection selected the model with the minimum \( C_p \) while the other methods used 10-fold CV. PLS and PCR used the OLS full model with 90% PI length 2395.74, forward selection used a constant and \( \text{mmen} \) with PI length 2114.72, ridge regression had PI length 20336.58, lasso and relaxed lasso used a constant, \( \text{mmen} \), and \( \text{pop} \) with lengths 8482.62 and 2226.53, respectively. Figure 1 shows the response plots for forward selection, ridge regression, lasso, and relaxed lasso. The plots for PLS=PCR=OLS full model were similar to those of forward selection and relaxed lasso. The plots suggest that the MLR model is appropriate since the plotted points scatter about the identity line. The 90% pointwise prediction bands are also shown, and consist of two lines parallel to the identity line. These bands are very narrow in Figure 1 a) and d).

Figure 1: Marry Data Response Plots
We used 5-fold CV with coverage and average 95% PI length to compare the forward selection models. All 4 models had coverage 1, but the average PI lengths were 2591.243, 2741.154, 2902.628, and 2972.963 for the models with 2 to 5 predictors.

For the model \( Y = X\beta + e \), and models \( I_1, \ldots, I_M \), we considered several methods for selecting the final submodel \( I_d \). Only method i) needs \( n/p \) large. Pelawa Watagoda (2017) has a large simulation.

i) Let \( I_d = I_{\text{min}} \) be the model that minimizes \( C_p \) for forward selection, relaxed lasso, or lasso.

ii) Let \( I_d \) use \( d = \min(\lceil n/J \rceil, p) \) variables, including a constant, where \( J \geq 5 \) is a positive integer. This method uses the OLS full model if \( n/p \geq J \). Hence large sample inference is simple if \( p \) is fixed. This method simulated well for \( d = p \) and \( J \geq 10 \), but prediction intervals were often unreliable, otherwise.

iii) Let \( I_d = I_{\text{min}} \) be the model that minimizes \( \text{EBIC} \) for forward selection or relaxed lasso. For forward selection, we used \( M = \min(\lceil n/5 \rceil, p) \).

iv) Choose \( I_d \) using \( k \)-fold cross validation (CV). We used 10-fold CV.

The following method is currently slow to simulate, but is a useful diagnostic. See the above example. When the model underfits, PI (2.4) tends to have coverage near or greater than the nominal 0.95 coverage, but the PI length is long. When the model severely overfits, the PI tends to have short length with coverage less than 0.95.

v) Modify \( k \)-fold cross validation to compute the PI coverage and average PI length on all \( M \) models. Then \( n \) PIs are made for \( Y_i^* \) using \( x_f = x_i \) for \( i = 1, \ldots, n \). The coverage is the proportion of times the \( n \) PIs contained \( Y_i^* \). For example, choose the model \( I_d \) with the shortest average PI length given that the nominal large sample 100(1 - \( \delta \))% PI had coverage

\[
\geq c_n = \max(1 - \delta - \frac{1}{3\sqrt{n}}, 1 - \delta - 0.02).
\]

If no model \( I_d \) had coverage \( \geq c_n \), pick the model with the largest coverage.

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 \( u_i \sim N_p(0, I) \) where the \( m = p - 1 \) elements of the vector \( u_i \) are iid \( N(0,1) \). Let the \( m \times m \) matrix \( A = (a_{ij}) \) with \( a_{ii} = 1 \) and \( a_{ij} = \psi \) where \( 0 \leq \psi < 1 \) for \( i \neq j \). Then the vector \( u_i = Au_i \) so that \( \text{Cov}(u_i) = \Sigma u = AA^T = (\sigma_{ij}) \) where the diagonal entries \( \sigma_{ii} = [1 + (m - 1)\psi^2] \) and the off diagonal entries \( \sigma_{ij} = [2\psi + (m - 2)\psi^2]/(1 + (m - 1)\psi^2) \) for \( i \neq j \) where \( x_i \) and \( x_j \) are nontrivial predictors. If \( \psi = 1/\sqrt{cp} \), then \( \rho \to 1/(c+1) \) as \( p \to \infty \) where \( c > 0 \). As \( \psi \) gets close to 1, the predictor vectors cluster about the line in the direction of \((1, \ldots, 1)^T\). Let \( Y_i = 1 + 1x_{i,2} + \cdots + 1x_{i,k+1} + e_i \) for \( i = 1, \ldots, n \). Hence \( \beta = (1, \ldots, 1, 0, \ldots, 0)^T \) with \( k + 1 \) ones and \( p - k - 1 \) zeros. The zero mean errors \( e_i \) were iid from five distributions: i) \( N(0,1) \), ii) \( t_3 \), iii) \( \text{EXP}(1) - 1 \), iv) uniform\((-1,1)\), and v) \( 0.9 \ N(0,1) + 0.1 \ N(0,100) \). Only distribution iii) is not symmetric.

The lengths of the asymptotically optimal 95% PIs are i) \( 3.92 = 2(1.96) \), ii) \( 6.365 \), iii) 2.996, iv) 1.90 = 2(0.95), and v) 13.490. The simulation used 5000 runs, so an observed coverage in [0.94, 0.96] gives no reason to doubt that the PI has the nominal coverage of 0.95. The simulation used \( p = 20, 40, 50, n, \) and \( 2n \). The simulation used \( \psi = 0, 1/\sqrt{p} \), and 0.9, and \( k = 1, 19, \) and \( p - 1 \).
Table 1 shows some simulation results for PI (2.4) where forward selection used $C_p$ for $n \geq 10p$ and EBIC for $n < 10p$. The other methods minimized 10-fold CV (with relaxed lasso applied to the selected lasso model). For methods other than ridge regression, the maximum number of variables used was approximately $\min([n/5], p)$. For $n \geq 10p$, coverages tended to be near or higher than the nominal value of 0.95. The average PI length was often near 1.3 times the asymptotically optimal length for $n = 10p$ and close to the optimal length for $n = 100p$. An exception was that lasso, relaxed lasso, and ridge regression often had too long lengths if both $k \geq 19$ and $\psi \geq 5$.

Table 1: Simulated Large Sample 95% PI Coverages and Lengths, $e_i \sim N(0,1)$

<table>
<thead>
<tr>
<th>$n$</th>
<th>$p$</th>
<th>$\psi$</th>
<th>$k$</th>
<th>FS</th>
<th>lasso</th>
<th>RL</th>
<th>RR</th>
<th>PLS</th>
<th>PCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>20</td>
<td>0</td>
<td>1</td>
<td>0.9644</td>
<td>0.9570</td>
<td>0.9438</td>
<td>0.9354</td>
<td>0.9438</td>
<td>0.9772</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.4490</td>
<td>4.3849</td>
<td>4.2595</td>
<td>4.1441</td>
<td>4.1419</td>
<td>5.5647</td>
</tr>
<tr>
<td>100</td>
<td>40</td>
<td>0</td>
<td>1</td>
<td>0.9654</td>
<td>0.9522</td>
<td>0.9286</td>
<td>0.8932</td>
<td>0.8832</td>
<td>0.9882</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.4294</td>
<td>4.3113</td>
<td>4.0899</td>
<td>3.8982</td>
<td>3.6104</td>
<td>4.0202</td>
</tr>
<tr>
<td>100</td>
<td>200</td>
<td>0</td>
<td>1</td>
<td>0.9648</td>
<td>0.9332</td>
<td>0.8632</td>
<td>0.9254</td>
<td>0.6616</td>
<td>0.9922</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.4268</td>
<td>4.1546</td>
<td>3.6263</td>
<td>4.9843</td>
<td>2.7695</td>
<td>12.412</td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>0</td>
<td>49</td>
<td>0.8996</td>
<td>0.8416</td>
<td>0.8440</td>
<td>0.8734</td>
<td>0.8448</td>
<td>1.0000</td>
</tr>
<tr>
<td>200</td>
<td>20</td>
<td>0</td>
<td>19</td>
<td>0.9788</td>
<td>0.9766</td>
<td>0.9788</td>
<td>0.9792</td>
<td>0.9550</td>
<td>0.9786</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.9613</td>
<td>4.9636</td>
<td>4.9613</td>
<td>5.0458</td>
<td>4.3211</td>
<td>4.9610</td>
</tr>
<tr>
<td>200</td>
<td>40</td>
<td>0</td>
<td>19</td>
<td>0.9742</td>
<td>0.9650</td>
<td>0.9600</td>
<td>0.9606</td>
<td>0.9324</td>
<td>0.9792</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.9285</td>
<td>4.8146</td>
<td>4.7164</td>
<td>4.8044</td>
<td>4.2152</td>
<td>5.3616</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
<td>0</td>
<td>19</td>
<td>0.9728</td>
<td>0.9124</td>
<td>0.7870</td>
<td>0.9696</td>
<td>0.3500</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.8835</td>
<td>4.3197</td>
<td>3.4075</td>
<td>16.589</td>
<td>2.1451</td>
<td>51.896</td>
</tr>
<tr>
<td>400</td>
<td>20</td>
<td>0.9</td>
<td>19</td>
<td>0.9664</td>
<td>0.9806</td>
<td>0.9686</td>
<td>0.9810</td>
<td>0.9554</td>
<td>0.9536</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.5121</td>
<td>11.711</td>
<td>4.8065</td>
<td>11.229</td>
<td>4.0017</td>
<td>3.9771</td>
</tr>
<tr>
<td>400</td>
<td>40</td>
<td>0.9</td>
<td>19</td>
<td>0.9674</td>
<td>0.9792</td>
<td>0.9710</td>
<td>0.9756</td>
<td>0.9482</td>
<td>0.9646</td>
</tr>
<tr>
<td>400</td>
<td>400</td>
<td>0.9</td>
<td>19</td>
<td>0.9348</td>
<td>0.9672</td>
<td>0.9564</td>
<td>0.9666</td>
<td>0.9462</td>
<td>0.9478</td>
</tr>
<tr>
<td>400</td>
<td>800</td>
<td>0.9</td>
<td>19</td>
<td>0.9268</td>
<td>0.9660</td>
<td>0.9554</td>
<td>0.9692</td>
<td>0.9438</td>
<td>0.9554</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.3427</td>
<td>67.405</td>
<td>4.7992</td>
<td>66.689</td>
<td>4.2965</td>
<td>4.6533</td>
</tr>
<tr>
<td>400</td>
<td>400</td>
<td>0</td>
<td>399</td>
<td>0.9486</td>
<td>0.2330</td>
<td>0.1302</td>
<td>0.8702</td>
<td>0.0948</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>78.411</td>
<td>7.3847</td>
<td>4.0371</td>
<td>47.590</td>
<td>1.1749</td>
<td>305.93</td>
</tr>
</tbody>
</table>

For $n \leq p$, good performance needed stronger regularity conditions. PLS tended to have severe undercoverage with small average length. The PCR length was often too long for $\psi = 0$. When there was $k = 1$ active predictor, forward selection, lasso, and relaxed lasso often performed well. For $k = 19$, forward selection performed well, as did lasso and relaxed lasso for $\psi = 0$. For dense models with $k = p − 1$ and $n/p$ not large, there was often undercoverage. Let $d − 1$ be the number of active predictors in the selected model. For $N(0,1)$ errors, $\psi = 0$, and $d < k$, an asymptotic population 95% PI has
length $3.92\sqrt{k - d + 1}$. EBIC occasionally had undercoverage, especially for $k = 19$ or $p - 1$, which was usually more severe for $\psi = 0.9$ or $1/\sqrt{p}$.

A small simulation was done, using the same type of data as for the prediction interval simulation, using $B = \max(1000, n, 20p)$ and 5000 runs. The regression model used $\beta = (1, 1, 0, 0)^T$ with $n = 100$ and $p = 4$. When $\psi = 0$, the full model least squares confidence intervals for $\beta_i$ should have length near $2t_{0.975, 9}/\sqrt{n} \approx 2(1.96)\sigma/10 = 0.392\sigma$ when the iid zero mean errors have variance $\sigma^2$. The simulation computed the Frey shorth(c) interval for each $\beta_i$ and used the prediction region method to test $H_0 : \beta_3 = \beta_4 = 0$. 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. Models with the first $k + 1 \beta_i = 1$ and the last $p - k - 1 \beta_i = 0$ were also considered.

The regression models used the residual bootstrap on the full model least squares estimator and on the forward selection estimator $\hat{\beta}_{I_{\min}, 0}$. Results are shown for when the iid errors $e_i \sim N(0, 1)$. Table 2 shows two rows for each model giving the observed confidence interval coverages and average lengths of the confidence intervals. The term “reg” is for the full model regression, and the term “vs” is for forward selection. The column for the “test” gives the length and coverage = $P(\text{fail to reject } H_0)$ for the interval $[0, D(U_{lb})]$ where $D(U_{lb})$ is the cutoff for the confidence region. These lengths do not give information about the volume of the confidence region, which will decrease to 0 as $n \to \infty$. The cutoff will often be near $\sqrt{\chi^2_{2, 0.95}}$ 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 test. The coverages were near 0.95 for the regression bootstrap on the full model.

Suppose $\psi = 0$. Then from Section 3, $\hat{\beta}_S$ has the same limiting distribution for $I_{\min}$ and the full model. Note that the average lengths and coverages were similar for the full model and forward selection $I_{\min}$ for $\beta_1$ and $\beta_2$ and $\beta_S = (\beta_1, \beta_2)^T$. Simulations for lasso were similar with $\psi = 0$.

Table 2: Bootstrapping OLS Regression and Forward Selection, $e_i \sim N(0, 1)$

<table>
<thead>
<tr>
<th>model</th>
<th>$\psi$</th>
<th>cov/len</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\beta_3$</th>
<th>$\beta_4$</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>reg</td>
<td>0</td>
<td>cov</td>
<td>0.9456</td>
<td>0.9474</td>
<td>0.9496</td>
<td>0.9474</td>
<td>0.9442</td>
</tr>
<tr>
<td></td>
<td></td>
<td>len</td>
<td>0.3961</td>
<td>0.3997</td>
<td>0.3988</td>
<td>0.3992</td>
<td>2.4503</td>
</tr>
<tr>
<td>vs</td>
<td>0</td>
<td>cov</td>
<td>0.9472</td>
<td>0.9470</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9936</td>
</tr>
<tr>
<td></td>
<td></td>
<td>len</td>
<td>0.3964</td>
<td>0.3991</td>
<td>0.3246</td>
<td>0.3233</td>
<td>2.6936</td>
</tr>
<tr>
<td>reg</td>
<td>0.5</td>
<td>cov</td>
<td>0.9432</td>
<td>0.9452</td>
<td>0.9498</td>
<td>0.9506</td>
<td>0.9436</td>
</tr>
<tr>
<td></td>
<td></td>
<td>len</td>
<td>0.3976</td>
<td>0.6642</td>
<td>0.6645</td>
<td>0.6637</td>
<td>2.4507</td>
</tr>
<tr>
<td>vs</td>
<td>0.5</td>
<td>cov</td>
<td>0.9458</td>
<td>0.9728</td>
<td>0.9976</td>
<td>0.9974</td>
<td>0.9926</td>
</tr>
<tr>
<td></td>
<td></td>
<td>len</td>
<td>0.3966</td>
<td>0.6598</td>
<td>0.5383</td>
<td>0.5383</td>
<td>2.7055</td>
</tr>
<tr>
<td>reg</td>
<td>0.9</td>
<td>cov</td>
<td>0.9432</td>
<td>0.9512</td>
<td>0.9500</td>
<td>0.9498</td>
<td>0.9442</td>
</tr>
<tr>
<td></td>
<td></td>
<td>len</td>
<td>0.3963</td>
<td>3.2621</td>
<td>3.2613</td>
<td>3.2611</td>
<td>2.4505</td>
</tr>
<tr>
<td>vs</td>
<td>0.9</td>
<td>cov</td>
<td>0.9422</td>
<td>0.9678</td>
<td>0.9944</td>
<td>0.9970</td>
<td>0.9914</td>
</tr>
<tr>
<td></td>
<td></td>
<td>len</td>
<td>0.3957</td>
<td>2.7640</td>
<td>2.7356</td>
<td>2.7430</td>
<td>2.7121</td>
</tr>
</tbody>
</table>
For $\psi > 0$ and $I_{min}$, the coverages for the active predictors were near 0.95, perhaps since $p - a_S = 2$ is small, but the average length could be shorter since $I_{min}$ tends to have less multicorrelation than the full model. For $\psi \geq 0$, the $I_{min}$ coverages were higher than 0.95 for the inactive predictors (and for the “test”) since zeros often occurred for $\beta_j^*$ for $j = 3, 4$. The average lengths were shorter for $I_{min}$ for the inactive predictors.

5 CONCLUSIONS

Let $p$ be fixed and $n \to \infty$. For forward selection, PCR, PLS, ridge regression, relaxed lasso, and lasso, if $P(d \to p) \to 1$ as $n \to \infty$ then the six methods are asymptotically equivalent to the OLS full model, and the PIs (2.4) and (2.5) are asymptotically optimal on a large class of iid unimodal error distributions. For PCR and some constants $\theta_i$, $\sum_{i=1}^n \theta_i \gamma_i^T x_i = \sum_{i=1}^p \beta_i x_i$ if $j = p$, but not if $j < p$ in general. Hence PCR tends to give inconsistent estimators unless $P(j = p) = P(\text{PCR uses the OLS full model})$ goes to one. Forward selection with $C_p$ produces a $\sqrt{n}$ consistent estimator $\hat{\beta}_{I_{min}, 0}$ of $\beta$. Using $d = \min(\lceil n/J \rceil, p)$ with forward selection, PCR, PLS, ridge regression, lasso, and relaxed lasso makes large sample inference easy since the selected model is the OLS full model if $n/p \geq J$.


If $n/p$ is large, the residual bootstrap with OLS full model residuals should work for lasso, relaxed lasso, and ridge regression if $\hat{\lambda}_{1,n} = o_P(\sqrt{n})$. Also see Knight and Fu (2000). Camponovo (2015) suggests that the nonparametric bootstrap does not work for lasso. Chatterjee and Lahiri (2011) state that the residual bootstrap with lasso does not work. Hall, Lee, and Park (2009) state that the residual bootstrap with OLS full model residuals does not work, but the $m$ out of $n$ residual bootstrap with OLS full model residuals does work. Rejchel (2016) is a good review of lasso theory. Fan and Lv (2010) review large sample theory for some alternative methods. See Lockhart, Taylor, Tibshirani, and Tibshirani (2014) for a partial remedy for hypothesis testing with lasso. Also see Ning and Liu (2017). Xu, Caramanis, and Mannor (2011) suggest that sparse algorithms are not stable.

Lei, G’Sell, Rinaldo, Tibshirani, and Wasserman (2016) and Wasserman (2014) suggest prediction intervals for estimators such as lasso. Also see Butler and Rothman (1980). Steinberger and Leeb (2016) use leave-one-out residuals to make a PI. Chao, Ning, and Liu (2014) assume that the $e_i$ are iid $N(0, \sigma^2)$. Denham (1997) gave a PI for PLS when the number of PLS components $V_j$ is selected in advance. Zhang (1992) has some $k$–fold CV theory.

With $n/p$ large, $C_p$ produced good PIs for forward selection and 10-fold CV produced good PIs for PCR and PLS. For lasso and ridge regression, 10-fold CV produced good PIs if $\psi = 0$ or if $k$ was small. If $k$ was larger than about 18 and the predictors were
highly correlated, 10-fold CV tended to underfit and the PI lengths were too long.

When $n/p$ is not large, inference is currently much more difficult. Zheng and Loh (1995) show that $BIC_S$ can work if $p = o(\log(n))$ and there is a consistent estimator of $\sigma^2$. Chun and Keleş (2010) show that PLS does not give a consistent estimator of $\beta$ unless $p/n \to 0$. Also see Cook, Helland, and Su (2013). Fan and Lv (2010) give large sample theory for some methods if $p = o(n^{1/5})$. Leeb, Pötscher, and Ewald (2015) suggests that the Berk et al. (2013) method does not really work. Also see Dezeure et al. (2015), Javanmard and Montanari (2014), Lu et al. (2017), Taylor et al. (2014), and van de Geer et al. (2014).

Response plots of the fitted values $\hat{Y}$ versus the response $Y$ are useful for checking linearity of the MLR model and for detecting outliers. Residual plots should also be made. When $n$ is large, the points within the pointwise PI bands can be omitted, eliminating a black band about the identity line.

The simulations were done in $R$. See R Core Team (2016). A much larger simulation study is in Pelawa Watagoda (2017). We used several $R$ functions including forward selection as computed with the `regsubsets` function from the `leaps` library, principal components regression with the `pcr` function and partial least squares with the `plsr` function from the `pls` library, and ridge regression and lasso with the `glmnet` and `cv.glmnet` functions from the `glmnet` library.

The collection of Olive (2017d) $R$ functions `slpack`, available from (http://lagrange.math.siu.edu/Olive/slpack.txt), has some useful functions for the inference. Table 1 was made with `mspisim` while Table 2 was made with `regbootsim` for the OLS full model and `vsbootsim3` for forward selection. The function `lassobootsim3` uses the prediction region method for lasso. For PI (2.5), the function `valvspisim` is for forward selection using the minimum $C_p$ model, and the function `valrelpisim` simulates the relaxed lasso model corresponding to the lasso model chosen with 10-fold CV. The function `pifold` can be used to do $k$-fold CV with PI coverage and average length.

6 References


Butler, R., and Rothman, E. (1980), “Predictive Intervals Based on Reuse of the Sam-


Olive, D.J. (2017d), *Prediction and Statistical Learning*, online course notes, see (http://lagrange.math.siu.edu/Olive/slearnbk.htm).