Bootstrapping Hypothesis Tests and Confidence Regions

David J. Olive*
Southern Illinois University

July 2, 2017

Abstract

It will be shown that under regularity conditions, applying the Olive (2013) large sample $100(1-\delta)\%$ prediction region to the bootstrap sample $T_1^*, ..., T_B^*$ gives a large sample $100(1-\delta)\%$ confidence region for an $r \times 1$ parameter vector $\boldsymbol{\mu}$, generalizing the percentile method for $r = 1$ to $r \geq 1$. This prediction region method will be compared to the Efron (2014) confidence interval for variable selection, and used to bootstrap a correlation matrix.

Consider testing $H_0 : \boldsymbol{\mu} = \boldsymbol{c}$ versus $H_1 : \boldsymbol{\mu} \neq \boldsymbol{c}$ where $\boldsymbol{c}$ is a known $r \times 1$ vector. Let $\hat{\boldsymbol{\mu}}$ be a consistent estimator of $\boldsymbol{\mu}$ and make a bootstrap sample $\boldsymbol{w}_i = \hat{\boldsymbol{\mu}}_i^* - \boldsymbol{c}$ for $i = 1, ..., B$. Make the prediction region for the $\boldsymbol{w}_i$ and determine whether $\boldsymbol{0}$ is in the prediction region.

Bootstrapping test statistics is well known, and the prediction region method can be regarded as a special case where the bootstrapped test statistic is the squared Mahalanobis distance $D^2_{\boldsymbol{\mu}_0} = (\overline{T^*} - \boldsymbol{\mu}_0)^T [\boldsymbol{S}_T^*]^{-1} (\overline{T^*} - \boldsymbol{\mu}_0)$ where the bagging estimator $\overline{T^*}$ is the sample mean and $\boldsymbol{S}_T^*$ is the sample covariance matrix of $T_1^*, ..., T_B^*$.

The material in this manuscript has been incorporated in Olive (2017c: section 5.3, d: section 2.3) and in abbreviated form in Olive (2017a, c: section 3.4.1). Applications include Pelawa Watagoda and Olive (2017), Rupasinghe Arachchige Don and Olive (2017), and Rupasinghe Arachchige Don and Pelawa Watagoda (2017).

KEY WORDS: bagging, confidence region, prediction region, variable selection

*David J. Olive is Professor, Department of Mathematics, Southern Illinois University, Carbondale, IL 62901, USA.

1 Introduction

Consider testing $H_0 : \boldsymbol{\mu} = \boldsymbol{c}$ versus $H_1 : \boldsymbol{\mu} \neq \boldsymbol{c}$ where $\boldsymbol{c}$ is a known $r \times 1$ vector. If a confidence region can be constructed for $\boldsymbol{\mu} - \boldsymbol{c}$, then fail to reject H_0 if $\boldsymbol{0}$ is in the confidence region, and reject H_0 if $\boldsymbol{0}$ is not in the confidence region. Given training data $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$, a large sample $100(1-\delta)\%$ prediction region for a future test value $\boldsymbol{w}_f$ is a set $\mathcal{A}_n$ such that $P(\boldsymbol{w}_f \in \mathcal{A}_n) \rightarrow 1-\delta$ as $n \rightarrow \infty$, while a large sample confidence region for a parameter $\boldsymbol{\mu}$ is a set $\mathcal{A}_n$ such that $P(\boldsymbol{\mu} \in \mathcal{A}_n) \rightarrow 1-\delta$ as $n \rightarrow \infty$.

The duality between hypothesis tests and confidence regions is well known, but for bootstrap samples, the percentile method is simultaneously a large sample confidence interval for μ, and also a large sample prediction interval (PI) for a future bootstrap statistic $T^*_{f,n}$. It is natural to try to extend the percentile method to vector valued statistics and parameters using a large sample prediction region for a future bootstrap statistic $T^*_{f,n}$ as a large sample confidence region for $\boldsymbol{\mu}$, but practical large sample prediction regions where $r > 1$ and the underlying distribution is unknown have only recently been developed. See Olive (2013, 2017ab) and Lei, Robins, and Wasserman (2013).

For $r = 1$, the percentile method uses an interval that contains $U_B \approx k_B = \lceil B(1-\delta) \rceil$ of the $T^*_{i,n}$ from a bootstrap sample $T^*_{1,n}, ..., T^*_{B,n}$ where the statistic T_n is an estimator of μ based on a sample of size n. Often the n is suppressed in the double subscripts. Here $\lceil x \rceil$ is the smallest integer $\geq x$, e.g. $\lceil 7.8 \rceil = 8$. Let $T^*_{(1)}, T^*_{(2)}, ..., T^*_{(B)}$ be the order statistics of the bootstrap sample. Then one version of the percentile method discards the largest and smallest $\lceil B\delta/2 \rceil$ order statistics, resulting in an interval $(\hat{L}_B, \hat{R}_B)$. Janssen and Pauls (2003) and Mammen (1992) suggest that the bootstrap works if there is a central limit theorem for the statistic T_n. Also see Beran (1988), Bickel and Freedman (1981), Horowitz (2001), Machado and Parente (2005), and MacKinnon (2009).

Olive (2014, p. 283) recommends using the shorth(c) estimator for the percentile method. Let $c = k_B$, and let $W_i = T^*_{i,n}$. Let $W_{(1)}, ..., W_{(B)}$ be the order statistics of the W_i. Compute $W_{(c)} - W_{(1)}, W_{(c+1)} - W_{(2)}, ..., W_{(B)} - W_{(B-c+1)}$. Let $[W_{(s)}, W_{(s+c-1)}]$ correspond to the closed interval with the smallest distance. Then reject $H_0 : \mu = \mu_0$ if μ_0 is not in the interval. The shorth interval tends to be shorter than the interval that deletes the smallest and largest $\lceil B\delta/2 \rceil$ observations W_i when the W_i do not come from a symmetric distribution. Frey (2013) showed that for large $B\delta$ and iid data, the shorth(k_B) PI has maximum undercoverage $\approx 1.12\sqrt{\delta/B}$, and used the shorth$(c)$ estimator as the large sample $100(1-\delta)\%$ prediction interval where $c = \min(B, \lceil B[1-\delta+1.12\sqrt{\delta/B}\,] \rceil)$. Hence if $B = 1000$, there may be about 1% undercoverage using $c = k_B$. We recommend using the Frey (2013) shorth(c) intervals for the percentile method. Hall (1988) discusses the shortest bootstrap interval based on all bootstrap samples.

Some notation is needed to give the Olive (2013) prediction region used to bootstrap a hypothesis test. Suppose $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$ are iid $r \times 1$ random vectors with mean $\boldsymbol{\mu}$ and nonsingular covariance matrix $\boldsymbol{\Sigma_w}$. Let a future test observation $\boldsymbol{w}_f$ be independent of the $\boldsymbol{w}_i$ but from the same distribution. Let $(\overline{\boldsymbol{w}}, \boldsymbol{S})$ be the sample mean and sample covariance matrix where

$$\overline{\boldsymbol{w}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{w}_i \text{ and } \boldsymbol{S} = \boldsymbol{S_w} = \frac{1}{\mathrm{n}-1}\sum_{\mathrm{i}=1}^{\mathrm{n}} (\boldsymbol{w}_\mathrm{i} - \overline{\boldsymbol{w}})(\boldsymbol{w}_\mathrm{i} - \overline{\boldsymbol{w}})^\mathrm{T}. \quad (1)$$

Then the ith *squared sample Mahalanobis distance* is the scalar

$$D^2_{\boldsymbol{w}} = D^2_{\boldsymbol{w}}(\overline{\boldsymbol{w}}, \boldsymbol{S}) = (\boldsymbol{w} - \overline{\boldsymbol{w}})^T \boldsymbol{S}^{-1} (\boldsymbol{w} - \overline{\boldsymbol{w}}). \tag{2}$$

Let $D_i^2 = D^2_{\boldsymbol{w}_i}$ for each observation $\boldsymbol{w}_i$. Let $D_{(c)}$ be the cth order statistic of $D_1, ..., D_n$. Consider the hyperellipsoid

$$\mathcal{A}_n = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{\boldsymbol{w}}, \boldsymbol{S}) \leq D^2_{(c)}\} = \{\boldsymbol{w} : D_{\boldsymbol{w}}(\overline{\boldsymbol{w}}, \boldsymbol{S}) \leq D_{(c)}\}. \tag{3}$$

If n is large, we can use $c = k_n = \lceil n(1-\delta) \rceil$. If n is not large, using $c = U_n$ where U_n decreases to k_n, can improve small sample performance.

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

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta r/n), \quad \text{otherwise.} \tag{4}$$

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

Let $D_{(U_n)}$ be the $100q_n$th percentile of the D_i. Then the Olive (2013) large sample $100(1-\delta)\%$ nonparametric prediction region for a future value $\boldsymbol{x}_f$ given iid data $\boldsymbol{x}_1, ..., , \boldsymbol{x}_n$ is

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D^2_{(U_n)}\}, \tag{5}$$

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

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq \chi^2_{p,1-\delta}\}. \tag{6}$$

Olive (2013) showed that (5) is a large sample $100(1-\delta)\%$ prediction region under mild conditions, although regions with smaller volumes may exist. Note that the result follows since if $\boldsymbol{\Sigma_w}$ and $\boldsymbol{S}$ are nonsingular, then the Mahalanobis distance is a continuous function of $(\overline{\boldsymbol{w}}, \boldsymbol{S})$. Let $D = D(\boldsymbol{\mu}, \boldsymbol{\Sigma_w})$. Then $D_i \xrightarrow{D} D$ and $D_i^2 \xrightarrow{D} D^2$. Hence the sample percentiles of the D_i are consistent estimators of the population percentiles of D at continuity points of the cumulative distribution function (cdf) of D, and (5) estimates the highest density region for a large class of elliptically contoured distributions. See Olive (2017ab) for more on prediction regions. The population percentile $D^2_{1-\delta}$ satisfies $P(D^2 \leq D^2_{1-\delta}) = 1 - \delta$.

The prediction region method makes a bootstrap sample $\boldsymbol{w}_i = \hat{\boldsymbol{\mu}}_i^* - \boldsymbol{c}$ for $i = 1, ..., B$. Make the prediction region (5) for the $\boldsymbol{w}_i$ and determine whether $\boldsymbol{0}$ is in the prediction region. As shown below, the prediction region method is a special case of the percentile method, and a special case of bootstrapping a test statistic.

Consider testing $H_0 : \boldsymbol{\mu} = \boldsymbol{c}$ versus $H_1 : \boldsymbol{\mu} \neq \boldsymbol{c}$, and the statistic $T_i = \hat{\boldsymbol{\mu}} - \boldsymbol{c}$. If $E(T_i) = \boldsymbol{\theta}$ and $\text{Cov}(T_i) = \boldsymbol{\Sigma}_T$ were known, then the squared Mahalanobis distance $D_i^2(\boldsymbol{\theta}, \boldsymbol{\Sigma}_T) = (T_i - \boldsymbol{\theta})^T \boldsymbol{\Sigma}_T^{-1} (T_i - \boldsymbol{\theta})$ would be a natural statistic to use if the percentile $D^2_{1-\delta}(\boldsymbol{\theta}, \boldsymbol{\Sigma}_T)$ was known. The prediction region method bootstraps the squared Mahalanobis distances, forming the bootstrap sample $\boldsymbol{w}_i = T_i^* = \hat{\boldsymbol{\mu}}_i^* - \boldsymbol{c}$ and the squared Mahalanobis distances $D_i^2 = D_i^2(\overline{T^*}, \boldsymbol{S}_T^*) = (T_i^* - \overline{T^*})^T [\boldsymbol{S}_T^*]^{-1} (T_i^* - \overline{T^*})$ where $\overline{T^*} = \dfrac{1}{B} \sum_{i=1}^{B} T_i^*$ and $\boldsymbol{S}_T^* = \dfrac{1}{B-1} \sum_{i=1}^{B} (T_i^* - \overline{T^*})(T_i^* - \overline{T^*})^T$ are the sample mean and sample covariance

matrix of $T_1^*, ..., T_B^*$. Then the percentile method that contains the smallest U_B distances is used to get the closed interval $[0, D_{(U_B)}]$. If H_0 is true and $E[\hat{\boldsymbol{\mu}}] = \boldsymbol{c}$, then $\boldsymbol{\theta} = \mathbf{0}$. Let $D_{\mathbf{0}}^2 = \overline{T^*}^T [\boldsymbol{S}_T^*]^{-1} \overline{T^*}$ and fail to reject H_0 if $D_{\mathbf{0}} \leq D_{(U_B)}$ and reject H_0 if $D_{\mathbf{0}} > D_{(U_B)}$. This percentile method is equivalent to computing the prediction region (5) on the $\boldsymbol{w}_i = T_i^*$ and checking whether $\mathbf{0}$ is in the prediction region.

Note that the percentile method makes an interval that contains U_B of the scalar valued T_i^*. The prediction region method makes a hyperellipsoid that contains U_B of the $r \times 1$ vectors $T_i^* = \boldsymbol{w}_i$, and equivalently, makes an interval $[0, D_{(U_B)}]$ that contains U_B of the D_i.

When $r = 1$, a hyperellipsoid is an interval. Suppose the parameter of interest is μ, and there is a bootstrap sample $T_1^*, ..., T_B^*$. Let $a_i = |T_i^* - \overline{T^*}|$. Let $\overline{T^*}$ and S_T^{2*} be the sample mean and variance of the T_i^*. Then the squared Mahalanobis distance $D_\mu^2 = (\mu - \overline{T^*})^2 / S_T^{2*} \leq D_{(U_B)}^2$ is equivalent to $\mu \in [\overline{T^*} - S_T^* D_{(U_B)}, \overline{T^*} + S_T^* D_{(U_B)}] = [\overline{T^*} - a_{(U_B)}, \overline{T^*} + a_{(U_B)}]$, which is an interval centered at $\overline{T^*}$ just long enough to cover U_B of the T_i^*. Hence the prediction region method is a special case of the percentile method if $r = 1$. Note that when $r = 1$, then S_T^* and $D_{(U_B)}$ do not need to be computed.

Bootstrapping test statistics is well known, and the prediction region is a special case of this method using $D_{\boldsymbol{\mu}_0}^2 = D_{\boldsymbol{\mu}_0}^2(\overline{T^*}, \boldsymbol{S}_T^*)$ as the test statistic. See Bickel and Ren (2001).

The point of the above discussion is that prediction region method can be thought of as a variant of two widely used methods. Polansky (2008, p. 73) summarizes a bootstrap percentile confidence region suggested by Efron and Tibshirani (1998). Section 2 explains an important relationship between prediction regions and confidence regions. Section 3 examines the method for multiple linear regression, Section 4 examines the method for variable selection, and Section 5 gives examples and simulations.

2 A Relationship Between Hyperellipsoidal Prediction and Confidence Regions

When teaching confidence intervals, it is often noted that by the central limit theorem, the probability that $\overline{Y}_n$ is within two standard deviations ($2SD(\overline{Y}) = 2\sigma/\sqrt{n}$) of μ is about 95%. Hence the probability that μ is within two standard deviations of $\overline{Y}_n$ is about 95%. Thus the interval $(\mu - 1.96S/\sqrt{n}, \mu + 1.96S/\sqrt{n})$ is a large sample 95% prediction interval for a future value of the sample mean $\overline{Y}_{n,f}$ if μ is known, while $(\overline{Y}_n - 1.96S/\sqrt{n}, \overline{Y}_n + 1.96S/\sqrt{n})$ is a large sample 95% confidence interval for the population mean μ. Note that the lengths of the two intervals are the same. Where the interval is centered, at the parameter μ or the statistic $\overline{Y}_n$, determines whether the interval is a confidence or a prediction interval.

The following theorem shows that the hyperellipsoid R_c centered at the statistic T_n is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\mu}$, but the hyperellipsoid centered at known $\boldsymbol{\mu}$ is a large sample $100(1-\delta)\%$ prediction region for a future value of the statistic $T_{f,n}$.

Theorem 1. Let the $100(1-\delta)$th percentile $D_{1-\delta}^2$ be a continuity point of the distribution of D^2. Assume that $D_{\boldsymbol{\mu}}^2(T_n, \boldsymbol{\Sigma}_T) \xrightarrow{D} D^2$, $D_{\boldsymbol{\mu}}^2(T_n, \hat{\boldsymbol{\Sigma}}_T) \xrightarrow{D} D^2$, and $\hat{D}_{1-\delta}^2 \xrightarrow{P} D_{1-\delta}^2$

where $P(D^2 \le D^2_{1-\delta}) = 1-\delta$. i) Then $R_c = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \hat{\boldsymbol{\Sigma}}_T) \le \hat{D}^2_{1-\delta}\}$ is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\mu}$, and if $\boldsymbol{\mu}$ is known, then $R_p = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\boldsymbol{\mu}, \hat{\boldsymbol{\Sigma}}_T) \le \hat{D}^2_{1-\delta}\}$ is a large sample $100(1-\delta)\%$ prediction region for a future value of the statistic $T_{f,n}$. ii) Region R_c contains $\boldsymbol{\mu}$ iff region R_p contains T_n.

Proof: i) Note that $D^2_{\boldsymbol{\mu}}(T_n, \hat{\boldsymbol{\Sigma}}_T) = D^2_{T_n}(\boldsymbol{\mu}, \hat{\boldsymbol{\Sigma}}_T)$. Thus the probability that R_c contains $\boldsymbol{\mu}$ is $P(D^2_{\boldsymbol{\mu}}(T_n, \hat{\boldsymbol{\Sigma}}_T) \le \hat{D}^2_{1-\delta}) \to 1-\delta$, and the probability that R_p contains $T_{f,n}$ is $P(D^2_{\boldsymbol{\mu}}(T_{f,n}, \hat{\boldsymbol{\Sigma}}_T) \le \hat{D}^2_{1-\delta}) \to 1-\delta$, as $n \to \infty$.

ii) $D^2_{\boldsymbol{\mu}}(T_n, \hat{\boldsymbol{\Sigma}}_T) \le \hat{D}^2_{1-\delta}$ iff $D^2_{T_n}(\boldsymbol{\mu}, \hat{\boldsymbol{\Sigma}}_T) \le \hat{D}^2_{1-\delta}$. $\square$

Hence if there was an iid sample $T_{1,n}, ..., T_{B,n}$ of the statistic, the Olive (2013) large sample $100(1-\delta)\%$ prediction region $\{\boldsymbol{w} : D^2(\overline{T}, \boldsymbol{S}_T) \le D^2_{(c)}\}$ for $T_{f,n}$ contains $E(T_n) = \boldsymbol{\mu}$ with asymptotic coverage $\ge 1-\delta$. To make the asymptotic coverage equal to $1-\delta$, use the large sample $100(1-\delta)\%$ confidence region $\{\boldsymbol{w} : D^2(T_{1,n}, \boldsymbol{S}_T) \le D^2_{(c)}\}$. The prediction region method bootstraps this procedure by using a bootstrap sample of the statistic $T^*_{1,n}, ..., T^*_{B,n}$. Centering the region at $T^*_{1,n}$ instead of $\overline{T}^*$ is not needed since the bootstrap sample is centered near T_n: the distribution of $\sqrt{n}(T_n - \boldsymbol{\mu})$ is approximated by the distribution of $\sqrt{n}(T^* - T_n)$ or by the distribution of $\sqrt{n}(T^* - \overline{T}^*)$. See equations (7), (12), and (13) below.

When the bootstrap is used, a large sample $100(1-\delta)\%$ confidence region for an $r \times 1$ parameter vector $\boldsymbol{\mu}$ is a set $\mathcal{A}_{n,B}$ such that $P(\boldsymbol{\mu} \in \mathcal{A}_{n,B}) \to 1-\delta$ as $n, B \to \infty$. Assume $n\boldsymbol{S}^*_T \xrightarrow{P} \boldsymbol{\Sigma}_A$ as $n, B \to \infty$ where $\boldsymbol{\Sigma}_A$ and $\boldsymbol{S}^*_T$ are nonsingular $r \times r$ matrices, and T_n is an estimator of $\boldsymbol{\mu}$ such that

$$\sqrt{n}\ \ (T_n - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{U} \tag{7}$$

as $n \to \infty$. Then

$$\sqrt{n}\ \boldsymbol{\Sigma}_A^{-1/2}\ (T_n - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{\Sigma}_A^{-1/2}\boldsymbol{U} = \boldsymbol{Z},$$

$$n\ (T_n - \boldsymbol{\mu})^T\ \hat{\boldsymbol{\Sigma}}_A^{-1}\ (T_n - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{Z}^T\boldsymbol{Z} = D^2$$

as $n \to \infty$ where $\hat{\boldsymbol{\Sigma}}_A$ is a consistent estimator of $\boldsymbol{\Sigma}_A$, and

$$(T_n - \boldsymbol{\mu})^T\ [\boldsymbol{S}^*_T]^{-1}\ (T_n - \boldsymbol{\mu}) \xrightarrow{D} D^2 \tag{8}$$

as $n, B \to \infty$. Assume the cumulative distribution function (cdf) of D^2 is continuous and increasing in a neighborhood of $D^2_{1-\delta}$ where $P(D^2 \le D^2_{1-\delta}) = 1-\delta$. If the distribution of D^2 is known, then a common bootstrap large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\mu}$ is

$$\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T[\boldsymbol{S}^*_T]^{-1}(\boldsymbol{w} - T_n) \le D^2_{1-\delta}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}^*_T) \le D^2_{1-\delta}\}. \tag{9}$$

Often by a central limit theorem or the multivariate delta method, $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} N_r(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$, and $D^2 \sim \chi^2_r$. Note that $[\boldsymbol{S}^*_T]^{-1}$ could be replaced by $n\hat{\boldsymbol{\Sigma}}_A^{-1}$. Machado and Parente (2005) provide sufficient conditions and references for when $n\boldsymbol{S}^*_T$ is a consistent estimator of $\boldsymbol{\Sigma}_T$.

Bickel and Ren (2001) use $n\hat{\boldsymbol{\Sigma}}_A^{-1}$ instead of $[\boldsymbol{S}^*_T]^{-1}$, and replace the D^2 cutoff in (9) by $D^2_{(k_B)}$ where $D^2_{(k_B)}$ is computed from $D^2_i = n(T^*_i - T_n)^T\hat{\boldsymbol{\Sigma}}_A^{-1}(T^*_i - T_n)$ for $i = 1, ..., B$.

If $n\boldsymbol{S}_T^* = \hat{\boldsymbol{\Sigma}}_A$, the (modified) large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\mu}$ is

$$\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \le D^2_{(U_B)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \le D^2_{(U_B)}\} \quad (10)$$

where $D^2_{(U_B)}$ is computed from $D_i^2 = (T_i^* - T_n)^T [\boldsymbol{S}_T^*]^{-1} (T_i^* - T_n)$ for $i = 1, ..., B$.

The prediction region method large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\mu}$ is

$$\{\boldsymbol{w} : (\boldsymbol{w} - \overline{T}^*)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - \overline{T}^*) \le D^2_{(U_B)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{S}_T^*) \le D^2_{(U_B)}\} \quad (11)$$

where $D^2_{(U_B)}$ is computed from $D_i^2 = (T_i^* - \overline{T}^*)^T [\boldsymbol{S}_T^*]^{-1} (T_i^* - \overline{T}^*)$ for $i = 1, ..., B$.

Given (7) and (8), a sufficient condition for (10) to be confidence region is

$$\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{U}, \quad (12)$$

while sufficient conditions for (11) to be confidence region are

$$\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{U}, \quad (13)$$

and

$$\sqrt{n}(\overline{T}^* - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{U}. \quad (14)$$

(We could replace $\boldsymbol{U}$ by $\boldsymbol{W}$ in (13) and (14), but $\boldsymbol{W} \sim \boldsymbol{U}$ works.) Note (13) and (14) follow from (12) and (7) if $\sqrt{n}(T_n - \overline{T}^*) \xrightarrow{P} \boldsymbol{0}$, so $T_n - \overline{T}^* = o_P(n^{-1/2})$.

Following Bickel and Ren (2001), let $\boldsymbol{\mu} = T(F)$, $T_n = T(F_n)$, and $T^* = T(F_n^*)$ where F is the cdf of iid $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, F_n is the empirical cdf, and F_n^* is the empirical cdf of $\boldsymbol{x}_1^*, ..., \boldsymbol{x}_n^*$, a sample from F_n using the nonparametric bootstrap. If $\sqrt{n}(F_n - F) \xrightarrow{D} \boldsymbol{z}_F$, a Gaussian random process, and if T is sufficiently smooth (Hadamard differentiable with a well behaved Hadamard derivative $\dot{T}(F)$), then (7) and (12) hold with $\boldsymbol{U} = \dot{T}(F)\boldsymbol{z}_F$. Note that F_n is a perfectly good cdf "F" and F_n^* is a perfectly good empirical cdf from $F_n =$ "F." Thus if n is fixed, and a sample of size m is drawn with replacement from the empirical distribution, then $\sqrt{m}(T(F_m^*) - T_n) \xrightarrow{D} \dot{T}(F_n)\boldsymbol{z}_{F_n}$. Now let $n \to \infty$ with $m = n$. Then bootstrap theory gives $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \lim_{n\to\infty} \dot{T}(F_n)\boldsymbol{z}_{F_n} = \dot{T}(F)\boldsymbol{z}_F \sim \boldsymbol{U}$.

To justify the prediction region method, assume that (7) and (12) hold where $\boldsymbol{U} \sim N_r(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$. Use $\boldsymbol{W}_n \sim AN_r(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ to indicate that a normal approximation is used: $\boldsymbol{W}_n \approx N_r(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$. Let $T_i^* = T_{i,n}^*$. Then $T_i^* \sim AN_r\left(T_n, \frac{\boldsymbol{\Sigma}_A}{n}\right)$. Fix n temporarily and let $\boldsymbol{W}_i = \sqrt{n}(T_i^* - T_n)$. Then with respect to the bootstrap distribution (so conditional on the data), $\boldsymbol{W}_1, ..., \boldsymbol{W}_B$ are iid, and $\sqrt{n}(\overline{T}^* - T_n) = \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{W}_i \sim AN_r\left(\boldsymbol{0}, \frac{\boldsymbol{\Sigma}_A}{B}\right)$ is a normal approximation. Hence $\sqrt{nB}(\overline{T}^* - T_n) \sim AN_r(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$. Now unfix n. Since the same normal approximation holds for n and B large (and $AN_r(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$ does not depend on n or B), it follows that $\overline{T}^* - T_n = o_P(n^{-1/2})$.

The prediction region method should often work if $E(\overline{T}^*) - T_n = o_P(n^{-1/2})$ and the asymptotic covariance matrix of $\sqrt{nB}(\overline{T}^* - T_n)$ is $\boldsymbol{\Sigma}_A$ as $n, B \to \infty$. Following Efron (2014), $\overline{T^*}$ is the bagging or smoothed bootstrap estimator of $\boldsymbol{\mu}$, which often outperforms

T_n for inference. See Büchlmann and Yu (2002) and Friedman and Hall (2007) for theory and references for the bagging estimator.

These results suggest that under reasonable conditions, (7), (12), (13), and (14) hold: $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{U}$, $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{U}$, $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{U}$, and $\sqrt{n}(\overline{T}^* - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{U}$. Stronger conditions are needed for $n\boldsymbol{S}_T^* \xrightarrow{P} \boldsymbol{\Sigma}_A$. The regularity conditions for the prediction region method are weaker when $r = 1$, since $\boldsymbol{S}_T^*$ does not need to be computed.

The following result is also informative. Let $T_i = T_{i,n}$, and assume $T_1, ..., T_B$ are iid where

$$\frac{n}{B}\sum_{i=1}^{B}(T_i - \boldsymbol{\mu})(T_i - \boldsymbol{\mu})^T \xrightarrow{P} \boldsymbol{\Sigma}_A \text{ and } \frac{\mathrm{n}}{\mathrm{B}}\sum_{\mathrm{i}=1}^{\mathrm{B}}(\mathrm{T}_\mathrm{i}^* - \overline{\mathrm{T}}^*)(\mathrm{T}_\mathrm{i}^* - \overline{\mathrm{T}}^*)^{\mathrm{T}} \xrightarrow{\mathrm{P}} \boldsymbol{\Sigma}_{\mathrm{A}}.$$

Then

$$\frac{n}{B}\sum_{i=1}^{B}(T_i - \boldsymbol{\mu})(T_i - \boldsymbol{\mu})^T - \frac{n}{B}\sum_{i=1}^{B}(T_i^* - \overline{T}^*)(T_i^* - \overline{T}^*)^T \xrightarrow{P} \mathbf{0}, \tag{15}$$

the $r \times r$ matrix of zeroes. The trace is a continuous linear function. Post multiply both sides of (15) by $[\boldsymbol{S}_T^*]^{-1}$, and take the trace of both sides to get

$$\frac{n}{B}\sum_{i=1}^{B}(T_i - \boldsymbol{\mu})^T[\boldsymbol{S}_T^*]^{-1}(T_i - \boldsymbol{\mu}) - \frac{n}{B}\sum_{i=1}^{B}(T_i^* - \overline{T}^*)^T[\boldsymbol{S}_T^*]^{-1}(T_i^* - \overline{T}^*) \xrightarrow{P} 0. \tag{16}$$

Now $(T_i - \boldsymbol{\mu})^T[\boldsymbol{S}_T^*]^{-1}(T_i - \boldsymbol{\mu}) - n(T_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}_A^{-1}(T_i - \boldsymbol{\mu}) \xrightarrow{P} 0$. Hence the first sum in (16) behaves like a sum of iid nonnegative terms that each converge in distribution to D^2. If n is fixed, then the T_i^* are iid with respect to the bootstrap distribution where $\overline{T}^* \approx E(T_i^*) = \boldsymbol{\mu}_n$ and $\boldsymbol{S}_T^* \approx Cov(T_i^*) = \boldsymbol{\Sigma}_n$ with respect to the bootstrap distribution. Hence the second sum in (16) behaves like a sum of iid nonnegative terms with respect to the bootstrap distribution.

The prediction region method will often simulate well even if B is rather small. Figure 1 shows 10%, 30%, 50%, 70%, 90%, and 98% prediction regions for a future value of T_f for two multivariate normal statistics. The plotted points are iid $T_1, ..., T_B$. If the T_i^* are iid from the bootstrap distribution, then $Cov(\overline{T}^*) \approx Cov(T)/B \approx \boldsymbol{\Sigma}_A/(nB)$. Consider the 90% region. Suppose many iid samples are generated to produce $\overline{T}^*$. By Theorem 1, if $\overline{T}^*$ is in the 90% prediction region with probability near 90%, then the confidence region should give simulated coverage near 90% and the volume of the confidence region should be near that of the 90% prediction region. If $B = 100$, then $\overline{T}^*$ falls in a covering region of the same shape as the prediction region, but centered near T_n and the lengths of the axes are divided by $\sqrt{B}$. Hence if $B = 100$, then the axes lengths are about one tenth of those in Figure 1. Hence when T_n falls within the 70% prediction region, the probability that $\overline{T}^*$ falls in the 90% prediction region is near one. If T_n is just within or just without the boundary of the 90% prediction region, $\overline{T}^*$ tends to be just within or just without of the 90% prediction region. Hence the coverage and volume of prediction region confidence region is near that of the nominal coverage 90% and near the volume of the 90% prediction region.

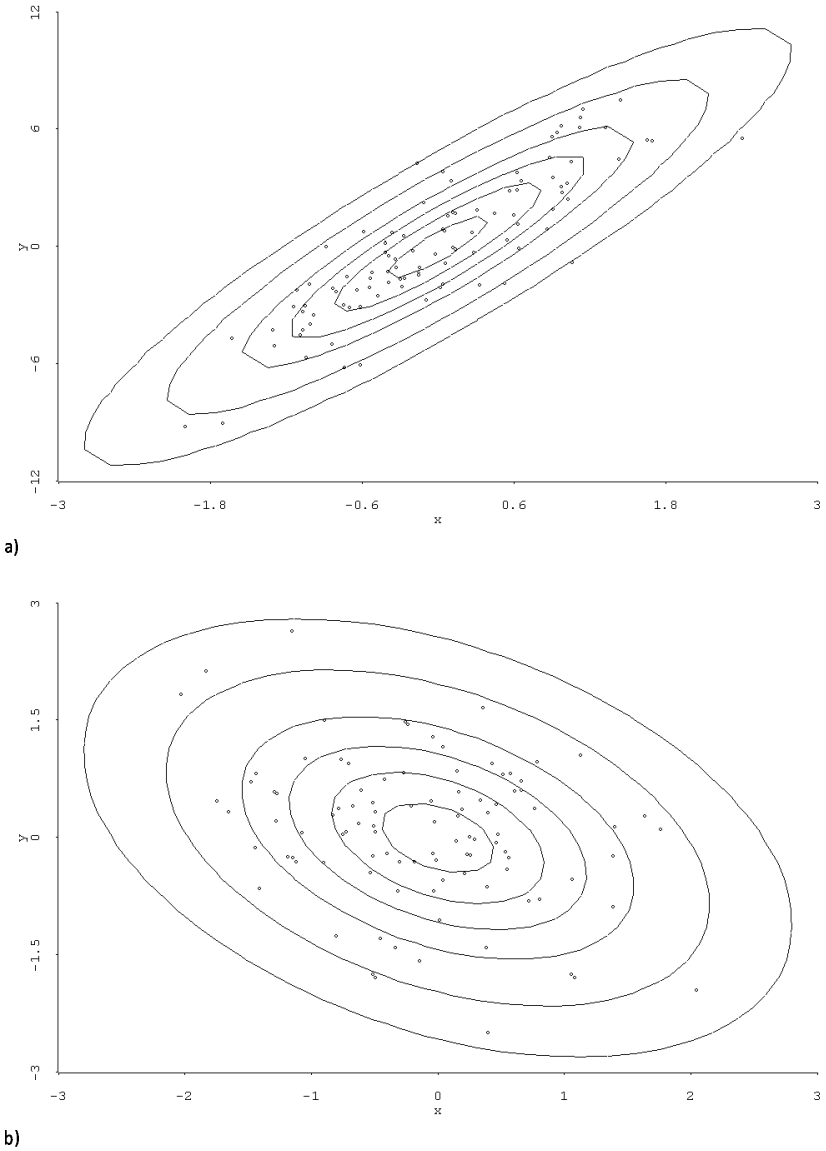

Figure 1: Confidence Regions for 2 Statistics with MVN Distributions

Hence B does not need to be large provided that n and B are large enough so that $S_T^* \approx Cov(T^*) \approx \mathbf{\Sigma}_A/n$. If n is large, the sample covariance matrix starts to be a good estimator of the population covariance matrix when $B \geq Jr$ where $J = 20$ or 50. For small r, using $B = 1000$ often led to good simulations, but $B = \max(50r, 100)$ may work well.

Often D^2 is unknown, and we use $D^2_{(U_B)}$ to estimate $D^2_{1-\delta}$ instead of assuming $D^2 \sim \chi^2_r$. Suppose the $T_i^* = T_{i,n}^*$ are iid from some distribution with cdf $\tilde{F}_n$. For example, if $T_{i,n}^* = t(F_n^*)$ where iid samples from F_n are used, then $\tilde{F}_n$ is the cdf of $t(F_n^*)$. Fix n, and let $E(T_{i,n}^*) = \boldsymbol{\mu}_n$ and $\mathrm{Cov}(T_{i,n}^*) = \mathbf{\Sigma}_n$. With respect to $\tilde{F}_n$, $\boldsymbol{\mu}_n$ and $\mathbf{\Sigma}_n$ are parameters, but with respect to F, $\boldsymbol{\mu}_n$ is a random vector and $\mathbf{\Sigma}_n$ is a random matrix. For example, using least squares and the residual bootstrap for the multiple linear regression model in Section 3, $\boldsymbol{\mu}_n = \hat{\boldsymbol{\beta}}$, $\mathbf{\Sigma}_n = \dfrac{n-p}{n} MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, and $\mathbf{\Sigma}_T = \sigma^2 \lim_{n\to\infty}(\boldsymbol{X}^T\boldsymbol{X}/n)^{-1}$. Then for fixed n, by the multivariate central limit theorem,

$$\sqrt{B}(\overline{T^*} - \boldsymbol{\mu}_n) \xrightarrow{D} N_r(\mathbf{0}, \mathbf{\Sigma}_n) \text{ and } \mathrm{B}(\overline{\mathrm{T}^*} - \boldsymbol{\mu}_\mathrm{n})^\mathrm{T}[\boldsymbol{S}_\mathrm{T}^*]^{-1}(\overline{\mathrm{T}^*} - \boldsymbol{\mu}_\mathrm{n}) \xrightarrow{\mathrm{D}} \chi^2_\mathrm{r}$$

as $B \to \infty$.

For $r = 1$, Efron (2014) uses confidence intervals $\overline{T^*} \pm z_{1-\delta} SE(\overline{T^*})$ where $P(Z \leq z_{1-\delta}) = 1 - \delta$ if $Z \sim N(0,1)$. Efron uses a delta method estimate of $SE(\overline{T^*})$ to avoid using the computationally expensive double bootstrap. The prediction region method, $\overline{T}^* \pm S_T^* D_{(U_B)}$, avoids assuming a normal limiting distribution and estimates the cutoff using quantiles of the Mahalanobis distances of the $T_{i,n}^*$ from $\overline{T^*}$. The shorth(c) estimator is recommended since it can be much shorter.

The following theorem will provide some intuition for why the percentile method works if T_n has a central limit theorem. Let Z_δ be the 100δth percentile of Z: $P(Z \leq Z_\delta) = \delta$, and let $P(Z_{\delta_L} \leq Z \leq Z_{\delta_U}) = 1 - \delta$. Let $T_{n,\delta}$ be the 100δth percentile of (the sampling distribution of) T_n. Then a population prediction interval for $T_{f,n}$ is $[T_{n,\delta_L}, T_{n,\delta_U}]$ which can be estimated by the sample percentiles $[T_{(c_L)}, T_{(c_U)}]$ when there is an iid sample $T_1, ..., T_n$. The shortest such interval can be estimated by the shorth.

Theorem 2. Suppose $r = 1$ and $\sqrt{n}(T_n - \mu) \xrightarrow{D} X$ and $\dfrac{\sqrt{n}(T_n - \mu)}{\sigma} \xrightarrow{D} \dfrac{1}{\sigma} X = W$. If the percentiles are continuity points of the distribution of W, then for each large sample $100(1-\delta)\%$ PI $[T_{(c_L)}, T_{(c_U)}]$ for $T_{f,n}$, there is a large sample $100(1-\delta)\%$ CI for μ $\left[T_n - W_{\delta_U}\dfrac{\hat{\sigma}}{\sqrt{n}}, T_n - W_{\delta_L}\dfrac{\hat{\sigma}}{\sqrt{n}}\right]$ with approximately the same length.

Proof. Note that $1 - \delta = P(T_{n,\delta_L} \leq T_n \leq T_{n,\delta_U}] \approx P(T_{(c_L)} \leq T_n \leq T_{(c_U)}) \approx$
$P(W_{\delta_L} \leq \dfrac{\sqrt{n}(T_n - \mu)}{\sigma} \leq W_{\delta_U}) = P(T_n - W_{\delta_U}\dfrac{\sigma}{\sqrt{n}} \leq \mu \leq T_n - W_{\delta_L}\dfrac{\sigma}{\sqrt{n}}) =$
$P(W_{\delta_L}\dfrac{\sigma}{\sqrt{n}} + \mu \leq T_n \leq W_{\delta_U}\dfrac{\sigma}{\sqrt{n}} + \mu)$. Hence $T_{n,\delta_L} \approx W_{\delta_L}\dfrac{\sigma}{\sqrt{n}} + \mu$ and $T_{n,\delta_U} \approx W_{\delta_U}\dfrac{\sigma}{\sqrt{n}} + \mu$.
Thus $T_{(c_U)} - T_{(c_L)} \approx \dfrac{\sigma}{\sqrt{n}}(W_{\delta_U} - W_{\delta_L}) \approx T_{n,\delta_U} - T_{n,\delta_L}$. $\square$

Theorem 2 suggests that the Frey (2013) shorth(c) interval applied to the bootstrap sample estimates the shortest large sample $100(1-\delta)\%$ CI $\left[T_n - W_{\delta_U}\dfrac{\hat{\sigma}}{\sqrt{n}}, T_n - W_{\delta_L}\dfrac{\hat{\sigma}}{\sqrt{n}}\right]$

based on the asymptotic pivot. Note that if $Z_i = T_n + \mu - T_i$ for $i = 1, ..., n$, then $P(Z_{(c_L)} \leq \mu \leq Z_{(c_U)}) \approx P(T_n + \mu - T_{n,\delta_U} \leq \mu \leq T_n + \mu - T_{n,\delta_L}) = P(T_{n,\delta_L} \leq T_n \leq T_{n,\delta_U}) = 1 - \delta$. Then the Z_i are centered at T_n with deviations equal to $\mu - T_i$. Note that the distribution of $T_n - \mu$ is the same as the distribution of $T_i - \mu$: $T_i - \mu \stackrel{D}{=} T_n - \mu$. Now the bootstrap approximation says that the distribution of $T_n - \mu$ can be approximated by the distribution of $T_i^* - T_n$. Thus $T_i - \mu \stackrel{D}{=} T_n - \mu \approx T_i^* - T_n$, or $T_i^* \approx T_i + T_n - \mu$. If the distribution of $T_n - \mu$ is approximately the same as the distribution of $\mu - T_n$ (asymptotic symmetry), then the percentile method should work. Since $\sqrt{n}(T_n - \mu) \stackrel{D}{\to} X$, we have $n^\gamma(T_n - \mu) \stackrel{D}{\to} 0$ if $0 < \gamma < 0.5$. The point mass at 0 is a symmetric distribution, and $n^\gamma(T_i + T_n - \mu) \approx n^\gamma \mu$ for large n.

3 Bootstrap Tests for Multiple Linear Regression

Consider the multiple linear regression model $Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i$ for $i = 1, ..., n$, written in matrix form as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{Y}$ is $n \times 1$ and $\boldsymbol{X}$ is $n \times p$. Consider testing $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{c}$ where $\boldsymbol{A}$ is an $r \times p$ matrix with full rank r and $\boldsymbol{\mu} = \boldsymbol{A}\boldsymbol{\beta}$. To perform the test, suppose a bootstrap sample $\hat{\boldsymbol{\beta}}_1^*, ..., \hat{\boldsymbol{\beta}}_B^*$ has been generated. Form the prediction region (5) for $\boldsymbol{w}_1 = \boldsymbol{A}\hat{\boldsymbol{\beta}}_1^* - \boldsymbol{c}, ..., \boldsymbol{w}_B = \boldsymbol{A}\hat{\boldsymbol{\beta}}_B^* - \boldsymbol{c}$. If $\boldsymbol{0}$ is in the prediction region, fail to reject H_0, otherwise reject H_0.

It is useful to compare the bootstrap tests with classical tests. Methods for bootstrapping this model are well known. The estimated covariance matrix of the (ordinary) least squares estimator is

$$\widehat{Cov}(\hat{\boldsymbol{\beta}}_{OLS}) = MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}.$$

The residual bootstrap computes the least squares estimator and obtains the n residuals and fitted values $r_1, ..., r_n$ and $\hat{Y}_1, ..., \hat{Y}_n$. Then a sample of size n is selected with replacement from the residuals resulting in $r_{11}^*, ..., r_{n1}^*$. Hence the empirical distribution of the residuals is used. Then a vector $\boldsymbol{Y}_1^* = (Y_{11}^*, ..., Y_{n1}^*)^T$ is formed where $Y_{j1}^* = \hat{Y}_j + r_{j1}^*$. Then $\boldsymbol{Y}_1^*$ is regressed on $\boldsymbol{X}$ resulting in the estimator $\hat{\boldsymbol{\beta}}_1^*$. This process is repeated B times resulting in the estimators $\hat{\boldsymbol{\beta}}_1^*, ..., \hat{\boldsymbol{\beta}}_B^*$. This method should have $n \geq 10p$ so that the residuals r_i are close to the errors e_i.

Efron (1982, p. 36) notes that for the residual bootstrap, the sample covariance matrix of the $\hat{\boldsymbol{\beta}}_i^*$ is estimating the population bootstrap matrix $\frac{n-p}{n} MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ as $B \to \infty$. Hence the residual bootstrap standard error $SE(\hat{\beta}_i^*) \approx \sqrt{\frac{n-p}{n}}\, SE(\hat{\beta}_{i,OLS})$.

If the $\boldsymbol{z}_i = (Y_i, \boldsymbol{x}_i^T)^T$ are iid observations from some population, then a sample of size n can be drawn with replacement from $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$. Then the response and predictor variables can be formed into vector $\boldsymbol{Y}_1^*$ and design matrix $\boldsymbol{X}_1^*$. Then $\boldsymbol{Y}_1^*$ is regressed on $\boldsymbol{X}_1^*$ resulting in the estimator $\hat{\boldsymbol{\beta}}_1^*$. This process is repeated B times resulting in the estimators $\hat{\boldsymbol{\beta}}_1^*, ..., \hat{\boldsymbol{\beta}}_B^*$. This nonparametric bootstrap uses the empirical distribution of the cases $\boldsymbol{z}_i$ where often the $\boldsymbol{z}_i^T$ are the rows of a matrix $\boldsymbol{Z}$.

Following Seber and Lee (2003, p. 100), the classical test statistic for testing H_0 is

$$F_R = \frac{(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{c})^T [MSE \;\; \boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{c})}{r},$$

and when H_0 is true, $rF_R \xrightarrow{D} \chi^2_r$ for a large class of error distributions. The sample covariance matrix $\boldsymbol{S_w}$ of the $\boldsymbol{w}_i$ is estimating $\frac{n-p}{n} MSE \;\; \boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T$, and $\overline{\boldsymbol{w}} \approx \boldsymbol{0}$ when H_0 is true. Thus under H_0, the squared distance $D_i^2 = (\boldsymbol{w}_i - \overline{\boldsymbol{w}})^T \boldsymbol{S}_{\boldsymbol{w}}^{-1} (\boldsymbol{w}_i - \overline{\boldsymbol{w}}) \approx$

$$\frac{n}{n-p}(\boldsymbol{A}\hat{\boldsymbol{\beta}}^* - \boldsymbol{c})^T [MSE \;\; \boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}}^* - \boldsymbol{c}),$$

and we expect $D^2_{(U_B)} \approx \frac{n}{n-p}\chi^2_{r,1-\delta}$, for large n and B and small p. Hence the prediction region method is closely related to the Mammen (1993) suggestion of bootstrapping the test statistic F_R for this model.

4 Bootstrapping the Variable Selection Estimator

Variable selection, also called subset or model selection, is the search for a subset of predictor variables that can be deleted without important loss of information. By treating a variable selection estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ as a shrinkage estimator, the bootstrap can be used to examine variable selection. Forward selection, backward elimination, stepwise regression, and all subsets variable selection can be used if there is a criterion that selects the submodel, such as AIC or C_p. Similar ideas can be used to bootstrap other shrinkage estimators.

Consider testing $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{c}$ where $\boldsymbol{A}$ is an $r \times p$ matrix with full rank r. Now let $\hat{\boldsymbol{\beta}}$ be a variable selection estimator of $\boldsymbol{\beta}$. To perform the test, suppose a bootstrap sample $\hat{\boldsymbol{\beta}}^*_1, ..., \hat{\boldsymbol{\beta}}^*_B$ has been generated. Form the prediction region (5) for $\boldsymbol{w}_1 = \boldsymbol{A}\hat{\boldsymbol{\beta}}^*_1 - \boldsymbol{c}, ..., \boldsymbol{w}_B = \boldsymbol{A}\hat{\boldsymbol{\beta}}^*_B - \boldsymbol{c}$. If $\boldsymbol{0}$ is in the prediction region, fail to reject H_0, otherwise reject H_0.

A *model for variable selection* in multiple linear regression can be described by

$$Y = \boldsymbol{x}^T\boldsymbol{\beta} + e = \boldsymbol{\beta}^T\boldsymbol{x} + e = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E + e = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + e \tag{17}$$

where e is an error, Y is the response variable, $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$ is a $p \times 1$ vector of predictors, $\boldsymbol{x}_S$ is a $k_S \times 1$ vector and $\boldsymbol{x}_E$ is a $(p - k_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$ and E denotes the subset of terms that can be eliminated given that the subset S is in the model.

Since S is unknown, candidate subsets will be examined. Following Olive and Hawkins (2005), let $\boldsymbol{x}_I$ be the vector of k terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). Then

$$Y = \boldsymbol{x}_I^T\boldsymbol{\beta}_I + \boldsymbol{x}_O^T\boldsymbol{\beta}_O + e. \tag{18}$$

The model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$ that uses all of the predictors is called the *full model*. A model $Y = \boldsymbol{x}_I^T\boldsymbol{\beta}_I + e$ that only uses a subset $\boldsymbol{x}_I$ of the predictors is called a *submodel*.

Suppose that S is a subset of I and that model (17) holds. Then

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_{I/S}^T\boldsymbol{\beta}_{(I/S)} + \boldsymbol{x}_O^T\boldsymbol{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I \tag{19}$$

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

For multiple linear regression, if the candidate model of $\boldsymbol{x}_I$ has k terms (including the constant), then the partial F statistic for testing whether the $p-k$ predictor variables in $\boldsymbol{x}_O$ can be deleted is

$$F_I = \frac{SSE(I) - SSE}{(n-k)-(n-p)} / \frac{SSE}{n-p} = \frac{n-p}{p-k}\left[\frac{SSE(I)}{SSE} - 1\right]$$

where SSE is the error sum of squares from the full model and SSE(I) is the error sum of squares from the candidate submodel. An important criterion for variable selection is the C_p criterion

$$C_p(I) = \frac{SSE(I)}{MSE} + 2k - n = (p-k)(F_I - 1) + k$$

where MSE is the mean square error for the full model. Olive and Hawkins (2005) show that submodels with $C_p(I) \leq \min(2k, p)$ are especially interesting. The AIC is criterion similar to C_p.

Other automated variable selection methods may work better than I_{min}. For the C_p criterion, find the submodel I_I with the fewest number of predictors such that $C_p(I_I) \leq C_p(I_{min}) + 1$. For AIC, Burnham and Anderson (2004) suggest that if $\Delta(I) = AIC(I) - AIC(I_{min})$, then models with $\Delta(I) \leq 2$ are good. Find the submodel I_I with the smallest number of predictors such that $\Delta(I_I) \leq 2$. It is possible that $I_I = I_{min}$ or that I_I is the full model. Do not use more predictors than model I_I to avoid overfitting.

Suppose model I is selected after variable selection. Then least squares output for the model $\boldsymbol{Y} = \boldsymbol{X}_I\boldsymbol{\beta}_I + \boldsymbol{e}$ can be obtained, but the least squares output is not correct for inference. In particular, $MSE(I)(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}$ is not the correct estimated covariance matrix of $\hat{\boldsymbol{\beta}}_I$. The selected model tends to fit the data too well, so $SE(\hat{\beta}_i)$ from the incorrect estimated covariance matrix is too small. Hence the confidence intervals for β_i are too short, and hypothesis tests reject $H_0 : \beta_i = 0$ too often.

Hastie, Tibshirani, and Friedman (2009, p. 57) note that variable selection is a shrinkage estimator: the coefficients are shrunk to 0 for the omitted variables. Suppose $n \geq 10p$. If $\hat{\boldsymbol{\beta}}_I$ is $k \times 1$, form $\hat{\boldsymbol{\beta}}_{I,0}$ from $\hat{\boldsymbol{\beta}}_I$ by adding 0s corresponding to the omitted variables. Then $\hat{\boldsymbol{\beta}}_{I,0}$ is a nonlinear estimator of $\boldsymbol{\beta}$, and the residual bootstrap method can be applied. For example, suppose $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{I_{min},0}$ is formed from model I_{min} that minimizes C_p from some variable selection method such as forward selection, backward elimination, stepwise selection, or all subsets variable selection. Instead of computing the least squares estimator from regressing $\boldsymbol{Y}_i^*$ on $\boldsymbol{X}$, perform variable selection on $\boldsymbol{Y}_i^*$ and $\boldsymbol{X}$, fit the model that minimizes the criterion, and add 0s corresponding to the omitted variables, resulting in estimators $\hat{\boldsymbol{\beta}}_1^*, ..., \hat{\boldsymbol{\beta}}_B^*$.

Suppose the variable selection method, such as forward selection or all subsets, produces K models. Let model I_{min} be the model that minimizes the criterion, e.g. $C_p(I)$

or $AIC(I)$. 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 $n \to \infty$. Since there are a finite number of regression models I that contain the true model, and each model gives a consistent estimator $\hat{\boldsymbol{\beta}}_{I,0}$ of $\boldsymbol{\beta}$, the probability that I_{min} picks one of these models goes to one as $n \to \infty$. Hence $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is a consistent estimator of $\boldsymbol{\beta}$ under model (17).

Note that if $S \subseteq I$, and $\boldsymbol{Y} = \boldsymbol{X}_I\boldsymbol{\beta}_I + \boldsymbol{e}_I$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I) \xrightarrow{D} N_k(\boldsymbol{0}, \sigma_I^2 \boldsymbol{W}_I)$ under mild regularity conditions where $n(\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1} \to \boldsymbol{W}_I$. Hence $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I,0} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma_I^2 \boldsymbol{W}_{I,0})$ where the $\boldsymbol{W}_{I,0}$ has a column and row of zeroes added for each variable not in I. Note that $\boldsymbol{W}_{I,0}$ is singular unless I corresponds to the full model. For example, if $p = 3$ and model I uses a constant and x_3 with

$$\boldsymbol{W}_I = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}, \quad \text{then} \quad \boldsymbol{W}_{\mathrm{I,0}} = \begin{bmatrix} W_{11} & 0 & W_{12} \\ 0 & 0 & 0 \\ W_{21} & 0 & W_{22} \end{bmatrix}.$$

Hence it is reasonable to conjecture that $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_{min},0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{U}$ where

$$\boldsymbol{U} = \sum_{i=1}^{K} \pi_i N_p(\boldsymbol{0}, \sigma_{I_i}^2 \boldsymbol{W}_{I_i,0}),$$

$0 \le \pi_i \le 1$, $\sum_{i=1}^{K} \pi_i = 1$, and K is the number of subsets I_i that contain S.

Before Efron (2014), inference techniques for the variable selection model have not had much success. Efron (2014) let $t(\boldsymbol{Z})$ be a scalar valued statistic, based on all of the data $\boldsymbol{Z}$, that estimates a parameter of interest μ. Form a bootstrap sample $\boldsymbol{Z}_i^*$ and $t(\boldsymbol{Z}_i^*)$ for $i = 1, ..., B$. Then $\tilde{\mu} = s(\boldsymbol{Z}) = \dfrac{1}{B}\sum_{i=1}^{B} t(\boldsymbol{Z}_i^*)$, a "bootstrap smoothing" or "bagging" estimator. In the regression setting with variable selection, $\boldsymbol{Z}_i^*$ can be formed with the nonparametric or residual bootstrap using the full model. The prediction region method can also be applied to $t(\boldsymbol{Z})$. For example, when $\boldsymbol{A}$ is $1 \times p$, the prediction region method uses $\mu = \boldsymbol{A}\boldsymbol{\beta} - c$, $t(\boldsymbol{Z}) = \boldsymbol{A}\hat{\boldsymbol{\beta}} - c$ and $\overline{T^*} = \tilde{\mu}$. Efron (2014) used the confidence interval $\overline{T^*} \pm z_{1-\delta} SE(\overline{T^*})$ which is symmetric about $\overline{T^*}$. The prediction region method uses $\overline{T^*} \pm S_T^* D_{(U_B)}$ which is also a symmetric interval centered at $\overline{T^*}$. If both the prediction region method and Efron's method are large sample confidence intervals for μ, then they have the same asymptotic length (scaled by multiplying by $\sqrt{n}$), since otherwise the shorter interval will have lower asymptotic coverage. Since the prediction region interval is a percentile interval, the shorth(c) interval could have much shorter length than the Efron interval and the prediction region interval if the bootstrap distribution is not symmetric.

The prediction region method can be used for vector valued statistics and parameters, and does not need the statistic to be asymptotically normal. These features are likely useful for variable selection models. Prediction intervals and regions can have higher than the nominal coverage $1 - \delta$ if the distribution is discrete or a mixture of a discrete distribution and some other distribution. In particular, coverage can be high if the $\boldsymbol{w}_i$ distribution is a mixture of a point mass at $\boldsymbol{0}$ and the method checks whether $\boldsymbol{0}$ is in the

prediction region. Such a mixture often occurs for variable selection methods and lasso. The bootstrap sample for the $W_i = \hat{\boldsymbol{\beta}}^*_{ij}$ can contain many zeroes and be highly skewed if the jth predictor is weak. Then the computer program may fail because $\boldsymbol{S_w}$ is singular, but if all or nearly all of the $\hat{\boldsymbol{\beta}}^*_{ij} = 0$, then there is strong evidence that the jth predictor is not needed given that the other predictors are in the variable selection method.

As an extreme simulation case, suppose $\hat{\boldsymbol{\beta}}^*_{ij} = 0$ for $i = 1, ..., B$ and for each run in the simulation. Consider testing $H_0 : \beta_j = 0$. Then regardless of the nominal coverage $1 - \delta$, the closed interval [0,0] will contain 0 for each run and the observed coverage will be $1 > 1-\delta$. Using the open interval (0,0) would give observed coverage 0. Also intervals $[0, b]$ and $[a, 0]$ correctly suggest failing to reject $\beta_j = 0$, while intervals $(0, b)$ and $(a, 0)$ incorrectly suggest rejecting $H_0 : \beta_j = 0$. Hence closed regions and intervals make sense.

5 Example and Simulations

Example. Cook and Weisberg (1999, pp. 351, 433, 447) gives a data set on 82 mussels sampled off the coast of New Zealand. Let the response variable be the logarithm $\log(M)$ of the *muscle mass* M, and the predictors are the *length* L and *height* H of the shell in mm, the logarithm $\log(W)$ of the *shell width* W, the logarithm $\log(S)$ of the *shell mass* S, and a constant. Table 1 shows results for the full model including the shorth(c) nominal 95% confidence intervals for β_i computed using the nonparametric and residual bootstraps. As expected, the residual bootstrap intervals are close to the classical least squares confidence intervals $\approx \hat{\beta}_i \pm 2SE(\hat{\beta}_i)$. The minimum C_p model used a constant, H, and $\log(S)$. Table 2 shows results for this variable selection model including the shorth(c) nominal 95% confidence intervals for β_i using the residual bootstrap. Note that the interval for H is right skewed and contains 0 when closed intervals are used instead of open intervals.

It was expected that $\log(S)$ may be the only predictor needed, along with a constant, since $\log(S)$ and $\log(M)$ are both log(mass) measurements and likely highly correlated. Hence we want to test $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ with the I_{min} model selected by all subsets variable selection. (Of course this test would be easy to do with the full model using least squares theory.) Then $H_0 : \boldsymbol{A\beta} = (\beta_2, \beta_3, \beta_4)^T = \boldsymbol{0}$. Using the prediction region method with the full model gave an interval [0,2.930] with $D_{\boldsymbol{0}} = 1.641$. 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} variable selection model had $[0, D_{(U_B)}] = [0, 3.293]$ while $D_{\boldsymbol{0}} = 1.134$. So fail to reject H_0.

A small simulation study was done in R using $B = \max(1000, n, 20r)$ and 5000 runs. The regression model used $\boldsymbol{\beta} = (1, 1, 0, 0)^T$ with $n = 100$, $p = 4$, and various zero mean iid error distributions. The design matrix $\boldsymbol{X}$ consisted of iid N(0,1) random variables. Hence the full model least squares confidence intervals for β_i should have length near $2t_{96,0.975}\sigma/\sqrt{n} \approx 2(1.96)\sigma/10 = 0.392\sigma$ when the iid zero mean errors have variance σ^2. The simulation computed the shorth(c) interval for each β_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.

Table 1: Bootstrapping the Full Model

variable	$\hat{\beta}$	OLS SE	OLS CI	rowboot	resboot
constant	−1.2493	0.8388	(−2.920,0.421)	[−2.93,−0.048]	[−3.138,0.194]
L	−0.0006	0.0023	(−0.005,0.004)	[−0.005,0.003]	[−0.005,0.004]
logW	0.1298	0.3738	(−0.615,0.874)	[−0.384,0.827]	[−0.555,0.971]
H	0.0075	0.0050	(−0.002,0.018)	[−0.002,0.018]	[−0.003,0.017]
logS	0.6404	0.1686	(0.305,0.976)	[0.188,1.001]	[0.276,0.955]

Table 2: Bootstrapping the Variable Selection Model

variable	$\hat{\beta}_{I_{min},0}$	OLS SE	resboot
constant	-0.9573	0.1519	[−2.769,0.460]
L	0		[−0.004, 0.004]
logW	0		[−0.595, 0.869]
H	0.0072	0.0047	[0.000, 0.016]
logS	0.6530	0.1160	[0.324, 0.913]

The regression models used the residual bootstrap on the full model least squares estimator and on the all subsets variable selection estimator for the model I_{min}. The residuals were from least squares applied to the full model in both cases. Results are shown for when the iid errors $e_i \sim N(0,1)$. Table 3 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 the all subsets variable selection. The column for the "test" gives the length and coverage = P(fail to reject H_0) for the interval $[0, D_{(U_B)}]$ where $D_{(U_B)}$ is the cutoff for the confidence region. The volume of the confidence region will decrease to 0 as $n \to \infty$. The cutoff will often be near $\sqrt{\chi^2_{r,0.95}}$ if the statistic T is asymptotically normal. Note that $\sqrt{\chi^2_{2,0.95}} = 2.448$ is very close to 2.4493 for the full model regression bootstrap test. The coverages were near 0.95 for the regression bootstrap on the full model. For I_{min} the coverages were near 0.95 for β_1 and β_2, but higher for the other 3 tests since zeroes often occurred for $\hat{\beta}^*_j$ for $j = 3, 4$. The average lengths and coverages were similar for the full model and all subsets variable selection I_{min} for β_1 and β_2, but the lengths are shorter for I_{min} for β_3 and β_4. Volumes of the hyperellipsoids were not computed, but the average cutoff of 2.69 for the variable selection test suggests that the test statistic was not multivariate normal, which is not surprising since many zeroes were produced for $\hat{\beta}^*_j$ for $j = 3, 4$.

Larger sample sizes n are needed as r increases. Olive (2013) suggested that for iid elliptically contoured data $\boldsymbol{x}_i$ where $\boldsymbol{x}_i$ is $p \times 1$, the prediction region coverage for a future value $\boldsymbol{x}_f$ started to get close to the nominal coverage when $n \geq 20p$, but volume ratios needed $n \geq 50p$. Hence we may need $B \geq 50r$ for the confidence region to have small volume. Consider testing whether correlations in a correlation matrix are 0.

Table 3: Bootstrapping Regression and Variable Selection

model	cov/len	β_1	β_2	β_3	β_4	test
reg	cov	0.9496	0.9430	0.9440	0.9454	0.9414
	len	0.3967	0.3996	0.3997	0.3997	2.4493
vs	cov	0.9482	0.9486	0.9974	0.9974	0.9896
	len	0.3965	0.3990	0.3241	0.3257	2.6901

Table 4: Bootstrapping the Correlation Matrix

n	ψ	cov/len	ρ_{12}	ρ_{13}	ρ_{14}	ρ_{23}	ρ_{24}	ρ_{34}	test
100	0	cov	0.943	0.939	0.942	0.937	0.940	0.941	0.848
		len	0.391	0.391	0.391	0.391	0.392	0.392	3.549
400	0	cov	0.944	0.948	0.943	0.946	0.950	0.952	0.923
		len	0.199	0.199	0.199	0.199	0.199	0.199	3.559
400	0.03	cov	0.950	0.950	0.948	0.949	0.948	0.951	0.441
		len	0.198	0.198	0.198	0.198	0.198	0.198	3.558
400	0.1	cov	0.947	0.949	0.952	0.949	0.952	0.951	0.000
		len	0.190	0.190	0.189	0.190	0.189	0.189	3.561

There are $r = p(p-1)/2$ correlations $\rho_{i,j} = cor(X_i, X_j)$ where $i < j$. The simulation simulated iid data $\boldsymbol{w}$ with $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{w}$ and $\boldsymbol{A}_{ij} = \psi$ for $i \neq j$ and $\boldsymbol{A}_{ii} = 1$. Hence $\text{cor}(X_i, X_j) = [2\psi + (p-2)\psi^2]/[1 + (p-1)\psi^2]$. Let $\boldsymbol{\mu} = (\rho_{12}, ..., \rho_{1p}, \rho_{23}, ..., \rho_{2p}, ..., \rho_{p-1,p})^T$.

Table 4 shows the results for multivariate normal data with $p = 4$ so $r = 6$ for testing $H_0 : \boldsymbol{\mu} = \boldsymbol{0}$. The nominal coverage was 0.95. For $n = 100$ and $\psi = 0$, the test failed to reject H_0 84.8% of the time, but 92% of the time for $n = 400$. Note that $\sqrt{\chi^2_{6,0.95}} = 3.548$. With $n = 400$ and $\psi > 0$, for the test the coverage $= 1 -$ power. For $\psi = 0.03$ the simulated power was 0.56, but 1.0 for $\psi = 0.1$.

6 Conclusions

Let $T_n = T_{1,n}$ where $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{U}$, and suppose there was an iid sample $T_{1,n}, ..., T_{B,n}$. Then standard inference techniques could be used to examine how the statistic T_n behaves. Usually there is only one sample and one value of the statistic T_n, but if the empirical distribution is well behaved, and if the statistic T_n is sufficiently smooth, then bootstrap sample of the statistic $T_1^*, ..., T_B^*$ is useful: $T_1^* - T_n, ..., T_B^* - T_n$ is pseudodata for $T_{1,n} - \boldsymbol{\mu}, ..., T_{B,n} - \boldsymbol{\mu}$, and applying the Olive (2013) large sample $100(1-\delta)\%$ prediction region to the $T_1^*, ..., T_B^*$ results in a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\mu}$. If T_n is asymptotically normal, then under regularity conditions, the large sample confidence region and equivalent hypothesis test are closely related to applying the Hotelling's T^2 test statistic and confidence region to the $T_1^*, ..., T_B^*$.

Applications of the prediction region method are numerous, but we may need $n \geq 50r$ and $B \geq \max(100, n, 50r)$ if the test statistic has an approximate multivariate normal distribution. Sample sizes may need to be much larger for other limiting distributions. An abbreviated version of this manuscript is Olive (2017a), and see Olive (2017b, ch. 5) for more on the prediction region method.

A similar technique can be used to estimate the $100(1-\delta)\%$ Bayesian credible region for $\boldsymbol{\theta}$. Generate $B = \max(100000, n)$ values of $\hat{\boldsymbol{\theta}}$ from the posterior distribution, and compute the prediction region (5). Olive (2014, p. 364) used the shorth estimator to estimate Bayesian credible intervals. The mussels data was obtained from (http://lagrange.math.siu.edu/Olive/lregdata.txt).

Theorem 1 says that the hyperellipsoidal prediction and confidence regions have exactly the same volume. We compensate for the prediction region undercoverage when n is moderate by using $D^2_{(U_n)}$. If n is large, by using $D^2_{(U_B)}$, the prediction region method confidence region compensates for undercoverage when B is moderate, say $B \geq Jr$ where $J = 20$ or 50. This result can be useful if a simulation with $B = 1000$ or $B = 10000$ is much slower than a simulation with $B = Jr$. The price to pay is that the prediction region method confidence region is inflated to have better coverage, so the power of the hypothesis test is decreased if moderate B is used instead of larger B.

Simulations were done in R. See R Core Team (2016). The collection of R functions *mpack*, available at (http://lagrange.math.siu.edu/Olive/mpack.txt), has some useful functions for the prediction region method. The function `vselboot` bootstraps the minimum C_p model from all subsets variable selection. The function `shorth3` can be used to find the Frey (2013) shorth(c) intervals for $\hat{\mu}_i$. The function `predreg` computes the prediction region and the Mahalanobis distance of the zero vector corresponding to $\boldsymbol{A\theta} - \boldsymbol{c} = \boldsymbol{0}$. The functions `rowboot` and `regboot` do the nonparametric and residual bootstrap for the full model. The functions `regbootsim` and `vsbootsim` can be used to simulate the bootstrap tests for multiple linear regression and for the all subsets variable selection model that minimizes C_p. The functions `corboot` and `corbootsim` can be used to bootstrap the correlation matrix.

R code for Tables 1 and 2 is below.

```
library(leaps)
y <- log(mussels[,5]); x <- mussels[,1:4]
x[,4] <- log(x[,4]); x[,2] <- log(x[,2])
out <- regboot(x,y,B=1000)
tem <- rowboot(x,y,B=1000)
outvs <- vselboot(x,y,B=1000) #get bootstrap CIs,
apply(out$betas,2,shorth3);
apply(tem$betas,2,shorth3);
apply(outvs$betas,2,shorth3)
ls.print(outvs$full)
ls.print(outvs$sub)
#test if beta_2 = beta_3 = beta_4 = 0
Abeta <- out$betas[,2:4]
#prediction region method with residual bootstrap
```

```
predreg(Abeta)
Abeta <- outvs$betas[,2:4]
#prediction region method with Imin
predreg(Abeta)
```

R code for Table 3 is below.

```
regbootsim(nruns=5000) #takes a while
library(leaps)
vsbootsim(nruns=5000)  #takes a long while
vsbootsim2(nruns=5000) #bootstraps forwards selection
```

R code for Table 4 is below.

```
corbootsim(type=1,n=100,nruns=5000)
corbootsim(type=1,n=400,nruns=5000) #takes a few minutes
corbootsim(type=1,n=400,psi=0.03,nruns=5000)
corbootsim(type=1,n=400,psi=0.1,nruns=5000)
```

7 References

Beran, R. (1988), "Prepivoting Test Statistics: a Bootstrap View of Asymptotic Refinements," *Journal of the American Statistical Association*, 83, 686-697.

Bickel, P.J., and Freedman, D.A. (1981), "Some Asymptotic Theory for the Bootstrap," *The Annals of Statistics*, 1196-1217.

Bickel, P.J., and Ren, J.–J. (2001), "The Bootstrap in Hypothesis Testing," in *State of the Art in Probability and Statistics: Festschrift for William R. van Zwet*, eds. de Gunst, M., Klaassen, C., and van der Vaart, A., The Institute of Mathematical Statistics, Hayward, CA, 91-112.

Büchlmann, P., and Yu, B. (2002), "Analyzing Bagging," *The Annals of Statistics*, 30, 927-961.

Burnham, K.P., and Anderson, D.R. (2004), "Multimodel Inference Understanding AIC and BIC in Model Selection," *Sociological Methods & Research*, 33, 261-304.

Cook, R.D., and Weisberg, S. (1999), *Applied Regression Including Computing and Graphics,* Wiley, New York, NY.

Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans,* SIAM, Philadelphia, PA.

Efron, B. (2014), "Estimation and Accuracy After Model Selection," (with discussion), *Journal of the American Statistical Association,* 109, 991-1007.

Efron, B. and Tibshirani, R.J. (1998), "The Problem of Regions," *The Annals of Statistics*, 26, 1687-1718.

Frey, J. (2013), "Data-Driven Nonparametric Prediction Intervals," *Journal of Statistical Planning and Inference*, 143, 1039-1048.

Friedman, J.H., and Hall, P. (2007), "On Bagging and Nonlinear Estimation," *Journal of Statistical Planning and Inference,* 137, 669-683.

Hall, P. (1988), "Theoretical Comparisons of Bootstrap Confidence Intervals," (with discussion), *The Annals of Statistics*, 16, 927-985.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd ed., Springer, New York, NY.

Horowitz, J.L. (2001), "The Bootstrap," in *Handbook of Econometrics,* Vol. 5, eds. Heckman, J.J., and Leamer, E., Elsevier Science, Amsterdam, ch. 52.

Janssen, A., and Pauls, T. (2003), "How Do Bootstrap and Permutation Tests Work?," *The Annals of Statistics,* 31, 768-806.

Lei, J., Robins, J., and Wasserman, L. (2013), "Distribution Free Prediction Sets," *Journal of the American Statistical Association*, 108, 278-287.

Machado, J.A.F., and Parente, P. (2005), "Bootstrap Estimation of Covariance Matrices Via the Percentile Method," *Econometrics Journal,* 8, 70–78.

MacKinnon, J.G. (2009), "Bootstrap Hypothesis Testing," in *Handbook of Computational Econometrics*, eds. Belsey, D., and Kontoghioghes, E., Wiley, Hoboken, NJ, ch. 6.

Mammen, E. (1992), "Bootstrap, Wild Bootstrap, and Asymptotic Normality," *Probability Theory and Related Fields,* 93, 439-455.

Mammen, E. (1993), "Bootstrap and Wild Bootstrap for High Dimensional Linear Models," *The Annals of Statistics*, 21, 255-285.

Nishi, R. (1984), "Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression," *The Annals of Statistics,* 12, 758-765.

Olive, D.J. (2013), "Asymptotically Optimal Regression Prediction Intervals and Prediction Regions for Multivariate Data," *International Journal of Statistics and Probability*, 2, 90-100.

Olive, D.J. (2014), *Statistical Theory and Inference*, Springer, New York, NY.

Olive, D.J. (2017a), "Applications of Hyperellipsoidal Prediction Regions," *Statistical Papers*, to appear. See (http://lagrange.math.siu.edu/Olive/pphpr.pdf).

Olive, D.J. (2017b), *Robust Multivariate Analysis,* Springer, to appear.

Olive, D.J. (2017c), *Linear Regression,* Springer, New York, NY.

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

Olive, D.J., and Hawkins, D.M. (2005), "Variable Selection for 1D Regression Models," *Technometrics*, 47, 43-50.

Pelawa Watagoda, L.C.R., and Olive, D.J. (2017), "Inference for Multiple Linear Regression After Model or Variable Selection," preprint at (http://lagrange.math.siu.edu/Olive/ppvsinf.pdf).

Polanski, A. M. (2008), *Observed Confidence Levels: Theory and Applications*. Chapman & Hall/CRC, Boca Raton, FL.

R Core Team (2016), "R: a Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org).

Rupasinghe Arachchige Don, H.S., and Olive, D.J. (2017), "Bootstrapping Analogs of the One Way MANOVA Test," preprint at (http://lagrange.math.siu.edu/Olive/ppmanova.pdf).

Rupasinghe Arachchige Don, H.S., and Pelawa Watagoda, L.C.R. (2017), "Bootstrapping Analogs of the Two Sample Hotelling's T^2 Test," *Communications and Statistics:*

Theory and Methods, to appear. See preprint at (http://lagrange.math.siu.edu/Olive/stwosample.pdf).

Seber, G.A.F., and Lee, A.J. (2003), *Linear Regression Analysis,* 2nd ed., Wiley, New York, NY.