Spring 4-6-2018

BOOTSTRAPPING FORWARD SELECTION WITH BIC

Charles Murphy
murphington11@siu.edu

BOOTSTRAPPING FORWARD SELECTION WITH BIC

by

Charles Murphy

B.S., Southern Illinois University, 2016

A Research Paper
Submitted in Partial Fulfillment of the Requirements for the
Master of Science

Department of Mathematics
in the Graduate School
Southern Illinois University Carbondale
May, 2018

RESEARCH PAPER APPROVAL

BOOTSTRAPPING FORWARD SELECTION WITH BIC

by

Charles Murphy

A Research Paper Submitted in Partial

Fulfillment of the Requirements

for the Degree of

Master of Science

in the field of Mathematics

Approved by:

David J. Olive

Michael C. Sullivan

Seyed Y. Samadi

Graduate School
Southern Illinois University Carbondale
April 3, 2018

AN ABSTRACT OF THE RESEARCH PAPER OF

CHARLES MURPHY, for the Master of Science degree in MATHEMATICS, presented on APRIL 3, 2018, at Southern Illinois University Carbondale.

TITLE: BOOTSTRAPPING FORWARD SELECTION WITH BIC

MAJOR PROFESSOR: Dr. David J. Olive

This paper presents a method for bootstrapping the multiple linear regression model $Y = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e$ using forward selection with the BIC criterion.

KEY WORDS: Bootstrap; Confidence Region; Forward Selection; Prediction Interval.

ACKNOWLEDGMENTS

I would like to take this opportunity to thank my research advisor, Dr. David Olive for overseeing my Master's project and Dr. Michael Sullivan and Dr. Seyed Samadi for sitting on my committee. I would also like to thank all of the other professors who have encouraged and challenged me over the past 5.5 years here to be the best student I could be. Finally, I would like to thank my family for their neverending love and support. I certainly couldn't have achieved all that I have without them!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 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 $\boldsymbol{x}_i^T = (x_{i,1}, ..., x_{i,p})$ and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ where β_1 corresponds to the intercept. Then the multiple linear regression (MLR) model is

$$Y_i = \beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{1.1}$$

for $i = 1, ..., n$. 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

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{1.2}$$

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

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

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S \tag{1.3}$$

where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$, $\boldsymbol{x}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p - a_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$ and E denotes the subset of terms that can be eliminated given that the subset S is in the model. Let $\boldsymbol{x}_I$ be the vector of a terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). Suppose that S is a subset of I and that model (1.3) holds. Then

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_{I/S}^T\boldsymbol{\beta}_{(I/S)} + \boldsymbol{x}_O^T\boldsymbol{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I \tag{1.4}$$

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 $S \subseteq I$.

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 $Q_2(I) = SSE(I) = RSS(I) = \boldsymbol{r}^T(I)\boldsymbol{r}(I) = \sum_{i=1}^n r_i^2(I) = \sum_{i=1}^n (Y_i - \hat{Y}_i(I))^2$. 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 $Q_j(I) = \boldsymbol{r}^T(I)\boldsymbol{r}(I) = \sum_{i=1}^n r_i^2(I) = \sum_{i=1}^n (Y_i - \hat{Y}_i(I))^2$. 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 $j = 2, ..., p$ where $n \geq 10p$ and p is fixed.

When there is a sequence of p submodels, the final submodel I_d needs to be selected. Let the candidate model I contains a terms, including a constant. For example, let $\boldsymbol{x}_I$ and $\hat{\boldsymbol{\beta}}_I$ be $a \times 1$ vectors. Then there are many criteria used to select the final submodel I_d. For a given data set, p, n, and $\hat{\sigma}^2$ act as constants, and a criterion below may add a constant or be divided by a positive constant without changing the subset I_{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 σ^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 σ^2 under mild conditions by Su and Cook (2012).

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 Nishi (1984), the probability that model I_{min} from C_p or AIC underfits goes to zero as $n \to \infty$. If $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$, form the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I,0}$ from $\hat{\boldsymbol{\beta}}_I$ 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 $\hat{\boldsymbol{\beta}}_{I,0}$ 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 $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ under model (1.3). See Pelawa Watagoda and Olive (2018) and Olive (2017a: p. 123, 2017b: p. 176).

Chapter 2 considers mixture distributions. Chapter 3 shows that a bootstrap confidence region can be formed by applying a prediction region to the bootstrap sample, and Chapter 4 gives a simulation.

CHAPTER 2

MIXTURE DISTRIBUTIONS

Mixture distributions are useful for variable selection since asymptotically $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is a mixture distribution of $\hat{\boldsymbol{\beta}}_{I_j,0}$ where $S \subseteq I_j$. See Equation (1.3). A random vector $\boldsymbol{u}$ has a mixture distribution if $\boldsymbol{u}$ equals a random vector $\boldsymbol{u}_j$ with probability π_j for $j = 1, ..., J$.

Definition 1. The distribution of a $g \times 1$ random vector $\boldsymbol{u}$ is a mixture distribution if the cumulative distribution function (cdf) of $\boldsymbol{u}$ is

$$F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t}) \tag{2.1}$$

where the probabilities π_j satisfy $0 \leq \pi_j \leq 1$ and $\sum_{j=1}^{J} \pi_j = 1$, $J \geq 2$, and $F_{\boldsymbol{u}_j}(\boldsymbol{t})$ is the cdf of a $g \times 1$ random vector $\boldsymbol{u}_j$. Then $\boldsymbol{u}$ has a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j.

Theorem 1. Suppose $E(h(\boldsymbol{u}))$ and the $E(h(\boldsymbol{u}_j))$ exist. Then

$$E(h(\boldsymbol{u})) = \sum_{j=1}^{J} \pi_j E[h(\boldsymbol{u}_j)]. \tag{2.2}$$

Hence

$$E(\boldsymbol{u}) = \sum_{j=1}^{J} \pi_j E[\boldsymbol{u}_j], \tag{2.3}$$

and $Cov(\boldsymbol{u}) = E(\boldsymbol{u}\boldsymbol{u}^T) - E(\boldsymbol{u})E(\boldsymbol{u}^T) = E(\boldsymbol{u}\boldsymbol{u}^T) - E(\boldsymbol{u})[E(\boldsymbol{u})]^T =$

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

$$\sum_{j=1}^{J} \pi_j Cov(\boldsymbol{u}_j) + \sum_{j=1}^{J} \pi_j E(\boldsymbol{u}_j)[E(\boldsymbol{u}_j)]^T - E(\boldsymbol{u})[E(\boldsymbol{u})]^T. \tag{2.4}$$

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

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

This theorem is easy to prove if the $\boldsymbol{u}_j$ are continuous random vectors with (joint) probability density functions (pdfs) $f_{\boldsymbol{u}_j}(\boldsymbol{t})$. Then $\boldsymbol{u}$ is a continuous random vector with pdf

$$f_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j f_{\boldsymbol{u}_j}(\boldsymbol{t}), \quad \text{and}$$

$$E(h(\boldsymbol{u})) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(\boldsymbol{t}) f_{\boldsymbol{u}}(\boldsymbol{t}) d\boldsymbol{t} = \sum_{j=1}^{J} \pi_j \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(\boldsymbol{t}) f_{\boldsymbol{u}_j}(\boldsymbol{t}) d\boldsymbol{t} = \sum_{j=1}^{J} \pi_j E[h(\boldsymbol{u}_j)]$$

where $E[h(\boldsymbol{u}_j)]$ is the expectation with respect to the random vector $\boldsymbol{u}_j$. Note that

$$E(\boldsymbol{u})[E(\boldsymbol{u})]^T = \sum_{j=1}^{J} \sum_{k=1}^{J} \pi_j \pi_k E(\boldsymbol{u}_j)[E(\boldsymbol{u}_k)]^T. \tag{2.5}$$

Definition 2. The *population mean* of a random $p \times 1$ vector $\boldsymbol{X} = (X_1, ..., X_p)^T$ is

$$E(\boldsymbol{X}) = (E(X_1), ..., E(X_p))^T$$

and the $p \times p$ *population covariance matrix*

$$Cov(\boldsymbol{X}) = E(\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{X} - E(\boldsymbol{X}))^T = (\sigma_{ij}).$$

That is, the ij entry of $Cov(\boldsymbol{X})$ is $Cov(X_i, X_j) = \sigma_{ij}$.

Note that $Cov(\boldsymbol{X})$ is a symmetric positive semidefinite matrix. The following results are useful. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $p \times 1$ random vectors, $\boldsymbol{a}$ a conformable constant vector, and $\boldsymbol{A}$ and $\boldsymbol{B}$ are conformable constant matrices, then

$$E(\boldsymbol{a} + \boldsymbol{X}) = \boldsymbol{a} + E(\boldsymbol{X}) \ \text{ and } \ \mathrm{E}(\boldsymbol{X} + \boldsymbol{Y}) = \mathrm{E}(\boldsymbol{X}) + \mathrm{E}(\boldsymbol{Y}) \tag{2.6}$$

and

$$E(\boldsymbol{AX}) = \boldsymbol{A}E(\boldsymbol{X}) \ \text{ and } \ \mathrm{E}(\boldsymbol{AXB}) = \boldsymbol{A}\mathrm{E}(\boldsymbol{X})\boldsymbol{B}. \tag{2.7}$$

Thus

$$Cov(\boldsymbol{a} + \boldsymbol{AX}) = Cov(\boldsymbol{AX}) = \boldsymbol{A}Cov(\boldsymbol{X})\boldsymbol{A}^T. \tag{2.8}$$

For the multivariate normal (MVN) distribution $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and

$$Cov(\boldsymbol{X}) = \boldsymbol{\Sigma}.$$

CHAPTER 3

BOOTSTRAPPING CONFIDENCE REGIONS

Inference will consider bootstrap confidence intervals and bootstrap confidence regions for bootstrap hypothesis testing. Applying the shorth prediction interval and the Olive (2013) prediction region to the bootstrap sample will give the bootstrap confidence intervals and regions.

Consider predicting a future test random variable Z_f given iid training data $Z_1, ..., Z_n$. A large sample $100(1-\delta)\%$ prediction interval (PI) for Z_f has the form $[\hat{L}_n, \hat{U}_n]$ where $P(\hat{L}_n \leq Z_f \leq \hat{U}_n) \to 1-\delta$ as the sample size $n \to \infty$. The shorth(c) estimator is useful for making prediction intervals. Let $Z_{(1)}, ..., Z_{(n)}$ be the order statistics of $Z_1, ..., Z_n$. Then let the shortest closed interval containing at least c of the Z_i be

$$\mathrm{shorth(c)} = [\mathrm{Z_{(s)}}, \mathrm{Z_{(s+c-1)}}]. \tag{3.1}$$

Let $\lceil x \rceil$ be the smallest integer $\geq x$, e.g., $\lceil 7.7 \rceil = 8$. Let

$$k_n = \lceil n(1-\delta) \rceil. \tag{3.2}$$

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

$$c = \min(n, \lceil n[1-\delta+1.12\sqrt{\delta/n}\ ]\ \rceil). \tag{3.3}$$

Example 1. Given below were votes for preseason 1A basketball poll from Nov. 22, 2011 WSIL News where the 778 was a typo: the actual value was 78. As shown below, finding shorth(3) from the ordered data is simple. If the outlier was corrected, shorth(3) = [76,78].

```
111    89    778    78    76
```

```
order data: 76 78 89 111 778

                    13 = 89 - 76

                        33 = 111 - 78

                            689 = 778 - 89
shorth(3) = [76,89]
```

We also want to use bootstrap tests. Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_0$ is a known $g \times 1$ vector. Given training data $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$, a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is a set $\mathcal{A}_n$ such that $P(\boldsymbol{\theta} \in \mathcal{A}_n) \to 1-\delta$ as $n \to \infty$. Then reject H_0 if $\boldsymbol{\theta}_0$ is not in the confidence region $\mathcal{A}_n$. For model (1.1), let $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta}$ where $\boldsymbol{A}$ is a known full rank $g \times p$ matrix with $1 \leq g \leq p$.

To bootstrap a confidence region, Mahalanobis distances and prediction regions will be useful. Consider predicting a future test value $\boldsymbol{z}_f$, given past training data $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$ where the $\boldsymbol{z}_i$ are $g \times 1$ random vectors. A *large sample* $100(1-\delta)\%$ *prediction region* is a set $\mathcal{A}_n$ such that $P(\boldsymbol{z}_f \in \mathcal{A}_n) \to 1-\delta$ as $n \to \infty$. Let the $g \times 1$ column vector T be a multivariate location estimator, and let the $g \times g$ symmetric positive definite matrix $\boldsymbol{C}$ be a dispersion estimator. Then the ith *squared sample Mahalanobis distance* is the scalar

$$D_i^2 = D_i^2(T, \boldsymbol{C}) = D^2_{\boldsymbol{z}_i}(T, \boldsymbol{C}) = (\boldsymbol{z}_i - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z}_i - T) \tag{3.4}$$

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

$$\overline{\boldsymbol{z}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{z}_i \quad \text{and} \quad \boldsymbol{S} = \frac{1}{\mathrm{n}-1} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\boldsymbol{z}_{\mathrm{i}} - \overline{\boldsymbol{z}})(\boldsymbol{z}_{\mathrm{i}} - \overline{\boldsymbol{z}})^{\mathrm{T}}. \tag{3.5}$$

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

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta g/n), \quad \text{otherwise.} \tag{3.6}$$

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

$$c = \lceil n q_n \rceil. \tag{3.7}$$

Let $(T, \boldsymbol{C}) = (\overline{\boldsymbol{z}}, \boldsymbol{S})$, and let $D_{(U_n)}$ be the $100 q_n$th sample quantile of the D_i. Then the Olive (2013) large sample $100(1-\delta)\%$ nonparametric prediction region for a future value $\boldsymbol{z}_f$ given iid data $\boldsymbol{z}_1, ..., , \boldsymbol{z}_n$ is

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{z}}, \boldsymbol{S}) \le D^2_{(U_n)}\}, \tag{3.8}$$

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

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{z}}, \boldsymbol{S}) \le \chi^2_{g,1-\delta}\}. \tag{3.9}$$

Definition 3. Suppose that data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ has been collected and observed. Often the data is a random sample (iid) from a distribution with cdf F. The *empirical distribution* is a discrete distribution where the $\boldsymbol{x}_i$ are the possible values, and each value is equally likely. If $\boldsymbol{w}$ is a random variable having the empirical distribution, then $p_i = P(\boldsymbol{w} = \boldsymbol{x}_i) = 1/n$ for $i = 1, ..., n$. The *cdf of the empirical distribution* is denoted by F_n.

Example 2. Let $\boldsymbol{w}$ be a random variable having the empirical distribution given by Definition 3. Show that $E(\boldsymbol{w}) = \overline{\boldsymbol{x}} \equiv \overline{\boldsymbol{x}}_n$ and $Cov(\boldsymbol{w}) = \dfrac{n-1}{n}\boldsymbol{S} \equiv \dfrac{n-1}{n}\boldsymbol{S}_n$.

Solution: Recall that for a discrete random vector, the population expected value $E(\boldsymbol{w}) = \sum \boldsymbol{x}_i p_i$ where $\boldsymbol{x}_i$ are the values that $\boldsymbol{w}$ takes with positive probability p_i. Similarly, the population covariance matrix

$$Cov(\boldsymbol{w}) = E[(\boldsymbol{w} - E(\boldsymbol{w}))(\boldsymbol{w} - E(\boldsymbol{w}))^T] = \sum (\boldsymbol{x}_i - E(\boldsymbol{w}))(\boldsymbol{x}_i - E(\boldsymbol{w}))^T p_i.$$

Hence

$$E(\boldsymbol{w}) = \sum_{i=1}^n \boldsymbol{x}_i \frac{1}{n} = \overline{\boldsymbol{x}},$$

and

$$Cov(\boldsymbol{w}) = \sum_{i=1}^n (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \frac{1}{n} = \frac{n-1}{n}\boldsymbol{S}. \ \square$$

Example 3. If $W_1, ..., W_n$ are iid from a distribution with cdf F_W, then the empirical cdf F_n corresponding to F_W is given by

$$F_n(y) = \frac{1}{n}\sum_{i=1}^{n} I(W_i \leq y)$$

where the indicator $I(W_i \leq y) = 1$ if $W_i \leq y$ and $I(W_i \leq y) = 0$ if $W_i > y$. Fix n and y. Then $nF_n(y) \sim$ binomial $(n, F_W(y))$. Thus $E[F_n(y)] = F_W(y)$ and $V[F_n(y)] = F_W(y)[1 - F_W(y)]/n$. By the central limit theorem,

$$\sqrt{n}(F_n(y) - F_W(y)) \xrightarrow{D} N(0, F_W(y)[1 - F_W(y)]).$$

Thus $F_n(y) - F_W(y) = O_P(n^{-1/2})$, and F_n is a reasonable estimator of F_W if the sample size n is large.

Suppose there is data $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$ collected into an $n \times p$ matrix $\boldsymbol{W}$. Let the statistic $T_n = t(\boldsymbol{W}) = T(F_n)$ be computed from the data. Suppose the statistic estimates $\boldsymbol{\theta} = T(F)$, and let $t(\boldsymbol{W}^*) = t(F_n^*) = T_n^*$ indicate that t was computed from an iid sample from the empirical distribution F_n: a sample $\boldsymbol{w}_1^*, ..., \boldsymbol{w}_n^*$ of size n was drawn with replacement from the observed sample $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. This notation is used for von Mises differentiable statistical functions in large sample theory. See Serfling (1980, ch. 6). The *empirical bootstrap* or *nonparametric bootstrap* or *naive bootstrap* draws B samples of size n from the rows of $\boldsymbol{W}$, e.g. from the empirical distribution of $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. Then T_{jn}^* is computed from the jth bootstrap sample for $j = 1, ..., B$.

Example 4. Suppose the data is 1, 2, 3, 4, 5, 6, 7. Then $n = 7$ and the sample median T_n is 4. Using *R*, we drew $B = 2$ bootstrap samples (samples of size n drawn with replacement from the original data) and computed the sample median $T_{1,n}^* = 3$ and $T_{2,n}^* = 4$.

```
b1 <- sample(1:7,replace=T)

b1

[1] 3 2 3 2 5 2 6
```

```
median(b1)
[1] 3
b2 <- sample(1:7,replace=T)
b2
[1] 3 5 3 4 3 5 7
median(b2)
[1] 4
```

The bootstrap has been widely used to estimate the population covariance matrix of the statistic $Cov(T_n)$, for testing hypotheses, and for obtaining confidence regions (often confidence intervals). An iid sample $T_{1n}, ..., T_{Bn}$ of size B of the statistic would be very useful for inference, but typically we only have one sample of data and one value $T_n = T_{1n}$ of the statistic. Often $T_n = t(\boldsymbol{w}_1, ..., \boldsymbol{w}_n)$, and the bootstrap sample $T^*_{1n}, ..., T^*_{Bn}$ is formed where $T^*_{jn} = t(\boldsymbol{w}^*_{j1}, ..., \boldsymbol{w}^*_{jn})$. The *residual bootstrap* is often useful for additive error regression models of the form $Y_i = m(\boldsymbol{x}_i) + e_i = \hat{m}(\boldsymbol{x}_i) + r_i = \hat{Y}_i + r_i$ for $i = 1, ..., n$ where the ith residual $r_i = Y_i - \hat{Y}_i$. Let $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$, $\boldsymbol{r} = (r_1, ..., r_n)^T$, and let $\boldsymbol{X}$ be an $n \times p$ matrix with ith row $\boldsymbol{x}_i^T$. Then the fitted values $\hat{Y}_i = \hat{m}(\boldsymbol{x}_i)$, and the residuals are obtained by regressing $\boldsymbol{Y}$ on $\boldsymbol{X}$. Here the errors e_i are iid, and it would be useful to be able to generate B iid samples $e_{1j}, ..., e_{nj}$ from the distribution of e_i where $j = 1, ..., B$. If the $m(\boldsymbol{x}_i)$ were known, then we could form a vector $\boldsymbol{Y}_j$ where the ith element $Y_{ij} = m(\boldsymbol{x}_i) + e_{ij}$ for $i = 1, ..., n$. Then regress $\boldsymbol{Y}_j$ on $\boldsymbol{X}$. Instead, draw samples $r^*_{1j}, ..., r^*_{nj}$ with replacement from the residuals, then form a vector $\boldsymbol{Y}^*_j$ where the ith element $Y^*_{ij} = \hat{m}(\boldsymbol{x}_i) + r^*_{ij}$ for $i = 1, ..., n$. Then regress $\boldsymbol{Y}^*_j$ on $\boldsymbol{X}$.

The Olive (2017ab, 2018ab) prediction region method obtains a confidence region for $\boldsymbol{\theta}$ by applying the nonparametric prediction region (3.8) to the bootstrap sample $T^*_1, ..., T^*_B$, and the theory for the method is sketched below. Let $\overline{T}^*$ and $\boldsymbol{S}^*_T$ be the sample mean and sample covariance matrix of the bootstrap sample. Assume $n\boldsymbol{S}^*_T \xrightarrow{P} \boldsymbol{\Sigma}_A$. See Machado and Parente (2005) for regularity conditions for this assumption. Following Bickel and Ren

(2001), let the vector of parameters $\boldsymbol{\theta} = T(F)$, the statistic $T_n = T(F_n)$, 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 (has a Hadamard derivative $\dot{T}(F)$), then $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ and $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$ with $\boldsymbol{u} = \dot{T}(F)\boldsymbol{z}_F$. Olive (2017b) used these results to show that if $\boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$, then $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{D} \boldsymbol{0}$, $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{u}$, $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, and that the prediction region method large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is

$$\{\boldsymbol{w} : (\boldsymbol{w} - \overline{T}^*)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - \overline{T}^*) \leq D^2_{(U_B)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{S}_T^*) \leq D^2_{(U_B)}\} \tag{3.10}$$

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 $i = 1, ..., B$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(\overline{T}^* - \boldsymbol{\theta}_0)^T [\boldsymbol{S}_T^*]^{-1} (\overline{T}^* - \boldsymbol{\theta}_0) > D^2_{(U_B)}$. The prediction region method for testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ is simple. Let $\hat{\boldsymbol{\theta}}$ be a consistent estimator of $\boldsymbol{\theta}$ and make a bootstrap sample $\boldsymbol{w}_i = \hat{\boldsymbol{\theta}}_i^* - \boldsymbol{\theta}_0$ for $i = 1, ..., B$. Make the nonparametric prediction region (3.10) for the $\boldsymbol{w}_i$ and fail to reject H_0 if $\boldsymbol{0}$ is in the prediction region (if $D_{\boldsymbol{0}} \leq D_{(U_B)}$), reject H_0 otherwise.

The modified Bickel and Ren (2001) large sample $100(1-\delta)\%$ confidence region is

$$\{\boldsymbol{w} : (\boldsymbol{w} - T)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B,T)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{(U_B,T)}\} \tag{3.11}$$

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

The Pelawa Watagoda and Olive (2018) hybrid large sample $100(1-\delta)\%$ confidence region shifts the hyperellipsoid (3.10) to be centered at T instead of $\overline{T}^*$:

$$\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{(U_B)}\}. \tag{3.12}$$

Hyperellipsoids (3.10) and (3.12) have the same volume since they are the same region shifted to have a different center. The ratio of the volumes of regions (3.10) and (3.11) is

$$\left(\frac{D_{(U_B)}}{D_{(U_B,T)}} \right)^g. \tag{3.13}$$

Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_0 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}$ is $g \times 1$. For example, let $\boldsymbol{A}$ be a $g \times p$ matrix with full rank g, $\boldsymbol{\theta} = \boldsymbol{A\beta}$, $\boldsymbol{\theta}_0 = \boldsymbol{0}$, and $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$. This section gives some theory for the bagging estimator $\overline{T}^*$, also called the smoothed bootstrap estimator. The theory may be useful for hypothesis testing after model selection if n/p is large. Empirically, bootstrapping with the bagging estimator often outperforms bootstrapping with T_n. See Efron (2014). See Büchlmann and Yu (2002) and Friedman and Hall (2007) for theory and references for the bagging estimator.

If i) $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, then under regularity conditions, ii) $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$, iii) $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, iv) $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{u}$, and v) $n\boldsymbol{S}_T^* \xrightarrow{P} Cov(\boldsymbol{u})$.

Suppose i) and ii) hold with $E(\boldsymbol{u}) = \boldsymbol{0}$ and $Cov(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}}$. With respect to the bootstrap sample, T_n is a constant and the $\sqrt{n}(T_i^* - T_n)$ are iid for $i = 1, ..., B$. Let $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{v}_i \sim \boldsymbol{u}$ where the $\boldsymbol{v}_i$ are iid with the same distribution as $\boldsymbol{u}$. Fix B. Then the average of the $\sqrt{n}(T_i^* - T_n)$ is

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

where $\boldsymbol{z} \sim AN_g(\boldsymbol{0}, \boldsymbol{\Sigma})$ is an asymptotic multivariate normal approximation. Hence as $B \to \infty$, $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \boldsymbol{0}$, and iii) and iv) hold. If B is fixed and $\boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{u}})$, then

$$\frac{1}{B}\sum_{i=1}^{B} \boldsymbol{v}_i \sim N_g\left(\boldsymbol{0}, \frac{\boldsymbol{\Sigma}_{\boldsymbol{u}}}{B}\right) \text{ and } \sqrt{\mathrm{B}}\sqrt{\mathrm{n}}(\overline{\mathrm{T}}^* - \mathrm{T_n}) \xrightarrow{\mathrm{D}} \mathrm{N_g}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{u}}).$$

Hence the prediction region method gives a large sample confidence region for $\boldsymbol{\theta}$ provided that the sample percentile $\hat{D}^2_{1-\delta}$ of the $D^2_{T_i^*}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(T_i^* - \overline{T}^*)^T(n\boldsymbol{S}_T^*)^{-1}\sqrt{n}(T_i^* - \overline{T}^*)$ is a consistent estimator of the percentile $D^2_{n,1-\delta}$ of the random variable $D^2_{\boldsymbol{\theta}}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(\boldsymbol{\theta} - \overline{T}^*)^T(n\boldsymbol{S}_T^*)^{-1}\sqrt{n}(\boldsymbol{\theta} - \overline{T}^*)$ in that $\hat{D}^2_{1-\delta} - D^2_{n,1-\delta} \xrightarrow{P} 0$. Since iii) and iv) hold, the sample percentile will be consistent under much weaker conditions than v) if $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ is nonsingular. For example, if $(n\boldsymbol{S}_T^*)^{-1} = \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1} + \boldsymbol{C} + o_p(1)$ for some $g \times g$ constant matrix $\boldsymbol{C}$. Olive (2017b $\oint$ 5.3.3) proved that the prediction region method gives a large sample confidence region under the much stronger conditions of v) and $\boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{u}})$, but the above proof is simpler.

Now suppose that T_n is equal to the estimator T_{jn} with probability π_{jn} for $j = 1, ..., J$ where $\sum_j \pi_{jn} = 1$, $\pi_{jn} \to \pi_j$ as $n \to \infty$, and $\sqrt{n}(T_{jn} - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}_j$ with $E(\boldsymbol{u}_j) = \boldsymbol{0}$ and $Cov(\boldsymbol{u}_j) = \boldsymbol{\Sigma}_j$. Then the cumulative distribution function (cdf) of T_n is $F_{T_n}(\boldsymbol{z}) = \sum_j \pi_{jn} F_{T_{jn}}(\boldsymbol{z})$ where $F_{T_{jn}}(\boldsymbol{z})$ is the cdf of T_{jn}. Hence

$$\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u} \tag{3.14}$$

where the cdf of $\boldsymbol{u}$ is $F_{\boldsymbol{u}}(\boldsymbol{z}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{z})$ and $F_{\boldsymbol{u}_j}(\boldsymbol{z})$ is the cdf of $\boldsymbol{u}_j$. Thus $\boldsymbol{u}$ is a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j, $E(\boldsymbol{u}) = \boldsymbol{0}$, and $Cov(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \sum_j \pi_j \boldsymbol{\Sigma}_j$.

For the bootstrap, suppose that T_i^* is equal to T_{ij}^* with probability ρ_{jn} for $j = 1, ..., J$ where $\sum_j \rho_{jn} = 1$, and $\rho_{jn} \to \pi_j$ as $n \to \infty$. Let B_{jn} count the number of times $T_i^* = T_{ij}^*$ in the bootstrap sample. Then the bootstrap sample $T_1^*, ..., T_B^*$ can be written as

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

where the B_{jn} follow a multinomial distribution and $B_{jn}/B \xrightarrow{P} \rho_{jn}$ as $B \to \infty$. Conditionally on the B_{jn} and with respect to the bootstrap sample, the T_{ij}^* are independent. Denote $T_{1j}^*, ..., T_{B_{jn},j}^*$ as the jth bootstrap component of the bootstrap sample with sample mean $\overline{T}_j^*$ and sample covariance matrix $\boldsymbol{S}_{T,j}^*$. Then

$$\overline{T}^* = \frac{1}{B}\sum_{i=1}^{B} T_i^* = \sum_j \frac{B_{jn}}{B}\frac{1}{B_{jn}}\sum_{i=1}^{B_{jn}} T_{ij}^* = \sum_j \hat{\rho}_{jn} \overline{T}_j^*.$$

Suppose $\sqrt{n}(T_i^* - E(T^*)) \xrightarrow{D} \boldsymbol{v}_i \sim \boldsymbol{v}$ where $E(\boldsymbol{v}) = \boldsymbol{0}$, $Cov(\boldsymbol{v}) = \boldsymbol{\Sigma}_{\boldsymbol{v}}$, and $E(T^*) = \sum_j \rho_{jn} E(T_{ij}^*)$ where often $E(T_{ij}^*) = T_{jn}$. With respect to the data distribution, suppose $\sqrt{n}(E(T^*) - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{w}$. Then by an argument similar to the one given for when T_n is not from a mixture distribution, $\sqrt{n}(\overline{T}^* - E(T^*)) \xrightarrow{P} \boldsymbol{0}$, $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{v}$, and $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{w}$.

Assume $T_1, ..., T_B$ are iid with nonsingular covariance matrix $\boldsymbol{\Sigma}_{T_n}$. Then the large sample $100(1-\delta)\%$ prediction region $R_p = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}, \boldsymbol{S}_T) \le \hat{D}^2_{(U_B)}\}$ centered at $\overline{T}$

contains a future value of the statistic T_f with probability $1-\delta_B \to 1-\delta$ as $B \to \infty$. Hence the region $R_c = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T) \le \hat{D}^2_{(U_B)}\}$ centered at a randomly selected T_n contains $\overline{T}$ with probability $1-\delta_B$. If i) holds with $E(\boldsymbol{u}) = \boldsymbol{0}$ and $Cov(\boldsymbol{u}) = \boldsymbol{\Sigma_u}$, then for fixed B,

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

Hence $(\overline{T} - \boldsymbol{\theta}) = O_P((nB)^{-1/2})$, and $\overline{T}$ gets arbitrarily close to $\boldsymbol{\theta}$ compared to T_n as $B \to \infty$. Hence R_c is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ as $n, B \to \infty$. We also need $(n\boldsymbol{S}_T)^{-1}$ to be "not too ill conditioned."

With a mixture distribution, the bootstrap sample shifts the data cloud to be centered at $\overline{T}^*$ where $\sqrt{n}(\overline{T}^* - \sum_j \rho_{jn} T_{jn}) \xrightarrow{P} \boldsymbol{0}$. The T_{jn} are computed from the same data set and hence correlated. Suppose $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{w}$, and $(n\boldsymbol{S}^*_T)^{-1}$ is "not too ill conditioned." Then

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

$$D_2^2 = D^2_{\boldsymbol{\theta}}(T_n, \boldsymbol{S}^*_T) = \sqrt{n}(T_n - \boldsymbol{\theta})^T (n\boldsymbol{S}^*_T)^{-1} \sqrt{n}(T_n - \boldsymbol{\theta}), \quad \text{and}$$

$$D_3^2 = D^2_{\boldsymbol{\theta}}(\overline{T}^*, \boldsymbol{S}^*_T) = \sqrt{n}(\overline{T}^* - \boldsymbol{\theta})^T (n\boldsymbol{S}^*_T)^{-1} \sqrt{n}(\overline{T}^* - \boldsymbol{\theta})$$

are well behaved in that there exist cutoffs $\hat{D}^2_{i,1-\delta}$ that would result in good confidence regions for $i = 2$ and 3. Heuristically, for a mixture distribution, the deviation $\overline{T}^* - \boldsymbol{\theta}$ tends to be smaller on average than the deviations $T_n - \boldsymbol{\theta} \approx T_i^* - \overline{T}^*$, while the deviation $T_i^* - T_n$ tends to be larger than the other three deviations, on average. Hence $\hat{D}^2_{2,1-\delta} = D^2_{(U_B)}$ gives coverage close to the nominal coverage for prediction region (3.12), but cutoffs $\hat{D}^2_{3,1-\delta} = D^2_{(U_B)}$ and $\hat{D}^2_{2,1-\delta} = D^2_{(U_B,T)}$ are slightly too large, and prediction regions (3.10) and (3.11) tend to have coverage slightly higher than the nominal coverage $1-\delta$ if n and B are large. In simulations for $n \ge 20p$, the coverage tends to get close to $1-\delta$ for $B \ge \max(400, 50p)$ so that $\boldsymbol{S}^*_T$ is a good estimator of $Cov(T^*)$.

To examine the bagging estimator, assume that each bootstrap component satisfies vi) $\sqrt{n}(T_{jn} - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}_j \sim N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_j)$, vii) $\sqrt{n}(T^*_{ij} - T_{jn}) \xrightarrow{D} \boldsymbol{u}_j$, viii) $\sqrt{n}(\overline{T}^*_j - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}_j$, ix) $\sqrt{n}(T^*_{ij} - \overline{T}^*_j) \xrightarrow{D} \boldsymbol{u}_j$, x) $n\boldsymbol{S}^*_{T,j} \xrightarrow{P} \boldsymbol{\Sigma}_j$, and xi) $\sqrt{n}(T_{jn} - \overline{T}^*_j) \xrightarrow{P} \boldsymbol{0}$ as $B_{jn} \to \infty$ and $n \to \infty$.

Consider the random vectors

$$Z_n = \sum_j \frac{B_{jn}}{B} T_{jn} \text{ and } \mathrm{W_n} = \sum_{\mathrm{j}} \rho_{\mathrm{jn}} \mathrm{T_{jn}}.$$

By xi)

$$\sqrt{n}(Z_n - \overline{T}^*) = \sqrt{n}(\sum_j \frac{B_{jn}}{B} T_{jn} - \overline{T}^*) = \sum_j \frac{B_{jn}}{B} \sqrt{n}(T_{jn} - \overline{T}^*_j) \xrightarrow{P} \mathbf{0}.$$

Also, $\sqrt{n}(Z_n - \boldsymbol{\theta}) - \sqrt{n}(W_n - \boldsymbol{\theta}) =$

$$\sum_j \left(\frac{B_{jn}}{B} - \rho_{jn} \right) \sqrt{n}(T_{jn} - \boldsymbol{\theta}) = \sum_j O_P(1) O_P(n^{-1/2}) \xrightarrow{P} \mathbf{0}.$$

Assume the $\boldsymbol{u}_{nj} = \sqrt{n}(T_{jn} - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}_j$ are such that

$$\sqrt{n}(W_n - \boldsymbol{\theta}) = \sum_j \rho_{jn} \sqrt{n}(T_{jn} - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{w} = \sum_j \pi_j \boldsymbol{u}_j.$$

Note that $E(\boldsymbol{w}) = \mathbf{0}$ and $Cov(\boldsymbol{w}) = \boldsymbol{\Sigma}_{\boldsymbol{w}} = \sum_j \sum_k \pi_j \pi_k Cov(\boldsymbol{u}_j, \boldsymbol{u}_k)$. Hence

$$\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{w}. \tag{3.15}$$

Since $\boldsymbol{w}$ is a weighted mean of the $\boldsymbol{u}_j \sim N_g(\mathbf{0}, \boldsymbol{\Sigma}_j)$, a normal approximation is $\boldsymbol{w} \approx N_g(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}})$. The approximation is exact if the $\boldsymbol{u}_j$ with positive π_j have a joint multivariate normal distribution.

Now consider variable selection for model (1.1) with $\boldsymbol{\theta} = \boldsymbol{A\beta}$ where $\boldsymbol{A}$ is a known full rank $g \times p$ matrix with $1 \leq g \leq p$. Olive (2017a: p. 128, 2018a) showed that the prediction region method can simulate well for the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I_{min},0}$. Assume p is fixed, $n \geq 20p$, and that the error distribution is unimodal and not highly skewed. The response plot and residual plot are plots with $\hat{Y} = \boldsymbol{x}^T \hat{\boldsymbol{\beta}}$ 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. If the error distribution is skewed or multimodal, then much larger sample sizes may be needed.

For the nonparametric bootstrap, cases are sampled with replacement, and the above conditions hold since each component bootstraps correctly. For the residual bootstrap, we use the fitted values and residuals from the OLS full model, but fit $\hat{\boldsymbol{\beta}}$ for a method such as forward selection, lasso, et cetera. Consider forward selection where each component uses a $\hat{\boldsymbol{\beta}}_{I_j}$. Let $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{Y}}_{OLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{H}\boldsymbol{Y}$ be the fitted values from the OLS full model where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Let $\boldsymbol{r}^W$ denote an $n \times 1$ random vector of elements selected with replacement from the OLS full model residuals. Following Freedman (1981) and Efron (1982, p. 36), $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} + \boldsymbol{r}^W$ follows a standard linear model where the elements r_i^W of $\boldsymbol{r}^W$ are iid from the empirical distribution of the OLS full model residuals r_i. Hence

$$E(r_i^W) = \frac{1}{n}\sum_{i=1}^n r_i = 0, \quad V(r_i^W) = \sigma_n^2 = \frac{1}{n}\sum_{i=1}^n r_i^2 = \frac{n-p}{n} MSE,$$

$$E(\boldsymbol{r}^W) = \boldsymbol{0}, \text{ and } \operatorname{Cov}(\boldsymbol{Y}^*) = \operatorname{Cov}(\boldsymbol{r}^{\mathrm{W}}) = \sigma_{\mathrm{n}}^2 \boldsymbol{I}_{\mathrm{n}}.$$

Then $\hat{\boldsymbol{\beta}}^*_{I_j} = (\boldsymbol{X}_{I_j}^T\boldsymbol{X}_{I_j})^{-1}\boldsymbol{X}_{I_j}^T\boldsymbol{Y}^* = \boldsymbol{D}_j\boldsymbol{Y}^*$ with $Cov(\hat{\boldsymbol{\beta}}^*_{I_j}) = \sigma_n^2(\boldsymbol{X}_{I_j}^T\boldsymbol{X}_{I_j})^{-1}$ and $E(\hat{\boldsymbol{\beta}}^*_{I_j}) = (\boldsymbol{X}_{I_j}^T\boldsymbol{X}_{I_j})^{-1}\boldsymbol{X}_{I_j}^T E(\boldsymbol{Y}^*) = (\boldsymbol{X}_{I_j}^T\boldsymbol{X}_{I_j})^{-1}\boldsymbol{X}_{I_j}^T\boldsymbol{H}\boldsymbol{Y} = \hat{\boldsymbol{\beta}}_{I_j}$ since $\boldsymbol{H}\boldsymbol{X}_{I_j} = \boldsymbol{X}_{I_j}$. The expectations are with respect to the bootstrap distribution where $\hat{\boldsymbol{Y}}$ acts as a constant.

For the above residual bootstrap with forward selection and C_p, let $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$ and $T_{jn} = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0} = \boldsymbol{A}\boldsymbol{D}_{j,0}\boldsymbol{Y}$ where $\boldsymbol{D}_{j,0}$ adds rows of zeroes to $\boldsymbol{D}_j$ corresponding to the x_i not in I_j. If $S \subseteq I_j$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \sigma^2\boldsymbol{V}_j)$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}_j \sim N_p(\boldsymbol{0}, \sigma^2\boldsymbol{V}_{j,0})$ where $\boldsymbol{V}_{j,0}$ adds columns and rows of zeroes corresponding to the x_i not in I_j. Then under regularity conditions, (3.14) and (3.15) hold where $\sqrt{n}(\sum_j \rho_{jn}T_{jn} - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{w}$, and the sum is over $j : S \subseteq I_j$. Thus $E(T^*) = \sum_j \rho_{jn}\boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0}$ and $\boldsymbol{S}_T^*$ is a consistent estimator of $Cov(T^*)$

$$= \sum_j \rho_{jn} Cov(T_{jn}^*) + \sum_j \rho_{jn}\boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0}\hat{\boldsymbol{\beta}}_{I_j,0}^T\boldsymbol{A}^T - E(T^*)[E(T^*)]^T$$

where asymptotically the sum is over $j : S \subseteq I_j$. If $\boldsymbol{\theta}_0 = \boldsymbol{0}$, then $n\boldsymbol{S}_T^* = \boldsymbol{\Sigma}_A + O_P(1)$ where

$$nCov(T_n) \xrightarrow{P} \boldsymbol{\Sigma}_A = \sum_j \sigma^2\pi_j\boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T.$$

Then $(n\boldsymbol{S}^*_T)^{-1}$ tends to be "well behaved" if $\boldsymbol{\Sigma}_A$ is nonsingular. The prediction region (3.10) bootstraps T_n, but uses $\overline{T}^*$ to increase the coverage for moderate samples.

Some special cases are also interesting. Suppose $\pi_d = 1$ so $\boldsymbol{u} \sim \boldsymbol{u}_d \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_d)$. This occurs for C_p if $a_S = p$ so S is the full model, and for methods like BIC that choose I_S with probability going to one. Knight and Fu (2000) had similar bootstrap results for this case. Next, if for each $\pi_j > 0$, $\boldsymbol{A}\boldsymbol{u}_j \sim N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\Sigma}_j\boldsymbol{A}^T) = N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$, then
$\boldsymbol{A}\boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.

In the simulations where S is not the full model, inference with forward selection with I_{min} using C_p appears to be more precise than inference with the OLS full model if $n \geq 20p$ and $B \geq 50p$. Higher than nominal coverage can occur because of the zero padding. It is possible that $\boldsymbol{S}^*_T$ is singular if a column of the bootstrap sample is equal to $\boldsymbol{0}$.

Examining $\hat{\boldsymbol{\beta}}_S$ and $\hat{\boldsymbol{\beta}}_E$ is informative for I_{min}. See Equation (1.3). First assume that the nontrivial predictors are orthogonal or uncorrelated with zero mean so $\boldsymbol{X}^T\boldsymbol{X}/n$
$\rightarrow diag(d_1, ..., d_p)$ as $n \rightarrow \infty$ where each $d_i > 0$. Then $\hat{\boldsymbol{\beta}}_S$ has the same multivariate normal limiting distribution for I_{min} and for the OLS full model. The bootstrap distribution for $\hat{\boldsymbol{\beta}}_E$ is a mixture of zeros and a distribution that would produce a confidence region for $\boldsymbol{A}\boldsymbol{\beta}_E = \boldsymbol{0}$ that has asymptotic coverage of $\boldsymbol{0}$ equal to $100(1-\delta)$%. Hence the asymptotic coverage is greater than the nominal coverage provided that $\boldsymbol{S}^*_T$ in nonsingular with probability going to one (e.g., $p - a_S$ is small), where $T = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{E,I_{min},0}$. For uncorrelated predictors with zero mean, the number of bootstrap samples $B \geq 50p$ may work well for the shorth confidence intervals and for testing $\boldsymbol{A}\boldsymbol{\beta}_S = \boldsymbol{0}$.

In the simulations for forward selection, coverages did not change much as the ρ was increased from zero to near one, where ρ was the correlation between any two nontrivial predictors. Under model (1.3), we still have that $\hat{\boldsymbol{\beta}}_{I_j,0}$ is a $\sqrt{n}$ consistent asymptotically normal estimator of $\boldsymbol{\beta} = (\boldsymbol{\beta}_S^T, \boldsymbol{\beta}_E^T)^T$ where $\boldsymbol{\beta}_E = \boldsymbol{0}$. Hence the limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_{min},0} - \boldsymbol{\beta})$ is a mixture of $N_p(\boldsymbol{0}, \sigma^2\boldsymbol{V}_{j,0})$ distributions, and the limiting distribution of $\sqrt{n}(\hat{\beta}_{i,I_{min},0} - \beta_i)$ is a mixture of $N(0, \sigma^2_{ij})$ distributions. For a β_i that is a component of

$\boldsymbol{\beta}_S$, the symmetric mixture distribution has a pdf. Then the simulated shorth confidence intervals have coverage near the nominal coverage if n and B are large enough.

Note that there are several important variable selection models, including the model given by Equation (1.3). Another model is $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_{S_i}^T\boldsymbol{\beta}_{S_i}$ for $i = 1, ..., J$. Then there are $J \geq 2$ competing "true" nonnested submodels where $\boldsymbol{\beta}_{S_i}$ is $a_{S_i} \times 1$. For example, suppose the $J = 2$ models have predictors x_1, x_2, x_3 for S_1 and x_1, x_2, x_4 for S_2. Then x_3 and x_4 are likely to be selected and omitted often by forward selection for the B bootstrap samples. Hence omitting all predictors x_i that have a $\beta_{ij}^* = 0$ for at least one of the bootstrap samples $j = 1, ..., B$ could result in underfitting, e.g. using just x_1 and x_2 in the above $J = 2$ example. If n and B are large enough, the singleton set $\{\boldsymbol{0}\}$ could still be the "100%" confidence region for a vector $\boldsymbol{\beta}_O$.

Suppose the predictors x_i have been standardized. Then another important regression model has the β_i taper off rapidly, but no coefficients are equal to zero. For example, $\beta_i = e^{-i}$ for $i = 1, ..., p$.

For $g = 1$, the percentile method uses an interval that contains $U_B \approx k_B = \lceil B(1-\delta) \rceil$ of the T_i^* from a bootstrap sample $T_1^*, ..., T_B^*$ where the statistic T_n is an estimator of θ based on a sample of size n. Note that the squared Mahalanobis distance $D_\theta^2 = (\theta - \overline{T^*})^2/S_T^{2*} \leq D_{(U_B)}^2$ is equivalent to $\theta \in [\overline{T^*} - S_T^* D_{(U_B)}, \overline{T^*} + S_T^* D_{(U_B)}]$, 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 $g = 1$. Efron (2014) used a similar large sample $100(1-\delta)\%$ confidence interval assuming that $\overline{T}^*$ is asymptotically normal. The Frey (2013) shorth(c) interval (3.1) (with c given by (3.3)) applied to the T_i^* is recommended since the shorth confidence interval can be much shorter than the Efron (2014) or prediction region method confidence intervals if $g = 1$. The shorth confidence interval is a practical implementation of the Hall (1988) shortest bootstrap interval based on all possible bootstrap samples. Note that if $\sqrt{n}(T_n - \theta) \xrightarrow{D} \boldsymbol{u}$ and $\sqrt{n}(T_i^* - \theta) \xrightarrow{D} \boldsymbol{u}$ where $\boldsymbol{u}$ has a symmetric probability density function, then the shorth confidence interval is asymptotically equivalent to the usual percentile method confidence interval that uses the central proportion of the bootstrap sample.

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 χ^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.

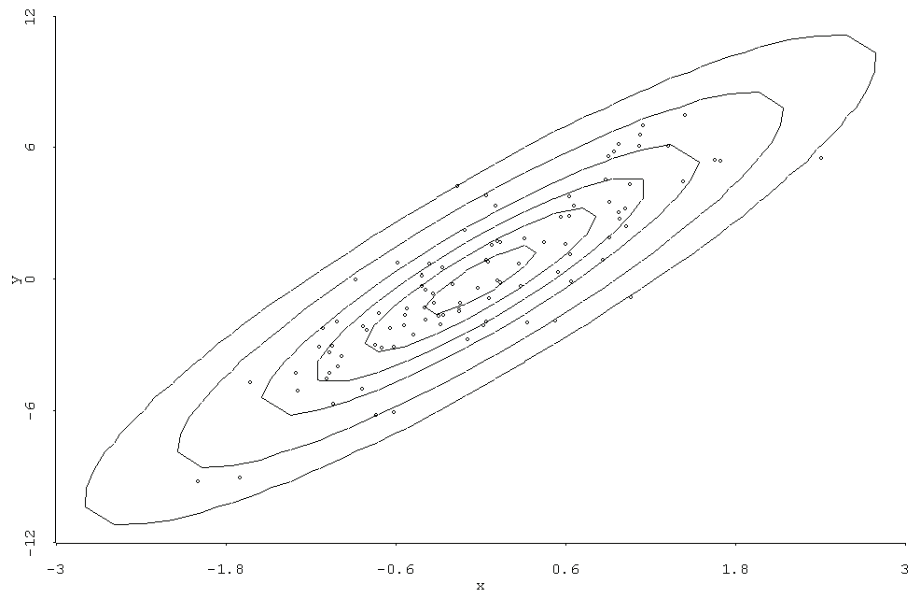
CHAPTER 4

EXAMPLE AND SIMULATIONS

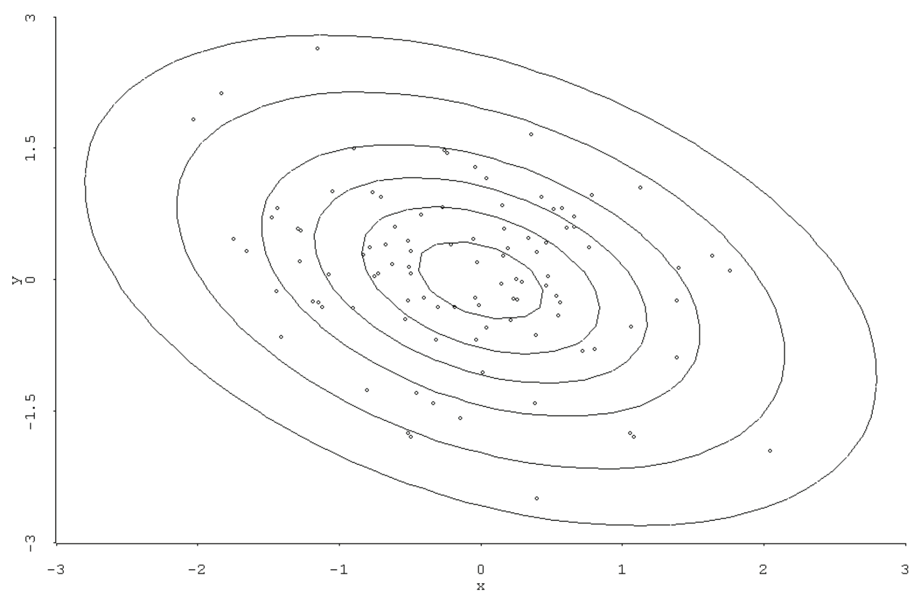
Figure 1 shows 10%, 30%, 50%, 70%, 90% and 98% prediction regions for a future value of T_f for two multivariate normal distributions. The plotted points are iid $T_1, ..., T_B$ with $B = 100$.

Example. The Hebbler (1847) data was collected from $n = 26$ districts in Prussia in 1843. We will study the relationship between $Y =$ the *number of women married to civilians* in the district with the predictors $x_1 =$ constant, $x_2 = pop =$ the *population of the district in 1843*, $x_3 = mmen =$ the *number of married civilian men* in the district, $x_4 = mmilmen =$ *number of married men in the military* in the district, and $x_5 = milwmn =$ 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. Forward selection with BIC selected the model a constant and *mmen*.

Let $\boldsymbol{x} = (1 \ \boldsymbol{u}^T)^T$ where $\boldsymbol{u}$ is the $(p-1) \times 1$ vector of nontrivial predictors. In the simulations, for $i = 1, ..., n$, we generated $\boldsymbol{w}_i \sim N_{p-1}(\boldsymbol{0}, \boldsymbol{I})$ where the $m = p - 1$ elements of the vector $\boldsymbol{w}_i$ are iid N(0,1). Let the $m \times m$ matrix $\boldsymbol{A} = (a_{ij})$ with $a_{ii} = 1$ and $a_{ij} = \psi$ where $0 \leq \psi < 1$ for $i \neq j$. Then the vector $\boldsymbol{u}_i = \boldsymbol{A}\boldsymbol{w}_i$ so that
$Cov(\boldsymbol{u}_i) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ where the diagonal entries $\sigma_{ii} = [1 + (m-1)\psi^2]$ and the off diagonal entries $\sigma_{ij} = [2\psi + (m-2)\psi^2]$. Hence the correlations are
$cor(x_i, x_j) = \rho = (2\psi + (m-2)\psi^2)/(1 + (m-1)\psi^2)$ 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 ψ gets close to 1, the predictor vectors cluster about the line in the direction of $(1, ..., 1)^T$. Let
$Y_i = 1 + 1x_{i,2} + \cdots + 1x_{i,k+1} + e_i$ for $i = 1, ..., n$. Hence $\boldsymbol{\beta} = (1, .., 1, 0, ..., 0)^T$ with $k + 1$ ones and $p - k - 1$ zeros. The zero mean errors e_i were iid from five distributions: i)



)



b)

Figure 4.1. Prediction Regions

N(0,1), ii) t_3, iii) 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.

A small simulation was done using $B = \max(1000, n, 20p)$ and 5000 runs. So an observed coverage in [0.94, 0.96] gives no reason to doubt that the CI or confidence region has the nominal coverage of 0.95. The simulation used $p = 4, 6, 7, 8$, and 10; $n = 25p$ and $50p$, $\psi = 0, 1/\sqrt{p}$, and 0.9; and $k = 1$ and $p - 2$.

When $\psi = 0$, 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 Frey shorth(c) interval for each β_i and used bootstrap confidence regions to test whether first $k+1$ $\beta_i = 1$ and the last $p-k-1$ $\beta_i = 0$. The nominal coverage was 0.95 with $\delta = 0.05$. Observed coverage between 0.94 and 0.96 would suggest coverage is close to the nominal value.

The regression models used the residual bootstrap on the forward selection estimator $\hat{\boldsymbol{\beta}}_{I_{min},0}$ with BIC. Table 1 gives results for when the iid errors $e_i \sim N(0, 1)$. Two rows for each model giving the observed confidence interval coverages and average lengths of the confidence intervals. The last six columns give results for the tests. The the length and coverage = P(fail to reject H_0) for the interval $[0, D_{(U_B)}]$ or $[0, D_{(U_B),T}]$ where $D_{(U_B)}$ or $D_{(U_B),T}$ is the cutoff for the confidence region. Volumes of the confidence regions can be compared using (3.13). The first two lines of the table correspond to the R output shown below, with $g = 2$.

```
library(leaps);Y <- marry[,3]; X <- marry[,-3]
temp<-regsubsets(X,Y,method="forward")
out<-summary(temp)
out$bic
[1] -239.4149 -236.3515 -233.1085 -229.8540
Selection Algorithm: forward
         pop mmen mmilmen milwmn
```

```
1  ( 1 ) " " "*"  " "     " "
2  ( 1 ) " " "*"  "*"     " "
3  ( 1 ) "*" "*"  "*"     " "
4  ( 1 ) "*" "*"  "*"     "*"

record coverages and ``lengths" for
b1, b2, bp-1, bp, pm0,  hyb0,   BR0,  pm1,   hyb1,  BR1,
library(leaps)
bicbootsim(n=100,p=4,k=1,nruns=5000,type=1,psi=0)

$cicov
 [1] 0.9478 0.9478 0.9996 0.9998 0.9992 0.9918 0.9996 0.9408 0.9418 0.9422
$avelen
 [1] 0.3948321 0.3973231 0.2153983 0.2145764 3.4006284 3.4006282 3.6963001
 [8] 2.4501023 2.4501437 2.45612555
$beta
[1] 1 1 0 0
$k
[1] 1
```

Table 4.1. Bootstrapping OLS Forward Selection with BIC Type 1

n,p,k,ψ	β_1	β_2	β_{p-1}	β_p	pm0	hyb0	br0	pm1	hyb1	br1
100,4,1,0	0.9478	0.9478	0.9996	0.9998	0.9992	0.9918	0.9996	0.9408	0.9418	0.9422
len	0.3948	0.3973	0.2154	0.2146	3.4006	3.4006	3.6963	2.4501	2.4501	2.4561
100,4,2,0	0.9396	0.9466	0.9462	0.9998	0.9998	0.9682	0.9998	0.9326	0.9326	0.9320
len	0.3950	0.3975	0.3984	0.2195	1.8434	1.8434	2.0855	2.8003	2.8003	2.8047
100,4,1,1/$\sqrt{p}$	0.9452	0.9742	1.0000	0.9998	0.9992	0.9960	1.0000	0.9764	0.9772	0.9838
len	0.3958	0.6261	0.3595	0.3573	3.4424	3.4424	3.7143	2.5574	2.5574	2.7219
100,4,2,1/$\sqrt{p}$	0.9442	0.9596	0.9618	0.9998	0.9996	0.9740	0.9996	0.9774	0.9774	0.9834
len	0.3962	0.6512	0.6500	0.3681	1.8300	1.8300	2.0645	2.9388	2.9388	3.0528
100,4,1,0.9	0.9428	0.9486	0.9976	0.9978	1.0000	0.8894	1.0000	0.9604	0.9272	0.9576
len	0.3956	2.1746	1.9488	1.9684	2.7434	2.7434	2.9890	2.5333	2.5333	2.6716
100,4,2,0.9	0.9466	0.9110	0.9104	0.9990	0.9990	0.8854	0.9994	0.9920	0.9826	0.9948
len	0.3968	2.3035	2.2987	2.1084	2.4007	2.4007	2.7693	3.2150	3.2150	3.4741
175,7,1,0	0.9514	0.9452	0.9998	0.9998	1.0000	1.0000	1.0000	0.9422	0.9432	0.9436
len	0.2945	0.3045	0.1334	0.1354	5.1894	5.1894	5.3111	2.4342	2.4343	2.4500
175,7,5,0	0.9498	0.9234	0.9226	0.9212	0.9222	0.9252	0.9994	0.9994	0.9492	0.9994
len	0.3004	0.3011	0.3021	0.1242	1.5442	1.5542	1.7002	3.6042	3.6042	3.6212
175,7,1,1/$\sqrt{p}$	0.9498	0.9234	0.9226	0.9212	0.9222	0.9252	0.9994	0.9994	0.9492	0.9994
len	0.2991	0.4386	0.1918	0.1958	5.0443	5.0443	5.2423	2.4869	2.4869	2.5553
175,7,5,1/$\sqrt{p}$	0.9498	0.9234	0.9226	0.9212	0.9222	0.9252	0.9994	0.9614	0.9614	0.9678
len	0.3001	0.4419	0.4119	0.2188	1.5232	1.5232	1.6532	3.5999	3.5999	3.6423
175,7,1,0.9	0.9450	0.9208	0.9996	0.9996	1.0000	0.9998	0.9998	0.9182	0.8652	0.9182
len	0.2992	2.0704	1.5432	1.5433	4.5547	4.5547	4.7647	2.5887	2.5887	2.6196
175,7,5,0.9	0.9498	0.9234	0.9226	0.9212	0.9222	0.9252	0.9994	0.9722	0.9270	0.9730
len	0.3038	2.5443	2.5379	1.6935	1.7776	1.7776	1.9763	4.2855	4.2855	4.5140

Table 4.2. Bootstrapping OLS Forward Selection with BIC Type 1(cont.)

250,10,1,0	0.9495	0.9413	1.0000	1.0000	0.9995	0.9986	1.0000	0.9388	0.9378	0.9398
len	0.2512	0.2516	0.1062	0.1069	6.2092	6.2092	6.4014	2.4322	2.4322	2.4577
250,10,1,1/$\sqrt{p}$	0.9438	0.9706	1.0000	1.0000	1.0000	1.0000	1.0000	0.9636	0.9636	0.9752
len	0.2505	0.3492	0.1428	0.1405	6.1386	6.1386	6.3393	2.4222	2.4222	2.6282
250,10,1,0.9	0.9434	0.9078	1.0000	0.9998	0.9998	1.0000	1.0000	0.8762	0.8156	0.8776
len	0.2503	1.9279	1.1796	1.1844	5.6734	5.6734	5.9500	2.5454	2.5454	2.6776
250,10,8,0.9	0.9464	0.9164	0.9224	0.9160	0.9198	0.9188	0.9154	0.9014	0.6748	0.8820
len	0.2556	2.3371	2.3157	1.2916	1.5965	1.5965	1.8022	4.7095	4.7095	5.1978
300,6,1,0	0.9494	0.9476	1.0000	1.0000	1.0000	0.9996	1.0000	0.9454	0.9462	0.9460
len	0.2298	0.2307	0.0888	0.0892	4.2355	4.2355	5.0332	2.3323	2.3326	2.7442
300,6,4,0	0.9454	0.9528	0.9468	0.9484	0.9502	0.9996	0.9996	0.9818	0.9452	0.9456
len	0.2300	0.2307	0.2342	0.0966	1.2304	1.2304	1.4493	3.3326	3.3326	3.3418
300,6,1,1/$\sqrt{p}$	0.9498	0.9816	1.0000	1.0000	1.0000	0.9998	1.0000	0.9998	0.9740	0.9746
len	0.2290	0.3354	0.1270	0.1292	4.8107	4.8107	5.0252	2.7557	2.7757	2.6979
300,6,4,1/$\sqrt{p}$	0.9466	0.9570	0.9498	0.9552	0.9536	1.0000	1.0000	0.9692	0.9696	0.9736
len	0.2300	0.3473	0.3470	0.1397	1.3121	1.3121	1.4444	3.2214	3.2214	3.4970
300,6,1,0.9	0.9470	0.9384	1.0000	0.9998	1.0000	0.9998	0.9976	0.9692	0.9978	0.9252
len	0.2297	1.7048	1.1010	1.1001	4.0613	4.0613	4.2221	2.4661	2.4661	2.6089
300,6,4,0.9	0.9530	0.9286	0.9292	0.9998	0.9998	0.9696	0.9998	0.9230	0.8770	0.9242
len	0.2315	2.1667	2.1546	1.2109	1.6445	1.6445	1.8543	3.7622	3.7622	4.1773
400,8,1,0	0.9488	0.9554	1.0000	0.9998	0.9998	1.0000	1.0000	0.9434	0.9450	0.9466
len	0.1987	0.1989	0.0716	0.0773	5.7824	5.7824	5.9973	2.5753	2.5753	2.6002
400,8,6,0	0.9468	0.9544	0.9444	1.0000	1.0000	0.9836	1.0000	0.9418	0.9428	0.9428
len	0.1994	0.1998	0.1999	0.0732	1.5532	1.5532	1.6665	3.1142	3.1142	3.2001

Table 4.3. Bootstrapping OLS Forward Selection with BIC Type 1(cont.)

400,8,1,1/$\sqrt{p}$	0.9488	0.9808	1.0000	1.0000	1.0000	1.0000	1.0000	0.9682	0.9712	0.9796
len	0.1986	0.2812	0.0933	0.1077	5.2232	5.2232	5.3090	2.4112	2.4112	2.6968
400,8,6,1/$\sqrt{p}$	0.9496	0.9554	0.9518	0.9518	0.9468	0.9444	0.9518	0.9604	0.9600	0.9644
len	0.1993	0.2854	0.2766	0.1043	1.2985	1.2985	1.3117	3.7884	3.7884	3.8965
400,8,1,0.9	0.9464	0.9438	1.0000	0.9996	1.0000	0.9852	0.9966	0.9154	0.8384	0.9162
len	0.1990	1.6517	0.8133	0.8129	5.0542	5.0542	5.2327	2.4098	2.4098	2.5056
400,8,6,0.9	0.9456	0.9108	0.9144	1.0000	1.0000	0.9666	1.0000	0.8976	0.7948	0.8774
len	0.2010	2.0326	2.0226	0.9537	1.4552	1.4552	1.6662	4.1299	4.1299	4.5120

Table 4.4. Bootstrapping OLS Forward Selection with BIC Type 2

n,p,k,ψ	β_1	β_2	β_{p-1}	β_p	pm0	hyb0	br0	pm1	hyb1	br1
100,4,1,0	0.9478	0.9484	0.9998	0.9998	0.9998	0.9936	0.9998	0.9518	0.9508	0.9522
len	0.6652	0.6778	0.3656	0.3615	3.3553	3.3554	3.6589	2.4765	2.4765	2.4848
100,4,2,0	0.9396	0.9348	0.9404	1.0000	1.0000	0.9704	1.0000	0.9450	0.9472	0.9498
len	0.6582	0.6884	0.6892	0.3687	1.8572	1.8572	2.1113	2.8692	2.8692	2.8835
100,4,1,1/$\sqrt{p}$	0.9414	0.9646	1.0000	0.9998	0.9996	0.9880	1.0000	0.9700	0.9556	0.9746
len	0.6580	1.1045	0.6889	0.6885	3.3182	3.3182	3.6039	2.5784	2.5784	2.7611
100,4,2,1/$\sqrt{p}$	0.9436	0.9334	0.9328	0.9998	0.9998	0.9652	0.9996	0.9706	0.9448	0.9672
len	0.6628	1.2667	1.2675	0.7583	2.2142	2.2412	2.5027	3.0391	3.0390	3.2039
100,4,1,0.9	0.9436	0.9336	0.9980	0.9966	0.9984	0.9286	0.9998	0.9754	0.9712	0.9790
len	0.6602	3.4802	3.3721	3.4000	3.0051	3.0051	3.2116	2.6446	2.6446	2.7773
100,4,2,0.9	0.9464	0.9034	0.9072	0.9976	0.9982	0.8782	0.9996	0.9892	0.9796	0.9922
len	0.6548	3.2829	3.3038	3.1870	2.0573	2.0574	2.4191	3.0484	3.0484	3.2597
175,7,1,0	0.9452	0.9448	0.9998	1.0000	1.0000	1.0000	1.0000	0.9416	0.9408	0.9418
len	0.5017	0.5115	0.2307	0.2320	5.0702	5.0702	5.2886	2.4680	2.4680	2.4764
175,7,5,0	0.9386	0.9492	0.9446	1.0000	1.0000	0.9774	1.0000	0.9498	0.9480	0.9498
len	0.5009	0.5129	0.5117	0.2394	1.5668	1.5668	1.7745	3.6649	3.6649	3.6723
175,7,1,1/$\sqrt{p}$	0.9472	0.9702	1.0000	0.9998	0.9998	0.9998	0.9998	0.9748	0.9748	0.9818
len	0.4998	0.7591	0.3363	0.3399	5.0069	5.0069	5.2248	2.5067	2.5067	2.6653
175,7,5,1/$\sqrt{p}$	0.9448	0.9352	0.9414	1.0000	1.0000	0.9776	1.0000	0.9834	0.9778	0.9846
len	0.5070	0.8026	0.8053	0.3565	1.5584	1.5584	1.7535	3.8685	3.8685	3.9477
175,7,1,0.9	0.9440	0.8778	0.9998	1.0000	1.0000	1.0000	1.0000	0.9696	0.9638	0.9724
len	0.5001	3.1455	2.7716	2.7721	4.6978	4.6978	4.9545	2.6784	2.6784	2.8209
175,7,5,0.9	0.9392	0.8666	0.8654	1.0000	1.0000	0.9574	1.0000	0.9986	0.9960	0.9992
len	0.5063	3.5194	3.5408	2.7855	1.9122	1.9122	2.1371	4.4440	4.4440	4.7823

Table 4.5. Bootstrapping OLS Forward Selection with BIC Type 2(cont.)

250,10,1,0	0.9422	0.9490	0.9996	1.0000	1.0000	1.0000	1.0000	0.9444	0.9448	0.9464
len	0.4239	0.4310	0.1768	0.1777	6.1685	6.1685	6.3665	2.4669	2.4669	2.4745
250,10,1,1/$\sqrt{p}$	0.9474	0.9738	0.9998	1.0000	1.0000	1.0000	1.0000	0.9636	0.9662	0.9756
len	0.4203	0.5964	0.2409	0.2441	6.0990	6.0990	6.3031	2.4802	2.4802	2.6346
250,10,1,0.9	0.9412	0.8430	0.9998	1.0000	1.0000	1.0000	1.0000	0.9350	0.9206	0.9360
len	0.4244	2.7286	2.1710	2.1765	5.7059	5.7059	5.9675	2.6515	2.6516	2.7959
250,10,8,0.9	0.9424	0.8612	0.8580	0.9998	0.9998	0.9458	0.9996	0.9974	0.9812	0.9958
len	0.4269	3.2147	3.2002	2.1729	1.6402	1.6402	1.8454	5.1395	5.1395	5.5527
300,6,1,0	0.9476	0.9454	1.0000	1.0000	0.9998	0.9996	1.0000	0.9488	0.9486	0.9490
len	0.3874	0.3920	0.1589	0.1566	4.7395	4.7395	4.9914	2.3908	2.3908	2.4545
300,6,4,0	0.9466	0.9472	0.9484	0.9998	0.9996	0.9822	0.9998	0.9552	0.9568	0.9564
len	0.3885	0.3966	0.3946	0.1554	1.3390	1.3390	1.4854	3.4049	3.4049	3.4086
300,6,1,1/$\sqrt{p}$	0.9512	0.9796	1.0000	1.0000	1.0000	1.0000	1.0000	0.9752	0.9742	0.9840
len	0.3874	0.5748	0.2257	0.2247	4.7823	4.7823	5.0229	2.5368	2.5368	2.6983
300,6,4,1/$\sqrt{p}$	0.9478	0.9474	0.9522	0.9506	0.9532	0.9998	0.9998	0.9788	0.9804	0.9838
len	0.3878	0.6050	0.6064	0.2379	1.3026	1.3026	1.4389	3.5264	3.5264	3.5912
300,6,1,0.9	0.9504	0.9004	0.9998	0.9998	0.9998	0.9998	1.0000	0.9398	0.8990	0.9386
len	0.3905	2.2815	1.8054	1.8165	4.1676	4.1676	4.3697	2.5782	2.5782	2.6982
300,6,4,0.9	0.9524	0.8918	0.8924	0.9996	0.9996	0.9672	0.9996	0.9886	0.9752	0.9912
len	0.3906	2.9795	2.9661	1.9804	1.7049	1.7049	1.9000	4.0476	4.0476	4.4212
400,8,1,0	0.9496	0.9474	1.0000	1.0000	1.0000	1.0000	1.0000	0.9434	0.9440	0.9444
len	0.3373	0.3407	0.1241	0.1247	5.7576	5.7576	5.9650	2.4643	2.4643	2.4682
400,8,6,0	0.9448	0.9506	0.9576	1.0000	1.0000	0.9864	1.0000	0.9542	0.9542	0.9542
len	0.3379	0.3420	0.3419	0.1275	1.2489	1.2489	1.3696	3.8506	3.8506	3.8540

Table 4.6. Bootstrapping OLS Forward Selection with BIC Type 2(cont.)

400,8,1,1/$\sqrt{p}$	0.9518	0.9792	1.0000	1.0000	1.0000	1.0000	1.0000	0.9750	0.9760	0.9834
len	0.3363	0.4812	0.1726	0.1686	5.7419	5.7419	5.9311	2.5085	2.5085	2.6737
400,8,6,1/$\sqrt{p}$	0.9482	0.9504	0.9514	1.0000	0.9998	0.9850	1.0000	0.9684	0.9694	0.9732
len	0.3379	0.4933	0.4916	0.1781	1.2154	1.2154	1.3313	3.9364	3.9364	3.9815
400,8,1,0.9	0.9478	0.8698	1.0000	1.0000	1.0000	0.9966	1.0000	0.8994	0.8362	0.8930
len	0.3369	2.1940	1.5615	1.5275	5.1174	5.1174	5.3412	2.5855	2.5855	2.7040
400,8,6,0.9	0.9522	0.8918	0.9018	1.0000	1.0000	0.9588	1.0000	0.9672	0.9232	0.9672
len	0.3399	2.7933	2.8184	1.5926	1.4750	1.4750	1.6518	4.5143	4.5143	4.9527

Table 4.7. Bootstrapping OLS Forward Selection with BIC Type 3

n,p,k,ψ	β_1	β_2	β_{p-1}	β_p	pm0	hyb0	br0	pm1	hyb1	br1
100,4,1,0	0.9412	0.9504	1.0000	1.0000	0.9998	0.9954	0.9998	0.9366	0.9372	0.9372
len	0.3916	0.3912	0.2150	0.2145	3.3608	3.3609	3.3609	3.6704	2.4538	2.4538
100,4,2,0	0.9378	0.9458	0.9478	1.0000	0.9998	0.9708	0.9998	0.9312	0.9316	0.9332
len	0.3915	0.3974	0.3978	0.2179	1.8355	1.8355	2.0802	2.8209	2.8209	2.8256
100,4,1,1/$\sqrt{p}$	0.9420	0.9770	0.9998	0.9998	0.9992	0.9954	0.9996	0.9672	0.9680	0.9764
len	0.3911	0.6210	0.3507	0.3496	3.4236	3.4236	3.6593	2.5516	2.5516	2.7117
100,4,2,1/$\sqrt{p}$	0.9382	0.9516	0.9460	0.9996	0.9996	0.9734	0.9996	0.9678	0.9672	0.9750
len	0.3913	0.6547	0.6578	0.3676	1.8384	1.8384	2.0840	2.9561	2.9561	3.0718
100,4,1,0.9	0.9360	0.9464	0.9980	0.9982	0.9986	0.9034	0.9996	0.9484	0.9222	0.9508
len	0.3909	2.1869	1.9452	1.9543	2.7413	2.7413	2.9875	2.5190	2.5190	2.6552
100,4,2,0.9	0.9402	0.9120	0.9172	0.9984	0.8966	0.9994	0.9822	0.9734	0.9822	0.9884
len	0.3911	2.2675	2.2986	2.0796	2.3662	2.3662	2.7231	3.2088	3.2088	3.4783
175,7,1,0	0.9364	0.9444	1.0000	0.9998	1.0000	1.0000	1.0000	0.9354	0.9338	0.9354
len	0.2966	0.2988	0.1376	0.1370	5.0766	5.0766	5.2946	2.4532	2.4532	2.4612
175,7,5,0	0.9376	0.9472	0.9408	1.0000	1.0000	0.9804	1.0000	0.9342	0.9358	0.9362
len	0.2991	0.3019	0.3021	0.1370	1.5203	1.5203	1.7048	3.6059	3.6059	3.6094
175,7,1,1/$\sqrt{p}$	0.9454	0.9790	0.9998	1.0000	1.0000	1.0000	1.0000	0.9622	0.9644	0.9748
len	0.2978	0.4381	0.1988	0.2018	5.0196	5.0196	5.2325	2.4873	2.4873	2.6507
175,7,5,1/$\sqrt{p}$	0.9410	0.9500	0.9490	0.9998	0.9996	0.9764	0.9996	0.9578	0.9578	0.9626
len	0.2987	0.4419	0.4420	0.2074	1.5506	1.5506	1.7450	3.6888	3.6888	3.7518
175,7,1,0.9	0.9444	0.9176	0.9998	0.9996	1.0000	0.9938	1.0000	0.9141	0.8626	0.9138
len	0.2979	2.0714	1.5217	1.5289	4.4997	4.4997	4.7467	2.5438	2.5438	2.6715
175,7,5,0.9	0.9432	0.9280	0.9208	0.9998	0.9998	0.9454	0.9998	0.9682	0.9278	0.9702
len	0.3015	2.5473	2.5437	1.6526	1.8231	1.8231	2.0587	4.1539	4.1539	4.5718

Table 4.8. Bootstrapping OLS Forward Selection with BIC Type 3(cont.)

250,10,1,0	0.9488	0.9480	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9410	0.9420
len	0.2495	0.2511	0.1046	0.1053	0.1006	6.2009	6.2009	6.3936	2.4516	2.4596
250,10,1,1/$\sqrt{p}$	0.9422	0.9764	1.0000	1.0000	1.0000	1.0000	1.0000	0.9640	0.9652	0.9736
len	0.2487	0.3473	0.1396	0.1390	6.1349	6.1349	6.3352	2.4672	2.4672	2.6317
250,10,1,0.9	0.9496	0.9038	0.9998	0.9998	0.9998	0.9992	1.0000	0.8796	0.8226	0.8812
len	0.2503	1.9361	1.1754	1.1703	5.6670	5.6670	5.9426	2.5317	2.5317	2.6659
250,10,8,0.9	0.9444	0.9212	0.9200	0.9996	0.9996	0.9474	0.9996	0.9114	0.7780	0.8890
len	0.2544	2.3389	2.3467	2.3290	2.3483	1.2941	1.6328	4.7296	4.7296	5.2096
300,6,1,0	0.9442	0.9528	1.0000	1.0000	1.0000	0.9996	1.0000	0.9424	0.9414	0.9420
len	0.2290	0.2302	0.0908	0.0920	4.7612	4.7612	5.0091	2.4517	2.4517	2.4550
300,6,4,0	0.9508	0.9512	0.9424	1.0000	0.9998	0.9842	1.0000	0.9388	0.9390	0.9394
len	0.2289	0.2300	0.2302	0.0915	1.3204	1.3204	1.4537	3.3613	3.3613	3.3637
300,6,1,1/$\sqrt{p}$	0.9478	0.9794	1.0000	1.0000	1.0000	1.0000	1.0000	0.9732	0.9748	0.9814
len	0.2285	0.3346	0.1296	0.1293	4.8306	4.8306	5.0424	2.5326	2.5326	2.6973
300,6,4,1/$\sqrt{p}$	0.9468	0.9516	0.9578	1.0000	0.9998	0.9844	1.0000	0.9652	0.9662	0.9708
len	0.2291	0.3470	0.3472	0.1425	1.3540	1.3504	1.4912	3.4517	3.4517	3.5168
300,6,1,0.9	0.9504	0.9386	0.9996	1.0000	0.9968	0.9734	0.9972	0.9208	0.8368	0.9228
len	0.2288	1.6864	1.0954	1.0822	4.0650	4.0650	4.2720	2.4643	2.4643	2.6102
300,6,4,0.9	0.9512	0.9316	0.9314	0.9994	0.9992	0.9612	0.9992	0.9292	0.8848	0.9292
len	0.2307	2.1519	2.1536	1.1358	1.6014	1.6014	1.8020	3.7569	3.7569	4.1761
400,8,1,0	0.9496	0.9500	1.0000	1.0000	1.0000	1.0000	1.0000	0.9194	0.9410	0.9412
len	0.1983	0.1989	0.0729	0.0727	5.7878	5.7878	5.9940	2.5409	2.4509	2.4543
400,8,6,0	0.9512	0.9504	0.9478	0.9998	0.9996	0.9856	0.9996	0.9404	0.9414	0.9412
len	0.1986	0.1995	0.1998	0.0737	1.2149	1.2149	1.3267	3.7947	3.7947	3.7973

Table 4.9. Bootstrapping OLS Forward Selection with BIC Type 3(cont.)

400,8,1,1/$\sqrt{p}$	0.9500	0.9814	1.0000	1.0000	0.9998	0.9998	1.0000	0.9708	0.9690	0.9780
len	0.1979	0.2816	0.0974	0.0994	5.7646	5.7646	5.9522	2.4991	2.4991	2.6660
400,8,6,1/$\sqrt{p}$	0.9496	0.9538	0.9466	0.9484	0.9560	0.9998	0.9998	0.9860	0.9998	0.9876
len	0.1986	0.2850	0.2851	0.1077	1.2427	1.2427	1.3585	3.8743	3.8743	3.9204
400,8,1,0.9	0.9490	0.9418	0.9970	0.9882	0.9974	0.9168	0.8432	1.0000	0.9164	0.9866
len	0.1983	1.6620	0.8137	0.8039	5.1044	5.1044	5.3029	2.4428	2.4428	2.5860
400,8,6,0.9	0.9500	0.9164	0.9160	0.9996	0.9996	0.9624	0.9996	0.8934	0.8014	0.8784
len	0.2003	2.0217	2.0326	0.9321	1.4769	1.4769	1.6566	4.1085	4.1085	4.6190

Table 4.10. Bootstrapping OLS Forward Selection with BIC Type 4

n,p,k,ψ	β_1	β_2	β_{p-1}	β_p	pm0	hyb0	br0	pm1	hyb1	br1
100,4,1,0	0.9486	0.9518	0.9998	0.9998	0.9994	0.9992	0.9424	0.9420	0.9422	0.9422
len	0.2284	0.2302	0.1257	0.1243	3.3881	3.3881	3.6908	2.4449	2.4449	2.4509
100,4,2,0	0.9474	0.9466	0.9420	0.9996	0.9996	0.9716	0.9996	0.9340	0.9356	.9368
len	0.2289	0.2307	0.2303	0.1277	1.8762	1.8762	2.1145	2.7925	2.7925	2.7964
100,4,1,1/$\sqrt{p}$	0.9430	0.9804	0.9994	0.9998	0.9988	0.9950	0.9990	0.9708	0.9686	0.9802
len	0.2284	0.3574	0.2035	0.2044	3.4673	3.4673	3.7354	2.5651	2.5651	2.7287
100,4,2,1/$\sqrt{p}$	0.9468	0.9560	0.9552	0.9992	0.9990	0.9678	0.9992	0.9722	0.9718	0.9770
len	0.2290	0.3724	0.3719	0.2141	1.8515	1.8515	2.0875	2.9239	2.9239	3.0373
100,4,1,0.9	0.9462	0.9596	0.9996	0.9992	0.9972	0.9934	0.9990	0.9504	0.8934	0.9534
len	0.2291	1.5321	1.1806	1.1810	2.9694	2.9694	3.2135	2.4636	2.4636	2.6350
100,4,2,0.9	0.9450	0.9350	0.9384	1.0000	0.9998	0.9672	0.9998	0.9412	0.9156	0.9498
len	0.2309	1.9054	1.9038	1.3513	2.1684	2.1684	2.4246	3.1921	3.1921	3.5055
175,7,1,0	0.9486	0.9502	0.9998	0.9998	1.0000	1.0000	1.0000	0.9404	0.9422	0.9432
len	0.1730	0.1736	0.0785	0.0786	5.1223	5.1223	5.3305	2.4481	2.4481	2.4552
175,7,5,0	0.9440	0.9516	0.9454	1.0000	1.0000	0.9786	1.0000	0.9312	0.9306	0.9310
len	0.1734	0.1743	0.1743	0.0825	1.5569	1.5569	1.7475	3.5505	3.5504	3.5543
175,7,1,1/$\sqrt{p}$	0.9494	0.9774	1.0000	0.9996	0.9996	0.9996	0.9998	0.9688	0.9676	0.9776
len	0.1728	0.2530	0.1153	0.1154	5.0389	5.0389	5.2498	2.4884	2.4885	2.6550
175,7,5,1/$\sqrt{p}$	0.9454	0.9450	0.9462	1.0000	1.0000	0.9794	1.0000	0.9558	0.9560	0.9618
len	0.1736	0.2552	0.2555	0.1192	1.5581	1.5581	1.7409	3.6406	3.6406	3.7037
175,7,1,0.9	0.9538	0.9664	1.0000	1.0000	0.9984	0.9898	0.9988	0.9466	0.8914	0.9500
len	0.1734	1.5276	0.8450	0.8397	4.7090	4.7090	4.9299	2.4204	2.4204	2.5734
175,7,5,0.9	0.9462	0.9332	0.9304	0.9998	0.9998	0.9580	0.9998	0.9142	0.8318	0.8992
len	0.1760	1.7847	1.7941	0.9309	1.7665	1.7665	1.9949	3.9187	3.9187	4.2770

Table 4.11. Bootstrapping OLS Forward Selection with BIC Type 4(cont.)

250,10,1,0	0.9486	0.9496	0.9998	1.0000	0.9994	1.0000	1.0000	0.9436	0.9436	0.9448
len	0.1446	0.1450	0.0591	0.0605	6.2170	6.2170	6.4070	2.4488	2.4488	2.4565
250,10,1,1/$\sqrt{p}$	0.9432	0.9782	1.0000	1.0000	1.0000	1.0000	1.0000	0.9608	0.9628	0.9698
len	0.1446	0.2017	0.0826	0.0813	6.1348	6.1348	6.3360	2.4660	2.4660	2.6299
250,10,1,0.9	0.9484	0.9780	0.9998	0.9998	0.9992	0.9986	0.9994	0.9498	0.9022	0.9550
len	0.1449	1.4437	0.5954	0.6013	6.0119	6.0119	6.2151	2.4261	2.4261	2.5819
250,10,8,0.9	0.9498	0.9288	0.9318	0.9336	0.9998	0.9996	0.9728	0.9996	0.8174	0.8844
len	0.1479	1.5961	1.5993	0.7035	1.5692	1.5692	1.7678	4.4792	4.4792	4.8229
300,6,1,0	0.9516	0.9490	1.0000	0.9998	0.9998	0.9994	0.9998	0.9478	0.9466	0.9470
len	0.1326	0.1327	0.0524	0.0533	4.7810	4.7810	5.0298	2.4491	2.4491	2.4527
300,6,4,0	0.9508	0.9416	0.9538	1.0000	0.9996	0.9812	1.0000	0.9450	0.9438	0.9444
len	0.1327	0.1231	0.1332	0.0525	1.2950	1.2950	1.4244	3.3317	3.3317	3.3347
300,6,1,1/$\sqrt{p}$	0.9490	0.9820	1.0000	1.0000	1.0000	1.0000	1.0000	0.9726	0.9748	0.9836
len	0.1326	0.1944	0.0765	0.0757	4.8152	4.8152	5.0303	2.5320	2.5320	2.6990
300,6,4,1/$\sqrt{p}$	0.9562	0.9472	0.9548	1.0000	1.0000	0.9828	1.0000	0.9662	0.9654	0.9700
len	0.1329	0.2004	0.2006	0.0820	1.3333	1.3333	1.4697	3.4292	3.4292	3.4938
300,6,1,0.9	0.9520	0.9800	1.0000	0.9998	0.9992	0.9946	0.9992	0.9740	0.9504	0.9814
len	0.1327	1.1535	0.5278	0.5256	4.6690	4.6690	4.8983	2.5783	2.5783	2.8073
300,6,4,0.9	0.9530	0.9420	0.9438	0.9998	0.9998	0.9780	0.9998	0.9580	0.9250	0.9594
len	0.1336	1.4494	1.4438	0.5901	1.4967	1.4968	1.6719	3.6865	3.6865	3.9079
400,8,1,0	0.9512	0.9456	1.0000	1.0000	1.0000	1.0000	1.0000	0.9466	0.9462	0.9468
len	0.1148	0.1150	0.0424	0.0423	5.8070	5.8070	6.0107	2.4489	2.4489	2.4527
400,8,6,0	0.9492	0.9450	0.9480	1.0000	1.0000	0.9868	1.0000	0.9410	0.9410	0.9424
len	0.1149	0.1152	0.1152	0.0429	1.2320	1.2320	1.3476	3.7631	3.7631	3.7652

Table 4.12. Bootstrapping OLS Forward Selection with BIC Type 4(cont.)

400,8,1,1/$\sqrt{p}$	0.9550	0.9822	1.0000	1.0000	1.0000	1.0000	1.0000	0.9708	0.9696	0.9794
len	0.1148	0.1632	0.0580	0.0590	5.7763	5.7763	5.9636	2.5016	2.5016	2.6677
400,8,6,1/$\sqrt{p}$	0.9498	0.9490	0.9488	1.0000	1.0000	0.9882	1.0000	0.9618	0.9638	0.9670
len	0.1149	0.1648	0.1647	0.0615	1.2425	1.2425	1.3549	3.8397	3.8397	3.8867
400,8,1,0.9	0.9518	0.9850	1.0000	1.0000	0.9998	0.9994	0.9998	0.9724	0.9616	0.9840
len	0.1148	1.0996	0.4184	0.4083	5.7252	5.7252	5.9246	2.5917	2.5917	2.8365
400,8,6,0.9	0.9522	0.9410	0.9402	0.9998	0.9996	0.9832	0.9998	0.9658	0.9418	0.9630
len	0.1158	1.2818	1.2789	0.4604	1.3524	1.3524	1.4912	4.1606	4.1606	4.3391

Table 4.13. Bootstrapping OLS Forward Selection with BIC Type 5

n,p,k,ψ	β_1	β_2	β_{p-1}	β_p	pm0	hyb0	br0	pm1	hyb1	br1
100,4,1,0	0.9432	0.9414	0.9998	1.0000	0.9992	0.9954	0.9996	0.9410	0.9288	0.9446
len	1.2681	1.3469	0.8314	0.8318	3.1723	3.1723	3.4764	2.5255	2.5255	2.5690
100,4,2,0	0.9422	0.9342	0.9362	0.9998	0.9998	0.9634	0.9998	0.9254	0.8716	0.9178
len	1.2735	1.3733	1.3800	0.7521	1.9297	1.9297	2.2047	2.9693	2.9693	3.1249
100,4,1,1/$\sqrt{p}$	0.9400	0.9604	0.9992	0.9992	0.9978	0.9720	0.9990	0.9536	0.9254	0.9578
len	1.2754	1.8040	1.4672	1.4725	3.0138	3.0138	3.2893	2.5095	2.5095	2.6738
100,4,2,1/$\sqrt{p}$	0.9394	0.9360	0.9330	1.0000	1.0000	0.9436	1.0000	0.9562	0.9338	0.9630
len	1.2724	2.0618	2.0483	1.4971	2.2581	2.2581	2.5647	3.0645	3.0645	3.3658
100,4,1,0.9	0.9332	0.9310	0.9976	0.9980	0.9986	0.9842	0.9998	0.9824	0.9834	0.9884
len	1.2691	6.8111	6.8183	6.8333	3.2946	3.2946	3.5151	2.7596	2.7596	2.9196
100,4,2,0.9	0.9376	0.9604	0.9014	0.9984	0.9976	0.9080	0.9970	0.9932	0.9900	0.9962
len	1.2677	6.3865	6.3837	6.3609	1.9602	1.9602	2.2601	3.3072	3.3072	3.4745
175,7,1,0	0.9462	0.9288	1.0000	1.0000	1.0000	1.0000	1.0000	0.9504	0.9466	0.9550
len	0.9739	1.0769	0.4739	0.4716	4.9193	4.9193	5.1507	2.5252	2.5252	2.5673
175,7,5,0	0.9424	0.9298	0.9296	1.0000	1.0000	0.9764	1.0000	0.9602	0.9208	0.9494
len	0.9877	1.1252	1.1210	0.4564	1.5369	1.5369	1.7446	3.9744	3.9744	4.0992
175,7,1,1/$\sqrt{p}$	0.9514	0.9640	0.9998	1.0000	0.9998	0.9992	0.9998	0.9468	0.9178	0.9530
len	0.9762	1.4899	0.7805	0.7824	4.7572	4.7572	4.9923	2.4946	2.4946	2.6505
175,7,5,1/$\sqrt{p}$	0.9432	0.9322	0.9386	0.9996	0.9996	0.9608	0.9996	0.9250	0.8688	0.9156
len	0.9937	1.6758	1.6747	0.8430	1.7289	1.7289	1.9676	3.9330	3.9330	4.2147
175,7,1,0.9	0.9444	0.8132	1.0000	0.9998	1.0000	0.9998	1.0000	0.9846	0.9836	0.9900
len	0.9700	6.0518	5.8893	5.8521	4.5535	4.5535	4.7975	2.8241	2.8241	3.0073
175,7,5,0.9	0.9424	0.7528	0.7412	0.9998	0.9998	0.9218	0.9998	0.9998	1.0000	1.0000
len	0.9724	5.2087	5.2194	4.8546	2.0096	2.0096	2.2791	4.5509	4.5509	4.8019

Table 4.14. Bootstrapping OLS Forward Selection with BIC Type 5(cont.)

250,10,1,0	0.9476	0.9410	0.9994	0.9998	1.0000	1.0000	1.0000	0.9596	0.9594	0.9616
len	0.8189	0.8923	0.3506	0.3452	6.1250	6.1250	6.3264	2.5109	2.5109	2.5361
250,10,1,1/$\sqrt{p}$	0.9434	0.9594	1.0000	1.0000	1.0000	1.0000	1.0000	0.9546	0.9338	0.9628
len	0.8208	1.2706	0.4983	0.4917	6.0051	6.0051	6.2162	2.5178	2.5178	2.6859
250,10,1,0.9	0.9438	0.7414	1.0000	1.0000	1.0000	1.0000	1.0000	0.9844	0.9844	0.9894
len	0.8188	5.0029	4.7703	5.3780	5.3780	5.6168	5.6168	2.8269	2.8269	3.0155
250,10,8,0.9	0.9464	0.7068	0.6962	0.9996	0.9996	0.9620	0.9996	1.0000	0.9998	1.0000
len	0.8264	4.7997	4.7404	3.9512	1.5842	1.5842	1.7615	5.6604	5.6604	5.9544
300,6,1,0	0.9494	0.9428	1.0000	1.0000	1.0000	0.9998	1.0000	0.9628	0.9638	0.9654
len	0.7523	0.7902	0.2986	0.2969	4.7093	4.7093	4.9580	2.4926	2.4926	2.5019
300,6,4,0	0.9488	0.9426	0.9360	1.0000	1.0000	0.9844	1.0000	0.9780	0.9728	0.9776
len	0.2300	0.2303	0.2305	0.2305	0.2304	0.0929	1.3467	1.3467	1.4834	3.3391
300,6,1,1/$\sqrt{p}$	0.9510	0.9690	1.0000	1.0000	0.9996	0.9998	0.9996	0.9708	0.9562	0.9782
len	0.7534	1.2346	0.5082	0.5040	4.6388	4.6388	4.8629	2.5457	2.5457	2.7269
300,6,4,1/$\sqrt{p}$	0.9428	0.9380	0.9444	0.9998	0.9998	0.9770	0.9998	0.9596	0.9306	0.9544
len	0.2300	0.3473	0.3473	0.3471	0.3469	0.1406	1.3240	1.3240	1.4625	3.4319
300,6,1,0.9	0.9478	0.8594	0.9996	0.9998	1.0000	0.9998	1.0000	0.9784	0.9764	0.9826
len	0.7515	4.1236	3.8798	3.8860	4.5491	4.5491	4.7538	2.7358	2.7358	2.8789
300,6,4,0.9	0.9464	0.7902	0.7826	0.9998	0.9998	0.9066	1.0000	0.9990	0.9976	0.9996
len	0.7500	3.8525	3.8329	3.4621	1.8744	1.8744	2.1509	4.2067	4.2067	4.4727
400,8,1,0	0.9526	0.9480	1.0000	1.0000	1.0000	1.0000	1.0000	0.9598	0.9612	0.9618
len	0.6511	0.6697	0.2387	0.2357	5.7319	5.7319	5.9392	2.4782	2.4782	2.4852
400,8,6,0	0.9492	0.9454	0.9438	1.0000	0.9998	0.9864	1.0000	0.9762	0.9758	0.9770
len	0.6546	0.6789	0.6791	0.2346	1.1894	1.1894	1.3040	3.9991	3.9991	4.0171

Table 4.15. Bootstrapping OLS Forward Selection with BIC Type 5(cont.)

400,8,1,1/$\sqrt{p}$	0.9516	0.9696	1.0000	1.0000	1.0000	1.0000	1.0000	0.9716	0.9676	0.9798
len	0.6530	1.0319	0.3563	0.3519	5.6889	5.6889	5.8853	2.5461	2.5461	2.7269
400,8,6,1/$\sqrt{p}$	0.9478	0.9398	0.9364	1.0000	1.0000	0.9836	1.0000	0.9796	0.9608	0.9748
len	0.6591	1.1202	1.1170	0.3938	1.3080	1.3080	1.4506	4.2212	4.2212	4.3306
400,8,1,0.9	0.9540	0.8134	0.9996	1.0000	0.9998	0.9998	0.9998	0.9794	0.9782	0.9818
len	0.6503	3.6476	3.3059	3.3015	5.2450	5.2450	5.4707	2.7237	2.7237	2.8710
400,8,6,0.9	0.9488	0.7232	0.7294	1.0000	1.0000	0.9610	1.0000	0.9996	0.9992	0.9666
len	0.6561	3.9995	4.0096	2.9919	1.4677	1.4677	1.6283	5.1005	5.1005	5.4079

CHAPTER 5

CONCLUSIONS

There is massive literature on variable selection and a fairly large literature for inference after variable selection. See references in Pelawa Watagoda and Olive (2018).

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.

The simulations were done in *R*. See R Core Team (2016). We used several *R* functions including forward selection as computed with the `regsubsets` function from the `leaps` library. The collection of Olive (2018b) *R* functions *slpack*, available from (http://lagrange.math.siu.edu/Olive/slpack.txt), has some useful functions for the inference.

The tables were made with `bicbootsim`. There was occasionally undercoverage, especially for the hybrid region and ψ=0.9.

REFERENCES

Akaike, H. (1973), "Information Theory as an Extension of the Maximum Likelihood Principle," in *Proceedings, 2nd International Symposium on Information Theory*, eds. Petrov, B.N., and Csakim, F., Akademiai Kiado, Budapest, 267-281.

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.

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.

Freedman, D.A. (1981), "Bootstrapping Regression Models," *The Annals of Statistics*, 9, 1218-1228.

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

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

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

Hebbler, B. (1847), "Statistics of Prussia," *Journal of the Royal Statistical Society, A*, 10, 154-186.

Knight, K., and Fu, W.J. (2000), "Asymptotics for Lasso-Type Estimators," *The Annals of Statistics*, 28, 1356–1378.

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

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. (2017a), *Linear Regression*, Springer, New York, NY.

Olive, D.J. (2017b), *Robust Multivariate Analysis*, Springer, New York, NY.

Olive, D.J. (2018a), "Applications of Hyperellipsoidal Prediction Regions," *Statistical Papers*, to appear, see (http://lagrange.math.siu.edu/Olive/pphpr.pdf).

Olive, D.J. (2018b), *Prediction and Statistical Learning*, online course notes, see (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. (2018), "Inference for Multiple Linear Regression After Model or Variable Selection," preprint at (http://lagrange.math.siu.edu/Olive /ppvsinf.pdf).

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

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics,* 6, 461-464.

Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics,* Wiley, New York, NY.

Su, Z., and Cook, R.D. (2012), "Inner Envelopes: Efficient Estimation in Multivariate Linear Regression," *Biometrika*, 99, 687-702.

VITA

Graduate School
Southern Illinois University

Charles Murphy

murphington11@gmail.com

Southern Illinois University
Bachelor of Science, Mathematics, December 2016

Research Paper Title:
Bootstrapping Forward Selection with BIC

Major Professor: Dr. David J. Olive