2017

Prediction Interval After Forward Selection Using EBIC

Mulubrhan Haile
murer563@yahoo.com

PREDICTION INTERVALS AFTER FORWARD SELECTION USING EBIC

by

Mulubrhan Haile

B.S., University of Asmara, 2005

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, 2017

RESEARCH PAPER APPROVAL

PREDICTION INTERVALS AFTER FORWARD SELECTION USING EBIC

by

Mulubrhan Haile

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

Bhaskar Bhattacharya

Kwangho Choiy

Graduate School
Southern Illinois University Carbondale
March 31, 2017

AN ABSTRACT OF THE RESEARCH PAPER OF

MULUBRHAN HAILE, for the Master of Science degree in MATHEMATICS,
presented on MARCH 31, 2017, at Southern Illinois University Carbondale.

TITLE: PREDICTION INTERVALS AFTER FORWARD SELECTION USING EBIC

MAJOR PROFESSOR: Dr. David J. Olive

This paper presents a prediction interval for the multiple linear regression model $Y = \beta_1 x_1 + \cdots + \beta_p x_p + e$ after forward selection, where the model is selected using the EBIC criterion.

KEY WORDS: Forward Selection; Prediction Interval; Relaxed Lasso.

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. David Olive for his invaluable assistance and insights leading to the writing of this paper. He was always available to answer all my questions. He encouraged me at every moment. Dr. David Olive, it was a pleasure to have you an advisor. My sincere thanks also goes to Dr. S. Yaser Samadi who helped me improve my statistical skills through the courses I had with him. I would like to thank Professors of my committee for their helpful comments and suggestions. I would also like to thank all Mathematics Department faculty and staff. You were so helpful to me.

TABLE OF CONTENTS

LIST OF TABLES

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 $\boldsymbol{x}_i^T = (x_{i,1}, ..., x_{i,p}) = (1 \quad \boldsymbol{u}_i^T)$ 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 the random variable e_i is the ith error. 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 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. 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). 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 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 $Q_2(I) = SSE(I) = RSS(I) = \boldsymbol{r}^T(I)\boldsymbol{r}(I) = \sum_{i=1}^n r_i^2(I) = \sum_{i=1}^n (Y_i - \hat{Y}_i(I))^2$, where RSS stands for residual sum of squares and SSE stands for sum of squared errors. 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 $j = 2, ..., M$. Often $M = \min(\lceil n/J \rceil, p)$ for some integer J such as $J = 5, 10$, or 20. Here $\lceil x \rceil$ is the smallest integer $\geq x$, e.g., $\lceil 7.7 \rceil = 8$.

When there is a sequence of M submodels, the final submodel I_d needs to be selected. Let $\boldsymbol{x}_I$ and $\hat{\boldsymbol{\beta}}_I$ be an $a \times 1$ vector. Hence the candidate model contains a terms, including a constant. Suppose the e_i are independent and identically distributed (iid) with variance $V(e_i) = \sigma^2$. Then there are many criteria used to select the final submodel I_d. 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 full OLS 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).

It is hard to get a good estimator of σ^2 when n/p is not large. The following criterion are described in Burnham and Anderson (2004), but still need n/p large.

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

$$AIC_C(I) = n \log\left(\frac{SSE(I)}{n}\right) + 2\frac{a(a+1)}{n-a-1},$$

and

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

Let I_{min} be the submodel that minimizes the criterion. 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 $n \to \infty$. If $\hat{\boldsymbol{\beta}}_I$ is an $a \times 1$ vector, form the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I,0}$ from $\hat{\boldsymbol{\beta}}_I$ by adding 0's corresponding to the omitted variables. Since there are a finite number of regression models I that 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).

An interesting BIC-type criterion is given in Luo and Chen (2012) that may work when n/p is not large. Let $0 \le \gamma \le 1$ and $|I| = a \le \min(n,p)$ if $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$. We may use $a \le \min(n/5, p)$. Then

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

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

Consider predicting a future test response variable Y_f given a $p \times 1$ vector of predictors $\boldsymbol{x}_f$ and training data $(\boldsymbol{x}_1, Y_1), ..., (\boldsymbol{x}_n, Y_n)$. A large sample $100(1-\delta)\%$ prediction interval (PI) has the form $[\hat{L}_n, \hat{U}_n]$, where $P(\hat{L}_n \le Y_f \le \hat{U}_n) \to 1 - \delta$ 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

$$\text{shorth(c)} = [\text{Z}_{(\text{s})}, \text{Z}_{(\text{s}+\text{c}-1)}]. \tag{1.5}$$

Let

$$k_n = \lceil n(1 - \delta) \rceil. \tag{1.6}$$

Frey (2013) showed that for large $n\delta$ and identically independent distributed (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

$$c = \min(n, \lceil n[1 - \delta + 1.12\sqrt{\delta/n}\ ]\ \rceil). \tag{1.7}$$

A problem with the prediction intervals that cover $\approx 100(1-\delta)\%$ of the training data cases Y_i (such as the shorth(k_n) PI), is that they have coverage lower than the nominal coverage of $1-\delta$ for moderate n. This result is not surprising since empirically statistical methods perform worse on test data. Increasing c will improve the coverage for moderate samples.

Example 1. (Example 5.3 from Olive (2017b).) 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]
```

Olive (2007) developed prediction intervals for the full MLR model. Olive (2013) developed prediction intervals for models of the form $Y_i = m(\boldsymbol{x}_i) + e_i$, and variable selection models for (1.1) have this form, as noted by Olive (2017a). Both these PIs need n/p large. Let c be given by (2.2) with d replaced by p, and let

$$b_n = \left(1 + \frac{15}{n}\right) \sqrt{\frac{n+2p}{n-p}}. \tag{1.8}$$

Compute the shorth(c) of the residuals $= [r_{(s)}, r_{(s+c-1)}] = [\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2}]$ where the ith residual $r_i = Y_i - \hat{Y}_i = Y_i - \hat{m}(\boldsymbol{x}_i)$. Then a 100 $(1-\delta)$% large sample PI for Y_f is

$$[\hat{m}(\boldsymbol{x}_f) + b_n \tilde{\xi}_{\delta_1}, \hat{m}(\boldsymbol{x}_f) + b_n \tilde{\xi}_{1-\delta_2}]. \tag{1.9}$$

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 2

PREDICTION INTERVALS AFTER FORWARD SELECTION

If n/p is large, the PI (1.9) can be used for the variable selection estimators with $\hat{m}(\boldsymbol{x}) = \boldsymbol{x}_{I_d}^T \hat{\boldsymbol{\beta}}_{I_d}$, where I_d denotes the index of predictors selected from the variable selection method. Hence $I_d = I_{min}$ is the model that minimizes C_p for forward selection. Now we want to minimize EBIC for forward selection, where n/p is not necessarily large.

PI (1.9) needs the shorth of the residuals to be a consistent estimator of the population shorth of the error distribution. Olive and Hawkins (2003) show that if the $\|\boldsymbol{x}_i\|$ are bounded and $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$, then $\max_{i=1,...,n} |r_i - e_i| \xrightarrow{P} 0$ and the sample quantiles of the residuals estimate the population quantiles of the error distribution. For OLS, each submodel I produces a $\sqrt{n}$ consistent estimator provided that $S \subseteq I$.

The Cauchy Schwartz inequality says $|\boldsymbol{a}^T \boldsymbol{b}| \leq \|\boldsymbol{a}\| \ \ \|\boldsymbol{b}\|$. Suppose $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = O_P(1)$ is bounded in probability. This will occur if $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, e.g. if $\hat{\boldsymbol{\beta}}$ is the OLS estimator. Then

$$|r_i - e_i| = |Y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}} - (Y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})| = |\boldsymbol{x}_i^T (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})|.$$

Hence

$$\sqrt{n} \max_{i=1,...,n} |r_i - e_i| \leq (\max_{i=1,...,n} \|\boldsymbol{x}_i\|) \ \ \|\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\| = O_P(1)$$

since $\max \|\boldsymbol{x}_i\| = O_P(1)$ or there is extrapolation. Hence OLS residuals behave well if the zero mean error distribution of the iid e_i has a finite variance σ^2.

Let d be a crude estimate of the model degrees of freedom. For forward selection with OLS, $\hat{\boldsymbol{\beta}}_{I_d}$ is a $d \times 1$ vector. For example, use $I_d = I_{min}$ where d is the number of nonzero coefficients, including a constant, in the submodel I_{min} that minimized a criterion such as EBIC.

The Olive (2017d) and Pelawa Watagoda and Olive (2017) PI that can work if $n >> p$ or $p > n$ is defined below. The PI is similar to the Olive (2013) PI with p replaced by d, but some care needs to be taken to that the PI is well defined and does not have infinite length. Let

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

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

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

$$c = \lceil n q_n \rceil, \tag{2.2}$$

and let

$$b_n = \left(1 + \frac{15}{n}\right) \sqrt{\frac{n + 2d}{n - d}}, \tag{2.3}$$

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

$$b_n = 5 \left(1 + \frac{15}{n}\right),$$

otherwise. Compute the shorth(c) of the residuals $= [r_{(s)}, r_{(s+c-1)}] = [\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2}]$. Then a 100 $(1 - \delta)$% large sample PI for Y_f is

$$[\hat{m}(\boldsymbol{x}_f) + b_n \tilde{\xi}_{\delta_1}, \hat{m}(\boldsymbol{x}_f) + b_n \tilde{\xi}_{1-\delta_2}]. \tag{2.4}$$

CHAPTER 3

EXAMPLES AND SIMULATIONS

Let $\boldsymbol{x} = (1 \ \boldsymbol{u}^T)^T$ where $\boldsymbol{u}$ is the $(p-1)\times 1$ vector of nontrivial predictors. For 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 $\boldsymbol{u} = \boldsymbol{A}\boldsymbol{w}_i$ so that $Cov(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$, 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 $(1, ..., 1)^T$. Then $Y_i = 1 + 1x_{i,2} + \cdots + 1x_{i,k} + e_i$ 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 of five types: i) N(0,1) errors, ii) t_3 errors, iii) EXP(1) - 1 errors, iv) uniform$(-1, 1)$ errors, and v) 0.9 N(0,1) + 0.1 N(0,100) errors.

The lengths of the asymptotically optimal 95% PIs are i) 3.92 = 2(1.96), ii) 6.365, iii) 2.996, iv) 1.90 = 2(0.95), and v) 13.490. Suppose that the simulation uses K runs and $W_i = 1$ if Y_f is in the ith PI, and $W_i = 0$ otherwise, for $i = 1, ..., K$. Then the W_i are iid binomial(1,$1 - \delta_n$) where $\rho_n = 1 - \delta_n$ is the true coverage of the PI when the sample size is n. Let $\hat{\rho}_n = \overline{W}$. Since $\sum_{i=1}^K W_i \sim$ binomial(K, ρ_n), the standard error $SE(\overline{W}) = \sqrt{\rho_n(1-\rho_n)/K}$. For $K = 5000$ and ρ_n near 0.9, we have $3SE(\overline{W}) \approx 0.01$. Hence an observed coverage of $\hat{\rho}_n$ within 0.01 of the nominal coverage $1 - \delta$ suggests that there is no reason to doubt that the nominal PI coverage is different from the observed coverage. So for a large sample 95% PI, we want the observed coverage to be between 0.94 and 0.96. Also a difference of 0.01 is not large. Coverage slightly higher than the nominal coverage is better than coverage slightly lower than the nominal coverage.

The forward selection used 2, 3, ..., $M = \min(\lceil n/J \rceil, p)$ variables in the MLR model, including a constant, with $J = 5$.

The simulation used 5000 runs with $p = 20, 40, n$ and $2n$. The simulation used $\psi = 0, 1/\sqrt{p}$, and 0.9, so an observed coverage in [0.94, 0.96] gives no reason to doubt that the PI has the nominal coverage of 0.95. The simulation used $k = 1, 19$, and $p - 1$.

Some R code is below. For 5000 runs of the nominal large sample 95% PI, the observed coverage was 0.963, the average length was 4.441, and variable selection on average used 2.1 variables, including a constant. We would like this number, recorded as dave, to be near but slightly larger than $k + 1$ when n/k is large.

```
library(leaps)

out<-evspisim(n=100,p=20,k=1,nruns=5000,psi=0,type=1)

out

$fselpicov

[1] 0.963

$fselpimenlen

[1] 4.441144

mean(out$dd)+1

[1]  2.0968
```

CHAPTER 4

SIMULATIONS FOR FIVE ERROR TYPES

Table 4.1. R-output for different values of n, p, k and ψ for error type-1

n	p	k	ψ	cov	len	dave
100	20	1	0	0.963	4.441	2.097
100	20	19	0	0.979	5.705	20.000
100	20	19	0.9	0.955	5.170	7.187
100	40	1	0	0.967	4.434	2.095
100	100	1	0	0.963	4.425	2.094
100	100	1	0.9	0.955	4.352	2.149
100	100	99	0	0.941	40.564	3.454
100	200	1	0	0.966	4.430	2.092
400	20	1	0	0.949	4.006	2.040
400	20	19	0	0.976	4.695	20.000
400	20	19	0.9	0.961	4.444	13.229
400	40	1	0	0.951	4.006	2.042

Table 4.2. R-output for different values of n, p, k and ψ for error type-1

n	p	k	ψ	cov	len	dave
400	40	39	0	0.975	4.900	40.000
400	400	1	0	0.956	4.009	2.028
400	400	1	0.05	0.958	4.008	2.023
400	400	399	0	0.946	78.458	2.292
400	800	1	0	0.954	4.007	2.027
800	20	1	0	0.953	3.947	2.024
800	20	1	0.9	0.953	3.945	2.013
800	20	1	0.224	0.954	3.946	2.023
800	20	19	0	0.964	4.251	20.000
800	40	1	0	0.952	3.946	2.025
800	40	1	0.9	0.950	3.943	2.009
800	40	39	0	0.979	4.673	40.000
800	800	1	0.035	0.949	3.949	2.014
800	800	19	0	0.965	4.250	20.185
800	800	799	0	0.946	110.364	2.179

Table 4.3. R-output for different values of n, p, k and ψ for error type-1

n	p	k	ψ	cov	len	dave
1000	20	1	0	0.953	3.937	2.023
1000	20	1	0.9	0.951	3.937	2.007
1000	20	19	0	0.963	4.177	20.000
1000	40	19	0	0.959	4.177	20.217
1000	40	1	0.9	0.952	3.935	2.007
1000	1000	1	0	0.952	3.937	2.019
1000	1000	999	0.9	0.750	15.787	198.991
2000	20	1	0	0.952	3.909	2.017
2000	20	1	0.9	0.951	3.909	2.007
2000	20	1	0.224	0.951	3.909	2.015
2000	20	19	0	0.956	4.033	20.000
2000	20	19	0.9	0.956	4.033	19.991
2000	40	19	0	0.957	4.033	20.130
2000	40	39	0	0.964	4.171	40.000
2000	40	39	0.224	0.964	4.171	40.000

Table 4.4. R-output for different values of n, p, k and ψ for error type-2

n	p	k	ψ	cov	len	dave
100	20	1	0	0.955	7.244	2.100
100	20	19	0	0.974	10.013	19.925
100	20	19	0.9	0.958	8.312	4.790
100	40	1	0	0.953	7.232	2.084
100	100	1	0	0.956	7.207	2.094
100	100	1	0.9	0.953	7.151	2.278
100	100	99	0	0.933	41.069	3.302
100	200	1	0	0.954	7.238	2.094
400	20	1	0	0.950	6.463	2.034
400	20	19	0	0.973	8.445	19.990
400	20	19	0.9	0.953	7.018	7.939
400	40	1	0	0.951	6.475	2.035

Table 4.5. R-output for different values of n, p, k and ψ for error type-2

n	p	k	ψ	cov	len	dave
400	40	39	0	0.976	8.751	39.986
400	400	1	0	0.948	6.462	2.030
400	400	1	0.05	0.949	6.462	2.027
400	400	399	0	0.947	78.618	2.291
400	800	1	0	0.948	6.453	2.029
800	20	1	0	0.942	6.366	2.024
800	20	1	0.9	0.941	6.358	2.012
800	20	1	0.224	0.942	6.367	2.021
800	20	19	0	0.953	7.190	19.994
800	40	1	0	0.945	6.368	2.025
800	40	1	0.9	0.943	6.356	2.011
800	40	39	0	0.971	8.464	39.993
800	800	1	0.035	0.951	6.370	2.017
800	800	19	0	0.963	7.186	20.187
800	800	799	0	0.947	110.480	2.169

Table 4.6. R-output for different values of n, p, k and ψ for error type-2

n	p	k	ψ	cov	len	dave
1000	20	1	0	0.951	6.349	2.024
1000	20	1	0.9	0.948	6.343	2.011
1000	20	19	0	0.963	6.982	19.996
1000	40	19	0	0.955	7.000	20.203
1000	40	1	0.9	0.946	6.348	2.009
1000	1000	1	0	0.951	6.355	2.016
1000	1000	999	0.9	0.760	16.768	193.686
2000	20	1	0	0.953	6.320	2.014
2000	20	1	0.9	0.954	6.319	2.009
2000	20	1	0.224	0.953	6.320	2.015
2000	20	19	0	0.958	6.646	20.000
2000	20	19	0.9	0.957	6.591	16.053
2000	40	19	0	0.955	6.636	20.129
2000	40	39	0	0.960	7.005	40.000
2000	40	39	0.224	0.960	7.006	39.998

Table 4.7. R-output for different values of n, p, k and ψ for error type-3

n	p	k	ψ	cov	len	dave
100	20	1	0	0.961	3.782	2.093
100	20	19	0	0.977	5.652	20.000
100	20	19	0.9	0.958	5.212	7.334
100	40	1	0	0.964	3.773	2.097
100	100	1	0	0.962	3.771	2.086
100	100	1	0.9	0.956	3.848	2.139
100	100	99	0	0.936	40.610	3.433
100	200	1	0	0.966	3.792	2.089
400	20	1	0	0.949	3.206	2.037
400	20	19	0	0.972	4.321	20.000
400	20	19	0.9	0.958	4.164	13.419
400	40	1	0	0.958	3.218	2.036

Table 4.8. R-output for different values of n, p, k and ψ for error type-3

n	p	k	ψ	cov	len	dave
400	40	39	0	0.979	4.671	40.000
400	400	1	0	0.955	3.217	2.032
400	400	1	0.05	0.956	3.215	2.024
400	400	399	0	0.944	78.414	2.294
400	800	1	0	0.955	3.214	2.028
800	20	1	0	0.952	3.121	2.024
800	20	1	0.9	0.952	3.155	2.011
800	20	1	0.224	0.952	3.120	2.025
800	20	19	0	0.961	3.681	20.000
800	40	1	0	0.953	3.119	2.021
800	40	1	0.9	0.952	3.168	2.011
800	40	39	0	0.973	4.315	40.000
800	800	1	0.035	0.950	3.119	2.017
800	800	19	0	0.963	3.694	20.195
800	800	799	0	0.942	110.359	2.201

Table 4.9. R-output for different values of n, p, k and ψ for error type-3

n	p	k	ψ	cov	len	dave
1000	20	1	0	0.952	3.101	2.023
1000	20	1	0.9	0.951	3.122	2.011
1000	20	19	0	0.960	3.563	20.000
1000	40	19	0	0.965	3.567	20.204
1000	40	1	0.9	0.956	3.129	2.008
1000	1000	1	0	0.950	3.099	2.016
1000	1000	999	0.9	0.748	15.801	198.984
2000	20	1	0	0.951	3.047	2.015
2000	20	1	0.9	0.951	3.048	2.008
2000	20	1	0.224	0.950	3.047	2.015
2000	20	19	0	0.954	3.323	20.000
2000	20	19	0.9	0.954	3.323	19.989
2000	40	19	0	0.956	3.330	20.135
2000	40	39	0	0.961	3.557	40.000
2000	40	39	0.224	0.961	3.557	40.000

Table 4.10. R-output for different values of n, p, k and ψ for error type-4

n	p	k	ψ	cov	len	dave
100	20	1	0	0.992	2.208	2.098
100	20	19	0	0.993	2.962	20.000
100	20	19	0.9	0.969	2.927	13.525
100	40	1	0	0.992	2.206	2.091
100	100	1	0	0.990	2.206	2.086
100	100	1	0.9	0.977	2.225	2.046
100	100	99	0	0.936	40.314	3.529
100	200	1	0	0.991	2.203	2.090
400	20	1	0	0.967	1.963	2.039
400	20	19	0	0.987	2.223	20.000
400	20	19	0.9	0.986	2.223	19.986
400	40	1	0	0.973	1.964	2.041

Table 4.11. R-output for different values of n, p, k and ψ for error type-4

n	p	k	ψ	cov	len	dave
400	40	39	0	0.980	2.411	40.000
400	400	1	0	0.966	1.963	2.033
400	400	1	0.05	0.967	1.963	2.024
400	400	399	0	0.940	78.376	2.277
400	800	1	0	0.966	1.962	2.023
800	20	1	0	0.957	1.926	2.021
800	20	1	0.9	0.960	1.926	2.017
800	20	1	0.224	0.960	1.926	2.020
800	20	19	0	0.972	2.038	20.000
800	40	1	0	0.957	1.926	2.027
800	40	1	0.9	0.959	1.926	2.014
800	40	39	0	0.981	2.200	40.000
800	800	1	0.035	0.956	1.925	2.016
800	800	19	0	0.970	2.042	20.191
800	800	799	0	0.945	110.424	2.170

Table 4.12. R-output for different values of n, p, k and ψ for error type-4

n	p	k	ψ	cov	len	dave
1000	20	1	0	0.959	1.919	2.025
1000	20	1	0.9	0.961	1.919	2.017
1000	20	19	0	0.973	2.006	20.000
1000	40	19	0	0.967	2.009	20.215
1000	40	1	0.9	0.961	1.919	2.017
1000	1000	1	0	0.964	1.919	2.018
1000	1000	999	0.9	0.741	15.542	199.511
2000	20	1	0	0.951	1.905	2.015
2000	20	1	0.9	0.950	1.905	2.012
2000	20	1	0.224	0.949	1.905	2.013
2000	20	19	0	0.956	1.945	20.000
2000	20	19	0.9	0.956	1.945	20.000
2000	40	19	0	0.962	1.945	20.128
2000	40	39	0	0.969	1.997	40.000
2000	40	39	0.224	0.969	1.997	40.000

Table 4.13. R-output for different values of n, p, k and ψ for error type-5

n	p	k	ψ	cov	len	dave
100	20	1	0	0.945	13.684	2.066
100	20	19	0	0.966	21.821	18.118
100	20	19	0.9	0.951	15.818	3.089
100	40	1	0	0.946	13.647	2.056
100	100	1	0	0.941	13.583	2.056
100	100	1	0.9	0.945	14.267	2.418
100	100	99	0	0.942	43.213	3.012
100	200	1	0	0.942	13.511	2.046
400	20	1	0	0.947	12.447	2.033
400	20	19	0	0.968	21.140	20.000
400	20	19	0.9	0.948	13.385	4.522
400	40	1	0	0.942	12.593	2.033

Table 4.14. R-output for different values of n, p, k and ψ for error type-5

n	p	k	ψ	cov	len	dave
400	40	39	0	0.968	21.737	40.000
400	400	1	0	0.943	12.565	2.031
400	400	1	0.05	0.943	12.563	2.025
400	400	399	0	0.943	79.409	2.254
400	800	1	0	0.946	12.558	2.028
800	20	1	0	0.947	12.617	2.024
800	20	1	0.9	0.947	12.593	2.024
800	20	1	0.224	0.948	12.617	2.026
800	20	19	0	0.959	16.562	20.000
800	40	1	0	0.946	12.648	2.026
800	40	1	0.9	0.945	12.623	2.037
800	40	39	0	0.971	22.071	40.000
800	800	1	0.035	0.949	12.620	2.017
800	800	19	0	0.962	16.559	20.196
800	800	799	0	0.945	111.196	2.163

Table 4.15. R-output for different values of n, p, k and ψ for error type-5

n	p	k	ψ	cov	len	dave
1000	20	1	0	0.949	12.684	2.023
1000	20	1	0.9	0.949	12.661	2.019
1000	20	19	0	0.959	15.780	20.000
1000	40	19	0	0.956	15.838	20.205
1000	40	1	0.9	0.947	12.676	2.027
1000	1000	1	0	0.947	12.682	2.019
1000	1000	999	0.9	0.818	22.299	153.057
2000	20	1	0	0.947	12.709	2.015
2000	20	1	0.9	0.947	12.695	2.008
2000	20	1	0.224	0.947	12.709	2.014
2000	20	19	0	0.952	14.380	20.000
2000	20	19	0.9	0.950	13.322	8.317
2000	40	19	0	0.953	14.384	20.128
2000	40	39	0	0.958	16.173	40.000
2000	40	39	0.224	0.958	16.173	40.000

CHAPTER 5

CONCLUSIONS

Several methods of prediction intervals after variable or model selection are considered for (1.1) by Olive (2017d), Pelawa Watagoda (2017) and Pelawa Watagoda and Olive (2017). Prediction intervals are also used in Olive (2017ac). EBIC could also be used for relaxed lasso Meinshausen (2007), which is OLS applied to the predictors that have nonzero lasso coefficients, including a constant.

The simulations were done in *R*. See R Core Team (2016). The collection of *R* functions *slpack*, available from (http://lagrange.math.siu.edu/Olive/slpack.txt), has some useful functions for the inference. The function `evspisim` was used to do the simulation.

The following points can be observed from the simulation tables.

1. When ψ=0.9 and $k > 1$, dave is sometimes too low, especially if $n/p \leq 20$.
2. The simulations took longer when n and p are large.
3. The dave, cov and len outputs were bad when we have k=p-1 and p is very large.
4. As the sample size increases the coverage is fairly close to 0.95.

REFERENCES

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

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

[3] Luo, S., and Chen, Z. (2013), "Extended BIC for Linear Regression Models With Diverging Number of Relevant Features and High or Ultra-High Feature Spaces," *Journal of Statistical Planning and Inference,* 143, 494-504.

[4] Meinshausen, N. (2007), "Relaxed Lasso," *Computational Statistics & Data Analysis*, 52, 374-393.

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

[6] Olive, D.J. (2007), "Prediction Intervals for Regression," *Computational Statistics & Data Analysis,* 51, 3115-3122.

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

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

[9] Olive, D.J. (2017b), *Robust Multivariate Analysis*, Springer, New York, NY, to appear.

[10] Olive, D.J. (2017c), *Linear Regression*, Springer, New York, NY, to appear.

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

[12] Olive, D.J., and Hawkins, D.M. (2003), "Robust Regression with High Coverage," *Statistics & Probability Letters,* 63, 259-266.

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

[14] Pelawa Watagoda, L.C.R. (2017), *Inference After Variable Selection*, Ph.D. Thesis, Southern Illinois University, to appear. See an early version at (http://lagrange.math.siu.edu/Olive/slasanthiphd.pdf).

[15] 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).

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

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

[18] 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

Mulubrhan Haile

gmulubrhan@gmail.com

University of Asmara
Bachelor of Science, Mathematics, July 2005

Major Professor: Dr. David J. Olive