Research Papers Graduate School

Spring 5-11-2017

Prediction Intervals After Forward Selection Using d Variables

Kosman W G D H Rajapaksha
drhansana@siu.edu

Recommended Citation

Rajapaksha, Kosman W. "Prediction Intervals After Forward Selection Using d Variables." (Spring 2017).

PREDICTION INTERVALS AFTER FORWARD SELECTION USING D VARIABLES

by

Kosman Rajapaksha

M.Sc. in Applied Statistics, University of Peradeniya, 2014

B.Sc. Faculty of Science, University of Peradeniya, 2011

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
August, 2017

RESEARCH PAPER APPROVAL

PREDICTION INTERVALS AFTER FORWARD SELECTION USING D VARIABLES

By

Kosman Rajapaksha

A Research Paper Submitted in Partial

Fulfillment of the Requirements

for the Degree of

Master of Sciences

in the field of Mathematics

Approved by:

David Olive, Chair

Bhaskar Bhattacharya

Kwangho Choiy

Graduate School
Southern Illinois University Carbondale
May 11, 2017

AN ABSTRACT OF THE RESEARCH PAPER OF

KOSMAN RAJAPAKSHA, for the Master of Science degree in MATHEMATICS, presented on MAY 11, 2017, at Southern Illinois University Carbondale.

TITLE: PREDICTION INTERVALS AFTER FORWARD SELECTION USING D VARIABLES

MAJOR PROFESSOR: Dr. David Olive

This paper presets a prediction interval for the multiple linear regression model $Y = \beta_1 x_1 + ... + \beta_p x_p + e$ after forward selection, where the model is selected using $d = min(\lceil n/J \rceil, p)$ variables for some positive integer J such as 5, 10, 20, 50, and $\lceil n/p \rceil$.

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

ACKNOWLEDGMENTS

I would first like to express my gratitude to my thesis adviser Prof. D. Olive at Southern Illinois University. Prof. D. Olive was helpful throughout my Master Research Paper. He answered all the questions I had related to this paper. Also, I would like to thank Prof. Samadhi, Prof. Bhattacharya and Prof. Choiy at SIU because they were important impact to develop my Statistical and Mathematical knowledge by several courses I had with them. I would like to thank Dr. Kanthi Perera who did encourage me to get into the Master program. Finally, I must thank my parents and my wife who gave me support and encourage me to do my work.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

The response variable is the variable that you want to predict. The predictor variables are the variables used to predict the response variable. The response variable will be denoted by Y and the p predictor variables will be denoted by $x_1, ..., x_p$ and collected in a vector $\boldsymbol{x}$. Then $\boldsymbol{x}^T$ is the transpose of $\boldsymbol{x}$.

Suppose that the response variable Y and at least one predictor variable x_i are quantitative. Then the multiple linear regression (MLR) model is

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{1.1}$$

for $i = 1, \ldots, n$. Here n is the *sample size* and the random variable e_i is the ith *error*. Suppressing the subscript i, the model is $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$. A constant will be in the model, so $x_{i,1} \equiv 1$ is sometimes called the trivial predictor. 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.

Variable selection, also called subset or model selection, is the search for a subset of predictor variables that can be deleted without important loss of information. 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

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

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, $\boldsymbol{\beta}_O = \mathbf{0}$ if $S \subseteq I$.

Many methods for variable selection have been suggested. We will consider forward selection as computed with the R function `regsubsets` function from the `leaps` library.

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 $Q_2(I) = SSE(I) = RSS(I) = \mathbf{r}^T(I)\mathbf{r}(I) = \sum_{i=1}^n r_i^2(I) = \sum_{i=1}^n (Y_i - \hat{Y}_i(I))^2$ 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 $Q_j(I) = \mathbf{r}^T(I)\mathbf{r}(I) = \sum_{i=1}^n r_i^2(I) = \sum_{i=1}^n (Y_i - \hat{Y}_i(I))^2$. 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 $a \times 1$. 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 σ^2 is the 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 $\hat{\sigma}^2$ 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 describe 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

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

Let I_{min} be the submodel that minimize the criterion. Following Seber and Lee(2003, p. 448) and Nishi(1984), the probability that model I_{min} from C_p or AIC under fit 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 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 in not large. Let $0 \le \gamma \le 1$ and $|I| = a \le min(n,q)$ if $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$. We may use $a \le min(n/5, p)$. Then

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

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

A simple method is to take the model that uses $d = M = min(\lceil n/J \rceil, p)$. This method that we will investigate. If p is fixed, the method will use the full model once

$n/J \geq p$. Hence the PI (2.4) described below will be asymptotically optimal for a wide class of zero mean error distributions.

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 $P(\hat{L}_n < Y_f < \hat{U}_n) \to 1 - \delta$ 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

$$\mathrm{shorth(c)} = [\mathrm{Z_{(s)}}, \mathrm{Z_{(s+c-1)}}]. \tag{1.5}$$

Let

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

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

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

A problem with the prediction intervals that cover $\approx 100(1-\delta)\%$ of the training data cases Y_i (such as (1.5) using $c = k_n$ given by (1.6)), 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 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
```

```
ordered 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 model 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), 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 residual $= [r_{(s)}, r_{(s+c-1)}] = [\hat{\xi}_{\delta_1}, \hat{\xi}_{1-\delta_2}]$ 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

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

Note that the correlation 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 VARIABLE SELECTION

If n/p is large, the PI (1.9) can be used for the variable selection estimators with $\hat{m}(\boldsymbol{x}) = \boldsymbol{x}_{I_d}^T \hat{\beta}_{I_d}$, where I_d denotes the index of predictors selected from the variable selection method. For example, $I_d = I_{min}$ is the model that minimizes C_p for forward selection. Now we want I_d to used $d = M = min(\lceil n/J \rceil, p)$ variables 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 $|\mathbf{a^T b}| \leq \|\mathbf{a}\|\|\mathbf{b}\|$. Suppose $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = O_p(1)$ is bounded in probability. This will occur if $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(0, \Sigma)$, e.g. if $\hat{\boldsymbol{\beta}}$ is the OLS estimator. Then

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

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

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

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

$$c = \lceil nq_n \rceil, \tag{2.2}$$

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

THE SIMULATION

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 correlation 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 $\rho \rightarrow \frac{1}{c+1}$ where $c > 0$. As ψ gets close to 1, the predictor vectors cluster about the line in the direction of $(1, ..., 1)^T$. Then $Y_i = 1 + 1x_{i,2} + ... + 1x_{i,k} + e_i$ 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) $EXP(1) - 1$ errors, iii) $uniform(-1,1)$ errors, and v) $0.9N(0,1) + 0.1N(0,100)$ 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 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, ..., $M = d = \min(\lceil n/J \rceil, p)$ variables in the MLR model, including a constant. We used J = 5, 10, 20, 50, and $\lceil n/p \rceil$ as long as $J \leq n/p$ since $n/J \geq p$ uses the full model. The selected model used the d variables. 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 $p - 1$.

Table 3.1 shows some simulations for the new large sample prediction interval (2.4)

Table 3.1. Simulated PI Coverages and Lengths

n	p	k	J	ψ	cov	len
100	20	1	20	0	0.9692	4.86813
1000	20	1	10	0	0.963	4.177

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 4.177, and variable selection used p=20 variables, including a constant.

```
library(leaps)

dvspisim(n=1000,p=20,k=1,j=10,nruns=5000,psi=0,type=1)

$fselpimenlen

[1]0.983

$fselpmenlen

[1]4.176784
```

CHAPTER 4

EXAMPLES

Table 4.1. Simulated PI Coverages and Lengths, Error type = i)

n	p	k	J	ψ	cov	len
100	20	1	20	0	0.9692	4.868132
100	20	1	20	$1/\sqrt{20}$	0.97	4.875998
100	20	1	50	0.9	0.9604	4.392484
100	20	19	5	0	0.9786	5.70508
100	40	1	50	0	0.968	4.434229
100	40	1	20	0.9	0.9624	4.735577
100	40	19	5	$1/\sqrt{40}$	0.9842	5.699041
100	40	19	10	0.9	0.9572	4.982567
100	40	39	10	0	0.928	22.25589
100	40	39	10	0.9	0.9268	5.827665
100	40	39	10	$1/\sqrt{40}$	0.9094	33.54649
100	100	1	50	$1/\sqrt{100}$	0.964	4.429076
100	100	1	50	0.9	0.9578	4.360497
100	100	99	5	0.9	0.8234	6.70691
100	200	1	50	0	0.9666	4.437201
100	200	1	50	0.9	0.96	4.356799
100	200	19	20	0.9	0.9174	5.219952
400	40	1	50	0	0.9506	4.124511
400	40	39	5	0	0.975	4.900493
400	400	19	20	$1/\sqrt{400}$	0.974	4.695523
400	800	19	20	0	0.9756	4.697523
400	800	19	20	$1/\sqrt{800}$	0.9752	4.696247
1000	20	1	5	0	0.963	4.176784
2000	20	1	5	0	0.9562	4.033074
2000	40	1	50	0	0.9636	4.171298
2000	2000	1	20	0	0.9228	4.104282

Table 4.2. Simulated PI Coverages and Lengths, Error type = ii)

n	p	k	J	ψ	cov	len
100	20	1	20	0	0.964	8.665205
100	20	1	20	$1/\sqrt{20}$	0.9654	8.673434
100	20	1	50	0.9	0.9528	7.148538
100	20	19	5	0	0.974	10.0023
100	40	1	50	0	0.953	7.21319
100	40	1	20	0.9	0.9578	8.345325
100	40	19	5	$1/\sqrt{40}$	0.9748	10.00913
100	40	19	10	0.9	0.9526	8.408594
100	40	39	10	0	0.93	23.05468
100	40	39	10	0.9	0.9494	8.670369
100	40	39	10	$1/\sqrt{40}$	0.913	33.93117
100	100	1	50	$1/\sqrt{100}$	0.9524	7.191489
100	100	1	50	0.9	0.9526	7.08956
100	100	99	5	0.9	0.8636	8.580055
100	200	1	50	0	0.9534	7.230835
100	200	1	50	0.9	0.9542	7.105399
100	200	19	20	0.9	0.9496	8.132494
400	40	1	50	0	0.9522	6.846537
400	40	39	5	0	0.976	8.747738
400	400	19	20	$1/\sqrt{400}$	0.974	8.438005
400	800	19	20	0	0.9714	8.440177
400	800	19	20	$1/\sqrt{800}$	0.971	8.437335
1000	20	1	5	0	0.9632	6.981182
2000	20	1	5	0	0.9578	6.646147
2000	40	1	50	0	0.96	7.005258
2000	2000	1	20	0	0.9466	7.246556

Table 4.3. Simulated PI Coverages and Lengths, Error type = iii)

n	p	k	J	ψ	cov	len
100	20	1	20	0	0.9664	4.725709
100	20	1	20	$1/\sqrt{20}$	0.967	4.726421
100	20	1	50	0.9	0.9572	3.802869
100	20	19	5	0	0.9772	5.652184
100	40	1	50	0	0.9642	3.73385
100	40	1	20	0.9	0.963	4.669649
100	40	19	5	$1/\sqrt{40}$	0.9816	5.647228
100	40	19	10	0.9	0.955	5.016633
100	40	39	10	0	0.9276	22.26106
100	40	39	10	0.9	0.9336	5.916098
100	40	39	10	$1/\sqrt{40}$	0.9044	33.55738
100	100	1	50	$1/\sqrt{100}$	0.962	3.734274
100	100	1	50	0.9	0.9578	3.787153
100	100	99	5	0.9	0.8352	6.768091
100	200	1	50	0	0.9666	3.759443
100	200	1	50	0.9	0.9642	3.812725
100	200	19	20	0.9	0.9396	5.333583
400	40	1	50	0	0.9578	3.679031
400	40	39	5	0	0.9788	4.671147
400	400	19	20	$1/\sqrt{400}$	0.9762	4.33107
400	800	19	20	0	0.9768	4.325581
400	800	19	20	$1/\sqrt{800}$	0.9778	4.325469
1000	20	1	5	0	0.9602	3.562779
2000	20	1	5	0	0.9544	3.322709
2000	40	1	50	0	0.9608	3.557465
2000	2000	1	20	0	0.9326	4.120581

Table 4.4. Simulated PI Coverages and Lengths, Error type = iv)

n	p	k	J	ψ	cov	len
100	20	1	20	0	0.9812	2.435685
100	20	1	20	$1/\sqrt{20}$	0.9826	2.43635
100	20	1	50	0.9	0.9844	2.219116
100	20	19	5	0	0.9926	2.962008
100	40	1	50	0	0.994	2.198207
100	40	1	20	0.9	0.9664	2.449893
100	40	19	5	$1/\sqrt{40}$	0.9904	2.961731
100	40	19	10	0.9	0.9408	3.083672
100	40	39	10	0	0.9298	21.92773
100	40	39	10	0.9	0.914	4.677512
100	40	39	10	$1/\sqrt{40}$	0.903	33.32117
100	100	1	50	$1/\sqrt{100}$	0.9954	2.200217
100	100	1	50	0.9	0.9792	2.222378
100	100	99	5	0.9	0.8068	6.01635
100	200	1	50	0	0.9936	2.197057
100	200	1	50	0.9	0.9762	2.215849
100	200	19	20	0.9	0.9074	4.008402
400	40	1	50	0	0.9524	2.053494
400	40	39	5	0	0.9804	2.411465
400	400	19	20	$1/\sqrt{400}$	0.9812	2.224335
400	800	19	20	0	0.9856	2.222391
400	800	19	20	$1/\sqrt{800}$	0.9864	2.222302
1000	20	1	5	0	0.9734	2.00578
2000	20	1	5	0	0.9562	1.944549
2000	40	1	50	0	0.9686	1.996759
2000	2000	1	20	0	0.9076	2.162859

Table 4.5. Simulated PI Coverages and Lengths, Error type = v)

n	p	k	J	ψ	cov	len
100	20	1	20	0	0.9614	19.79061
100	20	1	20	$1/\sqrt{20}$	0.962	19.78306
100	20	1	50	0.9	0.9448	13.55284
100	20	19	5	0	0.9688	22.55319
100	40	1	50	0	0.946	13.54309
100	40	1	20	0.9	0.962	18.88696
100	40	19	5	$1/\sqrt{40}$	0.9592	20.09148
100	40	19	10	0.9	0.9572	18.43204
100	40	39	10	0	0.9392	27.02255
100	40	39	10	0.9	0.9592	18.55625
100	40	39	10	$1/\sqrt{40}$	0.9142	36.18274
100	100	1	50	$1/\sqrt{100}$	0.9442	13.51805
100	100	1	50	0.9	0.9448	13.57073
100	100	99	5	0.9	0.9064	15.32787
100	200	1	50	0	0.9422	13.47423
100	200	1	50	0.9	0.944	13.55491
100	200	19	20	0.9	0.9494	17.46374
400	40	1	50	0	0.949	14.58944
400	40	39	5	0	0.968	21.73703
400	400	19	20	$1/\sqrt{400}$	0.969	21.18569
400	800	19	20	0	0.9692	21.06037
400	800	19	20	$1/\sqrt{800}$	0.9694	21.23808
1000	20	1	5	0	0.9586	15.78017
2000	20	1	5	0	0.9516	14.38024
2000	40	1	50	0	0.9578	16.17342
2000	2000	1	20	0	0.9638	17.7788

CHAPTER 5

CONCLUSION

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). The method described here can be used for many other methods, such as lasso and 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 `dvspisim` was used to do the simulation.

According to the simulation tables we can found that 1) If $\frac{n}{J} < k$, then the average length is a lot higher than the optimal length. Then ψ=0.9 sometimes worked better but sometime had undercoverage. 2) If $\frac{n}{J} > k$, then $\frac{n}{J}$ close to k with in $\frac{n}{k}$ large often work well.

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 ModelsWith 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,” *The 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.

[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

Kosman Rajapaksha

Email address (drhansana@gmail.com)

Post graduate Institute of Science, University of peradeniya
M.Sc. in Applied Statistics, 2014
Faculty of Science, University of Peradeniya
B.Sc. , 2011

Research Paper Title:
Prediction Intervals After Forward Selection Using d Variables

Major professor: Dr.David Olive