Summer 6-28-2017

Prediction Intervals For Partial Least Squares and Principal Component Regression

Sung-ho Kim
Southern Illinois University Carbondale, shkim@siu.edu

PREDICTION INTERVALS FOR PARTIAL LEAST SQUARES AND PRINCIPAL COMPONENT REGRESSION USING D VARIABLES

by

Sung-ho Kim

B.S., Southern Illinois University Carbondale, 2015

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 FOR PARTIAL LEAST SQUARES AND PRINCIPAL COMPONENT REGRESSION USING D VARIABLES

by

Sung-ho Kim

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
June 28, 2017

AN ABSTRACT OF THE RESEARCH PAPER OF

SUNG-HO KIM, for the Master of Science degree in MATHEMATICS, presented on JUNE 28, 2017, at Southern Illinois University Carbondale.

TITLE: PREDICTION INTERVALS FOR PARTIAL LEAST SQUARES AND PRINCIPAL COMPONENT REGRESSION USING D VARIABLES

MAJOR PROFESSOR: Dr. David J. Olive

This paper, taken from Olive (2017d), presents and examines a prediction interval for the multiple linear regression model $Y = \beta_1 x_1 + \cdots + \beta_p x_p + e$, where the partial least squares or principal component regression is selected using d = $\min(\lceil n/J \rceil, p)$ variables $v_1, v_2, ..., v_d$ for some positive integer J such as 10 or 20. Here v_1 corresponds to a constant and v_i is a PLS component or principal component for $i \geq 2$.

TABLE OF CONTENTS

CHAPTER PAGE

CHAPTERS

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 $\boldsymbol{x}_i^T = (x_{i,1}, ..., x_{i,p}) = (1 \quad \boldsymbol{u}_i^T)$ 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. Ordinary least squares (OLS) is often used for inference if n/p is large.

It is often convenient to use the centered response $\boldsymbol{Z} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$ and the $n \times (p-1)$ matrix of standardized nontrivial predictors $\boldsymbol{W} = (W_{ij})$. For $j = 1, ..., p-1$, let W_{ij} denote the $(j+1)$th variable standardized so that $\sum_{i=1}^n W_{ij} = 0$ and $\sum_{i=1}^n W_{ij}^2 = n$. Hence

$$W_{ij} = \frac{x_{i,j+1} - \overline{x}_{j+1}}{\tilde{\sigma}_{j+1}} \quad \text{where} \quad \tilde{\sigma}^2_{\mathrm{j+1}} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i=1}}^{\mathrm{n}} (\mathrm{x}_{\mathrm{i,j+1}} - \overline{\mathrm{x}}_{\mathrm{j+1}})^2.$$

Note that the sample correlation matrix of the nontrivial predictors $\boldsymbol{u}_i$ is

$$\boldsymbol{R}_{\boldsymbol{u}} = \frac{\boldsymbol{W}^T \boldsymbol{W}}{n}.$$

Then regression through the origin is used for the model

$$\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e} \tag{1.3}$$

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

There are many alternative methods for estimating $\boldsymbol{\beta}$, including forward selection with OLS, principal component regression (PCR), and partial least squares (PLS) due to Wold (1975). Forward selection, PCR, and PLS use variables $v_1 = 1$ (the constant or trivial predictor) and $v_j = \boldsymbol{\gamma}_j^T \boldsymbol{x}$ that are linear combinations of the predictors for $j = 2, ..., p$. Model I_i uses variables $v_1, v_2, ..., v_i$ for $i = 1, ..., M$ where $M \leq p$ and often $M \leq \min(p, n/10)$. Then M models I_i are used where OLS is used to regress Y (or Z) on $v_1, ..., v_i$. Then a criterion chooses the final submodel I_d from candidates $I_1, ..., I_M$. See James, Witten, Hastie, and Tibshirani (2013, ch. 6), Olive (2017d), Pelawa Watagoda (2017), and Pelawa Watagoda and Olive (2017) for more details about these three methods.

Partial least squares (PLS) uses variables $v_1 = 1$ and "PLS components" $v_j = \boldsymbol{\gamma}_j^T \boldsymbol{x}$ for $j = 2, ..., p$. Often k–fold cross validation is used to pick the PLS model from $I_1, ..., I_M$. If $M = p$, then the PLS I_p model is the OLS full model. Chun and Keleş (2010) show that PLS does not give a consistent estimator of $\boldsymbol{\beta}$ unless $p/n \to 0$. Also see Cook, Helland, and Su (2013), and Wold (1985, 2006). Denham (1997) suggested a prediction interval (PI) for PLS that assumes the number of components is selected in advance.

Some notation for eigenvalues, eigenvectors, orthonormal eigenvectors, positive definite matrices, and positive semidefinite matrices will be useful before defining principal component regression, which is also called principal components regression. See Olive (2017d, ch. 3).

Notation: Recall that a square symmetric $p \times p$ matrix $\boldsymbol{A}$ has an *eigenvalue* λ with corresponding *eigenvector* $\boldsymbol{x} \neq \boldsymbol{0}$ if

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}. \tag{1.4}$$

The eigenvalues of $\boldsymbol{A}$ are real since $\boldsymbol{A}$ is symmetric. Note that if constant $c \neq 0$ and $\boldsymbol{x}$ is an eigenvector of $\boldsymbol{A}$, then $c\ \boldsymbol{x}$ is an eigenvector of $\boldsymbol{A}$. Let $\boldsymbol{e}$ be an eigenvector of $\boldsymbol{A}$ with unit length $\|\boldsymbol{e}\| = \sqrt{\boldsymbol{e}^T \boldsymbol{e}} = 1$. Then $\boldsymbol{e}$ and $-\boldsymbol{e}$ are eigenvectors with unit length, and $\boldsymbol{A}$ has p eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), ..., (\lambda_p, \boldsymbol{e}_p)$. Since $\boldsymbol{A}$ is symmetric, the eigenvectors are chosen such that the $\boldsymbol{e}_i$ are *orthonormal*: $\boldsymbol{e}_i^T \boldsymbol{e}_i = 1$ and $\boldsymbol{e}_i^T \boldsymbol{e}_j = 0$ for

$i \neq j$. The symmetric matrix $\boldsymbol{A}$ is *positive definite* iff all of its eigenvalues are positive, and *positive semidefinite* iff all of its eigenvalues are nonnegative. If $\boldsymbol{A}$ is positive semidefinite, let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. If $\boldsymbol{A}$ is positive definite, then $\lambda_p > 0$.

Theorem 1. Let $\boldsymbol{A}$ be a $p \times p$ symmetric matrix with eigenvector eigenvalue pairs $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), ..., (\lambda_p, \boldsymbol{e}_p)$ where $\boldsymbol{e}_i^T \boldsymbol{e}_i = 1$ and $\boldsymbol{e}_i^T \boldsymbol{e}_j = 0$ if $i \neq j$ for $i = 1, ..., p$. Then the *spectral decomposition* of $\boldsymbol{A}$ is

$$\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T = \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1^T + \cdots + \lambda_p \boldsymbol{e}_p \boldsymbol{e}_p^T.$$

Using the same notation as Johnson and Wichern (1988, pp. 50-51), let $\boldsymbol{P} = [\boldsymbol{e}_1 \ \boldsymbol{e}_2 \ \cdots \ \boldsymbol{e}_p]$ be the $p \times p$ orthogonal matrix with ith column $\boldsymbol{e}_i$. Then $\boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I}$. Let $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_p)$ and let $\boldsymbol{\Lambda}^{1/2} = \text{diag}(\sqrt{\lambda_1}, ..., \sqrt{\lambda_p})$. If $\boldsymbol{A}$ is a positive definite $p \times p$ symmetric matrix with spectral decomposition $\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T$, then $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^T$ and

$$\boldsymbol{A}^{-1} = \boldsymbol{P}\boldsymbol{\Lambda}^{-1}\boldsymbol{P}^T = \sum_{i=1}^{p} \frac{1}{\lambda_i} \boldsymbol{e}_i \boldsymbol{e}_i^T.$$

Theorem 2. Let $\boldsymbol{A}$ be a positive definite $p \times p$ symmetric matrix with spectral decomposition $\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T$. The *square root matrix* $\boldsymbol{A}^{1/2} = \boldsymbol{P}\boldsymbol{\Lambda}^{1/2}\boldsymbol{P}^T$ is a positive definite symmetric matrix such that $\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2} = \boldsymbol{A}$.

Principal component regression (PCR) uses OLS regression on the principal components of the correlation matrix $\boldsymbol{R_u}$ of the $p-1$ nontrivial predictors $u_1 = x_2, ..., u_{p-1} = x_p$. Suppose $\boldsymbol{R_u}$ has eigenvalue eigenvector pairs $(\hat{\lambda}_1, \hat{\boldsymbol{e}}_1), ..., (\hat{\lambda}_K, \hat{\boldsymbol{e}}_K)$ where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_K \geq 0$ where $K = \min(n, p-1)$. Then $\boldsymbol{R_u}\hat{\boldsymbol{e}}_i = \hat{\lambda}_i \hat{\boldsymbol{e}}_i$ for $i = 1, ..., K$. Since $\boldsymbol{R_u}$ is a symmetric positive semidefinite matrix, the $\hat{\lambda}_i$ are real and nonnegative.

The eigenvectors $\hat{\boldsymbol{e}}_i$ are *orthonormal*: $\hat{\boldsymbol{e}}_i^T \hat{\boldsymbol{e}}_i = 1$ and $\hat{\boldsymbol{e}}_i^T \hat{\boldsymbol{e}}_j = 0$ for $i \neq j$. If the eigenvalues are unique, then $\hat{\boldsymbol{e}}_i$ and $-\hat{\boldsymbol{e}}_i$ are the only orthonormal eigenvectors corresponding to $\hat{\lambda}_i$. For example, the eigenvalue eigenvector pairs can be found using the singular value decomposition of the matrix $\boldsymbol{W}/\sqrt{n-a}$ where $\boldsymbol{W}$ is the matrix of the standardized nontrivial predictors $\boldsymbol{w}_i$, the sample covariance matrix

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \frac{1}{n-a}\sum_{i=1}^n (\boldsymbol{w}_i - \overline{\boldsymbol{w}})(\boldsymbol{w}_i - \overline{\boldsymbol{w}})^T = \frac{1}{n-a}\sum_{i=1}^n \boldsymbol{w}_i \boldsymbol{w}_i^T = \boldsymbol{R}_{\boldsymbol{u}},$$

and usually $a = 0$ or $a = 1$. If $n > K = p - 1$, then the *spectral decomposition* of $\boldsymbol{R}_{\boldsymbol{u}}$ is

$$\boldsymbol{R}_{\boldsymbol{u}} = \sum_{i=1}^{p-1} \hat{\lambda}_i \hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_i^T = \hat{\lambda}_1 \hat{\boldsymbol{e}}_1 \hat{\boldsymbol{e}}_1^T + \cdots + \hat{\lambda}_{p-1} \hat{\boldsymbol{e}}_{p-1} \hat{\boldsymbol{e}}_{p-1}^T,$$

and $\sum_{i=1}^{p-1} \hat{\lambda}_i = p - 1$.

Let $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$ denote the standardized vectors of nontrivial predictors. Then the K *principal components* corresponding to the jth case $\boldsymbol{w}_j$ are $P_{j1} = \hat{\boldsymbol{e}}_1^T \boldsymbol{w}_j$, ..., $P_{jK} = \hat{\boldsymbol{e}}_K^T \boldsymbol{w}_j$.

Principal components have a nice geometric interpretation if $n > K = p - 1$. If $n > K$ and $\boldsymbol{R}_{\boldsymbol{u}}$ is nonsingular, then the hyperellipsoid

$$\{\boldsymbol{w} | D^2_{\boldsymbol{w}}(\boldsymbol{0}, \boldsymbol{R}_{\boldsymbol{u}}) \leq h^2\} = \{\boldsymbol{w} : \boldsymbol{w}^T \boldsymbol{R}_{\boldsymbol{u}}^{-1} \boldsymbol{w} \leq h^2\}$$

is centered at $\boldsymbol{0}$. The volume of the hyperellipsoid is

$$\frac{2\pi^{K/2}}{K\Gamma(K/2)} |\boldsymbol{R}_{\boldsymbol{u}}|^{1/2} h^K.$$

Then points at squared distance $\boldsymbol{w}^T \boldsymbol{R}_{\boldsymbol{u}}^{-1} \boldsymbol{w} = h^2$ from the origin lie on the hyperellipsoid centered at the origin whose axes are given by the eigenvectors $\hat{\boldsymbol{e}}_i$ where the half length in the direction of $\hat{\boldsymbol{e}}_i$ is $h\sqrt{\hat{\lambda}_i}$. Let $j = 1, ..., n$. Then the first principal component P_{j1} is obtained by projecting the $\boldsymbol{w}_j$ on the (longest) major axis of the hyperellipsoid, the second principal component P_{j2} is obtained by projecting the $\boldsymbol{w}_j$ on the next longest axis of the hyperellipsoid, ..., and the $(p-1)$th principal component $P_{j,p-1}$ is obtained by projecting the $\boldsymbol{w}_j$ on the (shortest) minor axis of the hyperellipsoid. Examine Figure 1.1 for two ellipsoids with 2 nontrivial predictors. The axes of the hyperellipsoid are a rotation of the usual axes about the origin.

Let the random variable V_i correspond to the ith principal component, and let $(P_{1i}, ..., P_{ni})^T = (V_{1i}, ..., V_{ni})^T$ be the observed data for V_i. Then the sample mean

$$\overline{V}_i = \frac{1}{n}\sum_{k=1}^n V_{ki} = \frac{1}{n}\sum_{k=1}^n \hat{\boldsymbol{e}}_i^T \boldsymbol{w}_k = \hat{\boldsymbol{e}}_i^T \overline{\boldsymbol{w}} = \hat{\boldsymbol{e}}_i^T \boldsymbol{0} = 0,$$

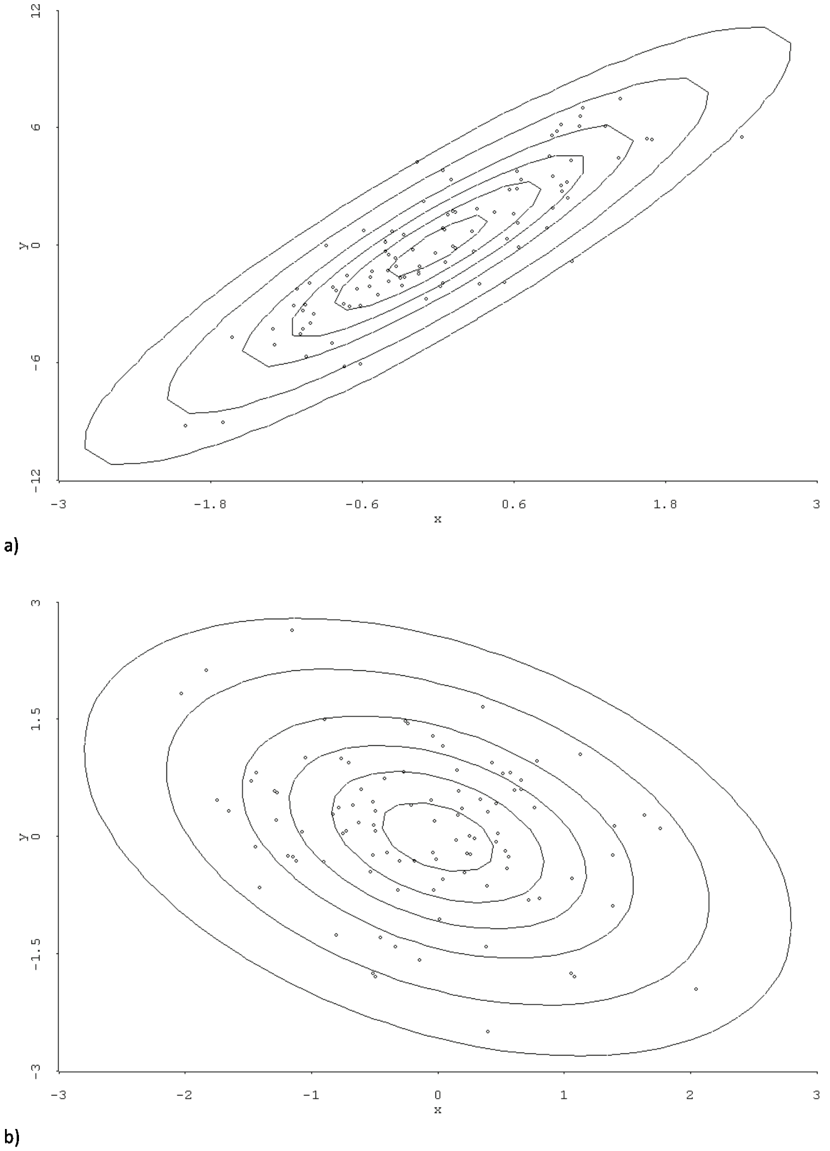

Figure 1.1. Population Prediction Regions for 2 MVN Distributions

and the sample covariance of V_i and V_j is

$$Cov(V_i, V_j) = \frac{1}{n}\sum_{k=1}^n (V_{ki} - \overline{V}_i)(V_{kj} - \overline{V}_j) = \frac{1}{n}\sum_{k=1}^n \hat{\boldsymbol{e}}_i^T \boldsymbol{w}_k \boldsymbol{w}_k^T \hat{\boldsymbol{e}}_j = \hat{\boldsymbol{e}}_i^T \boldsymbol{R}_{\boldsymbol{u}} \hat{\boldsymbol{e}}_j$$

$= \hat{\lambda}_j \hat{\boldsymbol{e}}_i^T \hat{\boldsymbol{e}}_j = 0$ for $i \neq j$ since the sample covariance matrix of the standardized data is

$$\frac{1}{n}\sum_{k=1}^n \boldsymbol{w}_k \boldsymbol{w}_k^T = \boldsymbol{R}_{\boldsymbol{u}}$$

and $\boldsymbol{R}_{\boldsymbol{u}} \hat{\boldsymbol{e}}_j = \hat{\lambda}_j \hat{\boldsymbol{e}}_j$. Hence V_i and V_j are uncorrelated.

PCR uses linear combinations of the standardized data as predictors. Let $v_1 = 1$ and $v_j = \boldsymbol{\gamma}_j^T \boldsymbol{w} = \hat{\boldsymbol{e}}_{j-1}^T \boldsymbol{w} = V_{j-1}$ for $j = 2, ..., K$. Let model I_i contain $v_1, ..., v_i$. Then for model I_i, PCR uses OLS regression of Y on $v_1, ..., v_i$.

Alternatively let $v_j = \hat{\boldsymbol{e}}_j^T \boldsymbol{w}$ for $j = 1, ..., K$ and let model I_i contain $v_1, ..., v_i$. Then for model I_i, use OLS regression of $Z = Y - \overline{Y}$ on $v_1, ..., v_i$ with $\hat{Y} = \hat{Z} + \overline{Y}$.

Generally there is no reason why the predictors should be ranked from best to worst by $v_1, v_2, ..., v_K$. Performing OLS forward selection or lasso on $v_1, ..., v_K$ may be more effective. There is one exception. Suppose $\sum_{i=1}^J \hat{\lambda}_i \geq q(p-1)$ where $0.5 \leq q \leq 1$, e.g. $q = 0.8$ where J is a lot smaller than $p-1$. Then the J predictors $V_1, ..., V_J$ capture much of the information of the standardized nontrivial predictors $w_1, ..., w_{p-1}$. Then regressing Y on $1, V_1, ..., V_J$ may be competitive with regressing Y on $1, w_1, ..., w_{p-1}$. This exception tends to occur when p is very small, and is an example of dimension reduction. PCR is equivalent to OLS on the full model when Y is regressed on a constant and all K of the principal components. PCR can also be useful if $\boldsymbol{X}$ is singular or nearly singular (ill conditioned). In general, PCR does not give a consistent estimator of $\boldsymbol{\beta}$ unless PCR is the full OLS model so all $p - 1$ principal components are used.

Variable selection is the search for a subset of predictor variables that can be deleted without important loss of information. Following Olive(2017c, p.99), a *model for variable selection* can be described by

$$\boldsymbol{x}^T \boldsymbol{\beta} = \boldsymbol{x}_S^T \boldsymbol{\beta}_S + \boldsymbol{x}_E^T \boldsymbol{\beta}_E = \boldsymbol{x}_S^T \boldsymbol{\beta}_S \tag{1.5}$$

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, $\boldsymbol{\beta}_E = \mathbf{0}$ 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 (1.5) holds. Then

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_{I/S}^T\boldsymbol{\beta}_{(I/S)} + \boldsymbol{x}_O^T\mathbf{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I \tag{1.6}$$

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 $S \subseteq I$.

When there is a sequence of M submodels, the final submodel I_d needs to be selected. 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. A simple method is to take the model that uses $d = \min(\lceil n/J \rceil, p)$ variables $V_1, ..., V_d$. This is the method that we will investigate. If p is fixed, the method will use the full OLS 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 $P(\hat{L}_n \leq Y_f \leq \hat{U}_n) \to 1 - \delta$ 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

$$\mathrm{shorth(c) = [Z_{(s)}, Z_{(s+c-1)}].} \tag{1.7}$$

Let

$$k_n = \lceil n(1-\delta).\rceil \tag{1.8}$$

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

$$c = \min(n, \lceil n[1 - \delta + 1.12\sqrt{\delta/n}\ ]\ \rceil). \tag{1.9}$$

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

$$b_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n+2p}{n-p}}. \tag{1.10}$$

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

$$[\hat{m}(\boldsymbol{x}_f) + b_n \tilde{\xi}_{\delta_1}, \hat{m}(\boldsymbol{x}_f) + b_n \tilde{\xi}_{1-\delta_2}]. \tag{1.11}$$

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 VARIABLE SELECTION

If n/p is large, the PI (1.11) can be used for the variable selection estimators with $\hat{m}(\boldsymbol{x}_f) = \hat{Y}_f = \boldsymbol{x}_{f,I_d}^T \hat{\boldsymbol{\beta}}_{I_d}$ 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 use $d = M = \min(\lceil n/J \rceil, p)$ variables where n/p is not necessarily large.

PI (1.11) 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, PCR, and PLS, $d = j$ is the number of components $V_1, ..., V_j$ in model I_j. 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 (1.11) 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 $Cov(\boldsymbol{u}) = \boldsymbol{\Sigma_u} = \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ 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 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$(\mathrm{K}, \rho_\mathrm{n})$, 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.

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. 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 1 shows some simulations for the new large sample prediction interval (2.4).

Table 3.1. Simulated PI Coverages and Lengths, Error type = i)

n	p	k	J	ψ	pcrcov	pcrlen	plscov	plslen
1000	20	1	10	0	0.960	4.175	0.960	4.175

Some R code is below. For 5000 runs of the nominal large sample 95% PI, the observed coverage for PCR and PLS was 0.960 and the average length was 4.175. Since $\min(n/J, p) = 20$, the OLS full model was fit for both PCR and PLS.

```
library(pls)
dpisim3(n=1000,p=20,k=1,J=10,nruns=5000,psi=0,type=1)
$pcrpicov
[1] 0.9604
$pcrpimenlen
[1] 4.174539
$plspicov
[1] 0.9604
$plspimenlen
[1] 4.174539  #PCR and PLS used full model OLS
```

CHAPTER 4

ERROR TYPE 1 EXAMPLES

Table 4.1. PI coverage and length for error type 1 (runs = 5000)

n	p	k	J	ψ	PCR-PIcov	PCR-PIlen	PLS-PIcov	PLS-PIlen
100	20	1	5	0	0.9838	5.6986	0.9838	5.6986
100	20	1	10	0	0.9800	6.1915	0.9648	4.9670
100	20	1	50	0	0.9636	6.1435	0.9352	4.3383
100	20	19	5	$1/\sqrt{p}$	0.9842	5.6968	0.9842	5.3696
100	20	19	5	0	0.9822	5.7227	0.9822	5.7227
100	20	19	10	0	0.9678	14.9895	0.9676	4.9840
100	20	19	10	0	0.9702	15.0622	0.9682	4.9871
100	40	1	5	0.9	0.9856	5.7251	0.9308	4.9458
100	40	19	5	0	0.9706	15.0521	0.9376	4.9373
100	40	19	10	$1/\sqrt{p}$	0.9802	12.2942	0.8922	4.3135
100	40	19	20	$1/\sqrt{p}$	0.9814	13.2225	0.9018	4.3601
100	40	19	50	0	0.9654	19.3259	0.8998	8.5771
100	40	39	5	0.9	0.9864	5.6988	0.9290	4.9310
100	40	19	5	0.9	0.9862	5.8007	0.9312	4.9389
100	100	19	5	0.9	0.9894	5.9683	0.1976	1.3784
100	100	1	50	0.9	0.9642	4.4555	0.9642	4.4555
100	100	19	5	$1/\sqrt{p}$	0.9816	16.6874	0.2158	1.4890
200	20	19	10	0	0.9764	4.9727	0.9764	4.9727
200	40	39	5	0	0.9812	5.3764	0.9812	5.3764
200	40	39	10	0	0.9584	19.2529	0.9586	4.6827
200	200	39	5	0	0.9764	26.2433	0.1044	0.9628

Table 4.2. PI coverage and length for error type 1 (runs = 5000)

n	p	k	J	ψ	PCR-PIcov	PCR-PIlen	PLS-PIcov	PLS-PIlen
200	200	19	10	0	0.9734	19.6847	0.1814	1.2602
200	200	1	50	0.9	0.9634	4.2975	0.6622	2.4768
200	400	19	5	0.9	0.9856	5.7063	0.0000	0+
200	400	19	10	0	0.9766	20.5782	0.0012	0.0093
400	400	1	20	$1/\sqrt{p}$	0.9754	5.8147	0.1714	1.1049
400	40	1	20	$1/\sqrt{p}$	0.9766	5.2346	0.9700	4.5699
400	400	19	20	0.9	0.9766	5.0590	0.1646	1.0945
400	400	19	10	$1/\sqrt{p}$	0.9756	18.0847	0.0960	0.8461
400	400	399	5	0	0.9740	81.0937	0.0518	0.7160
400	800	1	5	0.9	0.9892	5.3092	0.0000	0+
1000	1000	1	5	0	0.9834	6.6318	0.0170	3.9468
1000	2000	19	10	0	0.9746	20.1530	0.0000	0+
1000	1000	19	5	0.9	0.9894	5.5498	0.0192	0.3966
1000	1000	19	10	0	0.9744	19.2322	0.0320	0.5119
1000	1000	999	10	$1/\sqrt{p}$	0.9810	5.6068	0.0324	0.5037
1000	2000	19	5	0.9	0.9882	5.6161	0.0000	0.1948
2000	2000	19	10	0	0.9772	19.2329	0.0138	0.3587
2000	20	19	10	0	0.9584	4.0332	0.9584	4.0332
2000	40	19	50	0	0.9632	4.1714	0.9632	4.1714
2000	2000	19	50	0	0.9570	18.1827	0.0710	0.6969
2000	4000	19	20	0.9	0.9772	5.0651	0.0000	0.1042

CHAPTER 5

ERROR TYPE 2 EXAMPLES

Table 5.1. PI coverage and length for error type 2 (runs = 5000)

n	p	k	J	ψ	PCR-PIcov	PCR-PIlen	PLS-PIcov	PLS-PIlen
100	20	1	5	0	0.9792	9.9965	0.9792	9.9965
100	20	1	10	0	0.9720	9.9857	0.9604	8.7199
100	20	1	50	0	0.9566	8.2785	0.9334	6.7660
100	20	19	5	$1/\sqrt{p}$	0.9746	10.0399	0.9746	10.0399
100	20	19	5	0	0.9730	10.0729	0.9730	10.0729
100	20	19	10	0	0.9704	16.8327	0.9622	8.7501
100	40	1	5	0.9	0.9730	10.0105	0.9386	8.4738
100	40	19	5	0	0.9696	17.0405	0.9398	8.4748
100	40	19	10	$1/\sqrt{p}$	0.9756	14.3545	0.9080	7.3646
100	40	19	20	$1/\sqrt{p}$	0.9766	14.9748	0.9140	7.1425
100	40	19	50	0	0.9630	20.2452	0.9078	9.8723
100	40	39	5	0.9	0.9760	10.0731	0.9394	8.5256
100	40	19	5	0.9	0.9772	10.0500	0.9356	8.4479
100	100	19	5	0.9	0.9782	10.1202	0.1986	2.2876
100	100	1	5	0.9	0.9764	9.9913	0.1942	2.2722
100	100	19	5	$1/\sqrt{p}$	0.9798	18.5163	0.2162	2.3866
200	20	19	10	0	0.9706	8.7374	0.9706	8.7374
200	40	39	5	0	0.9712	9.2992	0.9712	9.2992
200	40	39	10	0	0.9632	20.4525	0.9596	8.1193
200	200	39	5	0	0.9760	27.3389	0.0986	1.5524

Table 5.2. PI coverage and length for error type 2 (runs = 5000)

n	p	k	J	ψ	PCR-PIcov	PCR-PIlen	PLS-PIcov	PLS-PIlen
200	200	19	10	0	0.9758	20.8502	0.1784	2.0256
200	200	1	50	0.9	0.9578	7.0348	0.7320	4.1288
200	400	19	5	0.9	0.9764	9.5186	0.0000	0+
200	400	19	10	0	0.9786	21.7175	0.0012	0.0109
400	400	1	20	$1/\sqrt{p}$	0.9660	9.2259	0.1674	1.8628
400	40	1	20	$1/\sqrt{p}$	0.9740	8.7070	0.9686	8.1353
400	400	19	20	0.9	0.9702	8.5641	0.1676	1.8438
400	400	19	10	$1/\sqrt{p}$	0.9756	19.3178	0.0932	1.3996
400	400	399	5	0	0.9764	81.4265	0.0522	1.1464
400	800	1	5	0.9	0.9774	9.3098	0.0000	0+
1000	1000	1	5	0	0.9800	10.0610	0.0220	0.6747
1000	2000	19	10	0	0.9778	21.2955	0.0000	0+
1000	1000	19	5	0.9	0.9788	9.4796	0.0170	0.6787
1000	1000	999	10	$1/\sqrt{p}$	0.9760	9.1170	0.0360	0.8613
1000	2000	19	5	0.9	0.9788	9.4974	0.0000	0.1736
2000	20	19	10	0	0.9532	6.6374	0.9532	6.6374
2000	40	19	50	0	0.9602	7.0086	0.9602	7.0086
2000	4000	19	20	0.9	0.9772	8.6916	0.0000	0.0960

CHAPTER 6

ERROR TYPE 3 EXAMPLES

Table 6.1. PI coverage and length for error type 3 (runs = 5000)

n	p	k	J	ψ	PCR-PIcov	PCR-PIlen	PLS-PIcov	PLS-PIlen
100	20	1	5	0	0.9828	5.6715	0.9828	5.6715
100	20	1	10	0	0.9746	6.2819	0.9684	4.9344
100	20	1	50	0	0.9564	6.0674	0.9364	4.1519
100	20	19	5	$1/\sqrt{p}$	0.9772	5.6342	0.9772	5.6342
100	20	19	5	0	0.9794	5.6661	0.9794	5.6661
100	20	19	10	0	0.9698	15.1213	0.9670	4.9244
100	40	1	5	0.9	0.9806	5.6583	0.9358	5.0030
100	40	19	5	0	0.9682	15.0609	0.9350	5.0092
100	40	19	10	$1/\sqrt{p}$	0.9804	12.3775	0.9054	4.3720
100	40	19	20	$1/\sqrt{p}$	0.9792	13.3088	0.9002	4.4148
100	40	19	50	0	0.9616	19.4041	0.8954	8.5805
100	40	39	5	0.9	0.9854	5.6216	0.9364	4.9688
100	40	19	5	0.9	0.9804	5.7460	0.9364	4.9729
100	100	19	5	0.9	0.9860	5.9587	0.2032	1.3571
100	100	1	5	0.9	0.9842	5.6614	0.2074	1.3740
100	100	19	5	$1/\sqrt{p}$	0.9808	18.4986	0.2198	2.3689
200	20	19	10	0	0.9744	4.7202	0.9744	4.7202
200	40	39	5	0	0.9750	5.2786	0.9750	5.2786
200	40	39	10	0	0.9610	19.2659	0.9598	4.6056
200	200	39	5	0	0.9732	26.2859	0.0964	0.9564
200	200	19	5	0	0.9728	18.7091	0.0982	0.9311

Table 6.2. PI coverage and length for error type 3 (runs = 5000)

n	p	k	J	ψ	PCR-PIcov	PCR-PIlen	PLS-PIcov	PLS-PIlen
200	200	19	10	0	0.9770	19.7015	0.1866	1.2476
200	400	19	5	0.9	0.9818	5.7142	0.0000	0+
200	400	19	10	0	0.9724	20.5824	0.0006	0.0093
400	400	1	20	$1/\sqrt{p}$	0.9760	6.3802	0.1696	1.1061
400	40	1	20	$1/\sqrt{p}$	0.9712	5.3134	0.9742	4.3537
400	400	19	20	0.9	0.9762	4.9638	0.1692	1.0918
400	400	19	10	$1/\sqrt{p}$	0.9688	18.1031	0.0936	0.8421
400	400	399	5	0	0.9756	81.0709	0.0518	0.7113
400	800	1	5	0.9	0.9818	5.2501	0.0000	0+
1000	1000	1	5	0	0.9822	6.7610	0.0184	0.3947
1000	2000	19	10	0	0.9810	20.1860	0.0000	0+
1000	1000	19	5	0.9	0.9854	5.5665	0.0180	0.3952
1000	1000	999	10	$1/\sqrt{p}$	0.9836	5.6595	0.0376	0.5020
1000	2000	19	5	0.9	0.9840	5.6525	0.0000	0.1976
2000	2000	19	10	0	0.9780	19.2504	0.0150	0.3563
2000	2000	19	50	0	0.9590	18.1872	0.0758	0.6978

CHAPTER 7

ERROR TYPE 4 EXAMPLES

Table 7.1. PI coverage and length for error type 4 (runs = 5000)

n	p	k	J	ψ	PCR-PIcov	PCR-PIlen	PLS-PIcov	PLS-PIlen
100	20	1	5	0	0.9898	2.9655	0.9898	2.9655
100	20	1	10	0	0.9764	4.3736	0.9728	2.5878
100	20	1	50	0	0.9646	4.9268	0.9346	2.6719
100	20	19	5	$1/\sqrt{p}$	0.9890	2.9630	0.9890	2.9630
100	20	19	5	0	0.9914	2.9664	0.9914	2.9664
100	20	19	10	0	0.9744	14.3614	0.9714	2.5855
100	40	1	5	0.9	0.9966	2.9863	0.9332	2.7193
100	40	19	5	0	0.9712	14.2618	0.9334	2.7160
100	40	19	10	$1/\sqrt{p}$	0.9758	11.5036	0.8866	2.3761
100	40	19	20	$1/\sqrt{p}$	0.9796	12.5995	0.8750	2.7480
100	40	19	50	0	0.9634	19.0424	0.8874	8.0006
100	40	39	5	0.9	0.9964	2.9637	0.9302	2.7160
100	40	19	5	0.9	0.9914	3.1739	0.9282	2.7096
100	100	19	5	0.9	0.9920	3.5293	0.2010	0.8019
100	100	19	5	$1/\sqrt{p}$	0.9812	16.0850	0.2258	0.9396
200	20	19	10	0	0.9872	2.4599	0.9872	2.4599
200	40	39	5	0	0.9860	2.7904	0.9860	2.7904
200	40	39	10	0	0.9594	18.8393	0.9566	2.4349
200	200	19	5	0	0.9768	18.2050	0.104	0.5671

Table 7.2. PI coverage and length for error type 4 (runs = 5000)

n	p	k	J	ψ	PCR-PIcov	PCR-PIlen	PLS-PIcov	PLS-PIlen
200	200	19	10	0	0.9704	19.2521	0.1858	0.8050
200	400	19	5	0.9	0.9904	3.5016	0.0000	0+
200	400	19	10	0	0.9748	20.1496	0.0006	0.0086
400	400	1	20	$1/\sqrt{p}$	0.9760	4.9445	0.1724	0.6484
400	40	1	20	$1/\sqrt{p}$	0.9694	3.5638	0.9698	2.2470
400	400	19	20	0.9	0.9788	3.1004	0.1672	0.6344
400	400	19	10	$1/\sqrt{p}$	0.9800	17.6391	0.0964	0.5085
400	400	399	5	0	0.9748	80.8251	0.0572	0.4646
400	800	1	5	0.9	0.9986	2.7852	0.0000	0+
1000	1000	1	5	0	0.9802	4.9959	0.0168	0.2277
1000	1000	19	5	0.9	0.9890	3.3199	0.0190	0.2291
1000	1000	999	10	$1/\sqrt{p}$	0.9796	3.8343	0.0334	0.2908
1000	2000	19	5	0.9	0.9910	3.4403	0.0000	0.2020

CHAPTER 8

ERROR TYPE 5 EXAMPLES

Table 8.1. PI coverage and length for error type 5 (runs = 5000)

n	p	k	J	ψ	PCR-PIcov	PCR-PIlen	PLS-PIcov	PLS-PIlen
100	20	1	5	0	0.9664	22.8287	0.9664	22.8287
100	20	1	10	0	0.9604	21.8894	0.9556	19.7716
100	20	1	50	0	0.9458	14.4248	0.9382	12.7511
100	20	19	5	$1/\sqrt{p}$	0.9674	22.6792	0.9674	22.6792
100	20	19	5	0	0.9664	22.7310	0.9664	22.7310
100	20	19	10	0	0.9654	25.1281	0.9586	19.7382
100	40	1	5	0.9	0.9676	22.5918	0.9508	18.1508
100	40	19	5	0	0.9716	26.0466	0.9444	18.3753
100	40	19	10	$1/\sqrt{p}$	0.9668	23.8813	0.9304	15.8560
100	40	19	20	$1/\sqrt{p}$	0.9666	23.6983	0.9364	15.1586
100	40	19	50	0	0.9596	23.5867	0.9194	14.1523
100	40	39	5	0.9	0.9704	22.6733	0.9482	18.2587
100	40	19	5	0.9	0.9678	22.6664	0.9486	18.1901
100	100	19	5	0.9	0.9652	22.6233	0.1998	4.3895
100	100	19	5	$1/\sqrt{p}$	0.9732	26.9037	0.1996	4.4913
200	20	19	10	0	0.9692	20.7071	0.9692	20.7071
200	40	39	5	0	0.9690	21.4478	0.9690	21.4478
200	40	39	10	0	0.9676	26.4430	0.9636	18.7647
200	200	19	5	0	0.9734	26.7558	0.0910	2.9437

Table 8.2. PI coverage and length for error type 5 (runs = 5000)

n	p	k	J	ψ	PCR-PIcov	PCR-PIlen	PLS-PIcov	PLS-PIlen
200	200	19	10	0	0.9680	26.5190	0.1872	3.7995
200	400	19	5	0.9	0.9704	21.5960	0.0000	0+
200	400	19	10	0	0.9726	27.2236	0.0012	0.0156
400	400	1	20	$1/\sqrt{p}$	0.9694	21.3949	0.1668	3.5956
400	40	1	20	$1/\sqrt{p}$	0.9676	21.3109	0.9666	20.2323
400	400	19	20	0.9	0.9686	21.2116	0.1674	3.5771
400	400	19	10	$1/\sqrt{p}$	0.9700	25.7260	0.0930	2.6704
400	400	399	5	0	0.9780	82.7513	0.0506	2.1254
400	800	1	5	0.9	0.9728	22.3380	0.0000	0+
1000	1000	1	5	0	0.9772	23.3596	0.0154	1.2894
1000	1000	999	10	$1/\sqrt{p}$	0.9716	22.7092	0.0364	1.6613
1000	2000	19	5	0.9	0.9736	23.2161	0.0000	0.1374

CHAPTER 9

CONCLUSIONS

0. When $n/J \geq p$, the method is doing a full OLS. In other words both PCR and PLS produce same coverage and lengths which are those of the OLS, as stated before.

1. When $p = n$, or $2n$, typically PLS coverage $\ll$ PCR coverage implying that PLS does not work for sufficiently large value of p. This was already stated before, Chun and Keleş (2010) show that PLS does not give a consistent estimator of $\boldsymbol{\beta}$ unless $p/n \rightarrow 0$. This can also be seen by the tables on previous pages, when $p = n$, or $2n$. Refer to table 9.1.

2. When $n > 2p$, $n/J < p$, and $k = 1$, PLS seems to work slightly better than PCR. This is seen by the coverage percentage and length, PCR gives us a longer coverage length. Refer to table 9.2.

3. When $n > 2p$, $n/J < k + 1$, or maybe when $k = p - 1$, PLS seems much more reliable than PCR. Refer to table 9.3 for examples. This is could also be due to the fact that $\psi = 0$ implying that there was no correlation between the predictors which usually lead to much longer PCR lengths than what was expected given the error types. Refer to table 9.3.

Table 9.1. Partial Least Squares ≪ Principal Component Regression, small n/p

n	p	k	J	ψ	error type	PCR-PIcov	PCR-PIlen	PLS-PIcov	PLS-PIlen
100	100	19	5	0.9	1	0.9894	5.9683	0.1976	1.3784
100	100	19	5	0.9	2	0.9782	10.1202	0.1986	2.2876
100	100	19	5	0.9	3	0.9860	5.9587	0.2032	1.3571
100	100	19	5	0.9	4	0.9920	3.5293	0.2010	0.8019
100	100	19	5	0.9	5	0.9652	22.6233	0.1998	4.3895
200	400	19	5	0.9	1	0.9856	5.7063	0.0000	0+
200	400	19	5	0.9	2	0.9764	9.5186	0.0000	0+
200	200	39	5	0	3	0.9732	26.2859	0.0964	0.9564
200	200	19	5	0	4	0.9768	18.2050	0.104	0.5671
400	400	19	10	$1/\sqrt{p}$	1	0.9756	18.0847	0.0960	0.8461
400	400	19	10	$1/\sqrt{p}$	2	0.9756	19.3178	0.0932	1.3996
400	800	1	5	0.9	3	0.9818	5.2500	0.0000	0+
400	800	1	5	0.9	4	0.9986	2.7852	0.0000	0+
1000	2000	19	10	0	1	0.9746	20.1530	0.0000	0+
1000	2000	19	10	0	2	0.9778	21.2955	0.0000	0+
1000	1000	1	5	0	3	0.9822	6.7610	0.0184	0.3947
1000	1000	19	5	0.9	4	0.9890	3.3199	0.0190	0.2291
2000	2000	19	10	0	1	0.9772	19.2329	0.0138	0.3587
2000	4000	19	20	0.9	2	0.9772	8.6916	0.0000	0.0960
2000	2000	19	50	0	3	0.9590	18.1872	0.0758	0.6978

Table 9.2. Partial Least Squares $\gg$ Principal Component Regression, $n/J < p$ and $k = 1$

n	p	k	J	ψ	error type	PCR-PIcov	PCR-PIlen	PLS-PIcov	PLS-PIlen
100	20	1	10	0	1	0.9800	6.1915	0.9648	4.9670
100	20	1	50	0	1	0.9636	6.1435	0.9352	4.3383
100	40	1	5	0.9	1	0.9856	5.7251	0.9308	4.9458
400	40	1	20	$1/\sqrt{p}$	1	0.9766	5.2346	0.9700	4.5699
100	20	1	10	0	2	0.9720	9.9857	0.9604	8.7199
100	20	1	50	0	2	0.9566	8.2785	0.9334	6.7660
100	40	1	5	0.9	2	0.9730	10.0105	0.9386	8.4738
400	40	1	20	$1/\sqrt{p}$	2	0.9740	8.7070	0.9686	8.1353
100	20	1	10	0	3	0.9746	6.2819	0.9684	4.9344
100	20	1	50	0	3	0.9564	6.0674	0.9364	4.1519
100	40	1	5	0.9	3	0.9806	5.6583	0.9358	5.0030
400	40	1	20	$1/\sqrt{p}$	3	0.9712	5.3134	0.9742	4.3537
100	20	1	10	0	4	0.9764	4.3736	0.9728	2.5878
100	20	1	50	0	4	0.9646	4.9268	0.9346	2.6719
100	40	1	5	0.9	4	0.9966	2.9863	0.9332	2.7193
400	40	1	20	$1/\sqrt{p}$	4	0.9694	3.5638	0.9698	2.2470
100	20	1	10	0	5	0.9604	21.8894	0.9556	19.7716
100	20	1	50	0	5	0.9458	14.4248	0.9382	12.7511
100	40	1	5	0.9	5	0.9676	22.5918	0.9508	18.1508
400	40	1	20	$1/\sqrt{p}$	5	0.9676	21.3109	0.9666	20.2323

Table 9.3. Partial Least Squares $\gg$ Principal Component Regression, $n/J < k + 1$ or maybe when $k = p - 1$

n	p	k	J	ψ	error type	PCR-PIcov	PCR-PIlen	PLS-PIcov	PLS-PIlen
100	20	19	10	0	1	0.9678	14.9895	0.9676	4.9840
200	40	39	10	0	1	0.9584	19.2529	0.9586	4.6827
100	20	19	10	0	2	0.9704	16.8327	0.9622	8.7501
200	40	39	10	0	2	0.9632	20.4525	0.9596	8.1193
100	20	19	10	0	3	0.9698	15.1213	0.9670	4.9244
200	40	39	10	0	3	0.9610	19.2659	0.9598	4.6056
100	20	19	10	0	4	0.9744	14.3614	0.9714	2.5855
200	40	39	10	0	4	0.9594	18.8393	0.9566	2.4349
100	20	19	10	0	5	0.9654	25.1281	0.9586	19.7382
200	40	39	10	0	5	0.9676	26.4430	0.9636	18.7647

Simulations were done in R. See R Core Team(2016).

REFERENCES

[1] Chun, H. and S. Keleş (2010), "Sparse Partial Least Squares Regression for Simultaneous Dimension Reduction and Predictor Selection," *Journal of the Royal Statistical Society, B,* 72, 325.

[2] Cook, R.D., Helland, I.S., and Su, Z. (2013), "Envelopes and Partial Least Squares Regression," *Journal of the Royal Statistical Society, B*, 75, 851-877.

[3] Denham, M.C., (1997), "Prediction Intervals in Partial Least Squares," *Journal of Chemometrics,* 11, 39-52.

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

[5] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An Introduction to Statistical Learning With Applications in R*, Springer, New York, NY.

[6] Johnson, R.A., and Wichern, D.W. (1988), *Applied Multivariate Statistical Analysis,* 2nd ed., Prentice Hall, Englewood Cliffs, NJ.

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

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

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

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

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

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

[13] Olive, D.J., and Hawkins, D.M. (2003), "Robust Regression with High Coverage,"

Statistics & Probability Letters, 63, 259-266.

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

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

[16] 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).

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

[18] Wold, H. (1975), "Soft Modelling by Latent Variables: the Nonlinear Partial Least Squares (NIPALS) Approach," in *Perspectives in Probability and Statistics, Papers in Honor of M.S. Bartlett,* ed. Gani, J., Academic Press, San Diego, CA, 117-144.

[19] Wold, H. (1985), "Partial Least Squares," *International Journal of Cardiology*, 147, 581-591.

[20] Wold, H. (2006), "Partial Least Squares," *Encyclopedia of Statistical Sciences*, Wiley, New York, NY.

VITA

Graduate School
Southern Illinois University

Sung-ho Kim

sunghkim418@gmail.com

Southern Illinois University
Bachelor of Science, Mathematics, August 2015

Research Paper Title:
Prediction Intervals for partial least squares and principal component regression Using D Variables

Major Professor: Dr. David J. Olive