A Sub-Model Theorem for Ordinary Least Squares

Lasanthi C. R. Pelawa Watagoda

Correspondence: Lasanthi C. R. Pelawa Watagoda, Department of Mathematical Sciences, Appalachian State University, Boone, NC, USA.

Received: October 7, 2018 Accepted: November 9, 2018 Online Published: November 19, 2018

doi:10.5539/ijsp.v8n1p40 URL: https://doi.org/10.5539/ijsp.v8n1p40

Abstract

Variable selection or subset selection is an important step in the process of model fitting. There are many ways to select the best subset of variables including Forward selection, Backward elimination, etcetera. Ordinary least squares (OLS) is one of the most commonly used methods of fitting the final model. Final sub-model can perform poorly if the variable selection process failed to choose the right number of variables. This paper gives a new theorem and a mathematical proof to illustrate the reason for the poor performances, when using the least squares method after variable selection.

Keywords: Multiple Linear Regression, Variance Co-variance Matrix of the OLS

1. Introduction

The use of OLS for multiple linear regression models after variable selection can results in poor models. First, we describe the Multiple Linear Regression (MLR) model in section 1. In section 2, we discuss the variable selection and in section 3 we introduce a new theorem and its proof to illustrate the reason for the poor performances of some OLS sub-models. This paper closely follows the author's related work Pelawa Watagoda (2017), Pelawa Watagoda and Olive (2018), Pelawa Watagoda and Olive (2018a).

1.1 Multiple Linear Regression Model

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

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

for $i = 1,...,n$. This model is also called the full model. Here n is the sample size, and assume that the random variables e_i are independent and identically distributed (iid) with variance $V(e_i) = \sigma^2$.

In matrix notation, these n equations become

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{2}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of response 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.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & x_{13} & \dots & x_{1p} \\ 1 & x_{22} & x_{23} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & x_{n3} & \dots & x_{np} \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \tag{3}$$

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.

It is often convenient to use the centered response $\boldsymbol{Z} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$ where $\overline{\boldsymbol{Y}} = \overline{Y}\boldsymbol{1}$, 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 $\sum_{i=1}^n W_{ij}^2 = n$. Note that the sample correlation matrix of the nontrivial predictors $\boldsymbol{u}_i$ is $\boldsymbol{R_u} = \boldsymbol{W}^T\boldsymbol{W}/n$. Then regression through the origin is used for the model

$$\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e} \tag{4}$$

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

There are many methods for estimating $\boldsymbol{\beta}$, including forward selection with OLS, principal components regression (PCR), partial least squares (PLS) due to Wold (1975), lasso due to Tibshirani (1996), and ridge regression (RR): see Hoerl and

Kennard (1970). Also, there are methods like variant of relaxed lasso that applies OLS to a constant and the predictors that had nonzero lasso coefficients, which is the LARS-OLS hybrid estimator of Efron et al. (2004), also called the relaxed lasso ($\phi = 0$) estimator by Meinshausen (2007).

These six methods produce M models and use a criterion to select the final model (e.g., C_p or 10-fold cross validation (CV)). The number of models M depends on the method. Lasso and ridge regression have a parameter λ. When $\lambda = 0$, the OLS full model is used. These methods also use a maximum value λ_M of λ and a grid of M λ values $0 \leq \lambda_1 < \lambda_2 < \cdots < \lambda_{M-1} < \lambda_M$ where often $\lambda_1 = 0$. For lasso, λ_M is the smallest value of λ such that $\hat{\boldsymbol{\eta}}_{\lambda_M} = \mathbf{0}$. Hence $\hat{\boldsymbol{\eta}}_{\lambda_i} \neq \mathbf{0}$ for $i < M$. For forward selection, PCR, and PLS, $M \leq p$. See James et al. (2013, ch. 6).

Consider choosing $\hat{\boldsymbol{\eta}}$ to minimize the criterion

$$Q(\boldsymbol{\eta}) = \frac{1}{a}(\boldsymbol{Z} - \boldsymbol{W\eta})^T(\boldsymbol{Z} - \boldsymbol{W\eta}) + \frac{\lambda_{1,n}}{a}\sum_{i=1}^{p-1}|\eta_i|^j \tag{5}$$

where $\lambda_{1,n} \geq 0$, $a > 0$, and $j > 0$ are known constants. Then $j = 2$ corresponds to ridge regression, $j = 1$ corresponds to lasso, and $a = 1, 2, n$, and $2n$ are common. The residual sum of squares $RSS(\boldsymbol{\eta}) = (\boldsymbol{Z} - \boldsymbol{W\eta})^T(\boldsymbol{Z} - \boldsymbol{W\eta})$, and $\lambda_{1,n} = 0$ corresponds to the OLS estimator $\hat{\boldsymbol{\eta}}_{OLS} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$.

2. Variable Selection

Variable selection is the search for a subset of predictor variables that can be deleted with little loss of information if n/p is large, and so that the model with the remaining predictors is useful for prediction. Following Olive and Hawkins (2005), a *model for variable selection* can be described by

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S \tag{6}$$

where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$, $\boldsymbol{x}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p - a_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \mathbf{0}$ 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 a terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). Suppose that S is a subset of I and that model (5) holds. Then

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_{I/S}^T\boldsymbol{\beta}_{(I/S)} + \boldsymbol{x}_O^T\mathbf{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I \tag{7}$$

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 $S \subseteq I$.

Forward selection forms a sequence of submodels $I_1, ..., I_M$ where I_j uses j predictors including the constant. Let I_1 use $x_1^* = x_1 \equiv 1$: the model has a constant but no nontrivial predictors. To form I_2, consider all models I with two predictors including x_1^*. Compute $Q_2(I) = SSE(I) = RSS(I) = \boldsymbol{r}^T(I)\boldsymbol{r}(I) = \sum_{i=1}^n r_i^2(I) = \sum_{i=1}^n (Y_i - \hat{Y}_i(I))^2$. Let I_2 minimize $Q_2(I)$ for the $p - 1$ models I that contain x_1^* and one other predictor. Denote the predictors in I_2 by x_1^*, x_2^*. In general, to form I_j consider all models I with j predictors including variables $x_1^*, ..., x_{j-1}^*$. Compute $Q_j(I) = \boldsymbol{r}^T(I)\boldsymbol{r}(I) = \sum_{i=1}^n r_i^2(I) = \sum_{i=1}^n (Y_i - \hat{Y}_i(I))^2$. Let I_j minimize $Q_j(I)$ for the $p - j + 1$ models I that contain $x_1^*, ..., x_{j-1}^*$ and one other predictor not already selected. Denote the predictors in I_j by $x_1^*, ..., x_j^*$. Continue in this manner for $j = 2, ..., M$. 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., $\lceil 7.7 \rceil = 8$.

Consider the six methods forward selection with OLS, PCR, PLS, lasso, relaxed lasso, and ridge regression. When there is a sequence of M submodels, the final submodel I_d needs to be selected. Let the candidate model I contain a terms, including a constant. For example, let $\boldsymbol{x}_I$ and $\hat{\boldsymbol{\beta}}_I$ be $a \times 1$ vectors for the methods excluding PCR and PLS. Then there are many criteria used to select the final submodel I_d.

3. OLS Sub Model Theorem and Proof

This section will prove Theorem 1 bellow and discuss its implications.

Theorem 1. *Suppose the usual linear model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$, with $E(\boldsymbol{Y}) = \boldsymbol{X\beta}$ and $E(\boldsymbol{e}) = \mathbf{0}$.*

Where, $Cov(\boldsymbol{Y}) = Cov(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$.

Define $\boldsymbol{X}$ and $\boldsymbol{\beta}$ as follows;

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_I & \boldsymbol{X}_o \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_I \\ \boldsymbol{\beta}_o \end{bmatrix},$$

Where $\boldsymbol{X_I}$ is the vector of a terms from a candidate subset indexed by I, $\boldsymbol{X_o}$ is vector of the predictors that are out of the candidate submodel and if $S \nsubseteq I$ then,

$$E(\hat{\boldsymbol{\beta}}_I) = \boldsymbol{\beta}_I + \left[\boldsymbol{X}_I^T \boldsymbol{X}_I\right]^{-1} \boldsymbol{X}_I^T \boldsymbol{X}_o \boldsymbol{\beta}_o$$

and

$$Cov(\hat{\boldsymbol{\beta}}_I) = \sigma^2 \left(\boldsymbol{X}_I^T \boldsymbol{X}_I\right)^{-1}.$$

Proof. Assume this is an arbitrary submodel, and I does not contain S. Then,

$\boldsymbol{X\beta} = \boldsymbol{X_I \beta_I} + \boldsymbol{X_o \beta_o}$ where $\boldsymbol{\beta}_I = \left[\boldsymbol{X}_I^T \boldsymbol{X}_I\right]^{-1} \boldsymbol{X}_I^T \boldsymbol{Y} = \boldsymbol{AY}$.

Now consider the expected value of $\hat{\boldsymbol{\beta}}_I$,

$$\begin{aligned}
E(\hat{\boldsymbol{\beta}}_I) &= E\left(\left[\boldsymbol{X}_I^T \boldsymbol{X}_I\right]^{-1} \boldsymbol{X}_I^T \boldsymbol{Y}\right) \\
&= E(\boldsymbol{AY}) \\
&= \boldsymbol{A}E(\boldsymbol{Y}) = \boldsymbol{AX\beta} \\
&= \boldsymbol{A}(\boldsymbol{X_I \beta_I} + \boldsymbol{X_o \beta_o}) \\
&= \left[\boldsymbol{X}_I^T \boldsymbol{X}_I\right]^{-1} \boldsymbol{X}_I^T (\boldsymbol{X_I \beta_I} + \boldsymbol{X_o \beta_o}) \\
&= \left[\boldsymbol{X}_I^T \boldsymbol{X}_I\right]^{-1} \boldsymbol{X}_I^T \boldsymbol{X_I \beta_I} + \left[\boldsymbol{X}_I^T \boldsymbol{X}_I\right]^{-1} \boldsymbol{X}_I^T \boldsymbol{X_o \beta_o} \\
&= \boldsymbol{\beta}_I + \left[\boldsymbol{X}_I^T \boldsymbol{X}_I\right]^{-1} \boldsymbol{X}_I^T \boldsymbol{X_o \beta_o} \neq \boldsymbol{\beta}_I
\end{aligned}$$

Now consider $Cov(\hat{\boldsymbol{\beta}}_I)$,

$$\begin{aligned}
Cov(\hat{\boldsymbol{\beta}}_I) &= Cov(\boldsymbol{AY}) = \boldsymbol{A}Cov(\boldsymbol{Y})\boldsymbol{A}^T \\
&= \boldsymbol{A}\sigma^2 \boldsymbol{I}\boldsymbol{A}^T = \sigma^2 \boldsymbol{AA}^T \\
&= \sigma^2 \left(\left[\boldsymbol{X}_I^T \boldsymbol{X}_I\right]^{-1} \boldsymbol{X}_I^T\right)\left(\left[\boldsymbol{X}_I^T \boldsymbol{X}_I\right]^{-1} \boldsymbol{X}_I^T\right)^T \\
&= \sigma^2 \left[\boldsymbol{X}_I^T \boldsymbol{X}_I\right]^{-1} \boldsymbol{X}_I^T \boldsymbol{X}_I \left(\left[\boldsymbol{X}_I^T \boldsymbol{X}_I\right]^{-1}\right)^T \\
&= \sigma^2 \left(\left[\boldsymbol{X}_I^T \boldsymbol{X}_I\right]^T\right)^{-1} \\
&= \sigma^2 \left(\boldsymbol{X}_I^T \boldsymbol{X}_I\right)^{-1}.
\end{aligned}$$

□

According to Theorem 1, when $S \nsubseteq I$, i.e. when the final submodel does not contain enough predictors, the $E(\hat{\boldsymbol{\beta}}_I) \neq \boldsymbol{\beta}_I$, and will produce a poor final submodel. On the other hand, following equations 6 and 7, when the submodel contains the set of predictors $S, \boldsymbol{\beta}_0 = 0$. Then $E(\hat{\boldsymbol{\beta}}_I) = \boldsymbol{\beta}_I + \left[\boldsymbol{X}_I^T \boldsymbol{X}_I\right]^{-1} \boldsymbol{X}_I^T \boldsymbol{X_o \beta_o} = \hat{\boldsymbol{\beta}}_I$.

4. Conclusions

This worked mathematically showed the reason for ordinary least squares to perform poorly when the submodel does not contain enough predictors.

Acknowledgements

Author thanks Dr. David J. Olive and Dr. Hasthika S. Rupasinghe Arachchige Don for guidance and support delivered to successfully complete this paper.

References

Efron B., Hastie, T., Johnstone, I., & Tibshirani R. (2004). Least Angle Regression. *The Annals of Statistics, 32*(2), 407-451.

Hoerl, A. E., & Kennard, R. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics, 12*, 55-67.

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

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

Pelawa, W. L. C. R. (2017). *Inference After Variable Selection. (PhD Thesis)*, Southern Illinois University, USA, at (http://lagrange.math.siu.edu/Olive/slasanthiphd.pdf).

Pelawa, W. L. C. R., & Olive, D. J. (2018). Inference For Multiple Linear Regression After Model or Variable Selection (Preprint) (http://lagrange.math.siu.edu/Olive/ppvsinf.pdf)

Pelawa, W. L. C. R., & Olive, D. J. (2018a). Comparing Shrinkage Estimators With Asymptotically Optimal Prediction Intervals (Preprint) (http://lagrange.math.siu.edu/Olive/pppicomp.pdf)

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, B*(58), 267-288.

Wold, H. (1975). Soft Modelling by Latent Variables: the Nonlinear Partial Least Squares (NIPALS) Approach. *Perspectives in Probability and Statistics, Papers in Honor of M.S. Bartlett, ed.* 117-144.

Copyrights