Prediction Intervals of Response Variables based on Quantiles in High Dimensional Regression Analyses

Septian Rahardiantoro*, Khairil Anwar Notodiputro, Anang Kurnia
Department of Statistics, Bogor Agricultural University
*email: rahardiantoro_14@apps.ipb.ac.id

Abstract  In regression analyses it is of interest to obtain prediction intervals of the response variables. However, such prediction intervals are not obvious if the number of explanatory variables exceeds the number of observations since the least square method cannot be used in this case. This paper discusses the problems of constructing prediction intervals in high dimensional regression models, in which the number of explanatory variables is greater than the number of observations. A quantile approach is proposed to construct such intervals and it has been evaluated by means of simulation. In this approach pairs of quantiles based on the certain probability are specified and followed by evaluation to obtain the shortest interval. Since the number of explanatory variables was large then several techniques to select the variables were employed. These techniques were the best subset regression, LASSO (Least Absolute Shrinkage and Selection Operator) regression, and model averaging. The simulation data was generated according to two different scenarios. The first scenario was designed for models having symmetric error distributions whereas the second scenario was designed for models with non-symmetric error distributions. The simulation results showed that in the case of symmetric error distributions all of the regression methods mentioned above produced similar prediction intervals, except the LASSO regression. However, in the case of non-symmetric error distributions it has been evidence that model averaging has provided the best prediction intervals when compared with the best subset and LASSO regressions although has wide range of intervals. This revealed that model averaging can be used to predict the response variables in high-dimensional regression analyses although the data is non-symmetrically distributed.

Keywords: best subset regression, error distribution, high dimensional data, LASSO, model averaging, skewed distribution, variable selection.

1. Introduction
High-dimensional data occurs when the number of predictor variables in modeling context exceeds the number of data observations. In regression analysis, there are some methods developed for handling high-dimensional data, such as best subset regression, LASSO (Least Absolute Shrinkage and Selection Operator) method, and also model averaging method. One of the most reasons using regression analysis is to get the prediction of response variable accurately.
The predictor variables in the data are used to predict the response of variable, which is in high-dimension condition, the researchers can use all of predictor variables, or the subset of predictor variables. Best subset regression has the idea to get the subset of predictor variables that give the best model for predicting the response variable based on some criterion values [1]. Different with best subset regression, LASSO and model averaging method use all of predictor variables to predict the response variable. LASSO shrinks some coefficients of predictor variables and sets others to 0 in application of regression process [2]. In the model averaging method, a class of candidate models with varying degrees of model complexity is specified first, than this approach advocates the pooling of predictions by giving higher weights to the better models [3].

This research focused on specifying the prediction intervals of response variable using three methods above. The propose approach that applied in this research for getting the prediction intervals is specified quartile approach. This approach has the main purpose to get the shortest interval in specified probabilities [4]. By simulation study in a thousand replications, this research also has the purpose to reveal the recommended method that have the best prediction intervals in two different scenarios, data have symmetric error distributions and non-symmetric error distributions. In practice, this research is using R software in implementing all of process above in high-dimensional data condition.

2. Materials
This research use the simulation study which generate the high-dimensional data in application. The high-dimensional data specified with number of predictor variables \( p \) 1000, and number of data observations \( n \) 100. It is designed by following the multivariate normal distribution with mean vector \( \mathbf{0} \), and variance-covariance matrix \( \mathbf{S} = \mathbf{I}_p \). The number of true regressors \( s = 50 \) and the true regressors \( x_i \) be spaced evenly, \( i = 20(j - 1) + 1, j = 1,2,...,50 \). In addition, the intercept coefficient of this data condition is set \( \beta_0 = 100 \).

The two scenarios of this research are implemented by generating the response variable that follows the symmetric distribution and non-symmetric distribution. In symmetric case, the normal distribution chosen, therefore the response variable follows normal distribution with expected value based on the data, and variance equal to 1. The log-normal distribution is used to present the non-symmetric condition of response variable distribution. In this scenario, the response variable is set to follow log-normal distribution with expected value based on the data, and the variance equal to 1.

Using the data that have been generated, the three methods; best subset regression, LASSO, and model averaging method applied. In practice, all of these process using 1000 replications to identify the prediction intervals of response variable with the shortest quantile approach. In addition, the simulation taken by dividing the data set to be two parts, training set and testing set. The modeling process is using training set, and the prediction process is using testing set. Training and testing set in this research determined 50% of data, therefore both training and testing set have 50 observations.

3. Methods
In this part would be described the methods to handle the high-dimensional data in regression cases, and also the propose method to specify the prediction intervals using quantile approach.

a. Best Subset Regression
The main purpose using best subset regression is to find the best model that has the most optimum criterion values. This method also could be used when the number of predictors more than number of observation, to verify the best model using some of predictors \( (q \) predictors). In best subset regression process, model specified need to be evaluated by \( R^2 \), \( S \), and \( C_p \) mallow statistics. The best model identified when it has the highest value of \( R^2 \), the lowest value of \( S \), and the value of \( C_p \) mallow statistics almost equivalent with the number of predictor variables [1].
b. LASSO Regression
For definition of LASSO [2], as in the usual regression set-up, assume that the observations are independent or that the $y_i$s are conditionally independent given the $x_{ij}$s, here also assume that the $x_{ij}$ are standardized that fulfill $\sum_i x_{ij}/N = 0, \sum x_{ij}^2/N = 1$.

Letting $\beta = (\beta_1, \beta_2, ..., \beta_p)$, the LASSO estimate $(\hat{a}, \hat{\beta})$ is defined by

$$
(\hat{a}, \hat{\beta}) = \arg \min \left\{ \sum_{i=1}^N \left( y_i - a - \sum_j \beta_j x_{ij} \right)^2 \right\}
$$

subject to $\sum |\beta_j| \leq t$. Here $t \geq 0$ is a tuning parameter. Now, for all $t$, the solution for $\alpha$ is $\hat{a} = \bar{y}$. The parameter $t \geq 0$ controls the amount of shrinkage that is applied to estimates. Let $\beta^0_j$ be the full least squares estimates and let $t_0 = \sum |\beta^0_j|$. Values of $t < t_0$ will cause shrinkage of the solutions towards 0, and some coefficients may be exactly equal to 0.

c. Model Averaging
Model averaging involves an attempt to combine several competing scientific models. Different models may be proposed by different research teams with possibly different ideas, computing systems, and background knowledge. Almost studies on model averaging have focused on situations in which the total number of predictors $p$ is much smaller than the observation size $n$.

Assume there is a matrix $X_{(nxp)}$, that indicates the matrix of predictor variables with $p >> n$. The first step is ordering every predictor variables by taking the marginal correlation between every predictor variables with the response variables. The matrix result of this process is called $\hat{X}_{(nxp)}$. Then the $k$ partitions taken from $\hat{X}_{(nxp)}$, that is in every partition has $m$ predictor variables. The OLS regression model of the response variable with the predictor variables matrix partition called model candidate. Finally, from each model candidates, the predictions can be specified, and the last step is taking weighted averaging of that [3]. This approach is mainly motivated by the attempt to address the dimensionality issue encountered in regression problems with $p >> n$. There are two steps involved, preparation the candidate models and optimize the model weights. In this research, the weight of models use AIC weight, also the specified model candidates use number of $k = 40$ and $m = 25$.

d. Prediction Intervals Approach
The method to specify the prediction intervals is using specified quartile approach. The main idea of this approach is taking some pairs of quartile and then it would be evaluated to get the shortest interval based on the probability used. The following parts would be described the algorithm of the method [5].

The algorithm of the method:
1. Define a probability of prediction intervals ($p_b$). This probability indicates the accuracy of intervals designed. In this study, the probability is 0.95, therefore there is 0.05 probability outside the intervals, that means 5% toleration of errors of predictions.
2. Define pairs of number of quantiles, $(Q_0, Q_{0.95})$, $(Q_{0.005}, Q_{0.995})$, $(Q_{0.01}, Q_{0.99})$, $(Q_{0.015}, Q_{0.985})$, $(Q_{0.02}, Q_{0.97})$, $(Q_{0.025}, Q_{0.975})$, $(Q_{0.03}, Q_{0.98})$, $(Q_{0.035}, Q_{0.985})$, $(Q_{0.04}, Q_{0.99})$, $(Q_{0.045}, Q_{0.995})$, $(Q_{0.05}, Q_{1})$. The value of each quartile calculated using the following formula:

$$Q_{pb} = \arg \sum_{i=1}^Z l_{x_i \leq Q} = pb$$

3. Count the difference of each pairs of quantiles to find the shortest difference that to be the solution of intervals.
4. Results and Discussion

Based on 1000 replications, Figure 1 demonstrates the result of prediction intervals of each individual observation using the shortest quantile approach in first scenario from the testing set. Based on the Figure, best subset regression and model averaging method have the prediction intervals that contain the all of actual values from the response variable. While, there are some actual values of the response variable outside of the prediction intervals in the LASSO method. In this case, the LASSO method seems less accurate to determine the prediction intervals using high-dimensional data, while the distribution of response variable have symmetric tailed distribution.

![Figure 1. Prediction intervals in first scenario using (a) best subset regression, (b) LASSO, and (c) model averaging method](image)

Different with the Figure 1, the following graphs in Figure 2 demonstrate the result of prediction intervals each of observations in testing set in second scenario. From 1000 replications, the prediction intervals result of best subset regression have the accuracy around 98%. It is very good result to determine the prediction of response variable in context the response variable have non-symmetric distribution. The prediction intervals of LASSO method seem not too good in this scenario. There are a lot of observations predicted not accurately. In addition, the model averaging method performs very good prediction intervals, while the intervals determined are wide enough.

![Figure 2. Prediction intervals in second scenario using (a) best subset regression, (b) LASSO, and (c) model averaging method](image)
Based on the result of prediction intervals above, best subset regression and model averaging have a good performance in the predicting case using high-dimensional data in both criterias. The model averaging method has the best prediction intervals, although the prediction intervals result in second criteria have very wide range. It is also can be concluded that LASSO method unpreferable method to predict the response variable.

5. Conclusion
The conclusions of this research are the shortest quantile approach would be a good alternative to specify the prediction intervals of prediction process in regression analysis. Based on the simulation process, the methods that have good performance to perform the prediction intervals are best subset regression and model averaging method, which model averaging has very good accuracy. This revealed that model averaging can be used to predict the response variables in high-dimensional regression analyses when the data is non-symmetrically distributed, although have wide range intervals.

References
[1] Myers RH 1990 Classical and Modern Regression with Applications. PWS-Kent: USA
[2] Tibshirani R 1996 Regression Shrinkage and Selection via the LASSO J. Royal Statistics Soc Series B 58 267-288
[3] Ando T. and Li KC 2014 A Model-Averaging Approach for High-Dimensional Regression J. Amer Statistic Assoc. 194 254-265
[4] Casella G & Berger RL (2002) Statistical Inference Second Edition. Duxbury Thomson Learning.
[5] Rahardiantoro S, Notodiputro KA, Kurnia A 2017 Prediction Intervals of Model Averaging Methods for High – Dimensional Data. Proceedings 2nd ISI Regional Statistics Conference 20-24 March 2017 Indonesia (Session CPS09)