|  |
| --- |
| Class Project |

|  |
| --- |
| Xinjie Li & Baoluo Sun  4-17-2020 |

**Model Selection – Different Validation Approaches**

**Abstract**

When we build a model, it is very important for us to examine the model performance in order to select the best model. The method is called “validation”, and there are different methods to examine it, such as cross validation or bootstrapping. By building various models and performing validations, we find the optimal model that could best illustrate the relationship between house pricing in Boston area and 9 key variables.

**Introduction**

Every time when we have a dataset, we want to find some patterns or features in the datasets. After performing some unsupervised learnings, patterns or relationships appeared. Next, we train the data set and fit the best model that can explain the data well. In selection of the model, we first need to choose which type of models should be used, such as linear regression, logistic regression. KNN or decision. In this experiment, we are going to perform linear regression on “Boston Housing Price” to show the procedure of training and validation the model.

The model should not only have an optimal low training error but also a low testing error. The optimal low training error does not necessarily mean the model with lowest training among other models. The selected model needs to have a tradeoff between bias and variance, which we can examine it by comparing testing MSE.

There are various techniques to do the validation. In the experiment, we have the four validation methods and by comparing the results, to select the best model that can be fitted into housing price equation. They are different how to resample our data set and how to divide them into training and testing set.

This first method is validation approach, which we directly divide the whole dataset into training set and validation set. In this way, by only create the model based on training dataset and evaluating the model on testing set, it can show the performance of various model. Even though this method is the most simple and direct way to validate, there are a few problems involved. Such as there will be only one training and validating set, which can reduce the accuracy of model.

The second method is leave on out cross validation approach (LOOCV). As the names suggested, for each time we only leave one out for testing and the rest for training. There are totally n times (n is totally number of data) needed to finish the validation. One advantage of this method is with less bias, since the model is trained base almost entire population. However, it also results higher variance, which the model is overfitting, and time consuming that we tried n times.

The third method is K fold cross validation. The dataset is first separate into k subsets, and each subset’s size is n/k. We trained the model on k-1 subsets and test on the last k subsets. In this way, we only need to finish validation in k times, which in much less time than LOOCV and the variance is lower, which the reduce the chance of overfitting.

The last method is bootstrapping. This method is totally different from the previous three, which instead of splitting the dataset, this method is doing resampling. Bootstrap randomly select number of samples from the dataset with replacement. Since we are doing resampling, one of big advantage is this method does not require large dataset, but there is significant overlap between the “new” training and validation sets.

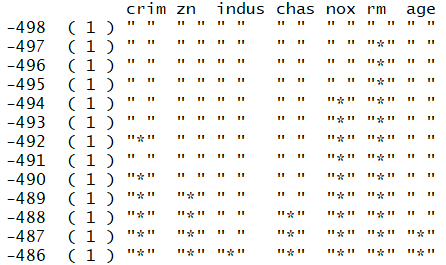
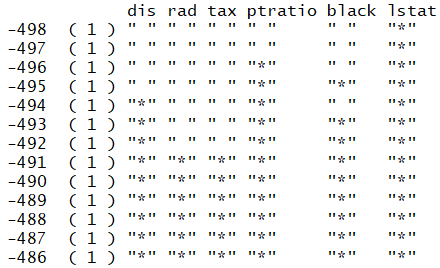
**Materials and Methods**

**1. Cleaned the dataset**

The “Housing Values in Suburbs of Boston” data from R database under Mass packages is being used to examine the validation methods. This database contains Boston housing price and 13. They are crime, proportion of residential land, proportion of non-retail business acres per town, whether close to Charles river or not, nitrogen oxides concentration, average number of rooms per dwelling, proportion of owner-occupied units built prior to 1940, mean distances to five Boston employment centers, index of accessibility to radial highways, property tax rate, pupil teacher ratio, proportion of blacks and lower status of population.

We first clean the whole database, by delete all “NA” values, and then delete all outliers that are greater than 3 standard deviations. By doing the linear regression of both before and after cleaned dataset, the adjusted-R squared does show an improvement, which increases from 0.7337897 to 0.7974304.

After first step cleaned the database, we then use the best subset selection method to pick the variable that would have significantly influence on the house price, in this way, we could decrease the possibility of overfitting the model by reduce the number of variables. As the result is showed in below, we decide to delete the proportion of residential land, proportion of non-retail business acres per town, whether close to Charles river or not and proportion of owner-occupied units built prior to 1940 and only leave the rest of 9 variables for next.

**2.Model Fitting**

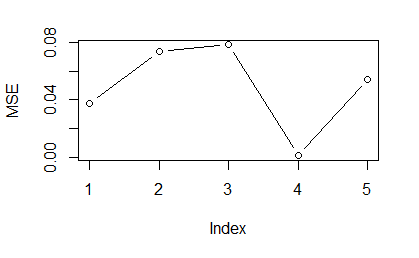
In choosing the model, we first decide to use linear regression model to predict housing price with the rest of 9 variables. Since there are too many variables, the variables with only one degree, which is a straight line, may not perfectly fit the overall dataset. There are few ways to fix the underfitting problem is to increase variables, increase variables’ degree, or using other methods such as logistic regress, random forest etc. In this experiment, we designed 5 models with different degrees. The function in R is called “poly” which helps us create models from first degree up to fifth degree.

**3. Model Selection**

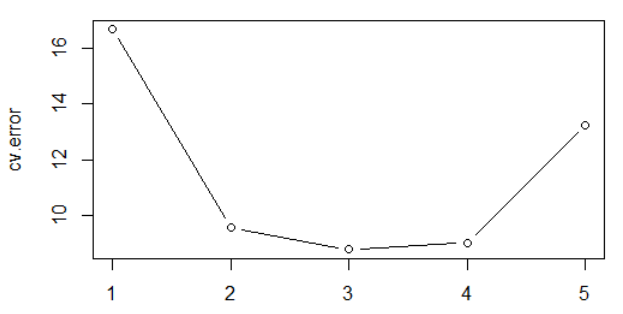
In choosing the model, we need to have a trade off between bias and variance. This process involved validation. We used four validation methods, and by combing the result to select the optimal model. For the validation approach, we split the data equally into two parts, one for training, and the other one for test. After each time finishing training, we then use R “predict” function to test the model. And by calculating the mean squared error to find the model with lowest training error. For LOOCV and K-folds, we use “cv.glm” to calculate the cross-validation error. In k-folds approach, we choose k=15 and k=20. The results are showed in the figures, which x-axis is degree, and y-axis is cross validation error, and use “which.min” method to find the model with the lowest training error. For the last one, bootstrapping, we used the library “boot” and implemented its function “boot.fn” to do the resample. The “boot” function shows the original, bias and standard error for each variable.

**Results**

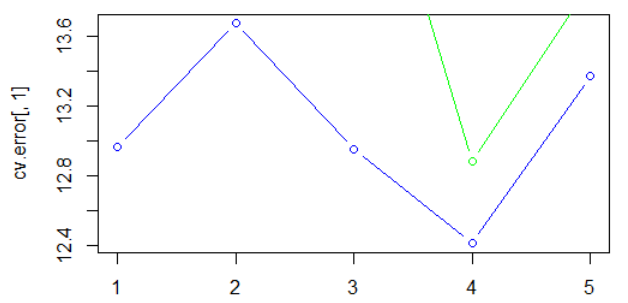
From the validation approach, we find that the model with degree 4 shows the lowest MSE.

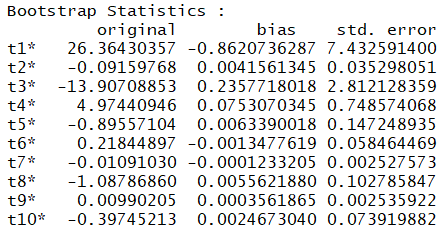


In LOOCV validation approach, we find that the model with degree 3 shows the lowest MSE.



In K-fold, the green is k=15, and blue is k=20. The lowest cross-validation error is at degree 4.



In bootstrap, the model with degree 4 show the best trade-off between bias and variance. 

**Discussion**

The first step by cleaning and delete a few unnecessary variables can help the model better fit on the data set. The next is evaluating the result on each model. All the four validation methods except LOOCV are pointing that the model with degree 4 has the lowest error or best trade-off between bias and variance. Even in LOOCV, the minimal cross validation error is at degree 3, the degree 4’s error is only a slightly larger than degree 3’s. After degree 5, the error goes up again, which mean the model increases it bias or variance. Therefore, we can conclude the model with degree 4 can well explain how the housing price in Boston area is affected by those 9 variables. In the further research, instead of increasing the all variables’ degree at the same time and same amount, we can train the model with different degree on different variable. Furthermore, we can also compare linear regression model with random forest or other regression methodologies and use validation approach to select the optimal model.

**Literature Cited**

C. Tantithamthavorn, S. McIntosh, A. E. Hassan and K. Matsumoto, "An Empirical Comparison of Model Validation Techniques for Defect Prediction Models," in IEEE Transactions on Software *Engineering*, vol. 43, no. 1, pp. 1-18, 1 Jan. 2017.

Max Kuhn, Kjell Johnson. 2018. Applied Predictive Modeling

John Fox, ” Bootstrapping Regression Models,” in *Stanford University statweb, Jan 2002*

Belsley D.A., Kuh, E. and Welsch, R.E. (1980) Regression Diagnostics. Identifying Influential Data and Sources of Collinearity. New York: Wiley.

Harrison, D. and Rubinfeld, D.L. (1978) Hedonic prices and the demand for clean air. J. Environ. Economics and Management **5**, 81–102.