International Journal of Engineering & Technology, 7 (4.30) (2018) 498-501

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Modelling House Price Using Ridge Regression and Lasso Regression

Seng Jia Xin[1], Kamil Khalid[2*]

[1,2]Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology,
Universiti Tun Hussein Onn Malaysia, Pagoh Educational Hub, 84600 Pagoh, Muar, Johor, Malaysia
**Corresponding author E-mail: kamil@uthm.edu.my*

Abstract

House price prediction is important for the government, finance company, real estate sector and also the house owner. The data of the house price at Ames, Iowa in United State which from the year 2006 to 2010 is used for multivariate analysis. However, multicollinearity is commonly occurred in the multivariate analysis and gives a serious effect to the model. Therefore, in this study investigates the performance of the Ridge regression model and Lasso regression model as both regressions can deal with multicollinearity. Ridge regression model and Lasso regression model are constructed and compared. The root mean square error (RMSE) and adjusted R-squared are used to evaluate the performance of the models. This comparative study found that the Lasso regression model is performing better compared to the Ridge regression model. Based on this analysis, the selected variables includes the aspect of house size, age of house, condition of house and also the location of the house.

Keywords: *Adjusted R-squared; Lasso Regression; Ridge Regression; Root mean square error (RMSE)*

1. Introduction

In the year of 1990s, the house price in United State has increased sharply and from the year onward the price has been increasing for seven percent at the nation level annually [1]. The bubble of house price has brought up a negative growth of the economy in United State since the people do not afford to buy a house. Therefore, it is significant to study the growth of the house price as the pattern and the fluctuation of the data of house price which will provide the important information to the people who related [2].

The house price often can be given a great impact by the factors. It is a multidimensional study in factors that contribute a great impact on the house pricing [3]. Since there are numerous factors affect the prices of the house, the relationship between those factors is complex and a mathematical model can be seen to summarize the relationship of the influential factors and the house price. With the house price model, it will contribute important information to the real estate market which will indirectly influence the economic growth of the country [4].

Multicollinearity is commonly occur in the high dimensional data where it will give a serious problem whenever analysis using multiple regression [5, 6, 7]. Therefore, Ridge regression and Lasso regression can be used to cope the multicollinearity problem [8, 9]. In this study, Ridge regression model and Lasso regression model are constructed to predict the house price in the United State. This will provide the important information to government, financial firm such as bank, or even a houseowner. Besides that, the constructed house price model can improve the growth of the real estate market [10]. There are quite considerable studies were carried out to study the application of statistics [11, 12, 13].

2. Material and Method

Two methods are used to construct the house price model which are Ridge regression and Lasso regression. Before constructing the model, data pre-processing is necessary to be carried out. Data pre-processing is used to clean up the outliers and handle the missing value in the data. The outliers will be removed from the observations and the missing value for the numeric variables will be replaced by zero while for the missing value of categorical variables will be replaced by none. Then the data has been split into two part which train data set for constructing the models and test data set is used for the model validation. The performance of the Ridge regression and Lasso regression are evaluated based on the root mean square error and adjusted R-squared.

2.1. Ridge Regression

Ridge regression is modifying the least squares method which to allow to have biased estimators of the regression coefficients in the regression model. Ridge regression put a particular form of constraints on parameters. $\hat{\beta}^{ridge}$ which to be used in minimizing the penalized sum of squares in the equation 3.1 [14].

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{k}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{k}\beta_j^2 \tag{3.1}$$

where y_i is the value of the response variable in the i^{th} trial, β_0 is the intercept coefficient, β_j is the regression coefficient for j = 1,…,k, x_{ij} is the j^{th} component of x_i which x_i is a known constant namely the value of the predictor variable in the i^{th} trial, λ is the constant value which shows the degree of bias in the estimators.

2.2. Lasso Regression

Least absolute shrinkage and selection operator which the short form name as Lasso. Lasso can be useful in estimating the regression coefficient and performing variable selection. This similar with the Ridge regression. However, there is an important characteristic which the coefficient of the Lasso regression can set to zero that this phenomenon will not happen in Ridge regression. The Lasso estimate minimizes the penalized sum of squares in the Equation 3.2 [14].

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{k}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{k}\left|\beta_j\right| \tag{3.2}$$

2.3. *K*-fold cross-validation

K-fold cross-validation is a method that use in estimating prediction error and smoothing parameters. The initial data will divide randomly into k mutually exclusive subsets which must be in an equal size. The prediction error of cross validation gives in the Equation 3.3 and Equation 3.4 [15].

$$\mathrm{CV}_k^{(\lambda)}=\frac{1}{n_k}\sum_{i\in J(k)}\left[y_i-\hat{f}_{-k}^{(\lambda)}(x_i)\right]^2 \tag{3.3}$$

$$\mathrm{CV}^{(\lambda)}=\frac{1}{K}\sum_{k=1}^{K}\mathrm{CV}_k^{(\lambda)} \tag{3.4}$$

From the equation, $y=\left[y_1,\ldots,y_K\right]$ represents one of a K-fold split of the $n\times 1$ data vector y. $J(k)$ is the set of element of $\{1,2,\ldots,n\}$ that correspond to the indices of data points within split k. $n_k=\left|J(k)\right|$ representing the number of element within split k with portion y_k removed. $\hat{f}_{-k}^{(\lambda)}(x_i)$ represents the i^{th} fitted value which computed from the fitting a model using y_{-k}.

2.4. Root Mean Square Error

Root mean square error (RMSE) is also known as root mean square error (MSE) which use to measure the different between the value of the actual observed value and the predicted value by the selected model. It is a useful method to determine whether the model is fit or not. The lower the value of RMSE meaning that the better fit of the model. The important feature of RMSE is the errors are weighted by means of squaring them. Therefore, the benefit of RMSE is penalizing the large errors which more strictly than any close by errors. The equation of the RMSE is shown in Equation 3.5 [16, 17, 18].

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{3.5}$$

where y_i is the observed response in i^{th} trial, $\hat{y}_i$ is the predicted response value in the i^{th} trial.

2.5. Lasso Regression

Adjusted R-squared is used to measure the proportionate the reduction of total variation in response variable associated the predictor variables. Adjusted R-squared may decrease or increase when another predictor variables is added in the model. The formula of adjusted R-squared is shown in Equation 3.8 [19].

$$SSE=\sum(y_i-\hat{y}_i)^2 \tag{3.6}$$

$$SSTO=\sum(y_i-\bar{y}_i)^2 \tag{3.7}$$

$$R_a^2=1-\left(\frac{n-1}{n-p}\right)\frac{SSE}{SSTO} \tag{3.8}$$

where SSE is the error sum of squares, $SSTO$ is total sum of squares, y_i is the observed response in the i^{th} trial, $\bar{y}_i$ is the mean of the response value in i^{th} trial.

3. Results and Discussion

Based on the Figure 1, the plot shows that many lines are converging to zero as the lambda increases. Each line in the plot represents the coefficient of the variable for the model and this has showed the function of the lambda as the regularization parameter. From the graph on the right, the different is the coefficients of the variables can exactly lie on the zero at horizontal axis. Therefore, Lasso regression can perform the variables selection.

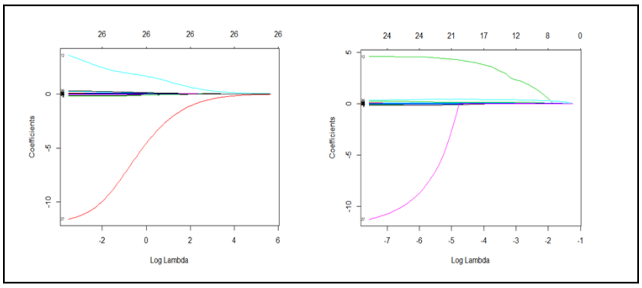
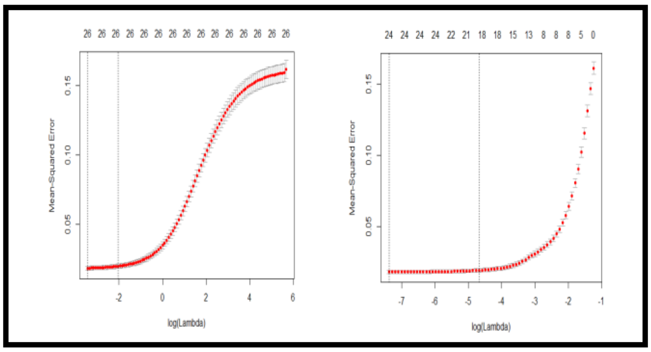


Fig. 1: Coefficient estimates for Ridge regression (on the left) and the Lasso regression (on the right) versus log λ.

Based on Figure 2, the graph on left shows log λ equal to 6, the value of lambda is big and the mean squared error is high. When the log λ value is approaching to -2 the mean squared error become small and stay flat. The optimal value of λ for Ridge regression is 0.03200447. The graph on the right shows log λ equal to -1, the value of lambda is big and the mean squared error is high. When the log λ value is approaching to -7, the mean squared error become small and stay flat. The optimal value of λ for Lasso regression is 0.0006282607.



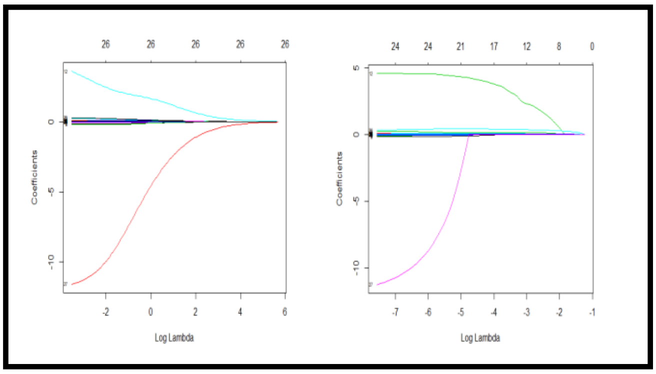


Fig. 2: Cross-validated estimate of the mean squared prediction error for Ridge (on the left) and Lasso (on the right), as a function of log λ

Table 1: Comparison for Ridge regression model and Lasso regression model

	Ridge regression	**Lasso regression**
Root Mean Square Error	0.1333643	0.1225798
Adjusted R-squared	0.8897418	0.9010351

Based on the Table 1, Lasso regression model is way better than the Ridge regression model which has low root mean square error and high adjusted R-squared value. Therefore, Lasso regression is the best method in determining the house price for this data set.

The model of the Lasso regression is shown as below. For X_1 is commercial zone, X_2 is medium residential zone, X_3 is lot frontage area, X_4 is lot area, X_5 is overall quality, X_6 is overall condition, X_7 is year built, X_8 is masonry veneer area, X_9 is external quality, X_{10} is basement quality, X_{11} is heating quality, X_{12} is central air, X_{13} is first floor area, X_{14} is ground living area, X_{15} number of bathroom at basement, X_{16} is number of bedroom, X_{17} is kitchen quality, X_{18} is capacity of garage car and X_{19} is the garage quality.

$$\begin{aligned}\text{Sale Price} = &– 25.8295 – 0.10196X_1 – 0003838X_2 + \\ &+ 0.01632X_3 + 0.067783X_4 + 0.031172X_5 + \\ &+ 0.045468X_6 + 4.191984X_7 + 0.000009X_8 + 0.069361X_9 \\ &+ 0.056703X_{10} + 0.009522X_{11} + 0.054173X_{12} + 0.14296X_{13} + \\ &+ 0.418605X_{14} + 0.048372X_{15} – 0.00219X_{16} + 0.04629X_{17} + \\ &+ 0.04426X_{18} + 0.004597X_{19}\end{aligned}$$

Since Lasso regression can perform variables selection, thus only the importance variables are included. Location of the house is a significant factor that affecting the house price which the zoning classification in the model that is related with the location. The house at the zone of commercial or residential with low medium density has a negatively related to the house price, meaning that the house price will be lower at these area. Commercial housing area is a place for the business purpose which supposed to have a good price for the house. However, not all the properties of the commercial housing area are expensive which the commercial area housing in Iowa, Ames, United States has cheaper price as Iowa is a rural area.

There are some variables which are positively related to the house prices, indicating that the house price will be higher. Lot frontage area, lot area, masonry veneer area, first floor area and garage car capacity and ground living area. Variables that closely related to size the house. Lot frontage is very costly as it requires money to pave the road or maybe to plant some grass or trees. The bigger the size of the house, the higher the price of the house. The bigger area of the house, the more land is occupied by the house, that is why the house price will high too. For the house with more space that consists of more than one floor, meaning that more materials are need to build the house, indirectly higher the house price.

For the variables year built, meaning that the age of the house, the latest brand new house will sell a better price compared to the old house as the old house has stood for a decades which more maintenance is need. Other characteristics of the house which are the number of the basement full bathroom show that the higher the number of the room, the more expensive for the house. Therefore, a house which is brand new, bigger in space and more rooms will have a high value of house price.

As United States is a four season country, central air conditioning and heating are necessary in a house. With central air conditioning is positively related to the house price. Houses with central air conditioning can make the temperature consistent and filter the air in the houses all year long, thus it will be a factor for higher house price. Heating quality is important as people can bath during winter or when is in a cold weather.

4. Conclusion

This comparative study found that the Lasso regression model is performing better compare to the Ridge regression model. Lasso regression has a lower root mean square error and higher adjusted R-squared compared to Ridge regression. The adjusted R-squared value of the Lasso regression is 0.90. This means that the sale price of a house is reduced by 90% when all the 18 predictor variables are considered. A high value of adjusted R-squared indicates the Lasso regression model is a better model.

In the Lasso regression model, the significant variables are selected which are mainly about the size of the house, the condition of the house, age of the house and also the location. Size of house, condition of house and age of house have positively related to the house price. Meaning that, the bigger the size of house either in space or area, the more expensive the house. The better the condition of the house, the higher the price of the house. For a new house that will be definitely with a high price compared to the old one. Location also play an important aspects for affecting the house price. The houses which are located at the commercial zone and medium density residual has cheaper price for the houses. United States is a four season country which the houses at Iowa, Ames would like to have the central air conditioning to maintain the temperature at house all year long. Least but not less, heating quality has to be good which uses to within to cold weather in United States.

Acknowledgement

The authors would like to express gratefully heartfelt thanks to the Universiti Tun Hussein Onn Malaysia and Office for Research, Innovation, Commercialization and Consultancy Management (ORICC) for the financial support under the TIER 1 research grant (U907).

References

[1] Bajari P, Benkard CL & Krainer J, “House prices and consumer welfare”, Journal of Urban Economics, 58(3), (2010), pp.474–487.

[2] Amri S & Tularam GA, “Performance of Mulitple Linear Regression and Nonlinear Neural Networks and Fuzzy Logic Techniques in Modelling House Prices”, Journal of Mathematics and Statistics, 8(4), (2012), pp.419–434.

[3] Mak S, Choy L & Ho W, “Quantile Regression Estimates of Hong Kong Real Estate Prices”, Urban Studies, 47(11), (2010), pp.2461–2472.

[4] Limsombunchai V, Gan C, & Lee M, “House Price Prediction : Hedonic Price Model vs Artificial Neural Network”, American Journal of Applied Sciences, 1(3), (2004), pp.193–201.

[5] Graham MH, “Confronting multicollinearity in ecological multiple regression”, Ecology, 84(11), (2003), pp.2809-2815.

[6] Kraha A, Turner H, Nimon K, Zientek & Henson RK, “Interpreting multiple regression in the face of multicollinearity”, Frontiers in Psychology, 3, (2012), pp.1–10.

[7] Bin Shafi MA, Bin Rusiman MS and Che Yusof NSH, “Determinants Status of Patient After Receiving Treatment at Intensive

Care Unit: A Case Study in Johor Bahru", I4CT 2014 - 1st International Conference on Computer, Communications, and Control Technology, 6914150, (2014), pp.80 – 82.

[8] Pasha GR & Shah MA, "Application of Ridge regression to multicollinear data", Journal of Research Science, 15(1), (2004), pp.97–106.

[9] Meinshausen N & Bühlmann P, "High-dimensional graphs and variable selection with the Lasso", The annals of statistics, 34(3), (2006), pp.1436–1462.

[10] Calhoun CA, "Property Valuation Methods and Data in the United States", Housing Finance International, 16(2), (2001), pp.12–23.

[11] Khalid K, Mohamed I and Abdullah NA, "An Additive Outlier Detection Procedure in Random Coefficient Autoregressive Models", AIP Conference Proceedings, 1682, (2015), 050017.

[12] Mohamed I, Khalid K And Yahya MS, "Combined Estimating Function for Random Coefficient Models with Correlated Errors", Communications In Statistics—Theory And Methods, 45(4), (2016), pp.967-975.

[13] Rusiman MS, Hau OC, Abdullah AW, Sufahani SF, Azmi NA, "An Analysis of Time Series for the Prediction of Barramundi (Ikan Siakap) Price in Malaysia", Far East Journal of Mathematical Sciences, 102(9), (2017) pp.2081-2093.

[14] Hastie T, Tibshirani R & Friedman J, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 1st ed. New York: Springer, (2001).

[15] Wakefield J, Bayesian and Frequentist Regression Methods, 1st ed. New York: Springer Science and Business Media, (2013).

[16] Chai T & Draxler RR, "Root mean square error (RMSE) or mean absolute error (MAE)–Arguments against avoiding RMSE in the literature", Geoscientific Model Development, 7(3), (2014), pp.1247–1250.

[17] Rusiman MS, Nasibov E and Adnan R, "The Optimal Fuzzy C-regression Models (OFCRM) in Miles per Gallon of Cars Prediction", Proceedings – 2011 IEEE Student Conference on Research and Development, SCOReD 2011, 6148760, (2011), pp.333-338.

[18] Shafi MA and Rusiman MS, "The Use of Fuzzy Linear Regression Models for Tumor Size in Colorectal Cancer in Hospital of Malaysia", Applied Mathematical Sciences 9 (56), (2015), pp.2749-2759.

[19] Kutner MH, Nachtsheim CJ & Neter J, Applied Linear Regression Models 4th ed., New York: McGraw-Hill Higher Education, (2003).