Final Project for Regression Analysis

Jie Gu

March 10, 2019

Contents

1 Introduction

Real estate is an industry full of opportunities and risks, while residential is the most concerned category of real estate. Therefore, it's practical to conduct research on residential building dataset.

The residential building dataset is comprised of 8 project physical and financial variables ($V1 \sim V8$), 19 economic variables and indices in 5 time-lag numbers ($V11 \sim V29$). The two output variables are construction costs ($V9$) and sale price ($V10$). There are 372 observations, which should be a small-sized dataset.

2 Model Building

I. Data Preprocessing

One of the Project Physical and Financial Variables is pretty special - $V1$ zip code, ranging from 1 to 20. Since situations may differ between locations, we defined $V1$ as a categorical variable. As we can see from the summary of the full model, the zip codes' P-values are large, so in the following steps, we did not use this variable into regression.

In order to apply ***k*-fold cross-validation** to check the model and its predictive ability in the final step, we split the dataset into two parts. By selecting 38 observations out randomly, the original database was split into training dataset (with 334 observations) and testing dataset (with 38 observations). In fact, we repeated the random selection performance for several times and found that the final models are almost indistinguishable.

Additionally, **correlation transformation** was applied to the remaining 26 predictors to help with controlling round off errors. Expressing the regression coefficients in the same units may be of help when these coefficients are compared. The standardization involves centering and scaling is as follows:

$$\frac{X_{ik} - \bar{X}_k}{sd_k}$$

II. Model Selection

Since there are so many as 27 variables in the original dataset, if we just build a simple linear regression model with all the variables, there must be some useless variables and multicollinearity. In order to improve efficiency and reduce noise, Lasso is the first choice to preliminarily screen the necessary predictors. We first study the regression formation of output $V9$.

Use the most important variable selected by **Lasso** as the variable of Model 1, then here comes the basic model of $V9$ - with the most important variables in the first-order form.

$$\begin{aligned} E\{Y_i\} &= \beta_0 + \beta_1 X_{i4} + \beta_2 X_{i7} + \beta_3 X_{i8} + \beta_4 X_{i16} + \beta_5 X_{i17} + \beta_6 X_{i18} \\ &+ \beta_7 X_{i20} + \beta_8 X_{i21} + \beta_9 X_{i23} + \beta_{10} X_{i26} + \beta_{11} X_{i28} + \beta_{12} X_{i29} \end{aligned}$$

Although we have sorted out the 12 most important variables, they are not necessarily linear. Drawing added-variable plots is a good method to determine the exact functional form of an independent variable in a multiple regression model. So the next step is to draw the **added-variable plot** of each variable in Model 1 against the other variables. Then we got 12 added-variable plots.

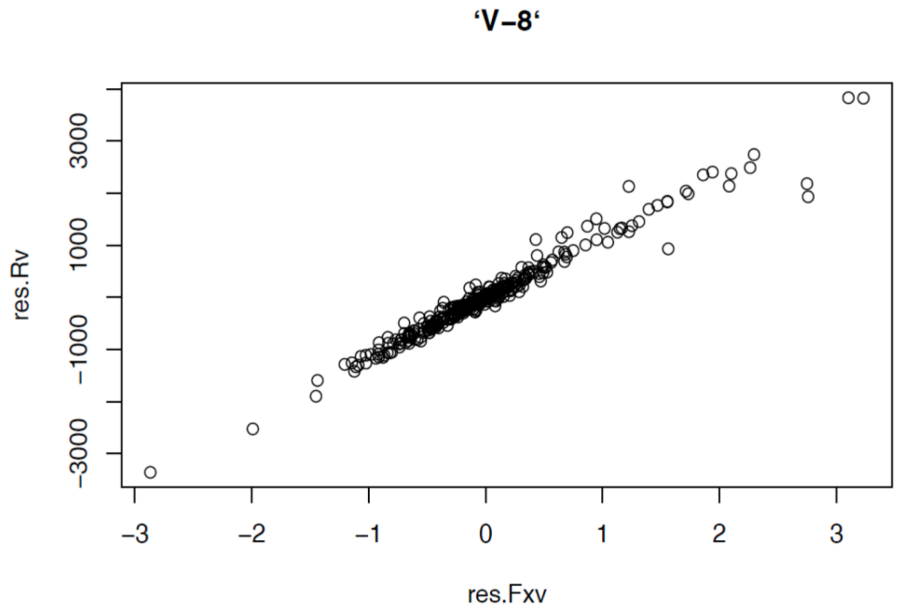

Figure 1: Added variable plot of $V8$

As can be seen from the plot above, the added-variable plot of $V8$ is exactly linear, however, the shape in the other plots seem curved. Then we decided to add the quadratic variables into Model 1 then use Lasso once again to sort out the Model 2.

Then here comes our Model 2 of V9 - with quadratic variables.

$$\begin{aligned} E\{Y_i\} &= \beta_0 + \beta_1 X_{i4} + \beta_2 X_{i7} + \beta_3 X_{i7}^2 + \beta_4 X_{i8} + \beta_5 X_{i17}^2 + \beta_6 X_{i18}^2 \\ &+ \beta_7 X_{i20}^2 + \beta_8 X_{i23} + \beta_9 X_{i23}^2 \end{aligned}$$

III. Diagnostics

We use the **Brown-Forsythe Test** to determine whether the errors have constant variance. Since there are a series of predictors in the dataset, it's impractical to use the median of X values to divide the observations into 2 groups. Hence, we did an innovation to the original method - use the median of Y values to divide the observations. The test outcome shows that the error variance of Model 1 is not constant.

$$|t^*_{BF}| = 5.81 > t(1 - \frac{.05}{2}; n - p - 1) = 1.97$$

Therefore, the **Weighted Least Squares Estimation** is necessary to remedy the unequal error variances. By calculating the weighted least square diagonal matrix $W^{1/2}_{n\times n}$, we got the weighted regression Model 1, which is stated below.

$$\boldsymbol{\gamma_W = X_W\beta + \epsilon_W}$$

After the weighted transformation, we use the Brown-Forsythe Test again to test the effect of Weighted Least Squares Estimation. The new P-value is 0.15, and now, the errors' variances of Model 1 are constant. Similarly, the errors' variances of Model 2 are not constant, either. We did the same operation on Model 2, making Model 2 also satisfactory.

Outliers will affect the accuracy of our models. **Bonferroni critical value** of $t[1 - \frac{.05}{2n}; n - p - 1]$ is used to determine Y outliers. **Cook's distance** with percentile value larger than 20% and **DFFITS** with the absolute value larger than $2\sqrt{p/n}$ are used to determine influential X outliers.

The Variance Inflation Factor is a formal measure of multicollinearity. A large **VIF** indicates the existence of multicollinearity. We calculated every variable in the models. The results show that our models do have slight multicollinearity, but the accuracy of the prediction will not be affected.

IV. Final Model

After removing the Y and influential X outliers, we refitted the coefficient value with the same terms in the two models. We used the Coefficient of Determination (R^2) to determine whether our models did a good job. To verify that our models don't have an overfitting problem, we recalculated the coefficients on the testing dataset with the exact same terms. The results show that our model has a high R^2 in both the training dataset and the testing dataset, and our models all performed well.

We performed the same regression process for output $V10$ as $V9$. The adjusted R^2 values of models are as below.

Table 1: Adjusted R-squared Values of Models

	Output $V9$				Output $V10$			
	Model 1		Model 2		Model 3		Model 4	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
R^2	99.74%	99.87%	99.55%	99.62%	97.68%	97.98%	88.05%	95.17%
R^2_α	99.73%	99.79%	99.54%	99.45%	97.65%	97.71%	88.01%	95.03%

By comparison, we decided to choose Model 1 and Model 3 as our final model. Because it has a higher R^2 value, the form is also more simple, reducing unnecessary noise.

The final model for output $V9$ is:

$$\begin{aligned}E\{Y_i\} &= 183.5 + 12.8X_{i4} + 82.9X_{i7} + 1183.7X_{i8} - 12.9X_{i16} - 285.1X_{i17} - 44.6X_{i18} \\ &- 17.8X_{i20} + 7.3X_{i21} + 15.2X_{i23} + 146.3X_{i26} + 30.3X_{i28} + 114.1X_{i29}\end{aligned}$$

The final model for output $V10$ is:

$$E\{Y_i\} = 226.8 + 3.5X_{i4} + 126.6X_{i5} + 13.7X_{i7} + 12.7X_{i23}$$

3 Conclusion

I. Results Interpretation

Based on our previous analysis and the models that we have built, we can find that the actual sales prices are highly positively related to the price of the unit a the beginning of the project. What's more, the land price index for the base year negatively influenced the actual sales prices. The CPI of housing, water, fuel power in the base year is also significantly influencing the actual sales prices. Other factors such as Total preliminary estimated construction cost based on the prices at the beginning of the project are also important, but the significance is not as important as the three factors that we mentioned before. Factors like lot project or total floor area of the building are not relevant to actual sales prices.

As for the second factors that we have predicted (actual construction costs), we find out that one variable is significant with our outcome, that is preliminary estimated construction cost based on the prices at the beginning of the project. Other factors such as duration of construction, total preliminary estimated construction cost based on the prices at the beginning of the project are also very important. According to the **stepwise selection** and observation of the connection between different observations, we believe this is reasonable and totally make sense.

II. Improvement

- Although our outcome is extremely great, $R2$ is pretty high, we still have some improvement space. For example, our date observations are only 372, which could be higher if we want a more accurate result.
- The data that we have received is contained 5 different lags, but we only used one of them, because this contains some time series problems that we cannot resolve based on our current knowledge. We believe that if we can all of the data and compared the results in different periods, our results might be better and more convincing.
- All the price is time sensitive data, and we can collect more data at different times to improve our prediction. We can still utilize feature engineering in order to create more significant variables to prove our results.

4 Reference

Kutner, M. H. (Ed.). (2005). Applied linear statistical models (5th ed). Boston: McGraw-Hill Irwin.

Hongmei, Jiang. (2019). Chapter3GraphsBrownTest. [Source code].
https://canvas.northwestern.edu/courses/89947/files/folder/R%20Codes?preview=6235941

Hongmei, Jiang. (2019). Chapter10Section6SurgicalUnit. [Source code].
https://canvas.northwestern.edu/courses/89947/files/folder/R%20Codes?preview=6558400

Hongmei, Jiang. (2019). Chapter12glmnet. [Source code].
https://canvas.northwestern.edu/courses/89947/files/folder/R%20Codes?preview=6235967

5 Appendix

R code for predicting V-9

R code for predicting V-10