

MLR Model

**Influence Factorsand Analysis of Automobile Price**

**Group member:**

**李瀚生 1930026065 ()**

**张宏熙 1930026155 ()**

**薛劭杰 1930026143 ()**

**朱奕卓 1930004048 ()**

**Abstract**

Nowadays cars are gradually becoming a necessary for every family. It also disturbs lots of family of the fluctuating car price. This report aims at analyze recent years of automobile price. Building a mathematical report, Using method includes Set dummy variables, Recursive feature elimination, Cross validation, Backward elimination, Transformation，Outlier Test. finding out the most five crucial factors.

**Introduction**

People care a lot about car price but seldom concern what factors are influencing price wave. To maximize the benefit. We would use a regression model to find out the five most contributed factors of the car price. It’s not only benefit for each family, but also helps when company doing strategy. The source of our data set comes from a website called Kaggle, which are posted on it in April 26th 2019. The data has also been processed by some data scientist first to hold a competition. This guarantees that it equipped with authority and aging. Then we do some basic cleaning and arrangement to start the analyzation.

**Methodology**

1. Set dummy variables: The qualitative variables were transformed into dummy variables so that the qualitative variables could be added to the linear regression model for evaluation. This improves the accuracy of the model.
2. Recursive feature elimination: To put it simply, give the value of each feature's contribution (significance) to the model, rank them by importance, and choose the appropriate features easily as they are ordered.
3. Cross validation: Based on the rankings from the RFE above, select the different number of features. Do the cross validation to the selected feature set. Select the best model (PRESS is minimum). It also facilitates the stability of the model.
4. Backward elimination: Reduce the number of features and simplify the regression model.
5. VIF: Remove the possible existence of multicollinearities.
6. Transformation: Let the distribution not concentrated in a certain range, so that the variance is going to be more stable.
7. Cook's distance: Detect influential points.
8. OutlierTest(): Function used to detect outliers (use studentized residual error).

**Data analysis**

Firstly, clean up the data. For example, remove useless information such as id and resolve NULL. Uniform formatting of data, such as unifying car manufactories’ names.

Secondly, transformations are performed on the data. By removing duplicate data and turning qualitative variables into dummy variables, the regression model is easier to build up.

After previous operations, perform descriptive analysis on the data. It is necessary look at the relationship between each independent variable and the dependent variable, the distribution of each independent variable, and the correlation between the independent variables. Following the basic assumptions of regression analysis, the independent variables should be independent of each other, and the independent and dependent variables should be linearly related to each other.

图表, 散点图

描述已自动生成图形用户界面

描述已自动生成

80% of the data was parted into the training set and the remaining 20% as the test set. The test set is not involved in training, in order to evaluate the stability of the model later. After RFE and cross-validation were done on the training set, low variance features were removed , appropriate model was also found. Up to this point, 10 features are included in the model. The following diagram shows the key to the model.

R횬剗|飇 VS 
Resets VS 

In an ideal linear regression model, one of the assumptions is that the residuals should be normally distributed and independent of the estimated value. At the “Residuals vs Fitted” picture, the residuals are basically independent of the estimated values. At the Q-Q plot, we can see that some of the residuals on the tail far from the normal distribution. Another assumption of linear regression model is that the variance in the prediction model is a fixed value. On the “Scale-Location” plot, we can find that the variance is growing at the beginning, which means the variance of error is not a constant. The last plot we use cook’s distance to find influential point.

表格

描述已自动生成表格

描述已自动生成

Next, we do the backward eliminate to the model, tried to simplify the model. We delete the features with P value less than 0.05. The remaining parameters have significant non-zero coefficients. Then, use variance inflation factor method to solve multicollinearity. We delete which the VIF value bigger than 5. The simplified model contains 5 features.

图示

描述已自动生成

Obviously, the first figure is a little bit fan-out and concentrated at the begining. Thus, use log to do the transform to the Y, which makes the large fitted values smaller and the residual to be more evenly distributed.

图表

描述已自动生成

Now, the residual plot looks like a horizontal band, and this is good for the regression model. The third picture looks even more like a horizontal line, which means that variance more constant.

图表, 折线图

描述已自动生成

The data do not have influential point. So, at the last step, deal with those outliers. From the plots, after treatment, residuals are getting closer to normal distributions. For model built up, deleting outliers is the most suitable way to deal with outliers since deleting outliers reserver the model’s stability.

表格

中度可信度描述已自动生成At last, using r2score to evaluate our model. For the training set, our model’s r2score is 0.91. For the testing set, the model’s r2score still have 0.81. Then, up to now, the regression model can consider both prediction and fitting. Although the model only contains a constant term and five features, it still has the interpretability of the model with fewer features.

**Conclusion**

After series of analyzation. Our training data set and modeling diagram shows the result have a good interpretation and prediction. The most five influence factors are const, horsepower, car width, hatchback, dohcv, highend. These are telling us to focus more on those why picking your cars. Since the contribution also defines the performance of a car. What’s more, company also need to figure out a way for a better work efficiency between those factor and others.

**Reference**