Linear Regression Analysis on Second-Handed Car Price

Aiwei Yin, Kaiyao Duan, Zhanyan Guo, Alexander Lee

2024-06-16

1. Introduction

The demand for cars in modern society has grown significantly, leading to a robust market for second-hand vehicles. Unlike the straightforward process of purchasing a new car from a dealership with fixed prices, buying a used car involves numerous factors that can complicate the decision-making process. Understanding these factors is crucial for consumers to make informed choices in the second-hand car market.

Previous research has extensively explored the determinants of second-hand car prices using various methods. (Pudaruth 2014) conducted a comparative study on multiple linear regression, k-nearest-neighbors (kNN), decision trees, and Naive Bayes to predict used car prices. The study found that while linear regression did not perform as well as kNN, the year of manufacture was a more significant predictor of car prices than mileage. Additionally, it was noted that logarithmic regression slightly outperformed linear regression in predictive accuracy. However, due to missing data, the application of linear regression was limited to using only the year and mileage as predictor variables.

Another study highlighted the critical role of data pre-processing in improving predictive models. By removing outliers, noisy values, and irrelevant columns, the coefficient of determination (R^2) improved from 0.62 to 0.73. This improvement underscores the importance of data quality and pre-processing techniques in enhancing the accuracy of predictive models for used car prices.(Muti & Yildiz 2023)

2. Method

This study finds the optimal linear regression model to predict the Price of second handed cars given the multiple predictor variables: Brand, Model, Year, Kilometers Driven, Fuel Type, Transmission Type, Owner Type, Mileage, Engine, Power, and Number of Seats.

We begin with preprocessing the data set, turning all the columns with text values that represent categorical variables into dummy variables. Then we will randomly shuffle the data, and partition it into 70% of training set and 30% testing set.

For all categorical columns, we first change the ones with multiple categories into multiple single column columns, then treat the variables as dummy variables. For categorical variables that have multiple categories, we create a dummy variable for each of the categories.

To find out the relation between each of the variables and the our response variable, we first analyze the scatter plot between each of the predictor variables and Price, and discard both the continuous variables that appears completely random and categorical variables that partition the data set into too many small chunks. Then we create a simple multilinear model without any modifications, to provide a baseline model for us to asses the outcomes of our methods. We first check violations of assumptions (linearity, uncorrelated errors, constant variance) with Normal Quantile-Quantile (Q-Q) plots and Residual vs Fitted Values plot. Ideally, the normal Q-Q plot should appear to be a straight diagonal line, and the Residual vs Predict value plot should look completely randomly distributed.

Based on the simple multilinear model, we calculate the Cook's Distance of all the observations in order to find the outliers. We choose to remove all the observations with Cook's distant $\geq 3/n$.

To resolve the violation of the above assumptions, we will use three methods. We will first use transformation on the response variable, choosing the optimal between box-cox transformation, log transformation and power transformation based on the normal QQ plot and residual plots.

After the transformations we will select the predictor variables based on the multi-colinearity with the VIF scores based on the privious simple multilinear model. We choose to analyze the variables with VIF > 0.3. There are two possible situations, one is that there is no significant linear relationship with price, and one is that there is a high correlation between two of the predictor variables. For the first situation, we choose to directly discard the predictor variable. For the second situation, we choose to discard the predictor variable with lower correlation with Price, and keep the predictor variable with higher covariance.

When we select the preferred model, we will again assess the assumptions above, and validate our model by examining the models with Analysis of Variance (ANOVA) tests, if the P-tests for all the predictor variable show a P-value of less than 0.05, we consider the model as significant. Furthermore we will compare the R^2 value of both models, if the R^2 value is indeed higher we choose to use the model we selected.

3. Results

Looking at the predictors, given the limited data set, we decided to discard brand and model as they partition the data into too many small subsets. There are also two categorical variables that are not binary, for Owner_type, they are split into three columns following our methodology section as there are a considerable amount of data points for each type. However for seats, due to four seats and seven seats having 6 data points in total, we decided to remove these two variable as well as the corresponding data points.

For continuous variables, box plots are used to look at the distributions of the data and find any leverage points, where we will use Cook's distance to eliminate bad leveraging points below.

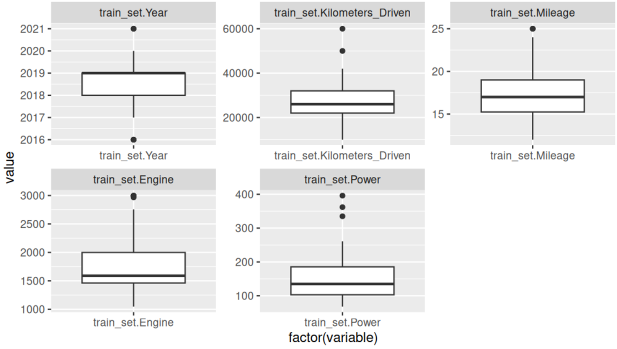

Figure 1: box plots for continuous variables

The following diagrams contains the box plots of each categorical variable graphed with respect to price, where we can see how the occurrence of a certain feature might affect the price.

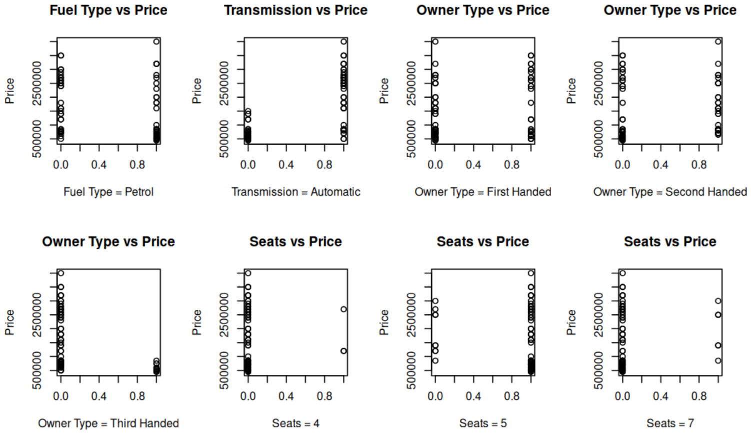

Figure 2: scatter plots of categorical variables

3.1 Simple Linear Model

Prior to pre-processing our data and applying various linear regression techniques, we will first generate a linear model to provide a baseline to observe any abnormality and compare the effectiveness of our methods.

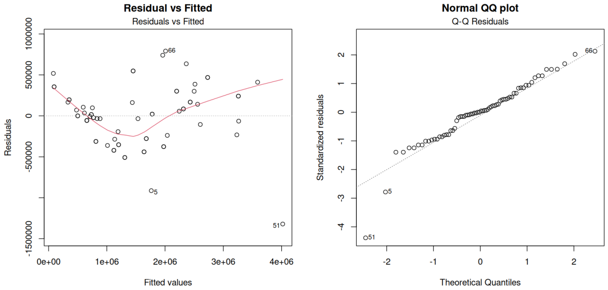

Figure 3: plots for simple linear models

From the Residual vs Fitted graph, it is immediate that there are patterns and the points are not scattered randomly this implies that there are exists non-linearity in the residuals which have to be dealed with. The spread of residuals also increases as the fitted values increase, suggesting a non-constant variance of errors, or heterosccedasity, moreover there also exists outliers with residual very far from 0. As for the QQ plot, normality of the residuals seems to be followed mostly except for the a few points.

3.2 Removing Outliers

We will first use cook's distance to identify and remove bad leverage points with $3/n$ as the threshold for outliers.

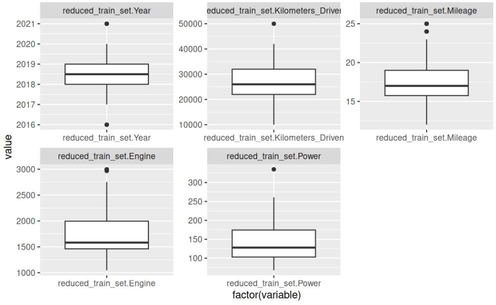

Figure 4: box plots of variables after removing outliers

We can now see that some outliers does removed.

Since linearity and constant variance were both violated according to the residual-fitted plot. We will apply **box cox transformation** to correct these. We find that the optimal lambda is $\lambda = -0.020202$

Next step is to detect whether there are any multicollinearty in our data by using Variance Inflation Factors:

Table 1: Variance Inflation Factors

	vif_scores
Year	3.767972
Kilometers_Driven	3.317379
Fuel_Type	1.762901
Transmission	1.633205
Owner_TypeFirst	1.268131
Mileage	2.356504
Engine	4.959014
Power	4.065378

Year and Kilometer have a very high VIF, similar results could be shown by inspecting the plot of these two variables, we choose to discard them. However for Power and Engine, we suspect the high VIF come from the fact that these two variables are highly correlated.

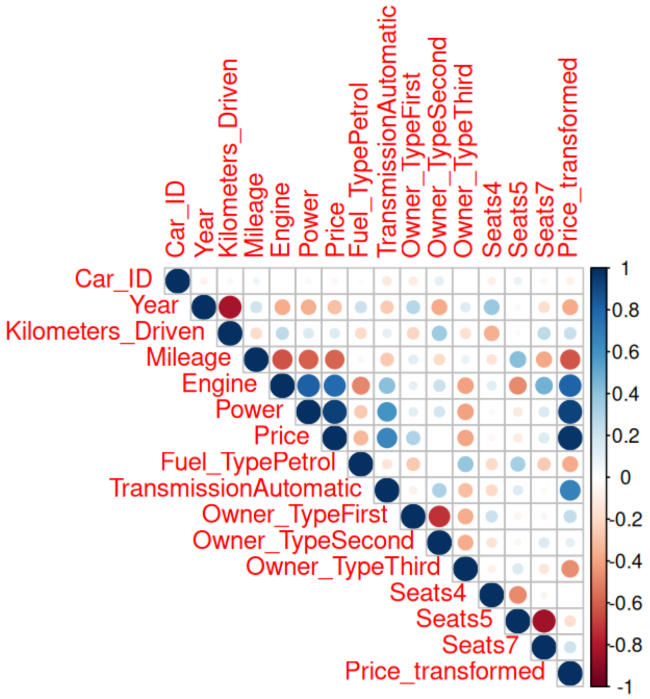

Figure 5: covariance matrix

Calculating the covariance, it seems that the pairs are year and mileage as well as engine and power, as they each have high covariances. For year and mileage, we decided to leave them both in as the VIF is not very high. However we chose to discard one of Engine and Power since they were very close to the threshold of five, and since Engine had a lower covariance with Price ($Cov(Engine, Price) = 0.8$, whereas $Cov(Power, Price) = 0.92$), we decided to remove Engine and keep Power in.

3.3 Model of choice

In the transformed model, we achieved a R-squared of 0.9400994, an increase of 0.0716579 compared to the baseline model (0.8684415).

Results show that our treatment of removing outliers with cook's distance, using box-cox transformation, and utilizing VIF to remove useless variables improves our model effectively. Despite the residual-fitted value plots still exhibits a slightly cubic polynomial pattern, it is now much flatter and thus random compared to the inital residual value plot of our original model.

We also provide the formula for our model:

$$\log(Price) = -0.108346*I(Fuel\_type = Petrol)-0.265362*I(Transmission\_Type = Mannual)+0.227912*I(Owner\_Type$$

3.4 Testing Dataset

We apply our model of choice to the testing dataset. The testing result show that aside from two outliers, out model satisfies the assumptions, so we confirm that we build a valid model.

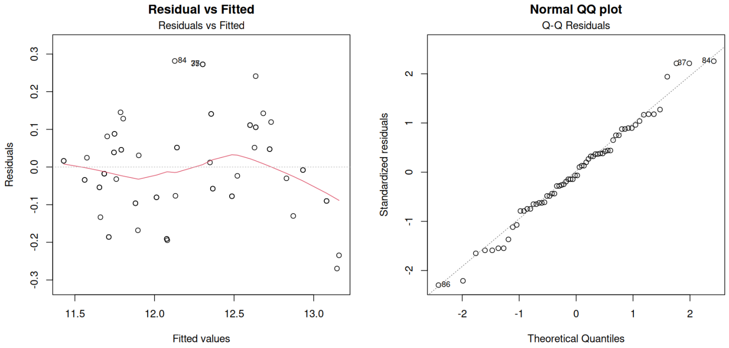

Figure 6: plots for the final model

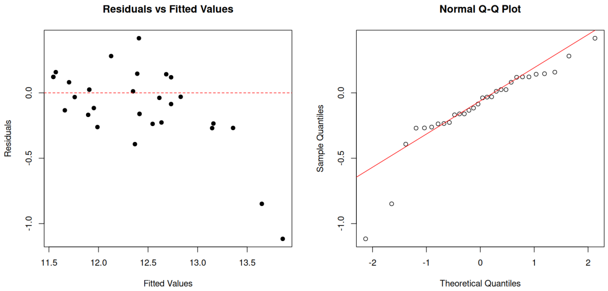

Figure 7: plots for the final model on testing set

4. Discussions

According to our final model, a one-unit increase Mileage is expected to cause a 0.039586 unit decrease in the log of Price, assuming that all other variables are constant. Additionally, an increase of Poser is expected to cause a 0.0044 of increase in the log of Price when all others hold. Similarly, each unit of increase in Year is expected to cause a 0.0486 decrease in log of year. Overall, our resarch shows that there are a strong linear relation between each of our selected predictors and the response variable (log of Price), with both positive and negative correlations with the log of Price.

Our model gives a credible and valuable insight into the relations between different situation of cars and their price in the second handed market. It also serves as a tool to predict the expected price in the market given a new observation. Although the assumption of linearity and normality is not fully met in the testing set, out model still gives a high R^2 score, and successfully describes most of the observations in the dataset.

However our model also have some drawbacks, the most significant one being the lack of observations in our dataset. Our entire data set contains only 100 observations, and thus might not serve as a statistically significant representative for the population. With this in mind, we might expect our model to be overfit when predicting further sets. This issue of overfit could not be resolved due to the lack of data.

Reference

Sharma, A. D., & Sharma, V. (2020, November 11). Used car price prediction using linear regression model. https://www.irjmets.com/uploadedfiles/paper/volume2/issue_11_november_2020/4868/1628083194.pdf

Pudaruth, Sameerchand. (2014). Predicting the Price of Used Cars using Machine Learning Techniques. International Journal of Information & Computation Technology. 4. 753-764.