Predicting the Price of Used Cars using Regression Analysis Techniques

Students: Wenzhao Wang (301326657)

Mengqi Xie (301377381)

Xuefei Li (301306324)

Kunpeng Wang (301326336)

Instructor: Dr. Derek Bingham

Course: STAT 350, SFU Burnaby (Fall, 2020)

Summary:

1 Abstract

In this project, we investigate the relationship between several factors and selling price of cars. Factors include vehicle manufacturers, mileage, engine displacement, max power, the years since it is registered and the number of miles it has been driven. The data we based on, contains more than seven thousand observations after sorting and preparation. After building multiple linear regression models and conducting ANOVA test on selected certain models, we obtained specific linear relationships between vehicle selling price and certain explanatory variables. Beside the investigation of linear relationships, we are also interested in constructing a function which can perform relative accurate prediction on the selling price of a vehicle given several explanatory variables. In order to confirm the accuracy of the prediction function, we ran a test by dividing the observation into two parts and building the function based on the first part, then test it on another one.

2 Introduction

Price prediction is the core essence of all market industries, especially in the resale market. According to data obtained from CNBC [1], Covid-19 has led to an increase in used car sales as people now avoid mass transportation and are more sensitive to auto budget in the downturn. The report shows due to the lockdowns, timeline to full auto factory restarts and a global consumer confidence decline, it expected a drop of 36 million new vehicle sales globally through 2022, compared to 2019 levels in the U.S. market. Meanwhile, the used car market was more than twice the size of the new car sales market even before Covid-19. The rise of used cars sales is rapidly increasing. Therefore, a need for a model that can predict the selling price of a used car by evaluating its features with the consideration of other cars in the

market is in high demand. For this project, we focus on studying the factors and their significance in the price prediction of used cars. It is important to note that from the choice of materials to manufacturing and production of cars, it will all contribute to the value of this car and with added factors for used cars. Likewise, the reselling price of these used cars will be one of the elements that is weighing on the car price in the first place. Hence, it is important to assess the price prediction of used cars and study the factors and find the significance of these factors.

Predicting the resale price of a used car is not an easy task. There are a number of factors that influence the car price. The most influential ones are usually the age of the car, the model of the car, the original country of the manufacturer, its mileage and its horsepower. [2] Other factors including, but not limited to, the type of fuel the car uses, the volume of its cylinders (measured in cc), the number of seats, the interior style, the braking system, whether it is automatic or manual transmission, whether it belonged to an individual or a company, how many owners does the car has, lien status, accident history and auto values. [3] Some special factors like if the car is customized, its color, if the previous owner smoked in the car, air conditioner, sound system, GPS navigator all may influence the price as well. [4] In this project, we have considered only a small subset of the factors listing above. More details are provided in the following sections.

In this research, we identify factors that affect secondhand car price from most influential order to least influential order. We wish to build our model based on these findings to conduct a multiple linear fit model that can predict the selling price of a secondhand car. We used the vehicle dataset from cardekho, an Indian online car dealer company for our analysis. [5] The data can be used for a variety of purposes, in our project, we accomplished

price prediction by the use of regression analysis. We constructed and determined an appropriate linear regression model as the final model by variable selection and data transformation . Then we checked the final model by robust regression analysis following with the interpretation of how variables affect the selling price. This paper is organized as follows. In the next section, details and descriptions of the data are presented by data visualizations. Section III describes the methodology, and all the analysis tools used. Following Section IV, results from the regression analysis. Finally, we end the paper with a conclusion with some pointers towards future work.

3 Data Description:

Dataset:

To accomplish the prediction of a used car's selling price, we used an open dataset "Vehicle dataset from cardekho" from Kaggle to build our model. The dataset is scraped from cardekho, an Indian online car dealer company. The dataset contains the selling prices and their features of 7906 used cars across 32 brands. Our dataset contains 12 unique attributes of a used car being sold out of which we removed one variable, "torque", that the variable has little impact on a car's price from our analysis.

Additionally, we perform the following preprocessing steps on the data set helping us have a better understanding of the dataset as well as narrowing down the features:

1. Keep only listings for cars with complete information.

2. Introduce one new additional data point into the dataset. The new data point we selected is a 5 seats Maruti Swift Dzire VDi car. We chose the data point randomly from the cardekho website with restrictions that the data point must have complete information with similar predictors as the original dataset, i.e. the car brand must choose from that 32 brands.
3. Convert qualitative variables to numeric variables by splitting the numeric parts and unit parts then strip off the unit columns.
4. Sort car manufacturers by origins.

We then consider a specific one variable named “selling price” as the response variable and others to be explanatory variables, details are listed below.

Response variables:

1. selling_price: The selling price of a used car

Explanatory variables:

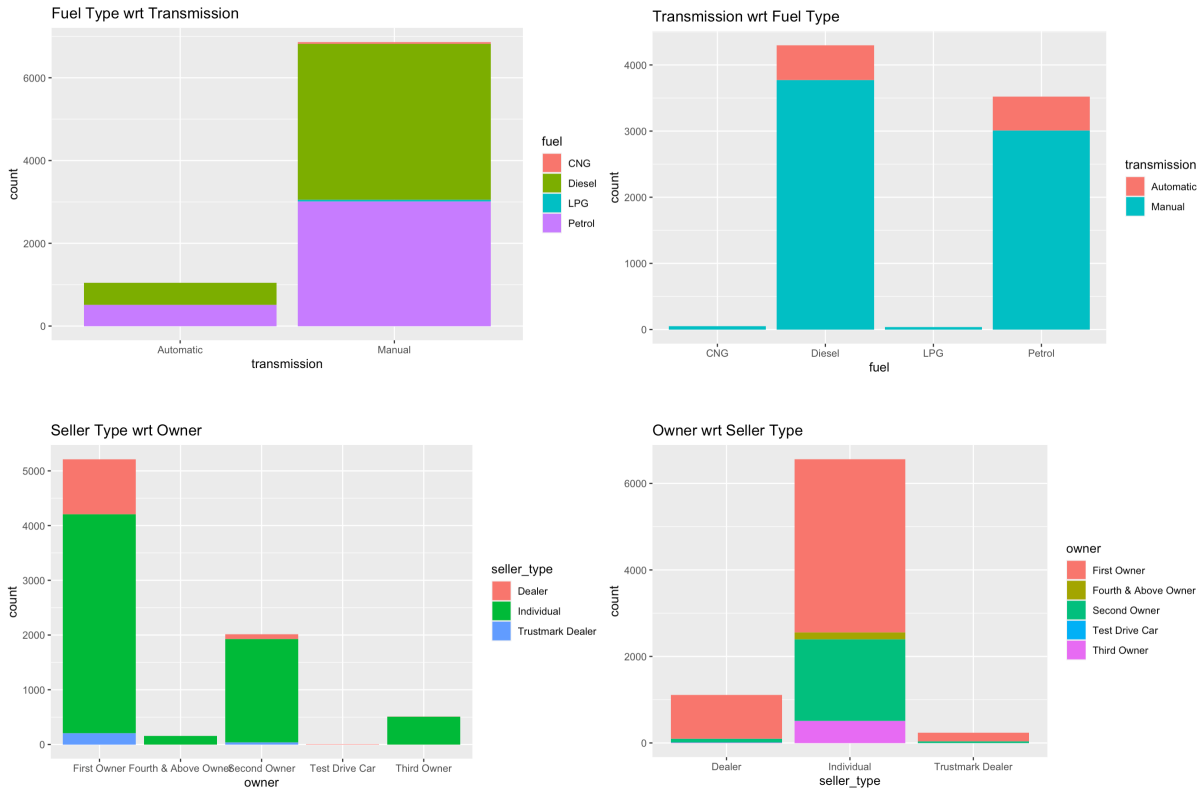
1. Manufacturer: The manufactured location of a car (Factor)
2. Years: The number of years since a car was produced (Integer)
3. Mileage: The mileage of a car (Integer)
4. Engine: The parameter of car engine, measured in cc (Integer)
5. km_driven: The distance that a car has travelled (Integer)
6. fuel: The fuel type of a car (Factor)
7. seller_type: whether the owner of a car sell it individually or through a dealer (Factor)
8. transmission: whether it is automatic or manual transmission (Factor)
9. owner: how many times the car has been trade (Factor)

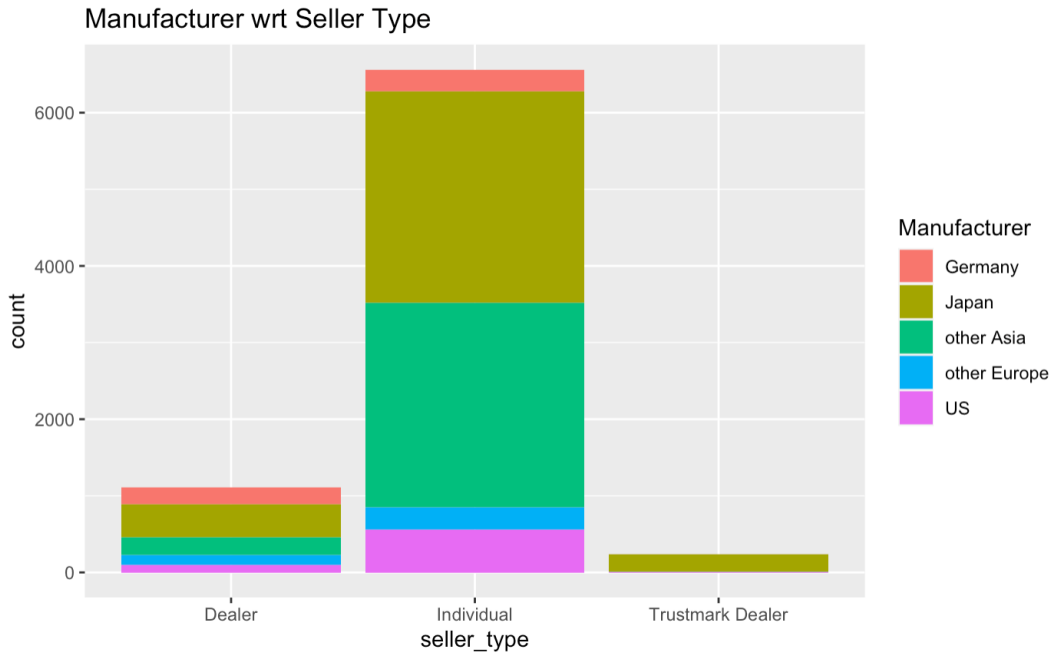
10. seats: number of seats of the car (Integer)
11. Max_power: the horsepower of the car (Integer)

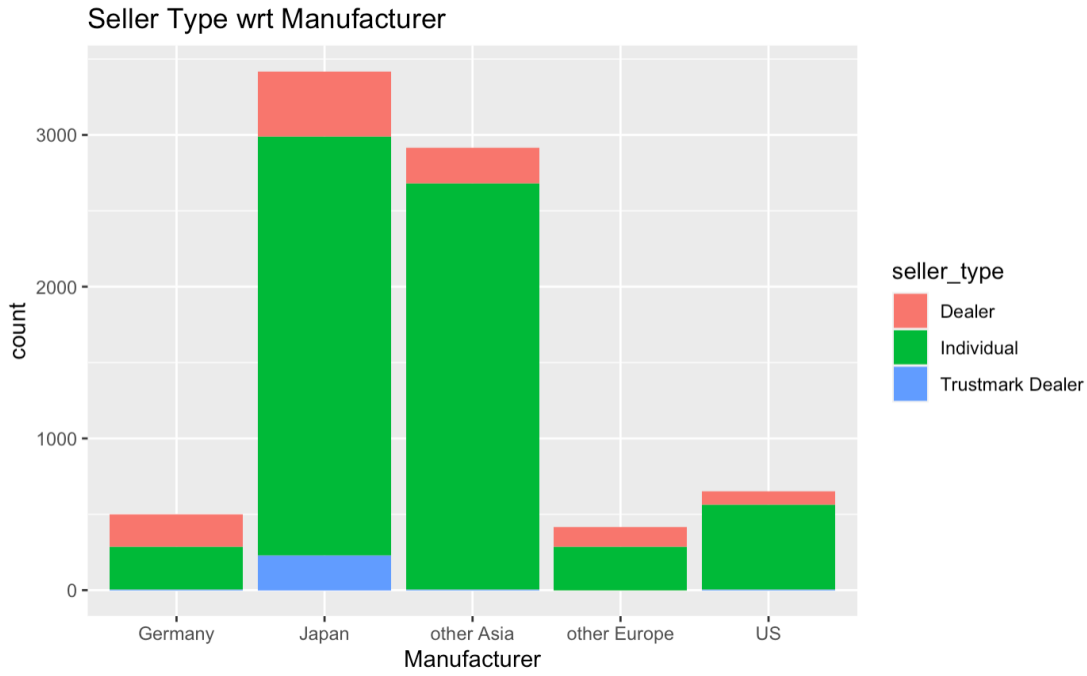
Data Visualizations:

Factor variables:

From nine explanatory variables, Manufacturer, fuel, seller_type ,transmission and owner are factor variables. We can understand them better through histograms below.

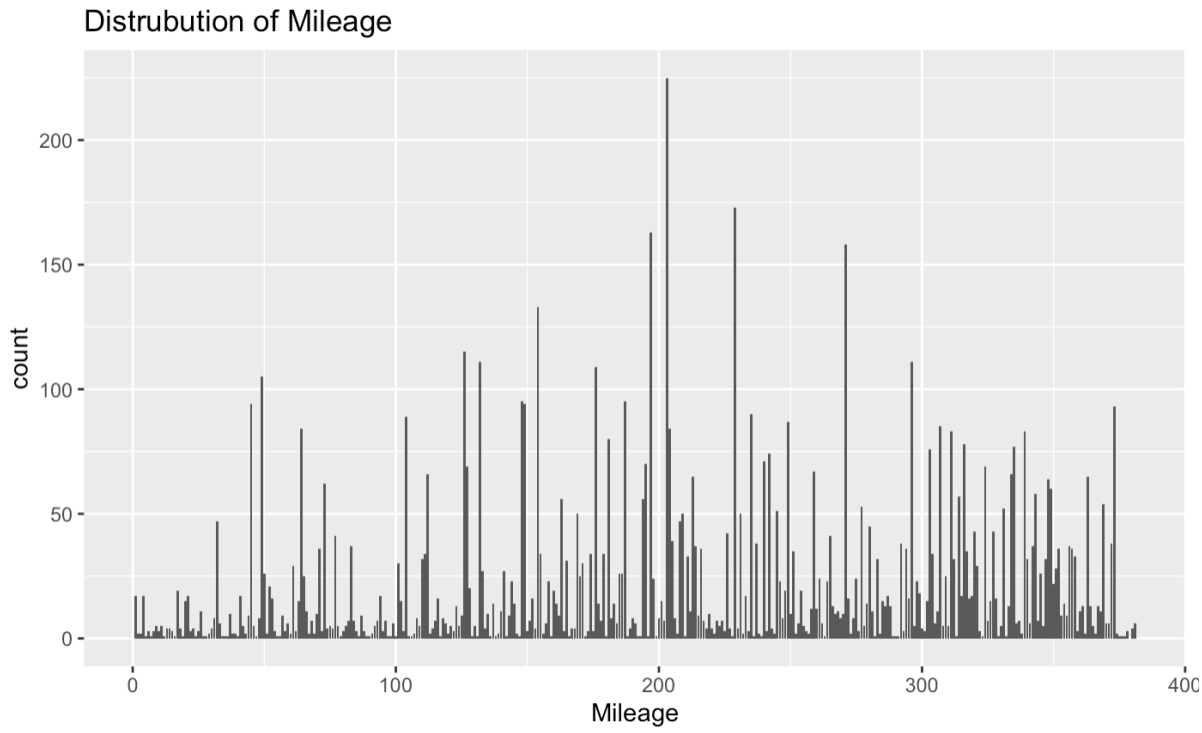


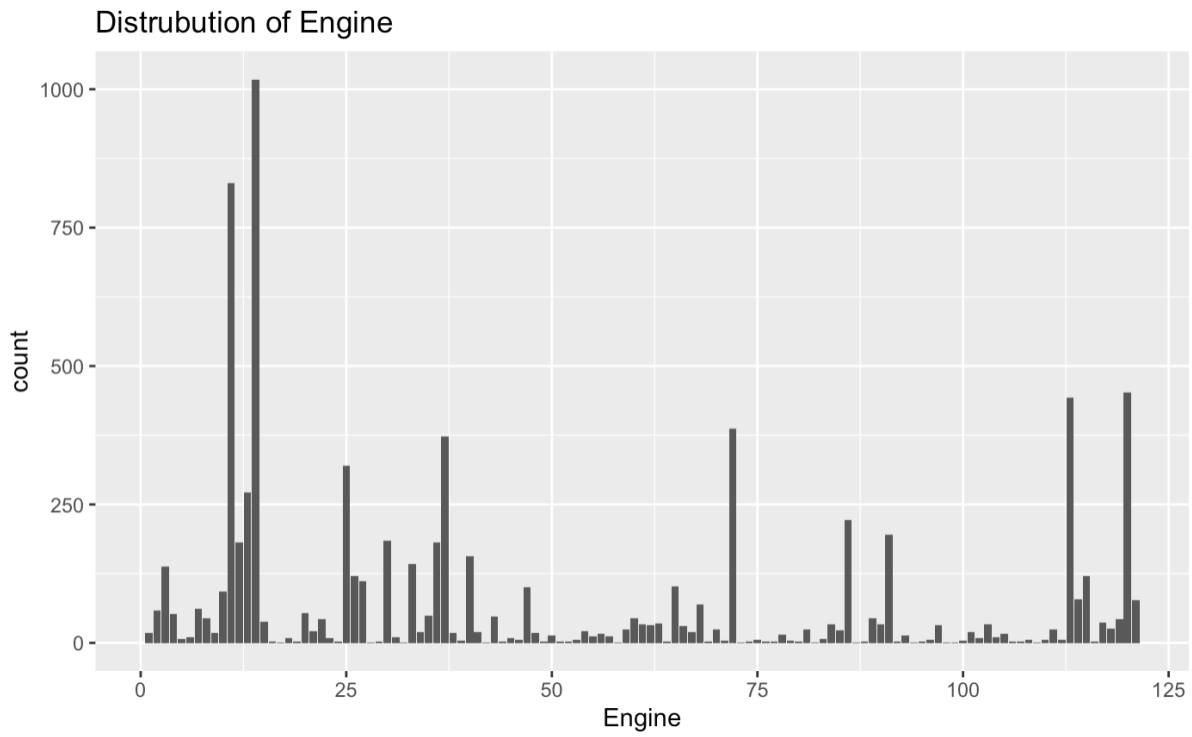

Manufacturer wrt Seller Type
6000
4000
2000
0
count
Dealer
Individual
Trustmark Dealer
seller_type
Manufacturer
Germany
Japan
other Asia
other Europe
US

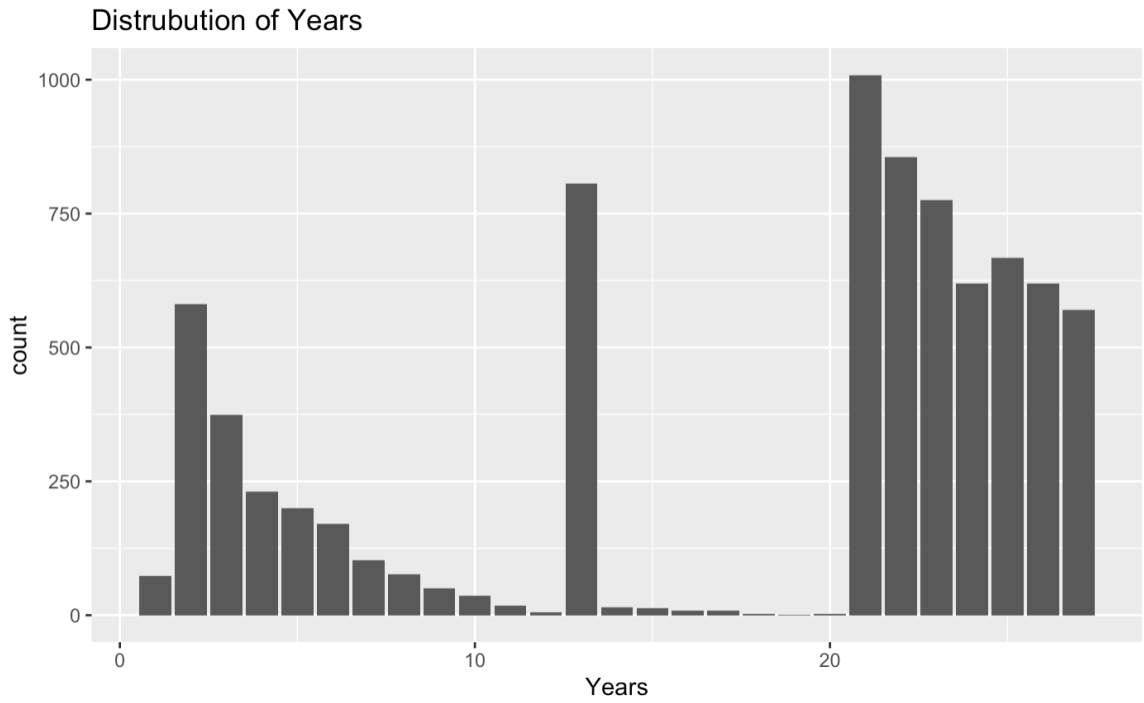

Seller Type wrt Manufacturer
3000
2000
1000
0
count
Germany
Japan
other Asia
other Europe
US
Manufacturer
seller_type
Dealer
Individual
Trustmark Dealer

Integer variables:

From nine explanatory variables, Years, Engine, Mileage and km_driven are integer variables. We can understand them better through histograms below. In this case, we choose Years, Mileage and Engine to make plots due to limited length of this paper.







4 Method

Model Diagnostic

There are two model assumptions for multiple linear regression that we have to check before we decide our final model:

1. The response is linearly related to the predictors.
2. The errors are i.i.d. normally distributed, with zero mean and constant variance.

First, we look at the residuals versus fitted values plot. With this plot we want to look for any pattern in the data. Ideally, we should see points distributed around zero, with the same variance throughout the whole space.

Next, we look at the Normal Q-Q (quantile-quantile) plot to check the normality of errors assumption. Ideally the points should fall on the straight dotted line.

Also, we look at the scale-location plot. This is used to check that the residuals display equal variance along the model plane. With this plot we want to see a horizontal line with points evenly spread above and below it.

Then, we look at the residuals vs. leverage plot. This is used to find influential observations in the dataset.

Lastly, we look at the residuals on the y-axis and their index on the x-axis to check for correlation between the residuals. Again here we want to see an evenly scattering around e=0.

According to the Scatterplot Matrix of Response Variables and Explanatory Variables, we can check whether or not the response variable is linearly independent from the predictors. Here, we want to see dots randomly spread in each block.

Moreover, we used VIF(Variance Inflation Factors) to detect the multicollinearity. The VIF for each x in the multiple linear regression model measures the combined effect of the linear dependencies among the regressors and that term. When VIF equals 1, it means the variance is not inflating at all. If one, or more, of the VIFs are large (say greater than 5 or 10) it is an indication that the associated regression coefficients are poorly estimated due to multicollinearity.

Model Building

After checking all the assumptions are met, we introduced three methods we have used to build our model:

1. Variable Selection (Stepwise Regression: forward selection, backward elimination)
2. Robust Regression (Huber's t function)

3. Cross Validation

Variable selection allows us to make the model as realistic and simple as possible. To achieve this goal, we used the stepwise regression methods, including forward selection and backward elimination.

Then with the Akaike's Information Criterion (AIC), we can estimate the relative amount of information lost by a specific model. The less information a model losses, the higher the quality of that model. Thus, we want to minimize the AIC to obtain the best fitted model. In the model selection step, we would set 2 base models, one full model with all the explanatory variables, one least model with least number of explanatory variables. Then we would use stepwise model selection to set 3 models with the AIC information criteria, one from stepwise model selection, one from backward direction, and one forward direction, and both direction.

Full Model

Model_Full E(selling_price) = β0 + β1*Maunufacturer + β2*years + β3*Mileage + β4*Engine + β5*Max_power + β6*km_drive + β7*fuel + β8*seller_type + β9*transmission + β10*owner + β11*seats

, where the variables Maunufacturer, fuel, seller_type, and transmission are categorical variables.

Forward Selection Model

Model_Forward. E(selling_price) = β0 + β1*Max_power + β2*seats

Backward Elimination Model

Model_Backward. E(selling_price) = β0 + β1*Engine + β2*mileage + β3*Years + β4*km_driven + β5*seller_type + β6*owner + β7*Manufacturer + β8*fuel + β9*transmission

, where the variable seller_type, Manufacturer, fuel, and transmission are categorical variables.

Stepwise Regression Model

Model_Both. E(selling_price) = β0 + β1*Max_power + β2*seats + β3*Engine + β4*mileage + β5*Years + β6*km_driven + β7*seller_type + β8*owner + β9*Manufacturer + β10*fuel + β11*transmission

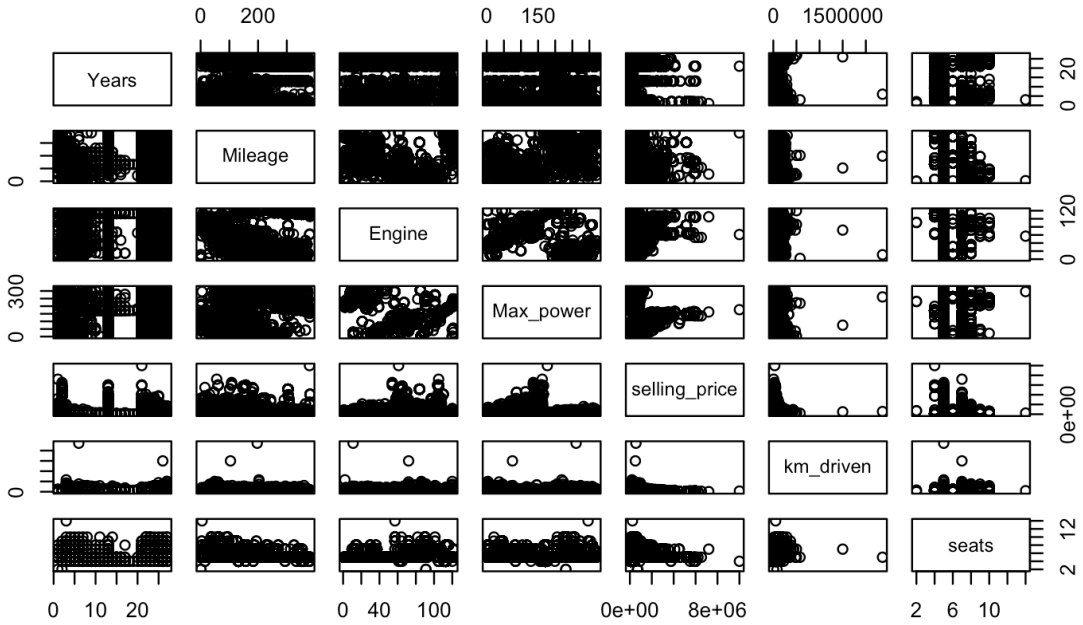
, where the variable seller_type, Manufacturer, fuel, and transmission are categorical variables.

Since our data contains many qualitative variables and the data is far from normal which have unequal variances and contain outliers, we cannot transform the data to follow the model assumptions. We decide to build a robust Regression using Huber's t function.

After building our model based on the robust regression, we will use cross validation to investigate the prediction performance of our model. We split data into 80% for training the model and 20% of the data for testing the model. Then we fit the model using the training data and predict the selling price on cars using the testing data. Finally we will evaluate the quality of our prediction.

5 Results

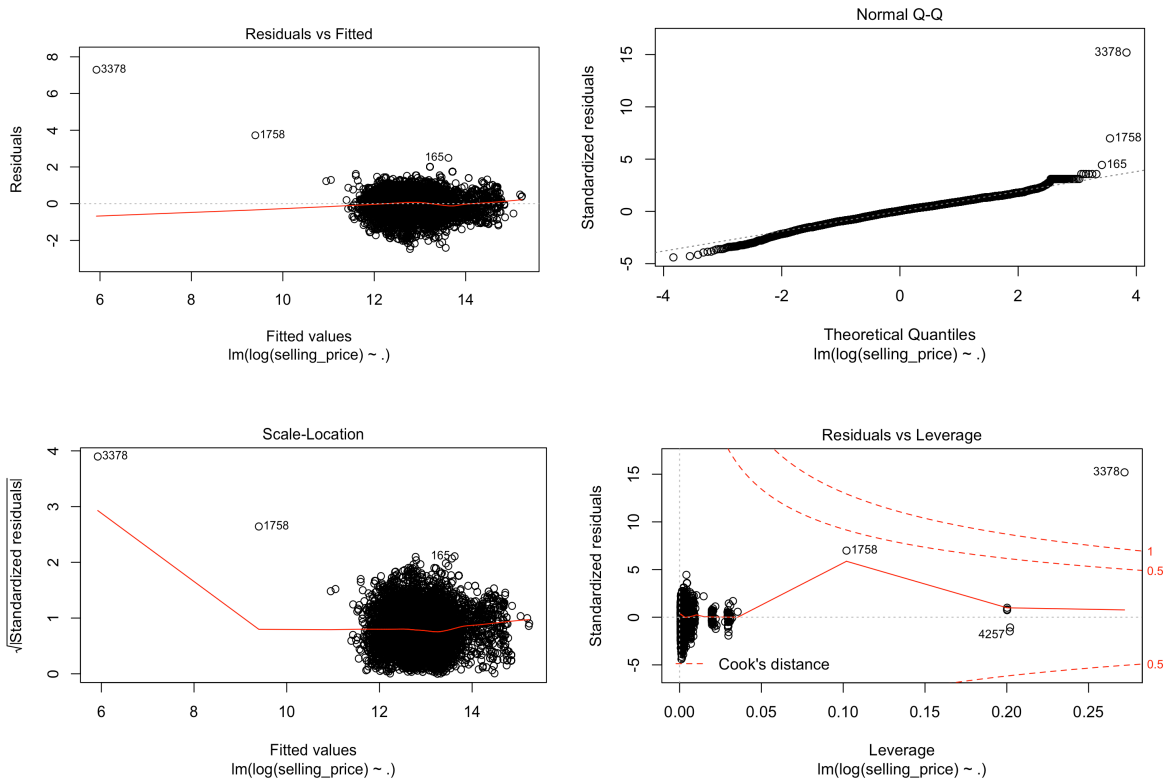
Scatterplot Matrix of Response Variables and Explanatory Variables



pairs(car2)

This plot is a scatterplot matrix of response variables and explanatory variables. We observe that all the small blocks showing no potential linear trend, meaning there is no obvious linear relationship between response variable (selling price) with each explanatory variable. All the variables are linearly independent from each other.

The Log Transformed Regression Model




```
log1.lm <- lm(log(selling_price) ~ . , data = car4)
```

First, we plot the linear regression with log transferred data.

Based on the residuals versus fitted values, our model seems to be adequate with respect to the linear relationship assumption. However, there are three problematic observations, 165th, 1758th and 3378th data points respectively.

Based on the Normal Q-Q plot, the residuals also meet the normality assumption except those three points.

The scale-location plot shows that the assumption of constant variance is met for the majority part of the data.

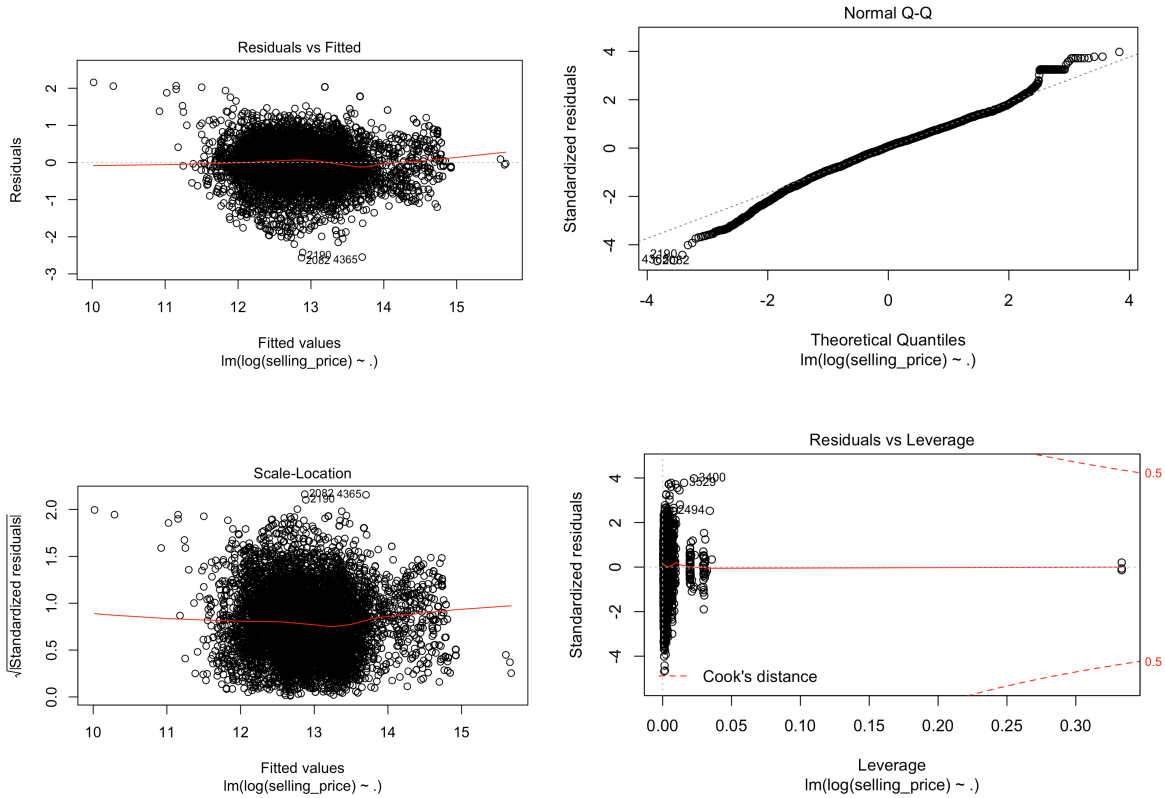
The residuals vs. leverage plot tells us point 3378 is an influential point since it appears outside of Cook's Distance line marked by the dotted red line.

VIF of the Log Transformed Regression Model with Outliers Omitted

```
vif(log.lm)
##                GVIF Df GVIF^(1/(2*Df))
## Manufacturer 1.542866  4        1.055701
## Years        1.109696  1        1.053421
## Mileage      1.312655  1        1.145712
## Engine       1.117918  1        1.057316
## km_driven    1.486285  1        1.219133
## fuel         1.333576  3        1.049147
## seller_type  1.349528  2        1.077818
## transmission 1.457660  1        1.207336
## owner        1.247930  4        1.028073
```

We observed that the VIF(Variance Inflation Factors) for all predictors approximately equals 1, indicating there is no multicollinearity between each predictors. The regression coefficients of our model are well estimated.

The Log Transformed Regression Model with Outliers Omitted




```
log.lm <- lm(log(selling_price) ~ ., data = car)
```

We have removed several problematic data points from the data set and constructed a new log transformed linear model.

Based on the residuals versus fitted values, the range of residuals becomes smaller, meaning our model is more linear and better than the previous one.

Based on the Normal Q-Q plot, the residuals meet the normality assumption.

The scale-location plot shows that the assumption of constant variance is met.

Based on the residuals vs. leverage plot, we no longer see any influential points or outliers.

The Summary of the Log Transformed Regression Model with Outliers Omitted

Table 1: Summary of Log Transformed Model

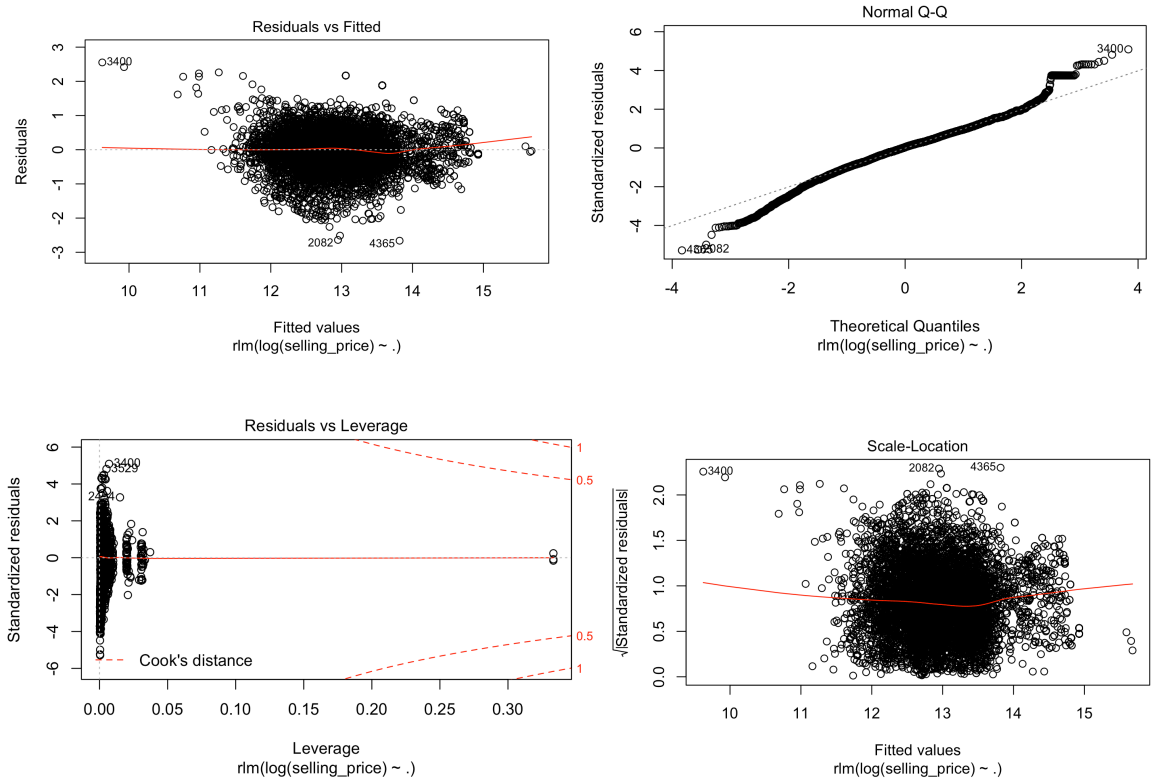
F-statistic	558.4
F-stat P-value	2.2E-16
R-square	0.5597

```
Call:
lm(formula = log(selling_price) ~ ., data = car)

## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.458e+01  8.531e-02 170.862  < 2e-16 ***
## ManufacturerJapan            -8.042e-02  2.896e-02  -2.777  0.00551 **
## Manufacturerother Asia       -2.371e-01  2.907e-02  -8.155 4.05e-16 ***
## Manufacturerother Europe     -5.318e-03  3.698e-02  -0.144  0.88565
## ManufacturerUS               -3.343e-01  3.445e-02  -9.702  < 2e-16 ***
## Years                         5.242e-03  7.528e-04   6.963 3.60e-12 ***
## Mileage                      -1.215e-03  7.326e-05 -16.588  < 2e-16 ***
## Engine                       -1.770e-03  1.648e-04 -10.740  < 2e-16 ***
## km_driven                    -4.565e-06  1.567e-07 -29.122  < 2e-16 ***
## fuelDiesel                    4.789e-01  7.726e-02   6.199 5.98e-10 ***
## fuelLPG                      -3.545e-01  1.202e-01  -2.949  0.00320 **
## fuelPetrol                   -2.231e-01  7.710e-02  -2.894  0.00382 **
## seller_typeIndividual        -2.022e-01  1.982e-02 -10.203  < 2e-16 ***
## seller_typeTrustmark Dealer  1.271e-02  4.063e-02   0.313  0.75439
## transmissionManual           -8.099e-01  2.203e-02 -36.758  < 2e-16 ***
## ownerFourth & Above Owner    -7.375e-01  4.500e-02 -16.389  < 2e-16 ***
## ownerSecond Owner            -3.724e-01  1.532e-02 -24.308  < 2e-16 ***
## ownerTest Drive Car           1.572e+00  3.179e-01   4.943 7.86e-07 ***
## ownerThird Owner             -5.561e-01  2.666e-02 -20.862  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We observed that the estimates of Manufacturer Asia, Manufacturer US, fuel type, seller type, transmission type, and owner type have relatively larger values indicating these variables have the most significant impacts on the selling price of used cars.

The Robust Regression Model




```
robust_huber.lm <- rlm(log(selling_price) ~., data = car, psi = psi.huber)
```

Then, we plot the robust regression using the Huber's t function.

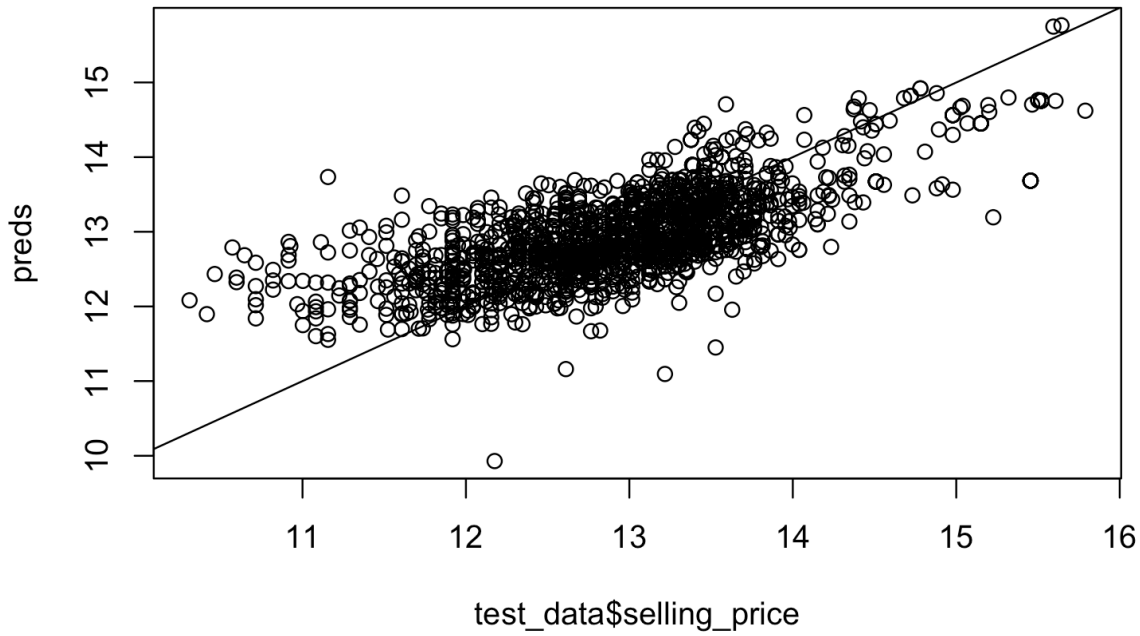
Based on the residuals versus fitted values, the linear relationship assumption is met but we still get several problematic observations.

The Normal Q-Q plot shows most of the residuals meet the normality assumption, but the left and right tails are still a little bit off.

The scale-location plot shows that the assumption of constant variance is met except for a couple data points.

Based on the residuals vs. leverage plot, we don't see any influential points or outliers.

The Testing Data v.s. the Predictions



This plot is the selling price from our testing data versus predictions of selling price we have made using our final model. We observe that most of our dots spread around this black line (test = preds). Thus, we suggest our model is good enough to predict the selling price of any cars.

6 Conclusions

According to the final model, we can conclude that the selling price of a used car is linearly related to:

1. The manufactured location of the car (Manufacturer)
2. The number of years since a car was produced (year)
3. The mileage of the car (Milage)
4. The parameter of car engine, measured in cc (Engine)
5. The distance that the car has travelled (km_driven)
6. The fuel type of a car (fuel)
7. Whether the owner of a car sell it individually or through a dealer (seller_type)

8. Whether it is automatic or manual transmission (transmission)
9. How many times the car has been trade (owner)

According to the evaluated quality of the prediction model, and based on the summary result of AIC model, we observed several significant predictors which heavily affects the selling price.

Firstly, we focus on favor variables, the variable “Manufacturer” directly affects selling price. We divided this into four categories, found vehicles produced in Europe tends to have the lowest selling price. Another important variable need to be mentioned is “owner”, it is clear that as the number of times the car has been trade increases, the selling price decreases. And the cars which relies on Diesel fuel have higher selling price than others.

For all the numeric variables, “years” and “km_driven” are the most influential variables. Vehicles is getting cheaper as the distance that the car has travelled and the age of them increases.

The prediction model we build was tested by using root-mean-square-prediction error(RMSPE) and root-mean-square error(RMSE), the results are 0.556752 and 0.427820, respectively. According to these numbers, the accuracy of this prediction can still be improved. We found one of the limitation of this study is the high number of variables that have been used. As future work, we intend to do variable selections using other machine learning techniques like k-nearest neighbors (kNN) to locate the best matches. Thus we can have a better presentation of which variables are the most influential affecting the selling price of a used car.

7 References

1. Erprose. (2020, October 16). The used car boom is one of the hottest, and trickiest, coronavirus markets for consumers. Retrieved December 08, 2020, from https://www.cnbc.com/2020/10/15/used-car-boom-is-one-of-hottest-coronavirus-markets-for-consumers.html
2. D'Allegro, J. (2020, August 28). Just What Factors Into The Value Of Your Used Car? Retrieved December 07, 2020, from https://www.investopedia.com/articles/investing/090314/just-what-factors-value-your-used-car.asp
3. C. (n.d.). What Should I Charge for My Used Car? Retrieved December 07, 2020, from https://www.carfax.ca/resource-centre/articles/how-much-should-i-charge-for-my-used-car
4. Threewitt, C. (2019, August 22). 10 Factors That Affect Your Car's Resale Value. Retrieved December 07, 2020, from https://auto.howstuffworks.com/buying-selling/car-resale-value.htm
5. Birla, N. (2020, October 24). Vehicle dataset from cardekho. Retrieved December 07, 2020, from https://www.kaggle.com/nehalbirla/vehicle-dataset-from-cardekho
6. Muir, A. (n.d.). Torque and BHP explained. Retrieved December 07, 2020, from https://www.howacarworks.com/technology/torque-and-bhp-explained

8 **Appendix**

All data and code in this link: https://github.com/kunpeng-wang-sfu/stat350-final-project-group8