HS5004 Econometrics – Term Paper

A study on predicting resale price of used cars

Subhash S – 111801042

Computer Science and Engineering, IIT Palakkad

Contents

[Abstract 3](#_Toc102595872)

[Motivation 3](#_Toc102595873)

[Dataset 3](#_Toc102595874)

[Dataset statistics 4](#_Toc102595875)

[Exploratory Data Analysis (EDA) 5](#_Toc102595876)

[Basic statistics 5](#_Toc102595877)

[Univariate analysis 6](#_Toc102595878)

[Bivariate Analysis 7](#_Toc102595879)

[Possible pre-processing steps: 8](#_Toc102595880)

[Gauss-Markov assumptions 8](#_Toc102595881)

[The model is linear in parameters and correctly specified 8](#_Toc102595882)

[The disturbance has zero expectation 8](#_Toc102595883)

[No exist an exact linear relationship among the regressors (No multicollinearity) 9](#_Toc102595884)

[The disturbance term is homoscedastic 9](#_Toc102595885)

[The values of disturbance term have independent distributions 10](#_Toc102595886)

[The disturbance term has a normal distribution 10](#_Toc102595887)

[Processing dataset 10](#_Toc102595888)

[Removing data outliers 10](#_Toc102595889)

[Dropping variables with high Variance Inflation Factor (VIF) 12](#_Toc102595890)

[Results 13](#_Toc102595891)

[References 14](#_Toc102595892)

# List of tables

[Table 1: Basic data statistics 4](#_Toc102637746)

[Table 2: An example row from the dataset 4](#_Toc102637747)

[Table 3: Statistics of numerical columns in the dataset 5](#_Toc102637748)

[Table 4: VIF of independent variables and possible reasons for high correlation 12](#_Toc102637749)

# List of figures

[Figure 1: Blot plots between frequency and each categorical variable 6](#_Toc102637750)

[Figure 2: Scatter plot between numerical features and the dependent variable 7](#_Toc102637751)

[Figure 3: Plot of residuals obtained after training 8](#_Toc102637752)

[Figure 4: Correlation plot for all pairs of variables 9](#_Toc102637753)

[Figure 5: Plot of residuals vs fitted values 9](#_Toc102637754)

[Figure 6: Plot of numerical data before removing the outliers 10](#_Toc102637755)

[Figure 7: Plots of numerical data after dropping the outliers 11](#_Toc102637756)

[Figure 8: OLS Regression results 13](#_Toc102637757)

[Figure 9: R-squared score on train and test data 13](#_Toc102637758)

[Figure 10: (1) Residuals are distributed according to the normal distribution, (2) Predicted and actual selling price have a near linear relationship, (3) Residuals are a constant distance from the line. 14](#_Toc102637759)

# Abstract

As the production of cars go up each year, the sales of used cars also contribute a significant portion of income for the automobile industry. With multiple start-ups focusing on the used car segments with renting and subscription models, it becomes an important part of their business to calculate the price of the vehicle after it has been used for a certain period. This study aims to achieve the task by using the *Ordinary Least Square (OLS)* regression over multiple independent variables.

# Motivation

The used car market is an ever-rising industry with rapidly growing market share and multiple companies like *Zoom Car*, *Car Dekho* etc. competing to make their mark in this segment. While these companies also have subscription models, those cars would eventually be sold at the end of a stipulated period.

Deciding the selling price of used cars and buyers ascertaining the worth of the posted prices on online sites can prove to be difficult due to various factors involved in finding the resale price. This study uses the OLS estimator to ascertain the retail value of the car on the date of sale.

# Dataset

For this study, we use the dataset from *Car Dekho* dataset on Kaggle that is referenced at the end of this document. This dataset consists of 9 columns that are discussed below:

1. **Car\_Name**: The name of the car that is under consideration. It is a text column with multiple values.
2. **Year**: This column represents the year in which the car was bought. It is an integer column of the form YYYY.
3. **Selling\_Price**: This is the selling price of the car. This variable is our dependent variable, and we need to predict this variable using the OLS estimator.
4. **Present\_Price**: This column tells the present price of the car i.e., the current price of the car when bought as the first-hand user.
5. **Kms\_Driven**: This column tells the total distance that the car has travelled from the date of purchase of the car.
6. **Transmission**: This column tells if the car is a manual of automatic transmission.
7. **Owner**: It tells the number of owners that the car has previously had.

# Dataset statistics

In this section, we will be seeing some important statistics about the raw dataset which will help us decide the pre-processing steps that has to be performed.

The dataset has a total of 301 rows and 9 columns. Various statistics about the 9 columns are listed in Table 1.

Table 1: Basic data statistics

|  |  |  |  |
| --- | --- | --- | --- |
| # | Column | Non-null count | Data type |
| 1 | Car\_Name | 301 | string |
| 2 | Year | 301 | integer |
| 3 | Selling\_Price | 301 | float |
| 4 | Present\_Price | 301 | float |
| 5 | Kms\_Driven | 301 | int |
| 6 | Fuel\_Type | 301 | Enum {Petrol | Diesel | CNG} |
| 7 | Seller\_Type | 301 | Enum {Dealer | Individual} |
| 8 | Transmission | 301 | Enum {Manual | Automatic} |
| 9 | Owner | 301 | int |

Table 2: An example row from the dataset

|  |  |
| --- | --- |
| Column | Value |
| Car\_Name | Bajaj Avenger 220 |
| Year | 2017 |
| Selling\_Price (in Lakhs) | 0.90 |
| Present\_Price (in Lakhs) | 0.950 |
| Kms\_Driven | 1300 |
| Fuel\_Type | Petrol |
| Seller\_Type | Individual |
| Transmission | Manual |
| Owner | 0 |

# Exploratory Data Analysis (EDA)

EDA is used in data science projects to get closer to certainty that the future results will be valid, correctly interpreted, and applicable to the desired problem statement. This is done by validating the raw data, checking for anomalies, and ensuring the dataset was collected without errors and noise.

## Basic statistics

Table 3: Statistics of numerical columns in the dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | selling\_price | present\_price | kms\_driven | age |
| count | 301.000000 | 301.000000 | 301.000000 | 301.000000 |
| mean | 4.661296 | 7.628472 | 36947.205980 | 8.372093 |
| std | 5.082812 | 8.644115 | 38886.883882 | 2.891554 |
| min | 0.100000 | 0.320000 | 500.000000 | 4.000000 |
| 25% | 0.900000 | 1.200000 | 15000.000000 | 6.000000 |
| 50% | 3.600000 | 6.400000 | 32000.000000 | 8.000000 |
| 75% | 6.000000 | 9.900000 | 48767.000000 | 10.000000 |
| max | 35.000000 | 92.600000 | 500000.000000 | 19.000000 |

The above statistics show a basic analysis for the range and the spread of the dataset. These values can be analysed and looking at the 75th percentile and max values, we can expect that the dataset has outliers which need to be removed.

## Univariate analysis

This dataset has five categorical variables, namely, *Car\_Name, Fuel\_Type, Seller\_Type, Transmission* and *Owner.* Just by some simple inspection, it becomes very apparent that having *Car\_Name* feature in the training set would lead the model to overfit since it can learn a one-one mapping from car name to the selling price, so I would not be analysing this variable further.

The first step involves plotting the bar-plots for all the categorical variables and deduce some conclusions from it. The below bar plots show the count of each categorical variable (on Y-axis) and the feature under observation (on X-axis).

Although the dataset is skewed, as we will be seeing in the bi-variate analysis, the *Present\_Price* variable has a good linear relationship with the *Selling\_Price,* so we can expect to get a good performance with the dataset.

Figure 1: Blot plots between frequency and each categorical variable

Chart, bar chart, waterfall chart

Description automatically generated

## Bivariate Analysis

In the previous section, we saw some bar-plots of each categorical variables, in this section we will be plotting the numerical variables against the *Selling\_Price* to see what kind of relationships they exhibit. To perform this bivariate analysis, we will be seeing some scatter plots of the numerical variables.

Figure 2: Scatter plot between numerical features and the dependent variable

Chart, line chart, scatter chart

Description automatically generated

***Note:*** The first image is plot with same features on both X and Y-axis, hence the straight line.

We can see that the *Present\_Price* has a near-perfect linear relationship with the *Selling\_Price*, this shows us that it will be influencing the *Selling\_Price* to a large extent. Looking at the numerical plots, we can spot the outliers here too, so removing outliers will have to be considered as a necessary pre-processing step.

## Possible pre-processing steps:

1. To make it uniform and to ease string comparisons (if any), I converted all the column names to lowercase, i.e., the column ***Owner*** becomes ***owner*** and the same applies to all other columns.
2. The ***Year*** column has values like 2014, 2015 etc., here it would make more sense to represent this column as the age of the vehicle as the price is influenced by the age of the vehicle. This can be done by subtracting the values from the current year (2022).
3. Since linear regression models take in numbers, it would be more appropriate to convert the enumerated columns (***Fuel\_Type, Seller\_Type, Transmission)*** to numbers.

For ex: The ***Transmission*** column has two values, hence the value *Manual* can be assigned a value of 0 and the value *Automatic* can be assigned a value of 1.

1. Since ***Car\_Name*** can be very different for each observation, it is not necessary to use it in the final training dataset, so this column can be dropped from the dataset.
2. From the analysis, we have seen that some columns have data outliers, and they will have to be removed from the dataset.

# Gauss-Markov assumptions

To justify the usage of multiple linear regression, we need to satisfy the Gauss-Markov assumptions. These assumptions and the justification are given below:

## The model is linear in parameters and correctly specified

This assumption holds in our case since our model is linear in all variables in the dataset i.e., the model can be represented as

## The disturbance has zero expectation

The below plot shows the plot of residuals vs their frequency. This plot shows that the residuals follow a normal distribution with 0 mean, hence this assumption is valid in our case.

Figure 3: Plot of residuals obtained after training

Chart, histogram

Description automatically generated

## No exist an exact linear relationship among the regressors (No multicollinearity)

This can be verified from the correlation plot. If two variables have a correlation factor of 1, then they are said to be in a perfect linear relation, but as we can see that none of the features are in perfect correlation with the *Selling\_Price*.

Figure 4: Correlation plot for all pairs of variables

A picture containing chart

Description automatically generated

## The disturbance term is homoscedastic

This assumption states that our error terms have a constant variance (no heteroscedasticity). The figure below shows that the residuals are at a constant distance from the line except for some outliers, hence we are safe to conclude that our assumption of homoscedasticity holds.

Figure 5: Plot of residuals vs fitted values

Chart, scatter chart

Description automatically generated

## The values of disturbance term have independent distributions

This dataset was published by the company *Car Dekho,* and we can assume that the data was randomly sampled based on the activities on their site, thus we satisfy this assumption. Also, from Figure 3, we can see that the residuals are randomly distributed along the horizontal line.

## The disturbance term has a normal distribution

As we can see from Figure 3, the residuals or the disturbance term has a normal distribution with mean 0, hence this assumption holds.

We have shown that all the six assumptions of the Gauss-Markov theorem hold true in our case, hence the use of a multiple linear regression or the OLS regression model is justified.

# Processing dataset

Before any training task, it is wise to process the dataset by removing outliners, reducing multicollinearity, and performing other suitable steps depending on the dataset.

## Removing data outliers

Outliers are those points in the dataset that are at the extremes with a very low density and usually indicate an erratic measurement, noise, or other unwanted factors. Such outliers disturb the accuracy of the model and lead to unreliable results.

Figure 6: Plot of numerical data before removing the outliers

Chart, bar chart

Description automatically generated

As we can see from Figure 6, there are some points that lie outside 1.5 x Inter Quartile range (IQR), we remove those points from the dataset to reduce noise.

The algorithm to remove outliers is as follows:

1. Find first quartile i.e., the 25th percentile (Q1).
2. Find the third quartile i.e., the 75th percentile (Q3).
3. The IQR is defined as Q3 – Q1.
4. Now, we remove all the points outside the range [Q1 – 1.5 \* IQR, Q3 + 1.5 \* IQR].

We can see from Figure 7 that almost all the outliers are removed from the desired columns. Thus, in this step, we have reduced unwanted discrepancies in the dataset by removing outliers.

Figure 7: Plots of numerical data after dropping the outliers

Chart, bar chart

Description automatically generated

## Dropping variables with high Variance Inflation Factor (VIF)

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable.

We assume a threshold of 10 for VIF and drop columns with higher threshold until all columns are below the threshold.

The algorithm is as follows:

1. Find VIF of all independent variables.
2. Drop the variable with highest VIF greater than the threshold and repeat from step 1.

Table 4: VIF of independent variables and possible reasons for high correlation

|  |  |
| --- | --- |
| Graphical user interface  Description automatically generated with medium confidence | Here, we can see that *fuel\_type\_Petrol* has the highest VIF of 22.5 and it is also above the desired threshold, so we will be removing this variable from the training set.  This is because *fuel\_type\_Petrol* is related to *fuel\_type\_Diesel* since vehicles are usually either of petrol or diesel. |
| Graphical user interface  Description automatically generated with medium confidence | After removing *fuel\_type\_Petrol*, the age variable still has a VIF of 14.8 which is also above the threshold.  This might be because *kms\_driven* and *age* might be in a strong linear relationship. |
| Table  Description automatically generated | After remove both the variables, we can see that are variables are below the required threshold. We can be sure that we have eliminated any possible multicollinearity effects on the dataset. |

# Results

We will be using the OLS estimator to perform a linear regression on the dataset. The dependent variable in the dataset is the *Selling\_Price* and the independent variables are show in the last image of Table 4.

Figure 8: OLS Regression results

A picture containing text, receipt, screenshot

Description automatically generated

From the above image, we can see that the R-squared is valued at 0.85 which is very close to the adjusted R-squared. This says that our model is not overfitting. We can further confirm this by checking the R-squared score on the train and test data respectively.

The below figure confirms our assumption that our model is not overfitting the dataset. We can also see the cross-validation results.

Figure 9: R-squared score on train and test data

Graphical user interface, text, application

Description automatically generated

# Plots

Figure 10: Plots obtained on the test dataset

Chart, scatter chart

Description automatically generated

1. The first plot (top-left) is the residual density plot. We can see that the residuals follow the normal distribution.
2. The second plot (top-right) shows the plot between predicted and actual values in the test dataset. Since there is a near linear relationship and there is not much deviation, we can infer that the model predicts *Selling\_Price* with a very good accuracy.
3. The third plot (bottom) shows the plot between the residuals and predicted *Selling\_Price*, we can see from the plot that the residuals are at a constant distance from the red line except for some outliers, thus we can conclude that residuals are randomly distributed, and the variance of error terms are equal.

# Interpretation

All inferences will be drawn from Figure 8.

## Practical significance

We have obtained an R-squared score of 0.84 on the dataset which is good for the size of data that we have. The *fuel\_type\_Diesel* has a positive value, which says that a diesel car usually adds to the *Selling\_Price* and this is intuitive since petrol cars are more affordable. The negative coefficient of *kms\_driven* is justified by the fact that cars with higher usage tend to be cheaper.

## Statistical significance

From the referenced figure, we can see the lack of statistical significance for the number of ownersand the transmission type of the car. This is because cars are usually priced based on the distance travelled and transmission type is merely a preference for individuals.

The high t-statistic of *Present\_Price* indicates that it has a strong significance in deciding the selling price which is also expected since current market value of a car is always taken into consideration to decide the resale value.

# Conclusion

Our study of analysing and predicting resale value of cars shows a promising model with good scores on test dataset. The t-tests and p-values complement the intuitive expectations on the resale value.

# References

1. [Used car details dataset by Car Dekho](https://www.kaggle.com/code/mohaiminul101/car-price-prediction/data?select=car+data.csv)
2. [Scikit-Learn Linear Regression](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html)
3. [Correlation plots using seaborn](https://datatofish.com/correlation-matrix-pandas/)
4. [Removing outliers](https://towardsdatascience.com/practical-implementation-of-outlier-detection-in-python-90680453b3ce)