ESTIMATION OF SELLING PRICE OF SECOND-HAND CAR

A report by Group 4

Animesh Gupta	2019B3AA0588H
Chinmay Goyal	2019B3AA1290H
Harsh Vardhan Gupta	2019B3A70630H
Hitesh Garg	2019B3A70466H
Sujay Nigam	2019B3AA1267H
Aryan Ramchandra Kapadia	2019B3A70412H
Dhruv Gupta	2019B3A70487H
Anand	2019B3A70613H

MARCH 19, 2022

BITS PILANI, HYDERABAD CAMPUS

Table of Contents

Section I: Introduction

Purchasing decision for second-hand cars is, in general, a tiring and tedious process. This is usually because in this market the seller has more information about the vehicle than the buyer. This asymmetrical information about the car is called the Lemon's Problem. The buyer does not want to pay more than the average price of the car, even if it is of premium quality. This benefits the seller if the car is a lemon but is a disadvantage if the car is of good quality. Thus, precise price estimates of second-hand cars are crucial for both buyers as well as sellers.

Though it has inherent problems, the market for used cars has been steadily growing. According to a report by Mordor Intelligence, the used car market in India is expected to grow at a CAGR of 15.1%. The market was valued at USD 32.14 Billion in 2021 and would reach USD 74.70 billion by 2027. With multiple companies like CarDekho, Cars24 and established players like Maruti True Value, Mahindra First Choice operating in the Indian market, the competition is heating up. Hence, it's essential to accurately predict these prices.

Forecasting of used car prices has been a hot topic of research. Various techniques like Regression Analysis, Machine Learning Algorithms, Neural Networks etc have been utilized to formulate models which give precise estimates of used car prices. The price is affected by various factors which include but are not limited to vehicle age, engine size, vehicle damage, type of vehicle, fuel efficiency, kilometres driven etc.

This research paper uses a multiple linear regression model with one dependent variable and nine independent variables. The relationship among the variables is linear.

Design: The research includes the formulation of a system explaining the linear relationship between X and Y which are price and other variables such as car age, engine size, mileage, etc.
Predict: The analysis seeks to predict the price of second-hand cars using a multiple linear regression model which looks for various patterns and accurately predicts their values.
Confirm: The analysis determines the most significant influencing variables for price.

The data for this research has been sourced from cardekho.com, and it consists of 15411 observations.

The organization of this paper is such that Section II is on related work, Section III discusses the methodology and Section IV presents the results and in-depth discussion. Section V concludes the paper and offers directions for future research.

Section II: Literature review

M.Richardson (2009), in his thesis, delved deeply into the resale market of cars and the factors that affect them. Its main focus was to study the effect of hybrid engines on resale value. It uses a comprehensive list of independent variables to accurately estimate the same. The paper concluded that age is negatively related to price since younger cars that are driven less will have a higher value. Secondly, mileage which is a measure of fuel efficiency is positively related to resale value because greater fuel efficiency will result in positive externalities and lower cost of ownership and operation will result in higher demand, thereby increasing prices. Thirdly, the manufacturer and safety ratings will play a significant role in positively affecting the prices. Due to this, the Japanese and European cars have added value to the resale price.

There are many components that affect the price of used cars. Genovese(1993) concludes **mileage** and **car condition** are important variables while deciding to purchase a car.

Gilmore and Lave(2013) concluded that the **size of the car engine** and its **type** has a significant impact on the price of the used vehicles. Engine sizes are usually measured in litres or cubic centimetres, while its type can be petrol engine, diesel engine, etc. A car's mileage and its age significantly affect the decision of the car-owner to sell the car, thus influencing the price of the car. The resale price is a function of fuel and vehicle type. In general, they concluded that the hybrid vehicles retained higher values as compared to their fuel counterparts. Based on similar lines, they also found that when the fuel prices were higher, higher fuel economy vehicles retained

greater value. Resale value differences were significantly associated with **future fuel savings** and the **cost of owning and operating** from the vehicle.

Meng et al. (2018) studied the used car market of Taiwan and conclude that having a **better engine** has a positive effect on the price though at a diminishing rate by incorporating a second-order engine realtor variable into the model. Moreover, the increase in the engine capacity by 1 litre, approximately increases the price of the car by 20.4%. The price is negatively impacted by the **age** of the car and its **mileage.** Change in age by one year affected the price by 14.7%**.** The resale value is positively impacted by **better body conditions**, good interiors. Additionally, type **of the car** (Dummy variable: SUV = 0, Sedan = 1) also affects the price: SUVs were on an average 16.9% more expensive than sedans. Finally, the resale value was also higher in the case of Toyota being the manufacturer.

Noor and Jan used multiple linear regression to predict used car values. They performed a variable-selection technique to determine the most important variables then leave out the remaining ones. The final data comprising of only selected variables are used to form the linear regression model. The primary aim of the paper was to develop an accurate model to predict the prices of used cars which accommodates various important attributes and features: model year, model, engine type and price, weeding out all the insignificant ones. The result was impressive with R-square = 98%.

Section III: Data and Methodology

Description of the dataset

The data was taken from cardekho.com to analyze how the selling price of secondhand vehicles depends on different parameters.

The dataset contains the following parameters:

- Car name - The name of the car that has been sold.
- Brand - The brand name of the car that has been sold.

- Model - The car model of a particular brand.
- Selling Price - The price at which the car was sold at cardekho.com.
- Min. Cost Price - The on-road price of the base model of the car.
- Max Cost Price - The on-road price of the top model of the car.
- Avg. Price - The average of min and max price of the car.
- Vehicle age - The number of years that the car has been used since first bought.
- Kms Driven - The number of kilometers that the car has been driven.
- Seller Type - The type of seller who sells the car. It can be an individual, a dealer or a trademark dealer.
- Fuel Type - The type of fuel that the car uses. It can be diesel, petrol or CNG.
- Transmission Type - The type of transmission of the car. It can be automatic or manual.
- Mileage - The distance(kms) that the car can travel in one liter of fuel type.
- Engine - Power output of the engine in CC.
- Max Power - The maximum power that the engine can generate.
- Seats - Number of seats in the car.
- Car Status - The status of the car depending on its price. It takes the value of luxury, medium and low.

Econometric Model

$$\mathbf{Y_{log(selling\_price)} = \beta_0 + \beta_1 log(avg\_price) + \beta_2 vehicle\_age + \beta_3 log(km\_driven) + \beta_4 seller\_type\_value + \beta_5 fuel\_type\_value + \beta_6 transmission\_type\_value + \beta_7 mileage + \beta_8 engine\_cc + \beta_9 seats + \beta_{10} car\_status\_value}$$

The selling price of the car depends on several factors. Some of the factors are directly related to the car price itself. For example, the price of the car at which it was bought, the age of the car and how much it has been driven. However, it might also depend on factors like seller type, fuel type, transmission type, the mileage of the car, the power of the engine, number of seats in the car and the luxury status of the car. In accordance with this view, a linear model has been estimated using OLS regression to predict cars prices for used cars. Here Y (the dependent variable) is the log of

the selling price of the car. A set of independent variables have been used to control for various observables. The variables and the reasoning are given below:

S. No.	Variable chosen	Description	Justification
1.	**log(selling_price)**	Dependent variable: log of car selling price.	To study the effect of different factors that affect car selling prices in India.
2.	**log(avg_price)**	The average of minimum and maximum price of the car.	The price of a secondhand car should depend on the original price at which it was bought.
3.	**vehicle_age**	The number of years that the car has been used since first bought.	It is expected that the price of the car decreases as it gets older.
4.	**log(km_driven)**	Log of the number of kilometers that the car has been driven.	The more a car has been driven, the worse its condition might get, so a negative relation on selling price is expected.
5.	**Individual, Dealer, Trustmark Dealer (Seller_type_value)**	There are dummy variables which indicate the type of seller.	People might prefer buying from a Dealer than an individual due to variety of options, and trust.
6.	**CNG, Diesel, Petrol, Electric, LPG (fuel_type_value)**	These are dummy variables which indicate the type of fuel in the car.	People may pay differently as the cost of different fuels are different.
7.	**Manual, Automatic (transmission_type_value)**	There are dummy variables which indicate whether a car is manual or automatic.	An automatic car may fetch more price than a manual car.
8.	**mileage**	The distance (in kms) that a car can travel in one liter of the fuel.	A car having more mileage should cost more.
9.	**seats**	Number of seats in the car	A car which has more seats should be more expensive.
10.	**Low-tier, Medium-tier, Luxury-tier (car_status_value)**	This dummy variable tells the status of the car depending on its price. It takes the value of luxury, medium and low.	A luxury car will cost more than the medium or low-level cars. This variable was used to control for the factor of brand value.

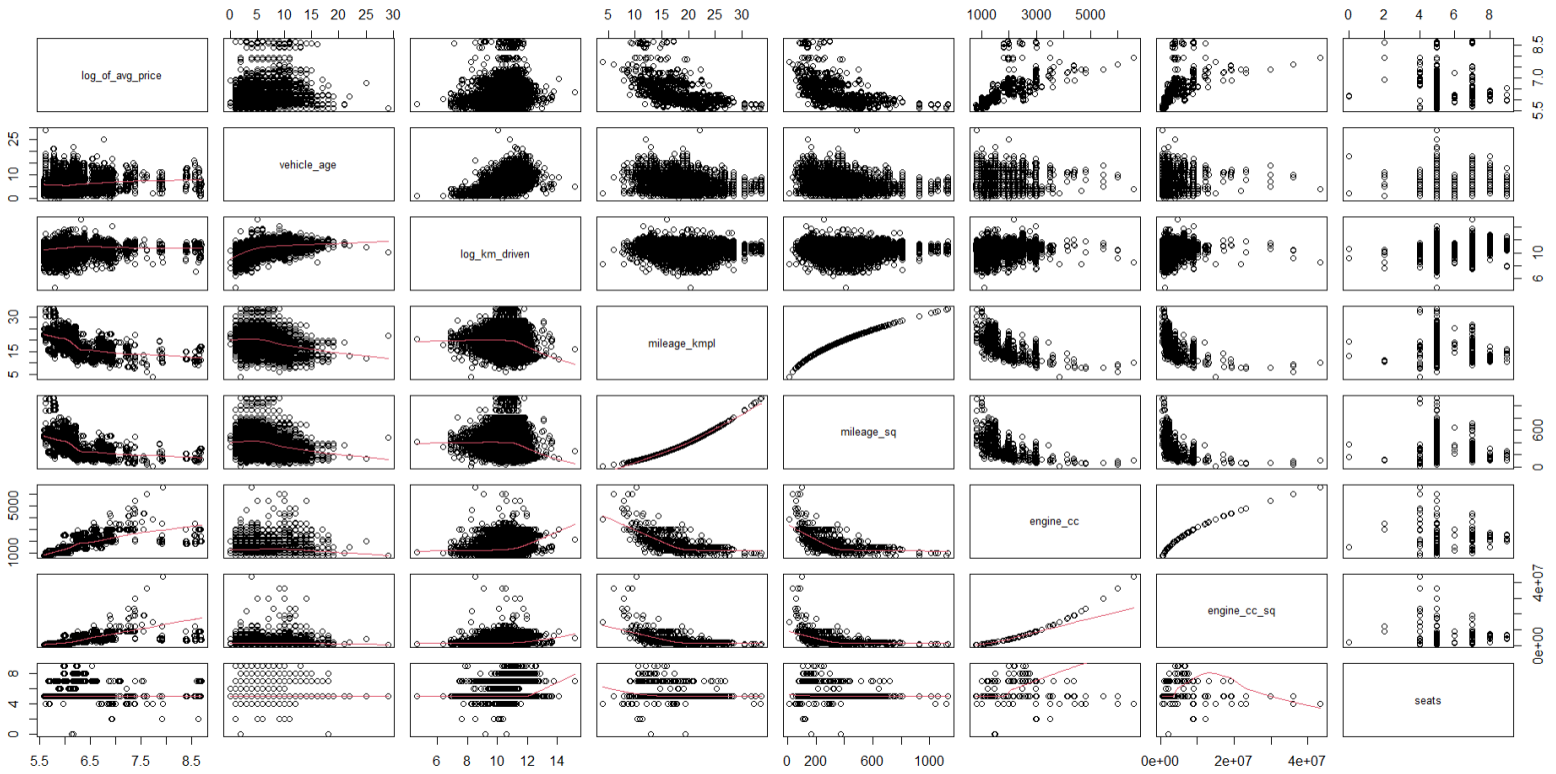
*It should be noted that the variable “Max Power” was not included in the model, since its effect would have been captured by the variable “engine_cc” that has been included in the model.

Steps to check the validity of the model

The model must satisfy the assumptions of the Multiple Linear Regression model. These will be checked using the following tests:

1. VIF test to check for multicollinearity.
2. Breusch-Pagan test for heteroskedasticity.
3. Jarque-Bera test for normality.
4. Ramsey RESET test for omitted variable bias.

Graph Matrix of the Regressors



The correlation between the explanatory variables is as shown above in the graph matrix. None of the variables are very strongly correlated with each other (except for engine_cc with engine_cc_sq, and mileage_kmpl and mileage_sq respectively). A few notable observations from this graph matrix are:

- Log_avg_price is positively correlated with engine_cc and engine_cc_sq.
- Vehicle age and mileage_kmpl are negatively correlated, which is intuitive since old cars would suffer from obsolete technology and depreciation of the car's components. A similar interpretation can be made for log_km_driven and mileage_kmpl.
- Log_km_driven and seats are positively correlated only at higher number of seats. This can be explained by the fact that bigger vehicles such as SUVs/MUVs are used extensively for long distance travelling, due to added benefit of carpooling.
- Mileage_kmpl and engine_cc are initially negatively correlated because there is a tradeoff between engine power and fuel efficiency in cars. However, this effect reduces after a particular value of mileage.
- Engine_cc and seats have a high positive correlation, which can be explained by the fact that bigger vehicles such as SUVs/MUVs are more powerful in general when compared to general purpose vehicles like sedans and hatchbacks.

Section IV: Results and Discussion

i) Summary statistics of the regression model

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.77751 -0.06043  0.00192  0.06372  0.85454

Coefficients:
                      Estimate Std. Error  t value Pr(>|t|)
(Intercept)           3.897e+00  4.297e-02   90.688  < 2e-16 ***
log_of_avg_price      3.108e-01  4.507e-03   68.955  < 2e-16 ***
log_km_driven        -1.600e-02  1.429e-03  -11.195  < 2e-16 ***
vehicle_age          -5.020e-02  3.684e-04 -136.244  < 2e-16 ***
seller_factor1        3.290e-02  1.834e-03   17.944  < 2e-16 ***
seller_factor2        1.639e-02  8.206e-03    1.998 0.045776 *
transmission_factor1  9.829e-02  2.574e-03   38.183  < 2e-16 ***
fuel_factor1          6.135e-02  3.034e-03   20.220  < 2e-16 ***
fuel_factor2          2.326e-02  6.958e-03    3.342 0.000833 ***
fuel_factor3          1.015e-02  1.609e-02    0.631 0.528037
fuel_factor4          6.672e-02  5.301e-02    1.259 0.208212
luxury_factor1        4.461e-02  3.950e-03   11.294  < 2e-16 ***
luxury_factor2        9.337e-02  4.713e-03   19.810  < 2e-16 ***
mileage_kmpl          1.421e-02  2.029e-03    7.005 2.58e-12 ***
mileage_sq           -3.629e-04  4.673e-05   -7.765 8.69e-15 ***
engine_cc             1.400e-04  1.081e-05   12.960  < 2e-16 ***
engine_cc_sq          5.336e-09  2.116e-09    2.522 0.011686 *
seats                -7.583e-03  1.559e-03   -4.865 1.16e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1055 on 15393 degrees of freedom
Multiple R-squared:  0.875,     Adjusted R-squared:  0.8749
F-statistic:  6340 on 17 and 15393 DF,  p-value: < 2.2e-16
```

Most of the explanatory variables are significant at less than 0.1% level of significance except fuel_factor3(LPG) and fuel_factor4(Electric) which are statistically insignificant. Also, seller_factor2(Trustmark Dealer) is significant at 5% and engine_cc_sq is significant at 1%.

R-squared is 0.8746 which means 87.46% of the dependent variable is explained by the model. Also, F-statistic is very high (6317) which means that all of the regression coefficients are jointly significant.

ii) Interpretation of the model

Most of the findings are as initially expected.

- **Intercept:** The intercept captures the case where all the dummy variables take the value 0. Thus, it represents the case where the car is a low-tier, manual, petrol car, sold by an individual. All the values of the coefficients and hence their ceteris paribus effects will be with respect to this base category.

- **Log_of_avg_price**: Average cost price of the vehicle is positively related to the selling price of the vehicle. 1% increase in cost price increases the selling price by 0.3113%

.

- **Log_Km_driven**: 1% increase in Km_driven will decrease the selling price of the vehicle by 0.016% which is intuitive because the more the car has been driven, the less the price for it.

- **Vehicle_age**: If the age of the vehicle increases by 1 unit, then the selling price of the vehicle will decrease by 5.02% which is also quite intuitive.

- **Seller_factor1(Dealer)**: The car sold by the dealer will have a value of 3.29% more as compared to a seller who is an individual.

- **Seller_factor2(Trustmark Dealer)**: The car sold by a Trustmark dealer will have a value of 1.639% more as compared to a seller who is an individual.

- **Transmission_factor1(Automatic)**: The car having an automatic transmission will have a value of 9.829% more as compared to a car whose transmission type is manual.

- **Fuel_factor1(Diesel)**: A car having fuel type diesel will have a value 6.135% more as compared to a car whose fuel type is petrol.

- **Fuel_factor2(CNG)**: A car having fuel type CNG will have a value 2.326% more as compared to a car whose fuel type is petrol.

- **Fuel_factor3(LPG)**: A car having fuel type LPG will have a value 1.015% more as compared to a car whose fuel type is petrol.

- **Fuel_factor4(Electric)**: An electric car will have a value 6.672% more as compared to a car whose fuel type is petrol.

- **Luxury_factor1(Mid-tier):** When the tier of the car changes from low-level to mid-level, the selling price increases by 4.461%.

- **Luxury_factor2(Luxury-tier):** When the tier of the car changes from low-level to luxury-level, the selling price increases by 9.337%.

- **Mileage_kmpl & Mileage_sq:** They both together signify the effect of mileage on the selling price of the car.
  It's of the form:

  $$\textbf{ß}_0\textbf{(mileage\_kmpl)} + \textbf{ß}_1\textbf{(mileage\_sq)}$$

  hence by differentiating the marginal effect of mileage on selling price can be shown as:

  1.421% - 2*0.03629*(mileage)%

It is observed that it is a diminishing effect which is also relevant in real life, because a consumer would be willing to spend more money if the mileage rises from 5 kmpl to 10 kmpl but won't spend the same amount for the rise in mileage from 40kmpl to 45kmpl.

- **Engine_cc and engine_cc_sq:** They both together signify the effect of engine's cc on the selling price of the car.

  It's of the form:

  $ß_0$(engine_cc) + $ß_1$(engine_cc_sq)

hence by differentiating the marginal effect of mileage on selling price can be obtained as:

0.014% - 2*5.336*10^(-7)*(engine_cc)%

Therefore, it is a diminishing effect which is also relevant in real life, because after a threshold of the engine's size, it becomes irrelevant to spend more on a bigger engine.

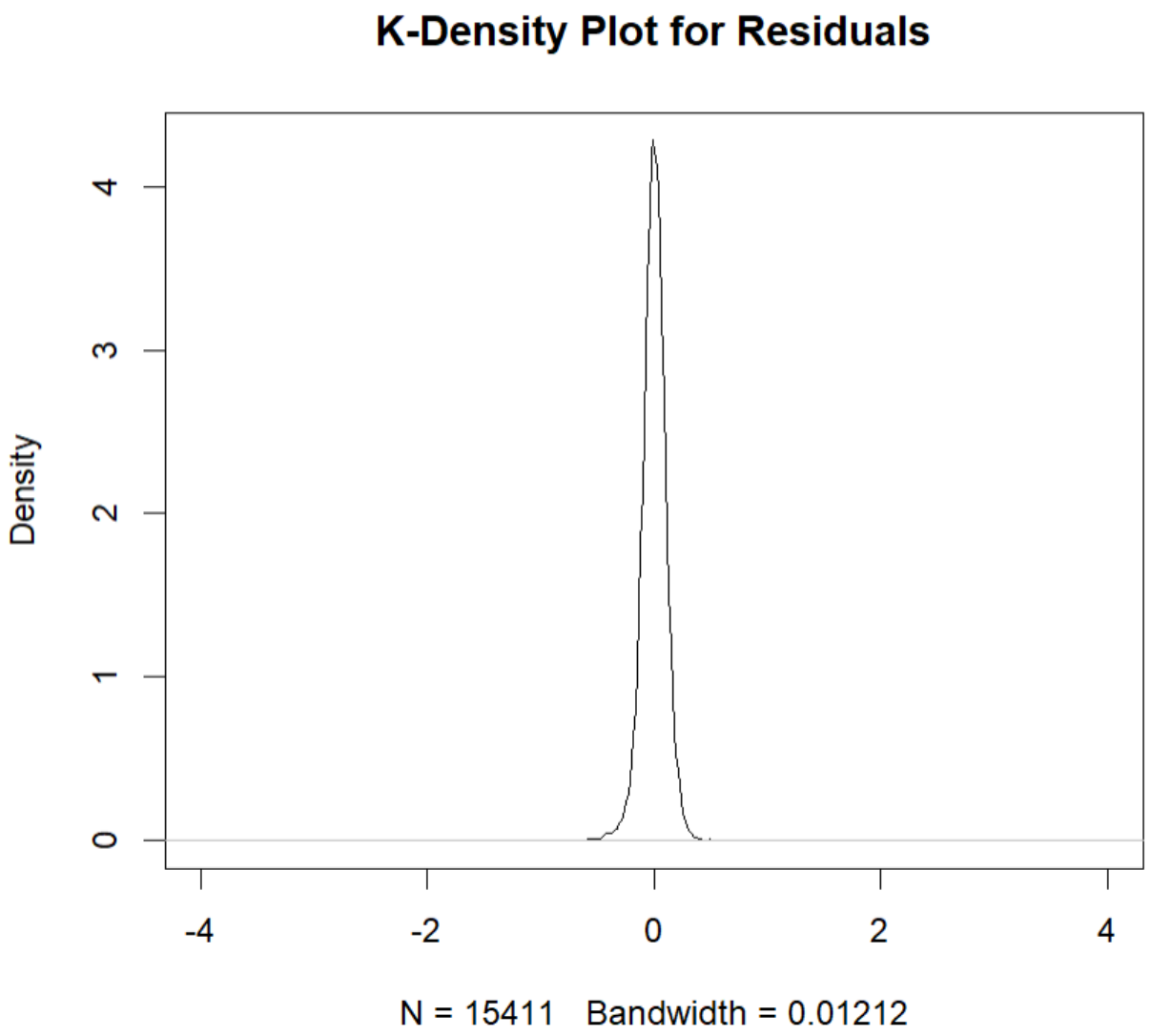
- **Seats:** It determines the effect of the number of seats in the car on its selling price. According to the regression model, the selling price of the used car changes by -0.7583% for every unit rise in the seats of the car. It's a rather unexpected finding as anyone would expect the price to rise with rising number of seats however, it can be explained by the fact that luxury and high-end cars have fewer seats (around 2-4) which results in a negative effect.

iii) Checking for the assumption of OLS

1) **Linear in parameters**: Since the assumed model has only a single $ß_i$ term for every explanatory variable/dummy variable, there is no product division or exponential form of the coefficient is present. Hence, it is linear in parameters.

2) **Random sampling of the data**: The data was collected by a car resale website (cardekho.com), with no specificity in the type of car, or any bias with respect to the data collected. It covers a wide range of observations, from the model of car to the fuel it uses and the type of dealer who sold the used car. Hence it can be concluded that the data was randomly sampled.

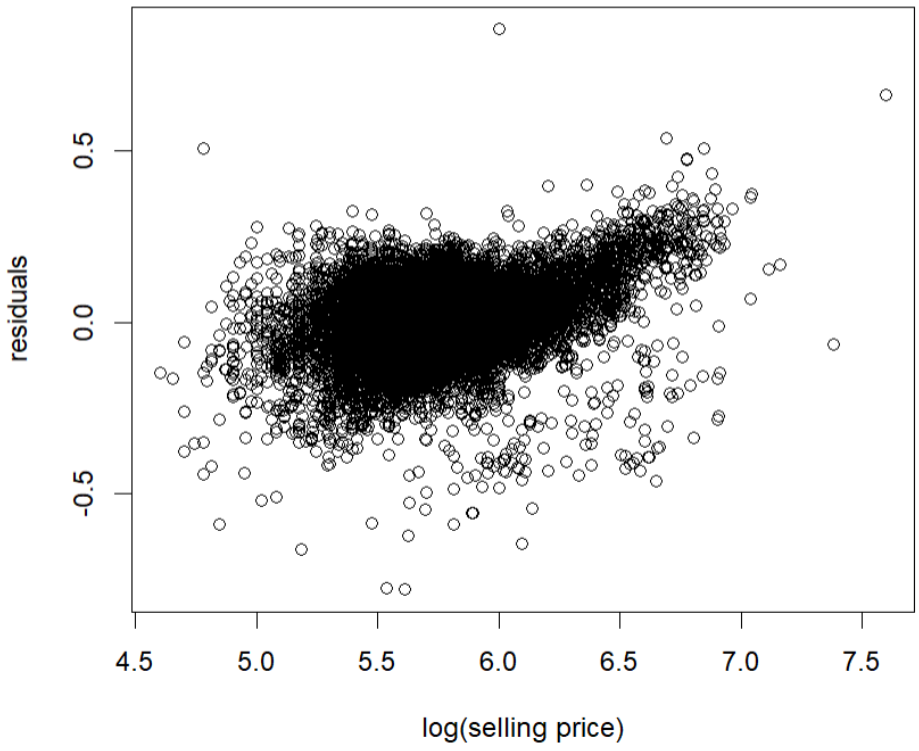
3) **Zero Conditional Mean** ($E(u_i|x_i) = 0$): This will be shown using kernel density plots.



We can observe that the plot of the residuals is centered at mean = 0 and is constant, hence it has a zero conditional mean.

4) **Homoscedasticity of error term**: The error term should have a constant variance over the interval of values of x_i. This can be checked via:

a) A scatter plot of residuals vs true values of dependent variable.



A funnel shape in the scatter plot signifies that variation is greater for lesser values of dependent variable and decreases for higher values. Hence, the model suffers from heteroscedasticity.

b) Breusch-Pagan test for heteroscedasticity

```
> (bp_test=bptest(reg1))

        studentized Breusch-Pagan test

data:  reg1
BP = 3627.1, df = 17, p-value < 2.2e-16
```

$H_{0:}$ The error variances are equal which means homoscedastic.

H_a: H_0 is false (Heteroscedastic)

Since the p-value in the test results is very close to 0, the null hypothesis of the BP test is rejected. Thus, heteroscedasticity exists in the model.

5) **Normality of Residuals**: It means that the population error is independent of independent variables and normally distributed with zero mean and variance sigma squared. This is one of the key assumptions of multiple regression model. The Jarque-Bera test can be used to verify this:

```
> (jbtest=jarque.bera.test(reg1$residuals))

        Jarque Bera Test

data:  reg1$residuals
X-squared = 4680.8, df = 2, p-value < 2.2e-16
```

$H_{0:}$ The error term is normally distributed.

H_a: H_0 is false (not normally distributed).

Since the p-value of the test result is very close to 0, the null hypothesis of Jarque Bera Test is rejected. Thus, error term is not normally distributed.

6) **Multicollinearity**: Multicollinearity occurs when the independent variables are correlated with each other in the model. In OLS estimation of coefficients, there should be no perfect correlation between two independent variables, partial correlations are allowed. To verify this the Variance Inflation Factor(VIF) test is performed.

```
> (multi=vif(reg1))
                        GVIF Df GVIF^(1/(2*Df))
log_of_avg_price    3.811185  1        1.952226
log_km_driven       1.697741  1        1.302974
vehicle_age         1.706611  1        1.306373
seller_factor       1.094453  2        1.022820
transmission_factor 1.504726  1        1.226673
fuel_factor         3.803582  4        1.181746
luxury_factor       2.388180  2        1.243131
mileage_kmpl       99.177774  1        9.958804
mileage_sq         85.176716  1        9.229123
engine_cc          43.898909  1        6.625625
engine_cc_sq       26.447782  1        5.142741
seats               2.194341  1        1.481331
```

Rule of thumb for the VIF test is that if VIF > 10, then the model suffers from multicollinearity.

The above table shows the VIF values for all the explanatory variables. All the variables except mileage_sq, mileage_kmpl, engine_cc, engine_cc_sq have VIF value < 10 hence, those variables don't have multicollinearity among them. The high VIF value can be explained for those 4 variables because they are multiples of each other thus, resulting in high multicollinearity.

They can be dropped from the model; however, the respective squared terms are needed to determine the diminishing/rising effect of them on dependent variable.

7) **Omitted Variable Bias**: Omitted variable bias occurs when a variable which actually belongs to the true population model is not included in the model. Omitting relevant and correlated independent variables from the model leads to misspecification of the model.

No test can detect omitted variables from thin air; however, the Ramsey test can detect if any higher order of the included variables is required in the model or not.

```
> (resettest(reg1,c(0,1,2),"fitted",data1))

        RESET test

data:  reg1
RESET = 105.23, df1 = 3, df2 = 15390, p-value < 2.2e-16
```

H_0 : Higher power not required

H_a: H_0 is false (Higher powers are required).

Since the p-value of the test result is very close to 0, the null hypothesis of the Reset test is rejected. Thus, the model needs to have more variables in higher powers. However, mathematically, the model can have all its explanatory variables in higher powers. The model in this research contains higher powers of only mileage (i.e., mileage_kmpl_sq) and engine power (i.e., engine_cc_sq) because based on previous research and review of literature, those are the variables that have been found to have a significant effect on the resale value of the used car.

iv) Model Diagnostics

a) Rectification of Heteroskedasticity

Since the model suffers from heteroskedasticity, it needs to be rectified using some transformations. For this research, the Box-Cox transformation was used to see if the heteroskedasticity issue can be solved.

For this, the BoxCoxTrans() function in the caret package was used to find the value of the transformation factor (lambda). Based on the value of lambda, the dependent variable is transformed as given below:

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log y, & \text{if } \lambda = 0. \end{cases}$$

The model provided the value of lambda= -0.2, hence the first transformation was attempted. Furthermore, the BP test was rerun to see if the heteroskedasticity issue was resolved.

```
> (lambda=BoxCoxTrans(data1$selling_price...1))
Box-Cox Transformation

15411 data points used to estimate Lambda

Input data summary:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
   40000   385000   556000   774971   825000 39500000

Largest/Smallest: 988
Sample Skewness: 10

Estimated Lambda: -0.2
With fudge factor, Lambda = 0 will be used for transformations

> y_transformed=((data1$selling_price...1)^(-0.2)-1)/(-0.2)
> data1=cbind(data1,y_transformed)
> reg_bc=lm(y_transformed~log_of_avg_price+log_km_driven+vehicle
actor+luxury_factor+ mileage_kmpl+ mileage_sq+engine_cc+ engine_
> (bptest(reg_bc))

        studentized Breusch-Pagan test

data:  reg_bc
BP = 2188.3, df = 17, p-value < 2.2e-16
```

It can be seen from the test results; the model still exhibits heteroskedasticity. Thus, the use of heteroskedasticity robust standard errors is a must in this scenario.

It is also to be noted that the final regression model still uses a transformed dependent variable (i.e., logarithmically transformed), which was done to remove the skewness in the data that tends to exist when it comes to prices/incomes in datasets.

After using the robust standard errors, tests of significance were rerun, and the results are as shown:

```
t test of coefficients:

                       Estimate  Std. Error   t value  Pr(>|t|)
(Intercept)          3.8967e+00  8.3967e-02   46.4072 < 2.2e-16 ***
log_of_avg_price     3.1079e-01  1.2487e-02   24.8896 < 2.2e-16 ***
log_km_driven       -1.6002e-02  1.5554e-03  -10.2882 < 2.2e-16 ***
vehicle_age         -5.0199e-02  4.8701e-04 -103.0752 < 2.2e-16 ***
seller_factor1       3.2904e-02  1.8188e-03   18.0910 < 2.2e-16 ***
seller_factor2       1.6393e-02  4.8287e-03    3.3949 0.0006883 ***
transmission_factor1 9.8293e-02  2.9058e-03   33.8262 < 2.2e-16 ***
fuel_factor1         6.1351e-02  3.5573e-03   17.2464 < 2.2e-16 ***
fuel_factor2         2.3255e-02  6.4888e-03    3.5839 0.0003396 ***
fuel_factor3         1.0153e-02  1.4492e-02    0.7006 0.4835647
fuel_factor4         6.6721e-02  2.2604e-02    2.9517 0.0031649 **
luxury_factor1       4.4610e-02  3.8704e-03   11.5261 < 2.2e-16 ***
luxury_factor2       9.3371e-02  5.4076e-03   17.2668 < 2.2e-16 ***
mileage_kmpl         1.4213e-02  2.9514e-03    4.8156 1.482e-06 ***
mileage_sq          -3.6288e-04  6.6177e-05   -5.4834 4.238e-08 ***
engine_cc            1.4004e-04  1.5140e-05    9.2498 < 2.2e-16 ***
engine_cc_sq         5.3357e-09  3.1097e-09    1.7159 0.0862087 .
seats               -7.5830e-03  2.1360e-03   -3.5501 0.0003862 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Thus, the significance results are same (if not better) than the previous case.

b) Rectification of Non-Normality

The Jarque-Bera test results indicated that the residuals were not normally distributed. However, as shown before in the kernel density plot of the residuals, **they maintain a normal curve shape**, although with a high kurtosis value. Thus, there is no need for any correction to account for the normality of the residual terms.

c) Rectification of Omitted Variable Bias

Mathematically, the model can have all its explanatory variables in higher powers. The model in this research contains higher powers of only mileage (i.e., mileage_kmpl_sq) and engine power (i.e., engine_cc_sq) because based on previous research and review of

literature, those are the variables that have been found to have a significant effect on the resale value of the used car.

Section V: Conclusion

In this research, an econometric study was conducted on the resale value of used cars in India. The Cross-sectional data used in this research was taken from cardekho.com a well-known vehicle exploration platform that assists users in matching new and used cars as per their needs and purchasing them.

This study has its limitations because of the limited scope of the used car market in India. Due to the electric vehicle segment being relatively new, there was barely any data available in the set for them, hence it was not found to be significant. The data for LPG was also not significant as its very rarely used as a fuel source in India, as was reflected in the dataset.

The data itself had to be statistically treated for relevant conclusions to be inferred. Robust standard errors were used in hypothesis testing (due to heteroscedasticity in data), some higher-order mathematically significant variables were also omitted from the model as they weren't observed in literature and had little physical meaning, but the important ones were retained despite slight multicollinearity issue since they had significant meaning. The error term also had a high kurtosis associated with it, which had to be worked with.
The rest of the assumptions for use of OLS were also verified and appropriate action was taken to ensure the results carry statistical meaning and significance.

The factors used to determine resale value are - Engine type, fuel type, Car Price in the primary market, kilometres-driven, Vehicle age etc., the effect of these variables was found to be in line with what is observed in the literature and significant at minimum 1% levels.
The factors that had the most effect on the pricing were found to be - Vehicle age, Log average price, Transmission factor (Auto or manual), fuel type and Luxury type.
Factors with medium effect on price were - mileage, some fuel types, seller factors and kilometres driven.

Other factors like engine and number of seats had little effect on price compared to the rest.
All these effects are in comparison to the base case i.e., the car is a low-tier, manual, petrol car, sold by an individual.

From the regressive analysis, it can be concluded that people are willing to pay more for luxury diesel cars with automatic transmission in the second-hand car market. While they are also mildly sensitive about the mileage, kilometres driven as well as from whom they are buying the used car, the specifications of the engine and the number of seats is of little consequence while determining their willingness to pay.

Section VI: References

1. Richardson, M. 2009. Determinants of Used Car Resale Value. Colorado College.
2. Noor, Kanwal & Jan, Sadaqat. (2017). Vehicle Price Prediction System using Machine Learning Techniques. International Journal of Computer Applications. 167. 27-31. 10.5120/ijca2017914373.
3. Meng, Shiang-Min & Liu, Li-Jen & Kuritsyn, Mikhail & Pechnikov, Vladislav. (2019). Price Determinants on Used Car Auction in Taiwan. International Journal of Asian Social Science. 9. 48-58. 10.18488/journal.1.2019.91.48.58
4. S. Pudaruth, "Predicting the Price of Used Cars using Machine Learning Techniques," International Journal of Information & Computation Technology, vol. 4, no. 7, pp. 753–764, 2014.
5. Genesove, D., 1993. Adverse selection in the wholesale used car market. Journal of Political Economy, 101(4): 644-665. Available at: https://doi.org/10.1086/261891
6. Gilmore, E.A. and L.B. Lave, 2013. Comparing resale prices and total cost of ownership for gasoline, hybrid and diesel passenger cars and trucks. Transport Policy, 27: 200-208. Available at: https://doi.org/10.1016/j.tranpol.2012.12.007