DEPARTMENT OF STATISTICS
ST. XAVIER'S COLLEGE (AUTONOMOUS), KOLKATA

Name: **ARPAN SAMANTA**

Roll Number: **20-300-4-07-0425**

REGISTRATION NUMBER: **A01-1112-0849-20**

SUPERVISOR: **Dr. AYAN CHANDRA**

TITLE: PREDICTION OF CAR PRICE USING MULTIPLE REGRESSION MODEL

DECLARATION:

"I affirm that I have identified all my sources and that no part of my dissertation paper uses unacknowledged materials."

_____

Signature of Student

CONTENTS

INTRODUCTION

The automotive industry is one of the most important and dynamic sectors of the global economy, with millions of cars sold every year worldwide. One of the key factors that influence the success of this industry is the pricing of cars. Accurately predicting the price of a car is a complex task that requires considering various factors such as brand, model, year of manufacture, engine capacity, fuel efficiency, and so on. In recent years, with the rapid growth of data analytics and machine learning technologies, predicting the price of a car has become easier and more accurate.

The purpose of this dissertation project is to develop a model that can predict the price of a car based on its various numerical parameters. The model will be trained on a dataset of cars with different specifications and prices, and it will use machine learning algorithms to identify the most relevant features that impact car prices. The primary goal of this research is to improve the accuracy of car price predictions and provide insights into the factors that affect car pricing.

The dissertation project will begin by collecting a dataset of cars with various numerical parameters such as horsepower, car weight, mileage, and so on. We will clean and pre-process the data, and perform exploratory data analysis to gain insights into the relationships between different variables and car prices.

We will then use various machine learning algorithms such as linear regression to develop a predictive model that can accurately forecast car prices based on the given numerical parameters. Finally, we will evaluate the performance of the model using various metrics such as mean squared error, root mean squared error, MAPE, and R-squared.

Overall, this dissertation project aims to contribute to the growing field of predictive analytics in the automotive industry and provide insights into the factors that impact car pricing. The results of this research can be valuable for car manufacturers, dealerships, and consumers looking to buy or sell cars.

OBJECTIVES

My project deals with a data set on 205 cars and their different numerical characteristics. These numerical characteristics are,

- Wheelbase of car
- Length of car
- Width of car
- Height of car
- The weight of a car without occupants or baggage.
- Size of car
- Boreratio of car
- Stroke or volume inside the engine
- Compression ratio of car
- Horsepower
- Car peak rpm
- Mileage in city
- Mileage on highway
- Price of car

My primary objective is to determine if there is a regression between the price of a car and its other parameters. I will also evaluate the goodness of fit, identify any outliers, assess the individual impact of each parameter on the car price, and examine the residual plot to determine it's randomness. Ultimately, I aim to create a regression equation that can predict the price of a car based on its useful characteristics.

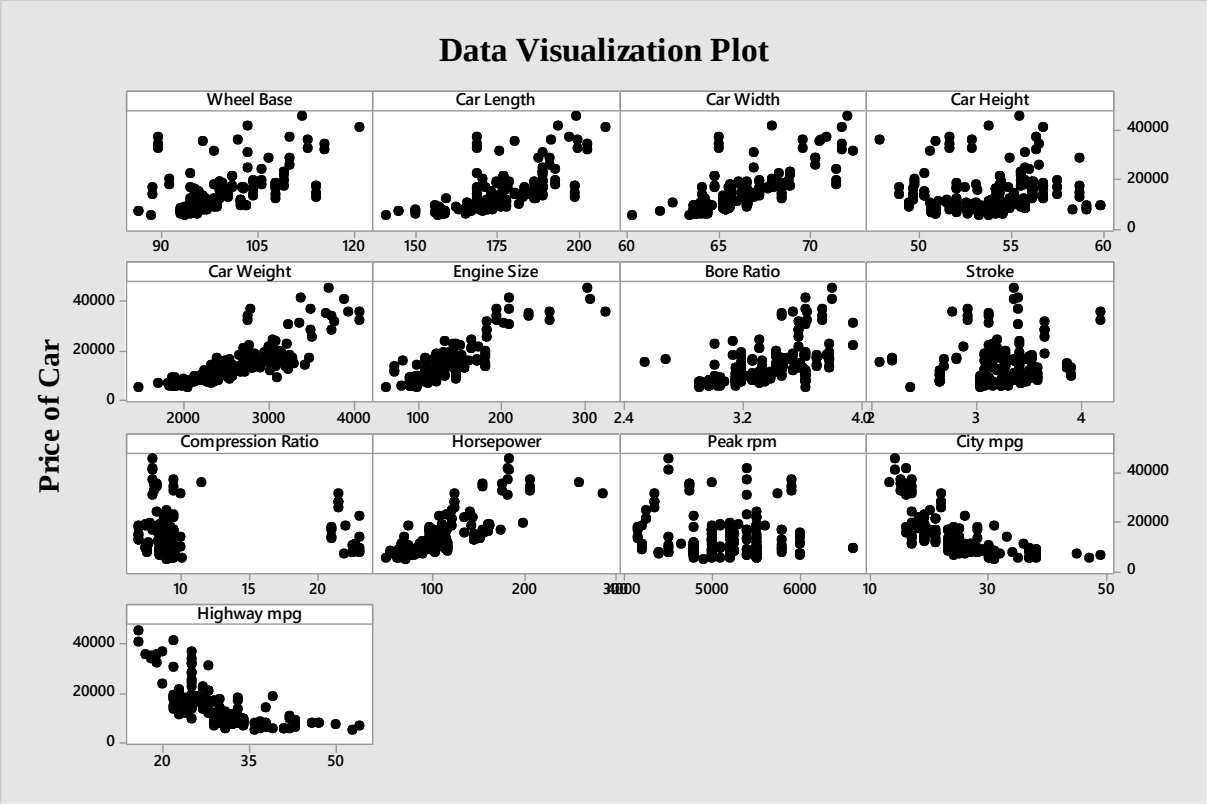
LIMITATIONS

Note that, it is not necessary that only these parameters will affect the price of the car. There exist some categorical parameters such as fuel type, number of doors, drive wheel location, etc. in the data set that has the ability to put some effect on the car price. However, since our response is a continuous random variable so in order to simplify our work we're only working with the numerical characteristics. It may affect our model but we will ignore that at the undergraduate level.

DATA VISUALIZATION

Data visualization is a technique where we represent the initial data in the form of some common graphics, charts, plots, animation, etc. and roughly try to analyze. In our case, we'll represent the data through a scattered plot. Since, we're working with multivariate data, a singular plot won't be enough to analyze. We'll plot the price of the car on the y-axis corresponding to each parameter on the x-axis. Thus, we'll get 13 plots. Since there are 13 numeric characteristics in the data set apart from the car price.

Graph | 1



Observing these plots we see that almost every parameter is more or less affecting the car price. The influence of wheelbase, car length, car width, car weight, bore ratio, and horsepower on the price individually is indicating a positive correlation. Whereas, city mpg and highway mpg are indicating a negative correlation. However, the plot of car height, Stroke, compression ratio, and peak rpm are not showing any significant pattern properly. So, we drop them and continue our analysis with the other parameters.

DATA DESCRIPTION

The data set I am working with is provided by Manish Kumar on Kaggle.com. The data set provides information on 205 cars and their various characteristics. Among these characteristics some are numeric and some are categorical in nature. However, as previously discussed I am working with numeric characteristics only. And using those characteristics as predictors that are showing some significant effect on the car price in the data visualization section.

In order to fulfill my purpose of simplifying the model mathematically, I assign different variables to each of those chosen car parameters.

- X_1 = Price of car
- X_2 = Weelbase of car
- X_3 = Length of car
- X_4 = Width of car
- X_5 = Height of car
- X_6 = The weight of a car without occupants or baggage.
- X_7 = Horsepower
- X_8 = Boreratio of car
- X_9 = Mileage in city
- X_{10} = Mileage on highway

From now on, instead of calling a particular characteristic by it's name, we'll simply denote it by it's corresponding variable.

Moreover, for multiple linear regression analysis, we'll call X_1 i.e. Price of car as ***the response*** and the rest of the X_j's, j = 2 (1) 10 as our ***predictors***. Clearly, there're 9 predictors corresponding to a single response, for each of which data on 205 cars are present.

CHOICE OF RESPONSE AND ITS SIGNIFICANCE

The response in a multiple regression analysis should be chosen very carefully. It should be such that there exists some possible effect of the predictors individually on it. From data visualization, it is clear that the 9 predictors are more or less affecting the car price individually. So, it is evident that they'll jointly affect the price. Also, in practical knowledge, we know that the price of a car is affected by it's numerous characteristics.

However, looking at our response from a different angle we see that car prices can significantly affect the global economy. Cars are one of the most noteworthy purchases in the common population. The price can put a direct effect on consumer spending, inflation, and economic growth.

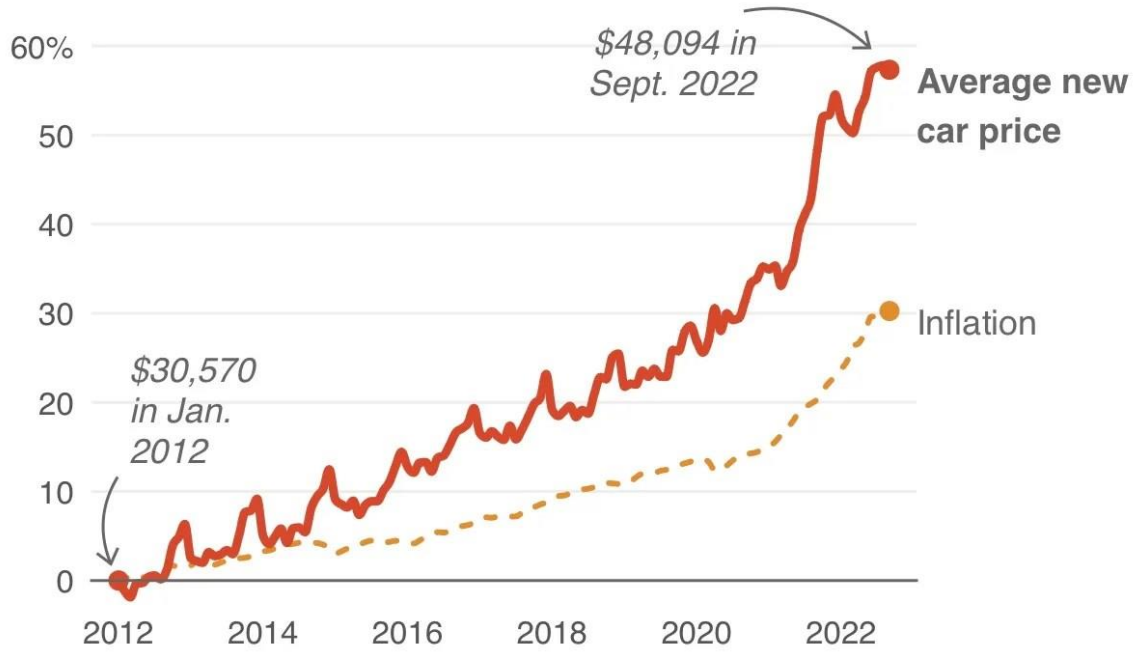
Increasing car prices can lead to a higher transportation cost that can result in a higher price of goods and services. This can cause a chain reaction as increased prices can lead to reduced consumer spending, lower sales for businesses, and ultimately, a slowdown in economic growth.

Moreover, the automotive industry is the main source of income in many countries economies such as Japan, Germany, the United States, South Korea, and China. Any harm in that sector can cause effect on revenue and profitability in those economies. A slowdown in the industry can cause in job losses and reduced tax revenue for governments. Moreover, car prices can also impact international trade. For instance, if a country's currency is weak, the cost of imported cars can go up, making them more expensive for consumers. This can lead to a shift in demand towards domestically-produced cars, which can help boost the local economy.

Taking these points in mind it is important to see how a key feature in a country's economy are affected by its own different features.

According to Kelley Blue Book and the Bureau Of Labor Statistics, the average new car price at the beginning of 2023 is $49,388 in the U.S. I have provided a chart given by them on the car price in the U.S. economy and how it is growing over the inflation curve.

Graph | 2



Source: Bureau of Labor Statistics and Kelley Blue Book

Credit: Daniel Wood

The graph shows the average price of a new car in the US from January 2012 to September 2022. The vertical axis shows the price of the car, while the horizontal axis shows the year and month.

As we can see from the graph, the average new car price has been steadily increasing over the years. In January 2012, the average new car price was around $30,570. By September 2022, the average price had increased to around $48,094.

We can also see that the rate of increase in new car prices has been much faster than inflation since 2014. The graph shows a steep upward trend in prices from around mid-2014, with a noticeable acceleration in 2021 and 2022. This acceleration was due to a decline in public transportation and supply shortages.

Overall, the graph shows a clear and steady increase in the average price of a new car in the US over the past decade, with significant spikes in recent years.

HISTORY OF MULTIPLE LINEAR REGRESSION

The history of multiple linear regression dates back to the early 19th century when French mathematician Adrien-Marie Legendre proposed a method to fit a line to a set of data points in 1805. This method was later used by the British scientist Sir Francis Galton for least squares multiple regression.

In the early 20th century, pioneers of modern statistics such as Karl Pearson and Ronald Fisher developed the statistical theory of multiple regression, which included the concepts of regression coefficients, standard errors, and hypothesis testing. With the introduction of computer technology in the mid-20th century, multiple linear regression analysis became possible for large data sets. This has led to the widespread use of regression analysis in various fields.

In recent years, the combination of multiple linear regression with other statistical and computational techniques has led to the development of advanced predictive models capable of handling large amounts of data and making accurate predictions. It is worth noting that in addition to simple linear and multiple linear regression, more advanced regression techniques have been developed in response to increasing data and computing power.

A statistical model which is linear in it's parameters is called a linear model. In multiple linear regression the relation between response and multiple predictors is represented by a suitable linear equation, say,

$$X_1 = \beta_1 + \beta_2 * X_2 + \beta_3 * X_3 + \ldots\ldots\ldots + \beta_p * X_p + \varepsilon$$

Where, X_1 is the response, X_j, j = 2 (1) p are the (p-1) predictors. Whereas ε is the error term and β_j , j = 2 (1) p are the (p-1) regression coefficients and β_1 is the intercept term.

The goal of the model is to determine the value of the coefficients that provide best fit to the data i.e. gives minimum error. This is typically done using a method called the method of least squares in which we minimize the sum of squares of errors with respect to the coefficients and thus solving various equations we get the values of βj's.

Multiple linear regression can be used for a variety of applications, such as predicting sales based on advertising spend, analyzing the relationship between multiple demographic variables and health outcomes, or predicting crop yields based on weather data.

However, it is important to note that MLR assumes that the relationship between the dependent and independent variables is linear and that there is no multicollinearity (high correlation) between the independent variables. If these assumptions are not met, the results of the analysis may be inaccurate.

MODEL AND ASSUMPTIONS

Let us consider the following linear regression model with n (= 205) sample observations and (p-1) (= 9) explanatory variables.

$$X_{1i} = b_1 + b_2X_{2i} + b_3X_{3i} \ldots\ldots\ldots\ldots\ldots\ldots\ldots + b_pX_{pi} + u_i \qquad i = 1\ (1)\ n$$

We re-write it in the matrix format as,

$$\underline{\boldsymbol{Y}} = \mathbf{X}\underline{\boldsymbol{\beta}} + \underline{\boldsymbol{u}}$$

With the assumptions,

- $\underline{u} \sim N_n(\underline{0}, \sigma^2 I_n)$
- Design matrix X is non – stochastic.
- Rank (X) = $(p - 1)$ $(< n)$

Where $\underline{\boldsymbol{Y}}$ is the response vector of order $n \times 1$, X is the design matrix of order $n \times p$, $\underline{\boldsymbol{\beta}}$ is the parameter vector of order $p \times 1$ and $\underline{\boldsymbol{u}}$ is the error vector of order $n \times 1$.

SOFTWARES USED

We will use Minitab – 17 and R – programming (version 4.1.2) for our necessary computation purpose.

REGRESSION ANALYSIS

Analyzing the data on 205 cars in the above-mentioned statistical softwares we get, the coefficient of determination as,

$$R^2 = 78.72\ \%$$

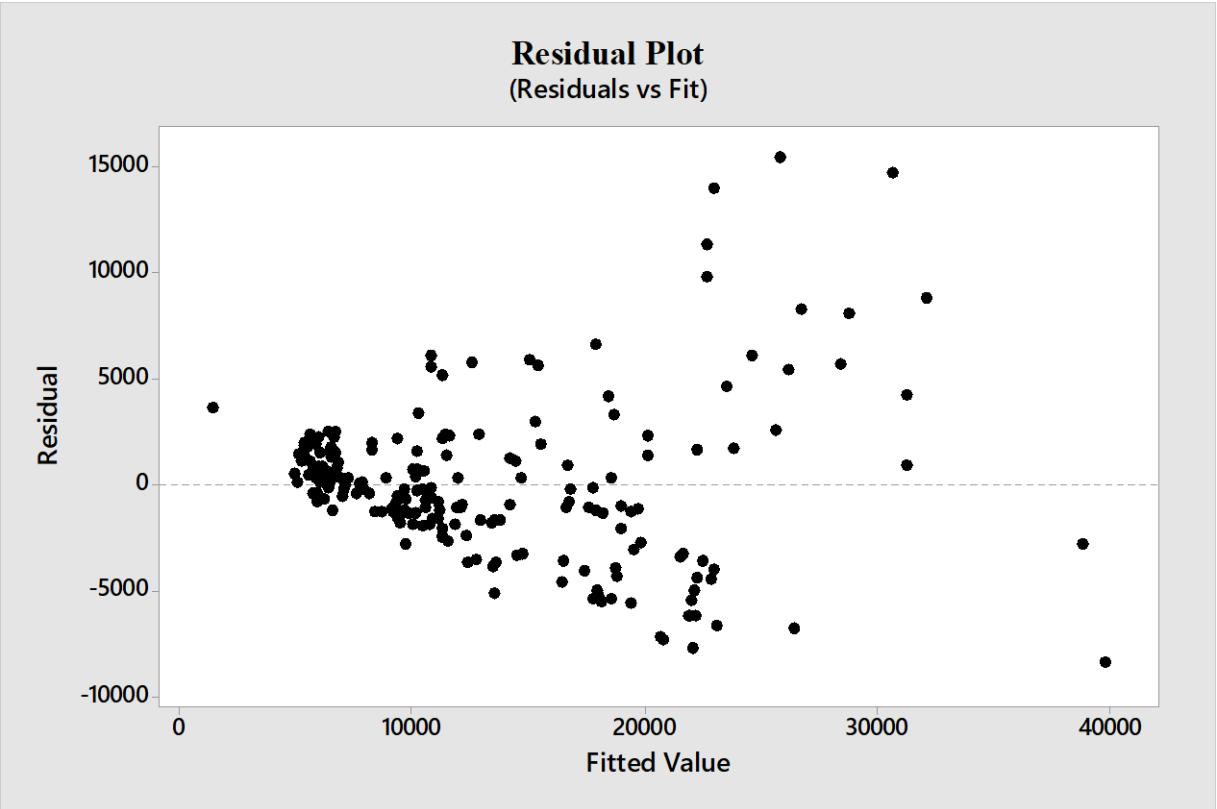
And, the multiple regression model after estimating the model parameters is given by,

$$X_{1.2345678910} = -54517 + 99 * X_2 - 71.6 * X_3 + 568 * X_4 + 58 * X_5 + 7.15 * X_6 - 839 * X_7 + 102.2 * X_8 + 90 * X_9 + 51 * X_{10}$$

Where, $X_{1.2345678910}$ is the predicted value of X_1 based on the predictors X_j , j = 2 (1) 10.

We would also like to check if there is any pattern in the residual plot. For that, we plot the residual values on y – axis taking the fitted response values of X_1 on the x – axis , thus we get the residual plot.

Graph | 3



From the residual plot, it is evident that the residuals are scattered around 0 and not much of a pattern can be observed. However, there may exist some outlier / influencial points that has to be omitted.

PRESENCE OF OUTLIERS

Dennis Cook (1977) introduced a distance measure for commonly used estimates to study the influence of a data point when performing least squares regression analysis. Data points with large residuals and/or high leverage may distort the outcome and accuracy of a regression. **Cook's distance** measures the effect of deleting a given observation. Points with a large Cook's distance are considered to merit closer examination of the analysis.

Cook's distance D_i of observation i (i = 1,2…..205) is defined as the sum of all the changes in the regression model when observation i is removed from it.

$$D_i = \sum_{i=0}^{n} \frac{\left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{ps^2}$$

Where,

$\hat{y}_{j(i)}$ is the fitted response value obtained when excluding i ,

p is the number of terms in the model + 1

s^2 is the means square error of the regression model.

CALCULATIONS

Using MINITAB, the Cook's Distance have been obtained for all the observations. The observations with comparatively high value (close to 1) are considered as influential points which distorts the regression equation. We remove such observations and re – fit the model.

After removing the unusual points we get our coefficient of determination as,

R^2 = 81.81 %

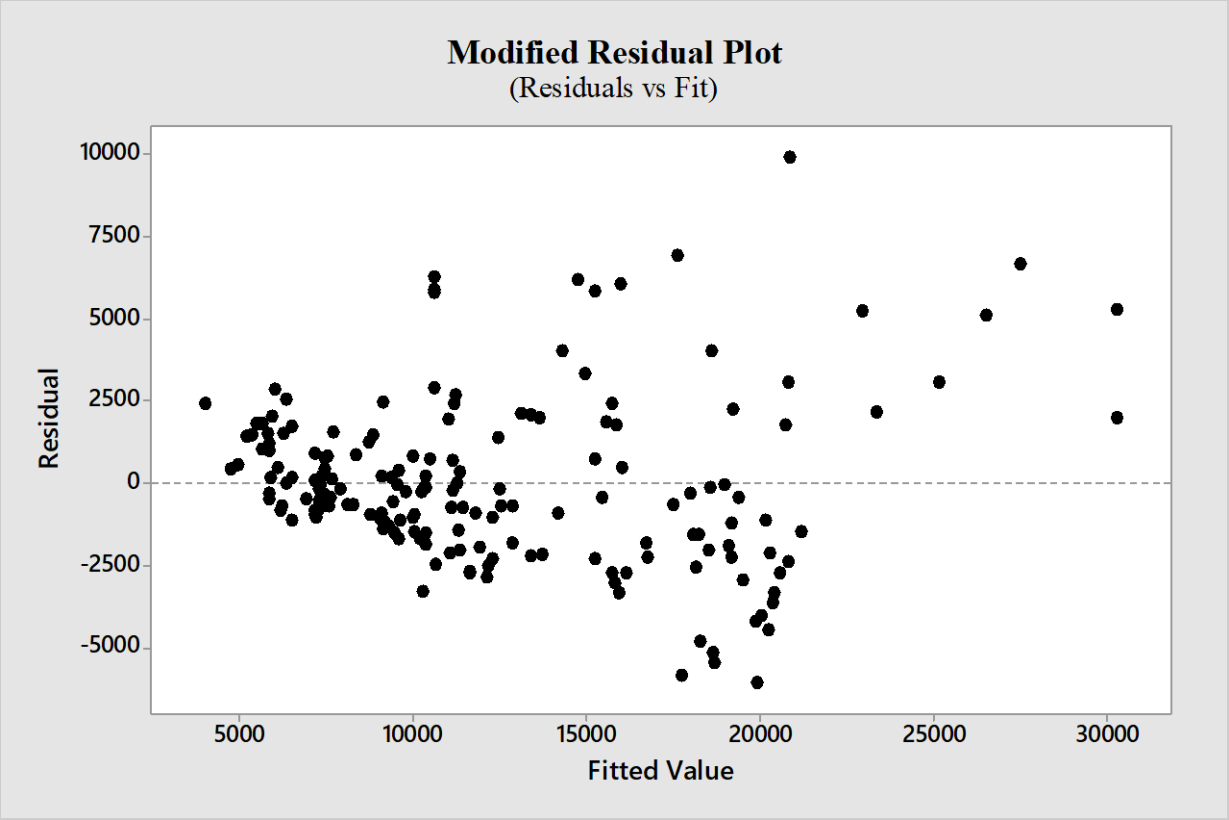
Corresponding to the model ,

$$X_{1.2345678910} =$$

$$-37924.685 + 177.460 * X_2 - 96.192 * X_3 + 462.614 * X_4 - 48.681 * X_5 + 8.932 * X_6 - 2284.362 * X_7 - 44.842 * X_8 - 127.425 * X_9 + 172.465 * X_{10}$$

Also, the new residual plot is coming as,

Graph | 4



We see that the scatteredness of the plot in the y – direction is much less in the new plot than in the previous. This is because we've omitted many influential points. However, the plot doesn't show any pattern that is observable.

SIGNIFICANT PREDICTORS

We'll now try to rebuild the model with the most useful predictors, in order to make the model more smooth. For this, we need the different p-values corresponding to different predictors that we've obtained while fitting the outlier-free model in R.

Predictor	p - value
X_2	0.05065
X_3	0.0442
X_4	0.03697
X_5	0.6767
X_6	$4.88 * 10^{-10}$
X_7	0.01782
X_8	0.00108
X_9	0.37822
X_{10}	0.21114

Lesser the p–value more significant the predictor. At 5% level of signicance, we will choose those predictors whose p–value is lesser than 0.05. That is our significant predictors are X_3 , X_4 ,X_6 , X_7 , X_8 . Among these, X_6 is the best significant predictor.

MULTICOLLINEARITY

Multicollinearity is a very common problem in multiple linear regression analysis. It might appear among the predictor variables that there exists a high correlation (more than 0.7 or 0.8) between one or more pairs of independent variables. It leads to unreliable or unstable estimates of regression coefficients affecting the accuracy of the model's prediction. In order to check for multicollinearity we've to first check the correlation in between the independent predictors. We give the correlation coefficient, r by the following table.

	X_3	**X_4**	**X_6**	**X_7**	**X_8**
X_3	1	0.8524	0.8899	0.6096	0.6014
X_4	0.8524	1	0.8696	0.5344	0.6239
X_6	0.8899	0.8696	1	0.6398	0.7676
X_7	0.6096	0.5344	0.6398	1	0.5139
X_8	0.6014	0.6239	0.7676	0.5139	1

We see there is a high correlation between X_3 and X_4 as well as between X_3 , X_6 and X_4 , X_6 . So in order to overcome this problem of multicollinearity we omit the predictors X_3 and X_4 .

And thus we get 3 predictors, X_6 , X_7 , X_8 corresponding to a single response X_1 in our ultimate model. With these information we form our ultimate modified model.

MODIFIED REGRESSION MODEL

The coefficient of determination for our ultimate model is,

$$\mathbf{R^2 = 80.43\ \%}$$

This value may look a negligible amount of smaller than the previous coefficient of determination, but the model is much more stable due to the absence of multicollinearity.

The model is now given by,

$$\boldsymbol{X_{1.678} = -7627.2234 + 10.2214 * X_6 - 2786.2447 * X_7 + 33.3958 * X_8}$$

Where, $X_{1.678}$ is the predictor of X_1 in the new model based on X_6 , X_7 , X_8 .

i.e. the model is expressed as,

$$\boldsymbol{Price\ of\ Car = -7627.2234 + 10.2214 * Car\ weight - 2786.2447 * Bore\ ratio + 33.3958 * Horse\ power}$$

Mean Sum of Square of Error (MSE)

The mean sum of square of error (MSE) is a statistical measure that is commonly used to evaluate the quality of a prediction or an estimator in a regression analysis.

It is calculated as the average of the squared differences between the predicted values and the actual values. The MSE is a way to quantify how far the predictions are from the actual values, on average.

The formula for calculating MSE is:

$$MSE = \frac{1}{n} * \sum_{i=1}^{n}(yi - \hat{y}i)^2$$

where n is the number of observations, yi is the actual value of the dependent variable, and ŷi is the predicted value of the dependent variable.

A smaller MSE indicates a better fit between the predicted values and the actual values, and therefore a better model or estimator.

MSE is our case is coming as,

MSE = 6984000

Considering, the price of a car can vary in a high scale of numbers, we can consider it as a moderately good fit. Also, the root MSE is coming as,

Root MSE = $\sqrt{MSE}$ = 2642.726

Mean Absolute Percentage Error (MAPE)

The MAPE (Mean Absolute Percentage Error) is a common metric used to evaluate the accuracy of a regression model. It measures the percentage difference between the predicted and actual values of the dependent variable. In multiple linear regression, which involves two or more independent variables, the MAPE can be calculated using the following formula:

$$MAPE = \frac{1}{n} * \Sigma\left|\frac{(y_{actual} - y_{predicted})}{y_{actual}}\right| * 100$$

where n is the number of observations, y_{actual} is the actual value of the dependent variable, and $y_{predicted}$ is the predicted value of the dependent variable.

Note that the MAPE is a relative measure of error, and it is expressed as a percentage. Therefore, it is useful for comparing the accuracy of different models or for evaluating the performance of a model against a certain benchmark. However, it has some limitations, such as being sensitive to outliers and not being defined when the actual value is zero.

The MAPE value in our model is coming as,

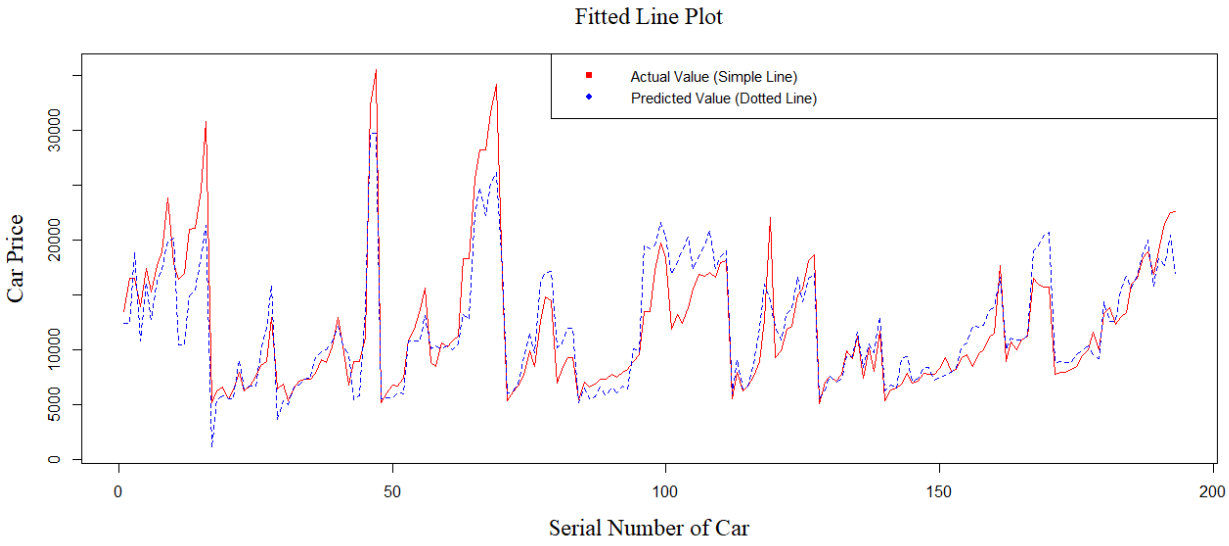
MAPE = 15.09 %

The MAPE value is showing a moderately lower value i.e percentage error between the predicted value and the actual value of the response variable is moderately low. This indicates a better fit of the model as it means that the predicted values are closer to the actual values.

PREDICTION

Putting the value of the corresponding predictors we can now easily predict the price of a given car. For prediction purposes we'ld like to see the fitted line plot where we've fitted the actual value of the car and its predicted value using this model in the same graph taking the car serial number on x – axis and car price on y – axis.

Graph | 5



From the plot, it is observable that though the fit isn't completely accurate, both plots are showing a similar kind of pattern. In the long run data analysis, this information might be helpful.

CONCLUSION

Apparently, it may look that the value of the coefficient of determination along with the other measure of accuracy should have expressed much better values, but we have to remember that we've only chosen the numerical characteristics of a car as our independent variables through limitations. There may exist some categorical parameters that affect the price of car. Working with those parameters alongside the numerical can explain the remaining correlation. However, that is a study of deeper machine learning algorithms and software (anaconda, jupyter note book etc.) that is beyond our scope of the study. We will not deal with that here.

Talking about the numerical characteristics with which we've worked only we can say that they are successfully affecting the car price and the ultimate picture of various predicted car prices can be shown via this model.

ACKNOWLEDGEMENT

I would like to thank Dr. Ayan Chandra, my supervisor and dissertation guide, for his guidance during the course of my dissertation. In addition, I would like to show all my appreciation to all my respected teachers of St. Xavier's College Statistics faculty, who have instilled in me a strong research mindset, a sense of curiosity, and the ability to continue in this field. Lastly, I would like to extend my gratitude to St. Xavier's College for the opportunity to present a dissertation project paper on a topic of my choice. I would also like to thank them for helping me to develop a research mindset in me.

REFERENCES

- “Fundamentals of Statistics – Vol 2” by A.M. Goon , M.K. Gupta and B. Dasgupta
- "Introduction to Linear Regression Analysis" by Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining.
- “Regression Analysis and Linear Models: Concepts, Applications, and Implementation” by Richard B. Darlington and Andrew F. Hayes
- Keggle.com
- Youtube.com
- financialsamurai.com
- Wikipedia.org