ABSTRACT

Accurate prediction of used car price is crucial for both buyers and sellers in the automotive market. This project leverages multiple linear regression analysis in R to develop a predictive model . Our results provide valuable insights for used car pricing strategies and demonstrate the effectiveness of multiple linear regression in predicting prices with a high degree of accuracy. This model can serve as a useful tool for industry professionals , buyers ,and sellers seeking to make informed decisions in the used car market.

CONTENTS

CHAPTER 1

1. INTRODUCTION

The used car market is a significant sector of the automotive industry, with millions of vehicles changing hands every year. Accurately predicting the prices of used cars is crucial for both buyers and sellers, as it directly impacts their financial decisions. With the rise of digital marketplaces and data analytics, it's now possible to develop data-driven models that can predict used car prices with high accuracy. This project aims to explore the development of a predictive model that can estimate the prices of used cars based on various factors .The data file "used cars1.csv" contains 11 variables. The dataset comprises 11 variables: location,year, kilometer_driven,fuel_type,transmission,owner type,mileage, engine, power, and price. Our objective is to leverage machine learning algorithms and historical data to create a tool that provides reliable price predictions for used cars. To achieve this goal, we employ different methods to build a regression model for predicting price. Multiple linear regression and stepwise analysis are the primary tools used to derive conclusions and insights from the dataset.
Stepwise analysis is a statistical technique that iteratively adds or removes predictors to a model, assessing their contribution to the regression equation. This approach helps to identify the most significant predictors, minimize multicollinearity, and improve model interpretability. By using stepwise analysis, we aim to identify the most important factors influencing used car prices and develop a predictive model.

CHAPTER 2

OBJECTIVE AND DATA DESCRIPTION

OBJECTIVE

1. To check that which variables are significant in predicting the price of a used car.
2. To check how well those variables describe the price of the car.
3. We are required to model the price of used car with available independent variables.

DATA DESCRIPTION

The original data is in csv form. It consists of 182 rows and 11 columns . The datafile “used cars1.csv” contains 11 variables ,they are Location,Year, Kilometers_driven, Fuel_type,Transmission,Owner_type, Mileage,Engine ,Power, Seat, New_price . Our response variable is the price of the used car. There are some categorical and numerical values are present in this data. Categorical variables are coded with numerical values ,such variables are called dummy variables.

Data Source:

`https://www.kaggle.com/datasets`

CHAPTER 3

Literature Review

The specific objective of this project is to predict the price of the used car with the available independent variables. We will learn to use linear regression model using dataset. We want to know the relationship among variables, especially between the used car price with other variable. We also want to predict the price of a used car based on the historical data. We are required to model the price of the cars with the available independent variables. It will be used by the management to understand how exactly the prices vary with the independent variables. They can accordingly manipulate the design of the cars, the business strategy etc. to meet certain levels. Further, the model will be a good way for management to understand the pricing dynamics of a new market.

CHAPTER 4

METHODOLOGY

Any data analysis needs use of statistical and computational methodologies. In this used car price data , to check how the variable price dependent of other variables. That is, here the response variable is new-price and all other variables are exploratory variables. And thus regression analysis is used to present study. A major activity in statistics is the building of statistical models that hopefully reflect the important aspects of the object of study with some degree of realism. In particular , the aim of regression analysis is to construct mathematical models which describe or explain relationships that may exist between variables. Also graphical methods such as boxplot, histogram and variance importance plot etc. Certain statistical software R, Microsoft excel are used for the purpose of data analysis. As well as internet facilities are also utilized for the purpose of data access.

4.1. LINEAR REGRESSION MODEL

If we denote the response variable by Y and the explanatory variables by $x_1, x_2, \ldots, x_k$, then a general model relating these variables is:

$$E[Y \mid X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k] = \Phi(x_1, x_2, \ldots, x_k) \quad (4.1)$$

Although, for brevity, we will usually drop the conditioning part write $E[Y]$. Here, we direct our attention to the important class of linear models, that is:

$$\Phi(x_1, x_2, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \quad (4.2)$$

which is linear in the parameter β_j.This restriction to linearity is not as restrictive as one might think. For example, many functions of several variables are approximately linear over sufficiently small regions, or they may be made linear by suitable transformation. Using logarithms for the gravitational model, we get the straight line

$$\text{LogF} = \log \alpha - \beta \log d \quad (4.3)$$

Also "categorical" models can be included under our umbrella by using dummy (indicator) x-variables.

4.2. MULTIPLE LINEAR REGRESSION

Let Y be a random variable that fluctuates about an unknown parameter η; that is,

$$Y = \eta + \epsilon \quad (4.4)$$

where ϵ is the fluctuation inherent in the experiment which gives rise to η , or it may represent the error in measuring η , so that η can be expressed in the form

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} \quad (4.5)$$

where the explanatory variables $x_1, x_2, \ldots, x_{p-1}$ are known constants and β_j ($j = 1, 2, \ldots, p-1$) are known parameters to be estimated.

If the x_j are varied and n values $Y_1, Y_2, \ldots, Y_n$ are observed, then

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i(p-1)} + \epsilon_i \quad (4.6)$$

where x_{ij} is the ith value of x_j. We call x_j an explanatory variable or regressor, and Y as the response variable.

4.2.1. Assumptions in multiple linear regression model

Some assumptions are needed in the model $Y = X\beta + \epsilon$ for drawing statistical inference. The following assumptions are made:

1. $E(\epsilon) = 0$
2. $E(\epsilon\epsilon') = \sigma^2 I_n$
3. Rank$(X) = k$
4. X is a non-stochastic matrix
5. $\epsilon \sim N(0, \sigma^2 I_n)$

4.3 MODEL ADEQUACY CHECKING

The fitting of the linear regression model, estimation of parameters testing of hypothesis properties of the estimator, is based on the following major assumptions :

1. The relationship between the study variable and explanatory variables is linear, at least approximately.

2. The error term has zero mean.

3. The error term has a constant variance .

4. The errors are uncorrelated.

5. The errors are normally distributed.

The validity of these assumptions is needed for the results to be meaningful .If these assumptions are violated ,the result can be incorrect and may have serious consequences. If these departure are small , the final result may not be changed significantly. But if the deviations are large ,the model obtained may become unstable in sense that a different sample could lead to an entirely different model with opposite conclusions. So such underlying assumptions have to be verified before attempting regression modelling. Such information is not available from the summary statistic such as t-statistics, F-statistics or coefficient of determination.

4.3.1 Checking of the linear relationship between study and explanatory variables

a)Case of one explanatory variable

If there is only one explanatory variable in the model, then it is easy to check the existence of the linear relationship between y and X by scatter diagram of the available data. If the scatter diagram shows a linear trend, it indicates the relationship between y and X is linear. If the pattern is not linear, then it suggest that the relationship between y and X is nonlinear. .

b)Case of more than one explanatory variables

To check the assumption of linearity between the study variable and the explanatory variables ,the scatter plot matrix of the data can be used. A scatterplot matrix is a two -dimensional array of two-dimension plots where each form contains a scatter diagram except for the diagonal . Thus, each scenario sheds some light on the relationship between a pair of variables. It gives more information than the correlation coefficient between each pair of variables ,because it provides a sense of linearity or nonlinearity of the relationship and some awareness of how the Individual data points are arranged over the region. It is a scatter diagram of (y versus X_1), (y versus X_2), ..., (y versus X_k).
Another option to present the scatterplot matrix is:

- Display the scatterplots in the upper triangular part of the plot matrix.
- Mention the corresponding correlation coefficient in the lower triangular part of the matrix.

4.3.2 Residual analysis

The residual is defined as the difference between the observed and fitted value of the study variable. The i-th residual is defined as the difference between the observed value and the fitted value for the i-th observation. Residuals can be viewed as the deviations between the data and the fit. Therefore, they measure the variability in the response variable that is not explained by the regression model. Residuals can also be thought of as the observed values of the model errors. Thus, any departure from the assumption of random errors should be reflected in the residuals. The analysis of residuals helps in identifying inadequacies in the model.

a)Standardized residuals:

The residuals are standardized based on the concept of subtracting the mean residual and dividing by its standard deviation. Since $\mathbb{E}(e_i) = 0$ and MS_{res} estimates the approximate average variance, the scaling of the residual is defined

as:

$$d_i = \frac{e_i}{\sqrt{\text{MS}_{\text{res}}}} \tag{4.7}$$

This is called the standardized residual, where:

$$\mathbb{E}(d_i) = 0 \quad \text{and} \quad \text{Var}(d_i) \approx 1$$

Thus, a large value of d_i potentially indicates an outlier.

b)Studentized residuals

The standardized residuals use the approximate variance of e_i . The studentized residuals use the exact variance of e_i.

4.4 INDICATOR VARIABLES

In general, explanatory variables in regression analysis are assumed to be quantitative in nature. For example, variables like temperature, distance, and age are recorded on a well-defined scale.

In many applications, variables cannot be defined on a well-defined scale and are qualitative in nature. For example, variables like sex (male or female), color (black, white), nationality, and employment status (employed, unemployed) are defined on a nominal scale. Such variables do not have a natural scale of measurement and usually indicate the presence or absence of a quality or attribute

(e.g., employed or unemployed, graduate or non-graduate, smokers or non-smokers, yes or no, acceptance or rejection). Therefore, they are defined on a nominal scale and can be quantified by artificially constructing variables that take values such as 1 and 0, where ”1” indicates the presence of the attribute and ”0” indicates the absence.

For instance, ”1” may indicate that a person is male, and ”0” indicates that the person is female. Similarly, ”1” may indicate that a person is employed, and ”0” indicates that the person is unemployed.

Such variables classify data into mutually exclusive categories and are called indicator or dummy variables.

Rule: When an explanatory variable leads to classification into ”m” mutually exclusive categories, use (m-1) indicator variables for its representation. Alternatively, use ”m” indicator variables but drop the intercept term.

4.5 MULTICOLLINEARITY

In many situations in practice, the explanatory variables may not remain independent due to various reasons. The situation where the explanatory variables are highly intercorrelated is referred to as multicollinearity.

Consequences of multicollinearity

In case of near or high multicollinearity, the following possible consequences are

encountered:

1. The OLSE remains an unbiased estimator of β, but its sampling variance becomes very large. Thus, OLSE becomes imprecise, and the property of Best Linear Unbiased Estimator (BLUE) does not hold anymore.

2. Due to large standard errors, the regression coefficients may not appear significant. Consequently, essential variables may be dropped.

3. Due to large standard errors, the confidence region may become large.

4. The OLSE may be sensitive to small changes in the values of explanatory variables. If some observations are added or dropped, OLSE may change considerably in magnitude as well as in sign. Ideally, OLSE should not change with the inclusion or deletion of variables. Thus, OLSE loses stability and robustness.

4.5.1 Multicollinearity diagnostics

An important question arises about how to diagnose the presence of multicollinearity in the data based on given sample information. Several diagnostic measures are available, each based on a particular approach. It is difficult to determine which diagnostic measure is best or ultimate. The detection of multicollinearity involves 3 aspects:

1. Determining its presence.

2. Determining its severity.

3. Determining its form or location.

There are many methods to detect the multicollinearity one of it is Variance inflation factor.

4.5.2 Variance Inflation Factor

In the presence of multicollinearity in the data, the matrix $X'X$ becomes ill-conditioned. The diagonal elements of $C(X'X)^{-1}$ help in detecting multicollinearity. If R_j^2 denotes the coefficient of determination obtained when X_j is regressed on the remaining $(k-1)$ variables excluding X_j, then the jth diagonal element of C is given by:

$$C_{jj} = \frac{1}{1 - R_j^2} \tag{4.8}$$

If X_j is nearly orthogonal to the remaining explanatory variables, then R_j^2 is small and consequently C_{jj} is close to 1. If X_j is nearly linearly dependent on a subset of the remaining explanatory variables, then R_j^2 is close to 1 and consequently C_{jj} is large.Based on this concept, the variance inflation factor (VIF) for the jth explanatory variable is defined as:

$$\text{VIF}_j = \frac{1}{1 - R_j^2} \tag{4.9}$$

In practice, a VIF greater than 5 indicates that the associated regression coefficients are poorly estimated due to multicollinearity. If regression coefficients are estimated by Ordinary Least Squares Estimation (OLSE) and its variance is $\sigma^2(X'X)^{-1}$, then VIF indicates that a part of this variance is influenced by VIF.

CHAPTER 5

VARIABLE SELECTION AND MODEL BUILDING

The complete regression analysis depend on the explanatory variables present in the model. it is understood in the regression analysis that only correct and important explanatory variables appear in the model. In practice, after ensuring the correct functional form of the model, the analyst usually has a pool of explanatory variables which possibly influence the process or experiment. Generally, all such candidate variables are not used in the regression modelling , but a subset of explanatory variables is chosen from this pool. How to determine such an appropriate subset of explanatory variables to be used in regression is called the problem of variable selection.

While choosing a subset of explanatory variables ,there are two possible options:

1. In order to make the model as realistic as possible, the analyst may include as many as possible explanatory variables.

2. In order to make the model as simple as possible, one way includes only a fewer number of explanatory variables.

Both approaches have their consequences. In fact, model building and subset selection have contradicting objectives. When a large number of variables are included in the model, these factors can influence the prediction of the study variable y. On the other hand, when a small number of variables are included, the predictive variance $\hat{Y}$ decreases. Additionally, collecting observations on a

larger number of variables involves more cost, time, labor, etc. A compromise between these consequences is necessary to select the "best regression equation."

The problem of variable selection is addressed assuming that the functional form of the explanatory variables (e.g., x^2, $\frac{1}{x}$, $\log x$, etc.) is known and no outliers or influential observations are present in the data. Various statistical tools like residual analysis,Identification of influential or high leverage observations, model adequacy, etc., are linked to variable selection. In fact, all these processes should be solved simultaneously. Usually, these steps are iteratively employed. In the first step, a strategy for variable selection is opted, and the model is fitted with selected variables. The fitted model is then checked for the functional form, outliers, influential observations, etc. Based on the outcome, the model is re-examined, and the selection of variables is reviewed again. Several iterations may be required before the final adequate model is decided.There can be two types of incorrect model specifications:

1. Omission/exclusion of relevant variables.

2. Inclusion of irrelevant variables.

5.1 EVALUATION OF SUBSET REGRESSION MODEL

A question arises after the selection of subsets of candidate variables for the model, how to judge which subset yields better regression model. Various criteria have been proposed in the literature to evaluate and compare the subset

regression models.

5.1.1Akaike's information criterion (AIC)

The AIC is an estimator of prediction error and thereby the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model relative to each of the other models. Thus, AIC provides a means for model selection.

The Akaike's Information Criterion (AIC) statistic is given by:

$$\text{AIC} = n \ln \left(\frac{\text{SSres}(P)}{n} \right) + 2P \tag{5.1}$$

where $\text{SSres}(P) = \mathbf{y}'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$ is based on the subset model:

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \boldsymbol{\epsilon} \tag{5.2}$$

The AIC is defined as:

$$\text{AIC} = -2(\text{maximized log likelihood}) + 2(\text{number of parameters})$$

Log likelihood is a measure of model fit. The higher the number, the better the fit. This is usually obtained from statistical output.

5.2 STEPWISE REGRESSION TECHNIQUES

This method is based on choosing the explanatory variables in the subset model in steps, which can either involve adding one variable at a time or deleting one variable at a time. Based on this, there are three procedures:

- **Forward selection**
- **Backward elimination**
- **Stepwise regression**

These procedures are computationally intensive and are typically executed using statistical software packages.

5.2.1 Forward selection procedure

This methodology assumes that there is no explanatory variable in the model except an intercept term. It adds variables one by one and tests the fitted model at each step using some suitable criterion. It has the following steps:

1. Consider only the intercept term and insert one variable at a time.
2. Calculate the simple correlations of x_i (for $i = 1, 2, \ldots$) with y.
3. Choose x_i which has the largest correlation with y.
4. Suppose x_1 is the variable which has the highest correlation with y.
5. Since the F-statistic is given by:

$$F_0 = \frac{(n-k)}{(k-1)} \cdot \frac{R^2}{1-R^2} \tag{5.2}$$

x_1 will produce the largest value of F_0 in testing the significance of regression.

6. Choose a prespecified value of the F value, say F_{1N} (F-to-enter). If $F > F_{1N}$, then accept x_1 and so x_1 enters into the model.

7. Adjust the effect of x_1 on y and re-compute the correlations of remaining variables with y to obtain partial correlations:

   - Fit the regression $y^* = \hat{\beta}_0 + \hat{\beta}_1 x_1$ and obtain the residuals.
   - Fit the regression of x_1 on other candidate explanatory variables as $x^* = \hat{\alpha}_0 + \hat{\beta}_1 x_1$ (for $j = 2, 3, \ldots, k$) and obtain the residuals.
   - Find the simple correlation between the two residuals. This gives the partial correlation.
   - Choose x_i with the second-largest correlation with y, i.e., the variable with the highest value of partial correlation with y.
   - Suppose this variable is x_2. Then the largest F-statistic is:

$$F = \frac{\text{SSres}(x_1/x_2)}{\text{MSres}(x_1, x_2)}$$

   - If $F > F_{1N}$, then x_2 enters into the model.
   - These steps are repeated. At each step, the partial correlations are computed, and the explanatory variable corresponding to the highest partial correlation with y is chosen to be added into the model. Equivalently, the partial F-statistics are calculated, and the largest F-statistic given the other explanatory variables in the model is chosen. The corresponding explanatory variable is added into the model if the partial F-statistic exceeds F_{1N}.

- Continue with such selection as long as either at a particular step, the partial F-statistic does not exceed F_{1N} or when the least explanatory variable is added to the model.

5.2.2 Backward elimination procedure

This methodology is contrary to the forward selection procedure. The backward elimination methodology begins with all k explanatory variables and keeps on deleting one variable at a time until a suitable model is obtained. It is based on the following steps:

- Consider all k explanatory variables and fit the model.
- Compute the partial F-statistic for each explanatory variable as if it were the last variable to enter the model.
- Choose a preselected value F_{out} (F-to-remove).
- Compare the smallest of the partial F-statistics with F_{out}. If it is less than F_{out}, then remove the corresponding explanatory variable from the model.
- The model will now have $(k-1)$ explanatory variables.
- Fit the model with these $(k-1)$ explanatory variables, compute the partial F-statistic for the new model, and compare it with F_{out}. If it is less than F_{out}, then remove the corresponding variable from the model.
- Repeat this procedure iteratively.
- Stop the procedure when the smallest partial F-statistic exceeds F_{out}.

CHAPTER 6

DATA ANALYSIS

We will learn to use linear regression model using used car price dataset. We want to know the relationship among variables , especially between the price with other variables. We also want to predict the price of a used car based on the historical data. Let's have a look at the data and see how features contributing in the data and price of car. In this data there is some categorical and numerical values ,and so we want change the categorical value to numerical value by substituting dummy variables.

Data Description

The dataset consists of 182 rows and 11 columns. The target variable is `new_price`, which signifies the price of the car. Before proceeding further, it is essential to ensure that the data is clean and usable.Firstly, we check for the presence of 'NA' values in the dataset. If any 'NA' values are found, appropriate handling steps will be taken.Next, we create a new data frame focusing on car prices and convert all character columns to numerical columns to facilitate analysis.

Variables	182 obs. of 13 variables
Location	"Mumbai" "Chennai" "Mumbai" "Mumbai"…
Year	2017 2014 2017 2012 2018 2016 2016 2018 2018 2017.
Kilometers_Driven	34000 70000 13500 66889 4000 70687 104350 7000 4708 14
Fuel_Type	"Diesel" "Diesel" "Petrol" "Diesel"…
Transmission	"Manual" "Manual" "Automatic" "Automatic" …
Owner_Type	"First" "Second" "First" "First" …
Mileage.kmpl.	13.7 23.6 14.8 22.5 15.4…
Engine.CC.	2393 1364 1598 1995 1598 1248 1364 998 1498 1199..
Power.bhp.	147.8 67.1 103.5 190 103.5…
Seats	7 5 5 5 5 5 5 5 5 5…
New_Price	"2527000" "9,27,000" "14,95,000" "70,43,000"…

Table 6.1

6.1 EXPLORATORY DATA ANALYSIS

Exploratory data analysis is a phase where we explore the data variables, see if there are any pattern that can indicate any kind of correlation between variables. And by drawing histogram of numerical variables and boxplot of categorical variables ,we get an initial idea about the variables and how this variables relate to explanatory variables.

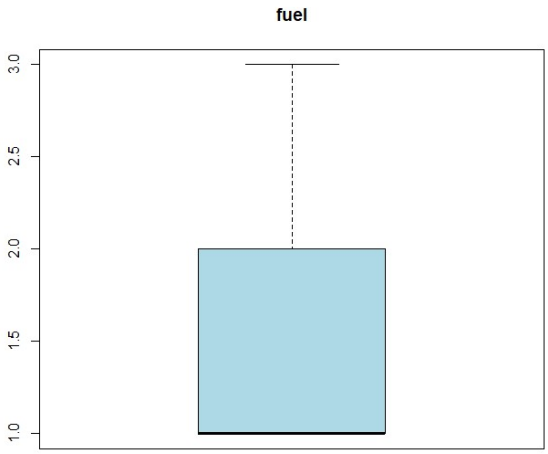

Figure 6.1:Boxplot of fuel

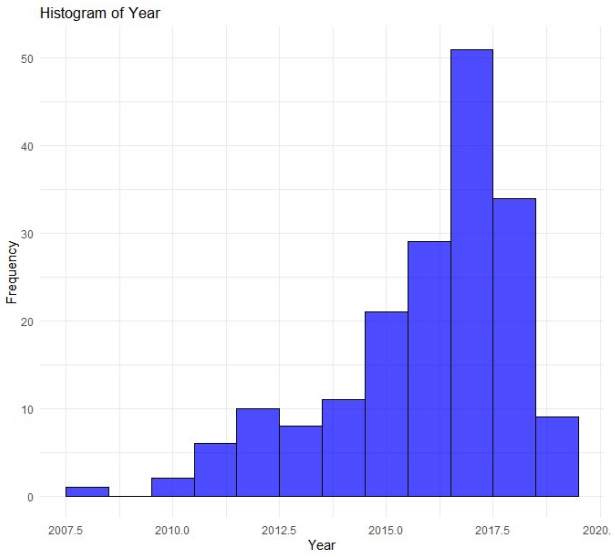

Figure 6.2:Histogram of Year

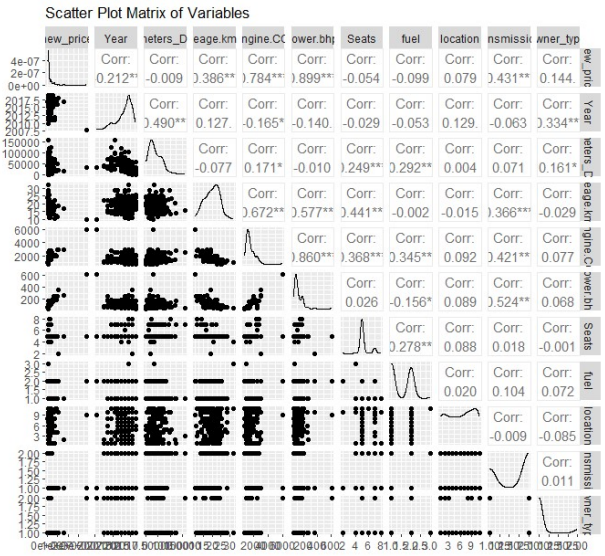

Figure 6.3

Exploratory data analysis give a visual effect of data.

From the correlation matrix (figure 6.4) we found that there are some correlation between explanatory variable. Thus we consider VIF of each variable for model building.Here using the corrplot command for displaying the pictorial representation of correlation matrix.It indicates the presence of correlations between variables.The circles with darker shades of blue shows the correlation between variables. The same can also be seen from a numerical point of view to get a clearer picture

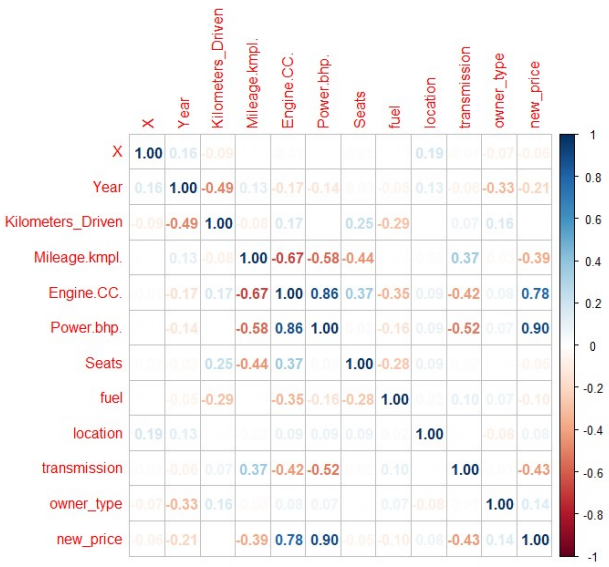

Figure 6.4 :Correlation Matrix

6.2 MODELING

6.2.1 Train and Test Split

Before we make the model , we need to split the data into train dataset and test dataset. We will use the train dataset to train the linear regression model. The test dataset will be used as a comparison and see if the model get overfit and can not predict new data that hasn't been seen during training phase. We will 70 percentage of the data as the training data and the rest of it as the testing data.

6.2.2 Variable Selection

Using stepwise technique and AIC method . After the stepwise iteration and removing the higher VIF variables the final model is figure 6.2

Variables	182 obs. of 13 variables
Location	"Mumbai" "Chennai" "Mumbai" "Mumbai"…
Year	2017 2014 2017 2012 2018 2016 2016 2018 2018 2017…
Kilometers_Driven	34000 70000 13500 66889 4000 70687 104350 7000 4708 14…
Fuel_Type	"Diesel" "Diesel" "Petrol" "Diesel"…
Transmission	"Manual" "Manual" "Automatic" "Automatic" …
Owner_Type	"First" "Second" "First" "First" …
Mileage.kmpl.	13.7 23.6 14.8 22.5 15.4…
Engine.CC.	2393 1364 1598 1995 1598 1248 1364 998 1498 1199…
Power.bhp.	147.8 67.1 103.5 190 103.5…
Seats	7 5 5 5 5 5 5 5 5 5…
New_Price	"2527000" "9,27,000" "14,95,000" "70,43,000"…

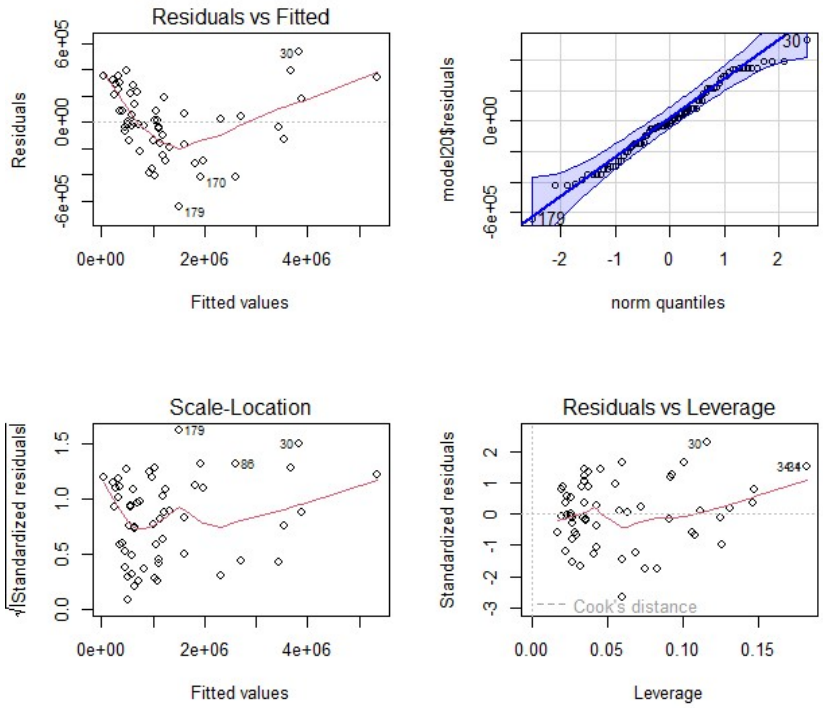

Figure 6.5

Table 6.2

6.2.3 Correlation

Correlation between the predicted price and the actual price is 0.97. And the R-squared value is 0.95 .

6.3 Model adequacy checking

We check the model adequacy by plotting the model

6.3.1 Test for normality

Shapiro-Wilk normality test

data: residuals(model20)

$W = 0.98196, \quad p\text{-value} = 0.2883$

With p-value > 0.05, we conclude that our residuals are normally distributed .

6.3.2 Test for Autocorrelation

lag	Autocorrelation	D-W Statistic	p-value
1	-0.03057568	2.047301	0.758

Alternative hypothesis: rho ! = 0

With p-value > 0.05, we can conclude that there is no autocorrelation present.

6.3.3 Test for Heteroscedasticity

Studentized Breusch-Pagan Test

Data: model20

BP = 6.254, df = 4, p-value = 0.181

With p-value > 0.05, we can conclude that variance of the residuals are constant .

6.3.4 Multicollinearity

Power.bhp.	transmission	Year	Mileage.kmpl.
2.353798	1.644978	1.394394	2.542247

CHAPTER 7

CONCLUSION

The study helped in understanding the relative importance of various factors responsible for pricing the car. Variables that are useful to describe the variances in car prices are `Power.bhp, transmission ,Year ,Mileage.kmpl` .And our final model has satisfied the classical assumptions .The R-squared of the model is high, with 95.3 percentage of the variables can explain the variances in the car price.

APPENDIX

Exploring and Preparing Dataset

```
data = read.csv("D:\\project 3\\used cars1.csv")
data
```

CHECKING FOR DUPLICATION AND NA VALUES

```
duplicate_rows = data[duplicated(data) | duplicated(data, fro

mLast = TRUE), ]
duplicate_rows


missing_values <- sapply(data, function(x) sum(is.na(x)))
missing_values
```

DATA PREPARATION AND DATA CLEANING

```
library(dplyr)
head(data)
dim(data)
str(data)
names(data)
```

Exploratory analysis

```
library(corrplot)
datamatrix = cor(data)
```

```
datamatrix
corrplot(datamatrix, method = "number")
```

DUMMY VARIABLE ANALYSIS

```
fuel = as.numeric(factor(data$Fuel_Type, levels = c("Diesel",

"Petrol", "CNG")))
data1 = cbind(data, fuel)
data1 = data1[, -which(names(data1) == "Fuel_Type")]


location = as.numeric(factor(data$Location, levels = c("Mumb

"Chennai", "Jaipur", "Hyderabad", "Pune", "Kolkata", "Ahmeda

bad", "Kochi", "Bangalore", "Coimbatore", "Delhi")))
data1 = cbind(data1, location)
data1 = data1[, -which(names(data1) == "Location")]


transmission = as.numeric(factor(data$Transmission, levels =

c("Automatic", "Manual")))
data1 = cbind(data1, transmission)
data1

data1 = data1[, -which(names(data1) == "Transmission")]
owner_type = as.numeric(factor(data$Owner_Type, levels = c("
```

```
First", "Second")))
data1 = cbind(data1, owner_type)
data1 = data1[, -which(names(data1) == "Owner_Type")]
column_values = as.numeric(data1$New_Price)
is.na(column_values)
rows_with_commas = grepl(",", column_values)
column_values[rows_with_commas] = as.numeric(gsub(",", "",

column_values[rows_with_commas]))
column_values = as.numeric(column_values)
data1$new_price = column_values
data1$new_price
```

Modeling

```
set.seed(121)
trainindices = sample(1:nrow(data1), 0.7 * nrow(data1), repla

ce = TRUE)
data1_train = data1[trainindices, ]
data1_test = data1[-trainindices, ]
library(MASS)
model = lm(new_price ~ ., data = data1_train)
model
stepwise_model = stepAIC(lm(new_price ~ 1, data = data1_train
stepwise_model
```

```
summary(stepwise_model)
Outliers Removal
cooks_dist = cooks.distance(model)
plot(cooks.distance(model), pch = 20, main = "OUTLIERS BY CO
OKS DISTANCE METHOD")
abline(h = 4 / nrow(data1_train), col = 'red', lty = 2)

cook_threshold = 0.5 / nrow(data1_test)
cook_threshold
outliers = which(cooks_dist > cook_threshold)
outliers
```

FINAL MODEL

```
model20 = lm(new_price ~ Power.bhp. + transmission + Year +
Mileage.kmpl., data = data1_train2)
summary(model20)
```

Model Adequacy

```
library(car)
library(lmtest)


shapiro.test(residuals(model20))
durbinWatsonTest(model20)
bptest(model20)
vif(model20)
```

```
residuals2 = residuals(model20)
mean_residuals = mean(residuals2)
mean_residuals

par(mfrow = c(2, 2))
plot(model20, which = 1)
qqPlot(model20$residuals)
plot(model20, which = 3)
plot(model20, which = 5)
```

REFERENCE

1. An Introduction to Multivariate Statistical Analysis , Third Edition by T.W. Anderson.

2. Applied Multivariate Statistical Analysis , Sixth Edition, Richard A.Johnson, Dean W.Wichern.

3. Introduction to Linear Regression Analysis , Fifth Edition by Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining.

4. Methods of Multivariate Analysis, Second Edition ,Alvin C. Rencher