TU Dortmund

Introductory Case Studies

Project 3: Regression Analysis

Lecturers:

Prof. Dr. Sonja Kuhnt

Dr. Birte Hellwig

Dr. Paul Wiemann

M. Sc. Hendrik Dohme

Author: Bhavesh Jain

Group number: 4

Group members: Akanksha Tanwar, Bhavesh Jain, Sohith Dhavaleswarapu, Akshatha Krishnananda Shanbhag

January 28, 2022

Contents

1 Introduction

In recent years, cars have become a product of necessity rather than a luxury. Buying used cars has become more affordable than purchasing a new one as there might be factors that might influence the price of the vehicle. These factors play a vital role in the purchase decision for the car. Generally, Cars which are of older technology may come with disadvantages, but they come with a lesser price tag, whereas cars which aren't that older would have a high price.

The goal of the project includes to determine the best multiple linear regression model estimating the price of the cars. The dataset contains information of the cars sold on a used car platform named Exchange and Mart Exchange Enterprises, Newsquest Media Group. (2022) based in the United Kingdom for the year 2020. To fit the linear model, the variables *model, price, year, mileage, fuel type, engineSize, tax, transmission* from the dataset are used.

For the first task, we transform the response variable price to the log of price, then we transform the mpg from using the scale as miles per gallon to liters per 100 kilometers and calculate the car's age with the help of the year it was first registered. For the second task, for choosing the best response variable as log price or the price, we make use of the model diagnostic tools on each linear model and found out that log price is best suited as the response variable. Then we used AIC and BIC as the best subset selection criterion to derive the subset of variables that impact the response variable. Lastly, on the model with the minimum value of BIC, we interpret the coefficients, statistical significance and discuss the confidence intervals and goodness of fit concerning the subset of the dependent variables.

In this report, section 2 has the details of the data set provided and the objectives of the project. In section 3, the statistical methods such as multiple linear regression, Least squares estimator, Model diagnostic tools, Test of significance, AIC, and BIC, which are used to accomplish the tasks, are stated and explained in detail. This section also includes details about how the linear model assumptions are examined and what are the criteria to choose the best variables having a better impact on the response. Section 4 consists of the application of the methods from section 3 to the dataset leading to interpretation of the best-fitted model. Finally, possible conclusions and views about further investigations are stated in the summary.

2 Problem statement

2.1 Data set and data quality

The official data set is provided by the instructors of the course Introductory Case Studies at TU Dortmund University in the winter semester 2021/2022 and is extracted from the website Kaggle.com Kaggle Inc. (2021). The data set contains 9 variables with 438 observations and provides information on cars sold on a used car platform named Exchange and Mart Exchange Enterprises, Newsquest Media Group. (2022), which is in the United Kingdom. The data was collected for the year 2020.

The variables *Model, Transmission, FuelType* are categorical and *price, mileage, mpg, tax, engine size, year* are numerical. *Price* tells about the price of each car and is given terms of 1000 GBP, *Year* refers to the year in which the car was first registered and *mileage* takes the value for the distance that can be traveled by car using one gallon (uk) of fuel, and this value is given in miles. The *fueltype* specifies which type of fuel is consumed by the car, and this variable has three categories: Petrol, Diesel, and Hybrid. The *engineSize* states the size of the car's engine and is measured in liters. *tax* is the amount of the annual tax (Vehicle Excise Duty), which has to be paid for the car. Lastly, the *transmission* has three categories: Semi-Auto, Manual, Automatic and tells about the type of gearbox used in the car. All this information is provided for the three car models manufactured by Volkswagen(VW): Up, Passat, and T-Roc Volkswagen AG (2022). There are no missing values in the dataset provided. Three new variables are created by transforming the existing variables. The first one is the *logprice*, which is the log-transformed values of *price*, the second one is the *lp100*, which is transformed using *mpg* and denotes the litres of fuel consumed per 100 kilometers. The third variable is the *cars_age*, which is calculate using *year* and tells how old the car is.

2.2 Project objectives

The primary objective of this project is to perform a regression analysis of the dataset provided. To perform the regression analysis, a multiple linear regression model is constructed predicting the price of the cars with the help of selected best features. Firstly, the three variables *logprice,lp100* and *cars_age* are created in order to perform further analysis. Then to choose the best among the two response variables(*price* and *logprice*), model diagnostic tools are used and the response variable is chosen whose assumptions

are more in line with the assumptions of the linear model. Then to choose the best subset of features that provide an impact to the response variable, AIC and BIC selection criteria are used. Lastly, on the model with the minimum value of BIC, interpretation of the coefficients, statistical significance, confidence intervals, and goodness of fit are discussed for the subset of the best features.

3 Statistical methods

3.1 Linear Regression

Linear regression is a statistical technique to model the relationship between a response variable y and an independent variable x. For a set of n observations in a data set, a response variable is the one that is to be predicted and the independent variables are the one that explains the prediction. The linear regression is given by the formula:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, ..., n.$$

Here, n denotes the number of observations and i denotes each single observation. β_0 represents the intercept and β_1 is the regression coefficient. The ε is the random noise (Rencher and Schaalje, 2008, p.127).

3.1.1 Multiple linear regression

In multiple linear regression, the prediction of the dependent variable is done on the basis of k different independent variables. The formula for multiple linear regression is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

It also assumes a linear relationship between the dependent variable and the independent variables. To determine the unknown regression coefficients $\beta_1 \ldots \beta_k$, it is summed up into a $k+1$ dimensional vector β. The independent variables are also summed up into a row vector, where the first column has the value of 1 unless specified and the error term ε is a scalar quantity. So, for n individual observations, the multiple linear regression

model can be written as follows:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + \varepsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + \varepsilon_n \end{aligned}$$

And, for the above n equations, the matrix form can be written as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

The above matrix notation can be summarized as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Here, $\boldsymbol{X}$ is called the design matrix. It is a $n \times (k+1)$ matrix and based on the assumptions, each column of the design matrix is linearly independent which means rank $(\boldsymbol{X}) = k + 1 = p$, where $n \geq p$. The β_0 is the constant term and the effect of x_1 on the expectation of y is shown by β_1, when all other independent variables are held constant. The ε is known as the residual and the value represents the difference of the prediction and the actual value for the dependent variable (Rencher and Schaalje, 2008, p.137-139).

3.1.2 Ordinary Least Squares (OLS)

Ordinary least squares is an estimation method that works on the principle of minimizing the squared distances between the value for the actual variable y and the predicted variable $\hat{y}$ for all n observations. It is denoted as follows:

$$\begin{aligned} OLS(\beta) = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_k x_{ik} \right)^2 . \end{aligned}$$

By differentiating the above equation with respect to each $\beta_j, j = 1 \dots k$ and equating it to zero, would yield to the values for every β_j that minimizes the equation. And, the

resulting OLS estimator for β is:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$$

With the help of the above equation, the estimated conditional mean of y is:

$$\widehat{E(y)} = \hat{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

Furthermore, substituting $\hat{\beta}$ in the above equation yields to $\hat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'y} = \boldsymbol{Hy}$. Here, $\boldsymbol{H}$ is known as the $n \times n$ hat matrix or the prediction matrix. This matrix is both symmetric and idempotent. The residuals can also be denoted in the matrix form, with the help of the hat matrix i.e. $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{Hy} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$.

With the OLS estimator, the error term or the residual $\hat{\varepsilon}$ is given by $\hat{\varepsilon} = y_i - x_i'\hat{\beta}$ and the unbiased estimator $\hat{\sigma}^2$ for σ^2 is given by:

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-p}$$

(Fahrmeir et al., 2013, p.105-109).

Considering $\hat{\sigma}^2$ as the unbiased estimator of variance with n observations, p covariates and h_{ii} as the diagonal elements of the hat matrix $\boldsymbol{H}$, the standardized residuals(r_i) are determined by (Fahrmeir et al., 2013, p.124):

$$r_i = \frac{\hat{\epsilon}}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

3.1.3 Assumptions

The multiple regression model is based on the following assumptions:

1. Linear relationship: There must exist a linear relationship between the dependent variable and the independent variables. This can be proven in 2 ways, the first method is to simply assume if all coefficients have no influence on the independent variable, this indicates the existence of a non linear relationship. Secondly, it can be proven with the help of a residual vs fitted plot (Fahrmeir et al., 2013, p.78).
2. No Multicollinearity: There must not exist any set of independent variables where exists linear relationship among them. Then those variables are not considered for the prediction of the dependent variable (Fahrmeir et al., 2013, p.158).

3. Independent and Identically distributed(i.i.d): It is assumed that the observations are independent and identically distributed i.e. $\{x_i, y_i\}_{i=1}^{n}$. For this assumption, random sampling of the observations makes sure of the distribution (Fahrmeir et al., 2013, p.79).

4. Homoscedasticity: It is assumed that the variation caused in the errors across the model must be similar i.e. $var(\varepsilon_i = \sigma^2)$ for $i = 1, 2, \ldots, n$. This assumption can be examined with the help of residual vs fitted plot (Fahrmeir et al., 2013, p.78).

5. Normality: It is assumed that the standardized residual errors are normally distributed i.e. $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. This assumption can be examined with the help of Normal Q-Q plot (Fahrmeir et al., 2013, p.80).

Model diagnostic tools

These tools aims to examine the above stated assumptions about the model. The tools used for examining the assumptions of a multiple linear regression model are:

1. Residual vs Fitted plot: It is a scatter plot, which is used to examine the assumption of Homoscedastic error variance in the model. In this plot, Residuals are plotted on the y-axis and fitted values by the model are plotted on the x-axis. The less the change in distance of the observations from the horizontal line which is drawn at residual value zero, provides more evidence for homoscedastic error variance. And, if the scattering of the observations lie across the horizontal line, it shows the sign of linearity (Fahrmeir et al., 2013, p.155-156).

2. Normal Q-Q plot: A QQ-plot is used to determine how well the distribution of the data given is in the form of standard normal distribution. We plot the values in our data set against the normal distribution, i.e., keeping theoretical quantiles on the $xaxis$ and the sample quantiles on the $yaxis$. When a reference line is imposed on the plot and if the plotted points lie on the line or close to the line, then the data is said to be normally distributed (Fahrmeir et al., 2013, p.156).

3.1.4 Dummy encoding for categorical variables

The process of conversion of categorical variable into dichotomous variables is known as dummy encoding. In this way, we indicate the presence or the absence of the particular

attribute to the model. If there are l number of categories in a particular categorical variable, then dummy encoding will convert those l variables into $l-1$ variables.

$$x_{i1} = \begin{cases} 1 & x_i = 1, \\ 0, & \text{otherwise} \end{cases} \quad \ldots \; x_{i,l-1} = \begin{cases} 1 & x_i = l-1, \\ 0, & \text{otherwise} \end{cases}, \text{for } i \; = \; 1, ..., n \text{ observations}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots \; + \beta_{i,l-1} x_{i,l-1} + \ldots \; + \epsilon_i$$

In the above representation, the dummy variable for category l is known as the reference category. The interpretation on the estimation is made by directly comparing it with the reference category (Fahrmeir et al., 2013, p.97).

3.1.5 R-Squared and adjusted R-Squared

To test if the model has the best fit to the data or not, a squared Pearson correlation coefficient is calculated between the observed values y and the predicted values $\hat{y}$. This value is also known as the coefficient of determination and is denoted by R^2. Its value ranges from 0 to 1. The more the value is closer to 1, it resembles a better fit to data and if the value is closer to 0, it under fits the data. The formula to calculate R^2 is given by:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n} \epsilon_i^2}{\sum_{i=1}^{n} (y_i - \hat{y})^2} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Here, SSE is calculated by $SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y}^2)$ and is is known as the sum of squared errors. SST is calculated by $SST = \sum_{i=1}^{n} (y_i - \bar{y}^2)$ and is known as the sum of squared total. For each i^{th} observation, $\hat{y}_i$ represents its fitted value and $\bar{y}$ is the mean of the values of y. The adjusted R squared value only considers the independent variables that help in prediction of the dependent variable and penalizes for the addition of independent variables that is of no help for the prediction. The value increases by the addition of an independent variable that is significant and affects the prediction. The adjusted R^2 value can be calculated by:

$$R^2_{adjusted} = \frac{(n-1)R^2 - k}{n - k - 1}$$

Here, k is the number of variables, n represents the number of observations and R^2 is derived from the previous formula (Rencher and Schaalje, 2008, p.162).

3.2 Test of significance

In order to perform the statistical tests and derive the confidence intervals, we assume the errors or the residuals be normally distributed i.e. $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

3.2.1 T-test and f-test

T-test: The null hypothesis for this test is assumed that the value for the regression coefficient is equal to zero and for the alternate hypothesis, it is assumed that the value is not equal to zero.

$$H_0 : \beta_j = 0; \quad H_1 : \beta_j \neq 0 \quad j = 1 \dots k$$

The t value obtained is the ratio of the estimated regression coefficient and its estimated standard deviation(se_j). The estimated standard deviation or the standard error is the average value of the variation of the parameter from the actual value and is given by $\widehat{\text{se}}_j = \sqrt{\widehat{\text{Var}\left(\hat{\beta}_j\right)}}, j = 0, 1, \dots, k$ (Fahrmeir et al., 2013, p.117). Therefore the value for t_j is given as:

$$t_j = \frac{\beta_j}{se_j}, j = 0, 1, \dots k.$$

A high t value provides significant evidence to reject the null hypothesis. The value t_j follows the t-distribution with having $n - p$ degrees of freedom, where $p = k + 1$. The critical value (|t|) of the rejection region of the null hypothesis is obtained and is the $(1-\alpha/2)$-quantile of the t-distribution with $n-p$ degrees of freedom. The p-value value of the two tailed t-test represents the probability of observing a greater value of t under the null hypothesis. The p-value is given by Pr(|T|>|t|), where T represents t_j value for all $j = 0, \dots, k$. The significance level (α=0.05) is already fixed before the test is performed and if the p-value is less than the α, the null hypothesis is not rejected. It provides enough evidence to conclude that, the particular β_j is statistically significant and does not contain the value 0. The null hypothesis is rejected if:

$$|t_j| > t_{n-p}(1 - \alpha/2), j = 0, 1, \dots, k.$$

(Fahrmeir et al., 2013, p.131).

F-test: The null and the alternate hypothesis are given by:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \quad H_1 : \beta_j \neq 0; \text{ for at least one value of } j$$

This test is used to check if there exists a linear relationship between the dependent variable and any regression coefficient β_j. The F-test performs the test considering all coefficients simultaneously and the F-statistic is determined by the following formula:

$$F = \frac{n-p}{k} \frac{R^2}{1-R^2}$$

If the value of R^2 is small, then the null hypothesis is retained as the value of F would also be smaller. But, if the value of F is larger or the R^2 value is close to 1, the null hypothesis is most likely to be rejected (Fahrmeir et al., 2013, p.131-133).

3.2.2 P value and Confidence interval

The P value is a probability value, where it measures the existence of the data by a random chance. The P value can take any value from 0 to 1. Smaller P value provides strong evidence to reject the null hypothesis and if the P value is higher it provides significant evidence to retain the null hypothesis (Levin, 2011, p.385).
The probability of obtaining the value for the regression parameter lying in certain range of values is known as confidence interval. It can be constructed on each individual regression coefficient β_j, $j = 0, \ldots k$. Under the conditions of normality, the confidence interval can be calculated with the help of t_j. The null hypothesis is rejected when $|t_j| > t_{n-p}(1-\alpha/2)$, where α is the significance level. After constructing the test in a way that the probability of rejection of H_0 equals α, when H_0 is true. The following is obtained:

$$P(|t_j| > t_{n-p}(1-\alpha/2)) = \alpha$$

Then, the probability of not rejecting the null hypothesis H_0, when it is true is given by:

$$P(|t_j| < t_{n-p}(1-\alpha/2)) = \; P(|(\hat{\beta}_j - \beta_j)/se_j| < t_{n-p}(1-\alpha/2)) = 1-\alpha$$

Which is equivalent to

$$P(\hat{\beta}_j - t_{n-p}(1-\alpha/2) \cdot se_j < \beta_j < \hat{\beta}_j + t_{n-p}(1-\alpha/2) \cdot se_j) = 1-\alpha$$

The following is obtained from the above equations and is the $(1-\alpha)$ interval for β_j (Fahrmeir et al., 2013, p.136).

$$[\hat{\beta}_j - t_{n-p}(1-\alpha/2) \cdot se_j, \hat{\beta}_j + t_{n-p}(1-\alpha/2) \cdot se_j]$$

3.3 Best subset selection

To construct an efficient regression model in terms of performance and accuracy, the best subset of independent variables are selected and the variables whose impact doesn't make provide significant evidence for prediction are not considered. Determining best subset leads to the reduction of computational effort and provides info about the variables that may not be necessary to predict the outcome. The OLS is used to fit the linear model and constructs $2^k - 1$ models for k independent variables. AIC and BIC are used as the model selection criteria (Fahrmeir et al., 2013, p.146).

3.3.1 Akaike Information Criterion(AIC)

It is the criteria for choosing a best model among all other models. Here, AIC value is calculated for each model and the models with smaller values of AIC are considered to be best models. AIC is calculated as follows:

$$AIC = -2l(\beta_M, \hat{\sigma}^2) + 2(|M| + 1)$$

Here, $l(\beta_M, \hat{\sigma}^2)$ is the maximum value of the log-likelihood function. This value is obtained when the maximum likelihood estimators $\hat{\beta}_M$ and $\hat{\sigma}^2$ are inserted into the log-likelihood function. $\hat{\sigma}^2$ is given by $\hat{\sigma}^2 = \hat{\varepsilon}'\hat{\varepsilon}/n$ Here, the total number of parameter is |M|+1 as the error variance σ^2 is also counted as a parameter. For a linear model with Gaussian errors, the above equation can be simplified as follows:

$$\begin{aligned} -2l\left(\hat{\beta}M, \hat{\sigma}^2\right) &= n\log\left(\hat{\sigma}^2\right) + \frac{1}{\hat{\sigma}^2}\left(y - XM\hat{\beta}M\right)'\left(y - XM\hat{\beta}_M\right) \\ &= n\log\left(\hat{\sigma}^2\right) + \frac{n\hat{\sigma}^2}{\hat{\sigma}^2} \\ &= n\log\left(\hat{\sigma}^2\right) + n \end{aligned}$$

And by substituting this in AIC, we get AIC $= n \cdot \log(\hat{\sigma}^2) + 2(|M| + 1)$ (Fahrmeir et al., 2013, p.148).

3.3.2 Bayesian Information Criterion(BIC)

BIC is also a best model selection criterion, where the process of selection is same as of AIC. The model with the smallest BIC value is preferred to be the best model. The formula for BIC is as follows:

$$\text{BIC} = -2 \cdot l\left(\hat{\beta}_M, \hat{\sigma}^2\right) + \log(n)(|M| + 1)$$

Assuming a linear model with Gaussian errors, the following is obtained.

$$\text{BIC} = n \cdot \log\left(\hat{\sigma}^2\right) + \log(n)(|M| + 1)$$

where n is the number of observations, (M+1) is the number of parameters and $\hat{\sigma}^2$ is the estimated error variance for the corresponding model. Penalizing the complex models more than what AIC does, is the main property which differentiates it from AIC (Fahrmeir et al., 2013, p.149-150).

4 Statistical analysis

In this section, the statistical methods stated above are applied to the processed data and the results are interpreted. For the purpose of complex calculations and visualization of plots, R (version 4.1.2) R Development Core Team (2020) is used with the help of its inbuilt functions and the `olsrr` Aravind, H. (2020), `psych` Revelle (2021) packages.

4.1 Descriptive analysis

A general descriptive analysis is performed describing how the data is distributed and various central tendency values are summarized in the table1. The data set contains 438 observations and 9 variables. The average value for the overall price of the car models and its standard deviation are 14.68 and 7.75 respectively. For the *mileage*, *mpg*, and *tax*; the average values are 25.11, 58.72, and 96.80 respectively and the standard deviations for the same are 25.04, 17.75, and 61.65 respectively. Petrol, Diesel, and Hybrid are three categories for the *fuelType* and Manual, Automatic, and Semi-Auto are the three types of transmission modes for the cars. The figure1 provides visual description using box plots.

4.2 Data preparation

To proceed with further analysis, some variables are transformed accordingly. For the *price* of the car, its logarithmic value is calculated and is stored in a new variable *logprice*. A new variable named *lp100* is created and has the ratio of 282.48 and the *mpg* i.e. $lp100 = 282.48/mpg$. Lastly, the age for the cars is calculated by subtracting the year in which the car was first registered with 2020 and is stored in the *cars_age*.

4.3 Analysis using Linear regression

Choosing best response variable

As there are two response variables, the best among them will be chosen for linear regression analysis. For this, two linear regression models are created with each having the response variable as *price* and *logprice* respectively. In the model generated, the categorical variables are dummy encoded automatically by the R software. To choose the best one, we take the help of model diagnostic tools i.e., the residual vs fitted plot and the normal Q-Q plot.

In the figure3, for the residual vs fitted plot, it can be noticed that the values are equally scattered and close to the reference line for the model with the response variable as *logprice* and proves the assumption of having homoscedastic error variance. And, in the figure5, which shows the q-q plot for the same model, it can be seen that the normal distribution of the standardized residual is more optimal. This is can proven as most of the observations are in line with the reference line(dotted line) in the normal q-q plot. The random sampling of the data makes sure that the data is identically and independently distributed and constructing the linear model with only the transformed variables assumed no multicollinearity among the independent variables.

The residual vs fitted plot in the figure2 for the model having response variable as *price* shows more variance in residuals when compared to the figure and it can be also seen that the observations are little apart from the reference line. In the figure4, we can see that the normality of the standardized residuals are not optimal when compared to the other model as the observations are not in line with the reference line. With this it can be said that the normality assumption is violated.

Finally, the *logprice* is chosen as the best response variable and will be considered as the dependent variable or the response variable for further analysis.

Choosing the best set of explanatory variables

To choose the best explanatory variables that impacts the value of the response variable, the function `ols_step _all _possible()` from `olsrr` package of the R software provides us with 255 models. Using the Akaike Information Criterion (AIC) for selecting the best subset of the explanatory variables, we obtain the following variables: *model, cars_age, mileage, lp100, fuelType, engineSize,* and *transmission*. The AIC value obtianed is -704.41.
Now, using the Bayesian Information Criterion (BIC) for selecting the best subset of the of the explanatory variables, we obtain the following variables: *model, cars_age, mileage, fuelType, engineSize* and *transmission*. The BIC value obtained is -659.34. The newly created variable $lp100$ was not considered by BIC. Now, this subset of explanatory variables are considered for further analysis, as these variables are chosen preferring to BIC.

Estimation of best linear model

After choosing the best subset of explanatory variables, we create a linear model with the builtin `lm()` function in R. The following model is generated:

$$\begin{aligned} logprice_i = 2.65 + 0.15 \cdot model(T-Roc)_i - 0.52 \cdot model(Up)_i - 0.09 \cdot carsage_i - 0.01 \cdot mileage_i \\ + 0.46 \cdot fuelType(Hybrid)_i + 0.12 \cdot fuelType(Petrol)_i + 0.29 \cdot enginesize_i \\ - 0.12 \cdot transmission(Manual)_i - 0.0 \cdot transmission(Semi - Auto)_i \end{aligned}$$

Firstly, we examine all the assumptions of the linear model to the above generated model. The residual vs fitted plot in the figure6 shows the the spread of observation points are constantly linear with respect to the reference line, which proves the assumption of linearity. In the same plot, it can be seen that the observations are all close to the reference line or the zero line, which proves the assumption of homoscedastic variance of the residuals. As the data is randomly sampled, the observations are independent and identically distributed. Lastly, for examining the normality of the standardized residuals, standardized residuals are calculated and plotted against the normal distribution in a normal q-q plot. Almost all the observations in the figure7 lie on the reference line, which proves that the standardized residuals follow a normal distribution.

For the generated model, the coefficient of Intercept is 2.65. The variables *model(T-Roc), FuelType(Hybrid), FuelType(Petrol)* and *enginesize* have a positive impact in the prediction of *logprice* and their value for the coefficients are 0.15, 0.46, 0.12 and 0.29 respectively. The impact on the *logprice* is maximum by the variable $FuelType(Hybrid)$, as it has the highest value for coefficient among all variables and $fuelType(Petrol)$ has the lowest impact on it. The variables showing negative impact on the response variable are *model(Up), cars_age, mileage* and $transmission(Manual)$ with having the value for coefficients as -0.52, -0.09, -0.01 and -0.12 respectively. Since the value for the $cars\_age$ is -0.09, it has the most negative impact on the prediction and $model(Up)$ has the lowest negative impact. The values are also tabulated in table2.

For the above model, we performed the t-test considering the null hypothesis H_0 as the value for the regression coefficient to be 0 and for the alternate hypothesis, we assume the value for the regression coefficient is not 0. With this, we test the statistical significance of each variable. This fails in rejection of null hypothesis for the variable *transmission(Semi-Auto)* as the p-value is 0.839 and is greater when compared to the significance level. Whereas, the p-value for rest other variables is <0.001. This shows that *transmission(Semi-Auto)* has no impact on the price of the car.

After testing the statistical significance of each variable. We determine the confidence interval for each variable considering 0.05 as the significance level. The confidence interval of the $fuelType(Petrol)$ is 0.08 to 0.16 and means that the average value for that variable can take any value in the derived interval with 95 percent confidence. It will always have a positive impact on the price as the confidence interval lies above 0. The confidence interval for the rest other variables is stated in the table2 and can be referred to for more information.

Lastly, to determine the goodness of fit, we calculate the value for R^2 and $R^2_{adjusted}$. The values obtained are 0.963 and 0.962 respectively. As both the values are close to 1, it shows that the model is best fitted to the data having *logprice* as the output variable and the variables derived by the BIC as the best explanatory variables.

5 Summary

In this report, we have been given a dataset which is an extract from a large dataset available on "www.Kaggle.com". The dataset contains information on price, year, mileage, miles per gallon, fuel type, engine size, tax, and transmission for three different models

of cars, which are manufactured by Volkswagen. The variables *model*, *transmission*, and $fuelType$ were categorical and the rest were numerical. There are 438 observations in the dataset. The information is about the cars which were sold in the year 2020 on a used car platform named Exchange and Mart, which is based in the UK. The dataset has 438 observations. The goal was to perform regression on the dataset and lastly obtain a best linear model fitting to data. Three new variables are created for further analysis. $lp100$ is created by transforming *mpg* denoting the liters of fuel consumed per 100 kilometers, *logprice* containing log-transformed values of *price* and age of the car is calculated using *year*.

Further, two multiple linear regression models were fitted with having response variable as *logprice* and *price* respectively. Using model diagnostic tools, we preferred to consider log price as the response variable for further analysis. The linear model with the *logprice* as response was more in line with the assumptions of the linear model.

After selecting the response variable, we obtained 255 different models. Here, we used OLS estimator for estimating the regression coefficients. Out of those 255 models, we had to choose the best model with AIC and BIC selection criteria. We obtained the AIC value as -704.41 and the following variables for the model: *model*, *age*, *mileage*, $lp100$, $fuelType$, *engineSize* and, *transmission*. And, using BIC, we obtained the following variables: *model*, *cars_age*, *mileage*, $fuelType$ *engineSize* and *transmission* with a BIC value of -659.34.

After having the best response variable and the best subset of the explanatory variables, a linear model is constructed using them. After that, a t-test is performed on the model and we found that the variable $transmission(Semi - Auto)$ is not significant. With the help of the sign of the coefficients, we found out that $model(T-Roc)$, $FuelType(Hybrid)$, $FuelType(Petrol)$ and *engineSize* had a positive impact on the price of the car, whereas the rest others had a negative impact on the price of the car. The confidence interval values of the various variables were also determined. Lastly, the goodness of fit is determined by calculating the value of R^2 and $R^2_{adjusted}$. And the value obtained for the same were 0.963 and 0.962 respectively, which were close to 1 and portrays that the model is of good fit to the data.

For further studies, it would be useful to study what other factors are responsible for the varying car prices in the other countries too and what changes in the factors can cause the value of a used car to deteriorate. Finally, we can determine how these changes affect the customer's car buying behavior on the whole.

Bibliography

Aravind, H. *olsrr: Tools for Building OLS Regression Models*, 2020. URL `https://CRAN.R-project.org/package=olsrr`. R package version 0.5.3.

Exchange Enterprises, Newsquest Media Group. Exchange and mart, 2022. URL `https://www.exchangeandmart.co.uk/`. (visited on 21st Jan 2022).

L. Fahrmeir, T. Kneib, S. Lang, and B.D. Marx. *Regression: Models, Methods and Applications*. Springer Berlin Heidelberg, 2013. ISBN 9783662638811.

Kaggle Inc. kaggle, 2021. URL `https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes?select=vw.csv`. (last visited on 19th Jan 2022).

R.I. Levin. *Statistics for Management*. Pearson Education, 2011. ISBN 9788177585841.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

Alvin C. Rencher and G. Bruce Schaalje. *Linear Models in Statistics*. John Wiley Sons, Ltd, 2008. ISBN 9780470192603.

William Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois, 2021. URL `https://CRAN.R-project.org/package=psych`. R package version 2.1.9.

Volkswagen AG. Volkswagen, 2022. URL `https://www.volkswagen.de/`. (last visited on 21st Jan 2022).

Appendix

A Additional figures and tables

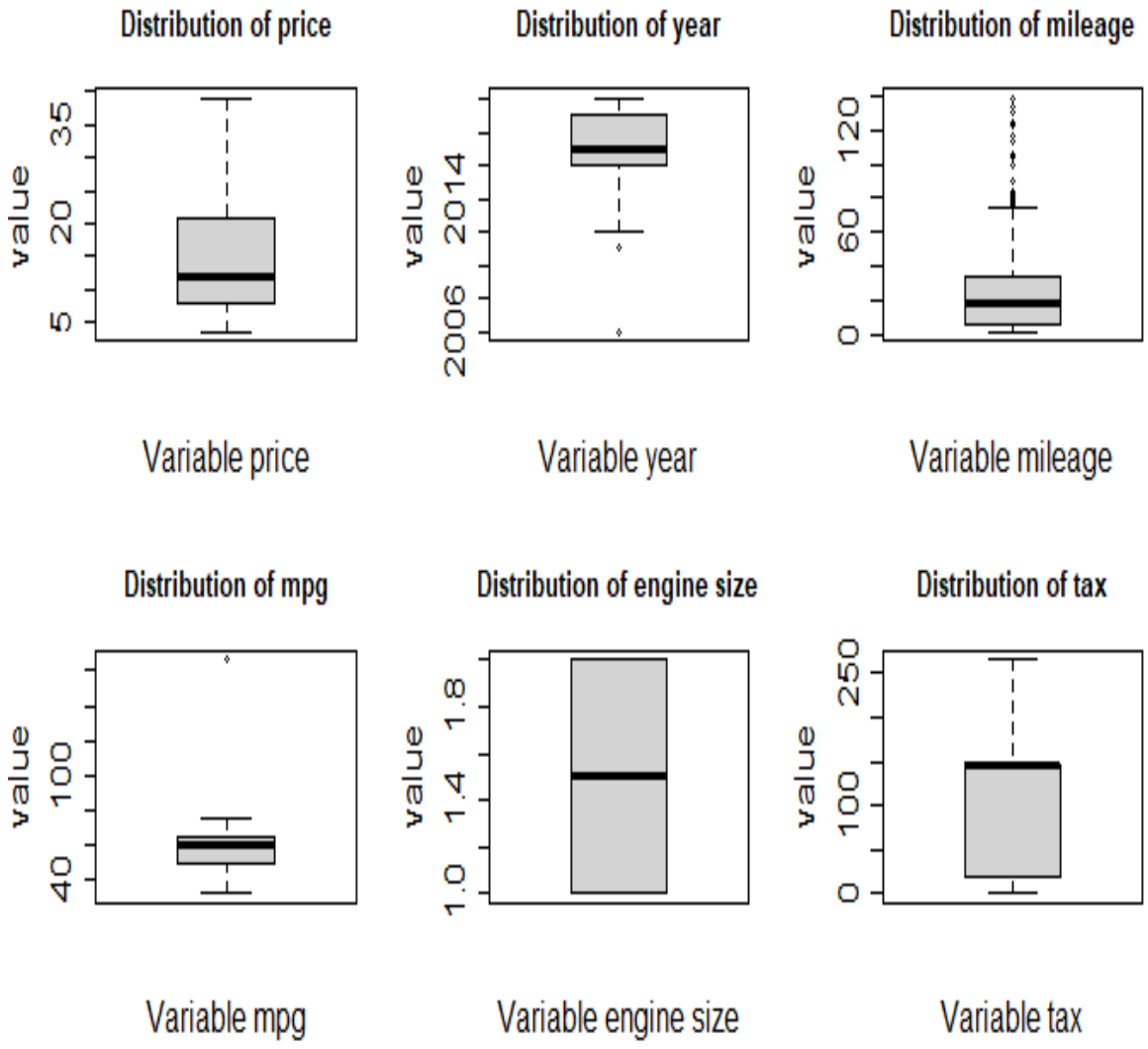

Figure 1: Boxplot for all variables

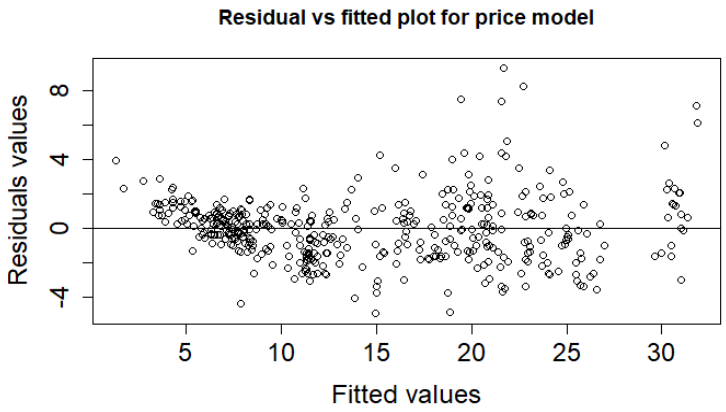

Figure 2: Residual vs fitted plot: Price as response variable

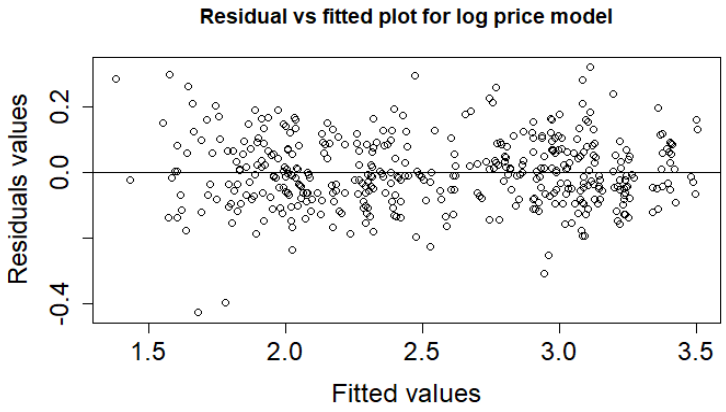

Figure 3: Residual vs fitted plot: Log price as response variable

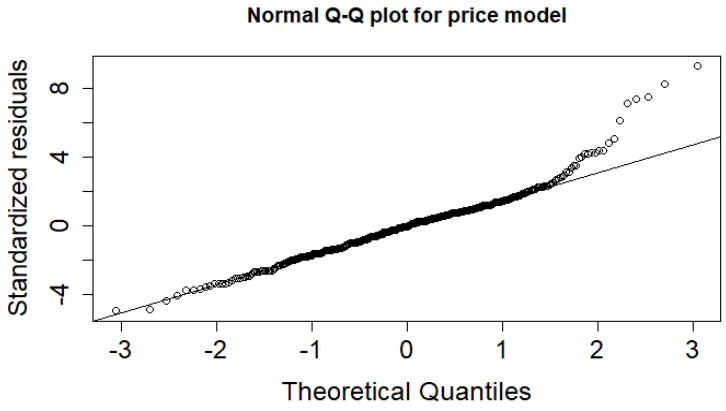

Figure 4: QQ plot: Price as response variable

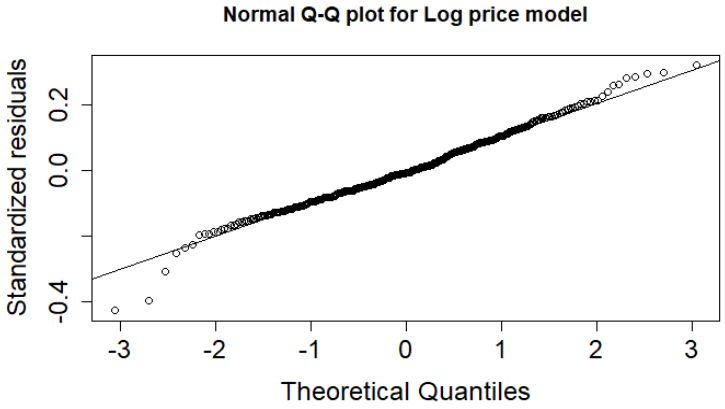

Figure 5: QQ plot: Log price as response variable

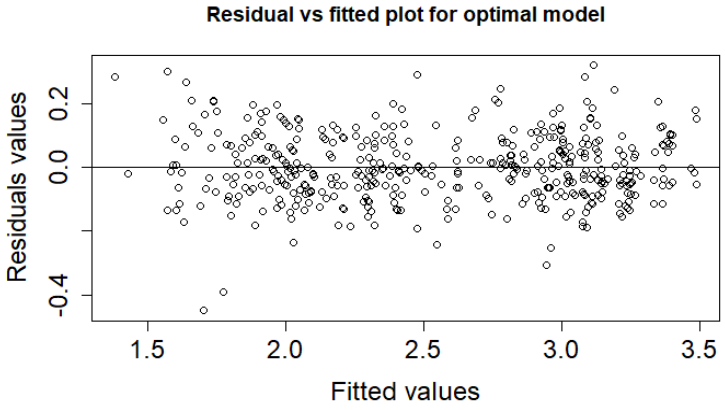

Figure 6: Residual vs fitted plot: Best model

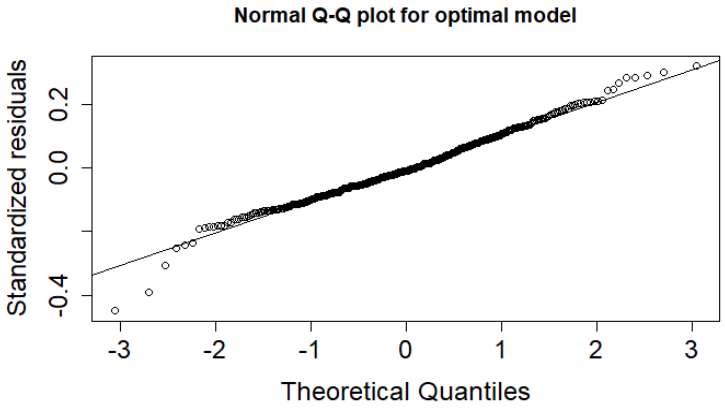

Figure 7: QQ plot: Best model

Table 1: Descriptive analysis of the numerical variables present in the dataset

	min	max	mean	median	sd
price	3.50	38.99	14.68	12.00	7.75
year	2006.00	2020.00	2017.21	2017.00	1.96
mileage	1.20	138.57	25.11	17.53	25.04
mpg	32.50	166.00	58.72	60.10	17.75
engineSize	1.00	2.00	1.47	1.50	0.42
tax	0.00	265.00	96.80	145.00	61.65

Table 2: Summary of results from best estimated linear model

	Estimate	Std. Error	t value	p value	Confidence interval
(Intercept)	2.65	0.06	41.88	<0.0001	2.53 - 2.77
modelT-Roc	0.15	0.02	8.59	<0.0001	0.12 - 0.19
modelUp	-0.52	0.02	-21.41	<0.0001	-0.57 - -0.47
carsage	-0.09	0.00	-22.70	<0.0001	-0.10 - -0.08
mileage	-0.01	0.00	-18.70	<0.0001	-0.01 - -0.01
fuelTypeHybrid	0.46	0.04	12.43	<0.0001	0.39 - 0.54
fuelTypePetrol	0.12	0.02	6.25	<0.0001	0.08 - 0.16
engineSize	0.29	0.03	9.77	<0.0001	0.23 - 0.34
transmissionManual	-0.12	0.02	-6.47	<0.0001	-0.16 - -0.09
transmissionSemi-Auto	-0.00	0.02	-0.20	0.83	-0.04 - 0.04
Observations	438				
R^2	0.963				
Adjusted R^2	0.962				