TU Dortmund

Introductory Case Studies

Project 3: Linear regression

Lecturer:
Dr. rer. nat. Maximilian Wechsung

Author: Kartik Kamboj

Group number: 1

Group members: Bidhan Chandra Roy, Gitto Benny, Hemani Chand, Jessia Erappakkal Joby, Kartik Kamboj

June 20, 2022

Contents

1 Introduction

Exchange and Mart is a UK-based e-commerce site where people can find great deals on new and used cars. It also offers a large selection of new and used four-wheelers and two-wheelers from manufacturers such as Audi, BMW, Volkswagen, Mercedes, Ford, and others. It also gives you the option of selling your automobile. The site provides a range of filters, including model, price, mileage, fuel type, etc., providing a user-friendly interface for searching and buying vehicles that meet customer needs.

The objective of this report is to build a linear regression model on a small extract of a large dataset of Exchange and Mart available online on the website of Kaggle. The dataset is limited to the used cars of the manufacturer Volkswagen (VW), which were advertised online in the year 2020. It includes some of the factors of used cars like model name, price of the car, year of registration, mileage, mpg (miles per gallon), fuel type, engine size, tax and transmission (type of gearbox). The report determines which factors or explanatory variables affect the price of a used car and by how much they influence it.

Firstly, the data is prepared by converting miles per gallon and year into litres per 100 km and age, respectively. After variable conversion, the log transformation of the variable, price, is compared with the untransformed one by estimating a full model to get an idea of which kind of response variable suits the study. Then the technique of best subset selection examines all possible combinations of explanatory variables and determines the most important set of explanatory variables, using the AIC, that predicts the price of the car. Finally, the selected set of explanatory variables are interpreted to get an idea of how each variable is impacting the price of the car and their significance, including the goodness of fit of the model.

The structure of the data contained in the dataset, as well as a discussion on data quality, are covered in greater depth in the second section. It also goes over the project's goals in greater depth. All of the statistical methodologies and techniques utilized in this research are detailed in the third part. The fourth segment discusses how to interpret results using statistical methods described in the third section. The fifth and final section concludes with a summary of the findings and recommendations for further data analysis.

2 Problem statement

2.1 Data set and data quality

The following dataset is an extract from a larger dataset accessible on the kaggle website (Kaggle, 2022) that contains information about 2532 used cars advertised on the UK-based e-commerce platform, Exchange and Mart (Exchange and Mart, 2022), in 2020. The dataset is restricted to Volkswagen (VW) used cars and includes different features like model, price, year, mileage, mpg, fuel type, engine size, tax and transmission. We have three categorical variables ($model$, $transmission$, $fuelType$) and six continuous variables ($price$, $mileage$, mpg, $year$, $engineSize$, tax). The car's $model$ is divided into three categories: Passat, T-Roc and Up. The $year$ denotes the car's first registration year, from 2006 till 2020. The variable $transmission$ refers to the car's gearbox, which can be automatic, manual, or semi-automatic. The $fuelType$ specifies the type of fuel consumed by the vehicle, which can be diesel, hybrid, petrol, or other. The $price$ of a car is measured in GBP (British Pound Sterling / £). The car's $mileage$ is the entire distance it has been driven, and it is expressed in miles. The mpg (miles per gallon) is a measurement of how far a car can go in miles on a single gallon of fuel. The $engineSize$ is the size of the car's engine, measured in litres, with a maximum capacity of 2 litres and is estimated to be one decimal place. The Vehicle Excise Duty in GBP, to be paid for the car is given in variable tax. Furthermore, the data quality is overall good, as there are no missing values in any of the dataset's 9 critical variables, discussed in this segment.

2.2 Project Objectives

The data is first preprocessed to make it more valuable and understandable for later use in the study. Then, using all the factors in the dataset, construct a full linear regression model that influences the car's price and another full linear model that influences the log-transformed price. As a result of all the assumptions checks for a linear model, the relevant response variable is determined and a suitable set of independent variables is chosen by utilizing best subset selection under the AIC criteria. The explanatory variables for the best fitted linear model are then interpreted using the t-test based on their regression coefficient values and p-values, as well as their confidence interval, to

determine their relevance to the response variable. Finally, the fitted linear regression model's goodness of fit is evaluated using R^2.

3 Statistical methods

The statistical approaches used to analyze the data collection are explained in this section. All the following visualizations were generated using the software R (R Core Team, 2020), version 4.0.5. The R packages used in the project for calculations and visualizations are ggplot2 (Wickham, 2016), olsrr (Hebbali, 2020), car (Fox and Weisberg, 2019a) and carData (Fox and Weisberg, 2019b).

3.1 Linear Regression

Linear regression is the process of using one or more explanatory factors to predict a dependent or response variable. The linear correlations between dependent and independent variables are described using a mathematical function in linear regression. A simple linear regression model is used when the value of the dependent variable is predicted using only one independent variable, while a multiple linear regression model is used when numerous independent variables are involved (Black, 2006, p. 469).

3.1.1 The linear regression model

Assume there is a dependent or response variable y and a collection of independent or regressors or explanatory variables, x_1, x_2, x_3, ..., x_k. The following is a representation of the relationship between these variables:

$$y = f(x_1, ..., x_k) + \epsilon$$

where, $f(x_1, ..., x_k)$ is the unknown mathematical function of covariates, x_1, ..., x_k and ϵ is a independent and identically distributed (i.i.d.) random error, which is the deviation of the observed value from the true value. The unknown function f can be determined by combining covariates, $\boldsymbol{x} = (1, x_1, ..., x_k)^{'}$ and unknown coefficients, $\boldsymbol{\beta} = (\beta_0, ..., \beta_k)^{'}$, where, β_0 signifies the intercept, and estimating each into p = k + 1 dimensional vectors,

$$f(\boldsymbol{x}) = \boldsymbol{x}^{'}\boldsymbol{\beta}.$$

Hence the linear model equation for observed data y_i can be predicted by using explanatory variables of each observation $x_i = (1, x_{i1}, ..., x_{ik})^{'}$ and the random error ϵ_i , $i = 1, ..., n$, occurred in predicting the i^{th} observation.

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \epsilon_i = \boldsymbol{x_i}^{'}\boldsymbol{\beta} + \epsilon_i.$$

To be summarized, $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$,

where, $\boldsymbol{y}$ and $\boldsymbol{\epsilon}$ are vectors and $\boldsymbol{X}$ is a design matrix of independent variables of a set of observed data, defined as,

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad and \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x_1}^{'} \\ \vdots \\ \boldsymbol{x_n}^{'} \end{pmatrix}.$$

We consider that design matrix, $\boldsymbol{X}$, has full column rank, which means that $rank(\boldsymbol{X}) = k + 1 = p$, implying that $\boldsymbol{X's}$ columns are linearly independent. The number of observations n must exceed the number of regression coefficients p.

There are some necessary assumptions for a linear model which have to be fulfilled before finalizing the model. Initially, the observations should be independent of each other, which means a predictor should not be a linear transformation of another predictor. Independent variables should not be highly correlated to each other or no multicollinearity in the data. The expectation of errors on covariates has to be equal to 0. Error variance should be constant across all observations, so $Var(\epsilon_i) = \sigma^2$ and also errors should be uncorrelated, $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$. The errors should be roughly normally distributed, $\epsilon \sim N(0, \sigma^2 I)$, where, 0 is the mean and $\sigma^2 I$ is the variance of the errors (Fahrmeir et al., 2013, p. 73-77).

3.1.2 Estimation of model parameters

The unknown value of regression coefficients $\boldsymbol{\beta}$ can be calculated using the principle of least squares (LS) and maximum likelihood (ML). The principle of least squares says

that the sum of the squares of random error is minimized. So, based on the equation, $\boldsymbol{y_i = x_i^{'}\beta + \epsilon_i}$, random error can be defined as the difference between each of the observed values to its fitted value.

$$\boldsymbol{\epsilon_i = y_i - x_i^{'}\beta}$$

Therefore,

$$LS(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \boldsymbol{x_i^{'}\beta})^2 = \sum_{i=1}^{n}{\epsilon_i}^2 = \boldsymbol{\epsilon^{'}\epsilon} = \boldsymbol{(y - X\beta)^{'}(y - X\beta)}$$

We minimize $LS(\boldsymbol{\beta})$ by differentiating it with respect to $\boldsymbol{\beta}$, setting it to zero and demonstrating that the matrix of the second derivative is positive definite. The final outcome is,

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

The least squares estimator of $\boldsymbol{\beta}$ is equivalent to the maximum likelihood estimator of $\boldsymbol{\beta}$ under the assumption of normally distributed errors. The log-likelihood function is

$$l(\boldsymbol{\beta, \sigma^2}) = -\frac{n}{2}log(2\pi) - \frac{n}{2}log(2\sigma^2) - \frac{1}{2\sigma^2}\boldsymbol{(y - X\beta)^{'}(y - X\beta)}.$$

Maximizing the log-likelihood with respect to $\boldsymbol{\beta}$ yields the same result as in minimizing least squares $\boldsymbol{(y - X\beta)^{'}(y - X\beta)}$.

We can estimate the response variable using the least squares estimator, $\hat{\boldsymbol{\beta}}$,

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

where, $\boldsymbol{H = X(X^{'}X)^{-1}X^{'}}$, known as the hat matrix or prediction matrix.

With the help of the hat matrix ($\boldsymbol{H}$), we can calculate the residuals. A residual ($\hat{\boldsymbol{\epsilon}}$) is the deviation of a point from the regression line. It is the difference between the observed value and predicted or estimated value of the response variable.

$$\boldsymbol{\hat{\epsilon} = y - \hat{y} = y - Hy = (I - H)y}$$

The estimated value of error variance is calculated by using maximum likelihood principles,

$$\hat{\sigma}^2 = \frac{1}{n-p}\boldsymbol{\hat{\epsilon}^{'}\hat{\epsilon}}$$

where, n is the number of observations while p is the number of parameters including the intercept (Fahrmeir et al., 2013, p. 104-109).

Residuals generally have unequal variability, which leads to a violation of the assumption of constant variance in errors across observations. Hence, standardization is used to solve such an issue. Standardized residuals (r_i) are obtained by dividing residuals ($\hat{\epsilon_i}$) by the estimated standard deviation of the residuals ($\hat{\sigma}$).

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

where, h_{ii} is the i^{th} diagonal element of the hat matrix (H) (Fahrmeir et al., 2013, p. 123-124).

The least squares estimator's ($\hat{\boldsymbol{\beta}}$) estimated covariance matrix is a symmetric matrix whose diagonal elements represent the least squares estimator's ($\hat{\boldsymbol{\beta_j}}$) estimated variances, $\widehat{Var(\hat{\boldsymbol{\beta_j}})}$,

$$\widehat{Cov(\hat{\boldsymbol{\beta}})} = \hat{\sigma}^2(X^{'}X)^{-1},$$

and the square root of the estimated variances are the estimated standard errors(se_j) (Fahrmeir et al., 2013, p. 116-117),

$$se_j = \widehat{Var(\hat{\boldsymbol{\beta_j}})}^{1/2}, j = 0, 1, ..., k.$$

3.1.3 T-test and confidence interval

The t-test is defined as the ratio between the estimated regression coefficient and the standard error of the coefficient. It tests the significance of the parameters $\hat{\boldsymbol{\beta}}$. The t-statistics is calculated as,

$$t_j = \frac{\hat{\beta_j}}{se_j} \sim t_{n-p},$$

where, n is the number of observations and p is the number of parameters. se_j is the estimated standard error of the estimated parameter $\hat{\beta_j}$.

The following hypotheses are being tested at 0.05 significance level , $H_0 : \beta_j = 0 \ vs$ $H_1 : \beta_j \neq 0$. Under the null hypothesis, the critical value for the rejection area is determined by the (1- $\alpha/2$) quantile of the t-distribution with $(n-p)$ degrees of freedom. If the critical value is less than the t-statistical value, and the p-value of observing this t-statistics under the null hypothesis is less than the significance level (α), the null

hypothesis is rejected (Fahrmeir et al., 2013, p. 135).

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

The confidence interval is the range in which the true value of the coefficients of the covariates can be found. Under normality, t-statistic is used and therefore the confidence interval of the parameter $\hat{\beta}_j$ with level $1 - \alpha$ is

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

The following hypotheses are being tested at 0.05 significance level, $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$. If the confidence interval contains 0, then H_0 cannot be rejected; otherwise, H_0 is rejected.

A 95% confidence interval is the common range that shows that 95% of the true value lies within that interval with a 5% significance level (Fahrmeir et al., 2013, p. 136).

3.1.4 Dummy coding for categorical variable

In order to use categorical variables having more than two characteristics in a regression model, we have to create dummy variables. For a covariate x with c characteristics, we define $c - 1$ dummy variables.

$$x_{i1} = \begin{cases} 1 & x_i = 1, \\ 0 & otherwise, \end{cases} \quad ... \quad x_{i,c-1} = \begin{cases} 1 & x_i = c - 1, \\ 0 & otherwise, \end{cases}$$

When, $i = 1, ..., n$, is used as a dependent variable in the linear regression model then,

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{i,c-1} x_{i,c-1} + ... + \epsilon_i.$$

Due to identifiability, the category that is more common in the dataset is excluded from the dummy variables. To interpret the estimated impacts of new dummy variables, the removed category is used as a reference category. (Fahrmeir et al., 2013, p. 97).

3.1.5 Coefficient of Determination

The coefficient of determination, R^2, is a measure of goodness of fit in a regression model. The goodness of fit tells us how well a statistical model fits with the data. The coefficient of determination is the proportion of variability of the dependent variable explained by the independent variable. It ranges from 0 to 1. When R^2 is close to 0, it means that the independent variables cannot explain any of the variability, and the model does not fit the data well. The closer R^2 is to 1, the better the model fits the data and the more accurate the prediction of variability by independent variables becomes.

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

where, y_i is an i^{th} response variable, $\bar{y}$ is the mean of all response variables, $\hat{y}_i$ is the estimated i^{th} response variable and $\hat{\epsilon}_i$ is the residual at i^{th} observation for n samples (Fahrmeir et al., 2013, p. 112-113).

3.1.6 Multicollinearity

Multicollinearity occurs when two or more predictors or independent variables have a strong correlation with each other. The fundamental problem with multicollinearity is that it reduces the goodness of fit of the model. There are numerous methods for detecting multicollinearity. One method is to perform a correlation test to look for any correlations between possible predictor variables, and another one is to use the Variance Inflation Factor (VIF).

In VIF, a regression analysis is used to predict the value of an independent variable based on the values of the other independent variables. In this scenario, the projected independent variable becomes the dependent variable. It is possible to discover whether any of the independent variable is a function of the other independent variables by repeating this method for each of the independent variables. The R^2 generated from the model is used to compute the VIF and if the VIF is greater than 10, then that is the case of multicollinearity.

$$VIF_i = 1/(1 - R_i^2),$$

where, R_i^2 is the coefficient of determination for the model fitted using i as a dependent variable (Black, 2006, p. 576-578).

3.2 Best Subset Selection

A set of good features is chosen to construct a fit model in a linear regression model. One of the ways is to do subset selection of all independent variables. A large computational effort is needed to perform best subset selection. Suppose there are k independent variables, then there are $2^k - 1$ distinct models for any set of predictors k. -1 is done because we typically ignore the intercept-only model. Finally, the best fitted model is selected with a good set of explanatory variables based on model selection criteria like AIC, BIC, etc. (James et al., 2013, p. 204). In this report, the AIC technique is used to find the best set of explanatory variables.

3.3 Model Selection Criteria

3.3.1 Akaike Information Criterion (AIC)

AIC is a model selection technique that is based on the concept of maximum likelihood. The model with the smallest value of AIC corresponds to a better fit. It is defined as

$$AIC = -2 \cdot l(\hat{\boldsymbol{\beta}}_M, \hat{\sigma}^2) + 2(|M| + 1),$$

where, $l(\hat{\boldsymbol{\beta}}_M, \hat{\sigma}^2)$ is the maximum value of the log-likelihood with parameters $\hat{\boldsymbol{\beta}}_M$ and estimated error variance $\hat{\sigma}^2$. $|M| + 1$ is the total number of parameters including error variance $\hat{\sigma}^2$. If a linear model has Gaussian errors, $-2 \cdot l(\hat{\boldsymbol{\beta}}_M, \hat{\sigma}^2) = nlog(\hat{\sigma}^2) + n$ then $AIC = n \cdot log(\hat{\sigma}^2) + 2(|M| + 1)$, by ignoring constant n (Fahrmeir et al., 2013, p. 148).

3.3.2 Bayesian Information Criterion (BIC)

In a similar form to AIC, BIC is also a model selection technique. Small values of BIC show a good fit of the model. It is also known as Schwartz criteria when multiplied by 1/2. It is defined as

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

where, log(n)(|M| + 1) is the added penalty depends on complexity of the model. If a linear model has Gaussian errors, then BIC is calculated as, $BIC = n \cdot log(\hat{\sigma}^2) +$

$log(n)(|M|+1)$. The BIC gives a more parsimonious model than the AIC as it penalizes complex models much more than the AIC (Fahrmeir et al., 2013, p. 149-150).

4 Statistical analysis

In this section, exploratory data analysis is performed on the dataset. First the data is manipulated and transformed to make the dataset more understandable. Secondly, through the best subset selection technique based on AIC criteria a good set of exploratory variables are analyzed. Then we fit a linear model with the selected set of predictors and interpret the result.

4.1 Data Preparation and Preprocessing

Initially, data is processed and prepared to get the data in a more informative and useful way to improve the performance of a model.

The variable *mpg* (miles per gallon) is converted into a new unit called $lp100km$ (litres per 100 km). Therefore, the new feature $lp100km$ is calculated as $lp100km = 282.48/mpg$. Furthermore, the variable *year* is replaced with the new variable *age*, which is the age of the car from the first registration year till 2020. As a result, the calculation is as follows : $age = 2020 - year$.

So the variables *mpg* and *year* are dropped from the dataset and are replaced by the new variables $lp100km$ and *age*, respectively, to avoid multicollinearity in the modeling.

After preprocessing, the dataset contains 2532 observations with 9 variables, which includes two newly added relevant variables, $lp100km$ and *age*. The $lp100km$, a continuous variable, is the litres of fuel the car needs to travel 100km and has a mean of 5.253 with a minimum value of 1.702 lp100km and a maximum value of 8.692 lp100km. The variable *age* is also continuous and varies from 0 to 14 years. The majority of cars *models* in the dataset is the Passat, i.e., 915, followed by the Up and T-Roc, which are 884 and 733, respectively. The minimum and maximum value of the variable *price* are 1495 GBP and 40999 GBP, respectively, with *mileage* from 1 mile to 176,000 miles. A large set (1821) of cars has a manual *transmission* and only 473 have a semi-auto, while 238 have an automatic *transmission*. 1488 of the listed cars run on petrol and 970 run on diesel. There are 58 hybrid cars in quantity, and only 16 cars run on other fuels. The minimum

amount of tax paid is 0.0 GBP while the maximum tax amount is 265.0 GBP. The minimum *engineSize* is 0 litres and the maximum *engineSize* is 2 litres. Furthermore, there are no missing values in any of the variables in the dataset. Table 1 in the Appendix on page 18 summarizes the minimum, median, mean and maximum of all the continuous variables including the count of categorical variables.

4.2 Response Variable Selection

In the following section, two linear models are compared: one with the original (raw) *price* variable as a dependent or response variable and another with the log transformation of *price* as a response variable, using all other covariates as explanatory variables.

Firstly, a new variable is computed, *log_price*, which is a logarithmic transformation of the variable *price*. The *log_price* is a continuous variable with a minimum value of 7.310 GBP and a maximum value of 10.621 GBP. The mean of the transformed variable is 9.504 GBP.

A linear model's assumptions act as a criterion for determining which response variable best fits the linear model. To begin with, it is believed that the explanatory variables in the dataset are generated separately; thus, one observation has no impact on other observations. We can assume that a car's *model* has no impact on its *mileage*, and that *engineSize* is unaffected by the kind of *transmission*. In Figure 1 in the Appendix on page 17, the Residuals vs Fitted plot, a form of scatter plot with residuals on the y-axis and fitted values on the x-axis, for variable *price* reveals some patterns (like a cone-shaped) around the regression line. In Figure 2 in the Appendix on page 18 the Residuals vs Fitted plot for variable *log_price* does not reveal any structure in the scatter points around the regression line. As a result, we may argue that for *log_price* but not for *price*, there is a linear relationship between covariates and the expected value of errors. This also means that the error terms' expectation based on explanatory factors is zero and that errors are uncorrelated. As a result, the assumptions of linearity and expectation of error terms equal to zero are met. The assumption of multicollinearity in the data is fulfilled for both of the variables, as we can see in Table 2 in the Appendix on page 19 that no independent variable has a VIF value greater than 10.

The normal Q-Q plot in Figure 1 in the Appendix on page 17 shows that variable *price* does not follow the assumption of normality, as the error points show large deviations from the reference line, in contrast to the normal Q-Q plot in Figure 2 in the Appendix on

page 18, of variable *log_price*, which shows fewer deviations in the error points (approx. normal). Overall, we can infer that the normality assumption appears to be violated in *price*, but the normality assumption is followed by the *log_price*. The Scale-location plot, a type of scatter plot with square root of standardized residuals on y-axis and fitted values on x-axis, in Figure 1 in the Appendix on page 17 for variable *price* shows that error terms are not evenly distributed and depict some pattern. This pattern appears to be cone-shaped, i.e., the error variances are not constant across observations, yet the error terms of *log_price* are evenly distributed and show no pattern as given in Figure 2 in the Appendix on page 18. As a result, in comparison to *log_price*, the response variable *price* violates the assumption of constant error variances.

In conclusion, the *log_price* fulfills all the assumptions of a linear model and fits the model better than *price*, and hence the variable *price* is dropped from the dataset for further analysis and the *log_price* is the perfect choice as a response variable.

4.3 Selection of best set of explanatory variables

In this section, to get the best set of explanatory variables, we use the Best Subset Selection technique using AIC criteria. Then based on the AIC values for all the models with best subset of explanatory variables, we select that set of the variables for which AIC value is the lowest among all.

Here, we take all eight independent variables into account and fit the best model using the set of one variable, two variables, and so on until we get a model with all eight variables. Now we employ AIC selection criteria and select the model with the lowest AIC value. We can see in Table 3 in the Appendix on page 19 that the lowest AIC value is for the model with all 8 variables. With an AIC value of -3664.49, the best subset of variables chosen are $model$, $mileage$, $fuelType$, $engineSize$, tax, $transmission$, $lp100km$, age.

4.4 Interpretation of selected set of explanatory variables

The model is fit again using the best subset of explanatory variables and the assumptions of a linear model are checked again. The resulting figures and tables are the same as what we have seen previously in Figure 2 and Table 2 in the Appendix on page 18 and 19, respectively. The predictors are independent due to the nature of the dataset and also the even distribution of errors around the regression line validates that it follows the

assumptions of linearity in covariates, full rank of the design matrix, and the expected value of errors conditioned on covariates is 0. The model follows the assumption of normality as depicted in the normal Q-Q plot. All standardized residuals fall on the reference line with very little deviation at the tail ends. The scale-location plot does not show any pattern and the standardized residual points are evenly distributed. Therefore, it satisfies the assumption of constant error variance across the observations. Table 2 in the Appendix shows that all of the variables have a VIF of less than 10, hence the assumption of multicollinearity is also fulfilled.

Table 4 in the Appendix on page 19 depicts the value of estimates, standard error, t-value and p-value of the fitted linear model after model selection using AIC. The regression model has generated two dummy variables ($modelT - Roc$ and $modelUp$), and $modelPassat$ is the default reference variable for the three $model$ categories (Passat, T-Roc, and Up). Similarly, $fuelType$ (diesel, petrol, hybrid, and other) is split into three dummy variables ($fuelTypeHybrid$, $fuelTypeOther$, $fuelTypePetrol$) and a reference variable ($fuelTypeDiesel$). Furthermore, $transmission$ has two dummy variables ($transmissionManual, transmissionSemi - Auto$) and one reference variable ($transmissionAutomatic$). In total, six covariates have a negative effect on the response variable, and the other six covariates have a positive effect on the response variable.

The model intercept is roughly 9.653 if we keep all other covariates at 0. But this has no physical meaning if we do not involve other covariates. As a result, if all other independent variables remain constant, cars of $modelT - Roc$ tend to increase the log_price of the car by 0.112 units, while cars of $modelUp$ tend to decrease the log_price of the car by 0.568 units. Similarly, if $fuelTypeDiesel$ is a reference variable, the log_price increases by 0.435 units for $fuelTypeHybrid$, 0.072 units for $fuelTypeOther$, and 0.076 units for $fuelTypePetrol$ while all other covariates remain constant. $transmissionManual$ and $transmissionSemi - Auto$ have an estimate of -0.120 and 0.000, respectively. With 1 unit increase in $transmissionManual$, log_price decreases by 0.120 units but with 1 unit increase in $transmissionSemi - Auto$, approximately no change in log_price. The variables, $mileage$ and tax also have an approximately negligible effect on log_price i.e., if there is a 1 unit increase in $mileage$ then log_price decreases by a very small unit (approx., 0) when all other covariates are constant, and similarly, log_price decreases by a very small unit (approx., 0) with a 1 unit increase in tax. When all other variables are constant, $engineSize$ tends to increase the log_price by a factor of 0.177, $lp100km$ has a 0.034 unit positive effect, and age has a 0.093 unit negative effect on log_price.

We can conclude that $fuelTypeHybrid$ has the greatest effect on increasing log_price and $modelUp$ has the greatest effect on decreasing log_price.

Now, to test the statistical significance of each variable, we obtain the corresponding t-value from the estimated value of the variable divided by its standard error and the corresponding probability (p-value) of observing this t-value under the null hypothesis with 2519 degrees of freedom. The null hypothesis to be tested is that the regression coefficient (β) of a variable is equal to 0 and the alternate hypothesis is that the regression coefficient is not equal to 0. The statistical significance value (α) is set at 0.05. We can see in Table 4 that only variable $transmissionSemi - Auto$ has a p-value of 0.983, which is greater than 0.05. So we fail to reject the null hypothesis and conclude that the coefficient of this variable is approximately equal to 0. All other variables, $modelT - Roc$, $modelUp$, $mileage$, $fuelTypeHybrid$, $fuelTypeOther$, $fuelTypePetrol$, $engineSize$, tax, $transmissionManual$, $lp100km$ and age, have the p-value less than the significance level (0.05). Therefore, we reject the null hypothesis and conclude that their estimates are not equal to zero and have a significant effect on the response variable, log_price.

We can see in the Table 5 in the Appendix on page 20 that $transmissionSemi - Auto$ is the only variable that has 0 in its 95% confidence interval and no other variable has 0 in their 95% confidence interval. So we fail to reject the null hypothesis for this variable and it has no significant effect on log_price at $\alpha = 0.05$ whereas all other variables of the regression model reject the null hypothesis and have a significant effect on the log_price at $\alpha = 0.05$. The coefficient of determination (R^2) is 0.954, indicating that the model accounts for approximately 95% of the variation in log_price. Hence the model fits better with the set of explanatory variables and is efficient.

5 Summary

For this study, we investigated the dataset of used cars by Volkswagen (VW) advertised on Exchange and Mart in 2020. The Introductory Case Studies lecturer compiled the dataset from the website of Kaggle. It contains information on 2532 used cars with 9 features. The main purpose of this report is to fit the best linear model with the best set of predictors.

Initially, we performed data preprocessing by converting miles per gallon into litres per 100 km by dividing 282.42 by mpg for a better understanding of the feature. The year

of the car's first registration is replaced by the age of the car by subtracting the year of the car's first registration from 2020. Also, we transformed the price of the car into its log form to compare it with the untransformed to get a perfect response variable for the linear model. The logarithmic transformation of price followed all the assumptions of a linear model more rigorous than an untransformed price, and hence log_price selected as the response variable or dependent variable.

The set of all 8 variables was the best set of explanatory variables among all the potential best subsets of predictors, as per AIC. The variables, $model$, $mileage$, $fuelType$, $engineSize$, tax, $transmission$, $lp100km$, and age are the best subset of explanatory factors chosen. Using log_price and the given set of explanatory variables, we estimated a fit model. The new fitted linear model followed all of the necessary assumptions for a linear model. As a result, we explained the model's coefficients and their importance, offered confidence intervals for the regression parameters, and assessed the model's goodness of fit. The variable $modelUp$ has the greatest negative impact on the log_price, while the $fuelTypeHybrid$ has the greatest positive impact. The $transmissionSemi - Auto$ has a p-value greater than the significance level (0.05), indicating that it failed to reject the null hypothesis, whereas all other variables rejected the null hypothesis. As a result, $transmissionSemi - Auto$ has no influence on $logprice$. We also deduced from the $transmissionSemi - Auto$ confidence interval at 95% that this variable has 0 in the interval, indicating that it was not statistically significant. Finally, the coefficient of determination (R^2), 0.954, was used to determine the goodness of fit and stated that the predictors account for nearly 95% of the variation. We got the best fitted model.

To improve the analysis and obtain more reliable and accurate results, we can extract more records and expand the dataset to include new non-collinear variables such as body type, color and number of airbags. In further analysis, we can use BIC as a selection technique to obtain a more parsimonious model.

Bibliography

Ken Black. *Business statistics for contemporary decision making*. John Wiley & Sons, Inc., 2006. ISBN 978-0470-40901-5.

Exchange and Mart. *Exchange and Mart*, 2022. URL `https://www.exchangeandmart.co.uk/`. (visited on 15th June 2022).

Ludwig Fahrmeir, Thomas Kneib, and Stefan Lang. *Regression: Models, Methods und Applications*. Springer New York, 2013. ISBN 978-3-642-34332-2.

John Fox and Sanford Weisberg. *car: Companion to Applied Regression*, 2019a. URL `https://CRAN.R-project.org/package=car`. R package version 3.5.0.

John Fox and Sanford Weisberg. *carData: Companion to Applied Regression Data Sets*, 2019b. URL `https://CRAN.R-project.org/package=carData`. R package version 3.5.0.

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

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *Regression: Models, Methods und Applications*. Springer New York, 2013. ISBN 978-1-461-47137-0.

Kaggle. *Kaggle*, 2022. URL `https://www.kaggle.com/`. (visited on 15th June 2022).

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4.

Appendix

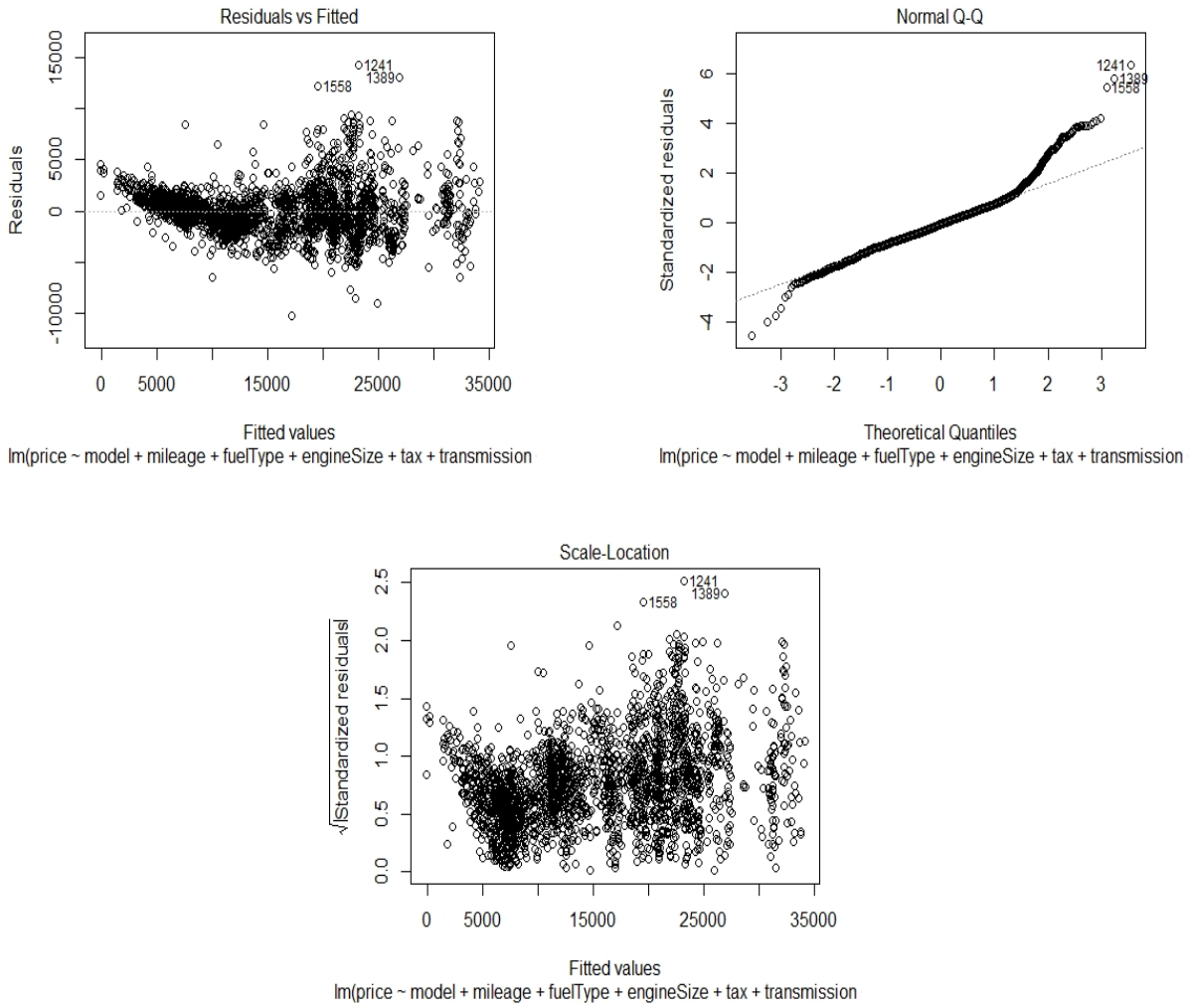

Figure 1: Plots of the fitted linear model using *price*

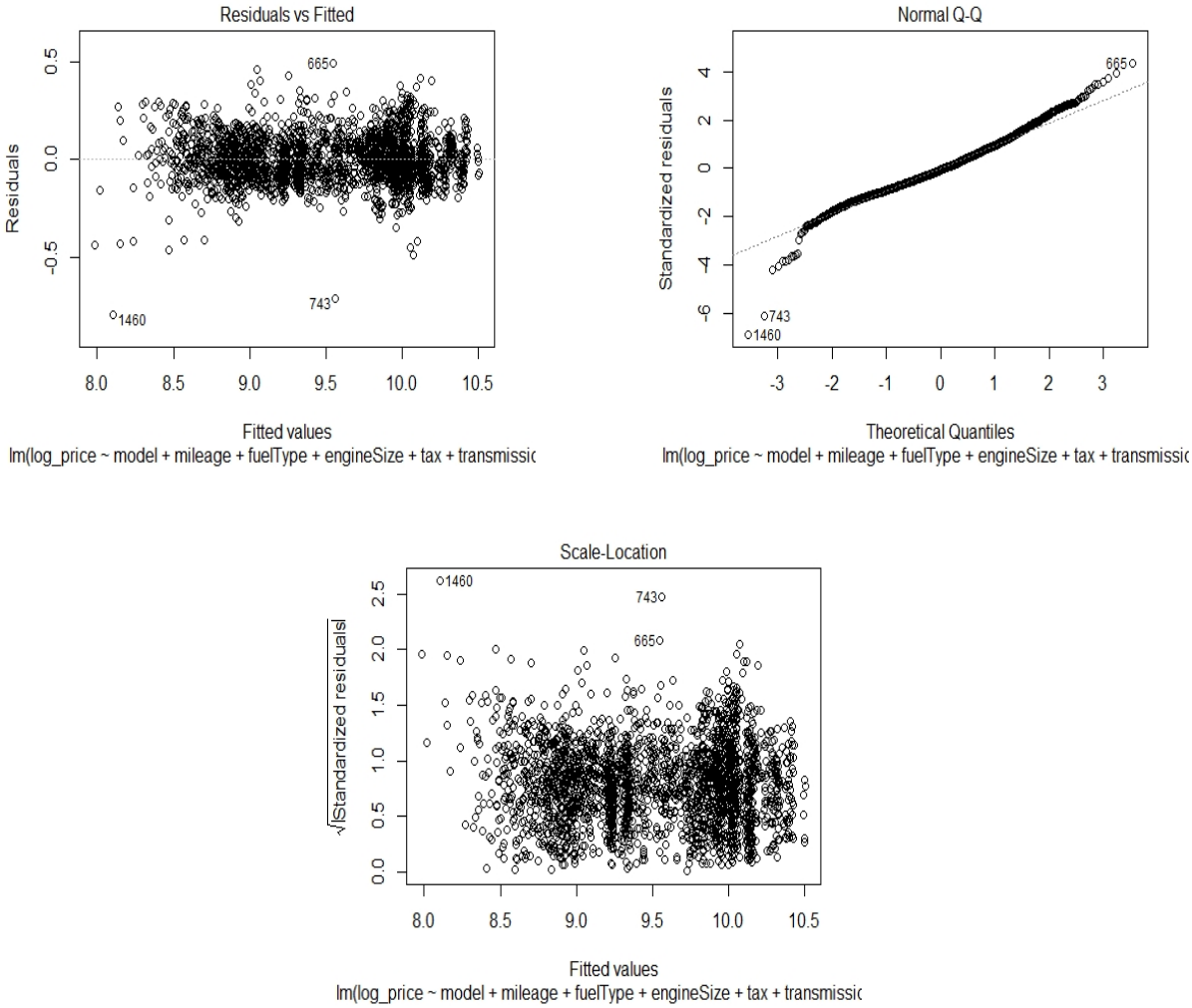

Figure 2: Plots of the fitted linear model using *log_price*

model	transmission	mileage	fuelType	tax
Passat:915	Automatic: 238	Min. : 1	Diesel: 970	Min. : 0.0
T-Roc :733	Manual :1821	Median : 12095	Hybrid: 58	Median :145.0
Up :884	Semi-Auto: 473	Mean : 21021	Other : 16	Mean :105.3
		Max. :176000	Petrol:1488	Max. :265.0

price	engineSize	lp100km	age	log_price
Min. : 1495	Min. :0.000	Min. :1.702	Min. : 0.000	Min. : 7.310
Median :13986	Median :1.500	Median :5.202	Median : 2.000	Median : 9.546
Mean :15445	Mean :1.466	Mean :5.253	Mean : 2.429	Mean : 9.504
Max. :40999	Max. :2.000	Max. :8.692	Max. :14.000	Max. :10.621

Table 1: Measures of central tendency for both categorical and continuous variables

	VIF
model	6.082
mileage	2.846
fuelType	5.241
engineSize	5.534
tax	2.423
transmission	1.745
lp100km	3.252
age	3.224

Table 2: Multicollinearity check using VIF for independent variables

	Predictors	AIC
1	model	1677.45
2	model age	-1315.05
3	model mileage age	-2307.01
4	model transmission mileage age	-2989.27
5	model transmission mileage fuelType age	-3278.91
6	model transmission mileage fuelType engineSize age	-3561.04
7	model transmission mileage fuelType engineSize lp100km age	-3620.60
8	model transmission mileage fuelType tax engineSize lp100km age	-3664.49

Table 3: Selected predictors with their AIC values

	Estimate	Std. Error	t-value	p-value
(Intercept)	9.653	0.003	318.687	2e-16
model T-Roc	0.112	0.001	14.858	2e-16
model Up	-0.568	0.001	-53.564	2e-16
mileage	0.000	0.000	-36.357	2e-16
fuelTypeHybrid	0.435	0.002	24.359	2e-16
fuelTypeOther	0.072	0.003	2.368	0.018
fuelTypePetrol	0.076	0.001	7.751	1.31e-14
engineSize	0.177	0.001	13.905	2e-16
tax	0.000	0.000	-6.788	1.42e-11
transmissionManual	-0.120	0.001	-12.825	2e-16
transmissionSemi-Auto	0.000	0.001	-0.021	0.983
lp100km	0.034	0.000	8.967	2e-16
age	-0.093	0.000	-44.746	2e-16
Residual standard error: 0.117 on 2519 degrees of freedom Coefficient of determination (R^2): 0.954				

Table 4: Summary of regression coefficients

	2.5%	97.5%
(Intercept)	9.594	9.713
model T-Roc	0.097	0.126
model Up	-0.589	-0.548
mileage	0.000	0.000
fuelTypeHybrid	0.400	0.470
fuelTypeOther	0.012	0.131
fuelTypePetrol	0.057	0.095
engineSize	0.152	0.202
tax	-0.001	0.000
transmissionManual	-0.138	-0.102
transmissionSemi-Auto	-0.019	0.018
lp100km	0.027	0.041
age	-0.097	-0.089

Table 5: Confidence Interval for regression coefficients