TU Dortmund

Introductory Case Studies

Project 3: Linear regression

Lecturers:

Dr. rer. nat. Maximilian Wechsung

Author: Ankita Sarkar - 230463

Group number: 22

Group members: Avisha Anilkumar Bhiryani, Janani Veeraraghavan, Kaushal Tajane, Shivam Shukla, Shubham Khochare

June 20, 2022

Contents

1 Introduction

The market for selling used cars has been quite popular for decades. Initially, the advertisement of selling the second-hand cars used to place in newspapers and other printed media. Soon the digital market place such as eBay, Facebook marketplace takes over the selling of the used cars. In this project we use a sample dataset of used car by VW advertised on the e-commerce platform Exchange and Mart in the UK in 2020. The goal of this project is to predict the car price based on numerous car features. The data is first pre-processed before fitting the model for prediction. The model assumptions are checked and the target variable is transformed to meet the model assumptions. Next, a preliminary best set of features are picked to fit the model based on the minimum value of AIC. Then, the model coefficients are estimated, checked for their statistical significance and lastly the model is evaluated based on the goodness of fit measure.

In section 2, the description of dataset, methods of data collections and data quality are described. Section 3 briefly explains the statistical methods involving the linear model, dummy coding for the categorical variables, best subset selection, AIC, statistical test, confidence interval, residual plot, goodness of fit measures and collinearity analysis as well as the software tools that are being used. Section 4 gives the detailed presentations of the results and their interpretations. The last section concludes the project and also discusses the possibilities of further analyses.

2 Problem statement

2.1 Description of the Dataset

The dataset used in this project is an extract from the original dataset available on Kaggle website. (Kaggle, 2020) The dataset contains the data about the used cars from the manufacturer Volkswagen (VW) being sold on the online platform Exchange and Mart in the United Kingdom in the year of 2020.

The sample size used in this project involves 2532 number of observations of used cars. It has three categorical variables model, transmission and fuelType. There are three types of model included namely, Passat, T-Roc and Up. There are 915 cars of Passat model, 733 of T-Roc model and 884 cars of Up model. The transmission categories are Automatic, Manual and Semi-Auto which are self-explanatory and it also includes

four types of fuel being used in the cars namely Diesel, Hybrid, Other and Petrol. The numeric variables in this dataset are year, price, mileage, tax, mpg and engineSize. The details about the used car ranges from the year of 2006 to 2020. The mileage of a car is the total number of distances (in 1000 miles) a car travels and miles per gallon (mpg) details the number of miles a car is driven using one gallon of fuel. The variable mpg where gallon is given in UK metric needs to be converted into litres per 100 kilometres (lp100km) for the purpose of this project. The engineSize is given in litres. Tax lists the amount of tax to be filled for a car annually. The age of the cars are also needed to be calculated. The dataset is checked for missing values and found to be containing no missig data.

2.2 Project objectives

The aim of this project is to predict the price (unit in pound) of the used cars in the given dataset. First, the dataset is being pre-processed for better results. The mpg is converted into lp100km and the current year is subtracted from the variable year to calculate the age of the cars. Next, three categorical variables are changed to factors. The summary for the numerical variables such as min, max, average values are listed in the corresponding tables and studied. To assess the linear dependencies between the variables for example age and mileage or age and price etc are examined using the scatter plots, correlation table and vif measurement tool. The model assumptions are checked to fit the linear regression model to the data. The logarithm transformation of price is considered as the response variable. The reason behind this transformation is listed as well. In the next step, AIC metric is being used to select the best subset combinations of the explanatory variables to be included to the model. Then the preliminary best model is regressed using the best subset. The model assumptions are again verified. After that, the coefficients of the covariates are estimated and checked using t-test for their statistical significance. Lastly, the confidence intervals are calculated for each of the regression parameters and the model is evaluated.

3 Statistical methods

In this section, various statistical methods and plots are described which are used to analyse the dataset. R programming language (R Development Core Team, 2020) version 4.1.0 is used.

3.1 Linear regression

The main aim of regression is to model the effects of a set of given variables $x_1, ..., x_k$ on a variable y. The variables $x_1, ..., x_k$ are called the covariates or explanatory variables or independent variables and y is called the response variable or dependent variable. The relationship between the response variable and the covariates are not deterministic rather random. Usually, a regression model tries to analyse the influence of the covariates on the mean value of the response variable. The linear regression model is specifically applicable if the y is the continuous variable and shows the normal distribution depending on the covariates which can be categorical or continuous. The most common class of the linear regression model is given by

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, ..., n$$

where $\beta_0 + \beta_1 x_1 + ... + \beta_k x_k$ is the linear combination of the covariates and ε_i is the the deviation from the mean value and is additive in nature. The coefficient parameters $\beta_0, \beta_1, ..., \beta_k$ are unknown and need to be estimated. β_0 is the intercept in the model. The errors are independent and identically distributed (i.i.d) with the expected value $E(\varepsilon_i) = 0$ and variance $Var(\varepsilon_i) = \sigma^2$. It is also assumed that the error is uncorrelated meaning $Cov(\varepsilon_i, \varepsilon_j) = 0 \quad \text{for} \quad i \neq j$ are normally distributed $\varepsilon_i \sim \ N(0, \sigma^2)$ (Fahrmeir et al., 2021, p. 21-24). The model can be written in the vector form[1] for the target variable and the error as below

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_0 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

and the design matrix and the coefficient matrix

[1] the vector or matrix form is denoted in bold

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_n \end{pmatrix}$$

So in the matrix form we can write $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the first column in $\boldsymbol{X}$ includes the intercepts. It is assumed that the columns of the design matrix $\boldsymbol{X}$ are independent of each other i.e no covariates x_i, for $i = 1, ..., k$ can be represented as a linear combination of other covariates and th number of observations n have to be equal to or greater than the number of coefficients k (full rank). This assumption is required to obtain the unique estimators of the regression coefficients $\boldsymbol{\beta}$. The linearity of the covariates is also assumed. (Fahrmeir et al., 2021, p. 74-76)

3.1.1 Parameter estimation

The method of least squares are used here to estimate the regression coefficients, denoted as $\hat{\boldsymbol{\beta}}$. So the estimated model is given by $\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_k x_{ik}$, for $i = 1, ..., k$. The sum of squared deviations between the true value y_i and the estimated value $\hat{y_i}$ are used to estimate the unknown regression coefficients $\hat{\boldsymbol{\beta}}$.

$$LS(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - x_i' \hat{\boldsymbol{\beta}})^2 = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon' \varepsilon \tag{1}$$

The below solution can be found by minimizing the $LS(\boldsymbol{\beta})$ by setting the first derivative of it to zero

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

The $\hat{\varepsilon_i}, i = 1, ..., k$ in equation (1), known as the residuals can be rewritten in the vector form as $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$. The residuals explains how far away the estimated value is from the true value which the covariates can not explain (Fahrmeir et al., 2021, p. 74-78). The $\hat{\boldsymbol{\beta}}$ is used to find the estimates of the mean of the response variable y

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

where $\boldsymbol{H}$ is called the hat matrix or the prediction matrix. The error variance is estimated as

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

where p is the number of coefficients and is given as $p = k + 1$. (Fahrmeir et al., 2021, p. 104- 108) The residuals are often standardized by dividing with the estimated standard deviation, thus the standardized residuals become

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

where h_{ii} is the diagonal elements of the hat matrix. (Fahrmeir et al., 2021, p. 124)

3.2 Dummy coding for the categorical covariates

To model the effects of any categorical covariate having c number of categories on the response variable, we define $(c-1)$ dummy variables. Dummy variables are covariates with new values with respect to the categories of the original covariates

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

for $i = 1, \dots, n$ (Fahrmeir et al., 2021, p. 95-97). [2]

3.3 Best subset selection

The least squared methods estimation is used to estimate the coefficients of the covariates in a linear model. We often want to have the best model out of a combination of models for better performance. This can be achieved by including different subset of covariates $k \in (0, 1, 2, \cdots, p)$ each time building the model. Then usually, the model with the minimum sum of squared residuals is selected. Software packages use various methods such as AIC or Mallow's c_p to select the best model. (Hastie et al., 2009, p. 57-60)

[2] For the purpose of this project, the factor function of R is used. Because R treats the factors as the dummy variables i.e it puts the value of 1 to the variable if true or else 0. (StackExchange, 2022)

3.3.1 Akaike Information Criterion

It is important for a model to better fit the data as well as to avoid the issue of overfitting which occurs when the model fits the data too well but does not generalize well meaning makes prediction error on new unseen data (Fahrmeir et al., 2021, p. 664). So the goodness of fit (explained later in this section) and likelihood inference [3] of the parameters of the model is modified using Akaike's Information Criterion AIC

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

where $l(\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}, \hat{\sigma}^2)$ is the maximum value of the log-likelihood (ML) using the estimated regression coefficient value of $\hat{\beta}_k$ and estimated error variance $\hat{\sigma}^2$. It is given $-2 \cdot l(\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}, \hat{\sigma}^2) = nlog(\hat{\sigma}^2) + n$. Thus AIC becomes by ignoring the constant term n

$$AIC = nlog(\hat{\sigma}^2) + 2(|k| + 1)$$

Usually the model with the smallest AIC is selected among the competing models. (Fahrmeir et al., 2021, p. 148)

3.3.2 Mallow's CP

Mallow's C_p is another model diagnostic tool and is defined as

$$C_p = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_{ik})^2}{\sigma^2} - n + 2|k|$$

where k is the number of covariates and σ^2 is the variance. (Fahrmeir et al., 2021, p. 148)

3.4 Statistical test

The hypotheses regarding the unknown regression parameter β are tested using the statistical t-test. The null hypothesis states that the values of the regression coefficients are zero and the alternative hypothesis states that the regression coefficients have some non-zero values as outlined in the next page

[3] the likelihood inference is the method of estimating the true unknown parameters of a population using the samples (Hastie et al., 2009, p. 265)

$$H_0 : \hat{\beta}_j = 0 \quad \text{against} \quad H_1 : \hat{\beta}_j \neq 0 \quad j = 1, \ldots, k.$$

The test statistic is constructed as

$$t_j = \frac{\hat{\beta}_j}{se_j},$$

where $se_j = \widehat{Var(\hat{\beta}_j)}^{1/2}$, the standard error of the estimated regression coefficient $\hat{\beta}_j$ which can be found in the diagonal elements of the covariance matrix $\widehat{cov(\hat{\beta})} = \hat{\sigma}^2(\boldsymbol{X'X})^{-1}$. The test statistic t_j is found to follow the t-distribution with $n - p$ degrees of freedom. We reject the null hypothesis at $(1 - \alpha/2)$ -quantile of the t-distribution with $(n - p)$ degrees of freedom i.e $|t_j| > t_{1-\alpha/2}(n - p)$. Here n is the total number of observations, p is the total number of coefficients and α is the significance level which is considered usually at 5%. So by rejecting the null hypothesis we confirm that at least one of the regression coefficient is not zero. (Fahrmeir et al., 2021, p. 131)

The $(1 - \alpha)$ confidence interval for the estimated regression coefficient $\hat{\beta}_j$ is given by (Fahrmeir et al., 2021, p. 136)

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

3.5 Residual plot

The residual plots are useful in detecting the issues of the heteroscedastic errors (when the variance of the error terms is not constant). The residuals $\hat{\varepsilon}_i$ are plotted against the predicted values of the response variable $\hat{y}_i$ obtained from fitting the model to the covariates $x_i, ..., x_n$ where $i = 1, ..., n$. All the covariates even which are not included in the model should also be considered during the plotting of the residual plot. In order to avoid the dependency structure appearing in the plot, the fiited values of $\hat{y}_i$ should be used over raw y_i as $\hat{\varepsilon}_i$ depends on the y_i. The residuals are standardized to avoid the heteroscedasticity problem as the standardized residuals fluctuate randomly around the line at zero indicating a constant variance, when this is not the case we conclude that the errors are heteroscedastic. (Fahrmeir et al., 2021, p. 183)

3.6 Coefficient of determination

The coefficient of determination is used as the goodness-of-fit measure of a model. It is defined as

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

The values of R^2 lie within 0 and 1 i.e $0 \leq R^2 \leq 1$. The values R^2 closer to 1 implies that the model is perfect in the sense of better fitting to data or the residual sum of squares $\sum_{i=1}^{n}\hat{\varepsilon}_i{}^2$ are closer to zero. If R^2 is closer to 0 then the residual sum of squares is relatively large and hence the values estimated by the model is quite far away from their original values indicating a poor model. Usually, a model with higher value of R^2 is chosen from a collection of models. But we need to keep in mind that the models being compared need to have the same response variable, same number of parameters and an intercept β_0. (Fahrmeir et al., 2021, p. 112-115)

The traditional R^2 gets only increased with the addition of the covariates to the model leading us to pick always the full model. The adjusted R^2 takes care of this problem by including a penalty term to the R^2 and is given by

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

where n is the total number of observations and k is the total number of covariates to the model. (Fahrmeir et al., 2021, p. 147-148)

3.7 Multicollinearity analysis

If the covariates of the model are highly correlated with each other which is called (multi)collinearity, then parameter estimation of the linear model becomes imprecise. It can be seen through the variance formula of $\hat{\beta}_j$

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{(1 - R_j^2)\sum_{i=1}^{n}(x_{ij} - \bar{x_j})^2}$$

where R_j^2 is the coefficient of determination of the covariate x_j which measures the linear dependence of it with the other covariates for $j = 1, ..., n$. When R_j^2 increases the $Var(\hat{\beta}_j)$ also increases. In extreme cases when $R_j^2 \to 1$, the $Var(\hat{\beta}_j)$ explodes to

infinity. The variance inflation factor is another diagnostic tool to measure the degree of collinearity based on the variance formula

$$VIF_j = \frac{1}{1 - R_j^2}.$$

As a general benchmark, if the value of the VIF_j exceeds 10 i.e $VIF_j > 10$, then collinearity problem is serious. We should consider dropping the dependent covariates from the model and rechecking for linear dependence of the model. (Fahrmeir et al., 2021, p. 157-158)

4 Statistical analysis

This section discusses the assumptions and the statistical tests performed on the given data set. It includes the interpretation of the results as well.

4.1 Data preparation

In the dataset, we have performed some data pre-processing to get better results. The given unit mpg has been converted to $l/(100km)$ as below [4]

$$l/(100km) = \frac{282.48}{\text{mpg}}.$$

The age of the car is calculated by subtracting the current year 2022 with the year for each car given in the dataset. The original mpg and year are replaced with the new calculates variables in the dataset.

4.2 Summary of the dataset

For the categorical variables such as the model, transmission, fuelType etc Table 3, Table 4 and Table 6 in Appendix give the details. We see that the maximum number of 915 cars are of Passat model and the model T-Roc has the lowest number of cars namely 733. We note that the costliest cars are of Passat models which cost £40999 and

[4] We get $l/(100km)$ by calculating the equation $\frac{4.54609 \cdot 100}{1.609344 \cdot mpg}$ where 1mile = 1.609344 km and 1gallon = 4.54609 litres.

the cheapest cars are the Up models cars. The price of the semi-automatic cars are the highest. For which the highest price is £40999. The cheapest cars are the manual driven cars which are also the highest number of cars being sold namely 1821. The least number of cars being sold are the automatic cars. From Table 6 (in Appendix), we observe that the maximum number of cars are driven on Petrol than any other fuel type. The hybrid type fuel driven cars cost the highest namely £40999 and the diesel driven cars costs the lowest which is £1495. All the categorical variables are changed into factors for the handling of dummy variables purposes.

Table 7 in Appendix describes the data description of the numeric variables in the dataset. From the table, we see that the recent car is 2 years old and the oldest is 16 years old. The average age of the cars in the dataset is 4 years. The cheapest cars cost £1495 and the costliest cars cost £40999. The average price for the cars is £15445.

4.3 Model assumptions check and choice of the response variable

To check the linear relationship between the numeric variables, scatter plots are used to plot the variables such as price and age, mileage and age. The scatter plots can be found in the Appendix under Figure 4. We observe some strong linear dependence between the variables. The correlation between the variables can be found in Table 8 in Appendix. It seems that the price decreases as the cars ages but the mileage seems to be positively increases with age. We still keep all the covariates despite the dependencies as the variance inflation factor in Table 1 shows that all the values for the numeric variables are below the threshold 10. We also assume that the data is independent based on the nature of data collection. The next assumption for a linear model is that the residuals have to have a constant variance.

Table 1: Summary table for the variance inflation factor

price	mileage	tax	engineSize	lp100km	age
4.8737	2.7651	2.3699	2.2127	2.5036	3.5018

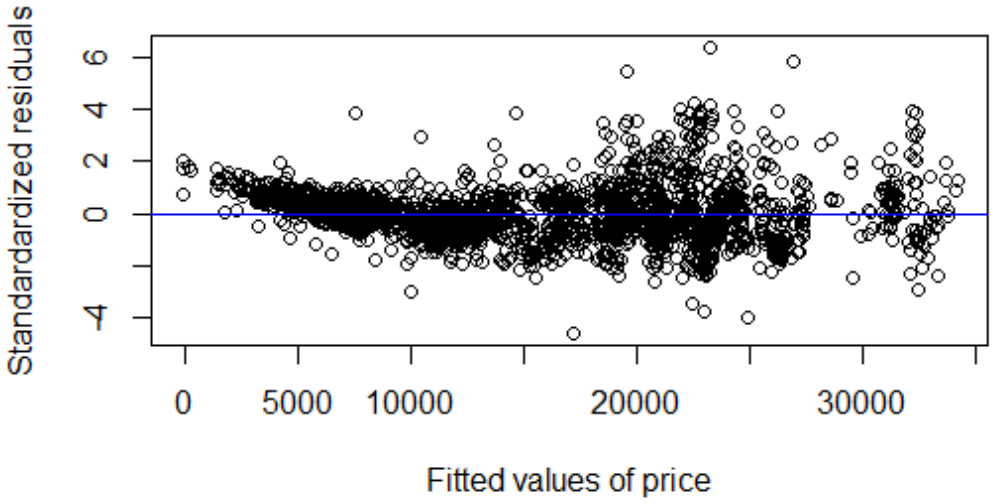


Figure 1: Residual plot for the price

To check the homoscedasticity, we check the residual plots for the original response variable price and the log transformed response variable $log(price)$ respectively after fitting the linear regression model in each case. From the Figure 1, we see that the residuals form a funnel shaped structure indicating heteroscedasticity or not a constant variance in the error terms whereas from Figure 2 we can observe a constant variance in the errors as the points are spread out almost evenly on the both side of the horizontal line. So in Figure 1, the assumption for homoscedasticity is clearly violated in case of the untransformed response variable.

The normal distribution assumption for the linear regression model can be verified using the QQ normal plot. Looking at both Figure 5 and Figure 6 in Appendix, we conclude that the data points fit the diagonal line better in Figure 6 (in Appendix) even though we observe the extreme values in both the cases. It seems that the normality assumptions are not met but due to the large size of the dataset, we can relax the normality assumptions in this project. So we see that the model assumptions are better maintained when we log transform the response variable price. Hence we choose log transformed price as the response variable and proceed with the further analysis.

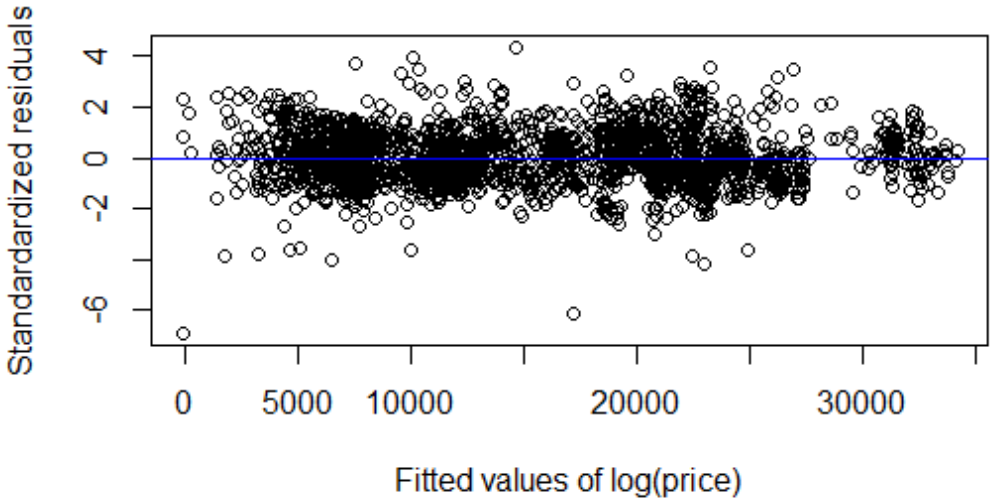


Figure 2: Residual plot for the transformed price

4.4 Selection of the best subset of the explanatory variables

We perform the best subset selection of the explanatory variables by adding different sets of covariates to the model. First only one of each of the covariates is added and next different sets of two covariates are added to the model. Thus the increasing number of covariates of different combinations are added to the model until all of the combinations are tried. Each time the AIC is calculated and recorded. Thus we have obtained total 255 models. Figure 7 in Appendix displays the 255 models under consideration. The AIC values are plotted on the y-axis and the number of covariates on the x-axis. We notice that the model with 8 covariates has the minimum AIC (marked in red). So we choose all the 8 covariates as the best subset of explanatory variables for the model to be regressed in the next subsection. The model attains -3664.49 as the AIC values.

4.5 Linear regression

In this subsection, the linear regression is performed on the best model selected from the above section. The covariates included in the model are model, transmission, mileage, fuelType, tax, engineSize, lp100km, age and the response variable is the log transformed

price variable. We can write the linear regression formula using the explanatory variables as displayed in equation (2) in Appendix.

In the above equation, $modelTRoc$, $modelUp$, $transmissionManual$, $transmissionSemiAuto$, $fuelTypeHybrid$, $fuelTypeOther$ and $fuelTypePetrol$ are dummy variables which takes the value of 1 if true or else 0. Table 9 in Appendix lists the values of $\hat{\beta}_0$ to $\hat{\beta}_{12}$.

But before assessing the estimated coefficients values of the model, it is again important to check if the selected best model approximately satisfies the necessary model assumptions. (Fahrmeir et al., 2021, p. 155). The standardized residuals of the model are calculated and plotted against the fitted values of the transformed response variable price in the residual plot as shown in Figure 3.

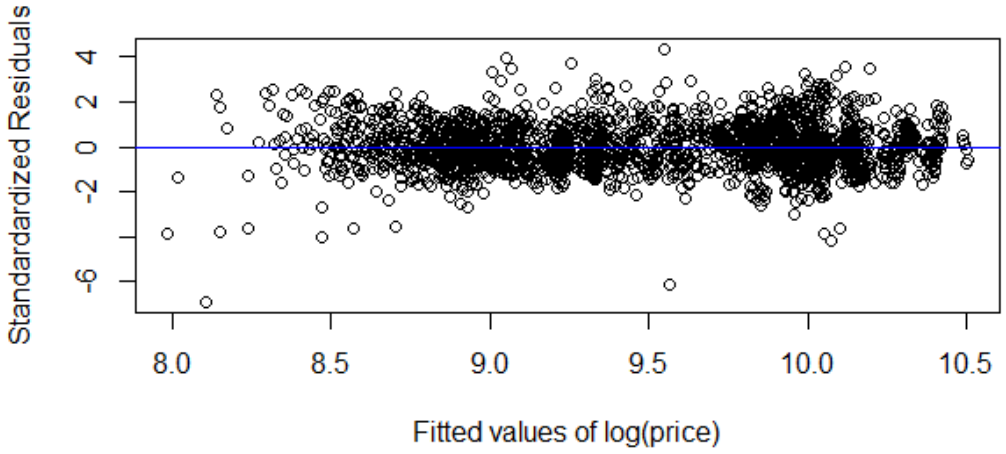


Figure 3: Residual plot for the best model

We observe that there is no clear pattern in the graph which indicates that the errors are uncorrelated. So we can be sure that the preliminary best model is not incorrectly specified. We also notice approximately even spread of data points around the horizontal line at zero on the y-axis. So we conclude that the assumptions of independence, linearity and homoscedasticity are preserved in the fitted model. If we look at the QQ normal plot in Figure 8 in Appendix we notice that the data points almost fall onto the 45° bisecting line. But there are deviations on the either end of the line. But due to the large sample size we relax the distribution assumptions of normality in this case as well.

Now the results of the estimation of the coefficients along with the confidence intervals are presented in Table 9 presented in Appendix.

From Table 9 (in Appendix), it is clear that all the estimated coefficient values significantly differ from zero at 5% level except for the covariate $transmissionSemiAuto$. We notice that the value of the intercept is 9.84 and it is significant as the p-value is very less than 0.05. But only keeping the intercept in the model doesn't add much to the interpretation of the regressors. So we move on with further analysis of the other covariates. The negative value of the coefficient of $ModelUp$ indicates that the car price is lower for $ModelUp$ as compared to that of $ModelTRoc$, the coefficient value of which is positive. The highest value of the coefficient for $fuelTypeHybrid$ among the other two types of fuelType tells us that the used car price is more which use the hybrid fuel type. The small negative coefficient values of mileage and tax describes that the little influences of these two covariates on the price of the used cars. Similarly, the engineSize and lp100km have very small positive coefficients for which the car price increases very little. Most notably, negative value for the coefficient of age indicates that as the car gets older the price decreases over the years. Finally, we get 0.9542 as the value for adjusted R^2. So the preliminary best model in this case can be concluded as a good model as the value of adjusted R^2 is closer to 1.

5 Summary

The main goal of this project was to predict the price of the used cars of VW advertised on the online platform Exchange and Mart in the UK in 2020. As people usually have to rely heavily on the current market trend to cross check the price when buying the second-hand cars, the prediction from this project will help them with the decision. The dataset involved in this project initially have the ten variables - serial number, model (Passat, T-Roc and Up), price (in pound) of the cars, year when the cars were first registered, mileage i.e the total distances (in 1000 miles) the car has been driven, mpg which is the the miles the car travels using one gallon (imperial) of fuel, fuelType (diesel, hybrid, other and petrol) in litres, engineSize in litres, tax (Vehicle Excise Duty) to be paid annually by the car and transmission, namely automatic, semi-automatic and manual which are the gearbox of the car. The data was first pre-processed to better fit the model. So the mpg was converted to lp100km and age of the cars were computed. The correlation between the numeric variables were checked using scatter

plots, computed correlation coefficients and the variance inflation factor (VIF). The variables such as age & price or lp100km & price etc exhibited correlation between them. But as the VIF for all of them is below the benchmark threshold 10, we retained all the covariates. The linear model was fit to the covariates and the model assumptions - linearity, homoscedasticity, uncorrelatedness, independence and normality were checked using residual plot and QQ normal plot. Price was log transformed as the errors showed heteroscedasticity variance.The QQ normal plot displayed skewed distribution. But we relaxed the normality assumptions due to large sample size. Afterwards, the serial number, year, mpg and price were dropped from the dataset. The different subsets of covariates were added to the model for best subset selection. The full model with all the covariates were chosen as the preliminary best model as it obtained the minimum AIC. The best model was regressed and the model assumptions were checked again. After fulfilling the model assumptions, the coefficients were estimated and checked using the t-test for their statistical significance. All the covariates except the variable transmission Semi-Auto were found to be significant at 0.05 level. We noted that the model T-Roc had positive coefficient and the model Up had negative coefficient confirming that the Up model was cheaper than than the T-Roc model. For petrol fuel type the coefficient value was positive and the highest among other two fuel types which indicated that the petrol driven cars are in general expensive. Moat notably, the negative coefficient for age indicates older cars are cheaper. The confidence interval for the covariates were calculated. The adjusted R^2 indicated a good model fit.

An alternative way to improve the prediction of the dataset could be to split the dataset into train-validation-test samples and perform the cross validation. In that way, the model generalizes well on the new unseen dataset. Thus the prediction errors could be minimized in great way. Similarly, cross validation score could be used as an alternative method to AIC for the variable selection. We ignored the influence of the outliers in this project which might impact the estimation of the model. We had also not considered the interactions of the covariates among themselves while fitting the model to the data. It would be interesting to see the prediction results without log transforming the target variable and instead using the weighted regression for example but it was out of the scope for this project. As no models are correct. So we always search for the best and try to improve the preliminary best model. (Fahrmeir et al., 2021, p. 155)

Bibliography

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. D. (2021). *Regression - models, methods and applications*. Springer.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer.

Kaggle (2020). Used car price prediction - vw. `https://www.kaggle.com/code/abhinavjhanwar/used-car-price-prediction-volkswagen-r2-score-96/data`. (Visited on 20th June 2022).

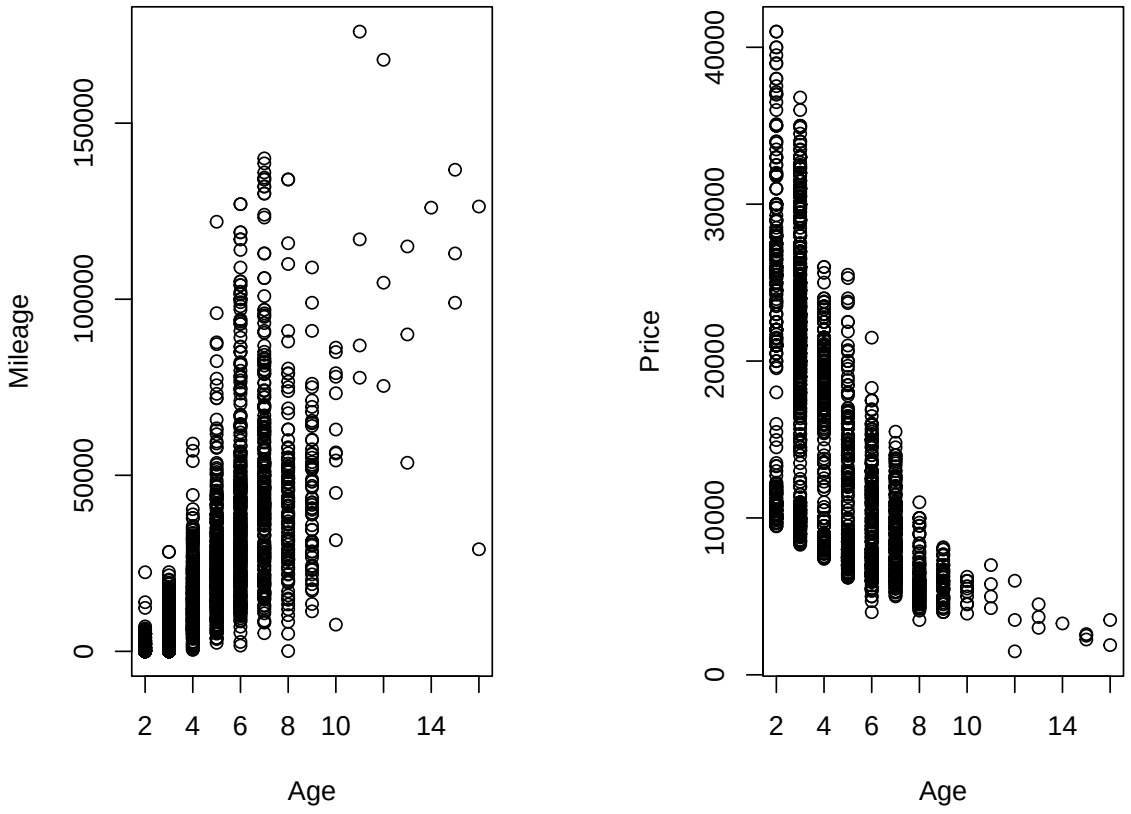
R Development Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

StackExchange (2022). Difference between the dummy variable and factors. `https://stats.stackexchange.com/questions/541924/difference-between-dummy-and-factor-variable`. (Visited on 20th June 2022).

Appendix

$$\widehat{log(price)} = \hat{\beta}_0 + \hat{\beta}_1 modelTRoc + \hat{\beta}_2 modelUp + \hat{\beta}_3 transmissionManual + \hat{\beta}_4 transmissionSemiAuto + \hat{\beta}_5 mileage + \hat{\beta}_6 fuelTypeHybrid + \hat{\beta}_7 fuelTypeOther + \hat{\beta}_8 fuelTypePetrol + \hat{\beta}_9 tax + \hat{\beta}_{10} engineSize + \hat{\beta}_{11} lp100km + \hat{\beta}_{12} age \quad (2)$$

A Additional figures



(a) Scatter plot between age and mileage (b) Scatter plot between age and price

Figure 4: Scatter plots between a) Age and Milegae and b) Age and price.

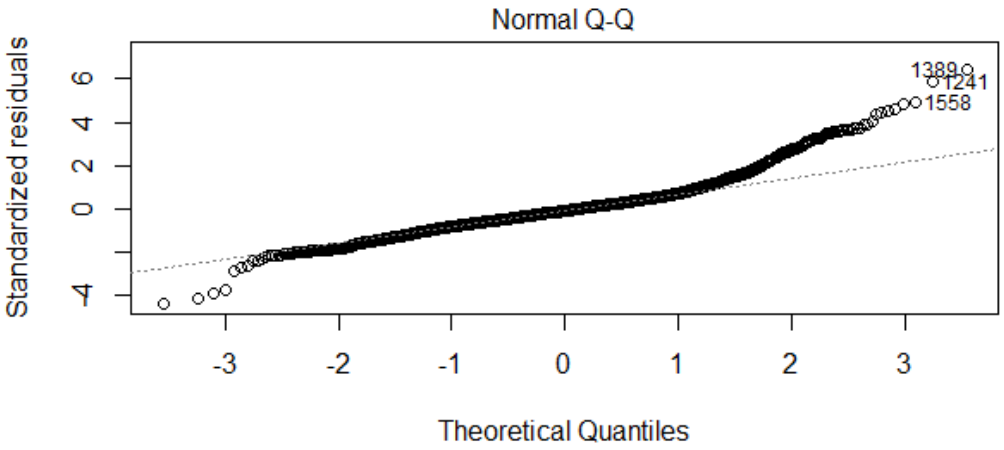


Figure 5: QQ normal plot for the price

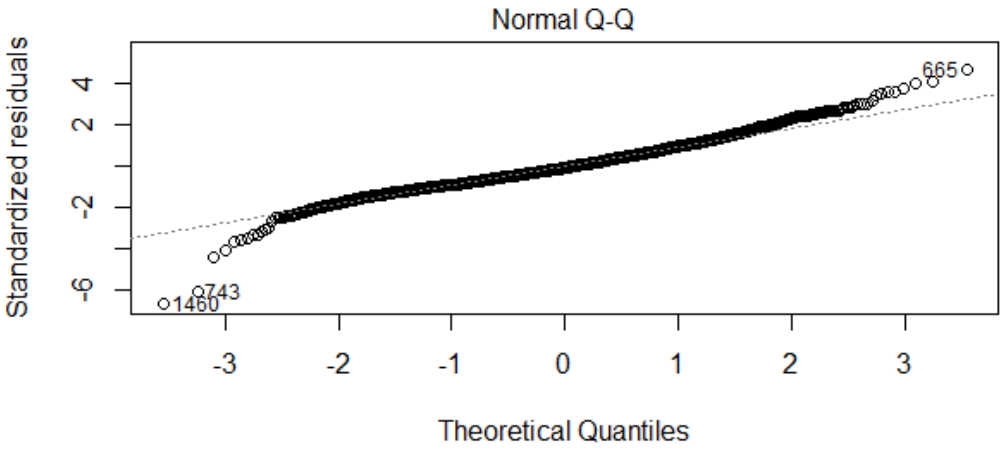


Figure 6: QQ normal plot for the transformed price

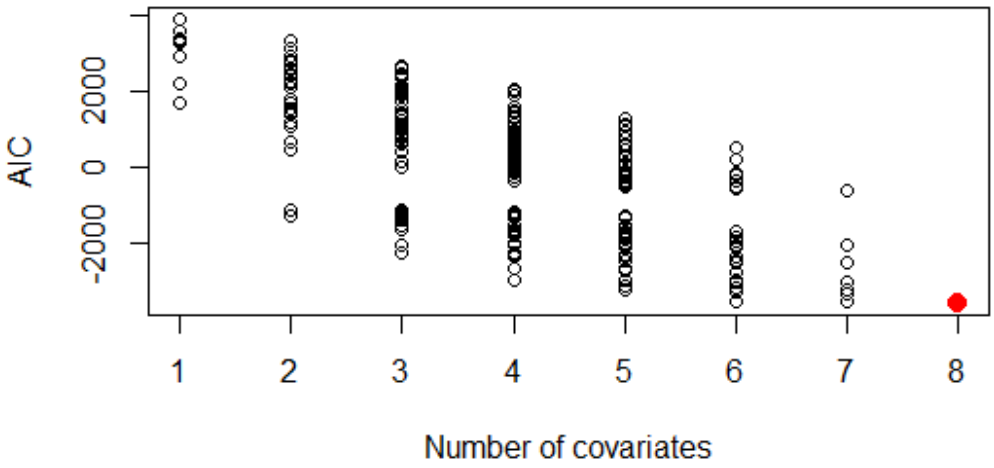


Figure 7: Residual plot for the transformed price

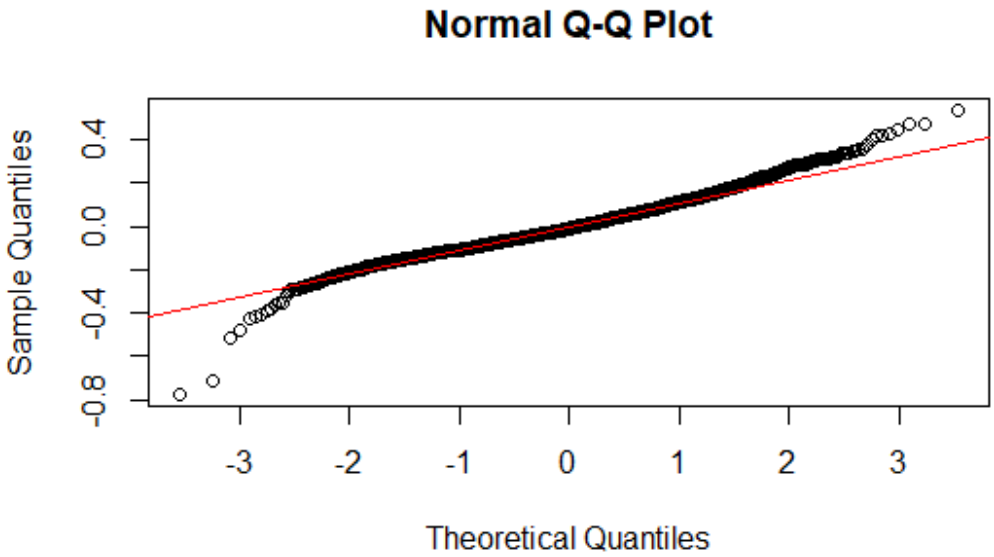


Figure 8: QQ normal plot for the best model

B Additional tables

Table 3 Summary table for the price of the models of the cars

Model	Count	Min.	Mean	Max.
Passat	915	11489	22839.39	40999
T-Roc	733	1495	16684.68	39989
Up	884	3495	8029.43	15991

Table 4 Summary table for the price of the transmission of the cars

Transmission	Count	Min.	Mean	Max.
Automatic	238	5495	22222.71	39989
Manual	1821	1495	12771.80	31895
Semi-Auto	473	6250	22324.15	40999

Table 6 Summary table for the price of the fuelType of the cars

	Count	Min.	Mean	Max.
Diesel	970	1495	16826.67	39989
Hybrid	58	3275	14015.94	40999
Other	16	6799	20380.25	32649
Petrol	1488	14498	27622.29	38000

Table 7 Summary table for the price of the numeric variables

	Min.	Mean	Max.
price	1495	15445	40999
mileage	1	21021	176000
tax	0	105.3	265
engineSize	0	1.47	2
lp100km	1.702	5.253	8.692
age	2	4	16

Table 8: Summary table for correlation between the numeric variables

	price	mileage	tax	engineSize	lp100km	age
price	1.0000	-0.4972	0.5871	0.5223	0.7380	-0.6658
mileage	-0.4972	1.0000	-0.5950	0.1782	-0.4946	0.7256
tax	0.5871	-0.5950	1.0000	0.1221	0.5779	-0.7272
engineSize	0.5223	0.1782	0.1221	1.0000	0.2295	-0.0833
lp100km	0.7380	-0.4946	0.5779	0.2295	1.0000	-0.5595
age	-0.6658	0.7256	-0.7272	-0.0833	-0.5595	1.0000

Table 9: Summary table for the estimation of coefficients

Variable	Estimate	95% Confidence interval	Std. error	t-value	p-value
(Intercept)	9.84	[9.77, 9.90]	0.03	311.57	0.00
Model T-Roc	0.11	[0.09, 0.13]	0.00	14.86	0.00
Model Up	-0.57	[-0.59, -0.55]	0.01	-53.56	0.00
transmissionManual	-0.12	[-0.14, -0.10]	0.01	-12.82	0.00
transmissionSemi-Auto	-0.00	[-0.02, 0.02]	0.01	-0.02	0.983
mileage	-0.00	[0.00, 0.00]	0.00	-36.36	0.00
fuelTypeHybrid	0.43	[0.40, 0.47]	0.02	24.36	0.00
fuelTypeOther	0.07	[0.01, 0.13]	0.03	2.37	0.018
fuelTypePetrol	0.08	[0.06, 0.09]	0.00	7.75	0.00
tax	0.00	[0.00, 0.00]	0.00	-6.79	0.00
engineSize	0.18	[0.15, 0.20]	0.01	13.90	0.00
lp100km	0.03	[0.03, 0.04]	0.00	8.97	0.00
age	-0.09	[-0.10, -0.08]	0.00	-44.74	0.00