TU Dortmund

Introductory Case Studies

# Project III: Linear Regression

Lecturers:
Dr. rer. nat. Maximilian Wechsung
M. Sc. Hendrik Dohme

Author: Supritha Palguna

Matriculation number: 229593

Group number: 11

Group members: Kunal Kochar, Dhanunjaya Elluri, Harshini Eggoni, Naveen Kumar Bhageradhi, Ashish Saini

June 20, 2022

# Contents

# 1 Introduction

According to current market estimates, the used automobile market is valued at the same level as the new car market. Because of the high price tag, most middle-class people prefer to acquire a used car than a new one, despite new features, technology, and endurance. However, evaluating the price of a used automobile can be difficult due to a variety of factors such as the car's age, the number of kilometers ridden, and the type of fuel it uses. As a result, analyzing these elements is of significant interest in order to help the average individual assess the cost of a used automobile before purchasing it.

The goal of this project is to apply linear regression to analyze data from the used cars dataset advertised on the e-commerce platform Exchange and Mart in UK in the year 2020. Initially, data preparation is accomplished to transform the data into a format that will be used later in the analysis. After preparing the data, descriptive analysis is carried out using statistical measures such as mean. By validating the assumptions provided on the linear model, a decision is made to opt between price and log-transformed price as the response variable. The best subset is then determined using the Akaike Information Criterion (AIC) statistic selection criteria. Finally, the model's coefficients and statistical significance are interpreted, and the model's goodness of fit is assessed.

Section 2 provides a summary of the dataset, its gathering process, and quality, as well as a description of the project's objective. Section 3 demonstrates the interpretation of the various statistical methods employed in this research. This includes statistical plots like the residual plot, as well as concepts like the linear regression model, Akaike Information Criterion (AIC). Section 4 presents the detailed analysis of the dataset that is performed using the methods indicated in section 3. The analysis' significant findings and takeaways are summarized in the final section.

# 2 Problem statement

The dataset and project objective are presented in detail in this section.

## 2.1 Data Set

The dataset examined in this project contains data on 2532 cars that were advertised on the e-commerce platform Exchange and Mart in the United Kingdom in 2020. Volk-

swagen (VW) models such as the Up, Passat, and T-Roc are included in this dataset. The information is obtained from the Kaggle (Goldbloom, 2010) website, which is intended solely for educational and research purposes. Professors from TU Dortmund of the 'Introductory Case Studies' course generated the dataset in the summer semester of 2022.

There are a total of nine variables in the dataset described as follows, *price* is a continuous numeric variable representing the price of the cars in GBP(£), *model* is a categorical variable which is the name of the car model (*Up, Passat*, and *T-Roc*) manufactured by Volkswagen (VW). *year*, a nominal variable, is the year when the car was first registered. *mileage* is a continuous variable, representing the total distance (in miles) for which the car has been driven. *mpg*, a continuous variable, gives the distance (measured in miles) a car can travel with one gallon (uk) of fuel. *fuelType*, a categorical variable, indicates the fuel type of the car i.e., petrol, diesel, hybrid or other. *engineSize* is a continuous numeric variable, measured in liters, which indicates the size of the car's engine. *tax*, a continuous numeric variable, is the amount of the annual tax (Vehicle Excise Duty) which to be paid for the car. *transmission* is a categorical variable with values manual, automatic and semi-automatic that tells which gearbox of the car contains. There are no null values in the dataset.

## 2.2 Project Objective

The objective of this project to fit the best linear regression model to predict the price of the used cars advertised on e-commerce platform in the UK. Initially, the data is preprocessed, the fuel consumption measurement is converted from miles per gallon (mpg) to l/(100km) (liters per 100 kilometers). Later, the age of the car is calculated using the variable *year*. Then, selection of the response variable is done between *price* and *logPrice* for linear regression using the linear model assumption. In addition, the best subset selection is done using AIC. The best model selected is fitted to the preprocessed data, after which it is analyzed. The model's coefficients are analyzed and tested for significance. The coefficients' confidence intervals are also presented, and the model's goodness of fit is assessed using the coefficient of determination.

# 3 Statistical methods

This section explains the various statistical tests, measures, and graphs used for the analysis of the dataset. The software R (R Development Core Team, 2020), version 4.0.5 is used for all the analysis and visualizations. Car (Fox and Weisberg, 2019) is the R package used.

## 3.1 Linear Regression Model

The process of developing a mathematical model or a function that can be used to predict the association between a dependent variable and one or more independent variables is known as regression analysis. The dependent or response variable is the variable that needs to be predicted. The independent or explanatory variable, often known as covariates, is the predictor. The fundamental purpose of regression is to figure out how covariates affect the response variable's mean value. The multiple linear regression model is employed when more than one independent variable is involved. The following is the equation for a linear regression model with the dependent variable $Y$ and $k$ independent variables $X_1, X_2, \ldots, X_k$,

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon,$$

where $\beta_0, \beta_1, ..., \beta_k$ are the model parameters, also known as coefficients. $\epsilon$ is the error which represents the deviation of the data points from the model's predicted values. The model can be alternatively represented by the $n$ observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$,

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

where $x_{ij}$ is the $i^{th}$ observation $(i = 1, \ldots, n)$ of $j^{th}$ covariate $(j = 1, \ldots, k)$.

A model is said to be linear if its parameters are linear, and it is said to be nonlinear if its parameters are nonlinear. The coefficient in multiple linear regression indicates how much the dependent variable is expected to increase when one of the independent variables is increased while the other independent variables remain constant (Black, 2006, p. 469), (Heumann et al., 2016, p. 251).

### Parameter Estimation

The regression parameters stated above must be calculated in order to perform the regression analysis. One way for estimating the coefficients is the least square approach. The coefficients are calculated by minimizing the sum of squared deviations between the actual and predicted values of the dependent variable using this method. The regression equation can be written in the matrix form as $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, where,

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_3 \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_3 \end{pmatrix},$$

$n$ represents the number of observations and $k$ represents the number of independent variables. The matrix $\boldsymbol{X}$ is the design matrix which includes a column of 1's that represents the intercept term and other columns that represents explanatory variables.

By taking the first derivative and equating it to zero, the squared deviations are minimized. $\boldsymbol{(X'X)^{-1}X'y}$ is the least square estimator of the coefficient $\hat{\boldsymbol{\beta}}$ in matrix form, where $\boldsymbol{X'}$ is the transpose of the design matrix $\boldsymbol{X}$. $\boldsymbol{\hat{y} = (X'X)^{-1}X'y = Hy}$ and $\boldsymbol{\hat{\epsilon} = y - \hat{y} = y - Hy = (I - H)y}$ are the least square estimates of the conditional mean of $\boldsymbol{y}$ and residuals $\boldsymbol{\epsilon}$, respectively. The hat or the prediction matrix is the matrix $\boldsymbol{H}$. The standardized residual is calculated by dividing the residual estimate $\epsilon_i$ by the residuals' estimated standard deviation:

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

where $h_{ii}$ is the $i^{th}$ diagonal element of the matrix $\boldsymbol{H}$ (Heumann et al., 2016, p. 251) (Fahrmeir et al., 2013, p. 105 & 108).

### Model Assumptions

The following are the assumptions of the regression model:

- Linearity - The response value and the independent variables have a linear relationship.
- Independence - The observations are independent from one another.

- Normality - The residuals follow a normal distribution.
- Homoscedasticity - The variances of the residuals are constant. (Black, 2006, p. 479).

### Hypothesis Testing

The null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$ and alternative hypotheses $H_1$: one or more regression coefficients are $\neq 0$, can be used to examine the overall significance of a model. If the null hypothesis is not rejected, it means that the regression model does not predict the dependent variable with any meaningful accuracy. If the null hypothesis is rejected, it means that at least one of the independent variables contributes significantly to the dependent variable's predictability. Setting a significance level $\alpha$ and utilizing the p-value produced from the F-statistic as explained in Project 2 can be used to test the hypotheses. The following formula is used to compute the F value,

$$F = \frac{n-p}{k} \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i \hat{\epsilon}_i{}^2},$$

where $n$ is the number of observations, $k$ is the number of independent variables, $p$ represents the number of parameters, $\hat{y}_i$ is the predicted value for the $i^{th}$ observation, $\bar{y}$ is the mean value of the dependent variable, $\hat{\epsilon}_i$ is the estimate of the error term for the $i^{th}$ observation. The F-statistic follows a F-distribution with $k$ and $(n-p)$ degrees of freedom.

To evaluate the statistical significance of the individual regression coefficients, the p-value obtained from the t-statistic can be used to do hypothesis testing. $H_1 : \beta_j = 0$ is the null hypothesis, while $H_1 : \beta_j = 0$ is the alternative hypothesis where $j = 1, \ldots, k$. The t-statistic is calculated as,

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

where $se_j = \widehat{Var(\hat{\beta}_j)}^{1/2}$ signifies the $\hat{\beta}_j$'s estimated standard deviation or standard error. With $n-p$ degrees of freedom, the t-value is t-distributed. When the population standard deviation is unknown and the data come from a normally distributed population, the t-distribution characterizes the standardized distances between sample means and the population mean. When the population standard deviation is unknown,

the t-distribution is employed instead of the normal distribution in inferential statistics (Heumann et al., 2016, p. 271-272) (Fahrmeir et al., 2013, p. 130-132).

A confidence interval is an estimate of a parameter of a population determined using a sample selected from the population over an interval (i.e., a range of values). If a large number of samples are obtained from the population and a confidence interval is computed for each one at the confidence level $(1-\alpha)$, then $100*(1-\alpha)\%$ of the confidence intervals will include the true population parameter. The level of confidence used in this project is $95\%((1-\alpha)\%)$. If the sample is collected repeatedly, and all model assumptions are true, the confidence intervals will overlap with the true value 95% of the time.

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

is the confidence interval for $\beta_j$ with level $(1-\alpha)$. Here, $p = k+1$ is the number of parameters and $t_{n-p}(1-\alpha/2)$ is the critical t-value obtained from t-table with significance level $\alpha$ and $n-p$ degrees of freedom (Black, 2006, p. 379) (Fahrmeir et al., 2013, p. 136).

### Dummy Variable Coding

For categorical independent variables, dummy variable coding is used. $c-1$ dummy variables are constructed for a categorical independent variable $X$ with $c$ categories as,

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

with $i = 1, \ldots, n$ and these are incorporated as explanatory variables in the regression model

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

One of the dummy variables is ignored for identifiability considerations, in this case the dummy variable for category c. This is referred to as the reference category. Direct comparison with the reference category is used to interpret the estimated impacts (Fahrmeir et al., 2013, p. 97).

### Goodness of Fit

The coefficient of determination, often known as $R^2$, indicates how close the data would be to the fitted model. It expresses the fraction of the dependent variable's variability that is explained by the independent variable. It is calculated by dividing the variance in the dependent variable $y$ into two parts: the sum of squares from the fitted model and the sum of squares from random data errors:

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

The closer $R^2$ is to 1 residual sum of squares, $\sum_i^n \hat{\epsilon}_i^2$ implies there is a better fit to the data. All residuals are zero and the fit to the data is excellent in the extreme case when $R^2$ is 1. The sum of squared residuals is quite big and the model fit is poor if $R^2$ is near to 0. As a result, the covariates have no explanatory power when it comes to the mean of $y$ (Heumann et al., 2016, p. 257) (Fahrmeir et al., 2013, p. 113).

## 3.2 Residual Plot

The residual plot is a technique for assessing the behavior of residuals. The residual plot is a form of graph in which the residuals for a given regression model are presented as an ordered pair with their corresponding $x$ value $(x, y - \bar{y})$. Examining the graphs can provide insight into how effectively the regression assumptions are satisfied by the specific regression model. With increasing sample sizes, residual graphs become more informative. Figure 1 shows a residual plot from a regression analysis that fits the assumptions—a healthy residual graph. The plot is fairly linear, the variances of the errors are roughly identical for each value of $x$, and the error terms do not appear to be connected to neighboring terms (Black, 2006, p. 480).

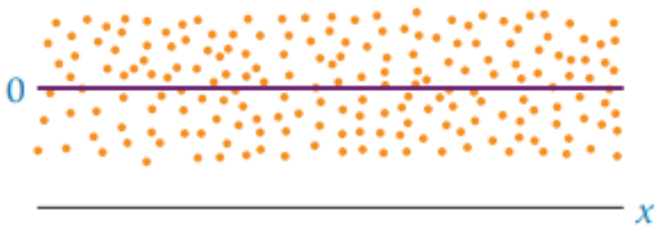



Figure 1: A healthy residual graph.

## 3.3 Collinearity Analysis

Collinearity occurs when two or more independent variables are correlated; multicollinearity occurs when three or more independent variables are correlated. Multicollinearity increases the variance of coefficients and produces type II errors, thus detecting and correcting it is critical. Variance inflation factor (VIF) is one of the methods to control multicollinearity.

In VIF, a regression analysis is used to predict an independent variable using the other independent variables. In this scenario, the predicted independent variable becomes the dependent variable. As this procedure is repeated for each of the independent variables, evidence of multicollinearity can be found if any of the independent variables are functions of the other independent variables. A VIF may be calculated using the results of such a model to detect if the standard errors of the estimates are inflated,

$$VIF = \frac{1}{1 - R^2}$$

where $R^2$ is the coefficient of determination, which is used to predict an independent variable using the other independent variables. VIF has a general benchmark of 10, which indicates that if the VIF value is larger than 10 then significant collinearity exists (Black, 2006, p. 576-578) (Fahrmeir et al., 2013, p. 158).

## 3.4 Information Criteria

The Best subset selection approach is utilized in this project, and it aims to determine the subset of independent variables that best predicts the outcome by taking into account all potential combinations of independent variables.

### Akaike Information Criterion

For the best model selection, the Akaike Information Criterion (AIC) is employed. The AIC penalizes models based on their number of parameters, decreasing data overfitting. When compared to other AIC values of models from the same dataset for the same dependent variable, a model with a lower AIC value is regarded to be preferable. The number of parameters $k + 1$ in a model and the maximum value of the model's log-likelihood function are used to determine the AIC value. Considering the maximum

likelihood estimator of variance $\hat{\sigma}^2 = \hat{\epsilon}'\hat{\epsilon}/n$, the AIC for a linear model with Gaussian errors is $n \cdot \log(\hat{\sigma}^2) + 2(k+1)$ (Fahrmeir et al., 2013, p. 148).

# 4 Statistical analysis

The statistical analysis of dataset is conducted in this section using the statistical methods outlined in Section 3.

## 4.1 Data Preparation

The original dataset given comprises 2532 observations and 9 variables, as mentioned in Section 2.1. The data preparation is covered in this section. For better analysis, the fuel consumption measurement is first transformed from miles per gallon (mpg) to liters per 100 kilometers (lp100) using the formula,

$$lp100 = \frac{282.48}{mpg}.$$

Later, the age of the cars is calculated by subtracting the year from 2020 (the year the dataset was created) and saved in the variable *age*. The variables *mpg* and *year* are removed after this. The *price* variable is transformed into a log format and saved in a variable named *logPrice*.

## 4.2 Descriptive Analysis

Table 1: Summary of continuous variables.

| Variable | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|
| price | 1495 | 8495 | 13986 | 15445 | 21422 | 40999 |
| mileage | 1 | 3803 | 12095 | 21021 | 29052 | 176000 |
| lp100 | 1.702 | 4.400 | 5.202 | 5.253 | 5.695 | 8.692 |
| age | 0 | 1 | 2 | 2.43 | 4 | 14 |

The univariate analysis of the variables is outlined in this section. The summaries of all the variables can be seen in Tables 1 and 2. It can be observed that the lowest *price* is £1,495, which is for a Passat model car that was originally registered in 2010. The maximum *price* £40,999 is of a T-Roc model car registered in 2020. A car sold on an

Table 2: Summary of categorical variables.

| | Variable | count | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|---|---|
| Model | T-Roc | 773 | 11489 | 19950 | 21990 | 22839.39 | 24590 | 40999 |
| | Passat | 915 | 1495 | 10989 | 14999 | 16684.68 | 20998.5 | 39989 |
| | Up | 884 | 3495 | 6495 | 7699 | 8029.43 | 9699.25 | 15991 |
| Fuel type | Diesel | 970 | 1495 | 11222.5 | 16495 | 16826.67 | 21499.5 | 39989 |
| | Petrol | 1488 | 3275 | 7400 | 10200 | 14015.94 | 19999.25 | 40999 |
| | Other | 16 | 6799 | 16896 | 21294.5 | 20380.25 | 22914.25 | 32649 |
| | Hybrid | 58 | 14498 | 23152.75 | 28995.5 | 27622.29 | 31999.5 | 38000 |
| Transmission | Automatic | 238 | 5495 | 15067.25 | 23075 | 22222.7 | 29771 | 39989 |
| | Manual | 1821 | 1495 | 7499 | 10299 | 12771.79 | 18950 | 31895 |
| | Semi-Auto | 473 | 6250 | 16795 | 22495 | 22324.15 | 26950 | 40999 |

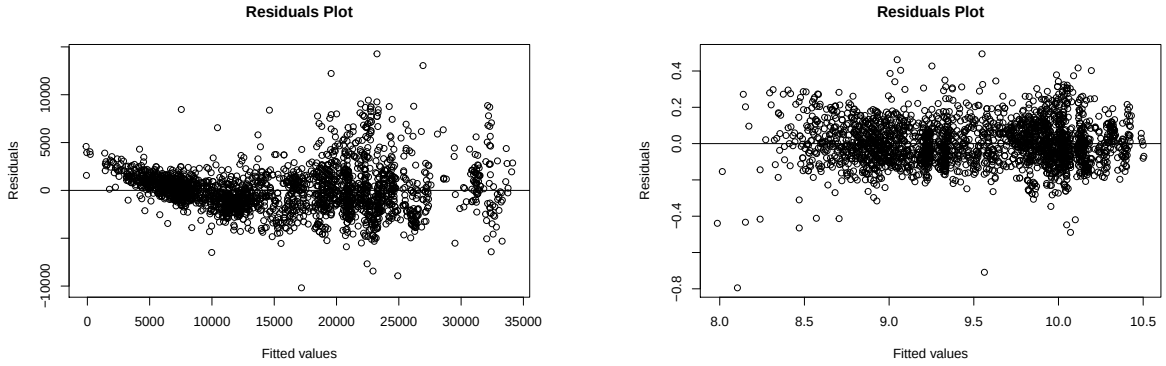
used car platform costs around £15,445 on an average. The minimum *mileage* of a car is one mile, whereas maximum *mileage* is 176,000 miles. An average *mileage* is about 21,021 miles. A car's lowest fuel consumption is 1.702 liters per 100 kilometers, while highest fuel consumption is 8.69 liters per 100 kilometers.

There are 773 *T-Roc* models, 915 *Passat* models and 884 *Up* different car models. *T-Roc* models are more expensive than *Passat* and *Up* models, with an average price of £22,839.39. Among the 2532 cars in the dataset, 970 have *Diesel*, 1488 have *Petrol*, 16 have *Other* and 15 have *Hybrid* fuel consumption type. On average, the price of car with *Hybrid* fuel type is higher than *Diesel*, *Petrol* and *Other* fuel consumption types. There are 473 cars with a *Semi-Auto* transmission, 238 with an *Automatic* transmission, and 1821 with a *Manual* transmission. The most expensive gearbox is the *Semi-Auto*, which costs £39,989 pounds, while the least expensive gearbox is the *Manual*, which costs £1495. *Automatic* is less expensive than *Semi-Auto* and *Manual*, with an average cost of £22,222.

## 4.3 Response variable selection

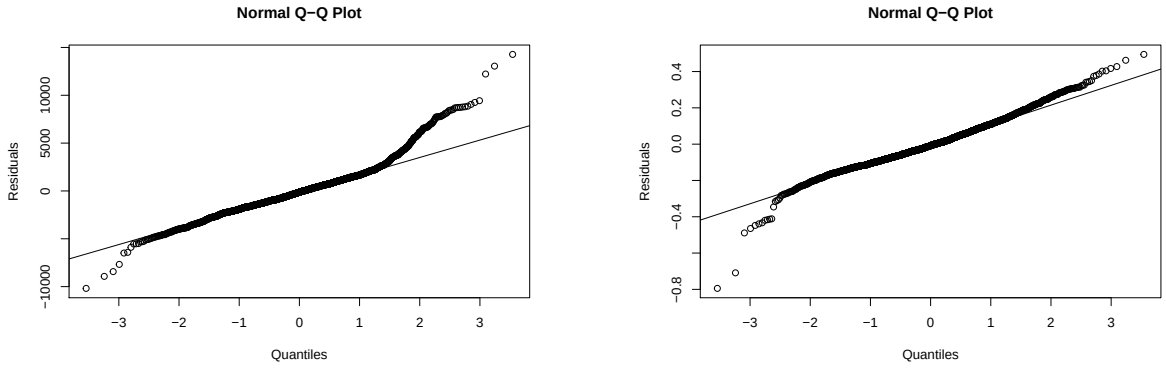
The aim of this subsection is to choose the response variable for linear regression between *price* and *logPrice*. Two linear regression models are fit with different response variables. The response variable for one model is *logPrice*, whereas the response variable for the other model is *logPrice*. For both models, the explanatory variables are the same. Figure 2 and Figure 3 show the residual plots and QQ plot for both models.

Many points in Figure 3(a) are far away from the straight line, however the points in Figure 3(b) are much closer to the straight line. This means that the linearity and homoscedasticity requirements are not violated for the model with *logPrice* as the response variable, but violated for the model with *price*. The points in Figure 3(a) are a little off the straight line towards the end, and there are a lot of outliers, but in Figure 3(b) the points are roughly in the straight line. Although there are a few outliers in the plots for the higher and lower quantiles, the points in general are near to the straight line, indicating that the residuals are close to the theoretical normal distribution. As a result, the normality assumption does not appear to be violated. Hence, *logPrice* can be used as the linear regression response variable.




(a) For *price* as the response variable. (b) For *logPrice* as response variable.

Figure 2: Residuals plots for models with *price* and *logPrice* as the response variable.




(a) For *price* as the response variable. (b) For *logPrice* as response variable.

Figure 3: QQ plots for models with *price* and *logPrice* as the response variable.

## 4.4 Best Subset Selection

The best models from all feasible subsets of independent variables determined by the Akaike Information Criterion (AIC) information criterion are discussed in this part. *model*, *mileage*, *fuelType*, *engineSize*, *transmission*, *lp100*, *tax* and *age* are the best subset of independent variables according to AIC, with an AIC value of -3664.49 when compared to all other subsets of independent variables.

## 4.5 Linear Regression Model

Table 3: Output of the Multiple Linear Regression Model.

| | Estimate | Std. Error | t value | Pr(>\|t\|) | 2.5 % | 97.5 % |
|---|---|---|---|---|---|---|
| (Intercept) | 9.65 | 3.03e-02 | 318.69 | < 2e-16 | 9.59 | 9.71 |
| model T-Roc | 1.12e-01 | 7.52e-03 | 14.86 | < 2e-16 | 9.69e-02 | 1.26e-01 |
| model Up | -5.68e-01 | 1.06e-02 | -53.56 | < 2e-16 | -5.89e-01 | -5.47e-01 |
| transmissionManual | -1.19e-01 | 9.35e-03 | -12.83 | < 2e-16 | -1.38e-01 | -1.01e-01 |
| transmissionSemi-Auto | -1.97e-04 | 9.41e-03 | -0.02 | 0.983 | -1.86e-02 | 1.82e-02 |
| mileage | -5.71e-06 | 1.57e-07 | -36.36 | < 2e-16 | -6.02e-06 | -5.40e-06 |
| fuelTypeHybrid | 4.35e-01 | 1.784e-02 | 24.36 | < 2e-16 | 3.99e-01 | 4.69e-01 |
| fuelTypeOther | 7.18e-02 | 3.04e-02 | 2.37 | 0.018 | 1.23e-02 | 1.31e-01 |
| fuelTypePetrol | 7.62e-02 | 9.83e-03 | 7.75 | 1.31e-14 | 5.69e-02 | 9.55e-02 |
| tax | -4.18e-04 | 6.15e-05 | -6.79 | 1.42e-11 | -5.38e-04 | -2.96e-04 |
| engineSize | 1.77e-01 | 1.28e-02 | 13.91 | < 2e-16 | 1.52e-01 | 2.02e-01 |
| lp100 | 3.39e-02 | 3.79e-03 | 8.97 | < 2e-16 | 2.65e-02 | 4.14e-02 |
| age | -9.32e-02 | 2.08e-03 | -44.75 | < 2e-16 | -9.73e-02 | -8.91e-02 |

This section performs the analysis of the linear regression model. The output of the regression analysis is presented in the Table 3. The coefficient estimates appear in the first column, the standard error of the estimates appears in the second column, and the t-value and p-value appear in the third and fourth columns, respectively. The 95% confidence interval for the coefficients is represented by the last two columns.

The intercept is 9.65, which indicates that if all of the variables are assumed to be zero, the expected value of the dependent variable *logPrice* is 9.65. When compared to the model Passat, the average value of *logPrice* for the model T-Roc is 0.11 higher and for the model Up it is 0.56 lesser, holding all other variables being constant. The average value of *logPrice* for cars with manual transmission and semi-auto transmission respectively lowers by 0.12 and 0.0002 when compared to cars with auto transmission. By keeping

all other variables constant, the average value of *logPrice* decreases by -5.71e-06 for the variable *mileage*, implying that the greater the distance a car has been driven, the lower is its price. The average value of *logPrice* increases by 0.43, 0.007 and 0.008 for cars using Hybrid, Other and Petrol fuel types, respectively. It can be interpreted that the average value of *logPrice* decreases by -0.0004 for the variable *tax* by keeping all other variables constant. The average value of *logPrice* increases by 0.17 as the engine size increases. When the fuel consumption of a car is increased, the average value of *logPrice* increases by 0.034. It can be inferred that as the age of the cars increases, the average value of *logPrice* lowers by 0.09.

The null and alternative hypotheses is used to test the overall model's statistical significance. The value of $\alpha$ is set to 0.05. The F-statistic value for degrees of freedom 12 and 2519 is 4395, and the associated p-value is 2.2e-16, which is less than 0.05, therefore the null hypothesis may be rejected. The overall model is statistically significant, showing that at least one of the independent variables contributes significantly to the mean value of the dependent variable *logPrice* predictability.

Individual regression coefficients are subjected to a hypothesis test using the p-value. The significance level is set to 0.05. The p-value for the variable *transmissionSemi-Auto* is larger than $\alpha$ value (0.05) in Table 3, indicating that the null hypothesis is rejected. The p-value for all of the other variables, on the other hand, is less than 0.05, indicating that changes in these independent variables have a statistical significance on the mean value of the dependent variable *logPrice*.

The coefficient estimates of variable *transmissionSemi-Auto*'s confidence interval has a value of 0, indicating that the coefficient estimate can be 0 and the null hypothesis cannot be rejected. It can be seen that the positive confidence interval variables *modelT-Roc*, *fuelTypeHybrid*, *fuelTypeOther*, *fuelTypePetrol*, *lp100* and *engineSize* increase the average value of *logPrice*, whereas the negative confidence interval variables *modelUp*, *transmissionManual*, *mileage*, *tax*, and *age* decrease the average value of *logPrice*.

The coefficient of determination $R^2$ is 0.95, indicating that model can explain 95% of the variation in the response variable around its mean. The obtained results are based on the homoscedasticity assumption, which is addressed in the next subsection.

## 4.6 Verification of Model Assumptions

The model assumptions provided in Subsection 3.1 are evaluated using residual and QQ plots in this subsection. A QQ plot is created for the normality test, as illustrated in Figure 3 in the Appendix on page 17. The theoretical quantiles of the normal distribution are represented on the x-axis, while the quantiles of the standardized residuals are represented on the y-axis. The points are roughly aligned in a straight line, as can be seen. Although there are a few outliers in the plots for the higher and lower quantiles, the points in general are near to the straight line, indicating that the residuals are close to the theoretical normal distribution. As a result, the normality assumption does not appear to be violated.

The residual versus fitted plot of the model is depicted in Figure 4 in Appendix on page 18. The points are randomly spread around the straight line and do not follow any discernible pattern. It is also worth noting that the distribution of the residuals is almost the same across all of the anticipated values. This means that the assumptions of linearity and homoscedasticity are not violated. As a result, the observations can be considered to be independent of one another. Table 3 in Appendix on page 17 shows that the VIF values of the independent variables are less than 10, implying that there is no collinearity between the variables and that none of the variables needs be removed.

# 5 Summary

The goal of this project is to do regression analysis on a dataset containing data on used cars advertised on an e-commerce platform in the United Kingdom in 2020. Volkswagen (VW) models such as the Up, Passat, and T-Roc are included in this dataset. The data is solely available for educational/research purposes on the website Kaggle (Goldbloom, 2010). Professors of 'Introductory Case Studies' course at TU Dortmund have compiled the dataset. There are 2532 observations and 9 variables in the dataset. Data preparation is done at first by transforming few variables that were later utilized for analysis. Following that, the best subset selection method AIC selection criteria was used to determine the best models from all feasible subsets of independent variables. Finally, the coefficients were evaluated, significance tests were conducted, and the 95% confidence intervals and goodness of fit were discussed for the model chosen based on the best subset.

During data preparation, the *price* was log transformed and saved in *logPrice*, which was then utilized as a response variable. The metric for fuel consumption was changed from miles per gallon (mpg) to liters per 100 kilometers (l/100km). The age of the cars was derived by subtracting *year* from 2020.

Following that, using the linear model assumption, log-transformed price was chosen as the response variable for linear regression. The best subset selection was then done using AIC, with the lowest AIC value of -3664.49. The variables picked by AIC were *model*, *mileage*, *fuelType*, *engineSize*, *transmission*, *lp100*, *tax* and *age*. Then, using the best subset determined by AIC, a linear regression model was fitted to the dataset. It was found that *modelT-Roc*, *fuelTypeHybrid*, *fuelTypeOther*, *fuelTypePetrol*, *engineSize*, *modelUp*, *transmissionManual*, *mileage*, *tax*, and *age* were all statistically significant during the significance test, indicating that changes in these factors will have an influence on the dependent variable *logPrice* on average. The confidence interval for the *transmissionSemi-Auto* variable had 0 within its range and hence failing to reject null hypothesis. The model's coefficient of determination was 0.95, indicating that it explains 95% of the variation in the response variable *logPrice* around its mean.

For further analysis, other variables or characteristics of a car, such as its history, cooling and heating systems, mechanical inspection, tire condition, safety features, and so on, can be included in the model to estimate a car's pricing. Cross validation model selection criteria may also be used, in which the data is split into a test set for parameter estimation and a validation set for evaluating prediction quality.

# Bibliography

Black, K. (2006), *Business statistics : for contemporary decision making.*, Hoboken, New Jersey : Wiley.

Fahrmeir, L., Kneib, T., Lang, S. and Brian, M. (2013), *Regression Models, Methods and Applications.*, Springer-Verlag Berlin Heidelberg.

Fox, J. and Weisberg, S. (2019), *An R Companion to Applied Regression*, third edn, Sage, Thousand Oaks CA.

Goldbloom, A. (2010), 'Kaggle - your machine learning and data science community.'. URL: `https://www.kaggle.com` (Visited on 17th December 2022).

Heumann, C., Schomaker, M. and Shalabh (2016), *Introduction to Statistics and Data Analysis.*

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

# Appendix

## A Additional tables

Table 4: VIF values of the variables

| Variable | VIF |
|---|---|
| model | 1.570 |
| mileage | 1.687 |
| fuelType | 1.317 |
| engineSize | 2.352 |
| age | 1.795 |
| lp100 | 1.803 |
| transmission | 1.149 |
| tax | 1.556 |

## B Additional figures

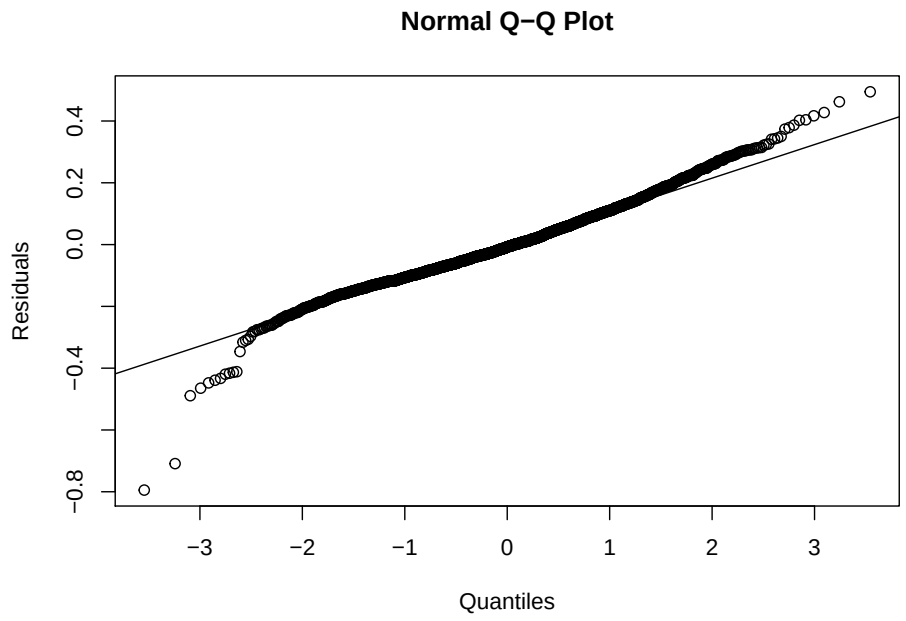



Figure 4: QQ-plot to check the normality assumption of the model.

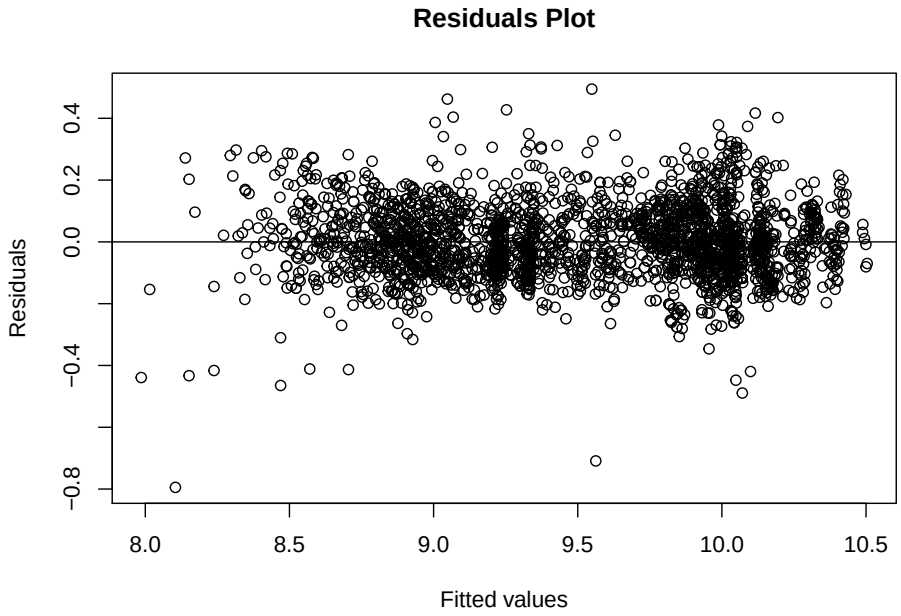



Figure 5: Residual vs Fitted Plot to check the linearity and homoscedasticity assumption of the model.