Contents

1 Introduction

The price for a used car can be estimated based on a variety of criteria such as the model of the car, the year of registration, type of transmission, mileage, type of fuel the car consumes, and other factors. In the present era, where people prefer to buy a used car rather than a new one because of its advantages (Betterton, 2022), analyzing the factors that affect the price of a used car can be useful for both buyers and sellers. Also, predicting the price of a used car can help buyers compare the price to a brand new car and make the best decision.

The aim of this report is to analyze the used car data of the UK from the year 2020 and to fit a linear regression model. First, we prepare the dataset for analysis by performing necessary transformations on the given dataset. Second, the best response variable will be determined by analyzing the different plots related to the response variable. Third, the best subset of explanatory variables will be determined by applying one of the best subset selection methods. Fourth, we build a linear regression model on the dataset based on the best subset of explanatory variables and check if the model satisfies certain assumptions of linear regression. Finally, we report the results obtained from the model.

In Section 2, a brief description of the dataset is provided. In addition, the variable definitions and the project objectives are presented in more detail. In Section 3, different statistical methods relevant to this project are elaborated on and explained. In Section 4, we apply the statistical methods explained in the previous section to the project dataset and interpret the results. The final section (Section 5) will summarize our findings and ideas for further possible analysis of the given dataset.

2 Problem statement

2.1 Description of dataset

The project makes use of data related to used cars in the UK from the year 2020, which were advertised on the e-commerce platform Exchange and Mart. The dataset is a small extract from the huge dataset available on Kaggle and it contains only the models of a particular manufacturer namely Volkswagen (VW) (Jhanwar, 2020). The dataset includes a total of 2532 observations and 9 columns. Each observation corresponds to information about a particular used car in the UK. The explanatory variable *price*

(numeric - continuous) refers to the price of a particular used car in GBP (Great British Pound). Further, the predictor variables like the *year* (numeric - continuous) refer to the particular year in which the car was first registered. The *model* refers to the name of the particular car model. The *mileage* (numeric - continuous) refers to the total distance the car has travelled in terms of miles. The *mpg* (numeric - continuous) refers to how many miles the car can run with one gallon of fuel. The *fuelType* (categorical) refers to the type of combustion fuel used to run the car. The engineSize (numeric - continuous) refers to the car's engine area in liters. *Tax* (numeric - continuous) refers to the total tax that has to be paid for a particular car annually. *Transmission* (categorical) refers to the category of the gearbox present in the particular car. (Manual, semi-auto, automatic).

The data set is a small extract from the standard webpage, it contains no missing values and also contains a sufficient number of observations. This will make the analysis a little easier and may help to get good results. Hence, the overall data quality is good.

2.2 Project objectives

The project's purpose is to build the regression model on the provided dataset. To achieve this, first, pre-processing step is performed on the dataset. Second, a brief descriptive analysis is performed on the modified dataset and the summary of the results is presented in Table 1 & 2. Third, we determine the appropriate response variable for our model by comparing the residual and Q-Q plots for the variables price and logprice (logarithmic value of price). Fourth, after choosing the response variable we perform the best subset selection using the Akaike information criteria (AIC), and the linear regression model is built using this best subset of covariates. Fifth, we check for the assumptions of linear regression. Finally, the results related to model estimates, confidence intervals, p-value, and coefficient of determination are interpreted in detail.

3 Statistical methods

The statistical methods essential to perform regression analysis on the dataset are explained in detail, along with their formulas. All the following graphs and calculations are created using the software R in version 4.0.5. (R Core Team, 2021) and the package car is used (Fox and Weisberg, 2019).

3.1 Maximum likelihood estimation

Given the random sample, the process of estimating the population parameters can be defined as the maximum likelihood estimation. It is used when the distribution is known but the parameters of the distribution are unknown.

Let $X_1, X_2..., X_n$ be a sample with n random variables. The density function of a random variable can be denoted as $f(x, \theta)$, $\theta \in \Omega$. Here Ω denotes parameter space. The likelihood function of the joint density of the sample can be represented as follows:

$$L(\theta;\ x_1, ..., x_n) = f(x_1, \theta) \cdot f(x_2, \theta) \cdot ... \cdot f(x_n, \theta)\ .$$

To find the maximum likelihood estimator of θ the log function is applied to the $L(\theta;\ x_1, ..., x_n)$. Applying the logarithmic function simplifies the likelihood term. The log is an increasing function, hence, maximizing the log-likelihood function is the same as maximizing the likelihood function. Next, taking the partial derivative with respect to θ and setting this to zero gives the maximum likelihood of the estimator $\hat{\theta}$. It is represented as follows:

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{\partial \log L(\theta;\ x_1, ..., x_n)}{\partial \theta} = 0\ ,$$

(Dodge, 2008, p. 334).

3.2 Linear regression model

The method to determine the relationship between one or more covariates and the response variable can be defined as a linear regression model. Let $(x_1, ..., x_k)$ be the set of covariates and y be the response variable, the function $f(x_1, ..., x_k)$ is used to model the relationship between them. The relationship is not deterministic because it also includes random errors ϵ. Hence, the response variable (y) is said to be a random variable. This can be represented as follows:

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

The function $f(x_1, ..., x_k)$ is a linear combination of covariates and represented as $f(x_1, ..., x_k) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$.

Where $\beta_0, \beta_1, ..., \beta_k$ are parameters that are not known and are required to be estimated. The model's intercept is represented by β_0. The combination of vector representation of covariates $\boldsymbol{x} = (1, x_1, ..., x_k)'$ and unknown parameters $\boldsymbol{\beta} = (\beta_0, ..., \beta_k)'$ is formulated as $f(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}$.

If the vectors are defined as follows:

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} ,$$

and the design matrix $\boldsymbol{X}$,

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix} ,$$

The matrix notation can be formulated as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} .$$

(Fahrmeir et al., 2013, p. 73-75).

3.2.1 Assumptions of linear regression model

To fit the linear regression model, the data should satisfy the following assumptions:

1.) There exists a linear relationship between the covariates and the response variable.

2.) The expectation of each random component or the error term should be zero, i.e., $E(\epsilon_i) = 0$. The error variance of each observation should be equal to σ^2 which is constant, i.e., $Var(e_i) = \sigma^2$. It can also be said as the error terms should satisfy the property of homoscedasticity. In addition, the assumption of error terms are not related to each other should also hold true, i.e., $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$.

3.) The exists no linear relationship between the covariates (No multi-collinearity).

4.) The error of each observation should be normally distributed, this can be represented as $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ (Fahrmeir et al., 2013, p. 75-76).

3.2.2 Estimation of regression coefficients

One of the efficient methods to estimate the regression coefficients can be the method of the least squares. In this method, the value of the sum of squared deviations is minimized with respect to regression coefficients beta, to determine the best beta values.

$$LS(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 = \sum_{i=1}^{n}\varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \ .$$

Because of its easily differentiable property, this technique is most often used for estimating the regression coefficients. However, this technique is more sensitive to outliers.

The process of minimizing the least-squares includes equating the vector of first derivatives to zero and proving the matrix of second derivatives is positive definite. By doing this, beta can be represented as follows:

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

Using the formula of estimation of regression coefficients, the formula to estimate the conditional mean of y can be derived as $\widehat{E(\boldsymbol{y})} = \hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$.

Here, the matrix term $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. It represents the hat matrix or prediction matrix (Fahrmeir et al., 2013, p. 104-107).

3.2.3 Residuals and standerdized residuals

As the $\hat{y}_i$ refers to the estimated value for a particular observation, the residual for an observation can be defined as the difference between the true value y_i and the estimated value $\hat{y}_i$. It can be represented as $\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$ (Fahrmeir et al., 2013, p. 77).

To obtain standardized residuals, the residuals are divided by the estimated standard deviation of residuals. It can be calculated as follows:

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

Here r_i represents the standardized residual of observation i and h_{ii} are the ith diagonal element of the hat matrix (Fahrmeir et al., 2013, p. 124).

3.2.4 Coefficient of determination

The percentage of the total variance in the response variable that can be explained by the predictor variables can be defined as the coefficient of determination. It can also be used as a goodness-of-fit measure and it is nearly associated with the empirical correlation coefficient. It can be represented as follows:

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

The values of R^2 range from 0 to 1. If the value is close to 1, it denotes the smaller value of the addition of the squares of the residuals, and the model fits the data better. If the R^2 value is close to 0, it denotes the model is not a better fit for the data (Fahrmeir et al., 2013, p. 112).

3.2.5 Variance inflation factor (VIF)

Collinearity can be defined as the close association of two or more covariates with each other (the increase or decrease effect appear jointly). The situation of collinearity can cause a problem in estimating the values for the regression parameters because it is hard to determine the effects of these variables on the response variable individually. To maintain the proper estimates for the regression coefficients, the effect of collinearity should be avoided and hence, VIF can be used to address this (multi) collinearity problem. The VIF for each covariate can be calculated as follows:

$$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{x_j|x_{-j}}} \ .$$

Where $R^2_{x_j|x_{-j}}$ denotes the coefficient of determination of a particular covariate (x_j) with respect to all other covariates (x_{-j}). If this value is near to 1, it denotes high collinearity.

If the VIF value is 1, it indicates that there is small collinearity or no collinearity at all. In general, for a variable, if the VIF is greater than 10, this suggests huge collinearity and it is preferred to drop one of the problematic variables (James et al., 2013, p. 99-102).

3.3 Dummy coding for categorical variables

The covariates can be continuous or categorical variables. In the case of categorical variables, the covariate x can contain many categories $\{1, ..., c\}$, a total of c categories. In the case of dummy encoding, $c-1$ dummy variables are created for the c categories. These dummy variables will contain only binary values (0 or 1). It can be represented as

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

for $i = 1, ..., n$. We then include these encoded variables in the regression model as:

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

In this approach, one dummy variable is excluded to maintain the identifiability. According to the above representation, category c is excluded in this case and it is termed as the reference category. The values of estimates from the other categories can be explained simply by comparing them with the reference category (Fahrmeir et al., 2013, p. 97).

3.4 Hypothesis testing and confidence intervals

The hypothesis testing is performed to examine the effect of covariates on the response variable. To define hypothesis statements, it is assumed that the errors are normally distributed i.e $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$. The null and the alternative hypotheses of the t-test are as follows:

H_0: None of the covariates have significant effect on the response variable ($\beta_j = 0$).

H_a: At least one covariate has significant effect on the response variable ($\beta_j \neq 0$) .

Here $j = 0, ..., k$. The test statistic t_j can be defined as

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

where

$se_j = \widehat{Var(\hat{\beta}_j)}^{1/2}$ represents the standard error estimate of $\hat{\beta}_j$,

t_j follows t-distribution with $n - p$ degrees of freedom.

The formula for variances of estimated coefficients ($\hat{\beta}_j$):

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

where

R_j^2 denotes the coefficient of determination for the regression model with x_j as response variable and all other variables (except x_j) as explanatory variables.

The $t_{n-p}(1 - \alpha/2)$ is obtained from the t-table using the values of $(1 - \alpha/2)$-quantile and $n - p$ degrees of freedom. The null hypothesis is rejected if the absolute value of t_j is greater than $t_{n-p}(1 - \alpha/2)$ ($|t_j| > t_{n-p}(1 - \alpha/2)$).

Correspondingly, the null hypothesis can also be rejected based on p-value. From the value of t_j, the associated p-value is obtained from the t-table. If the p-value is less than α the null hypothesis is rejected. (Fahrmeir et al., 2013, p. 125-131)

The probability for β_j to exist within a particular interval can be defined as the confidence interval for β_j. The $(1 - \alpha)$-confidence interval for β_j under the assumption of normality can be constructed as follows:

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

(Fahrmeir et al., 2013, p. 136)

3.5 Best subset selection

The process of selecting the subset of explanatory variables that explain the model best from the p explanatory variables can be defined as the best subset selection. To determine the best subset, 2^p models are fit including all possible combinations of explanatory variables. First, this process fits p models, which contain only one explanatory variable. Next, it fits $p(p - 1)/2$ models which contain only two explanatory variables, and so on. To find the best subset selection, methods like Akaike information criteria (AIC), Mallow's C_p, Bayesian information criteria (BIC), and other methods are used. However,

in this report, this process is performed using method of AIC (James et al., 2013, p. 205).

3.5.1 Akaike information criteria (AIC)

The AIC is one of the above-mentioned methods for selecting the best model from various competing models with different covariates and estimates. It determines the amount of information lost by a model. It is mainly used in the case of likelihood-based inference, the lower the AIC value, the better the model. AIC can be expressed using the following formula

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

where

$l(\hat{\boldsymbol{\beta}}_k, \hat{\sigma}^2)$ refers to the maximum log-likelihood value of unknown parameters in case of normal distribution of errors,

k denotes the number of covariates present in the model,

$k+1$ denotes the total number of parameters because error variance σ^2 is also considered as the parameter.

Hence, the term $l(\hat{\boldsymbol{\beta}}_k, \hat{\sigma}^2) = n \log(\hat{\sigma}^2) + n$. By substituting it in above equation and removing the constant value n, it changes as follows:

$$\text{AIC} = n \log(\hat{\sigma}^2) + 2(k+1) \ .$$

(Fahrmeir et al., 2013, p. 148).

3.6 Residual plots

A plot with ordered pairs of predicted values on the horizontal axis and residual values on the vertical axis can be defined as a residual plot. The residual plots are used to determine whether or not the model is a better fit for the data. If the points on the plot are randomly distributed across the mean line, the errors are homoscedastic, and the model fits the data better. If there are any patterns in the plot, the errors are heteroscedastic and the model needs to be improved. (Frost, 2019, p. 196)

4 Statistical analysis

The statistical methods described in the previous section were applied on the dataset used in this report and the results are interpreted in this section.

4.1 Preperation of the dataset

To prepare the dataset for analysis, the following changes will be made to the provided dataset.

First, we create a new variable *lp100*, which will have the converted value of the variable *mpg*. The value of *lp100* is calculated by dividing the number 282.48 by the value of the *mpg* variable.

$$lp100 = 282.48/mpg.$$

Second, the new variable *age* is created for each observation to hold the *age* of the car. As the dataset belongs to the *year* 2020, the *age* is calculated by subtracting 2020 from the variable *year* (*age* = 2020 - *year*). After calculating the variable *age*, the variable *year* in the dataset is replaced by *age*.

Third, we compute the logarithmic value of the response variable (*price*) using the formula $logprice = \log(price)$. The computed variable *logprice* is added to the dataset.

In conclusion, the new variable *lp100* will be used instead of *mpg* in the analysis. The variable *year* is replaced by *age*. The variables *price* and *logprice* are compared to analyze which response variable fits the data better.

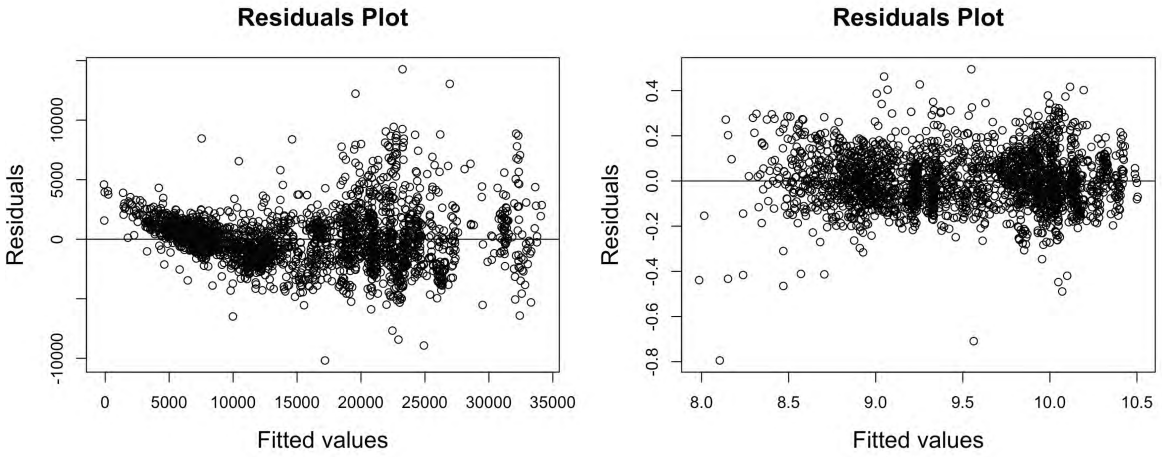
4.2 Descriptive analysis of the dataset

Table 2 in the appendix gives information about the summary of numerical variables. The average *price* and the average *logprice* of the used car are 15445 and 9.504 respectively. The new variable *age* ranges from 0 to 14.0, whereas lp100 ranges from 1.702 to 8.692. The max value of the *tax* paid for a particular car is £265.0. The mean value of *mileage* shows that on average the cars travelled a distance of 21021 miles. The *engineSize* ranges from 0 to 2.0.

Table 3 in the appendix gives information about the summary of categorical variables. The variable *model* contains 3 categories and each category has a nearly equal number of observations. The *model T-Roc* contains the highest price (£40999) and the *model passat* contains the lowest price (£1495). The *fuel type petrol* contains the highest number of observations when compared to other types and most of the advertised cars contain a manual *transmission* followed by *semi-auto* and *automatic*. Moreover, the data is imbalanced between the categories of *fuel type* and *transmission*

4.3 Response variable selection

After preparing the dataset, we have two response variables, namely, *price* and *logprice*. To determine which response variable fits the data better the residual plots and Q-Q plots with respect to both the variables are visualized and analyzed in this section.

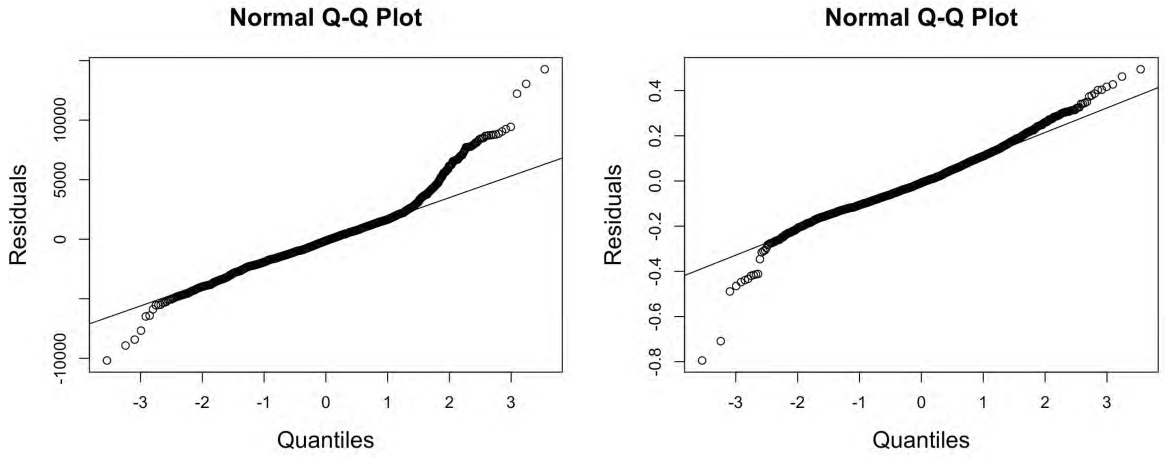


(a) Price (b) logPrice

Figure 1: Comparison of residual plots for linear model with reponse variable as price and logPrice

Assuming *price* as the response variable, we can notice from Figure 1(a), the points of the residual plot are not randomly distributed. Also, from Figure 2(a), it is evident that the data doesn't satisfy the property of normality as the data points largely deviate from the linear line for smaller and larger values.

Assuming *logprice* as the response variable, the points in the residual plot (Figure 1(b)) seem to be randomly distributed across the mean line. Moreover, from the Q-Q plot



(a) Price (b) logPrice

Figure 2: Comparison of Normality Q-Q plots for linear model with reponse variable as price and logPrice

(Figure 2(b)), the data points seem to deviate from the linear line for both smaller and larger values but it looks better when compared to the Q-Q plot of the *price* variable.

From the above statements, we can conclude that the model with *logprice* as the response variable fits the data better when compared to the model with the response variable *price*.

4.4 Analysis of assumptions

1.) From the Figure 1(b), we can notice that the errors are randomly distributed. Therefore, there is a linear relationship between the covariate and the response variable. We can also say that these errors are not related to each other.

2.) From the figure 1(b), the data points are equally distributed above and below the mean line. Hence, the mean value of the data points will be equal to zero i.e $E(\boldsymbol{\epsilon}) = 0$. Also, from the same figure, we can notice that the errors follow homoscedasticity. This proves the error terms contains a constant variance (σ^2).

3.) Table 3 provides information on the VIF values of each covariate. From this table, we can notice that all the VIF values are less than 10. This shows that there exists no linear relationship between the covariates (i.e there is no multicollinearity).

4.) From the figure 2(b), we can notice that the data points are closely aligned to the linear line. This proves that the errors are normally distributed.

4.5 Best model selection

In this subsection, the best set of covariates that explain the response variable *logprice* is determined by using the best subset selection procedure. For all the possible combinations of covariates, the linear regression model is fit and AIC values are computed. The model with the smallest AIC value will contain the largest amount of information. For our dataset, the minimum AIC value is - 3644.49, with *model, transmission, mileage, fuelType, tax, engineSize, lp100,* and *age* as covariates.

AIC model: $logPrice \sim model + transmission + mileage + fuelType + tax + engineSize + lp100 + age$.

4.6 Estimation, confidence intervals, and R-squared

Table 4 in the appendix gives information about the summary of estimates and confidence intervals. It contains continuous variables and categorical variables including the dummy encoding for categorical variables. The column "estimate" denotes the effect of the particular explanatory variable on the response variable. It provides estimates for 12 explanatory variables. Among these variables half of them contain positive values whereas the other half of them contain negative values. Hence, the variables *model T-Roc, fuelTypeHybrid, fuelTypeOther, fuelTypPetrol, engineSize,* and *lp100* has a positive effect on the *logprice* whereas the variables *model Up, transmissionManual, transmissionSemi-Auto, mileage, tax*, and *age* have a negative effect.

If all the covariate values are equal to zero, then the average value of *logprice* is equal to the intercept (9.6535). Let all covariates are assumed to be zero except *fuelTypePetrol*, then the average value of *logprice* is increased by a factor of 0.0762. Similarly, if *age* is considered to be the only non-zero variable, then the average of *logprice* is decreased by a factor of 0.0932. Moreover the variables *mileage, transmissionSemi-Auto* and *tax* have values almost zero and hence they have no significant effect on *logprice*.

The column "Pr($>$ |t|)" gives the p-value for all the explanatory variables in the model. The null hypothesis states that the variables have no significant effect on the response variable and the alternate hypothesis states that there exists at least one variable that has a significant effect. Using these hypotheses and the level of significance as 5%, the p-values are interpreted. From Table 4, we can notice that all the p-values are near 0 except for the variable *transmissionSemi-Auto*. This gives enough evidence to reject

the null hypothesis. Therefore, all the variables except *transmissionSemi-Auto* have a significant effect on the *logprice*.

The column "confidence interval" gives the possible range of values with upper and lower bounds for the estimates of variables. We can notice from Table 4 the confidence interval for the *transmissionSemi-Auto* is [-0.2. 0.2]. Using the same hypothesis as mentioned above, if the confidence interval includes the value 0, then there is not enough evidence to reject the null hypothesis. Hence, the variable *transmissionSemi-Auto* has no significant effect on *logprice*. All the other variables have significant effect.

The goodness of fit determines how well the model fits the provided dataset. In our model, we have the R-squared value as 0.9544. This shows that our model with the selected covariates explains the 95.4% of the variation in the response variable *logprice*. As the value is close to 1, we can conclude that our model fits the data better.

5 Summary

In this report, linear regression analysis was performed on the dataset and the results are interpreted. The dataset used in this report was compiled by the instructors of the course Introductory to case studies at TU Dortmund University in the summer semester 2022. The dataset was created using the information of various models of used cars in the UK from the year 2020. The data was restricted to a particular company namely Volkswagen (VW) and advertised on the e-commerce platform Exchange and Mart. Also, this dataset is a small part of the large dataset available on Kaggle (Jhanwar, 2020). It contained a total of 2532 observations belonging to the year 2020, with *price* as the response variable and 8 other predictor variables namely *model, year, transmission, mileage, fuelType, tax, mpg, engineSize*.

To achieve the goal, first, the dataset was transformed by changing *mpg*'s measuring unit, calculating the car's *age* and replacing it with the variable *year*, and computing the new column *logprice* as a logarithmic function of price. Second, descriptive analysis was performed on the transformed dataset and the results are included in Table 1 & 2. Third, after fitting the two models with *price* and *logprice*, a comparison between the *price* and *logprice* was performed using Q-Q plots and residual plots to understand which model fits the data better. From this, a conclusion is made that the response variable *logprice* fitted the data better. Fourth, to determine the best subset of covariates Akaike information criteria was used and it resulted in the model $logPrice \sim model +$

$transmission + mileage + fuelType + tax + engineSize + lp100 + age$ with an AIC value of -3644.49. Fifth, the linear regression model was fit using the best response variable and the best subset of covariates. From this model, the assumptions of linear regression are validated and it holds true.

Furthermore, from the estimate values the variables *transmissionSemi-Auto, mileage* and *tax* are almost zero, these results proved that these variables have no significant effect on the response variable. From the p-value results, all the variables have values of nearly zero except for *transmissionSemi-Auto* which has a p-value 0.9833. This statement provided evidence to reject the null hypothesis.

For further analysis, including more observations will provide more information for the analysis. In addition, including more explanatory variables such as the interior and exterior condition of the car, accident history, color may help explain the response variable more accurately. We can also split the data into train and test to assess the performance of the model.

Bibliography

Betterton, R. (2022), *Top advantages to buying a used car.* URL: https://www.bankrate.com/loans/auto-loans/advantages-to-buying-a-used-car/ (visited on 19 June 2022).

Dodge, Y. (2008), *The Concise Encyclopedia of Statistics*, Springer New York, New York, NY.

Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. (2013), *Regression Models, Methods and Applications*, Springer Berlin, Heidelberg.

Fox, J. and Weisberg, S. (2019), *An R Companion to Applied Regression*, Sage, Thousand Oaks CA. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/ (visited on 14th June 2022).

Frost, J. (2019), *Regression Analysis, An Intuitive Guide for Using and Interpreting Linear Models*, Statisics By Jim Publishing.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013), *An Introduction to Statistical Learning with Applications in R*, Springer New York, NY.

Jhanwar, A. (2020), *100,000 UK Used Car Data set*, Kaggle. URL: https://www.kaggle.com/code/abhinavjhanwar/used-car-price-prediction-volkswagen-r2-score-96/data (visited on 14 June 2022).

R Core Team (2021), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Version: 4.0.5, Vienna, Austria.

Appendix

A Additional tables

Variable	Min	Q1	Median	Mean	Q3	Max
price	1495	8495	13986	15445	21422	40999
logPrice	7.310	9.047	9.546	9.504	9.972	10.621
mileage	1	3803	12095	21021	29052	176000
age	0	1.0	2.0	2.429	4.0	14.0
lp100	1.702	4.4	5.202	5.253	5.695	8.692
engineSize	0	1.0	1.5	1.466	2.0	2.0
tax	0.0	20.0	145.0	105.3	145.0	265.0

Table 1: Measures of central tendency and dispersion of the continuous variables

Variable	Sub categories	count	Min	Q1	Median	Mean	Q3	Max
Model	T-Roc	773	11489	19950	21990	22839.392	24590	40999
	Passat	915	1495	10989	14999	16684.683	20998.5	39989
	Up	884	3495	6495	7699	8029.428	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

Table 2: Measures of central tendency, dispersion and count of the categorical variables

Variable	VIF
model	6.08
mileage	2.85
fuelType	5.24
engineSize	5.53
tax	2.42
transmission	1.75
lp100	3.25
age	3.22

Table 3: VIF values for covariates

	Estimate	Pr(> | t |)	Confidence interval
(Intercept)	9.6535	0.0000	[9.59, 9.71]
model T-Roc	0.1117	0.0000	[0.10, 0.13]
model Up	-0.5684	0.0000	[-0.59, -0.55]
transmissionManual	-0.1198	0.0000	[-0.14, -0.10]
transmissionSemi-Auto	-0.0002	0.9833	[-0.02, 0.02]
mileage	-0.0000	0.0000	[-0.00, -0.00]
fuelTypeHybrid	0.4346	0.0000	[0.40, 0.47]
fuelTypeOther	0.0719	0.0180	[0.01, 0.13]
fuelTypePetrol	0.0762	0.0000	[0.06, 0.10]
tax	-0.0004	0.0000	[-0.00, -0.00]
engineSize	0.1774	0.0000	[0.15, 0.20]
lp100	0.0339	0.0000	[0.03, 0.04]
age	-0.0932	0.0000	[-0.10, -0.09]

Table 4: Values of estimates, p-value and confidence intervals