Contents

1 Introduction

For over a century, the automobile industry has made a difference in people's lives. Every person has come across the automobile platform at some point in their life, be it for buying or selling motor vehicles. Determining if a new or a used car is worth the advertised price is a challenging task. As a result, being able to accurately predict the price of cars is of economic importance to both sellers and purchasers. The age and model of the car, the engine type, fuel type, total kilometers driven, and other factors all influence the price of a car. The main goal of this project is to focus on the price of used cars and present a linear regression model for estimating them.

The purpose of this report is to analyze the data set of Volkswagen (VW) cars that are extracted from the website (Exchange and Mart, 2020), which contains a listing of the cars sold in the United Kingdom in the year 2020. To begin, we undertake data pre-processing steps that include basic variable transformations and deriving new variables from the existing ones in the data set. Secondly, we estimate the entire model and decide whether the target or response variable in the regression analysis should be the raw price or the log-transformed price. Following that, using the best subset selection approach, such as the Akaike information criterion (AIC), the best set of explanatory variable for the response variable is identified. Finally, we compute and interpret the estimate results, confidence intervals and goodness of fit for our model.

Apart from the introduction, there are four other sections in this report. Section 2 provides a summary of the given data set as well as an explanation of all of the report's tasks. In Section 3, the statistical methods used in data analysis are defined in detail. The classical linear regression model, best subset selection, and related methods along with the properties and interpretation are all discussed here. The presented statistical methods and tests in Section 3 are applied to the given data set in Section 4, and the findings are analyzed. Section 5 concludes with a summary of the findings as well as key conclusions and suggested future research.

2 Problem statement

2.1 Description of the data set and quality

The data set was provided by the lecturers of TU Dortmund University's Introductory Case Studies course in the summer semester of 2022. The data set used in the analysis of this report contains a listing of the Volkswagen (VW) cars that were sold on a car platform (Exchange and Mart, 2020) in the United Kingdom in the year 2020. It is a subset of a larger data set that is available on Kaggle (Jhanwar, 2020).

The data set contains 2532 observations and ten features or variables. Three out of ten features are categorical type variables and they are, *model* which indicates whether the sold car is a Passat, T-Roc, or Up car model, *fuelType* which indicates the kind of fuel the car consumes and they could be Petrol, Diesel, Hybrid or Others, and the last categorical variable is *transmission* which indicates if the car has Manual, Automatic or Semi-Auto gearbox. Six out of ten features are numerical type variables and they are, *price* which indicates the price of the sold car in GBP (£), *mileage* indicates the total distance (in miles) the car was driven, *mpg* (miles per gallon) that mentions the distance (in miles) the car can travel with one gallon of fuel, *engineSize* which refers to the size of the car's engine in liters, *year* which indicates the year of the first registration of the car and *tax* which refers to the amount of the annual tax (Vehicle Excise Duty) to be paid for the car. There exists an unknown and unspecified variable in the data set which tends to reflect as *id* of the car which can be eliminated for this report as it is irrelevant. Furthermore, there are no missing values or NA values in the data set, and the data quality is appropriate for the analysis.

2.2 Project objective

There are three main objectives of this report. The first objective is to perform data preparation to find the useful collection of data needed for the analysis of the upcoming objectives. The second objective is to apply model diagnostics tools for finding the best response variable of the model and performing subset selection criteria using the AIC method to find the best set of explanatory variables for the response variable. Finally, the coefficients of the best linear model identified by the AIC in the previous step are calculated and explained. In addition to the goodness of fit, the confidence intervals for the regression parameters are calculated and analyzed.

3 Statistical methods

This section introduces statistical methods that will be used to analyze and evaluate the data set in the next section of the current report. R (R Development Core Team, 2020) version 4.0.5 with package *car* (Fox and Weisberg, 2019) and default loaded packages are used for all data processing and visualizations.

3.1 Classical linear regression model

In linear regression model, our main goal is to investigate the relationship and impact of a given set of explanatory variables (also known as independent variables, predictors, or regressors) $x_1, x_2, ..., x_k$ on the target variable y. In general, a linear function $f(x_1, x_2, ...x_k)$ could be used to model the connection between the target and the explanatory variables. The linear function is considered to have an error or noise (ε) in the model because the connection is not accurate, and thus we get the form as below.

$y = f(x_1, x_2, ...x_k) + \varepsilon$, where ε is the error or the noise in the model and f is the unknown function and represented as a linear combination of covariates given as, $f(x_1, x_2, ...x_k) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$.
Hence, the linear function can be rewritten as,
$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon$, where $\beta_0, \beta_1, ..., \beta_k$ are the unknown regression parameters or coefficients which needs to be determined.(Fahrmeir et al., 2013, p. 73-74)

For $i = (1, 2, ...n)$ samples, the matrix form of the regression model is given as,

$$\mathbf{y} = \begin{pmatrix} y_0 \\ \vdots \\ y_i \end{pmatrix} \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_0 \\ \vdots \\ \varepsilon_i \end{pmatrix}$$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{i1} & \cdots & x_{ik} \end{pmatrix} \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}$$

or, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
where $\mathbf{y}$ is the response vector, $\mathbf{X}$ is the design matrix, $\boldsymbol{\beta}$ is the unknown parameter vector and $\boldsymbol{\varepsilon}$ is the error vector.(Fahrmeir et al., 2013, p. 73-75)

3.1.1 Model assumption

There are a few assumptions that must be addressed before using the linear regression model. A linear regression model requires that the number of observations in the data set must be equal to or larger than the number of regression coefficients, which is an important criterion.

- The target and the explanatory variables must have a linear relationship. Residual vs fitted plot is used to verify this assumption.
- No perfect multicollinearity, which implies that any explanatory variable that is a linear modification or transformation of other explanatory variables should be avoided. This assumption could be validated using the variance inflation factor (VIF) which is discussed in subsection 3.1.4 of this report.
- Errors are assumed to be normally distributed which can be validated using a QQ-plot. If the residuals lie on the straight reference line then it can be said that the errors follow a normal distribution Thus, $E(\varepsilon_i) = 0$ and constant error variance $Var(\varepsilon_i) = \sigma^2$ are considered accordingly.
- The observations or records must be independent and identically distributed (i.i.d).

(Fahrmeir et al., 2013, p. 75-76)
All the above-mentioned assumptions with respect to the data set are verified in the next section of this report.

3.1.2 Estimation of parameters

The unknown regression coefficients can be estimated either by minimizing the sum of squared deviation (least squares approach) or by maximizing the likelihood estimation.

Least squares method The least squares approach helps in estimating the unknown regression parameters by minimizing the sum of squared deviations.

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

The sum of squared deviations is minimized by setting the vector of first derivative to zero. As a result, the obtained coefficient estimator is given as:
$\hat{\boldsymbol{\beta}} = \boldsymbol{(X'X)^{-1}X'y}$. where $\hat{\boldsymbol{\beta}}$ is the estimated coefficient vector.

Maximum likelihood estimation The least square estimate does not need any specific distributional assumptions for the error term whereas the maximum likelihood estimation assumes errors to be normally distributed $\varepsilon \sim N(0, \sigma^2 I)$. Assumption of normally distributed errors, we get the response variable vector, y, to be distributed normally, i.e., $y \sim N(X\beta, \sigma^2 I)$ which gives the log likelihood as,

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

Maximizing the above log likelihood with respect to β is equivalent to minimizing the sum of squared deviation and gives the same result of estimated coefficient vector as seen above. As a result, the maximum likelihood estimator of β is equal to the least squares estimator.

From the coefficient estimator, it is also possible to compute the predicted response or target variable using the below equation.

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

(Fahrmeir et al., 2013, p. 105-107)

Residuals Residuals can be defined as the difference between the true y_i and the predicted value $\hat{y}_i$ of the target variable. Additionally, the residuals are the estimates of errors or noise ε_i and denoted as $\hat{\varepsilon}_i$.

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$
$$or,\ \hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}.$$

(Fahrmeir et al., 2013, p. 77)
By dividing the estimated residual by the estimated standard deviation, we can obtain the standardized residual, which can then be plotted against the predicted value to see if the assumption of homoscedastic variance is broken or not. The standardized residual is computed as below:

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

where h_{ii} are the diagonal elements of the Hat matrix H which is given as,

$$\boldsymbol{H} = \mathbf{X}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')$$

Hence, the predicted response variable can be rewritten as,

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

(Fahrmeir et al., 2013, p. 124,107)

3.1.3 Dummy encoding for categorical variables

If the data set contains categorical variables $x_i \in 1, ..., c$ with c categories, the dummy encoding is used to model the influence of these variables by introducing c-1 dummy variables in the regression model. These dummy variables can have one of the two values i.e., 0 or 1.

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

To make the model identifiable, one of the dummy variable is omitted. For the above equation, the dummy variable for category c is removed and this category is then known as the reference category. Direct comparison with the excluded reference category is then used to interpret the effects of other dummy variables.(Fahrmeir et al., 2013, p. 97)

3.1.4 Collinearity analysis

As mentioned in subsection 3.1.1, multicollinearity occurs when two or more independent or explanatory variables are correlated or linearly dependent on each other. Multicollinearity can affect the model fitting and the regression results as the model parameters become extremely sensitive to small changes in the model and as a result, it becomes hard to check explanatory variables that are statistically significant for the response variable. In order to avoid the multicollinearity issue, we use the variance influence factor (VIF) in this report.

Variance influence factor Variance influence factor (VIF) is a technique to detect multicollinearity in regression analysis. It calculates how much multicollinearity in the model has inflated the variance of a regression coefficient. The higher the correlation between covariate x_j with other covariates, the higher the coefficient of determination

${R^2}_j$ and hence the larger the variance $Var\hat{\beta}_j$. VIF is given by the below equation:

$$VIF(\hat{\beta}j) = \frac{1}{1 - {R_j}^2} \, .$$

where j=1, 2, ..k is the number of covariates and ${R^2}_j$ is the coefficient of determination of covariate x_j.

When the variance inflation factor is larger than 10, i.e., $VIF_j > 10$, we state that there is a major collinearity problem.(Fahrmeir et al., 2013, p. 157-158)

3.1.5 Goodness of fit measure

The extent to which the obtained model fits the data is measured by its goodness of fit. Coefficient of determination (R^2) is one of the goodness of fit statistics used in this project to assess the model's quality. It is computed as:

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

(Fahrmeir et al., 2013, p. 113)

Its value varies from 0 to 1. If the R^2 is near to one, it means the residual sum of square is low and the model fits the data better and $R^2 = 1$ implies that the residuals are zero with a perfect fit to the data. Additionally, we can use *Adjusted* R^2, as an alternative of R^2 to avoid the false interpretation of increased R^2 as the number of explanatory variables increases. As a result, we include the number of explanatory variables in the computation for *Adjusted* R^2.

$$\bar{R}^2 = 1 - \frac{n-1}{n-k}(1 - R^2); \quad \text{where k is the number of covariates.}$$

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

3.2 Hypothesis testing and test statistic

3.2.1 Null and alternate hypothesis

- Test of Significance (Using t-test)
  The null hypothesis H_0 of the t-test statistic indicates that the estimated param-

eter's corresponding value is equal to zero, whereas the alternative hypothesis H_1 suggests that the estimated parameter's value is not equal to zero. It can be written as, $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$ for $j = (1, 2, ..k)$.(Fahrmeir et al., 2013, p. 127)

3.2.2 Test statistic

T-statistic The T-statistic can be calculated using the below formula:

$$t_j = \frac{\hat{\beta}_j}{s\hat{e}_j}$$

and $s\hat{e}_j = \widehat{Var(\hat{\beta}_j)}^{1/2}$.
where $\hat{\beta}_j$ is the estimated coefficient, $s\hat{e}_j$ is the estimated standard error or standard deviation of $\hat{\beta}_j$ and $\widehat{Var(\hat{\beta}_j)}^{1/2}$ is the diagonal element of the estimated covariance $\widehat{\text{Cov}(\hat{\beta})}$ $= \sigma^2(X'X)^{-1}$.

After computing the T-statistic, we use the t-distribution table to determine the corresponding t-value. If the calculated t-value is greater than the t-value received from the t-distribution table, we reject the null hypothesis, else we fail to reject it. As a result, if the t-value is higher, we can say that the coefficient would be statistically significant and less probable to be equal to zero. The null hypothesis that the associated coefficient has no effect on the dependent variable is tested by the p-value of each covariate's t-statistic. This procedure is already explained in project 2. We can reject the null hypothesis with a p-value less than or equal to 0.05. A large p-value, on the other hand, implies that the change in the predictor variable is unrelated to the change in the dependent variable.

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

Confidence interval We know that estimates of regression coefficients are prone to sampling error. As a result, we will never be able to correctly estimate the real value of these parameters from sample data. However, we create confidence intervals for the coefficients to evaluate the fitted value's predictions for the observed values of the variables. We use the T-statistic and obtain $(1 - \alpha)$ confidence interval for β_j as,

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

where n is the number of observations, $p = k+1$ is the number of β coefficients including the intercept, $(n-p)$ is the degree of freedom and $\hat{\beta}_j$ is the estimated coefficient.(Fahrmeir et al., 2013, p. 136)

3.3 Model choice criteria

Best subset selection is a strategy that considers all possible combinations of independent variables in order to discover the subset of independent variables that best predicts the results. For k different explanatory or independent variables, 2^{k-1} model combinations are possible (excluding null model). All of these models are then evaluated and the best model is selected. There exists several different techniques to select the best subset of explanatory variables such as Akaike information criteria (AIC), Bayesian information criteria (BIC), Mallow's complexity parameter etc. In this report, Akaike information criteria (AIC) is used as the model choice criteria.

Akaike information criteria (AIC) Within the scope of likelihood-based inference, the Akaike information criterion (AIC) is used for model choice. It is computed as:

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

Here, $|k\,| + 1$ is the total number of parameters and $l(\hat{\boldsymbol{\beta}}_k, \hat{\sigma}^2)$ is the maximum log-likelihood (ML) value. The AIC for each regression model is calculated in order to compare them. Models with low AIC values are considered to be a better model fit.
In case of linear models with gaussian errors, the maximum likelihood is calculated as, $-2 \cdot l(\hat{\boldsymbol{\beta}}_k, \hat{\sigma}^2) = n\ log(\hat{\sigma}^2)$.
and the AIC is then computed as,

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

(Fahrmeir et al., 2013, p. 146-148)

4 Statistical analysis

In this section, the statistical methods outlined above are applied to the data set provided, and the results are evaluated and interpreted.

4.1 Data preparation

Before we begin our analysis, we must first prepare our data. Preparation includes reformatting and finding the informative set of data that will help us improve the performance of our models. Firstly, we transform the variable *mpg* from fuel consumption in miles per gallon to *lp100* fuel consumption in liters per 100 kilometers using the conversion formula $lp100 = \frac{242.48}{mpg}$. The new variable *lp100* is then added to our data set. Furthermore, we calculate the *age* of the car as a new variable in our data set by subtracting 2020 (year of purchase) from the variable *year* which refers to the year of the initial registration of the car. Finally, we also compute the logarithm of the car price as *logprice* and store it into a new column in our data set which will help in determining whether to use *logprice* or raw *price* as a response variable in linear regression after estimating the full model. To prevent multicollinearity, the data set is reduced by deleting irrelevant variables or those that are a transformation of other variables. *mpg* and *year* are removed from the data set, since they are replaced with the variables *lp100* and *age*.

4.2 Descriptive analysis

Following the data preparation step, the final data set has the same number of observations (2532) and a total of 10 variables. Table 2 in the Appendix of this report provides a descriptive summary of all the continuous variables in the data set. It can be observed from the Table 2 that, the variable *price* varies from 1,495£ minimum to 40,999£ maximum, with the converted *logprice* ranging from 7.310£ minimum to 10.621£ maximum. The total distance traveled by car is recorded as variable *mileage*, with a minimum value of one mile and a maximum value of 176,000 miles. Following prepossessing, the car's recorded *age* spans from 0 to 14 years, with an average of around 2 years. *EngineSize* has a minimum and maximum value of 0 liters and 2 liters, respectively. The annual tax (Vehicle Excise Duty) payable on the car ranges from 0£ to 265£, with an average of 105.3£. The variable *lp100* records the fuel needed to travel 100kms distance and the minimum recorded is 1.702 liters/100kms whereas the maximum is 8.692 liters/100kms.

Table 3 in the Appendix shows a descriptive summary of all the categorical variables present in the data set. There are 773 T-Roc, 915 Passat and 884 Up *models*. When compared to the other two models, the model T-Roc has a higher average price (22,839.39£). The used car's maximum price is 40,999 £ and it belongs to the T-Roc model, while the lowest price belongs to the Passat model. There are 58 hybrid, 970 diesel, 1488 petrol

and 16 other *fuelType* cars. The average cost of a Hybrid fuelType car (27,622.29£) is higher than the cost of a Diesel (16,826.67£) or Petrol (14,015.94£) fuelType car. The car has three different types of *transmission*, with the semi-automatic gearbox having a higher average price (22,324.15£) than the automatic (22,222.7£) and manual (12,771.79£) gearboxes.

4.3 Response variable selection

In this section, we will determine whether to use the raw price or the log-transformed price as the response variable in the linear regression analysis. Two models were employed to reach this conclusion, each having the same and all appropriate explanatory variables but different response variables. The first model (Model 1) contains 'raw price' as the response variable, whereas the second model (Model 2) contains 'log-transformed price' as the response variable. After fitting both the models, two plots (QQ-plot and Residual plot) are shown as a model diagnostic tool to determine which model is better in line with the linear model assumptions. The comparison between Model 1 and Model 2 is shown in Figure 1 of the Appendix.

Assumption verification

- Figure 1(a) and 1(c) depicts the residual plot for Model 1 and Model 2 respectively. It can be observed that Figure 1(a) shows a horizontal line with unequally dispersed points forming a clear cone-shaped structure which indicates that the assumption of linearity of the data and homogeneity of residual variances are violated. Figure 1(c) on the other hand, displays no pattern in the residual plot and a horizontal line with equal spread points, indicating residual variance homogeneity and a linear connection between the response and the independent variables. As a result, there is no assumption violation in Figure 1(c).
- Figure 1(b) and 1(d) depicts the QQ-plot for Model 1 and Model 2 respectively. The QQ-plot in Figure 1(b) shows significant deviations from normality at its tails, indicating that the normality of the error distribution, which is one of the assumptions for classical linear regression, is broken. Figure 1(d), on the other hand, shows that all the points approximately fall along the reference line with slight acceptable deviations at its tails, implying that the linear model assumption is valid in Figure 1(d).

- It is assumed that the observations are independent and identically distributed since the data is randomly sampled from the population.
- There exist no multicollinearity in the data set since a few explanatory factors were removed from the data set during the data preparation step because they were transformations of other explanatory variables. As a result, each variable is distinct. The assumption of multicollinearity is further validated by the variance inflation factor (VIF). The VIF values for all the variables are less than 10, indicating that there is no multicollinearity between any explanatory variables, as shown in Table 4 in the Appendix. As a consequence, no explanatory variable is dropped.

Based on the above considerations, it can be stated that Model 2 is more in line with the linear model's assumptions. Hence, log-transformed price is chosen as the response variable and the raw price variable can be safely removed from the data set.

4.4 Best subset selection

In this subsection, we will use the best subset selection technique to select the best set of explanatory variables for the response variable *logprice*. The final data set for this task has 2532 observations and 8 explanatory variables. The best subset selection approach fits 255 ($2^8 - 1$) linear models for 8 explanatory variables. The obtained best model has an AIC value of -3664.49 and it is given as,

- **AIC best model**: logprice $\sim$ model + mileage + fuelType + engineSize + transmission + lp100 + age + tax

4.5 Estimation of the best linear model

The best model for the best AIC value obtained in the above subsection is fitted and it is interpreted using Table 1. It can also be seen that the regression process generated dummy variables for all categorical variables. The 'estimate' column of Table 1 shows the intercept value is 9.65, which is the value of the *logprice* when all the other variables remain 0. 6 out of 12 parameter estimations are negative, implying that with a unit increase in these parameter coefficients, a reduction in the value of *logprice* by the estimated value of these coefficients can be observed, whereas, a unit increase in positive parameter coefficients causes an increase in *logprice* by the estimated value of these

coefficients. For example, when *engineSize* is increased by 1 unit, the *logprice* increases by 0.177 units. Similarly, if the *age* is increased by 1 unit, the *logprice* decreases by 0.093 units and a similar interpretation could be made for other continuous variables. To understand the interpretation of categorical variables, let us take an example of the variable *model*, the model Passat is taken here as the reference category and the estimate of model T-Roc is 0.112 which indicates that the *logprice* of model T-Roc is greater by 0.112 units to that of model Passat, considering all other covariates remains constant. Likewise, *logprice* of model Up is lower by 0.568 units to that of model Passat. A similar interpretation could be made for other categories of categorical variables.

Table 1 further shows the p-value for each parameter estimate that is obtained using the corresponding t-statistics which follow t-distribution with 2519 degrees of freedom. With the exception of *transmission Semi-Auto*, all of the other coefficients have p-values less than 0.05 which means that the null hypothesis can be rejected and as a result, all of these coefficients deviate considerably from 0 and have a relationship with the response variable. The variable *transmission Semi-Auto*, on the other hand, with a p-value of 0.983 indicates that we fail to reject the null hypothesis and this variable does not show a proper association with the response variable and can be eliminated.

Confidence Interval The 95% confidence intervals of the parameters are shown in Table 1. Confidence intervals that do not include zero, according to the test results, indicate that the coefficient is significant and that the relevant explanatory variable has an effect on the response variable. It can be observed that, with the exception of *transmission Semi-Auto*, value 0 does not fall within the 95 percent confidence range for all other coefficients, indicating that these variables have an effect on the response variable *logprice*. Because the confidence interval for transmission Semi-Auto is [-0.01,0.01], which contains 0, the related coefficient for the response variable is not significant and may be deleted.

Goodness of fit The value of the R^2 or also known as the coefficient of determination is 0.954, indicating that the model explains roughly around 95 percent of the variation in the response variable. Being close to 1, we can say that the model fit is good and efficient. But, as discussed the drawbacks of R^2 in Section 3.1.5, it is more practical to look at the value of *Adjusted* R^2 value. The *Adjusted* R^2 score is 0.954, which is the same as R2 and close to 1, indicating that the fitted model is very effective.

Table 1: Estimates of the best linear model

	Estimate	**Std. Error**	**t value**	**P-value**	**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

5 Summary

The data set used in the analysis is an extract of a larger data set available on Kaggle (Jhanwar, 2020) that comprises data from cars sold on an online used car selling platform (Exchange and Mart, 2020) in the United Kingdom in 2020. The data set includes 2532 observations and 10 variables. Initially, we did a data pre-processing task by computing *logprice* from the variable *price* and then converting fuel consumption to liters per 100 kilometers *lp100* instead of miles per gallon *mpg*. Finally, the car's *age* is estimated by subtracting 2020 (year of purchase) from the car's *year* of registration. To avoid multicollinearity in our data set, two variables, *mpg* and *year* are removed from the data set. Following that, we used model diagnostic tools such as the residuals vs fitted plot and the QQ-plot to compare *price* and *logprice* as response variables, concluding that the model with *logprice* should be used as the response variable as it is more consistent with the assumptions made in the linear model. The next task was to use the AIC method to select the best set of explanatory variables for the linear regression study. The model quality of all possible combinations of variables was evaluated using the model selection criteria. AIC discovered the best set predictors *(model, mileage, fuelType, engineSize, transmission, lp100, tax and age)* with a score of -3664.49. Using regressors determined by the AIC, a linear model was generated for the response variable *logprice*. The findings demonstrate that, with the except *transmission Semi-Auto*, all other coefficients have a p-value of less than 0.05, and that the zero value is not included in the 95 percent

confidence interval for these variables. As a result, these explanatory variables have a significant effect on the response variable *logprice*. *Transmission Semi-Auto* has a p-value larger than 0.05 and a confidence interval containing 0, thus it can be removed from the model. Finally, R^2 and *Adjusted* R^2 were used to determine the goodness of fit. The R^2 and *Adjusted* R^2 values revealed that the model explains roughly 95 percent of the response variable variance, and the high number implies that the model fit is good and efficient.

For future research, we can integrate data from other manufacturers and not only from Volkswagen, to see if the car's manufacturer plays an important factor in determining the price of a used car. In further analysis, it would be of great interest to include other selection criteria such as BIC (Bayesian information criteria) or Mallow's Cp criteria to check and compare the subset of regressors selected by these criteria.

Bibliography

Exchange and Mart (2020), *ExchangeandMart.co.uk*. [Online; accessed 19-June-2022].
URL: *https://www.exchangeandmart.co.uk/*

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

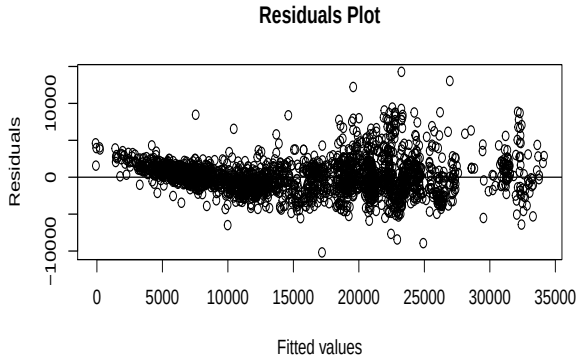
Fox, J. and Weisberg, S. (2019), *An R Companion to Applied Regression*, third edn, Sage, Thousand Oaks CA.
URL: *https://socialsciences.mcmaster.ca/jfox/Books/Companion/*

Jhanwar, A. (2020), *100,000 UK Used Car Data set*, Kaggle. [Online; accessed 19-June-2022].
URL: *https://www.kaggle.com/code/abhinavjhanwar/used-car-price-prediction-volkswagen-r2-score-96/data/*

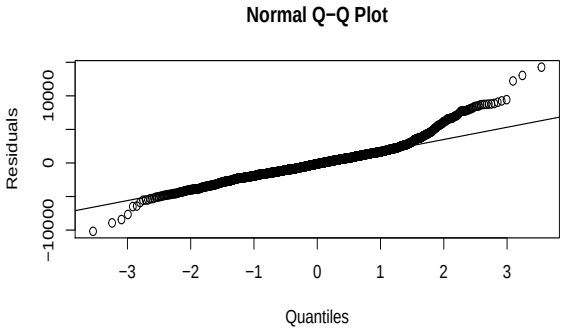
R Development Core Team (2020), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Appendix

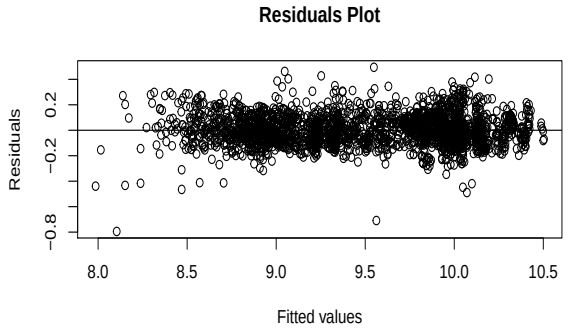
A Additional figures



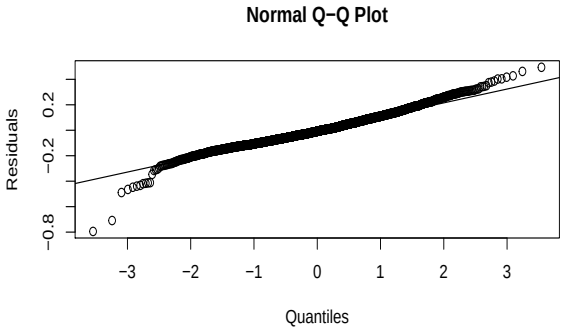
(a) Model 1 : Raw-price as response variable



(b) Model 1 : Raw-price as response variable



(c) Model 2 : Log price as response variable



(d) Model 2 : Log price as response variable

Figure 1: Model Diagnostic Plots

B Additional tables

Table 2: Descriptive analysis of the continuous variables

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 3: Descriptive analysis of the categorical variables.

	Variable	**count**	**Min**	**Q1**	**Median**	**Mean**	**Q3**	**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

Table 4: VIF values of the explanatory variable

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