TU Dortmund

Introductory Case Studies

Project 3: Regression Analysis

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

Author: Akshay Choudhary (224437)

Group number: 9

Group members: Nidhi Kiritbhai Patel

June 20, 2022

Contents

1 Introduction

The world has changed after the internet was born. It changes the way people buy and sell products. Nowadays, anyone from anywhere can buy or sell products without meeting in-person. It encouraged the automobile industry as well, and multiple e-commerce websites are competing in this area. Exchange and Mart (Exchange Enterprises, Newsquest Media Group, 2022) is one among them and it's data set is used in this project work. The data consists of information related to the car models Up, Passat, and T-Roc which are manufactured by Volkswagen (VW) company (Volkswagen AG, 2022). These models have price, year, mileage, miles per gallon (mpg), fuel type, engine size, tax, transmission as the features. The main objective of this project is to find the best set of features that impacts the price of the car. The optimal regression model is developed to predict the price based on the best-selected features.

To perform a brief statistical analysis, the data is prepared accordingly. We check model fit for the response variable price with and without log transformation, and liters per 100 kilometers are considered instead of miles per gallon (mpg). The year variable is converted to age. By taking all the features, we perform linear regression analysis. Residual versus fitted plots and QQ plots are used to check the assumptions of the linear model. Further, the best subset of features is determined using the Akaike information criterion (AIC) or Mallow's Cp. The AIC subset selection method is used in this report. Regression coefficients with their statistical significances, confidence interval, and the goodness of fit for the model will be interpreted.

The data set is described in detail in Section 2 and the objectives of the report are also briefly mentioned. Section 3 explains the statistical methods such as multiple linear regression, assumptions and definition of the linear model, tests for linear model, and best subset selection methods. Section 4 reflects the implementation of the approaches using the methods presented. Finally, Section 5 outlines the critical finding and address further studies that can be conducted on this dataset.

2 Problem statement

The data set is provided by the instructors of the course Introductory Case Studies at TU Dortmund University for the summer semester 2022. The original data set is extract from a large data set accessible on kaggle.com (Kaggle Inc., 2021) that consists of information on used automobiles sold via an e-commerce platform Exchange & Mart in the UK in 2020. The size of the sample dataset is 2,532 observations containing information about car models Up, Passat, and T-Roc manufactured by Volkswagen (VW) Company (Volkswagen AG, 2022) and their features are considered as the objects of study in this analysis.

2.1 Dataset description and data quality

The data set includes nine variables, with three categorical i.e., model, transmission, fuelType, and five variables are numerical i.e., price, mileage, tax, mpg, engineSize and one variable is ordinal i.e., year. The categorical variable model covers three car model types: T-Roc, Passat, and Up. Similarly, the transmission variable describes the three types of gearboxes available on the automobile, which are Manual, Semi-Auto, and Automatic, and the variable fuelType indicates the four types of fuel the car uses, i.e. Diesel, Petrol, Other, and Hybrid. Whereas, the numerical variable price reflects the car's price in 1000 GBP (i.e., the British Pound). The variable mileage represents the total distance travelled in 1000 miles. The mpg variable stands for miles per gallon and estimates the distance a car can drive in miles with one gallon (imperial) of fuel. The variable tax is the amount of the yearly tax (Vehicle Excise Duty) paid for the car. The engineSize variable indicates the engine size in litres.

The data set's quality is good as there are no missing values, and the whole data set is used for statistical analysis in this report. As a part of data preparation for statistical analysis, new variables are created, i.e., variable age by calculating the years from first registered date to 2020, and the mpg variable unit is converted into l/(100km) (litres per 100 km).

2.2 Project objective

In this project, the price or log-transformed price log price is considered as the response variables, and the remaining features are considered as exploratory variables for initial regression analysis.

The main objective of the project is to find the best set of features that impacts the price of the car. An optimal regression model is developed to estimate the price based on the best-selected features. In this process, a brief statistical analysis is performed.

The assumptions of the linear model for generated regression models having all the features as covariates and the raw price values and the log-transformed price values as response variables are verified. Model diagnostics are performed on both the models using residual versus fitted plots and QQ plots. Further, The best subset of features is determined using the AIC or Mallow's Cp. Based on the subset of features with minimum AIC value, we construct the best model. And based on the model created, the regression coefficients, are estimated along with their statistical significance, confidence interval, and the goodness of fitness.

3 Statistical methods

In this section, statistical methods are presented which are used for analysing the given sample data set. For all calculations and visualizations, the software R (R Development Core Team, 2020) version 4.1.2, `olsrr` library (Aravind, H., 2020) are used.

3.1 Linear regression

3.1.1 Multiple linear regression

Multiple linear regression is a statistical approach that models the relationship between the response variable and set of covariates. Let y_i be the continuous response variable, with $i = 1, .., n$ observations and x_{ij} is the value of the jth covariate, where $j = 1, ..., k$, for the ith observation. For every single observation the model is defined as

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

where β_0 is the intercept, β_i, $i = 1, ..., k$ are the regression coefficients and, ϵ_i is the error variable reflecting the difference between the observed and predicted values of ith observation (Fahrmeir et al., 2013, p.26).

3.1.2 Assumptions of linear model

In order to perform a multiple linear regression, the sample data must meet the linear model's assumptions. Model diagnostic tools such as the residual versus fitted plot and the QQ plot are used to verify these assumptions.

1. **Linear relationship:** The individual covariates and the response variables should have a linear relationship. The residual versus fitted plot can be used to determine the degree of linearity of the variables. The $x - axis$ shows the fitted values, while the $y - axis$ shows the residuals. Along the reference line that passes through zero residual, the data points should be evenly distributed horizontally (Fahrmeir et al., 2013, p.78).
2. **Homoscedastic error variances:** The variance of residuals or the error term ϵ_i should be constant across the observations, known as homoscedastic error variance. This can be checked using residual versus fitted charts. In real-world circumstances, we rely on regression parameter estimators (Fahrmeir et al., 2013, p.78).
3. **Independent and identically distributed:** All of the observations should have the identical distribution and should not be dependent on one another. Random sampling ensures this assumption in real-world data scenarios (Fahrmeir et al., 2013, p.79).
4. **No multi-collinearity:** Any dependency relationship between the covariates produces imprecise regression parameter estimation in a linear model. The Variation Inflation Factor(VIF) is used to check this. VIF method is discussed in later section (Fahrmeir et al., 2013, p.158).
5. **Normality:** We assume that errors $\epsilon_i, i = 1, ..., n$ for n observations are approximately normally distributed $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, for the computation of confidence intervals and hypothesis tests. QQ plot helps to check the normality of residuals where theoretical quantiles of standard normal distribution on the $x - axis$ and standardized residuals of regression on the $y - axis$. When all the data are displayed and generally follow a straight line, it is presumed that residuals are normally distributed (Fahrmeir et al., 2013, p.80).

3.2 QQ plot

QQ plot is used to provide a comparison between the sample distribution against the Normal distribution. In here, the dots represent the dependent variable and the line represents the Normal distribution.

In Figure 5 in Appendix, the y-intercept and the slope of the reference line are the sample mean and standard deviation respectively. The plotting points for are used to estimate theoretical quantiles, x_i, $i = 1, .., n$ points, corresponding to each sorted sample quantile, y_i. In idealized QQ plot, all the points closely follow a linear trend which suggests that the underlying random variable is plausibly normally distributed (Hay-Jahans, 2019).

3.3 Model definition

3.3.1 Design matrix

In multiple linear regression, to estimate the unknown regression coefficient β we have data of y_i and $x_i = (1, x_{i1}, ..., x_{ik})', i = 1, ..., n$ observations and for every observation, the following equation is used:

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

Let y is a n-dimensional column vector, X is a $n \times (k+1)$ design matrix and ϵ is a n-dimensional column vector of error terms then,

$$y = \begin{pmatrix} y_1 \\ . \\ . \\ . \\ y_n \end{pmatrix}, \ X = \begin{pmatrix} 1 & x_{11} & . & . & . & x_{1k} \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ 1 & x_{n1} & . & . & . & x_{nk} \end{pmatrix} = \begin{pmatrix} x_1' \\ . \\ . \\ . \\ x_n' \end{pmatrix} \text{and } \epsilon = \begin{pmatrix} \epsilon_1 \\ . \\ . \\ . \\ \epsilon_n \end{pmatrix}$$

The above equation can be summarized as $y = X\beta + \epsilon$. The X has a full column rank $rk(X) = k + 1 = p$ which means that the columns of the matrix X are linearly independent. The other property is, n number of observation should be equal or greater than the p. (Fahrmeir et al., 2013, p.75).

3.3.2 Model estimation and residuals

The method of least squares is used to derive the procedure for estimating regression coefficients for multiple linear models. The unknown regression coefficients are calculated using the least squares technique by minimizing the sum of the squared deviations between the real variable y_i and the estimated response variable $x_i'\beta$ of $i = 1, ..., n$

observations. The least squares (LS) equation represented as:

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

The unique solution equation getting the ordinary least squares estimator,

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

where X is the design matrix of covariates (Fahrmeir et al., 2013, p.105).

The residual denoted by $\hat{\epsilon_i}$, where $i = 1, ..., n$ for n observations is the estimated error calculated by deviation between the true value y_i and the estimated value $\hat{y_i}$, then $\hat{\epsilon_i} = y_i - \hat{y_i} = y_i - x_i'\hat{\beta}$. The vector of the residual is defined by $\hat{\epsilon_i} = (\epsilon_1, ..., \epsilon_n)'$, can be written as $\hat{\epsilon_i} = y_i - X\hat{\beta}$ (Fahrmeir et al., 2013, p.77).

3.3.3 Variable transformation and dummy encoding

The method for changing the measurement scale of a variable is known as variable transformation. Sample data for a linear model should meet the homoscedastic error variances assumption. Some observations may cause heteroscedastic errors, and to diagnose this situation linear transformation has been done (Fahrmeir et al., 2013, p.187).

Dummy encoding

In a sample, some explanatory variables or covariates can be categorical. Categorical variables are coded as dummy variables in the model by adding new covariates. Variables are coded with a '1' for attribute indication and a '0' otherwise. Let $x \in 1, ..., c$ categorize in a variable with c, $(c - 1)$ dummy variables are generated.

$x_{i,1} = \begin{cases} 1 & x_i = 1, \\ 0, & \text{otherwise} \end{cases} \quad . . . \quad x_{i,c-1} = \begin{cases} 1 & x_i = c-1, \\ 0, & \text{otherwise} \end{cases}$, for $i = 1, ..., n$ observations, then

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

The dummy variable for category c is called the reference category. The estimation effect can be interpreted by reference category for direct comparison (Fahrmeir et al., 2013, p.97).

3.4 Best subset selection

The sample data set for multiple linear regression may contain a significant number of explanatory variables. Some factors may have a minor impact on the predicted response variable, but they may necessitate a significant amount of calculation time when calculating regression coefficients. An effective regression model can be developed by selecting only the optimal subset of variables. By combining all of the covariates and using the least square estimator for k number of covariates, the $2^k - 1$ number of regression models may be constructed. AIC and Mallow's Cp are two of the model selection measurement criteria. (Yanjun, Wang. and Qun, Liu., 2006).

3.4.1 Akaike information criterion (AIC)

The Akaike information criterion (AIC) is widely used for the model selection. The comparison of AIC scores of several possible models has been done to select the best fit model for the data. For one or more fitted model objects, a log-likelihood value is generated, and the model with the lowest AIC value is considered the best model. The AIC of any set of covariates can be represented by the following equation:

$$AIC = -2.\ell(\hat{\beta_M}, \hat{\sigma^2}) + 2(|M| + 1).$$

Where $\ell(\hat{\beta_k}, \hat{\sigma^2})$ is the maximum value of log-likelihood and M represents the subset of covariates from $M \subset \{1, 2, 3, 4....k\}$. The maximum likelihood was chosen because the log-likelihood is a measure of how likely it is to see one's observed data reach its maximum value for $\beta = \hat{\beta}$. The lower the value of AIC, the better the fit of the model. In case of two models with the same AIC value we select the model with less number of parameters (Fahrmeir et al., 2013, p.148).

3.4.2 Mallow's Cp

The Mallow's Cp is another way to select the best covariates for the model. This method use the following formula to asses the model,

$$C_p = \frac{\sum_{i=1}^{n} (y_i - \hat{y_{iM}})^2}{\hat{\sigma^2}} - n + 2|M|.$$

Here, n is the number of observations and M is the subset of covariates. $\hat{y_{iM}}$ is the estimated value with M covariates and $\hat{\sigma^2}$is the estimated error variance. The lower the value of C_p, the better the model. So, model with the lowest C_p has the best covariates for the model (Fahrmeir et al., 2013, p.148).

3.5 Tests of Liner Model

3.5.1 T-test

A t-test is used in multiple regression to determine the significance of the regression parameter of individual covariate. This is accomplished by providing null and alternative hypotheses for each covariate and calculating the test statistic for each.

Let H_0, H_1 be the null and alternative hypotheises and β_j, $j = 1...,k$ be the regression parameters of k covariates. Then,

$$H_0 : \beta_j = 0 \quad \text{versus} \quad H_1 : \beta_j \neq 0$$

The test statistic is called as "t statistic" or "t value" t_j is obtained by

$$t_j = \frac{\hat{\beta_j}}{se_j}, \text{ where } se_j = \sqrt{Var\hat{(}\beta_j)} \text{ is the standard error of } \beta_j$$

The derived t statistic is compared to the critical value for the null hypothesis rejection region as $(1-\alpha)$ quantile from a conventional t distribution with $n-k$ degrees of freedom. The number of observations is n, and the covariates are k. We reject the null hypothesis when the observed t_j is greater than the $t_{1-\alpha/2}(n-k)$ value. We also calculate the p-value for the t statistic and reject the null hypothesis if the resultant p values are lower than the significance threshold.

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

3.5.2 F-test

With a more general hypothesis, the F test can be used in multiple linear regression to test the linear connection between the response variable and any of the covariates. Let H_0,H_1 be the null and alternative hypotheises and β_j, $j = 1, ..., k$ be regression parameters of covariates then,

$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$ versus $H_1 : \beta_j \neq 0,\ j$ be the atleast one $j = 1, ..., k$ covariates

The residual sum of sqaure $SSE_{H_0} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ under the null hypothesis model and full model SSE for test statistic. The difference is stated as $\Delta SSE = SSE_{H_0} - SSE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2$ and F statistic value can be calculated as

$$F = \frac{n-p}{k} \frac{\Delta SSE}{SSE_{H_0} - \Delta SSE}$$

Let α be the considered significance level. If the test statistic F is larger than $F_{k,n-p}(1-\alpha)$ the $(1-\alpha)$ quantile of the corresponding F-distribution with k covariates and $n-p$ degree of freedom, we reject the null hypothesis (Fahrmeir et al., 2013, p.133).

3.5.3 Confidence interval

In regression analysis, a confidence interval, also known as a confidence area, explains how the derived regression parameter value is likely to fall within a range of preset values. For a single parameter $\beta_j, j = 1, ..., k$, or a confidence ellipsoid for a subvector β_1 of $beta$, a confidence interval or confidence ellipsoid can be created. In the case of t value, it is determined using the test statistic t_j.

The probability of rejecting null hypothesis H_0 when it is equal to significance level α is $P(|t_j| > t_{n-p}(1-\alpha/2)) = \alpha$. The probability of not rejecting null hypothesis H_0 when it is true in nature is $P(|t_j| < t_{n-p}(1-\alpha/2)) = \ P(|(\hat{\beta}_j - \beta_j)/se_j| < t_{n-p}(1-\alpha/2)) = 1-\alpha$. The following equation can be rewritten as $P(\hat{\beta}_j - t_{n-p}(1-\alpha/2) \cdot se_j isthe < \beta_j < \hat{\beta}_j + t_{n-p}(1-\alpha/2) \cdot se_j) = 1-\alpha$. The $(1-\alpha)$ confidence interval for β_j can be obtained as

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

where $\hat{\beta}_j$ is estimated regression parameter, $t_{n-p}(1-\alpha/2)$ is value from t-distribution at $n-p$ degree of freedom, α is the considered significance level and se_j is the standard deviation of the estimator (Fahrmeir et al., 2013, p.136).

3.5.4 Goodness of fit

The coefficient of determination, or R^2, is used to assess the regression model. The value of R^2 is used to interpret the goodness of fit. It is calculated using the residual sum of squares as

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

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

The corrected coefficient of determination, or adjusted R^2, is used to calculate the more precise R^2 value in a regression model with many covariates. It calculates the ideal R^2 value by taking into account the number of covariates in the regression model. (Fahrmeir et al., 2013, p.146).

$$R^2_{adj} = 1 - \frac{(n-1)}{(n-p)}(1 - R^2), \text{for } n \text{ observations and } p \text{ covariates}$$

If the R^2 or R^2_{adj} is closer to 1 it is considered that data is a good fit for the model. When the value of R^2 is close to zero it implies larger value for residuals and the data is considered to be a poor fit to the model (Fahrmeir et al., 2013, p.113).

3.5.5 Variation inflation factor (VIF)

In the linear model assumption, we define that there should be no multi-collinearity. To verify this assumption the VIF is used. Let there is a estimated regression coefficient β_j then the formula is to calculate the VIF is as follows,

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

Here, the R_j is the coefficient of determination of j^{th} covariate and the VIF_j value is the factor by which the β_j is inflated by the existence of correlation among the predictor variables in the model. The factor value 1 shows that there is no multi-collinearity, factor value more than 4 is case of concern and factor value 10 or more shows that there is serious multi-collinearity (PennState, 2018).

4 Statistical analysis

In this section, the statistical methods explained above are applied to the data, and the results of the analysis are interpreted using software (R Development Core Team, 2020). For this report, the significance level α value is 0.05.

4.1 Descriptive analysis

Firstly, a descriptive analysis is performed on the given dataset which gives a general overview. The sample dataset contains 2532 observations with nine variables and, without any missing value. Table 1 in the Appendix shows a descriptive overview of quantitative variables, such as the price variable having a minimum value of 1495 GBP and an average value of 15445 GBP. The average mileage of the sample dataset is 21021 miles, whereas the minimum value is 1 mile, and the maximum is 176000 miles. The standard deviation of miles per gallon(mpg) is 14.26, and the maximum value is 166. The tax paid for the cars ranges from 0 GBP to 265 GBP.

4.1.1 Data preparation

As the first step towards the data preparation, the variable miles per gallon (mpg) is converted into the liters per 100 km. A new variable name LP100Km is calculated by dividing 282.48 by the values of the mpg. The variable year is also transformed into the new variable name age. This variable is calculated by subtracting the year variable by 2020. Both of these transformed variables are used for the analysis. The variable price is used as a response variable in the task, this variable is also transformed into the log price and the linear regression model is tested on both transformed and not transformed variables.

4.1.2 The optimal response variable

Then, we perform two individual linear regression models. One models is based on price as the response variable and log price in another. All remaining variables are considered as exploratory variables in these models, and the categorical variables among those dummy encoded automatically. To choose the optimal model, we first verify the assumptions of the linear model. The collected sample has generated from the random collection, so we assume our data is independent and identically distributed, and there is no-multicollinearity between the variables is observed.

From Figures 1 and 2 in Appendix, residual versus fitted plots, we observe residuals in both the models are plotted horizontally along the horizontal line at zero that matches the condition for linearity. The same residual versus fitted plot also helps to check the assumption of homoscedasticity error variance. When compared between models, the figure 2 price model shows consistent error variance across all the observations since its residuals are almost in line with the horizontal line at zero residual error and some with minimal distance. Figure 1 price model shows bit deviation for some observations from the horizontal line at zero, indicating heteroscedastic error variance.

The normality assumption is checked by the QQ plot from Figures 3 and 4 in Appendix. Figure 4 shows that the log price model shows normality in the plot since almost all the observations plotted as a straight line. On the other hand, the price model shows a bit of deviation for some observations. Overall, we observe that model with log price as a response variable is more in line with the assumptions of the linear model and considered it as an optimal response variable in further analysis.

4.2 Best subset of covariates

Furthmore, we obtain the best subset of covariates for construction of the optimal linear model. The `olsrr` library (Aravind, H., 2020) is used to select the best subset of covariates from a linear model. As there are eight covariates in the model, it calculates the AIC, measuring values for 255 linear models generated by multiple combinations of covariates. The linear model with model, Age, mileage, LP100Km, fuelType, engineSize, tax, and transmission combination of covariates has the minimum value of AIC equal to -3664.491. This shows that the above mentioned covariates are the best set of covariates for the model. So, here the minimum AIC value for the best subset model is with all the available covriates.

4.3 Best linear model

Log price as a response variable and the best subset of covariates obtained using AIC value, used as exploratory variables for the best linear model construction. The model with selected covariates is as follows,

$$
\begin{gathered}
logprice_i = 9.65 + 0.11 \cdot model(T - Roc)_i - 0.56 \cdot model(Up)_i - 0.009 \cdot age_i \\
- 5.71e - 06 \cdot mileage_i + 0.03 \cdot LP100Km_i + 0.43 \cdot fuelType(Hybrid)_i \\
+ 0.07 \cdot fuelType(Petrol)_i + 0.07 \cdot fuelType(other)_i \\
+ 0.17 \cdot enginesize_i - 4.17e - 04 \cdot tax_i \\
- 0.11 \cdot transmission(Manual)_i - 1.97e - 04 \cdot transmission(Semi - Auto)_i
\end{gathered}
$$

Now, we check the linear model assumptions. First, the linearity assumption is validated by the figure 2 in Appendix. The observations are scattered horizontally alongside the reference line at zero residuals, satisfying the linearity assumption. Similarly, it also represents that residual variance is equal to zero. Most of the plotted observations lie on the reference line satisfying the homoscedastic error variance assumption. The QQ plot in figure 4 in Appendix characterizes that all the data points illustrated as almost straight lines specifying the standardized residuals are normally distributed. The VIF values are shown in table 2 in Appendix. This represents that there is no multicollinearity between the covariates. Although, the variable model has the highest value of VIF 6.08, which might be a concern for the model but the VIF value is not near to the threshold, which is 10. Hence, the model satisfies all the required assumption for the linear model.

The results in table 1 present the output of the linear model with estimated regression parameters, test statistics, significance level, and confidence interval. The result represents that the variables model T-roc, LP100Km, fuel Type Hybrid, fuel Type Other, fuel Type Petrol,and engine size have a positive relation with response variable log price having estimated coefficients as 0.11, 0.03, 0.43, 0.07, 0.07, and 0.17 respectively. On the other hand, model Up, age, mileage, tax, transmission with manual, and transmission with semi-auto variables has negative relation with log price. For the estimated coefficients, cars with fuel type hybrid have the 0.43 times highest impact on the log price.

The null hypothesis estimates that the values of the coefficients are equal to zero, and the alternative hypothesis, that there is significant difference from zero. We reject the null hypothesis as the test statistic's p-values for the degree of freedom 2519 is less than the significance level of 0.05 for model T-Roc, model Up, age, mileage, LP100Km, fuel type hybrid, petrol and others, engine size, tax, and transmission manual covariates.

Table 1: Summary of results from best estimated linear model

	Estimate	Std. Error	t value	p value	CI(2.5%-97.5%)
(Intercept)	9.653	0.0302	318.68	<0.0001	9.59 - 9.71
modelT-Roc	0.117	0.0075	14.858	<0.0001	0.009 - 0.12
modelUp	-0.568	0.0106	-53.564	<0.0001	-0.58 - -0.54
age	-0.093	0.002	-44.746	<0.0001	-0.09 - -0.089
mileage	-6e10^{-6}	0.00	-36.579	<0.0001	0-0
Lp100KM	0.0339	0.00378	8.967	<0.0001	0.026-0.04
fuelTypeHybrid	0.434	0.018	24.36	<0.0001	0.39 - 0.46
fuelTypeOther	0.0718	0.031	2.368	0.018	0.012 - 0.13
fuelTypePetrol	0.0762	0.0098	7.751	<0.0001	0.05 - 0.09
engineSize	0.1774	0.0128	13.905	<0.0001	0.15 - 0.20
transmissionManual	-0.1198	0.00945	-12.825	<0.0001	-0.13 - -0.10
transmissionSemi-Auto	-1.97e^{-4}	0.00941	-0.021	0.983	-0.01 - 0.01
tax	-4.17e^{-4}	-6.15e^{-5}	-6.788	<0.0001	0 - 0
Observations	2532				
R^2	0.9544				
Adjusted R^2	0.9542				

The estimated regression parameters of these variables are significantly different from zero and impact the price variable. In contrast, we do not reject the null hypothesis for the transmission semi-auto variable because the obtained p-value is 0.983 respectively. This concludes that transmission semi-auto has an estimated regression parameter equal to zero and does not impact the price of the cars in its predictions.

The confidence interval for the linear model is 95 percent confidence level ($\alpha = 0.05$) for the hypothesis test. This interprets that when a value of covariate is provided, results in the population mean of the response variable lies in the estimated range of 95 percent confidence. For instance, let's take the fuel type as a hybrid then the regression model predicts that the average log price ranges from 0.39 to 0.46 with 95 percent confidence. The output of the regression model provides the adjusted R^2 results as 0.954, is closer to 1 concluding that the data provided had a good fit for the model.

5 Summary

The data set examined in this report was provided by the instructors of the course Introductory Case Studies at TU Dortmund University in the summer semester of 2022. The data set is extracted from the used car platform Exchange and Mart (Exchange Enterprises, Newsquest Media Group, 2022) in the United Kingdom for the year 2020. There are total of 2532 observations with nine variables price, year, model, mileage, miles per gallon(mpg), fuel type, engine size, tax, and transmission.

In the data preparation, two new variables, LP100Km, and age created. Also, the variable price was transformed into the log price for better results. The goal is to find the best set of features that impacts the price of the car. An optimal regression model was developed to estimate the price based on the best-selected features.

In the analysis, the response variable price is log-transformed as the optimal response variable. The set of explanatory variables obtained by the best subset selection criteria has a minimum AIC value of -3665.49. This results in the selection of all the available covariates of data set as the best explanatory variables.

An optimal regression model was constructed using the log price as the response variable and best set of covariates. The results show that the fuel type hybrid has a 0.43 times impact on price, and cars age has a -0.09 times impact on the price of cars. The confidence interval was also determined for each covariate. The goodness of the fit by provided dataset for the constructed linear model was 0.954 with 2519 degree of freedom.

The size of the sample was small, so there is the possibility that the price of the cars might be impacted by other factors as well. This opens a scope for further investigation with the larger data set.

Bibliography

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

Exchange Enterprises, Newsquest Media Group. Exchange and mart, 2022. URL `https://www.exchangeandmart.co.uk/`. (visited on 20th June 2022).

L. Fahrmeir, T. Kneib, S. Lang, and B.D. Marx. *Regression: Models, Methods and Applications*. Springer Berlin Heidelberg, 2013. ISBN 9783662638811.

Christopher Hay-Jahans. *R Companion to Elementary Applied Statistics*. Taylor and Francis Group, 2019. ISBN 9781138329164.

Kaggle Inc. kaggle, 2021. URL `https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes?select=vw.csv`. (visited on 19th June 2022).

PennState. Detecting multicollinearity using variance inflation factors, 2018. URL `https://online.stat.psu.edu/stat462/node/180/`. (visited on 19th June 2022).

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

Volkswagen AG. Volkswagen, 2022. URL `https://www.volkswagen.de/`. (visited on 18th June 2022).

Yanjun, Wang. and Qun, Liu. Comparison of akaike information criterion (aic) and bayesian information criterion (bic) in selection of stock–recruitment relationships. *Fisheries Research*, 77(2):220–225, 2006. ISSN 0165-7836. doi: https://doi.org/10.1016/j.fishres.2005.08.011.

Appendix

A Additional figures and tables

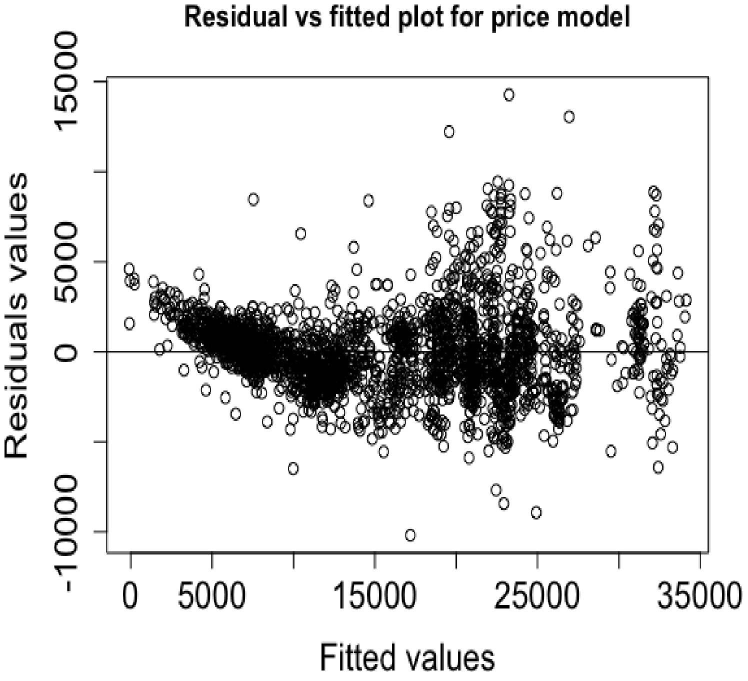

Figure 1: Residual vs fitted plot: Price as response variable

Table 1: Descriptive analysis of the numerical variables

	min	max	mean	median	sd
price	1495	40999	15445	13986	7890.26
mileage	1	176000	21021	12095	24981.2
mpg	32.50	166.00	56.35	54.30	14.26
engineSize	1.00	2.00	1.47	1.50	0.43
tax	0.00	265.00	105.3	145.00	58.85

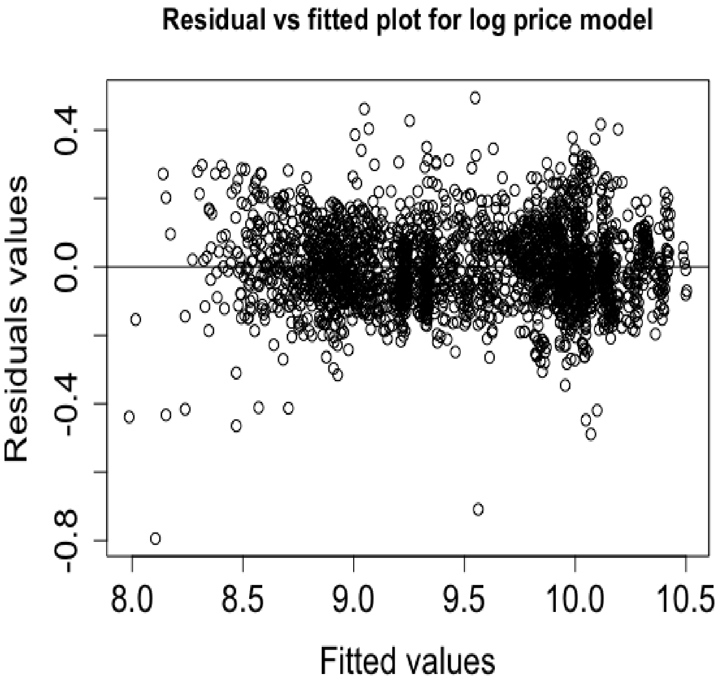

Figure 2: Residual vs fitted plot: Log price as response variable with best subset of covariates

Table 2: VIF values of each variable

model	age	mileage	LP100Km	fuelType	engineSize	tax	transmission
6.08	3.22	2.84	3.25	5.24	5.53	2.42	1.74

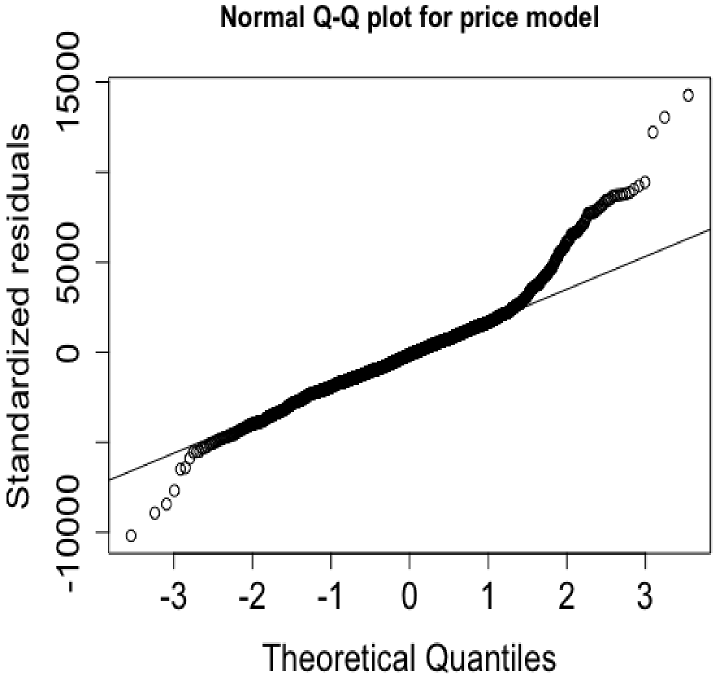

Figure 3: QQ plots: Price as response variable

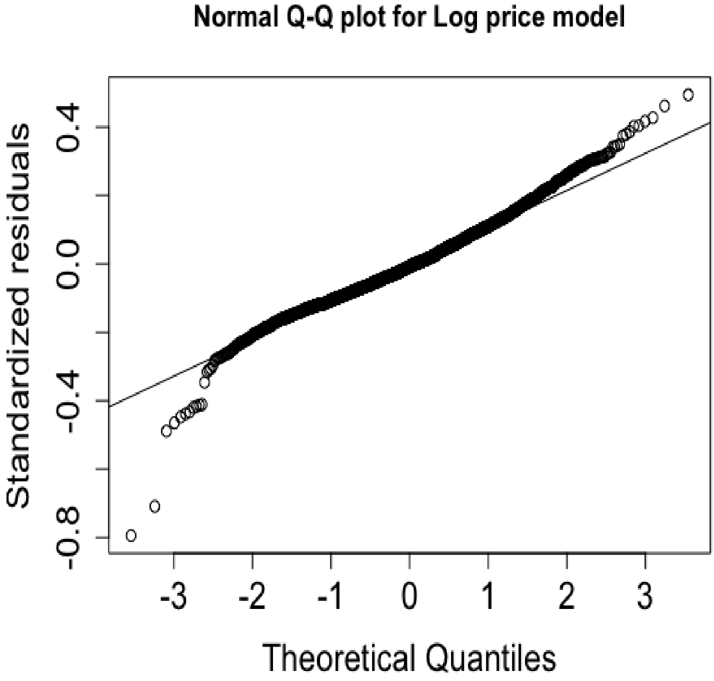

Figure 4: QQ plots: Log price as response variable

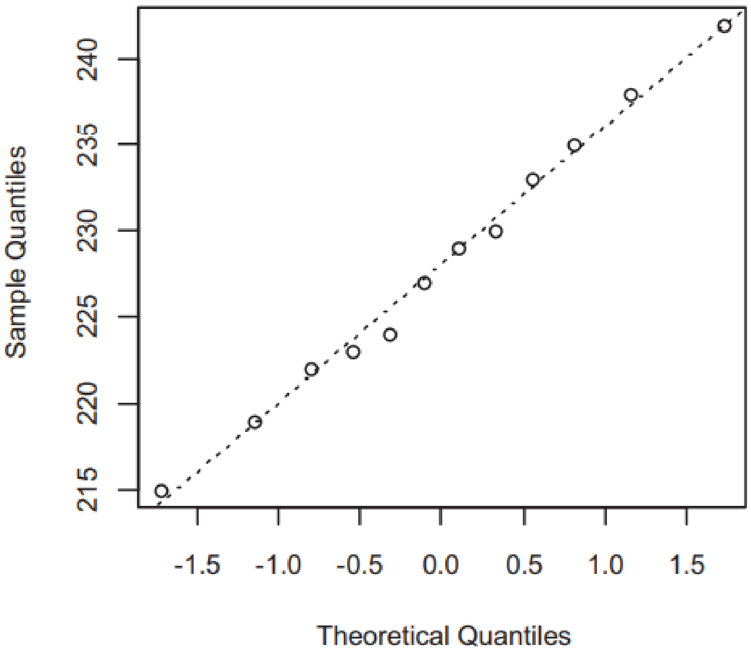

Figure 5: QQ-plot of the sample data.