TU Dortmund

Introductory Case Studies

Project 3: Regression Analysis

Lecturers:

Prof. Dr. Sonja Kuhnt

Dr. Birte Hellwig

Dr. Paul Wiemann

M. Sc. Hendrik Dohme

Author: Amirreza Khamehchin Khiabani

Group number: 5

Group members: Amirreza Khamehchin Khiabani, Rama Kassoumeh, Elchin Latifli, Shubham Gupta

January 28, 2022

Contents

1 Introduction

For calculation of the price of a used car, several factors need to be concerned by the sellers and the customers. According to Edmunds.com, an online automotive review site, each year about 40 million used automobiles are traded. In this case, some characteristics such as mileage, conditions, and so on are regarded. Some customers prefer low mileage, but others might select the factors of the age of vehicle or fuel options (Joe D'Allegro).

This report is closely related to consideration of the important factors for calculation of the used cars in the United Kingdom dealt on a used car platform (Exchange and Mart) in 2020. The purpose of the report is to train and assess the above data set and build linear regression models for the used car price in the United Kingdom. Several statistical methods are applied for this model construction, mostly focusing on statistical linear regression and the best model choice.

In Section 2, an overview concerning the presented data set is provided and the quality of the data is discussed. Furthermore, all tasks requested for this report are argued. The methods that would solve these tasks are debated in this section.

Section 3 comprises different parts. The techniques employed for the statistical analysis are introduced. Here, the multiple linear regression model and its properties are presented in addition to its requirements and assumptions that have to be fulfilled for these methods. Moreover, the concept of the AIC and BIC, their properties and interpretation of them are mathematically described. Furthermore, the definition of the confidence interval in multiple linear regression, as well as coefficient of determination, is introduced. In the statistical analysis Section, after data preprocessing and adding or modifying variables, all these statistical methods are practised to present the results. The concluding section includes a summary of the findings, a discussion thereof and an outlook on further possible analyses.

2 Problem statement

The data is provided by the website (kaggle.com), contains the data of vehicles traded on a used car platform (Exchange and Mart) in the United Kingdom in the year 2020. This data is obtained from a large data set including nine variables:

- *price*: the price of the cars, calculated in 1000 GBP (£).

- *year*: the year that the car for the first time was registered.
- *model*: There are three different model cars of Volkswagen (VW) company.
- *mileage*: the total distance (computed in 1000 miles) the car has been driven.
- *mpg*: the distance (computed in miles) that the car is able to be driven with one gallon (uk) of fuel.
- $fuelType$: the fuel type of the car that it uses.
- $engineSize$: the engin size of the car, measured in liters.
- *tax*: the yearly tax (Vehicle Excise Duty) to be paid for each car.
- *transmission*: the type of the car's gearbox that it has.

There are six numerical variables, four of them are continuous variables including: *price*, *mileage*, *mpg*, and $engineSize$. The two discrete reminded variables are *year* and *tax*. There are three categorical variables in the form of nominal, *model*, $fuelType$, and *transmission*. The *model* is included from Passat, T-Roc, and up. The variable $fuelType$ is divided between 3 different types such as Diesel, Hybrid, and Petrol. For the variable *transmission*, there are three kinds as well: Manual, Semi-Auto, and Automatic.

There is no missing value in the whole data set. Therefore, no statistical method is employed for this problem. There are 438 different used cars in whole give data set. These attributes are independent and there is no correlation among the parameters.

The objectives of the report are, first of all, to preprocess the data by the fuel consumption measures litres per 100 kilometres ($lp100$) instead of the miles per gallon (mpg) and using this new variable. Two new variables are added to the data set, log of the car price as *logprice* and *age* by using variable year. Thereafter, model testing tools to determine if a linear model with *price* or *logprice* one is more in line with the premises of the linear model are applied. Afterwards, the best set of explanatory variables for the price is selected by two different methods. These two methods, namely the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) as the selection criteria, are applied. Subsequently, the best subsets of two different methods are chosen and they are compared. Thereafter, a linear model is built for the select model of BIC method. The coefficients of the model and their statistical significance, as well as confidence intervals for the regression parameters are interpreted. Finally, the goodness of fit of the chosen model is determined to see how model is defined as a linear line.

3 Statistical methods

The following statistical methods and mathematical formulas are used. The statistical software R (Version 4.1.0, R Development Core Team) has been used for analysis with applying two extra packages from R, including *ggplot*2 (Hadley Wickham), *dplyr* (Hadley Wickham, Romain François, Lionel Henry,and Kirill Müller), *leaps*(Thomas Lumley based on Fortran code by Alan Miller), *olsrr*(Taravind Hebbali), *skimr*(Elin Waring et al), and *tidyverse*(Hadley Wickham).

3.1 Multiple Linear Regression Model

There are four assumptions for building linear models. If one or more of them are disobeyed, the final model might become unreliable and incorrect. They are listed as follow (Statology):

- Linearity: A linear relationship between independent and dependent variable needs to be concerned.
- Independency: The residuals are independent, i.e., there is no correlation between residuals.
- Homoscedasticity: The variance of the residuals is considered the same.
- Normality: The distribution of the residuals is Normal.

Multiple linear regression models are the result of a vector of k independent variables or covariates, $X_1, ..., X_k$, on a dependent or response variable y. Here, the covariates can be either continuous or properly coded categorical variables, but the response variable is continuous. The response variable is not a deterministic function $f(x_1, ..., x_k)$ of the covariates, rather this relationship indicates random errors. This indicates this formula:

$$f(X_1, ..., X_k) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k.$$

The linear function f is called the systematic component of the model. To model a categorical covariate, $X \in 1, ..., c$ with c categories, category c can be considered as a reference. Afterwards, c - 1 dummy variables can be described and then included in the

model, i.e.

$$X_{i1} = \begin{cases} 1 & X_i = 1 \\ 0 & otherwise \end{cases} \qquad ... \qquad X_{i,c-1} = \begin{cases} 1 & X_i = c-1 \\ 0 & otherwise \end{cases}$$

The parameters $\beta_0, ..., \beta_k$ are unknown. There is another fundamental assumption of the linear model, additivity of errors formulated as bellow:

$$y = f(\boldsymbol{X}) + \epsilon = \boldsymbol{X'\beta} + \epsilon.$$

The covariates and the parameters can be mixed into separate p = k + 1 dimensional vectors, $\boldsymbol{X} = (1, X_1, ..., X_k)'$ and $\boldsymbol{\beta} = (\beta_0, ..., \beta_k)'$, the systematic component can be expressed as a vector product. For estimation the parameters, data $(y_i, x_{i1}, ..., x_{ik})$ is collected where $i = 1, ...n$. The vectors y and and the design matrix X are defined as follows:

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

and

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}.$$

In this model, the errors are normally distributed, with the value of zero for mean and constant variance (i.e. homoscedastic errors), $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. The design matrix $\boldsymbol{X}$ is presumed to have full column rank, indicating that all columns are linearly independent and $rk(X) = k + 1 = p$, where p is the number of the parameters. Hence, n equations can be formed as

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

For estimation of the unknown parameters $\boldsymbol{\beta}$ method of least squares applied. The estimates $\hat{\boldsymbol{\beta}}$ are the minimizers of the sum of squared deviations,

$$LS(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}).$$

Assigning the derivative of this equation concerning $\boldsymbol{\beta}$ equal to zero results to the unique solution of the least-squares estimator, i.e.,

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

Estimatation of the error variance is obtained by using

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{n - p'},$$

where $\hat{\boldsymbol{\epsilon}}'$ represents the residuals of the model and $\hat{\epsilon}_i = y_i - \hat{y}_i$, as $\hat{y}_i$ for $i = 1, ..., n$ are its estimated values (Ludwig Fahrmeir et al, p. 74-107).

To test for the significance of a parameter β_j, the hypotheses are $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$, where the estimated t-statistic is obtained as:

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

where $s\hat{e}_j = [\widehat{var(\beta_j)}]^{1/2}$ is the estimated standard error of $\hat{\beta}_j$. the absolute value of the above statistic is compared to the $(1 - \alpha/2)$th quantile of the $t - distribution$ with $n - p$ degrees of freedom. H_0 is rejected if (Ludwig Fahrmeir et al, p. 135):

$$|\hat{t_j}| > t_{1-\alpha/2}(n - p).$$

3.2 Akaike Information Criterion (AIC)

The Akaike information criterion (AIC) is considered one of the most applied chosen models among competing ones. A smaller value of the index indicates the preferred model. The index is defined as follows:

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

where $l(\hat{\boldsymbol{\beta}}_{\boldsymbol{P}}, \hat{\sigma}^2)$ is the maximum log-likelihood from the computation of the model parameters $\hat{\boldsymbol{\beta}}_{\boldsymbol{P}}$ and the variance of residuals with the number of parameter 2(|P|+1). In a linear model when residuals are normally distributed with constant n the formula would

be (Ludwig Fahrmeir et al, p. 148):

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

The smallest value for AIC would considered for the model selection.

3.3 Bayesian Information Criterion (BIC)

The Bayesian information criterion (BIC) is another model choice of linear models. In comparison to AIC, although there is similarity between both models in modelization, BIC has more restriction when the model has more complexity. Therefore, the best models are obtained when using the BIC rather than the AIC. It is formulated as follow:

$$BIC = -2 \ . \ l(\hat{\boldsymbol{\beta}}_{\boldsymbol{P}}, \hat{\sigma}^2) + log(n) \ . \ (|P| + 1).$$

Similar to AIC in case of a linear model with normally distributed residuals, it is formulated by: (Ludwig Fahrmeir et al, p. 149)

$$BIC = n \ . \ log(\hat{\sigma}^2) + log(n) \ . \ (|P| + 1).$$

3.4 Confidence Intervals for β

For the construction of confidence intervals, it is useful to apply the equivalence of the definition of two-sided tests and confidence regions. For each parameter $\beta_i, i = 1, 2, ..., k$, it is essential to consider the premise of normality. Thus, by application of the test statistics, $t_j = (\hat{\beta}_j - dj)/se_j$ with the assumption of $H_0 : \hat{\beta}_j - dj$. The probability of rejecting H_o, i.e. α value, is as bellow:

$$P(|t_j| > t_{n-p}(1 - \alpha/2)) = \alpha.$$

Since the value of rejecting of the null hypothesis is obtained and we can define $1 - \alpha$ as the probability of accepting the null hypothesis and formalized as

$$P(\hat{\beta}_j - t_{n-p}(1 - \alpha/2).se_j < \beta_j < \hat{\beta}_j + t_{n-p}(1 - \alpha/2).se_j) = 1 - \alpha.$$

According to this formula, the confidence interval would be

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

for a large sample size or the assumption of normally distributed (Ludwig Fahrmeir et al, p. 136).

3.5 Coefficient of Determination

The coefficient of determination R^2 for a linear regression model is defined as

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

where y is the mean value of the response variable and $\hat{y}_i$ for $i = 1, ..., n$ are its estimated values as above (Ludwig Fahrmeir et al, p. 115). The coefficient of determination has the value ranged between 0 and 1, i.e., $0 \leq R^2 \leq 1$. As its value is closer to 1, the residual sum of squares is smaller. Therefore, the fit to the data is better. If R^2 is closer to 0, it means that the sum is getting larger and the regression model is poorly fitted. R2 is a measure of the proportion of the variance in the response variable that is predictable from the covariates. For model comparison, the coefficient of determination is of limited value. The corrected coefficient of determination mitigates its weaknesses by containing a correction term in the formula to account for the number of parameters in the model.

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

(Ludwig Fahrmeir et al, p. 147-148).

4 Statistical analysis

In this section, the statistical methods described in the previous part are applied to the granted data set to interpret the results.

4.1 Data Preprocessing

The response variable for the linear regression model is the car price. In the first task, it is required to compute the log of the car price and it is stored as a new random variable called 'logprice'.

In the second step, since in the data set, the fuel consumption measure is miles per gallon (recorded as 'mpg'), a new variable is defined by using the formula as bellow;

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

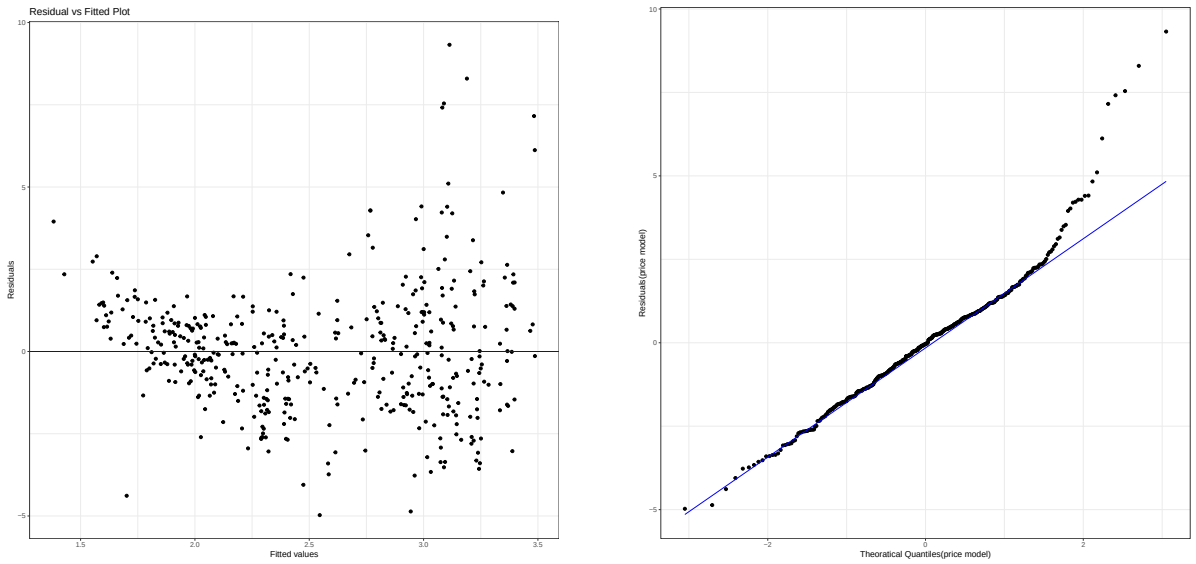
where 'lp100' is the new variable for the liters per 100 kilometers consuming of fuel. This new column is added to data set for further assumptions.

In the final part of this task, by applying variable 'year', current year of this report, and via programming with R, the new variable, called 'age', is represented in data set for the age of each car.

4.2 Applying raw price variable or the log-transformed price

After adding two new variables besides some changes, in this subsection, two different dependable variables are considered to distinguish model assumptions. The result of this section would be more suitable for further processing. All variables are considered for building the linear models.

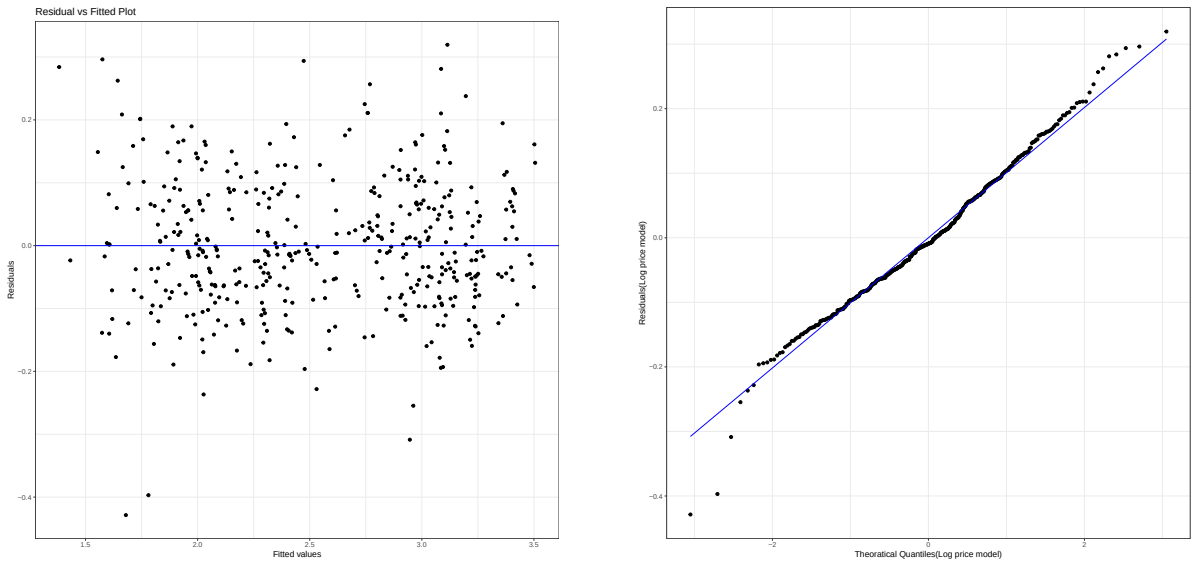
The graph 1 fitted vs. residuals comparisons as well as Q-Q plots for the raw price are shown. It can be concluded that the hypothesis of homoscedasticity for the variance of residuals do not hold and the residuals do not reflect around the zero line. Furthermore, the linearity assumption is disregarded. By reviewing the Q-Q plot for this model, although for the first quantiles, the normality is not violated, for the last quantile, residual points are not normally distributed and the distance between points and line is getting larger.



(a) (b)

Figure 1: Residual vs. fitted and Q-Q plots of raw price

On the other hand, the graph 2 for log price shows a better model according to the assumptions. It can be observed that in the fitted vs. residuals graph, the assumption of homoscedasticity for the variance of residuals are barely concerned and the residuals are approximately mirrored the zero line. The graph 2 for log price shows a better model according to the assumptions. It can be observed that in the fitted vs. residuals graph, the assumption of homoscedasticity for the variance of residuals are barely concerned and the residuals are approximately mirrored the zero line. For normality assumption, by checking the Q-Q plot for this model, for the first quantiles, there is minor violation in the normality. However, it can be concluded that residual points are normally distributed. In conclusion, log price is a more suitable model for further studying in this report.



(a) (b)

Figure 2: Residual vs. fitted and Q-Q plots of log price

4.3 Model Selection by AIC and BIC

In this subsection, the best models are chosen by applying AIC and BIC methods and then compared. Thus, the 'logprice' model is selected and there are 255 possible models, each of which contains a various mixture of variables, and then by the calculation of AIC, the model with the smallest value is selected as the best model. The best subset of variables according to the Akaike Information Criterion would be: $model, age, lp100, fuelType, engineSize, transmission, and'mileage'$ with the value of -704.4099, smallest amount in comparison to other model selections.

Therefore, the Bayesian Information Criterion (BIC) is involved. The lowest value for the linear model would be with the value of -659.3426 and the variable listed as: $model, age, fuelType, engineSize, transmission, andmileage$. As can be seen, both criteria removed the 'tax' as extra dimension. As a result, it is not necessary to apply this variable for construction of the model. On the other hand, BIC penalized the model with deleting $lp100$. Consequently, the Bayesian Information Criterion compare to another model reduce the parameters of the model with extra effort.

4.4 Results of the Linear Regression Model

After selection the model by using BIC, the linear is constructed by the parameters mentioned in the previous section. The table 1 displays the descriptive distribution of the residuals.

Table 1: Descriptive distribution of the residuals

variable	min	1Q	median	3Q	max
value	-0.45031	-0.06866	-0.00957	0.07009	0.32021

The normality assumption of the linear regression is checked using the Q-Q plot in figure 3. The graph shows that the points in the center lie on the line as well as last quantile. Only for the first quantile data points are divergent from the line of checking normality. It can be assumed that residuals are normally distributed according to the Q-Q plot.

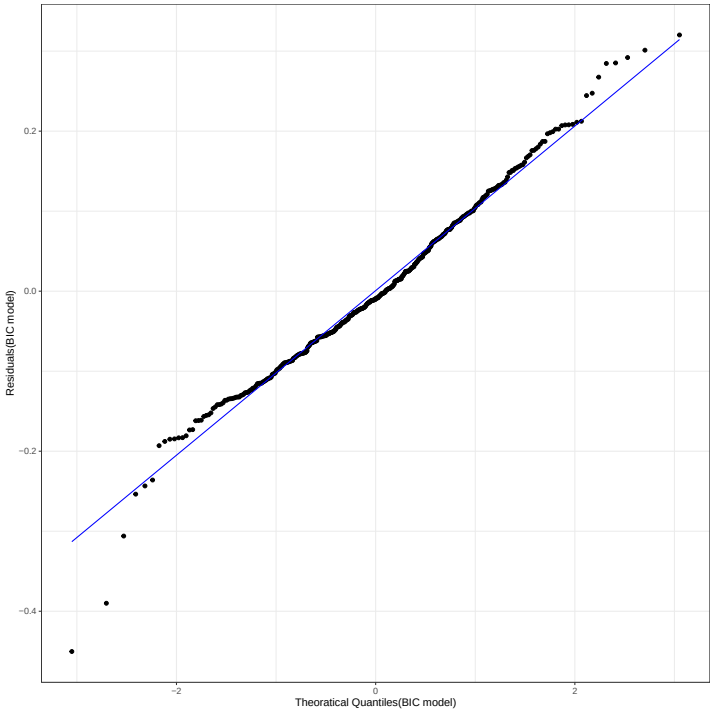


Figure 3: Q-Q plot of the model residuals

Plot 4 illustrates residuals versus fitted values. Since the expected value of the residuals according to the assumptions needs to be zero, approximately it can be seen that although the values of residuals in this linear model are not completely centered to zero,

but the spread of the residuals are formed roughly around the zero line. Therefore, the assumption of homoscedasticity for the variance of residuals seems to be met. Additionally, the residuals bounce randomly around the 0 line and this does not violate the linearity assumption.

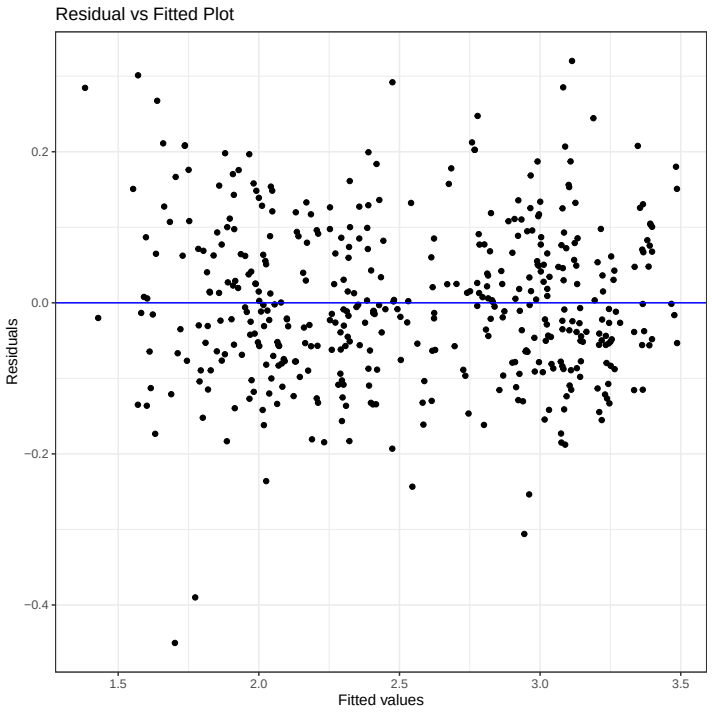


Figure 4: Scatter plot of fitted values vs. residuals

After setting up the model with all the assumptions checked as described above, the Table 2 presented the result of setting the model.

Table 2: Linear regression parameter estimates and their p-values

Covariate	Estimated parameter value	$(Pr(>\|t\|)$
(Intercept)	2.8265478	< 2e-16
modelT-Roc	0.1533899	< 2e-16
modelUp	-0.5180043	< 2e-16
age	-0.0884400	< 2e-16
fuelTypeHybrid	0.4623469	< 2e-16
fuelTypePetrol	0.1227253	1.01e-09
engineSize	0.2865238	< 2e-16
transmissionManual	-0.1220892	2.64e-10
transmissionSemi-Auto	-0.0040519	0.839
mileage	-0.0058255	< 2e-16

For $\alpha = 0.05$, the coefficients of the linear model are almost statistically significant when tested for the null hypothesis, i.e, $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$. The only exception is the dummy variable $'transmissionSemi - Auto'$. In this case, this variable can be omit from the assumption of the model, since it is not statistically significant in the model. Keeping variables with this property might reduce the precision of the model. As can be seen, model type *Passat*, fuel type of *Diesel*, and transmission *Automatic* are taken as the references in their related categorical variables.

While interpreting any estimated coefficient, the values for the other ones are assumed to be constant. According to the trained linear model, it is investigated that chosen coefficient sets its covariates value equal to one and keep the rest covariates equal to zero, then see how it affects the response variable. If the car model as a dummy variable is considered as $T - Roc$ in compare to Up, it can increase the price of the car 0.15 times, while the other model type decreases the price with the value of 0.51.

With the same assumption for setting the values of covariates and coefficients, while both *age* and *mileage* decreases the dependable variable with the values 0.088 and 0.005, it can be concluded that age has more effect on the final price, i.e., by increasing the age of the car, it is anticipated that the price of the car would observe a decline. For the dummy variable $fuelType$ in the case of hybrid or petrol, both increase the car price 0.462 and 0.122 times, respectively. As for the dummy variable *transmission*, the covariate of $transmission\ semi - auto$ is not statistically significant, only *manual* type would decrease the final price of car with the value of 0.122 times if all other covariates would considered constant.

It is interpreted that a 95% confidence interval for β_j means that the interval has the probability of 95% to contain true value of β_j, i.e., in 95% of all samples could be drawn, the confidence interval would cover the value of β_j. The Table 3 represented the result of the confidence intervals for each selected covariates.

As can be seen, each estimated values lie between their confidence intervals. For almost of them, there is no confidence interval such that it contains the value of zero. In that case, the null hypothesis is accepted by this value, i.e., the covarites has no significant influence in the model calculation. There is one confidence interval with this property, and already it is shown that its p-value is higher than α and the null hypothesis is accepted.

Table 3: Confidence intervals of the model coefficients

Covariate	2.5%	97.5%
(Intercept)	2.699831300	2.953264211
modelT-Roc	0.118308988	0.188470886
modelUp	-0.565550146	-0.470458420
age	-0.096098655	-0.080781305
fuelTypeHybrid	0.389239149	0.535454714
fuelTypePetrol	0.084107963	0.161342631
engineSize	0.228906588	0.344141088
transmissionManual	-0.159162009	-0.085016395
transmissionSemi-Auto	-0.043305998	0.035202297
mileage	-0.006437704	-0.005213362

4.5 Model Assessment

The adjusted R-squared value for the linear model is about 0.96. It means it is able to explain around 96% of the total variance in the car price is explained by fitted value in the linear model. The value is statistically close to 1, which can be concluded that with applying BIC and then with selected parameters, the models is flourishing in explaining inputs. There is no divergence from the straight line for higher quantiles indicating that the assumption of normally distributed errors may not be compromised.

5 Summary

The goal of this report is to describe and employ statistical methods for the analysis of a data set represented by the (kaggle.com) for the used cars and their prices by the manufacturer Volkswagen (VW) in the United Kingdom in the year 2020. The data includes nine variables, including price, models, mile age, and so on. There are 438 rows in the data set, for 3 different car models, Passat, T-Roc, and Up. Firstly, some data modifications such as including new variables and defining new measurements for consumption of the fuel are involved.

For answering these tasks, statistical methods are applied with their introduction and assumptions. Firstly, it is determined whether to utilise the raw price variable or the log-transformed price as the response variable in the linear regression analysis. As a result of using model assumption of linear models, Q-Q plots and fitted vs. residuals comparison, it is decided to use log-transformed price for building linear models.

After deciding on the response variable, the best set of explanatory variables are chosen by applying two different methods, called AIC and BIC, for the best subset selection. Their values are concerned and selected variables are compared together. Both method dropped the variable tax, while Bic model removed transformed fuel consumption variable as well. Afterwards, a regression model is built for the dependent variable w.r.t. the BIC, with estimation of 10 coefficients.

As the result of the interpretation of p-values, the coefficient of transmission of Semi-Auto is not statistically significant. Confidence intervals of coefficients are calculated and discussed. By applying model evaluation adjusted coefficient of determination, it is concluded that there is a strong relationship between the exploratory variables and the log price, as a response variable.

It can be concluded that it is practical to conduct statistical methods for building linear models, in this case for the price of the different used cars. For further study, it is useful to consider the different number of samples and other related attributes of selling used cars.

Bibliography

Elin Waring et al. *skimr: Compact and Flexible Summaries of Data*, 2021. URL `https://docs.ropensci.org/skimr/`.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*, 2016. URL `https://ggplot2.tidyverse.org`.

Hadley Wickham. *tidyverse: Easily Install and Load the 'Tidyverse'*, 2021. URL `https://tidyverse.tidyverse.org`.

Hadley Wickham, Romain François, Lionel Henry,and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2021. URL `https://CRAN.R-project.org/package=dplyr`. R package version 1.0.7.

Joe D'Allegro. Factors into the value of your used car. URL `https://www.investopedia.com/articles/investing/090314/just-what-factors-value-your-used-car.asp`. (visited on 25/01/2022).

Ludwig Fahrmeir et al. *Regression: Models, Methods and Applications*. Springer-Verlag Berlin, Heidelberg, Germany, Jan. 2013.

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

Statology. The four assumptions of linear regression. URL `https://www.statology.org/linear-regression-assumptions/`. (visited on 25/01/2022).

Taravind Hebbali. *olsrr: Tools for Building OLS Regression Models*, 2020. URL `https://olsrr.rsquaredacademy.com/`.

Thomas Lumley based on Fortran code by Alan Miller. *Regression subset selection*, 2020. URL `https://cran.r-project.org/web/packages/leaps/index.html`.