TU DORTMUND

INTRODUCTORY CASE STUDIES

Project 3: Regression analysis for price of VW cars

Lecturers:
Prof. Dr. Sonja Kuhnt
Dr. Paul Wiemann
Dr. Birte Hellwig
M. Sc. Hendrik Dohme

Author: Tadeo Hepperle

Group number: 2

Group members: Prem Kant Shekhar, Minjae Ok, Ishan Singh Dhapola, Fyalisia Amanda Putri, Tadeo Hepperle

January 19, 2022

Contents

1 Introduction

Every year millions of cars are bought and sold around the globe. Despite the pandemic, there were 2.62 million cars newly registered in Germany in the year 2020, according to Statista (2022). Almost 10% of Germany's GDP comes from automobile companies and their suppliers and about 930,000 people work in the car industry (DW, 2020). The massive size of this economic sector and its impact on our daily life makes it important to talk about the prices of cars. Purchased for 36,300 EUR on average (Statista, 2020), cars can lose their value pretty quickly. Therefore many private Buyers rather buy used cars. According to Handelsblatt (2017), about 2 out of three newly registered cars are not new. Therefore in this report we will look at the prices of used cars sold on the british car internet platform "Exchange and Mart" (Exchange & Mart, 2022). Our dataset contains prices and predictors such as age, tax, mileage, transmission type, fuel type fuel consumption and engine size for 438 VW cars of the 3 popular models Passat, Up and T-Roc. The goal is to be able to predict the price of a car as accurately as possible using the other variables provided. For this we are using multiple linear regression on the logarithmic price as the variable we want to predict. We found that the logarithmic price can be better modeled using the predictor variables. After conducting best subset selection we found that a regression model containing the six predictor variables car model, age, mileage, transmission type, engine size and fuel type was best at predicting the price of VW cars. More than 96 percent of the variation in the price of a car was rooted in these variables. Overall we found that greater mileage and age reduce a cars price while higher engine size increases it. Hybrid cars sell for much more than Petrol and Diesel cars, while automatic transmission gets you a higher price tag on your car as well. There were also huge differences between the different models. In section 2 the goal of the report is exyplained and we explain the VW cars data set in detail, while the statistic methods around multiple linear regression are explained in section 3. Section 4 holds the results and as a part of that the final model with all its coefficients, that are interpreted properly to explain them to the untrained eye. Finally section 5 summarizes the project and its implications for what to consider when looking for buying or selling a used car.

2 Problem statement

We analyzed a dataset containing prices and other features of 438 cars from the german car brand VW to determine which factors influence the price of a car and to what extent they do so. This makes it possible to predict the value of a car by knowledge about its features.

2.1 Description of the dataset

The dataset used in this report is a sclice of a bigger dataset from kaggle.com and contains data about 438 VW cars, which were originally scraped in 2020 from the british car internet platform "Exchange and Mart" (Exchange & Mart, 2022) (https://www.exchangeandmart.co.uk The slicing was done in a way to only include the VW models "Passat", "T-Roc" and "Up". For each of the 438 cars information on the following variables is provided:

- the **price** a car is selling for in 1000 GBP (£), metric variable
- the **year** a car was first registered, metric variable
- the **mileage** as the total distance a car has been driven in 1000 miles, metric variable
- the **fuel consumption** in mpg, the number of miles a car can drive with one gallon of fuel, metric variable
- the **fuel type** a car uses, categorical variable with 3 levels: "Diesel", "Hybrid", "Petrol"
- the **engine size** as the volume of fuel in liters and air a car can fit in its engines cylinders, metric variable
- the annual **tax**, also known as "Vehicle Excise Duty" that has to be paid for the car annually
- the type of **transmission** a car has, categorical variable with 3 levels: "Manual", "Semi-Auto", "Automatic"
-
- the **model** of the car, categorical variable with 3 levels: "Passat", "T-Roc", "Up"

From those 9 original variables, three more are computed:

- the **logprice** as the natural logarithm of the **price** of a car
- the **fuel consumption** in liters per 100 kilometers, computed from the fuel consumption in mpg as $\frac{282.48}{mpg}$
- the **age** of the car in 2020 when the data was taken, computed as $2020 - year$

From here on, when fuel consumption is mentioned, the liters per 100 kilometers value is meant. The quality of the data seems good, "Exchange and Mart" is a large and well known site and no data is missing. Although we have to rely on the data given to us and cannot check if for example the mileage and tax are truely correct.

2.2 Objectives of the report

The goal of this report is to predict the price of a car as accurately as possible. To achieve this we fit a multitude of linear regression models and using best subset selection will choose as the final model the one with the best indicators, namely the BIC (Bayesian information criterion). Interactions between the 8 predictors or nonliner relationships will not be considered.
Also we will figure out if it makes more sense to predict the price directly through a linear regression model or if it is better to use the logprice as the dependent variable. After making predictions this can be retransformed to normal prices so no information is lost, but it could be that the logarithmic price fits the data better. Sadly the range of car models in the data is quite limited and therefore we will only be able to predict prices of those models later on with our regression models.

3 Statistical Methods

We give a brief overview about the math behind multiple linear regression, how to select the best model in subset selection via certain indicators and how to assess how good a regression model fits the data.

3.1 classical linear regression

Multiple linear regression is a method for predicting a metric variable y on the basis of k metric variables $x_1, \ldots, x_k$. A linear regression model consists of kcoefficients that have to be fitted to the data and can be represented by the following formula:

$$y = \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

y is also called dependant variable or outcome, while x_1 to x_k are called the independent variables or predictors. Since the coefficients $\beta_1, \ldots, \beta_k$ are just multiplied by the predictors $x_1, \ldots, x_k$, a linear regression can only detect linear relationships in the data. In the formula ϵ represents the error term, because it is likely that even the best linear combination of the multiple $\beta_j x_j$ will not add up to the real value of y. But what do we mean by "the best linear combination"? Typically the coefficients in linear regression are estimated by minimizing the residual square sum (RSS). If we have all x_1 to x_k and β_1 to β_k we can estimate y as $\hat{y}$:

$$\hat{y} = \beta_1 x_1 + \cdots + \beta_k x_k$$

The error term ϵ is missing in this equation. $y = \hat{y} + \epsilon$ or $\epsilon = y - \hat{y}$. Suppose we have a data set consisting of n objects i with values y_i and $x_{1,i}$ to $x_{k,i}$. Each object has its own error term ϵ_i, sometimes we overestimate y with $\hat{y}$, sometimes we underestimate it. After fitting the model coefficients by minimizing the RSS, the mean error $(\sum_{i=1}^{n} \epsilon_i)$ will be zero.

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (\epsilon_i)^2$$

To estimate the coefficients, we introduce some matrix notation. Let Y be a $n \times 1$ column vector containing y_1 to y_n. X is a $n \times k$-matrix consisting of n row vectors that contain the values $x_{1,i}$ to $x_{k,i}$ for each object. Then we can define $\mathcal{B}$ as a $k \times 1$ vector containing the coefficients $\beta_1, \ldots, \beta_k$ and we define $\mathcal{E}$ as a $n \times 1$ column vector containing ϵ_1 to ϵ_n. In this setting we can then write the formula as the following:

$$Y = X\mathcal{B} + \mathcal{E}$$

Here the RSS depends on $\mathcal{B}$ can be rewritten as follows (Fahrmeir et al., 2013, p. 105):

$$\begin{aligned} RSS(\mathcal{B}) &= \mathcal{E}^T \mathcal{E} \\ &= (Y - X\mathcal{B})^T (Y - X\mathcal{B}) \\ &= Y^T Y - B^T X^T Y - Y^T X\mathcal{B} + \mathcal{B}^T X^T X\mathcal{B} \\ &= Y^T Y - 2Y^T X\mathcal{B} + \mathcal{B}^T X^T X\mathcal{B} \end{aligned} \tag{1}$$

Taking the derivative with respect to $\mathcal{B}$ now yields:

$$\frac{\partial}{\partial \mathcal{B}} RSS(\mathcal{B}) = -2X^T Y + 2X^T X \mathcal{B}$$

It can be shown that the second derivative is positive and therefore we can find a minimum for $\mathcal{B} = \hat{\mathcal{B}}$ by setting the first derivative to zero (Fahrmeir et al., 2013, p. 106).$\hat{\mathcal{B}}$ represents the least squares estimator for $\hat{\mathcal{B}}$.

$$\begin{aligned} -2X^T Y + 2X^T X \hat{\mathcal{B}} &= 0 \\ 2X^T X \hat{\mathcal{B}} &= 2X^T Y \\ \hat{\mathcal{B}} &= (X^T X)^{-1} X^T Y \end{aligned} \tag{2}$$

So this is how we calculate our estimated coefficients $\hat{\beta}_1, \ldots, \hat{\beta}_k$ as entries in the vector $\hat{\mathcal{B}}$. This classical linear model is missing an intercept though. Setting all predictors to zero would always yield $\hat{y} = 0$, no matter the coefficients. This is obviously not an optimal estimate. To accomodate for this, often a predictor x_0 is introduced that gets a value of 1 assigned for each object in the data. As a part of the predictor matrix X the method described above will then estimate a coefficient $\hat{\beta}_0$, known as the intercept.

A coefficient $\hat{\beta}_i$ can be interpreted in the following way: Holding all other predictor variables constant, an increase of one unit in x_i will on average result in an increase of $\hat{y}$ by$\hat{\beta}_i$. The intercept β_0 represents the expected value of y when $x_1 = \cdots = x_k = 0$. Sometimes logarithmic transformations are applied to the variables before fed into in the linear regression. When the outcome y is actually a proxy for $ln(y_{original})$, than a one unit increase in x_i leads to a $\hat{\beta}_i$ increase in y which means a multiplication of $y_{original}$ by $10^{\hat{\beta}_i}$. For small $\hat{\beta}_i$ this means: a one unit increase in x_i leads to a $100 \cdot \hat{\beta}_i\%$ increase in $y_{original}$.

To assess how good a model is, a statistic R^2 can be calculated. It represents how much variance in y can be explained by $x_1, \ldots, x_k$, or in other words how much variance y shares with $\hat{y}$. Therefore it can be calculated as 1 minus the fraction between RSS and

TSS where TSS is the empirical variance of y multiplied by n (James et al., 2013, p. 234).

$$
\begin{aligned}
R^2 &= 1 - \frac{RSS}{TSS} \\
&= 1 - \frac{RSS}{\sum (y_i - \overline{y})^2} \\
&= 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \overline{y})^2} \\
&= \frac{\sum (\hat{y}_i - \overline{y})^2}{\sum (y_i - \overline{y})^2}
\end{aligned} \tag{3}
$$

It can also be calculated as the squared correlation between the real outcome y and the predicted outcome $\hat{y}$ (Fahrmeir et al., 2013, p. 113).

$$R^2 = r^2_{y\hat{y}}$$

Therefore R^2 can take values between 0 and 1. A greater R^2 means a more accurate prediction. What a good R^2 value is, highly depends on the application though, while in physics often values close to 1 can be found, in "biology, psychology, marketing,and other domains" values below $R^2 \approx 0.10$ are to be expected according to James et al. (2013, p. 70).

A linear regression should only be used if some assumptions are fulfilled. First, the number of observations should be greater than the number of predictors. The residuals should be independent of the predicted values. For this, a scatter plot for $\hat{y}$ vs. ϵ should not show a specific pattern. Their variance should not change as predicted values increase or descrease. This is also known as equal variance or homoscedasticity (Prabhakaran, 2017). Also the correlation between residuals and each predictor should be zero. This however is fulfilled when minimizing the RSS, because otherwise, the predictors "know" something about the residuals, which means they know more about y and are therefore not optimal. In addition to that the residuals should be approximately normally distributed which can be checked by looking at a QQ-plot between quantiles of standardized residuals and the standard-normal distribution (Prabhakaran, 2017).

3.2 dummy coding of categorical variables

The question arises how to use categorical variables as predictors in multiple linear regression. This can be resolved by a coding system. A coding system creates for every categorical variable a number of derived metric variables that can then be used in linear regression. One such coding system is dummy coding. citeXXXXX Assuming we have a categorical variable x_i with $m \geq 2$ levels with labels $l_1, \ldots, l_m$. We can then create a dichotomous variable x_{l_j} for each label $l_j \in \{l_2, \ldots, l_m\}$, such that:

$$x_{l_j} = \begin{cases} 1, & \text{if } x_i = l_j \\ 0, & \text{otherwise} \end{cases}$$

In this way all information about the categorical variable x_i is contained in the variables $x_{l_2}, \ldots, x_{l_m}$. A datapoint with label $x_i = l_1$ can be recognized by having $x_{l_2} = \cdots = x_{l_m} = 0$, l_1 is also called the baseline accoring to James et al. (2013, p. 86). Those newly created numerical variables can then be used in regression analysis as predictors.

3.3 model diagnostics

Assessing how well a linear model fits the data, is aided by some graphical and numerical methods. First, we can look at R^2 as a measure of goodness of fit. Also it is important to check the assumptions of a linear regression. For this some plots can help (Kim, 2015)):

- Q-Q-plot of standardized residuals and a standard normal distribution check if residuals are approximately normally distributed
- Q-Q-plot of y and $\hat{y}$ to check if they follow approximately the same distribution
- scale-location-plot: a scatter plot of $\hat{y}$ on the x-axis and standardized residuals on the y-axis to see if they are evenly distributed along $\hat{y}$.

3.4 best subset selection

The more predictors go into a linear regression model, the more accurate it will perform on the training data and the greater its R^2 value will be. But taking more predictors into account is not always useful. It makes the model harder to interpret and does not highlight which variables are actually important for the outcome. Moreover a lot of times variance in y that can be explained by an additional predictor is already explained by other predictors, such that the gain in variance explanation through R^2 is marginal and could even just be a product of overfitting XXXXjameschapter2:Rsquared on train data is higher than on test data. To determine which predictors should actually be taken into account there is a method called "best subset selection" as described by James et al. (2013, p. 227).

To perform best subset selection on data with k predictors, for each $i \in 0, 1, ..., k$ all possible linear models with a number of exactly i predictors are computed. There are $\binom{k}{i}$ different models for each i. Among those models for every i the model with least RSS is selected. This results in k + 1 models (including the empty model) that need to be considered to choose the best model. Finally from those k + 1 models the one with the best score on some indicator is chosen. possible indicators can be:

3.4.1 adjusted R^2

While R^2 explains how much variance in the training data can be explained by the predictors it is an overestimation for the performance on actual test data. Also it gets only larger with more predictors which might be a problem. Therefore an $AdjustedR^2$ can be calculated (James et al., 2013, p. 234).

$$AdjustedR^2 = 1 - \frac{RSS/(n-k-1)}{TSS(n-1)}$$

The term above the division line contains k and punishes models with more predictors. When used as an indicator in best subset selection the model with the greatest adjusted R^2 should be chosen.

3.4.2 Akaike information criterion (AIC)

According to James et al. (2013, p. 234) the AIC can be calculated using the following formula, where k denotes the number of predictors and $\hat{\sigma}^2$ is the variance of ϵ as $VAR(y-\hat{y})$ when looking at the model using all available predictors.

$$AIC = \frac{1}{n}(RSS + 2k\hat{\sigma}^2)$$

Please note that for simplicity normalizing constants have been left out as they do not matter when comparing two models (James et al., 2013, p. 234). A better model is characterized by a lower AIC.

3.4.3 Bayesian information criterion (BIC)

The Bayesian information criterion can be computed quite similarly to the AIC:

$$BIC = \frac{1}{n}(RSS + ln(n)k\hat{\sigma}^2)$$

A low BIC characterizes a good model. The only difference is that the BIC uses the natural logarithm of n in the formula instead of 2. Because $ln(n) > 2$ for $n \geq 8$ and we usually have more than 8 datapoints in our data, the BIC penalizes more predictors more heavily than the AIC and will therefore result in models with less predictors compared to the AIC when used in best subset selection (James et al., 2013, p. 234).

3.5 inference and linear models

To check if a linear regression model with k predictors can predict a significant portion of the variance in the dependent variable an F statistic can be calculated (James et al., 2013, p. 76). Under the assumtion that the H0 (=model cannot predict variance in outcome) is true, F follows an $F_{a,b}$ distribution with $a = n - k$ and $b = k - 1$.

$$F = \frac{(TSS - RSS)/k}{RSS/(n - k - 1)}$$

Each coefficient $\hat{\beta}_j, j \in 1, ..., k$ can also be tested for statistical significance. For this, according to James et al. (2013, p. 67),we can compute a t statistic t_j by dividing the difference between $\hat{\beta}_j$ and the coefficient assumed under the null hypothesis (β_{H_0}) by the standard error of $\hat{\beta}_j$:

$$t_j = \frac{\hat{\beta}_j - \beta_{H_0}}{SE(\hat{\beta}_j)}$$

Under H_0 this would follow a t-distribution with n-k-1 degrees of freedom, which can then be compared to the 1-$\frac{1}{2}\alpha$-quantile of said distribution as a critical value. The standard error $SE(\hat{\beta}_j)$ can be computed with the following formula (James et al., 2013, p. 66), where $\hat{\sigma}^2$ is the residual standard error that can be calculated from the residuals.

$$SE(\hat{\beta}_j) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

$$\hat{\sigma}^2 = \sqrt{RSS/(n - 1 - k)}$$

This also allows for computing a 1-α-confidence interval for each $\hat{\beta}_j$, that signify that $\hat{\beta}_j$ differs significanctly from 0 if they do not contain the 0. a 1-α-confidence interval for a parameter $\hat{\beta}_j$ has the property, that if we would take many random samples from the same population, 1-$\alpha \cdot 100\%$ of confidence intervals will contain the true unknown parameter β_j (James et al., 2013, p. 66).

4 Results

First some descriptive statistics about the dataset is provided, then we evaluete which outcome will be used in the regression (price vs logprice) and finally the best model is computed and evaluated.

4.1 Descriptive Statistics

Table 1 provides a Five-number summary, mean and standard deviation for the relevant metrix variables used in the regression.

Table 1: Relevant Metric variables in the dataset

Rental Price	price	logprice	age	mileage	fuel comsumption	tax	engineSize
mean	14.68	2.54	3.79	25.11	5.11	96.80	1.47
sd	7.75	0.55	1.96	25.04	1.19	61.65	0.42
minimum	3.50	1.25	1.00	1.20	1.70	0.00	1.00
Q1	7.78	2.05	2.00	6.05	4.40	20.00	1.00
Q2 (median)	12.00	2.48	4.00	17.53	4.70	145.00	1.50
Q3	20.99	3.04	5.00	33.37	5.60	145.00	2.00
maximum	38.99	3.66	15.00	138.57	8.69	265.00	2.00

In addition to that we have 3 categorical variables with 3 levels each: model, fuel consumption and transmission. Of the 438 cars in total, there were 161 Passat, 127 T-Roc and 150 Up. The majority of cars had manual transmission (320), 66 cars had Semi-Auto and 52 automatic transmission. Most Cars used Petrol (256) or Diesel (169) with just 13 Hybrid cars that all where from the model Passat and had either automatic (4) or Semi-Auto transmission (9). The 150 Up models all had just manual transmission and ran on Petrol.

4.2 Determining the dependent variable

Two full linear regression models with all 8 predictor variables (age, mileage, fuel consumption, tax, engineSize, model, fuelType and transmission) were fitted to predict price and logprice. For the categorical variables dummy coding was utilized, where Passat was the reference category for model, Diesel for fuelType and Manual for transmission. This will stay constant for all following regressions in this report.

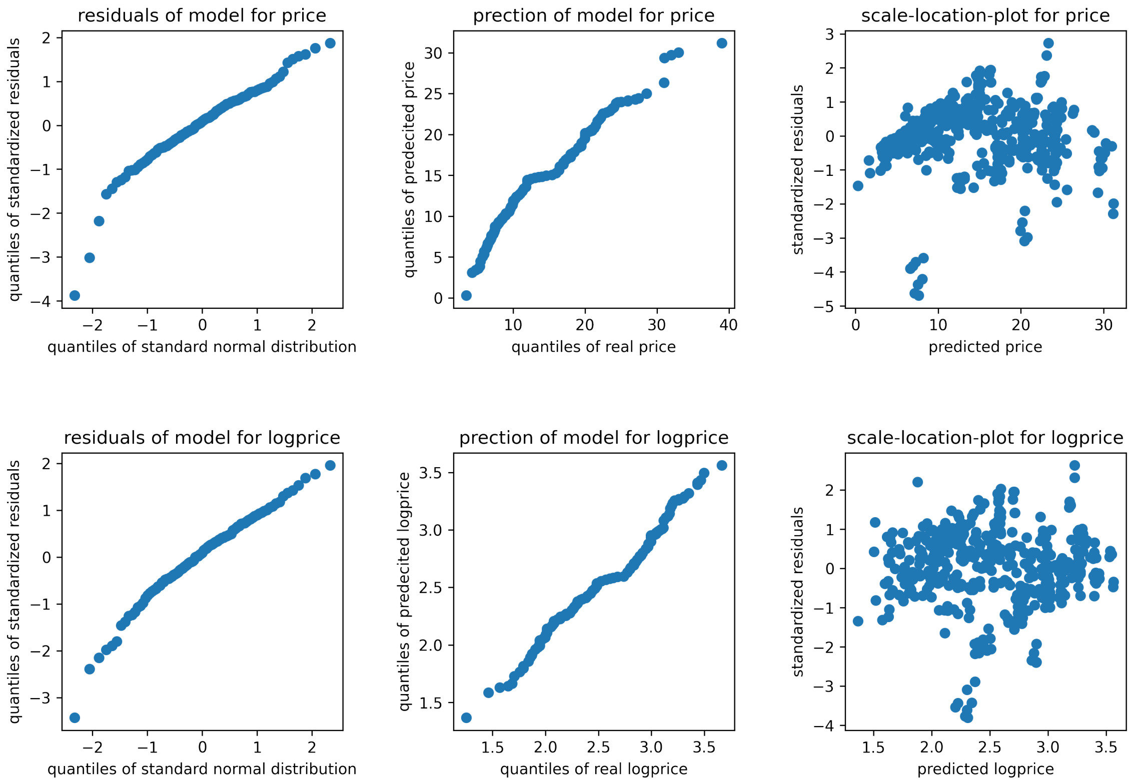


Figure 1: q-q-plots comparing quantiles of standardized residuals with standard normal distribution and predicted values with real values, scale-location-plots for relationship between predicted value and standardized residuals.

As Figure 1 shows, the residuals of the logprice model are closer to a normal distribution than those of the model with price as the dependent variable, because the line of points in the leftmost two q-q-plots is straighter for logprice. Also when predicting logprice instead of price, predicted values and real values seem to share a more similar distribution, as the middle two q-q-pots show. Also there seems to be a curvilinear relationship between predicted values and residuals when using price directly, which does not weem to be the case when using logprice. Finally also the adjusted R^2 value is higher for the logprice model ($R^2 = 0.830$) than for the price model ($R^2 = 0.802$), which suggests a better prediction. Therefore we settle at the logprice as our dependent variable.

4.3 best subset selection

With our 8 predictor variables, best subset selection is performed to determine which model can predict the logprice best. In total, 255 models ($= \sum_{i=1}^{8} \binom{}{8i} = 2^8 - 1$) have been fitted in iterations $i \in \{1, \ldots, 8\}$ to the data. Technically the number of predictors k used for models in iteration i was not always the same, since categorical variables were split up into two dichotomous variables each (dummy coding), before performing model fitting. The models of each iteration were then sorted for their BIC and AIC values and the best value from each iteration i can be seen in Table 2

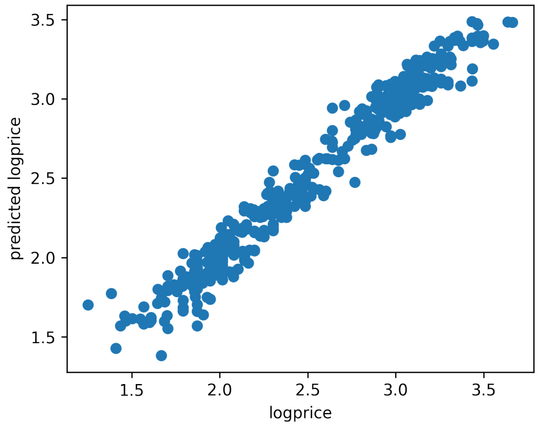
Table 2: Best subset selection: lowest AIC and BIC in each iteration and the respective model predictors added on top of predictors in all table cells above

i	AIC	BIC	predictors	models calculated
1	213.46	225.71	model	8
2	-215.15	-198.82	age	28
3	-414.14	-393.73	mileage	56
4	-564.22	-535.65	transmission	70
5	-651.78	-619.12	engineSize	56
6	-706.25	**-665.42**	fuelType	28
7	**-706.41**	-661.51	fuel consumption	8
8	-706.35	-657.37	tax	1

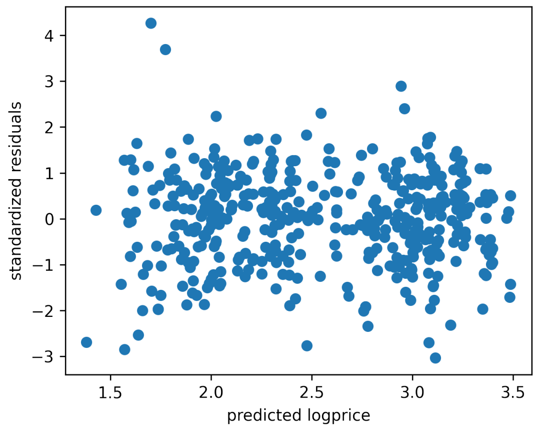
The AIC suggests that the model with 7 predictor variables (all but tax) is the best, while the BIC wants us to stop at iteration 6 and rises again for $i \geq 7$. So the BIC suggests that in addition to tax, also fuel consumption should be excluded from the final model. We will decide for the model the BIC suggests, which will include the predictors: model, age, mileage, transmission, engineSize and fuelType.

4.4 final model

Table 3 shows the regression coefficients and their confidence intervals for the final model with 6 predictors (technically 9, if we count the dummy variables as single predictors). The model exhibits an adjusted R^2 value of 0.962 which is quite impressive and signifies a very good model fit. The predictor can explain 96.2% of the variation in the logarithmic price. The good model fit can also be seen in the scatter plot in Figure 2(a) exhibiting a strong linear relationship between predcited logprice and real logprice. Figure 2(b) also shows that the residuals are spread out quite evenly which is good.



(a) predicted logprice and real logprice



(b) predicted logprice and standardized residuals

Figure 2: plots for the final model with the lowest BIC

All regression coefficients differ from 0 significantly since their 95%-confidence intervals ($[CI_l, CI_u]$) do not contain 0. We will now interpret the coefficients and their meaning for the real price, not the logprice. The intercept of $\hat{\beta}_0 = 2.6160 = log(13.681)$ tells us that the model assumes, for a Passat with Manual Transmission and Diesel an average price of 13,681 GPB, all other predictors held 0 (which is obviously not possible, only electric cars have an engineSize of 0 liters).

Table 3: The final regression model predicting logprice

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	CI_l	CI_u
intercept	2.6160	0.056	2.506	2.726
model: TRoc	0.1534	0.018	0.118	0.188
model: Up	-0.5180	0.024	-0.566	-0.470
age	-0.0884	0.004	-0.096	-0.081
mileage	-0.0058	0.000	-0.006	-0.005
transmission: Automatic	0.1221	0.019	0.085	0.159
transmission: Semi-Auto	0.1180	0.018	0.083	0.153
engine size	0.2865	0.029	0.229	0.344
fuelType: Petrol	0.1227	0.020	0.084	0.161
fuelType: Hybrid	0.4623	0.037	0.389	0.535

A T-Roc seems to cost about 15.34% more on average than Passat, while the model Up sells for less then half of a Passats price on average ($\hat{\beta} = -51.80\%$). An additional year of age seems to lower a cars price by almost 9% ($\hat{\beta} = -8.84\%$). It is the second strongest predictor of a cars value (after model). Additional mileage also reduces a cars price: each additional 1000 miles on a car lower its value by approximately 0.58%. Cars with automatic or semi-automatic transmission sell for 12.21% or 11.80% more than cars with manual transmission according to the model. Mind that this holds for all other variables held constant, so the huge amount of cheap Up models that all just come with manual transmission cannot be the reason for this. Engine size also matters: each additional liter of engine capacity increases a cars price by 28.65%. The least important predictor is the fuel type. Not that it would not matter by itself, but the other predictors already explain a lot of variance so it was added last to the model in best subset selection. Diesel cars are the cheapest. Compared to Diesel, the model predicts that petrol cars cost 12.27% more and hybrid cars even 46.23% more. All percentages provided have to be viewed in the light of their respective confidence intervals depicted in Table 3.

4.5 Summary

The question of this report was to determine

Bibliography

DW. Germanys car industry struggles with transformation amid coronavirus crisis | business | economy and finance news from a german perspective | dw | 08.09.2020. https://www.dw.com/en/what-germanys-car-summit-hopes-to-achieve/a-54841263#:~:text=The%20mixture%20is%20especially%20toxic,is%20made%20by%20the%20industry., 9 2020. (Accessed on 01/19/2022).

Exchange & Mart. Exchange & mart: New & used cars for sale near you. https://www.exchangeandmart.co.uk/, 2022. (Accessed on 01/16/2022).

Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian Marx. Regression models. In *Regression*. Springer, 2013.

Handelsblatt. Immer weniger neuwagen auf privatkunden zugelassen. https://www.handelsblatt.com/mobilitaet/motor/studie-immer-weniger-privatleute-kaufen-neuwagen/20073502.html?ticket=ST-1686404-mmCBHUATN5uCrdf3iizn-ap5, 2017. (Accessed on 01/19/2022).

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

Bommae Kim. Understanding diagnostic plots for linear regression analysis | university of virginia library research data services + sciences. https://data.library.virginia.edu/diagnostic-plots/, 2015. (Accessed on 01/16/2022).

Selva Prabhakaran. 10 assumptions of linear regression - full list with examples and code. http://r-statistics.co/Assumptions-of-Linear-Regression.html, 2017. (Accessed on 01/16/2022).

Statista. Durchschnittliche neuwagenpreise in deutschland | statista. https://de.statista.com/statistik/daten/studie/36408/umfrage/durchschnittliche-neuwagenpreise-in-deutschland/#:~:text=Neuwagenpreise%20in%20Deutschland%20bis%202020&text=Erneut%20ist%20der%20durchschnittliche%20Preis,eines%20Neuwagens%20rund%2036.300%20Euro., 2020. (Accessed on 01/19/2022).

Statista. Statistiken zu kfz-neuzulassungen | statista. https://de.statista.com/themen/1423/kfz-neuzulassungen/, 2022. (Accessed on 01/19/2022).

Appendix

A Additional figures