ACADEMIC METHODS - QUANTITATIVE APPROACHES

Determinants of prices of the German second-hand automobile market: The case of BMW

Otto-von-Guericke-University Magdeburg

Faculty of Economics and Management

Chair of Banking and Finance

Prof. Dr. Peter Reichling

Minh Anh Vu

Matriculation number: 235613

Course of studies: Academic methods - Quantitative approaches

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

WW FACULTY OF ECONOMICS AND MANAGEMENT FT

TABLE OF CONTENTS

CHAPTER 1. AN INTRODUCTION TO RESEARCH TOPIC

1.1. Overview of the automobile market in Germany

Automobile is one of the dominating sectors in Germany and cars are the most well-known German product in the world market. Today the total production of German manufacturer like Volkswagen, Opel, BMW, Audi, Porsche, Mercedes-Benz and others counts more as 10 million vehicles a year. Besides a robust new car market, the second-hand car market is also active. According to information from Deutsche Automobile Treuhand (DAT), the total market size of second-hand cars in Germany up to 2019 is 84.7 billion euros and is expected to continue grow.

Together with Volkswagen, Mercedes-Benz and Audi, BMW is one of the four largest automobile manufacturers in Germany with respect to sales volume. From 1970s up to now, BMW has introduced seven model generations. The 3-series-model which was deputed in 1990 has outweighed the other generations in terms of number of cars listed in the online used car markets.

1.2. Objective of the research

The main objective of this research is to identify the main specifications that determine car prices in German markets. Using a data set of BMW cars, the purpose of this paper is to seek empirical evidence of the relationship between car specifications and used car prices.

This study also aims to examine the effect magnitude on prices of each factors, thereby giving meaningful implications to help customers compare value of different BMW used cars as well as understand the main causes of the price difference.

CHAPTER 2. SAMPLE CONSTRUCTION AND DESCRIPTIVE STATISTICS

2.1. Data collection process

Overall, a cross-sectional sample consisted of prices of 395 pre-owned cars and 8 specifications for each observation have been gathered. The data collection process was composed of four major steps: experimental objects specification, data source selection, data collection and data cleaning. The details for each step are as following:

Step 1: Specifying experimental objects

Our analysis investigates the price variation of second-hand BMW cars in German market. More specifically, the study targets at its third-generation models (3 series), 4 seats, with four popular body types including sedan, coupé, compact and touring.

Step 2: Selecting source of data

The data of this study was collected from *autoscout24.com* on 22nd of January 2021 for several reasons. With 30 million users per month and over 43,000 affiliated dealers, *autoscout24.com* is one of the largest online car dealerships in Germany. In terms of popularity, *autoscout24.com* together with *mobile.de* are the two websites that recorded the highest search volume on Google over the last 12 months relative to other German online car markets.[1]

The significant number of active users and frequent access to *autoscout24.com* indicates a trustworthy source of market data that closely reflects the demand and supply forces. Moreover, the vast number of used cars listed on *autoscout24* also allows a larger sample size, therefore, produces more precise estimations.

Step 3: Collecting required data

There are 9 information were obtained for each observation, including: car price, mileage (number of kilometers driven), year of first registration, horse power, number of previous owners, gear type (automatic/manual), fuel type (gasoline/diesel), fuel consumption per 100km and CO_2 emission per km.

Step 4: Cleaning data

Raw data were then cleaned by removing inaccurate data and removing observations that present less than 3 out of nine information required. For example, 1/396 observation has the amount of fuel consumption per 100km and CO2 emission per km are zero and 1/396 observation has only 2 information available.

[1] See Annex 1

2.2. Descriptive statistics

Table 1: Descriptive statistics of car specifications

Variable	Observation	Mean	Std. Dev.	Min	Max
price	395	21.72	9.95	0.999	61.4
mlg	395	83.29	60.59	3.17	364.29
year	395	2015.37	4.79	1986	2020
age	395	5.63	4.79	1	35
hp	395	195.08	52.72	105	374
owner	356	1.31	0.6	1	4
con	343	5.4	2.198	0	10.6
ems	348	136.47	51.2	0	255
gear	395				
fuel	395				

As observed from Table 1, the data set contains information of prices of 395 used cars with 8 specifications for each observation. There are three variables presents missing data as follow: 39 cars in our sample were missing data on the number of previous owners *(owner)*, 52 were missing data on the amount of fuel consumption variable *(con)* and 47 were missing data on the amount of CO2 emission *(ems)*. However, as the data is missing randomly it does not threat the precision of estimation.

Car price, the predicted variable, reports an average value of 21,720 euro. The most expensive car in the sample is 61,400 euro, while the cheapest car is roughly 1,000 euro.

With respect to explanatory variables, a number of prominent patterns shall be brought into discussion. *First*, 70% of the cars in the sample are from 1- to 6-year-old, equivalent to an average age of 5.63 years. The oldest car reported was registered 35 year ago. *Second*, as expected, the older the auto is, the longer distance it has run. In our sample, the furthest a car has run was 364,290 km of a 16-year-old car and the shortest run was 3,172km with respect to a 1-year-old car. *Third*, 75% of the cars are resold for the first time. Only 7% of the cars in the sample have more than 2 previous owners. The finding supports the fact that majority of the listed second-hand cars are relatively new (1- to 6-year-old). *Fourth*, 74% of the pre-owned cars are automatic. It is apparent that newly manufactured cars prefer automatic transmission than manual one when the average age of automatic car group (4.4 years) and manual group (8.8 years) are brought into comparison. *Finally*, two-third of the second-hand cars run by diesel instead of gasoline. Also, diesel cars consumes less fuel as well as produce less CO2 emission than gasoline cars for the same distance.

2.3. Correlation matrix

Table 2: Correlation matrix of car specifications

	mlg	**age**	**hp**	**owner**	**dgear**	**dfuel**	**con**	**ems**
mlg	1							
age	0.6288	1						
hp	-0.0095	-0.1952	1					
owner	0.3979	0.5115	-0.0528	1				
dgear	-0.2746	-0.4091	0.3789	-0.2697	1			
dfuel	-0.0654	-0.3869	-0.0469	-0.1615	0.1363	1		
con	0.2751	0.5737	0.2799	0.2395	-0.1069	-0.7315	1	
ems	0.3228	0.5839	0.3107	0.2327	-0.0825	-0.5609	0.9542	1

The correlation matrix between explanatory variables can be found in table 2. In general, the correlation coefficients' absolute value and direction reflect the findings stated in the descriptive statistics section. For clarification, "dgear" represents automatic car and "dfuel" presents diesel car.

As expected, age possesses high correlation with mileage and number of previous owners, amounting to 0.63 and 0.51, respectively. Car age also has high correlation with amount of fuel consumption and CO2 emission. This can be explained by the fact the newly manufactured cars prefer diesel more than gasoline, therefore the pollution and energy waste are lower. Most notably, fuel consumption and CO2 emission are almost perfectly correlated (95%). The multicollinearity might pose a threat to the significant of regressed coefficient. A multicollinearity test will be conducted for this data set in the later section.

CHAPTER 3. METHODOLOGY AND REGRESSION RESULT

3.1. Model specification

Our analysis exploits Ordinary Least Square (OLS) multivariate regression model to identify factors that have significantly explanatory power to the variation of used car prices as well as to estimate the magnitude and direction of the effects. Our assumption is that all independent variables have linear relation with dependent variable.

$$Price_i = \beta_0 + \beta_1 mlg_i + \beta_2 age_i + \beta_2 hp_i + \beta_3 dgear_i + \beta_4 dfuel_i + \beta_5 dow1_i + \beta_6 dow2_i + \beta_7 dow3_i + \beta_8 ems_i + \beta_9 con_i + \epsilon$$

$Price_i$: price of the second-hand car *i* *(unit: thousand euro)*

mlg_i : mileage – number of kilometers driven by car *i* *(unit: thousand km)*

age_i : number of years that car *i* has been used since its first registration *(unit: year)*

hp_i : engine's horsepower of car *i* *(unit: horsepower)*

dgear : dummy variable

dgear = 1 if car *i* is automatic, dgear = 0 if car *i* is manual

dfuel : dummy variable

dfuel = 1 if car *i* uses diesel, dfuel = 0 if car *i* uses gasoline

dow1, dow2, dow3: dummy variables

dow1 = 1 if car *i* has 1 previous owner and dow1 = 0 otherwise

dow2 = 1 if car *i* has 2 previous owner and dow2 = 0 otherwise

dow3 = 1 if car *i* has 3 previous owner and dow3 = 0 otherwise

ems : CO2 emission per km *(unit: gam)*

con : fuel consumption *(unit: litter/100km)*

As our model uses OLS regression, it is important to test the homoskedasticity assumption. If this assumption is violated, OLS would present unreliable significant levels of estimations. The white test has shown heteroskedasticity problem in our model and our analysis uses robust standard errors to heal this problem. Besides, multicollinearity is also be examined. High multicollinearity increases standard errors and makes significant coefficient less likely. The study uses Variance inflation factor (VIFs) to detect whether near-perfect multicollinearity presents. The high correlation of fuel consumption and emissions has shown that their coefficients might be not reliable. The standard errors of coefficients of the remaining variables, however, are not affected by this problem [2]

[2] Details of the result and interpretation of the two tests can be found in Annex 2

3.2. Regression result

Table 3: Regression results of the linearity model

Adjusted R-squared: 0.8106

price	**mlg**	**age**	**hp**	**dgear**	**dfuel**	**dow1**	**dow2**	**dow3**	**con**	**ems**	**_cons**
Coef.	0.076***	0.86***	0.058***	2.104***	1.699*	1.423*	0.695	0.462	1.563	-0.02	12.295
Robust SE	0.007	0.088	0.007	0.499	0.893	0.769	0.742	1.147	0.926	0.037	2.086

Table 3 reports the result of the effects of 8 car specifications on the car prices. Overall, the model demonstrates strong explanatory power to the variation prices with R-squared amounted at 81%.

Four factors that have significant impacts on car prices at 1% level are distance driven, car age, horsepower and gear type. The coefficients of fuel type variable and 1-previous owner dummy variable are significant at 10%. The remaining factors are statistically insignificant.[3]

In details, mileage and car age have negative influence on car prices. Hold other factors fixed, every thousand kilometers the car run drives down its price by 76 euro. Other factors equal, the BMW car that one year older is 860 euro cheaper. On the other hand, automatic transmission and engine power have positive influence on car prices. Comparing two cars with similar profile, an automatic BMW car is on average 2,104 euro more expensive than a BMW manual car. Similarly, a car engine with more 1 horsepower will increase the car price by 58 euro.

Regarding the impact of fuel type and the number of previous owners, we are 90% confident that a BMW car run by diesel are priced 1,699 euro higher than a gasoline fueled car. At the same confident level, we can also assert that a car that is resold first time is 1,423 euro worth more than a car has been repurchased four times.

The directions of the effects are aligned with economic norms. For instance, the older the car is, the less customers are willing to pay for it due to the higher risk of malfunction (guarantee service might have expired). Another example is, diesel cars are by far more energy efficient than gasoline fueled cars, hence, customers tend to value them higher.

[3] The symbol (***), (**), (*) indicates 1%, 5% and 10% significant levels, respectively

CHAPTER 4: ROBUSTNESS CHECK

The model is run on the basis of key OLS assumptions, this subsection will discuss issues could alter the original results, which mainly focus on the endogeneity and functional misspecification problems

4.1. Selection bias

Selection bias occurs when the sample data set does not reflect the true relationship in the population. Here the target population is the second-hand BMW cars in German market, 3-series-model (4 seats). Our analysis mitigated this problem by retrieving the entire available information of this type of cars on autoscout24.com, one of the two largest online car dealers in Germany (together with mobile.de).

Provided that the used car market in Germany is liquid and competitive, it is logical to assume that the prices quoted on both sides for identical cars are relatively similar, hence, selection bias in our study is unlikely. However, it is recommended for further research to employ data of both mentioned car dealers, increasing sample size to get closer to the true relation in the population

4.2. Omitted variable bias

Omitted variable bias occurs when the regression fails to control for more additional variables due to data unavailability, therefore, the error terms correlate with the regressors. Our study mitigated this risk by incorporating all 8 crucial auto specifications into the model that are believed having no correlation with the remaining features of a car, such as, interior and exterior colors, additional furnishing (speakerphone, Bluetooth, light sensor, etc.). Also, the latter features are considered to have no significant influence on car prices. Hence, it is less likely that our model present biased result due to lacking important car specifications

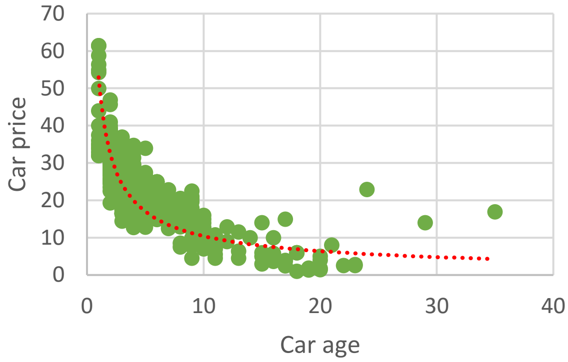
However, there is another form of omitted variable problem that our analysis might face: functional form misspecification.

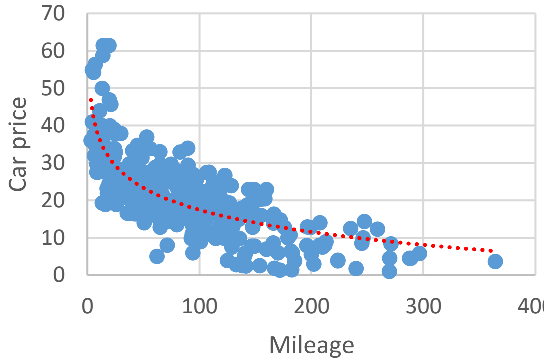
4.3. Functional form misspecification

One of our assumptions for the original model is that all the independent variables have linear relation with the dependent variable. If this assumption is violated, the coefficients are biased. The Author suspects this assumption might be violated for some reasons.

(1) As observed from the graph, the slope of the fitted line of price and age is not constant along the age variable.

Figure 1: Price variation with respect to car age and mileage





Within the first ten year of a car's life, its price decreases rapidly as each year passes. However, after 10 years, the fall in price seems to stagnate. A similar pattern is found in the relationship between price and mileage. These findings signify a possibility of non-linear relationship between dependent and independent variables, that might require different functional form.

(2) One of the tools to test functional form is the regression specification error test (RESET) of Ramsey's (1969). Conducting RESET test our original model reports p-value of 0.0000 for the value of an $F_{3,328}$. This is an evidence of functional form misspecification in original model.

One solution suggested is to add quadratic terms of any significant variables from the original model, and test whether the coefficients are jointly significant. If the additional quadratics are significant, they can be used in the model. Logarithm functional form is also one of the options.[4]

Our study has run two alternate functional forms to compare the result with the original model:

- *Quadratic functional form:*

$$Price_i = \beta_0 + \beta_1 mlg_i + \beta_2 mlg_i^2 + \beta_3 age_i + \beta_4 age_i^2 + \beta_5 hp_i + \beta_6 dgear_i + \beta_7 dfuel_i + \beta_8 dow1_i + \beta_9 dow2_i + \beta_{10} dow3_i + \beta_{11} ems_i + \beta_{12} con_i + \epsilon$$

- *Logarithm functional form:*

$$Price_i = \beta_0 + \beta_1 lnmlg_i + \beta_2 lnage_i + \beta_2 hp_i + \beta_3 dgear_i + \beta_4 dfuel_i + \beta_5 dow1_i + \beta_6 dow2_i + \beta_7 dow3_i + \beta_8 ems_i + \beta_9 con_i + \epsilon$$

[4] See Wooldridge (2008), pp. 301-302

Table 4: Comparison of regression results of linearity and non-linearity models

	Linear OLS model Ad_R2: 0.8106	**Quadratic model** Ad_R2: 0.8794	**Logarithm model** Ad_R2: 0.9048
mlg	(0.075) ***	(0.139) ***	-
mlg^2	-	0.00032 ***	-
ln(mlg)	-	-	(4.17) ***
age	(0.860) ***	(2.509) ***	-
age^2	-	0.0799 ***	-
ln(age)	-	-	(7.531) ***
hp	0.059 ***	0.074 ***	0.068 ***
dgear	2.104 ***	1.554 ***	1.545 ***

Table 4 compares the significant coefficients from the linearity models and two non-linear models. Some vital conclusions can be drawn the comparison as follow:

First, non-linear functional models are fitted to the sample data more than the linearity model. The adjusted R-squared of quadratic, logarithm and linear model are 0.81, 0.87, 0.90, respectively. Visually, the fitness of the three models can also be seen in annex 3 where actual prices are plotted against predicted values.

Second, the residuals of non-linear models are more randomly scattered around zero than the linearity model, showing less concern for the bias to the significant level of coefficients.[5]

Finally, but most importantly, irrespective of the functional form, all three models confirm the significant influence of mileage, car age, horsepower and gear type to used car prices at 1% level. This finding reinforces our analysis objective, which is to study the find the true determinants to used car prices.[6]

[5] See Annex 4

[6] See detail results in Annex 5

CHAPTER 5. CONCLUSION

Inspired by the active second-hand auto market in Germany, our study wishes to identify the true determinants of prices of the German used cars market: using the case of BMW, one of the largest car manufacturers in Germany.

The analysis first demonstrates the process of sampling data, pointing out and explaining the characteristics of the sample. Subsequently, eight car specifications and car prices are examined using multivariate regression model. Heteroskedasticity and multicollinearity are checked to ensure the consistency of the results. Finally, a detail discussion in the robustness of the model has reinforced some crucial findings as follow:

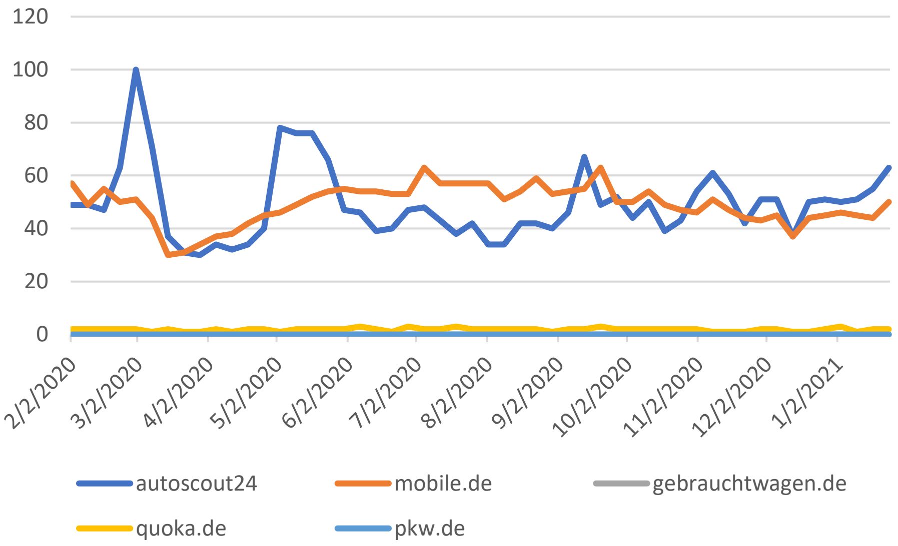
(1) Irrespective of linearity or non-linearity relationship between car prices and dependent variables, four factors including mileage, car age, horsepower and gear type are consistently significant at 1% level. The directions of the effects are also aligned with the economic norms.

(2) Compared to linearity model, quadratic and logarithm models demonstrate slightly stronger explanatory power to the variation of car prices for our data, yet, at the cost of complicating the interpretation of the models.

The results provide meaning implications for customers to determine a reasonable price to a second-hand BMW car (3-series model) as well as be able to explain the causes of price differences. For further research, the study suggests to conduct the methodology in a larger sample size (compiled data of various car dealerships) to avoid selection bias and to investigate further an appropriate functional form if the researcher wishes to compute the marginal effect of each factor on car prices.

APPENDIXES

Annex 1: Google search volume of online car dealers in Germany

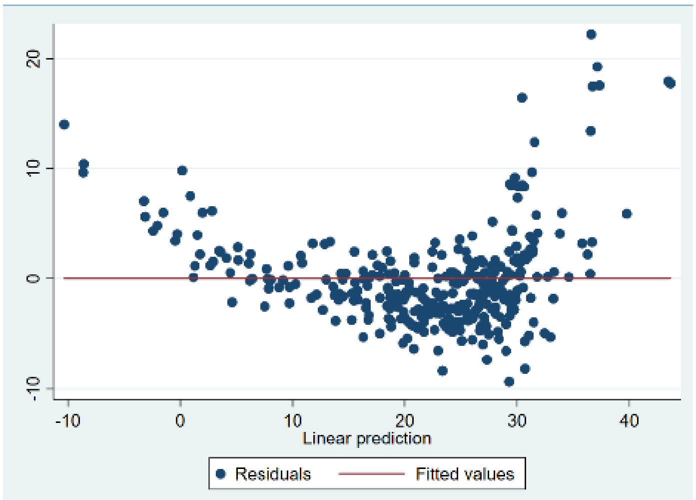


Annex 2: Heteroskedasticity and Multicollinearity tests

White test for heteroskedasticity

One of the vital assumptions of OLS model is homoskedasticity standard errors, which means the variance of residuals does not vary with independent variables. If this assumption is violated, the p-value of coefficients are driven downwards, hence, OLS would present unreliable significant levels of estimations. Our analysis tests this problem using residual plotted graph and Breusch-Pagan test.

Figure 2: Residuals against fitted values of the linearity model



Graphically, if the model is well-fitted, there should be no pattern to the residual plotted against the fitted value. However, our analysis shows a convergence tendency of residuals towards the right-hand side of fitted values, which affirms the problem of heteroskedasticity. Moreover, deploying Breusch-Pagan / Cook-Weisberg test with p-value = 0.0000 (<5%), the test also rejects the hypotheses of constant variance of the error terms.

Multicollinearity test

Table 5: Variance inflation factors of independent variables

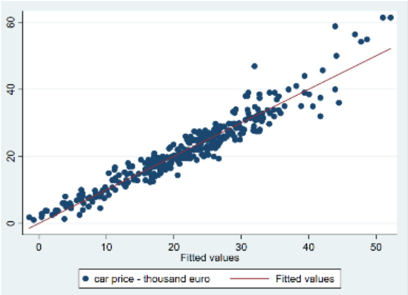
Var	VIF	1/VIF
con	28.04	0.04
ems	19.8	0.051
age	4.32	0.231
dfuel	3.78	0.264
dow1	3.33	0.3
dow2	2.35	0.426
mlg	2.18	0.458
hp	1.81	0.552
dow3	1.44	0.694
dgear	1.38	0.724
Mean VIF	6.84	

Multicollinearity occurs when there is correlation between two or more independent variables. High multicollinearity increases standard errors and makes significant coefficient less likely. Our analysis uses Variance inflation factor (VIFs) to detect whether near-perfect multicollinearity presents. As a rule of thumb, multicollinearity is a problem if VIFs are higher than 10.

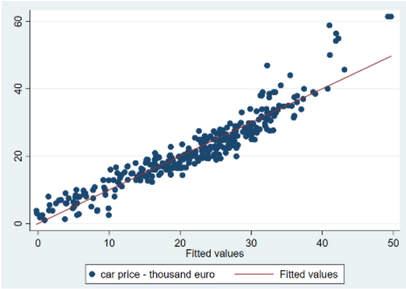
VIFs test presents a mean VIF of 6.84 <10 since most of the variables have VIF value lower than 5. However, fuel consumption and CO_2 emission variables are highly correlated evidenced by considerably high VIF values at 28 and 19.8, respectively. The coefficients of fuel consumption and emissions, therefore, are less reliable. The standard errors of coefficients of the remaining variables are not affected by this problem.

Annex 3: Actual values plotted against fitted values

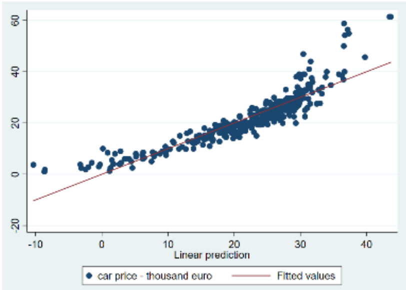
Logarithm model



Quadratic model

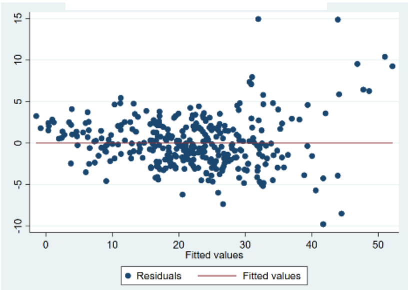


Linear model

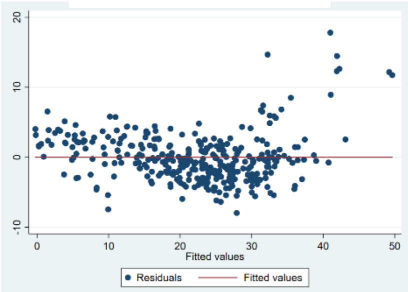


Annex 4: Residuals plotted against fitted values

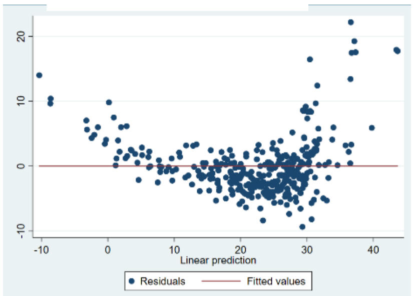
Logarithm model



Quadratic model



Linear model



Annex 5: Regression result of linearity and non-linearity models

(1) Linearity model

```
. reg price mlg age hp dgear dfuel dow1 dow2 dow3 con ems, robust

Linear regression                               Number of obs     =        342
                                                F(10, 331)        =      92.31
                                                Prob > F          =     0.0000
                                                R-squared         =     0.8106
                                                Root MSE          =     4.5552
```

price	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
mlg	-.0756324	.0069868	-10.83	0.000	-.0893765	-.0618883
age	-.8596066	.0879356	-9.78	0.000	-1.03259	-.6866235
hp	.0578308	.0068716	8.42	0.000	.0443132	.0713484
dgear	2.10415	.4987958	4.22	0.000	1.122941	3.08536
dfuel	1.699309	.8929788	1.90	0.058	-.0573199	3.455939
dow1	1.423392	.7686033	1.85	0.065	-.0885713	2.935355
dow2	-.6947219	.7415457	-0.94	0.350	-2.153459	.7640149
dow3	.4620203	1.146845	0.40	0.687	-1.794004	2.718045
con	1.562941	.9261121	1.69	0.092	-.2588667	3.384749
ems	-.0195453	.0370271	-0.53	0.598	-.0923835	.0532929
_cons	12.29479	2.08596	5.89	0.000	8.191379	16.3982

```
. di e(r2_a)
.80488128
```

(2) Logarithm model

```
Linear regression                               Number of obs     =        342
                                                F(10, 331)        =     220.95
                                                Prob > F          =     0.0000
                                                R-squared         =     0.9076
                                                Root MSE          =     3.1822
```

price	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnmlg	-4.170235	.320137	-13.03	0.000	-4.799994	-3.540475
lnage	-7.531961	.5146354	-14.64	0.000	-8.544329	-6.519592
hp	.0681139	.0040913	16.65	0.000	.0600657	.076162
dgear	1.54541	.3916138	3.95	0.000	.7750438	2.315775
dfuel	-.3578392	.6463776	-0.55	0.580	-1.629365	.9136868
dow1	.16569	.5033017	0.33	0.742	-.8243834	1.155763
dow2	.2313074	.5099466	0.45	0.650	-.7718374	1.234452
dow3	.3504277	.5985181	0.59	0.559	-.8269511	1.527807
con	.1457218	.6801352	0.21	0.830	-1.192211	1.483654
ems	.0101899	.0268307	0.38	0.704	-.0425904	.0629702
_cons	33.29132	1.625379	20.48	0.000	30.09394	36.4887

```
. di e(r2_a)
.90477803
```

(3) Quadratic model

```
. reg price mlg mlg2 age age2 hp dgear dfuel dow1 dow2 dow3 con ems, robust

Linear regression                               Number of obs     =        342
                                                F(12, 329)        =     149.99
                                                Prob > F          =     0.0000
                                                R-squared         =     0.8837
                                                Root MSE          =     3.5807
```

price	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
mlg	-.1390325	.011645	-11.94	0.000	-.1619405	-.1161245
mlg2	.0003188	.0000431	7.40	0.000	.000234	.0004036
age	-2.509722	.2235203	-11.23	0.000	-2.949431	-2.070012
age2	.0798867	.0101408	7.88	0.000	.0599378	.0998357
hp	.0740627	.006214	11.92	0.000	.0618385	.0862869
dgear	1.554254	.467966	3.32	0.001	.6336707	2.474837
dfuel	.451473	.752097	0.60	0.549	-1.028053	1.930999
dow1	-.3447161	.6341286	-0.54	0.587	-1.592174	.902742
dow2	.042126	.6448522	0.07	0.948	-1.226428	1.31068
dow3	1.031244	.7826211	1.32	0.189	-.5083286	2.570817
con	.7938222	.7401946	1.07	0.284	-.6622891	2.249934
ems	-.0151292	.0291349	-0.52	0.604	-.0724434	.042185
_cons	21.87721	1.50823	14.51	0.000	18.91022	24.84421

```
. di e(r2_a)
.87943688
```

REFERENCE

[1] Wooldridge, J.M. (2008*): Introductory Econometrics: A Modern Approach*, 4th ed., South-Western.

[2] Henk Bekker (2020), *2020 (Q3) Germany: Best-Selling Car Brands and Models*, <https://www.best-selling-cars.com/germany/2020-q3-germany-best-selling-car-brands-and-models/> [Accessed 31.01.2021].

[3] Thomas Thieme (2019), *Große Nachfrage nach Benzinern* <https://www.stuttgarter-nachrichten.de/inhalt.dat-report-gebrauchtwagenmarkt-auf-allzeithoch.6a83fd1d-1dba-4157-9163-f199680c5de5.html> [Accessed 05.02.2021].