Analysis on used cars - vol 2

Zsombor Hegedus

2020 december 18

Executive summary

This document is to uncover a pattern of association between the age and price of used Renault Megane cars. The following paper have two main parts, where the first one introduces important features of the data and focuses on simple linear regression analysis, whereas the second part will focus on non-linear patterns in the data. My final model choice is to use a log-level model, which not only has a very good fit, but also provides a clear interpretation. At the end, I will also showcase a few used cars that are proposed to be underpriced and can be considered a good deal. Of course, the truth is not always that simple, and there are many factors to consider when buying a car, other than its age.

This analysis is the continuation of an earlier project of mine, which is aviailable in this github repo. The analysis was done as homework for chapter 7 exercise 4 and chapter 8 exercise 5 - files for the current analysis are stored in this github repo

Summary statistics

I will make an attempt to uncover a pattern of association between the price and the age of given Renault Megane model (Price is expressed in million HUF throught the document). A trivial expectation is that older cars cost less than newer ones, but let's see if this can be proven by regression analysis. First, let's look at the most important statistics of the two variables, the *Age* and the *Prices* which are summarised in Table 1.

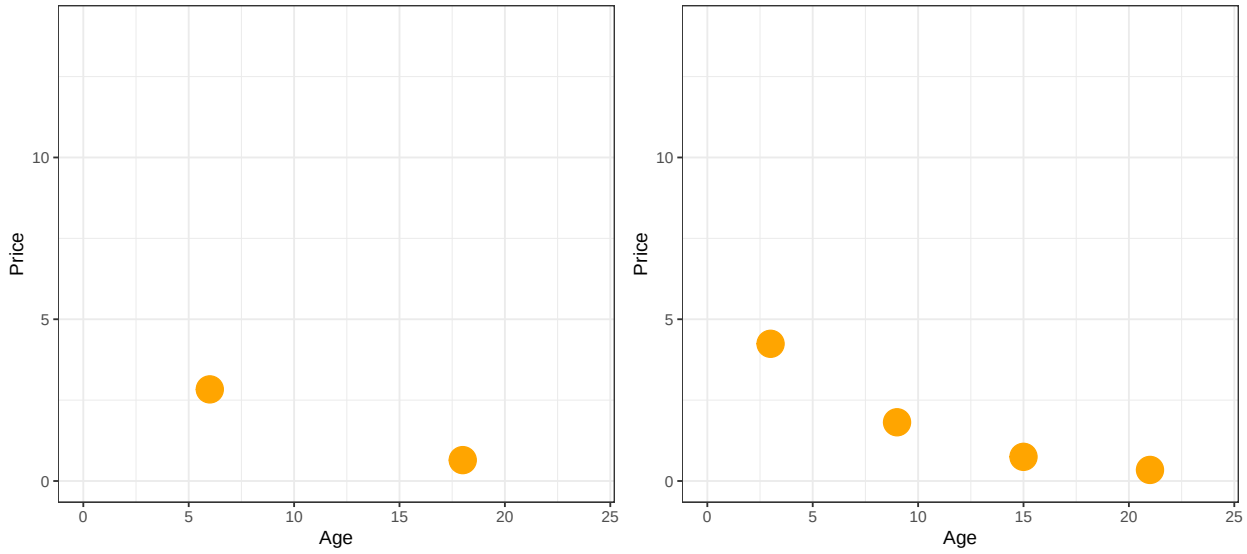
variable	mean	median	std	iq_range	min	max	skew	numObs
Age	12.41	13.00	5.78	8.00	0.00	24.00	-0.27	1124
Price in HUF(mm)	1.49	0.90	1.57	1.34	0.04	13.90	2.74	1124

Table 1: Summary statistics of examined variables

We have a little bit more than a 1000 observations in our data. Age variable shows that we already have cars made in the current year (2020) that are already being sold, and we can also see at least one 24 years old car which probably doesn't worth much now. The the standard deviation of the prices is quite high, and from the mean and median we can also suspect that the distribution of the variable resembles lognormal.

Simple linear regression without transformation

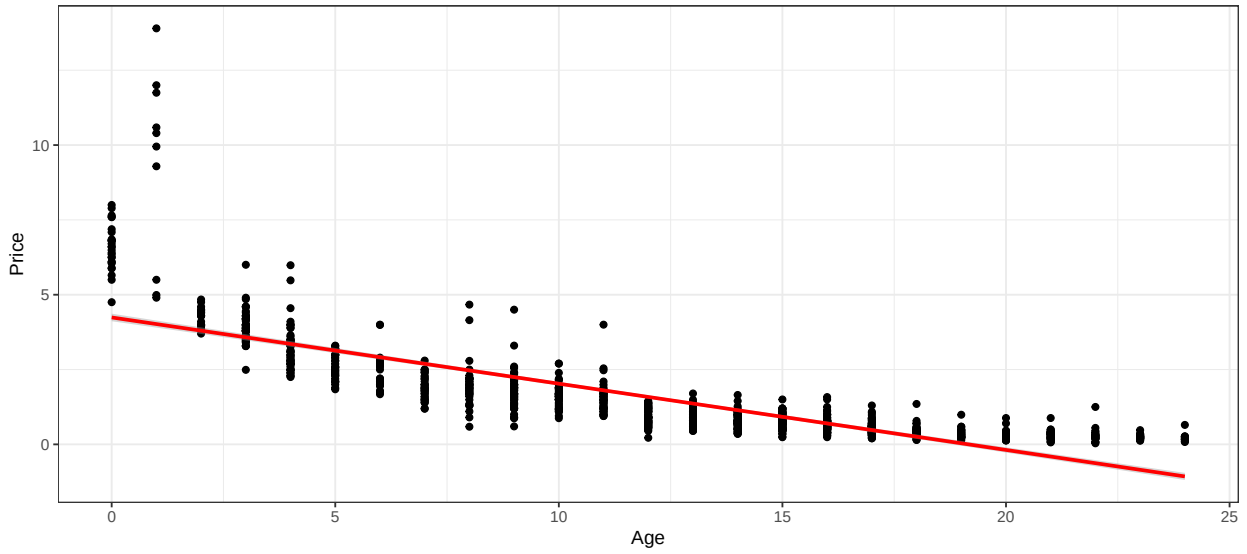
I will use a linear regression for analysing the pattern of association, but before that I will show two bin scatter plots so that we can have a sense of the relationship. First I used two bins that indicate a decreasing trend between our variables - this is in line with the expectations; older cars have lower price in the data. A bin scatter with 4 bins also enforces this message, the trend is more visible and perhaps it is even non-linear.



But perhaps bin scatter is not the best to uncover patterns. When I grouped together x variables, a lot of information was lost which might be very valuable. I can see some pattern from the bins as well, but there is a chance that I am missing out on e.g.: a few spikes in the data. I also want to generalise my findings and have a quantifoable impact of this association. To overcome this issue I ran the following linear regression.

$$y^E = \alpha + \beta x$$

The below figure illustrates how such a linear regression line fits on the scatter plot between the two variables.



We can see that it is pretty good at capturing the decreasing pattern between price and age. Estimated coefficients can be seen in Table 2. I will also do hypothesis testing and examine if they are significant at 5% - since I don't really want to see if this pattern exists without reasonable doubt, 5% looks like a good enough choice for this research question.

The intercept is at 4.24mm HUF, this is the expected price of a car with age = 0, which means that for cars that were made in 2020. The estimated slope coefficient is -221.3k HUF which means that we expect our price to be 221.3k HUF lower for a car that is one year older in our data. My pre-set confidence was 5%, and appereantly both of the estimated coefficient are significant at that level. P-values suggest that I could

variable	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	4.24	0.12	35.21	0.00	4.00	4.48
Age	-0.22	0.01	-27.42	0.00	-0.24	-0.21

Table 2: Hypothesis testing for the slope of the regression

have chosen an even lower significance level, but nevertheless, I will reject the H_0, and say that none of the estimated coefficients are zero.

It is worth taking a look at the residuals. If we were to choose the best deals based on the linear model, Table 3 shows what we would get:

ID	Age	Prices	Km	Power_kW	Pred	Residuals
16291520	12	0	299000	78	2	-1
16217935	8	1	228000	66	2	-2
16241170	9	1	252900	66	2	-2
16291795	9	1	353000	81	2	-1
16211561	9	1	305400	66	2	-1
16082831	9	1	399000	81	2	-1
16260283	8	1	320064	66	2	-2
16197756	8	1	236476	81	2	-1
16260660	7	1	350000	81	3	-2
16157086	7	1	431000	81	3	-1

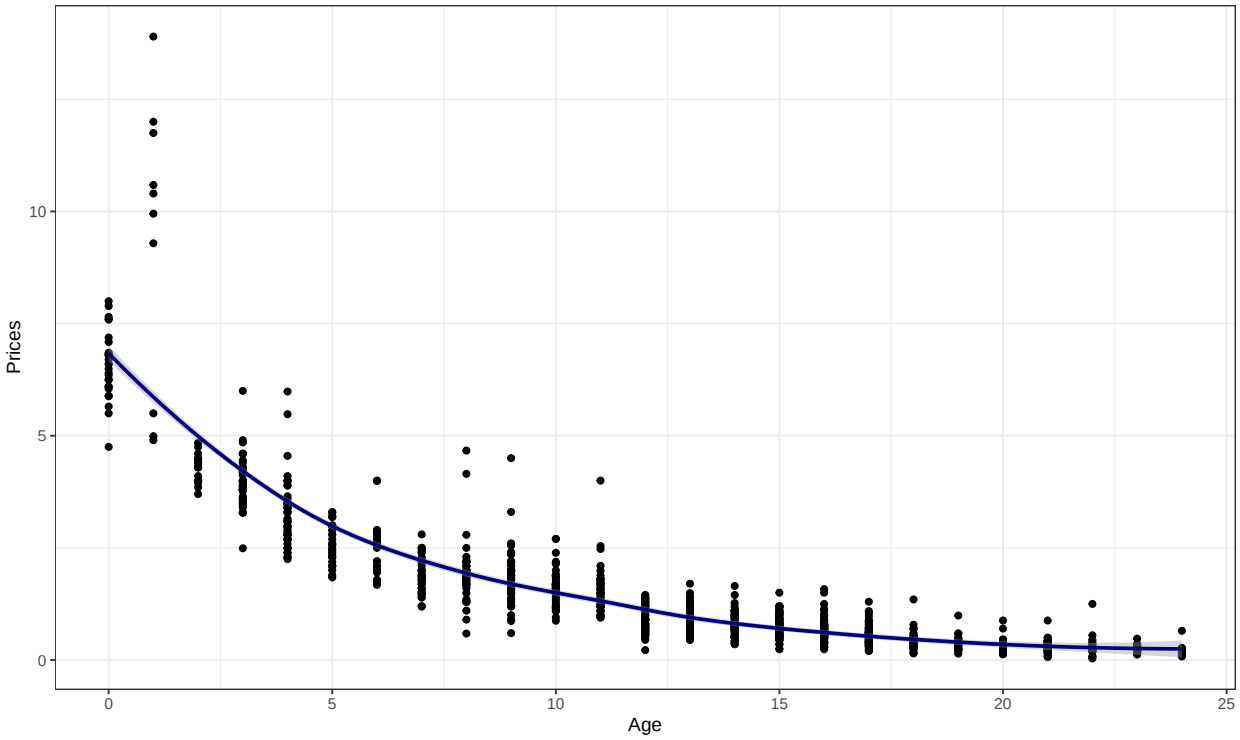
Table 3: 10 Best Renault Megane deals proposed by simple linear OLS based on Age-Price relationship

We could have seen from the figure with the regression line that the majority of the $\hat{y}$s with high negative residuals are between 5 and 13 years. Even though it is worth taking a look at the above ten cars, one should keep in mind that there are multiple other factors to consider when buying a car such as the Kms it run, or whether there are any extras included (like sensors for parking).

But when looking at the scatter plot I can see some non-linearity, so it is worth to continue the analysis with transformations that can help capturing this pattern.

Modeling non-linearity

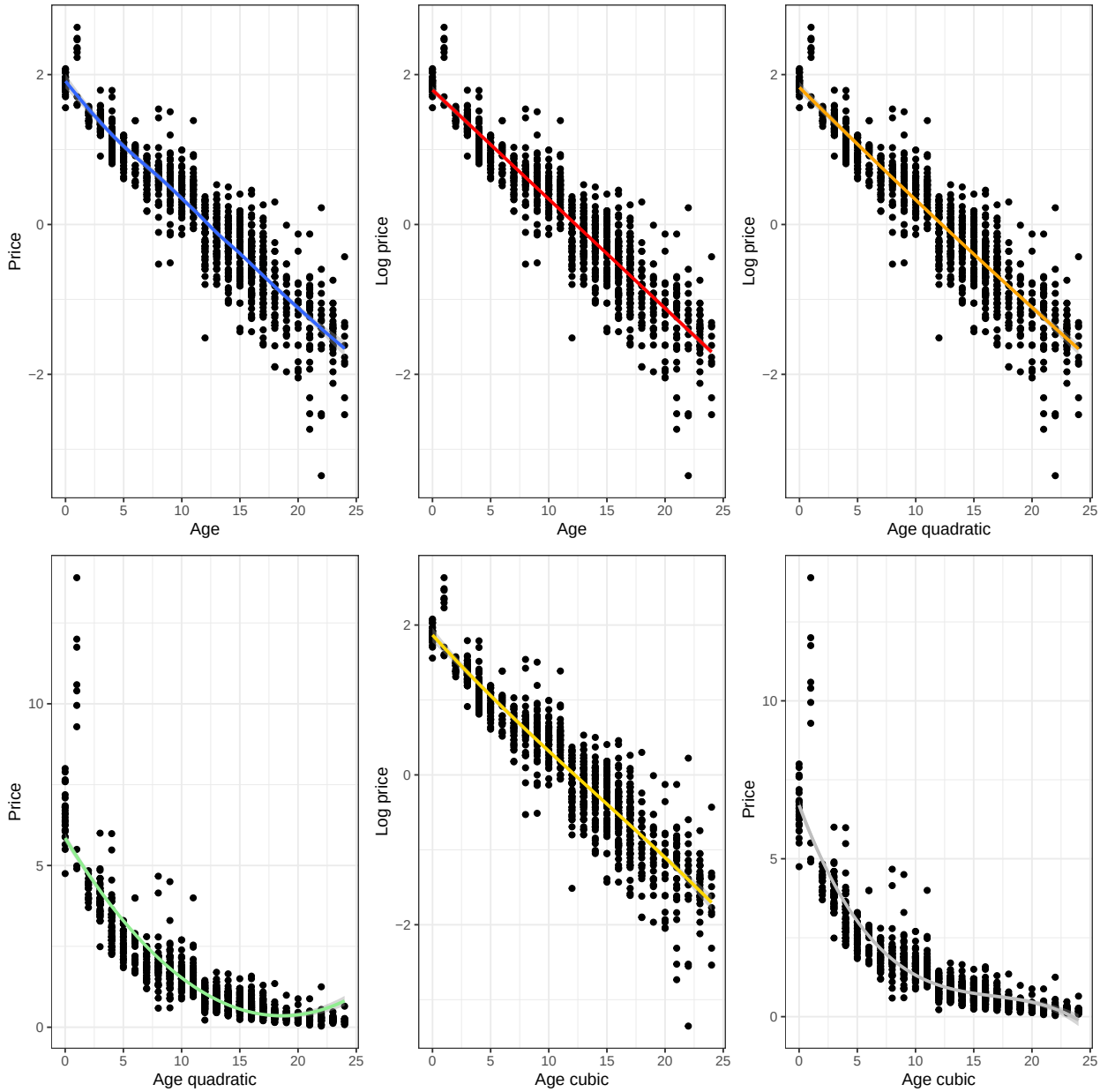
As a first step, I wanted to see if I can visually prove that there is a non-linear patter. One of the best ways to do that is by employing a non-parametric regression model, called lowess. Lowess will dynamically calculate the average ys in a symmetrical interval around different x values. The figure below shows the lowess function on the scatter plot with blue color.



The smooth line in the data is exactly what I wanted to see, there is clear non-linear pattern in our data that the lowess captured pretty well. Looks like the decreasing trend is not as steep as the first model implies, and another very interesting behaviour is around cars that are 1 years old. They are completely outside the pattern and can be considered to be extreme values - it is advisable to look into them a little bit more.

They are all *RENAULT MEGANE 1.8 TCe R.S RS* models which is a high end luxury version of the Renault Megane with a displacement of ca.1800 cm^3, which explains why they don't fit into our pattern that much. I decided to keep these variables as they are not a result of error and still in the population that I want to examine, which is Renault Megane used cars.

My next goal is to introduce a linear regression model that can capture this non-linearty. I thought of two approaches that can be of help for this exercise: polynomials and log transformations. Given the lognormal looking distribution of the *Price* variable (which is also a ratio variable), I decided to get the natural log of *Price*, and for the polynomial case I will try to see how well the quadratic and cubic form fares. I experimented visually with a few model choices, which is visible in the below figure:



Based on the above I narrowed down the potential model choices to 3, which were the following :

- Quadratic regression
- Cubic regression
- Log-level regression

I also kept the first linear regression model so that we can see the improvement compared to that as a baseline. The regression model summaries are collected in the below table.

	Price Age	Price Age - quadratic	Price Age - cubic	Log Price Age
(Intercept)	4.24*	5.85*	6.70*	1.79*
	[4.00; 4.48]	[5.51; 6.19]	[6.23; 7.16]	[1.75; 1.84]
Age	−0.22*	−0.59*	−0.99*	−0.15*
	[−0.24; −0.21]	[−0.64; −0.54]	[−1.11; −0.86]	[−0.15; −0.14]
Age_sq		0.02*	0.06*	
		[0.01; 0.02]	[0.05; 0.07]	
Age_cb			−0.00*	
			[−0.00; −0.00]	
R^2	0.66	0.81	0.84	0.85
Adj. R^2	0.66	0.81	0.84	0.85
Num. obs.	1124	1124	1124	1124
RMSE	0.91	0.68	0.63	0.35

* Null hypothesis value outside the confidence interval.

We can see a pretty substantial imrpovement in the R^2 for each new model, that captures non-linearity. We are now capable of explaining 80%-85% of the variation of y with the new models, compared to 66% of the level-level OLS. The β_2 coefficient in the quadratic form shows that the parabola used for this regression is convex. Even though the cubic case improves our model fit sligthly, itt brings extra complexity to the model, which we don't really need when uncovering patterns of association. On the other hand the log-level model provides a great fit, we could also see how well it fits on the scatterplot, and it has a clear interpretation as well. All in all the log-level model is going to be my model choice, which can be described in the below functional form:

$$ln(y)^E = \alpha + \beta x$$

The intercept can't really be interpreted as such, since our y variable is log transformed. The estimated slope coefficient is -0.15 which means that we expect our price to be 15% lower for a one year older car. My pre-set confidence was again set at 5%, and appereantly my slope coefficient is significant at that level. P-values suggest that it is highly significant so, I will reject the H_0, and my 95% CI will state that I can be 95% confident that the expected percentage decrease by a year older model is between 14% and 15%.

To get the best deals, we will look at the residuals another time in Table 4.

ID	Age	Prices	Km	Power_kW	Pred	Residual
16291520	12	0	299000	78	0	0
16295510	12	0	223189	78	0	0
16217935	8	1	228000	66	1	-0
16241170	9	1	252900	66	0	0
16291795	9	1	353000	81	0	0
16211561	9	1	305400	66	0	0
16082831	9	1	399000	81	0	0
16260283	8	1	320064	66	1	0
16260660	7	1	350000	81	1	0
16157086	7	1	431000	81	1	0

Table 4: Best Renault Megane deals proposed by linear OLS based on Log price - Age relationship

Looks like the winner here is the same as in the level-level case, so now two models suggest that we should look at that Renault used car with careful eyes. All in all, other than the first car, there is lot of common elements between the two tables, so even though we captured the non-linear pattern, the overall picture doesn't seem to be very much different between the two cases.

Summary

I set out to analyse the used car market of Renault Megane models to uncover a pattern of association between the age of a car and their prices. I used linear regression models for my analysis and experimented with log transformation and polynomials, since the pattern between my models seemed non-linear. Out of 4 models (quadratic, cubic, level-level, log-level) my final model choice was a log-level regression, which not only fitted the scatter plot of the variables most, but also produced a spectecular 85% R^2. When looking at residuals the level-level and log-level regression produced similar results, the model implies that there are some underpriced cars in the segment of 5-13 years. But I also noticed that some cars in the model don't fit into the big picture, as they are in a quasi luxurious category, so I can see further improvments down the line that can be done to the model, such as including multiple regressors (e.g. displacement) to further improve interpretation of residual analysis.