MA2822:
Advanced Statistics

Ecole Centrale Paris

Regression Analyisis of Used Car Prices

Professor
Christine Keribin

Authors
Erik Hendey Bröte
Patrick von Platen

March 2, 2017

Contents

1 Introduction

In this project we study the price of used cars based on a number of relevant charectaristics. Using regresion analysis tools implemented in the open source language *R* (which is widely used for statistical anlysis), we aim to model the sales price as well as detect correlations between factors. The dataset includes 172 observations of cars with a 12 charecteristics, including the following: *price*, in thousands of Euros; *age*, in months; *km*, the total milage in kilometers; *TIA*, the number of months untill the next vehicle inspection. The data also includes dummy variables (which are equal to 1 or 0 depending on whether the factor is true or false) such as *ABS* and *SunRoof* representing the presence of ABS-technology and a sunroof repsectively.

2 Charectaristics of and Correlations in the Data

Before procededing it is a good idea to get a general idea of the correlations between the variables. Figure 1 is a scatter plot showing the dependencies between the *price*, *age*, *km*, and *TIA*. It shows a relavitely strong postive correlation between *price* and *age* and some negative correlation between *age* and *km* as well as negative correlation between *price* and *km*. These relationships are in accord with our existing knowledge of the used car market. *TIA* seems to explain the other variables in this model poorly.

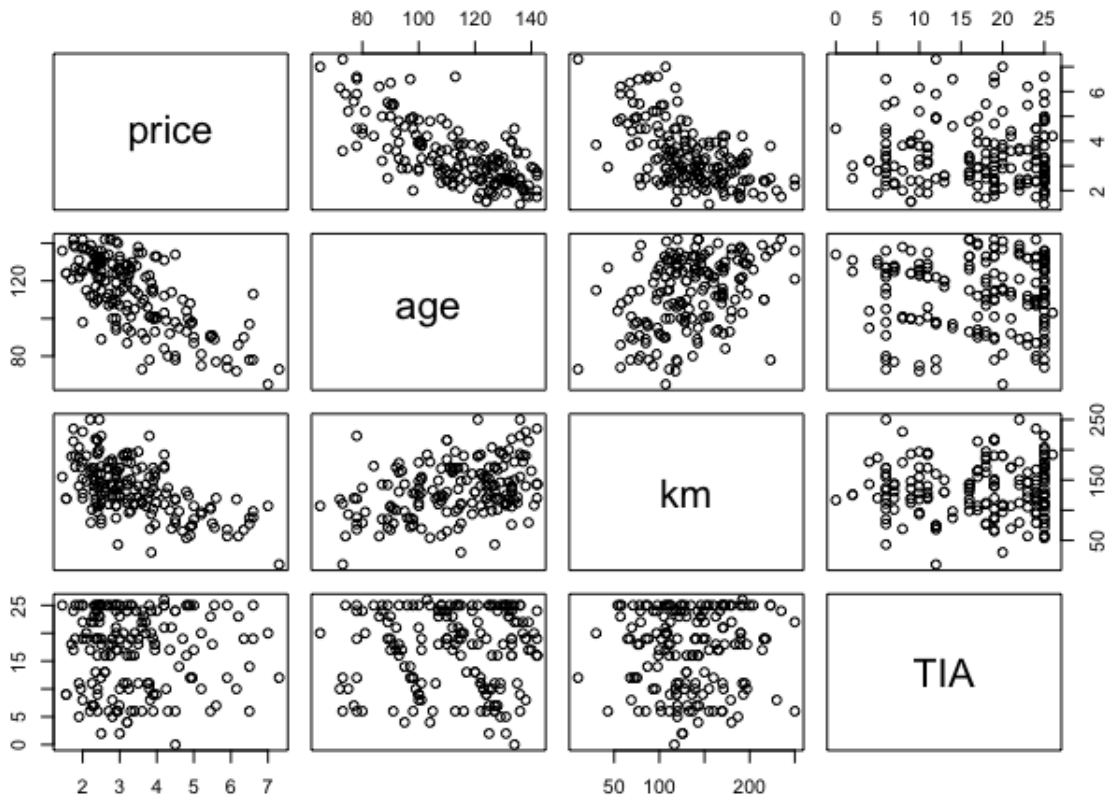

Figure 1: Scatter plot of the variables *price*, *age*, *km*, and *TIA*.

In addition to this we can use boxplots to analyze whether the presence of ABS technology or a sunroof, represented here by the dummy variables *ABS* and *SunRoof*, have an effect on price. Figure 2 displays this analysis for each variable. To the left the boxplots for vehicles with and without ABS show a few perhaps expected characteristics of the dataset. The median price of a car with ABS is significantly higher than their counterparties (compare 3 200 EUR to 2 625 EUR). Also, the variance of the price of cars having an ABS-technology is much lower than for those that without it. Since 50% of all "ABS" cars have a price that ranges between 2.6 and 4 with pretty much no "ABS" car having a price higher than 6, where as 50% of all "Non-ABS" cars have a price that ranges between 2.5 and 4.5 and a maximum of

7. It can be concluded that the price is somewhat correlated to the dummy variable *ABS*, but not very strongly.

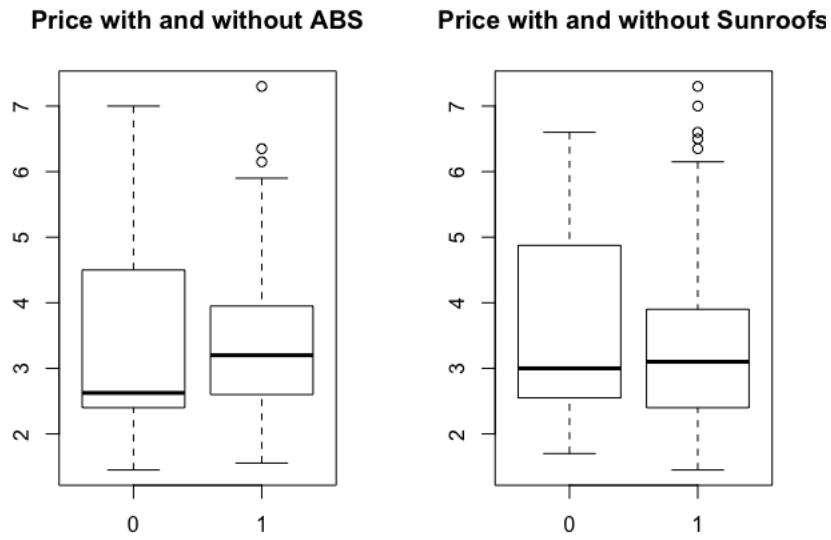

Figure 2: Pairs of boxplots of price of vehicles with and without ABS (left) and sunroofs (right).

A similar analysis can be done for the presence of a sunroof. The second set of boxplots in Figure 2 show values for vehicles with and without them. The medians are quite similar for "Non-OpenRoof" and "OpenRoof" cars. The "box range" around the median is smaller for cars with sunroofs, but apart from that they are no visible differences. The correlation between price and OpenRoof seems to be rather small; further analysis is needed to confirm this observed relationship.

3 Study of price as a function of the prescene of ABS-technology

As seen in Section 2 cars with ABS seem to generaly have a higher price that cars without this technology. We will further analyze how well *price* can be explained by the *ABS*.

To better understand how the *R* software works we created two simple regression models with all variables provided with *ABS* as qualative and quantitative variable. In implementation qualitative means that the variable is given a numerical value, 1 representing the presence of ABS or 0 otherwise. The quantative variable is a logical data type meaning it can be either *TRUE* or *FALSE*. No difference can be seen in the *summary* function (which provides values and different significance metrics for each coefficient) between those two models. It may be pertinant remember that the logical variable cannot be calculated upon, e.g. we cannot calculate the it's mean value. This should not make a difference in the final model as regression requires that the logical variable be implemented as a dummy variable (i.e. qualitavitely) anyways.

The simplest of regressions can be a good way to analyze the explanatory power of *ABS*. We created a regression model of *price* as only a function of *ABS* corresponding to the following model.

$$price = \beta_0 + \beta_1 \times ABS + \epsilon \tag{1}$$

Here β_0 is the intercept and β_1 is the slope of the equation. ϵ is the residual variable. Regression on our dataset resulted in the following estimations. $\hat{\beta}_0 = 3.45363$ and $\hat{\beta}_1 =$ -0.08321. It is clear that the variable *ABS* does not account for any of the variance of the price, since it's coefficent of deterimination, R^2 is very low (0.00097). That means that the model does not do a very good job predicting the price. This fact is also reflected by the p-value for *ABS*, $p = 0.686$. the p-value is the probility of obtaining a result as more extreme given the null hypothisis (which in this case is that the coefficient $\beta_1 = 0$.) For these reasons it is clear that *ABS* on it's own explains the price very poorly. Arguably, the intercept alone would be a better model as the model would just create a nearly constant line.

4 Study of Price as a Function of Mileage

4.1 Simple Regression

A simple regression of *price* on milage represented by the variable *km* shows quite different results. As in Section 3, we begin by performing a simple regression using the equation

$$price = \beta_0 + \beta_1 \times km + \epsilon \tag{2}$$

where β_0 is the intercept, β_1 is the slope and ϵ is the residual term. Applying this to our data results in β_0 and β_1 values of 5.559202 and -0.016051. Using the same analysis and reasoning as in 3, *km* is obviously a better parameter to describe the price of a car. The coefficent of detirmination R^2 is much higher (0.3317) and the p-value is much lower ($< 2e^{16}$), suggesting that *km* should be in model explaining the price.

4.2 Inference from Simple Model

We will now using the simple model from the previous section create confidence intervals of prices for cars with difference milages. This implemented in *R* using the *predict* function. For a cars with a milages of 50 000 km and 135 000 km a confindence intervals at the defult significance level 95% are shown in Table 1.

Mileage	fit	lower	upper
50 000 km	4.756639	4.426392	5.086886
136 000 km	3.392284	3.238505	3.546063

Table 1: Confindence intervals at 95% confidence level for cars with milages of 50 000 km and 135 000 km

We see that the confidence intervall for a car that has 135 000 km milage is much smaller than the other at the same sig. level. This implies that the variance of the price for cars having 135 000 km is lower that can be justified by the fact that the price of cars already having a high usage (in terms of km) are naturally lower and don't vary as much. Whereas for cars having lower usage (50 000 km) the price can vary much more since other factors play a more important role in determining its price (car type, car performance.) Also, it should be noted that 135 000 km is the mean of all of the data. Since the sample follows a t-distribution, it is normal that there is more data around 135 000 km (the distribution is denser). Therefore the 95% confidence intervall has a shorter lenght than the one at 50 000 km.

4.3 Determination of Mystery Variable *kop1*

The data set given includes the variable *kop1* given wit hout explaination of what it charectarizes. In fact, it reduces to the variable *km* as it is a centered and reduced version of it.

$$kop1 \approx \frac{km - mean(km)}{SD(km)} \tag{3}$$

In fact, simple regression models based on *kop1* and *km* result in the exact same model. It can be shown that the models are logically the same since R^2 and other important functions are equal. Algerbrically this can be show through the following reasoning. The regression model is given by the following equation:

$$price = \gamma_0 + \gamma_1 \times kop1 + e \tag{4}$$

Where γ_i are the coefficents and e is the residual. Using relationship 3 the regression model reduces to

$$\begin{aligned} price &\approx \gamma_0 + \gamma_1 \times \frac{km - mean(km)}{SD(km)} + e \\ &= (\gamma_0 + \frac{\gamma_1 \times mean(km)}{SD(km)}) + \frac{\gamma_1}{SD(km)} \times km + e. \end{aligned} \tag{5}$$

Given that $mean(km)$ and $SD(km)$ are constants, we note that this is this approximately equivant with the regression model 2 as $\beta_0 \approx \gamma_0 + \frac{\gamma_1 \times mean(km)}{SD(km)}$, $\beta_1 \approx \frac{\gamma_1}{SD(km)}$ and the residuals $\epsilon \approx e$. It is obvious that centering and reducing shouldn't change anything. The regression looks at how changes in the explainatory variables effect the response variable (here *price*). As simple regression models with *km* and *kop1* are essentially regressing on the same random variable they should result in equally good models with the same metrics such as R^2 which are equivelant (0.3317) and the p-value for the slope ($< 2e^{16}$).

4.4 Polynomial Regression Model

We will now analyze the performance of a polynomial regression model. Given the third degree equation 11 and our data, the results are:

$$price = \beta_0 + \beta_1 \times km + \beta_2 \times km^2 + \beta_1 \times km^3 + \epsilon \tag{6}$$

This is a linear model as it is defined given a random sample

$$(Y, X_1, \ldots, X_N)$$

the relation between the observations Y and the independant variables $X_1, \ldots, X_N$ is formulated as $Y = b_0 + b_1 * f_1(X_1) + \ldots + b_N * f_N(X_N)$, where F_i may be non-linear funcions.

It is interesting to note that the p-values for β_1, β_2 and β_3 are quite high 0.132, 0.839 and 0.766 respectively. The $R^2 = 0.366$ value is slightly better than the simple model. Despite this, it can be concluded that the the third degree polynomial explains the price of cars relatively poorly.

4.5 ANOVA

ANOVA or Analysis of Variance is now performed on the polynomial model. It compares the means of several grouop in order to among other things conclude the significance of each variable. We have implemented this on two regression models. The polynomial model form the previous section as well as a model in which the order of the polynomial terms is reversed. We will refer to them here as Model 1 and Model 2 respectively. The results are shown in the Table 2.

$$price = \alpha_0 + \alpha_1 \times km^3 + \alpha_2 \times km^2 + \alpha_1 \times km + e \tag{7}$$

Variable	km	km^2	km^3
Model 1	$< 2e^{-16}$	0.00313	0.76573
Model 2	0.132	$7.444e^{-9}$	$2.104e^{-12}$

Table 2: P-values for each variable in Model 1 and Model 3

When applying the anova method to the two models, we can see different results even though the three input variables are the same. This is due to the order the input varibles are taken in as an input. In our case, *anova()* determines how much variance is explained by the first entry (km e.g.) and tests its significance, then what portion of the remaining variance is explained by the next variable (km^2) and tests its significance and so forth. Thus, the remaining portion will differ depending on the first variable being inserted and therefore different significances (p-values) is the result.

4.6 Variance Inflation Factors

In order to calculate the vif of a model, we have to take the three input variables being in our case km, km^2 and km^3 and do a linear regression for all each using the other remaining two variables as the input variables (e.g. for km: $km\ I(km^2) + I(km^3)$). Then we take the coefficient of determination of each model (R^2) and apply the formula $1/(1 - R^2)$. This results in the variance inflation factor (VIF) of each variable. For our dataset the *vif* function from the *car* package resulted in the values shown in Table 3.

Variable	km	km^2	km^3
VIF	165.3163	787.3501	262.5217

Table 3: VIF values for each variable in the polynomial model.

It can clearly be seen that all the VIF values are high meaning that the R^2 is relatively close to one. That implies that the function predicts the actual values quite well. The largest VIF is that of km^2, which is logical because km^2 can easily be modeled by summing up $(factor * km)$ and $(km^3/factor)$ where as it is harder to model km by taking its square and cube because their functions are both bigger than km.

4.7 Regression on Orthogonal Base

In order to check whether the "kop variables" are centered and reduced, we wrote a small function that checks wheter the mean is close to zero and the standard deviation close to one.

Listing 1: check center and scale funcion in R

```
centReduc <- function(var){
  mean = all.equal(1,sd(var))
  vari = all.equal(0,mean(var))
  ls <- list(mean, vari)
  return (ls)
}
```

Executing this function, we can see that all three variables are centered and reduced. Additionally, we checked whether the covariances between each of the variables is 0 proving that they are orthogonal.

We will now analyse the performance of the obtained new polynomial regression model. The result is:

$$price = \beta_0 + \beta_1 \times kop_1 + \beta_2 \times kop_2 + \beta_1 \times kop_3 + \epsilon \tag{8}$$

Despite this model having the same R^2 value as the model in subsection 4.4 0.366, we have to note that the p-values for β_1, β_2 are much lower than in the previous polynomial regression model. As a conclusion, it can be said that this model will probably do a slightly better job predicting the price than the one before, but has a bias that is still too high ($R^2 = 0.366$).

Finally we will show, that all three "kop variables" can be constructed from the original variable "km". Since "kop1" derives from the same vector space as "km", "kop3" from the same as "km^2" and "kop3" from the same as "km^3" we know that we can write the three variables as the following functions:

$$kop_1 = \alpha + \beta_1 \times km \tag{9}$$

$$kop_2 = \alpha + \beta_1 \times km + \beta_2 \times km^2 \tag{10}$$

$$kop_3 = \alpha + \beta_1 \times km + \beta_2 \times km^2 + \beta_3 \times km^3 \tag{11}$$

Using the "lm-function" in R, we calculated the α and the β_i for every "kop variable" and showed afterwards, that the "kop variables" created by us using the above are equal to the "kop variables" of the data set.

5 Study of Price as a Function of All Variables

In this section, we will first calculate a linear regression model that takes all variables of the dataset into account in order to predict the price of a used car. In the following, we will use different methods do decide on which variables will be used in our "final model". Therefore, different factors will be taken into consideration, such as:

- the *"R-squared"* value showing how well the predictions fit the actual data and detected *high biais*,
- factors making sure the data is not *overfitting*, such as *BIC*, *AIC* and *adjusted "R-squared"*
- and factors using cross-validation methods in order to check whether the model is useful for data the model was not overfitted on our dataset (*BIC*, *AIC*).

5.1 Prelimenary Selection of Different Models

/noindent The first model we want to test later is simply the model that uses all variables in order to predict the price:

$$M_1 = lm(price \sim \ .) \tag{12}$$

/noindent The second model is obtained by looking at the p-values of every input variable using:

Listing 2: summary() in R

```
1        summary(Model1)
```

/noindnet There are three variables that have a significant lower p-value than the other variables, *age, km and ageop2.* Therefore our second model is as follows:

$$M_2 = lm(price \sim \ age + km + ageop2) \tag{13}$$

/noindnet Our third model was achieved by using stepFunction: "stepAIC". This function calculates the *AIC* value of the model at every step and removes all the variables, so that the remaining model has a lower AIC and is therefore a "better fit". Using this method, we get the following model:

$$M_3 = lm(price \sim \ age + km + ageop2 + kop2) \tag{14}$$

/noindnet Our fourth and last model was created using a model of preselected variables and executing the "stepAIC" function on this model in order to choose additional variables. The It is in some way a mixture of the methods used to obtain model2 and model3. We preselect the variables *age, km* since they show by far the highest p-values (see listing 2). Applying the "stepAIC" function, we get:

$$M_4 = lm(price \sim \ age + km + ageop2 + kop2) \tag{15}$$

/noindent This is the same model as model 3 and can therefore be discarded.

5.2 Validation of the Models

/noindnet First, we want to calculate the different *"R-squared"* values to select models having low biases when predicting the data. For all three models we get similiar results.

	M_1	M_2	M_3
R^2	0.65833	0.64338	0.65581

/noindent The closer *"R-squared"* is to 1, the better it is. Obviously, the first model is the closest to one since it uses all the variables for its model. We can see though that all *"R-squared"* are very close making it difficult to pick a clear winner. Further analysis is needed.

/noindnet To get in-depth analysis, we will use the *CV* function of the package *forecast*. This function calculates the *BIC*, *AIC*, *AICc* and *adjusted "R-squared"* for every model and returns a so-called *PRESS* value that takes into consideration all the factors mention above in the first part of Section 5 (especially the second and third factor). The lowest *PRESS* suggest that the corresponding model is best relative to the others. The result of the *CV* function is displayed in the table below:

```
> CV(Mfirst)
        CV         AIC        AICc         BIC       AdjR2
 0.5970033 -88.0296131 -86.3796131 -53.4071739   0.6393558
> CV(Msecond)
        CV         AIC        AICc         BIC       AdjR2
 0.5768241 -92.6608708 -92.2994250 -76.9233984   0.6370135
> CV(Mthird)
        CV         AIC        AICc         BIC       AdjR2
 0.5615731 -96.7643674 -96.2552765 -77.8794006   0.6475706
```

Figure 3: *PRESS* analysis of the models.

As we can see, we get the lowest CV for the third model. Our model from the preliminary selection in section 5.1.

6 Conclusion

Having taking into consideration model evaluation factors, such as:

- *R-squared*
- *R-squared adjusted*
- *BIC*
- *AIC*
- *AICc*
- and others

we chose **model3** as our model that will best predict the price of used cars.

$$M_3 = lm(price \sim \ age + km + ageop2 + kop2) \tag{16}$$

To conclude, we would like to note that even if this project is rather "small", it gave valuable experience in how to strucuture an approach to a modelling problem and how to arrive at a result that is logically well founded.