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 charecteristics. 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 or 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 postitive 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.

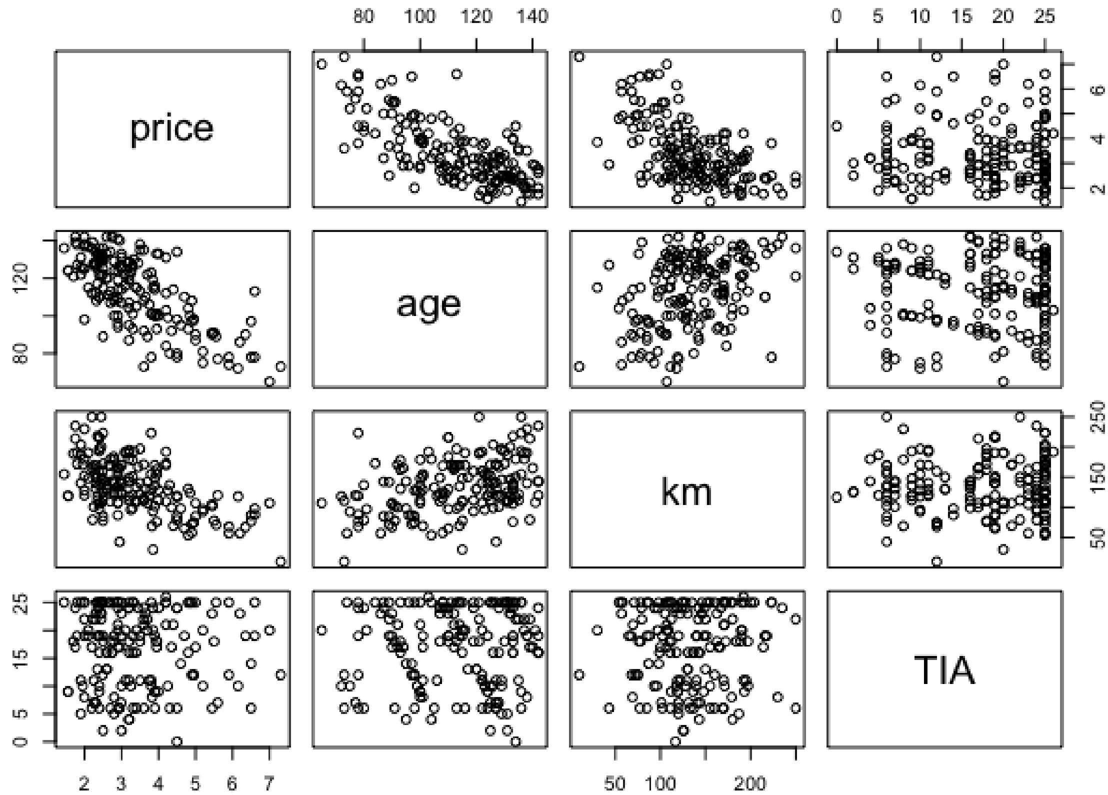

Figure 1: Scatter plot of the variables *price*, *age*, *km*, and *TIA*.

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 charectaristics 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.

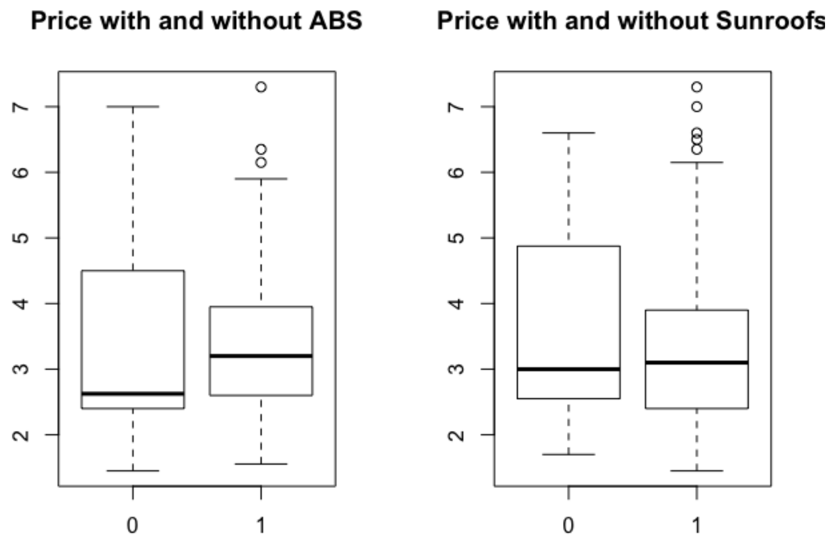

Figure 2: Pairs of boxplots of price of vehicles with and without ABS (left) and sunroofs (right).

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 term. 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

$$price = \beta_0 + \beta_1 \times km + \epsilon \quad (2)$$

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.

$$kop1 = km - mean(km) \tag{3}$$

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

$$price = \gamma_0 + \gamma_1 \times (km + mean(km)) + e = (\gamma_0 + \gamma_1 \times mean(km)) + \gamma_1 \times km + e \tag{5}$$

Given that $mean(km)$ is a constant, we note that this is this equivant with the regression model 2 as $\beta_0 = (\gamma_0 + \gamma_1 \times mean(km))$, $\beta_1 = \gamma_1$ and the residuals $\epsilon = e$.

It is obvious that centering and reducing shouldn't change anything. The regression looks at how changes in the X value affect the Y vaule