MA2822:
Advanced Statistics

Ecole Centrale Paris

Regression Analyisis of Used Car Prices

Professor

Christine Keribin

Authors

Erik Bröte

Patrick von Platen

March 2, 2017

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 distance traveled 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 correlation between *price* and *age* and some correlation between *age* and *km* as well as *price* and textitkm. *TIA* seems to explain the other variables in this model poorly.

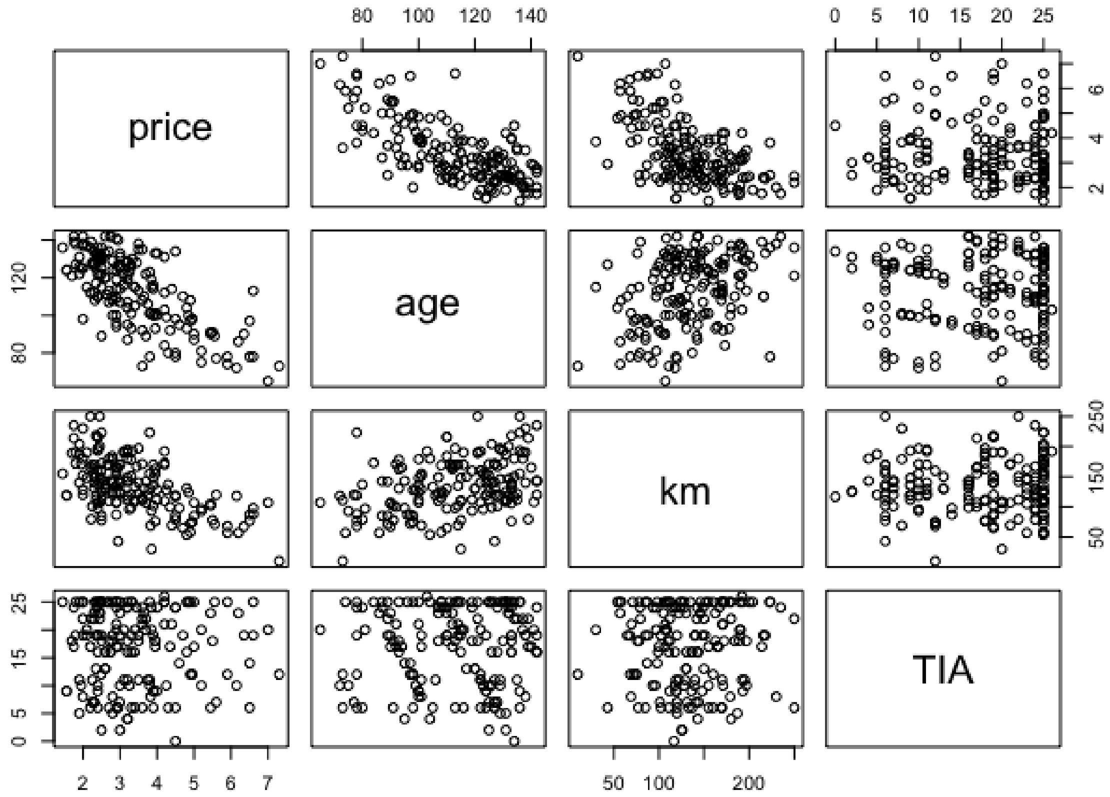

Figure 1: Scatter plot of the variables *price*, *age*, *km*, and *TIA*.

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.

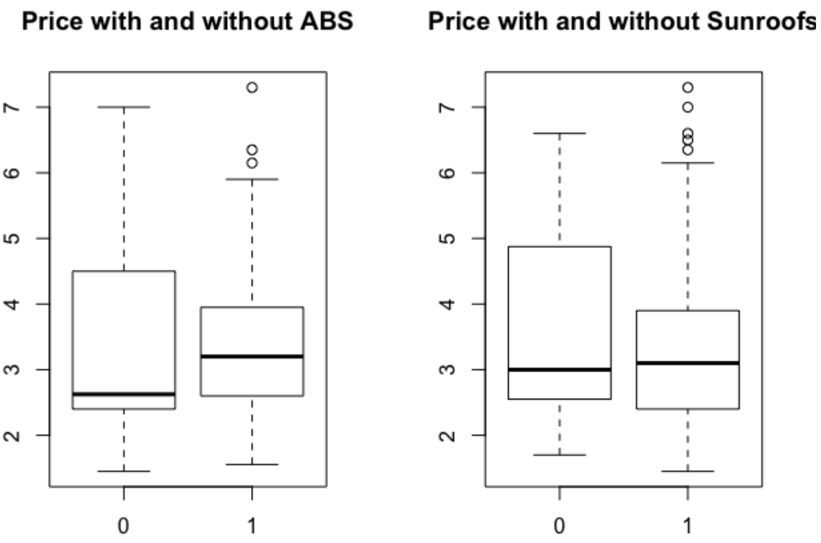

Figure 2: Pairs of boxplots of price of vehicles with and without ABS (left) and sunroofs (right).

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

COMMENTS: No difference can be seen between those two models, so there doesn't seem to be a difference. It may be pertinant remeber that the logical variable cannot be calculated upon, e.g. we cannot calculate the average of *loglDf$ABS*. Though this should not make a difference in the final model as regression requires the logical variable to implemented as a dummy variable (i.e. qualitavitely) anyways.

COMMENTS: It clearly can be seen that the variable "ABS" does not account for any of the variance of the price, since it's R^2 is very low (0.00097). That means in general that the model does not do a very good job predicting the price Also, it can be seen that our "Intercept" value nearly completely explains the price (look at the parameter values or $Pr(>|t|)$). Therefore, our prediction would just create a constant line (when plotting our pred), showing that the biais is too high.

COMMENT: WE SHOULD EXPLAIN THE MEANING OF EACH OF THE NUMBERS WE GET WHEN DOING LINEAR REGRESSION! THEN IT IS CLEARER AND EASIER TO UNDERSTAND THE CONCLUSION WE ARE ARRIVING AT TAKING INTO ACCOUNT A CERTIAN LINEAR REGRESSION

4 Étude du prix en fonction du kilométrage