PakWheels Data - Final Project

Khawaja Hassan Abbas

12/14/2021

Overview

The purpose of this Project was to analyze how the prices of used Honda Civic Oriel 1.8 i-VTEC variate on the bases on their total mileage.Moreover, to determine a real relationship between these two variable without exaggeration we introduced confounding variables to our statistical model.The confounding variable that we took into our different models are the registered city,color and distinct features. The description of the variables are as follows:

- Price: The Re-sale value of car in Pak Rupees [Exchange rate 1 PKR = 0.005 Euros]
- Mileage: The number of Kilo-meters the car has done at the time of sale [Km]
- Registered City: The city were the car is registered
- Features: The distinct features which come along with each cars
- Color: The 4 most common colors which were being sold

Lastly the data set used in our analysis is Pak-Wheel Used Cars Data from **Kaggle**.

Data Cleaning & Munging

Our data consisted of more than 56000 observation and to narrow down our analysis we had to munge the data into a useful and use-case-specific form.Therefore, several forms of cleaning procedure were preformed to make sure that our data is ready for our downstream analysis. Now we will be discussing the pipeline for cleaning process that we opted before we could use it in our statistical model.

Like we have mention above that raw data consisted of 50000 plus observation and 150+ different cars.The first step in our cleaning process was to filter out only specific variant and then used that particular variant in our statistical model. After initial filtering we were left with 382 observations but the data was still unstructured and required us to individually scrutinize the independent and the confounding variables. Once we went down the funnel of cleaning process we also dropped the unnecessary columns in our data set.

- **Price & Features** To start off our data munging we first had to check the number of NA values in each variable and see decide if we can drop them or not. Fortunately, in total there were only 10 NA values (9 in features and 1 in price) so we drop them.One of the reasons for dropping them was that other variables were also missing for these specific observation.

The problem that we had with the features were that almost every car had more or less the same features. Therefore, we had to look and see what were the distinct features which come out to be the `Climate Control` and `Navigation Feature`.Lastly, with the price variable had to be converted into integer since they were initially in character format.

- **Registered city & Color** After this we filtered our data on two main criteria , the specific cities and color of the cars. For cities we took the federal capital and two metropolitan cities **(Islamabad, Karachi, Lahore)** & for the color we took the top 4 colors which were being sold. The reason for filtering our data based on these factors was that more than 90 % of our observation belonged to these specific cities and had these colors.However, one additional cleaning that was done to remove the only observation where the car was unregistered.
- **Creating Binary variables** Before we moved to the next process we dropped the columns that were not relevant to our analysis. The next process was to creating binary for our confounding variables so we could use them as dummy variables in our statistical model.However, for features variable we first had to use to `str_detect` function to display logical argument if our distinct features lies in those observation or not. After creating binary variable we run `datasummary_skim` to check if we needed to take log or our given variables or not.Moreover, we also used scatter plot to see if the observation were clustered or well spread. After seeing the result from Exhibit 1 we decided to take log on mileage since the data was right skewed and we had to normalize the distribution. (Exhibit 2 shows scatter plot with log_mileage)

Table 1: Descriptive statistics

	Mean	Median	SD	Min	Max	P5	P95
Price	3 302 258.77	3 300 000.00	188 785.64	1700000	3800000	3 150 000.00	3 550 000.00
Mileage	49 291.58	47 000.00	24 112.04	4500.00	320 000.00	19 000.00	80 050.00
Mileage(log)	10.71	10.76	0.46	8.41	12.68	9.85	11.29
Karachi Register	0.16	0.00	0.37	0.00	1.00	0.00	1.00
Islamabad Register	0.37	0.00	0.48	0.00	1.00	0.00	1.00
Lahore Register	0.47	0.00	0.50	0.00	1.00	0.00	1.00
Climate control	0.08	0.00	0.27	0.00	1.00	0.00	1.00
Navigation	0.24	0.00	0.43	0.00	1.00	0.00	1.00

Correlation Matrix

To get an overview of how our variables are associated with one another we created matrix to extract the correlation coefficient for each of them. The correlation matrix shows the level of association of price with our dependent and confounding variables.Looking at the matrix we were able to comprehend that there was positive association of price with Islamabad registered car & White color car.On the other hand, the price was negatively associated with the other two cities and the remaining colors.One interesting find that this matrix indicated was that there was no as such correlation between the the distinct features and the price of the color. You can refer to the matrix in the appendix as Exhibit 3.

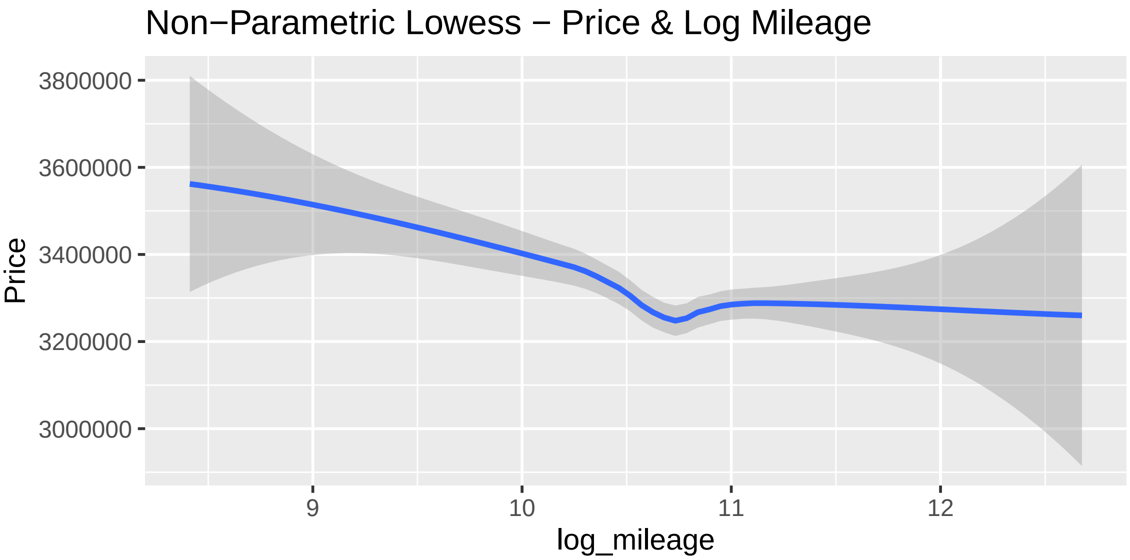
Regression model

Before running the regression model we make a hypothesis that the expected association with Registered cities was that the car prices in Karachi's will be relatively low as compare to Islamabad.The primary reason is that Karachi being a coastal city, the level of humidity is high leading to the car parts to erode at much faster rate. Moreover, the city infrastructure and traffic condition are not optimal in comparison to other two cities which can also cause the re-sale value to deplete at higher rate.

The second hypothesis was about cars having black color will be having prices comparatively lower. The rationale behind this is that due to dusty atmosphere in the these cities and warm weather condition there is general prejudice in the overall Pakistani population against this color.

Lastly, Climate control and Navigation systems are some luxury features that will be creating some value addition in the price of the car. Therefore the hypothesis is that cars having these features will create a positive change in the re-sale price.

Moving on, using Non-Parametric Lowess method we first checked if the we had to incorporate splines in our model for log mileage. However,as we can see in the curve below the curve was moving in a similar direction without any unusual variation so there was no need for splines.Now we run 4 different regression models and one by one adding the confounding variable and seeing the significance and level of magnitude.



Price VS Log Mileage

$$Price := \beta_0 + \beta_1 log(Mileage)$$

The first regression model is level-log model showing if our Mileage change by one percent our independent variable will negatively change by PKR 930 on an average. The R squared in our model is approximately around 5% , indicating that around this percentage of variation in price is explained by the log mileage & remaining is left for the residual variation.Moreover, Confidence interval suggest that at 95% significant level, the true beta coefficient in the population will be between -1677.67 to -184.34 on an average which is significantly different from zero. This means that with one percent of a km change the price will be lower within the given range on an average.

Price VS Log Mileage + Registered City

$$Price := \beta_0 + \beta_1 log(Mileage) + \beta_2 Lahore + \beta_3 Karachi$$

The second model introduce the dummy variable in the regression equation based on the variable Registered city. Since we initially had an expected relationship that prices of cars with Islamabad registration will be higher so we took it as the base in our statistical model.The beta coefficient of this model states that keeping other factors constant the price of cars which are registered in Karachi will be PKR 107533 lower as comparison to car registered in Islamabad on an average. The confidence interval for these two dummy variable are 90% & 99% respectively.

Price VS Log Mileage + Registered City + Color

$$Price := \beta_0 + \beta_1 log(Mileage) + \beta_2 Lahore + \beta_3 Karachi + \beta_4 White + \beta_5 Grey + \beta_6 Silver$$

The third model accounts for the second confounding variable which is the color of the car. Here we have taken black color as our base and the run our statistical model. The beta coefficient for the white states that keeping everything else constant,in comparison to black color,the prices of white will tend to be PKR 67000 higher on an average. The confidence interval for white is 99% , where as for the other colors it tends reflect zero level of confidence.

Price VS Log Mileage + Registered City + Color + Features

$$Price := \beta_0 + \beta_1 log(Mileage) + \beta_2 Lahore + \beta_3 Karachi + \beta_4 White + \beta_5 Grey + \beta_6 Silver + \beta_7 Climate + \beta_8 Navigation$$

The last model in our analysis is adding our distinct features as confounding variable and analyzing the overall association and changes in our beta coefficients. The result shows that cars having navigation and climate control feature will be having a price PKR 15500 and PKR 280 higher respectively. However, if we see the model the confidence interval is zero which that we are uncertain whether there is a effect or not. Having zero in confidence interval implies that this confounding could have a negative or positive effect on the outcome. The summary table for all models are mentioned in the appendix as Exhibit 4.

Conclusion

To conclude, we saw how some of our hypothesis were validated by our regression models and how some were rejected.The model 2 authenticate our hypothesis that cars which were registered in Karachi had a comparatively lower price as compare to other cities.Moreover, when comparing model 1 and model 2 we witness it was useful to add Registered cities as the R squared increased, also the coefficient of these two variable are significant at 99% confidence interval.

Moving forward as we add the dummy variable for colors we were able to deduce from model 3 that the price of white cars will be comparatively higher from black car,at a 99% Confidence interval.However, since we have zero confidence interval for the remaining colors we cannot validate that our initial hypothesis we made about the black color cars.

In our last model we were aiming to validate our hypothesis about features creating positive association with the re-sale price of the car.However,our regression model stated that there is no significant relationship between these variables and the price of the car.

Our preferred model is ;

$$Price := \beta_0 + \beta_1 log(Mileage) + \beta_2 Lahore + \beta_3 Karachi + \beta_4 White + \beta_5 Grey + \beta_6 Silver$$

Generally, you choose the models that have higher adjusted and predicted R-squared values.The adjusted R squared increases only if the new term improves the model more than would be expected by chance and it can also decrease with poor quality predictors.Our preferred model is Model 3 where we accounted for the Registered City & the Color confounding variable. The reason for the selection of this model is that with this model our Adjusted R squared comparatively increased from 8% to 9%. Moreover,our beta coefficient for the dependent variable also changed showing a better magnitude of association.Lastly, the question maybe be raised that why R2 is below 10 % for all these models.The reasons for this can be that the variation in prices can be explained much better if we consider other variables such as price of petrol, prices of their competitors and even the prices of previous model of Honda.

Appendix

Exhibit 1

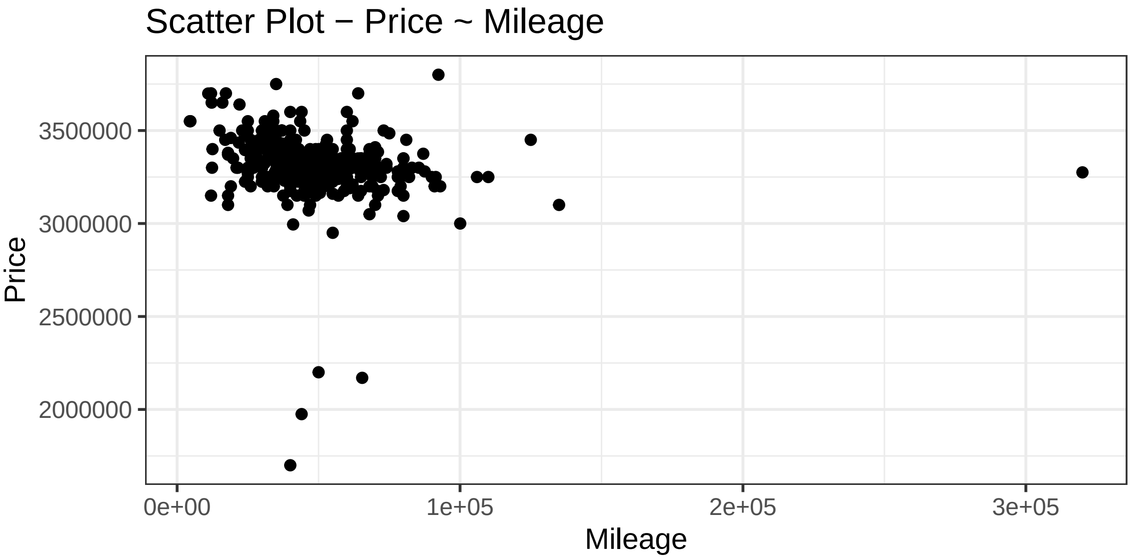


Exhibit 2

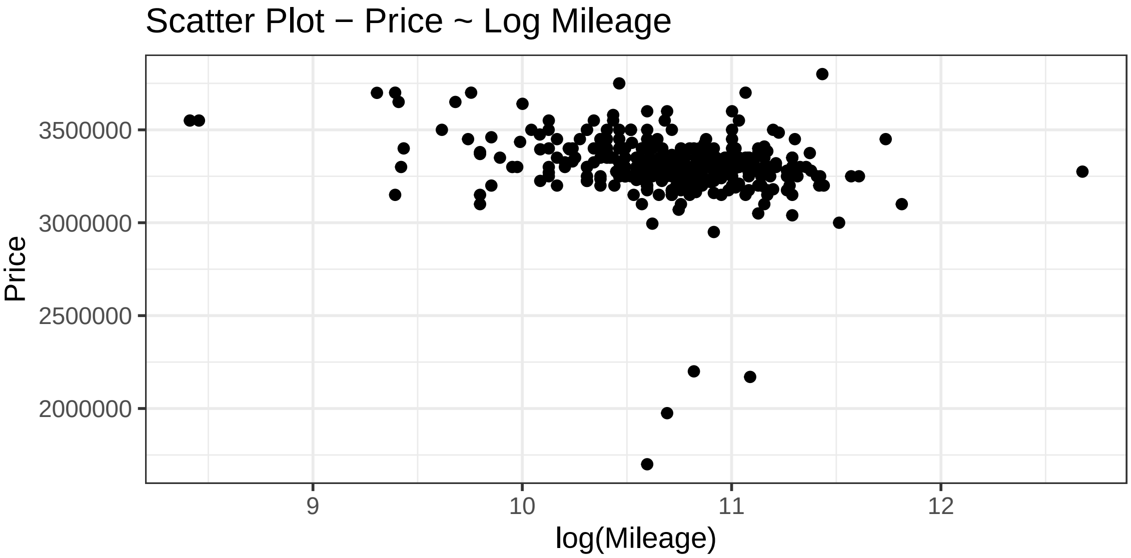


Exhibit 3

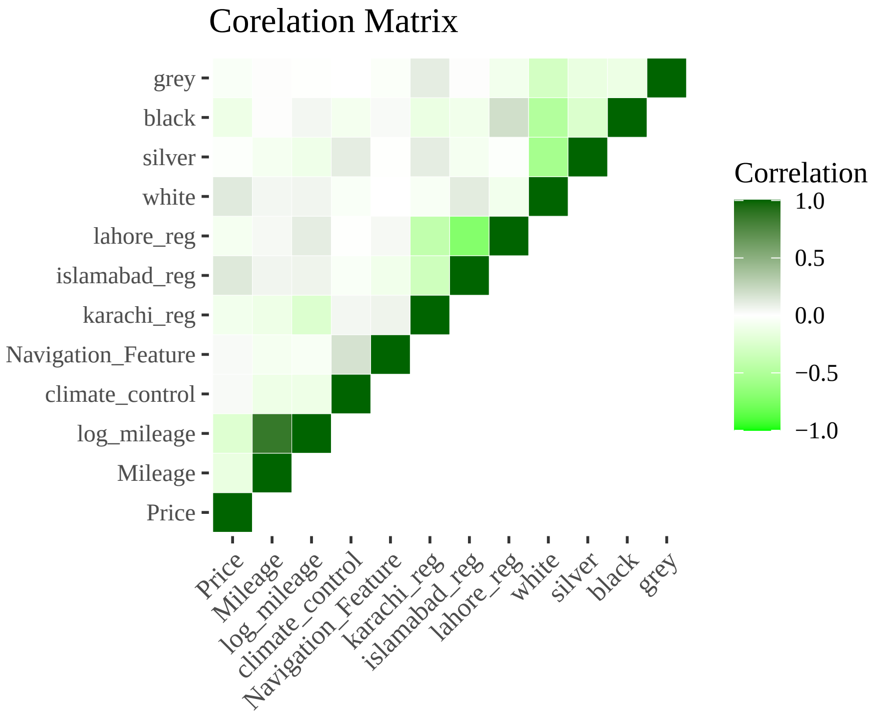


Exhibit 4

Table 2: Regression Model Summary

	Model 1	Model 2	Model 3	Model 4
Intercept	4 298 652**	4 501 742**	4 464 341**	4 455 865**
	(200 178)	(204 040)	(216 338)	(216 270)
Mileage(Log)	−93 055**	−108 339**	−109 506**	−109 027**
	(18 678)	(18 955)	(19 504)	(19 425)
Lahore Registered		−46 835*	−38 409	−39 482
		(20 911)	(23 098)	(23 554)
Karachi Registered		−107 534**	−105 698**	−107 666**
		(38 341)	(36 342)	(37 265)
White Color			67 624*	67 954*
			(32 099)	(32 325)
Grey Color			28 638	29 915
			(37 117)	(37 453)
Silver Color			35 526	36 317
			(35 156)	(35 480)
Climate Control				282
				(28 235)
Navigation Feature				15 556
				(23 243)
Num.Obs.	340	340	340	340

* $p < 0.05$, ** $p < 0.01$