Estimating the Impact of Mileage on Used Car Pricing in Indian Market

Srila Maiti

Introduction

As today's automobile industry is facing the car-chip shortage across the world and the entire globe is facing the unstable economic condition caused by pandemic and inflation, the buyers are shifting their focus towards the pre-owned cars.

The car experts provide some broad guidelines and data-based approaches to estimate the factors affecting the market value of the used cars. It has been traditionally thought that mileage is the main driver for used car valuation, but I seek to examine other factors that may influence the price of used cars, such as the car manufacturers, and regular or luxury cars, fuel type, region.

In this study, I would like to answer the below question:-
How much effect the mileage has on the used car valuation in Indian car market?

Data and Methodology

The original observational data set is sourced from Carsdekho.com via scraping through selenium, scrapy and beautifulsoup and was made available in Kaggle as a public data set, linked here UsedCars_Combined.csv. Each row in the data set represents the market value of a pre-owned car in Indian market, aging between 2004 to 2021 in Indian Rupees (INR). There are a total of 1715 rows and 9 columns in the auto dataset and 30% exploration data set is created by stratification sampling based on year and fuel type and remaining 70% records are used as the confirmation data set to generate the statistics in this report.

Table 1: Excluded Auto Records

Filter	Records Removed	Justification
Drop prior 2000 cars	1	Very old car
Drop CNG fuel type	8	Rare fuel in India
Drop Hybrid fuel type	1	Rare fuel in India
Drop cars with less than 2000 Km mileage	7	Farley new car

Table 2: Transformation Applied

Transformation	Justification
price transformed in unit of 100,000 INR	Most car prices are above 100,000 INR(known as lakhs).
distance transformed in unit of 10K km	Scale down mileage in smaller numbers
fuel_type to fuel_type_cleaned	'Petrol + 1' to 'Petrol' and rest fuel type as is
city transformed to region	Merged suburbs to larger geographic region

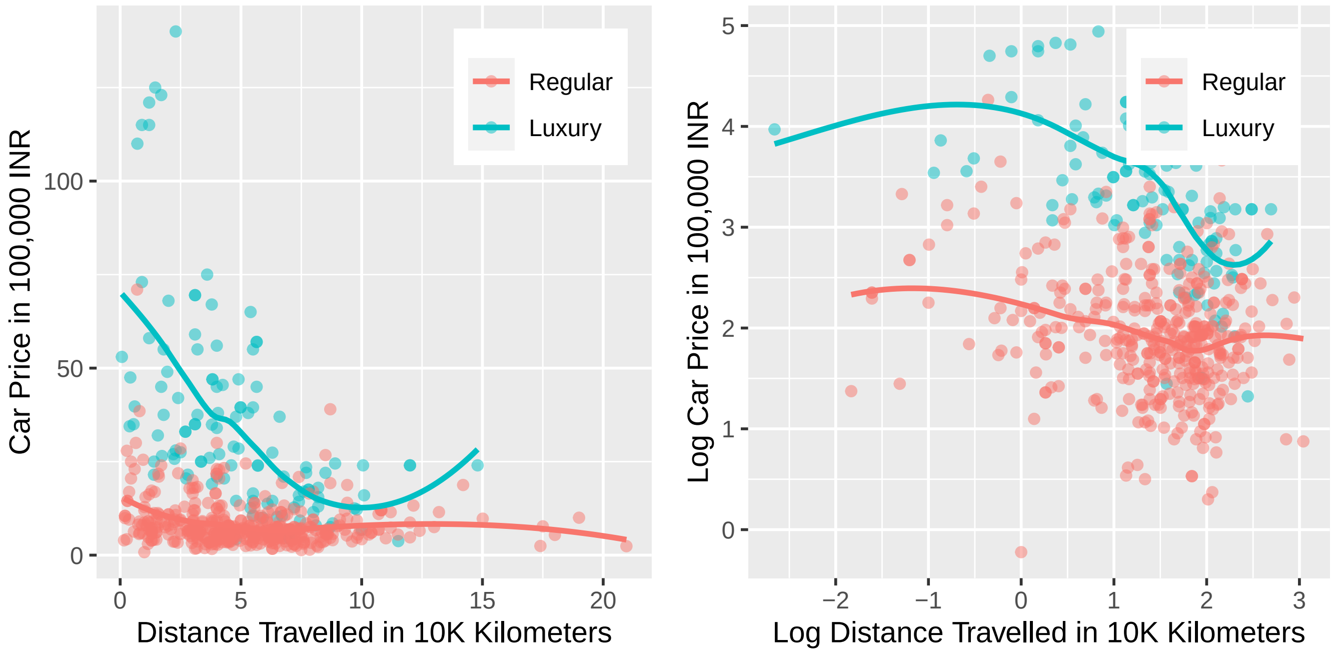
Transformation	Justification
brand to generate car_manufacturer and car_brand	Car manufacturer and car brand can be different
Added column luxury_vs_regular	Based on background knowledge of the car brand.
Log transformation on price_in_100k_inr	Skewed distribution
Log transformation on distance_travelled_in_10k_km	Skewed distribution
Added column car_age_in_years_during_sale	Car age information
Added column immediate_sold	Denotes if the car is immediately sold or not

- Operationalization:-

Table 3: Operationalization

Operationalize	Actual Column Used	Type
Used car market value	log_price_in_100k_inr	Outcome variable, represented as Y
Distance travelled	log_distance_travelled_in_10k_km	Independent variable, represented as X

  - I decided to take the log transformations for both explanatory and outcome variables due to the skewness of the distributions.
  - Excluded Features:-
    - ∗ From the original data set with 9 features (id, year, brand, full_model_name, model_name, price, city, distance_travelled_kms, fuel_type)
      - · I have dropped id column as it does not add any business value.
      - · I have also dropped full_model_name column as model_name is already present.



I am interested in the difference in value between two counterfactuals: used car's market valuation with distance traveled, and used car's market valuation for luxury vs regular car.

I created a base regression model using car price and mileage to see the effect of mileage on the price. I also built few other regression models to see the effects of other factors like fuel type, luxury vs regular car, car age and car manufacturers and region.

$$Z\ is\ a\ row\ vector\ of\ additional\ covariates\ andgamma\ is\ a\ column\ vector\ of\ coefficients.$$

$$log(\widehat{price\ in}\ 100k) = \beta_0 + \beta_1 \cdot log(distance\ in\ 10k\ km) + \mathbf{Z}\gamma$$

$$log(\widehat{price\ in}\ 100k) = \beta_0 + \beta_1 \cdot log(distance\ in\ 10k\ km) + \beta_2 \cdot petrol + \beta_3 \cdot regular + \mathbf{Z}\gamma$$

$$log(\widehat{price\ in}\ 100k) = \beta_0 + \beta_1 \cdot log(distance\ in\ 10k\ km) + \beta_2 \cdot petrol + \beta_3 \cdot regular + \beta_4 \cdot car\ age + \mathbf{Z}\gamma$$

$$log(\widehat{price\ in}\ 100k) = \beta_0 + \beta_1 \cdot log(distance\ in\ 10k\ km) + \beta_2 \cdot petrol + \beta_3 \cdot regular + \beta_4 \cdot car\ age + \beta_5 \cdot manufacturer + \mathbf{Z}\gamma$$

$$log(\widehat{price\ in}\ 100k) = \beta_0 + \beta_1 \cdot log(distance\ in\ 10k\ km) + \beta_2 \cdot petrol + \beta_3 \cdot regular + \beta_4 \cdot car\ age + \beta_5 \cdot manufacturer + \beta_6 \cdot region + \mathbf{Z}\gamma$$

Results

Table 4: Estimated Car Market Value

	Outcome Variable: Log Car Price in 100K INR				
	(1)	(2)	(3)	(4)	(5)
Log Distance in 10K Km.	-0.19^{***} (0.03)	-0.34^{***} (0.02)	-0.03 (0.03)	-0.01 (0.02)	-0.01 (0.02)
Petrol		-0.58^{***} (0.04)	-0.32^{***} (0.03)	-0.21^{***} (0.03)	-0.21^{***} (0.03)
Regular		-1.09^{***} (0.04)	-1.24^{***} (0.03)	-1.45^{***} (0.09)	-1.43^{***} (0.08)
Car Age			-0.13^{***} (0.01)	-0.13^{***} (0.005)	-0.13^{***} (0.005)
Constant	2.59^{***} (0.05)	3.87^{***} (0.05)	4.27^{***} (0.04)	4.32^{***} (0.08)	4.30^{***} (0.08)
Car Manufacturer				✓	✓
Region					✓
Observations	1,200	1,200	1,200	1,200	1,200
R^2	0.03	0.56	0.70	0.83	0.83
Residual Std. Error	0.86 (df = 1198)	0.58 (df = 1196)	0.48 (df = 1195)	0.37 (df = 1168)	0.37 (df = 1164)

Note: *p<0.05; **p<0.01; ***p<0.001
HC_1 robust standard errors in parentheses.

- Table 4 shows the results of five representative regressions.
  - Coefficient for log Distance in 10K Km. is significant in the first model with very little explanatory power of 3 %. Every 100 km increase in mileage leads to a price drop by INR 190, using small number approximation.
  - On average, holding all other factors constant, petrol cars show less depreciation over diesel cars according to model 2, 3, 4 and 5. Model 5 shows, on average, holding all other factors constant, petrol cars will have INR. 810.58 less depreciation than diesel cars.
  - On average, holding all other factors constant, regular cars show less depreciation over luxury cars according to model 2, 3, 4 and 5. Model 5 shows, on average, holding all other factors constant, regular cars will have INR. 239.30 less depreciation than luxury cars.
  - On average, holding all other factors constant, according to model 5, with every 1 year of aging, causes 13% car price depreciation, that is INR 13,000.
  - And finally, car manufacturer plays a major role in the used car's market valuation. Adding this factor in model 4, brings back model explanability power to 83% and also reducing the residual standard error to .37.
  - Coefficient for log Distance in 10K Km. is no longer significant after adding the regular car indicator, car age and car manufacturer and region.

Limitations

Consistent regression estimates follow the assumption of independent and identically distributed (IID) observations. Car price can fluctuate in various regions. So, there is a possibility of geographical clustering. I partly accounted for geographic clustering in model 4 and 5, by including a fixed effect for each region that is interacted with car price estimate. In other words, each region has a unique slope and linear trend over car price. I am not able to account for minute geographical clustering within each region.

Because car sales happen over a significantly long-time frame, there is a further possibility of temporal autocorrelation. Car price in recent years will be more than that of past car prices, in other words, the same car worth more in 2022 than it 5 year back.

Both the car mileage and car price have skewed distribution, violating the normal distribution. As a result, I have taken log distribution for both variables.

Consistent regression also requires unique BLP and to satisfy unique BLP (Best Linear Predictor), we need to have finite mean and variance. However, based on the histogram for both distance traveled and car price, I see that they are largely right-skewed (Pareto Distribution) and also contains several outliers. With and without outliers, the distribution look very different, which implies that there may not be a unique BLP.

Car data set may have some inherent problems. More international cars will be available in recent years than that of in the past and year wise car data can vary widely. To cater this problem, I used a stratified sampling to build a more representative exploration set and used the remaining as the confirmation set.

As far as structural limitations, several omitted variables may bias my used car's market value estimates. In a classic omitted variables framework, the omitted variable is assumed not to interact with the key variable in the true model. An example of such variable is car accident history. More mileage can cause more accidents thus affecting the outcome variable or because of more accidents, car mileage will be less (common ancestor / classic omitted variable). Other omitted variables can be maintenance, the more a car is maintained, the better resell value it offers. Thus, the depreciation value would be lower in such cases. Number of past car owners can affect the car market value. If the car purchase history shows fewer ownership changes, it has a better chances of getting good resell value. If the car is used for business purpose (Ex. Uber), car market value will be affected. I predict that all these omitted variables likely to have negative effect on the car market value in the true model, I predict a positive omitted variable bias on the key variables and the main effect is therefore being driven away from zero, making my hypothesis tests overconfident.

Conclusion

This study estimates how the various factors affect the market value of the pre-owned car in India. We notice that in Indian car market, petrol cars have lesser depreciation compared to diesel cars and luxury cars depreciate much faster than regular economy cars. Alhough, more car mileage causes the car depreciate faster, we see the mileage effect fades off with the factors of car fuel and car manufacturer. In other words, car manufacturer, car type (luxury or regular), car fuel cause car depreciation more significantly. So, while determining the price of the used cars, more emphasis should be put on the car manufacturer, car type, and car fuel type, rather than just the number of miles a car has been driven.

In future research, new data sets may be generated to better estimate the effects on car valuation. Buyers and sellers may want to know, for example, the benefit of car maintenance like upgraded sound system, new tire installation, fog lights on the car valuation. The ultimate hope of this analysis is to provide accurate tools what affects car price for both buyers and sellers while buying and selling a car in the used car market in India.