Econometrics Project

Introduction:

Individuals in developing nations such as Pakistan earn lesser wages. As a result, making a car purchase takes additional research. The key decision-making considerations are the availability of automobile components in local marketplaces, the cost of car parts, and the brand's resale value. Therefore, my research question is "**How does the price of a used model vs. a new model vary in the top four brands: Toyota, Suzuki, Honda, and Daihatsu**?" This is my focus because I want to see the reason for the price changes in these brands. The reasons can be the penetration strategy and how it influences people from diverse economic backgrounds. How are companies gaining an advantage over the size of the car and specializing in the market-oriented approach by producing vehicles shaped by customer experiences and expectations? How is the inflation rate playing a role in consumer preferences, and why do they want to buy a product that justifies its price with the experience? Finally, the policy outcomes of expanding the automobile industry can also be estimated based on this research proposition. The dataset I have used is representative, cross-sectional, and relevant to my research. My main findings concluded that there are significant differences between the prices of used and new cars of the four selected automobile brands.

Literature review:

The model of consumer buying choice was extensively examined in a study article that used the object of an MPV automobile as the research site in Bandung, Indonesia. Brand image, brand trust, product quality, and price were the four independent factors in the study, with purchase decisions as to the dependent variable. Although the findings of this study revealed that the four independent factors could impact a consumer's choice, the variable of price has a more significant effect and a higher significance level than the other variables in influencing a consumer's decision to buy MPV automobiles. As a result, company executives should consider product features and comfort in the provision of facilities and infrastructure because consumers regard the two components as the essential factors in deciding to purchase MPV cars, aside from the fact that the car's fuel consumption is economical. This outcome also implies that marketers should constantly provide a reasonable explanation about the quality of the products and affordable costs to generate a strong idea that an MPV automobile is a comfortable family car (Amron). Although I can observe that the source of competitiveness has switched from functional to emotional value, a comparison of the impacts remains insufficient. How does each brand's image affect corporate brand preference in the car industry?

As a consequence of analyzing the Japanese car industry, it is evident that many emotional value pictures had a significant impact, with innovation having the greatest impact. It is more effective to convey an emotional impression through technology than through communication. The functional value, on the other hand, is never unneeded. Regardless of how good the design is, people will not accept the product (Kato).

Model

The regression equation of the model is:

$$\widehat{\log}(Price) = 12.613 + .383\ NewSuzuki + .648\ UsedHonda + 1.044\ NewHonda + .070\ UsedDaihatsu + .515\ NewDaihatsu + .947\ UsedToyota + 1.270\ NewToyota + 0.033\ Log(Kmsdriven); N = 18{,}352, R^2 = 0.325$$

The response variable is log(price), where price corresponds to the market value of the top four car brands (Suzuki, Honda, Daihatsu, Toyota) based on the independent variables: brand type and log(kmsdriven).

Brand type is a categorical variable with categories (UsedSuzuki, NewSuzuki, UsedHonda, NewHonda, UsedDaihatsu, NewDaihatsu, UsedToyota, and NewToyota) label encoded from 0-7 with the reference category as *UsedSuzuki*. Log(kmsdriven) is a numeric variable representing the distance traveled by the top four brands in km.

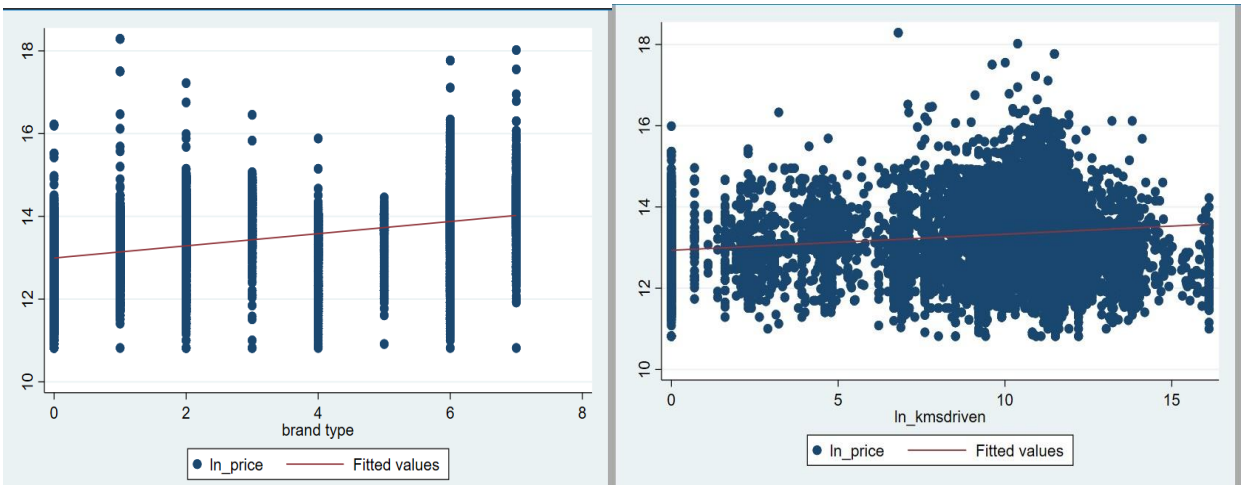
The specification of price as log(price) is done because the price is not normally distributed, and thus, from there, its relationship with other features is not linear. Log transformation is a common way to handle such issues. The main regressor of interest is brand type, as choosing any one of the categories in the equation reflects how much more or less expensive that brand is relative to the reference category. Similarly, the choice of reference category can be formulated based on the research proposition, such as if we want to compare UsedSuzuki with UsedToyota, then either of those could be used as a reference category.

Brand type is not a pre-defined variable in the metadata. Instead, it is an interaction of two variables in the dataset: brand and condition. Because the brand type depends on both variables, I cannot construct a t statistic by using these variables separately. Therefore, taking the

interaction term removes this constraint and accounts for the Omitted variable bias (OVB) of interaction terms if these variables had been introduced separately.

The control variable in my model is log(kmsdriven). Kmsdriven of a car has a significant influence on the price of the car. Thus, adding kmsdriven in the model will help to generalize our model and impact adjusted R2. As adj. R2 is higher for the non-nested model containing log(kmsdriven) than kmsdriven; thus, the functional form of kmsdriven is kept in log form. The quadratic term [log(kmsdriven)]2 was not included as the turning point of partial effect found was inconsistent with the data on hand. (Following are the figures of the best fit of both independent variables)

Figure 1: Best fit line of independent variables (brand type, log(kmsdriven))



The regressors that are not controlled for include: Model of the brand, registeredcity, transaction-type, and year. Variables such as the Model of brand and year have been treated as mechanism variables as the brand's models can easily be determined if the brand type is known. Alternatively, the brand type can also imply the car's age. E.g., UsedSuzuki will have a higher age

than NewSuzuki, and the model of the car under this brand is potentially known. Thus, including these variables can reduce the impact of brand type on price. Transactiontype and registeredcity have been excluded from the model due to their predictable nature. Since the majority of the cars are registered in Karachi and the common transaction type is cash, including these variables could have led to biasedness in estimates. Lastly, for simplicity, I kept the top 4 brands from the dataset as their cumulative frequency was very significant conditional to other brands.

The results from the model satisfied the assumptions of MLR 1-3. Since the dataset is large (18,352 observations) and variance inflation factor (VIF) values are less than 2, there is sufficient variation in x and no perfect collinearity. However, the Breush-pagan test for homoskedasticity and Reset test for OVB showed heteroskedasticity and Biasedness of OLS estimates.

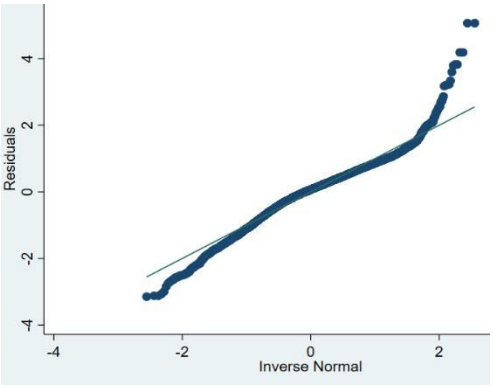
Potential reasons for biases include Measurement errors/Outliers in the data and omission of relevant control variables such as fuel average. Outliers have not been excluded from the model as it is often good to train your model on outliers so that the model can efficiently handle any test points. But, in my case, it is identifiable that the points are incorrectly marked, such as a used car cannot have one kmsdriven, so removing the outliers was one of the appropriate steps in data cleaning alongside removing rows containing missing values. Now, as removing these outliers messed up the analysis and statistical significance of the model, a log transformation of kmsdriven was done (scaled-down large number) to solve this issue.

Heteroskedasticity of the model was treated using robust standard errors. As corr(fuel avg, xi) will be >0 for some variables, <0 for others, and close to 0 for some variables, biasedness based on the omission of variables is very difficult to accurately estimate due to the MLR model. E.g., the sign of bias for log(kmsdriven) is positive as corr(fuel avg, log(kmsdriven)) > 0 as a brand with higher fuel avg can travel a more significant number of kms compared to a brand with smaller fuel

avg on the same fuel tank and $\beta > 0$ as again better fuel average is an indicator of higher price. Alternatively, I could assign bias for other variables based on the assumptions on the fuel avg of the brand types.

At last, the model is consistent as $Cov(xi,\ u) = 0\ \forall\ i$ and the assumption of normality is satisfied due to the Central limit theorem. The residuals are also normally distributed as most of the residuals fall along a roughly straight line in the qq-plot of residuals (shown below) except slight violation from the tails (test for normality of residuals).

Figure 2: QQ-plot of residuals



All things considered, after treating biasedness and heteroskedasticity as discussed above, we can safely assume that the OLS estimates satisfy the Gauss-Markov conditions and thus, are also asymptotically efficient.

Dataset

My data set for the research is made up of the car market on the online marketplace OLX in Pakistan. I chose this data set because of its versatility regarding the number of brands.

The number of variables allows space for many multiple linear regressions, for example, cross-brand analysis, cross-model analysis, and more. Since it was collected during a single time

period, the data set is cross-sectional, which was relevant for our research purpose. My data set consists of 8 variables: Brand, Condition, KMs Driven, Model, Price, Registered City, Transaction Type, and Year. Of these 8, Brand, Condition, KMs drove, and Price is relevant to our research question.

The number of observations in the data set is huge such that every feature can be assumed to have a random distribution which will maintain the assumption of normality based on the central limit theorem. While it is true that any online marketplace such as OLX will have biased points as not every consumer is rational, several numerical methods can be applied to counter those drawbacks and produce effective outcomes.

Regression results

I cycled through four OLS estimates to increase the authenticity and accuracy of the regression results and decided upon "c4" as the most accurate representative of the data.

Table 1: Distance travelled (km), brand type, and price of top 4 brands

VARIABLES	(1) c1 Price	(2) c2 ln_price	(3) c3 ln_price	(4) c4 ln_price
brand type = 1, NewSuzuki	415,851***	0.346***	0.338***	0.383***
	(110,016)	(0.0206)	(0.0206)	(0.0201)
brand type = 2, UsedHonda	523,595***	0.672***	0.672***	0.648***
	(19,567)	(0.0156)	(0.0155)	(0.0154)
brand type = 3, NewHonda	936,520***	1.034***	1.028***	1.044***
	(43,458)	(0.0314)	(0.0313)	(0.0309)
brand type = 4, UsedDaihatsu	87,304***	0.0798***	0.0843***	0.0704***
	(10,780)	(0.0199)	(0.0198)	(0.0196)
brand type = 5, NewDaihatsu	314,947***	0.508***	0.507***	0.515***
	(19,386)	(0.0328)	(0.0325)	(0.0311)
brand type = 6, UsedToyota	931,350***	0.975***	0.974***	0.947***
	(29,691)	(0.0142)	(0.0141)	(0.0142)
brand type = 7, NewToyota	1.413e+06***	1.250***	1.244***	1.270***
	(94,295)	(0.0238)	(0.0238)	(0.0228)
KMs Driven			-9.54e-08***	
			(7.83e-09)	
ln_kmsdriven				0.0333***
				(0.00148)
Constant	492,233***	12.93***	12.95***	12.61***
	(4,056)	(0.00690)	(0.00697)	(0.0157)
Observations	18,352	18,352	18,352	18,352
R-squared	0.079	0.306	0.311	0.325

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Main regressor

Results suggest that my main regressor, "brand type," has a positive relationship with the dependent variable, the natural log of price, throughout our case. The positive relationship is evident by the positive values of all the coefficients relating to our main regressor. The coefficient sizes are also mostly not extremely small, which indicates that the positive relationship is not weak but is relatively more substantial and notable. Furthermore, the independent variables associated with my main regressor are also all statistically significant with $p < 0.01$: if we repeated our experiment randomly 100 times, assuming that the null hypothesis (no statistical significance) is true, we would see the result only one time. Therefore, I have substantial evidence against the null hypothesis leading us to reject it and claim that my main regressor is statistically significant.

Other Regressor

The other regressor, the natural log of kilometers driven, is also statistically significant with $p < 0.01$ albeit with a weak positive relationship with our independent variable, as is evident by the small but positive coefficient size.

Economic significance

In order to gauge the economic significance of the coefficients, I first noted the type of relationship that exists between the dependent variables and the independent variable.

For my main regressor, we can see a log-level relationship. Therefore, the following relationship is established: *Keeping all other variables constant, the brand type in question is X % more/less expensive than a Used Suzuki (our reference category); X* = $100(\exp(\beta) - 1)$

Table 2: Economic Significance within top 4 brands relative to Old Suzuki (Log-level relationship)

Brand type	**X**
NewSuzuki	46.67
UsedHonda	91.17
NewHonda	184.06
UsedDaihatsu	7.29
NewDaihatsu	67.36
OldToyota	157.80
NewToyota	256.09

For my other regressor, we can see a log-log relationship. Therefore, the following relationship is established: *Keeping all other variables constant, a 1% increase in kms driven will lead to 0.0333% increase in the price*

Results in relation with the research question

My results show that there are notable and significant changes in the prices of old and new cars among the 4 car manufacturers selected for analysis. My model employed a used Suzuki as the reference category to gauge the change relative to other categories, but the model is flexible, and therefore any other reference category can also be used to demonstrate the changes.

Conclusion and Policy implications:

To conclude, the brand type of a car does influence the overall pricing strategy of a vehicle. Old Suzuki is relatively less expensive than new Suzuki, old/new Honda, old/new Daihatsu, and old/new Toyota. Similarly, the model is flexible by choosing any category from brand type to be a reference category. The R^2 of the model is 0.352 on 18,352 observations. Overall, the model is found to be asymptotically efficient.

The automobile industry today is one of the most lucrative industries. An increase in disposable income in rural and urban areas means a greater purchase of vehicles. In Pakistan alone, the industry saw a 32% growth between the sales of February 2021 to February 2022 (The News). Therefore, it is essential to examine the variables behind the price changes of these brands because they inevitably affect consumer behavior.

With increasing inflation rates, consumers are determined to buy products that justify the price with the experience. Suzuki, the most prominent shareholder in Pakistan's automobile industry, dominates the market through its customer-centric approach and low product prices. Owing to their large market size, vast distribution network, and cheap parts, the resale of a Suzuki in Pakistan is guaranteed. To achieve that resale level, Toyota, Daihatsu, and Honda should aim to position their product in line with consumer demand. Instead of refurbished cars at high prices, such as in the case of Daihatsu, the brand should aim to supply zero-mileage vehicles like its local competitors. Additionally, Toyota and Honda are well-loved automobile brand names in Pakistan, yet their premium prices deter even resale. The brands should attempt to set up manufacturing plants in Pakistan that can, in the future, also support the production of Toyota's innovative low-emission and hybrid models so that prices remain economical for consumers.

Work Cited

Amron, Amron. “The Influence of Brand Image, Brand Trust, Product Quality, and Price on the Consumer’s Buying Decision of MPV Cars.” *European Scientific Journal, ESJ*, vol. 14, no. 13, 2018, p. 228., doi:10.19044/esj.2018.v14n13p228.

Kato, Takumi. “Functional Value vs Emotional Value: A Comparative Study of the Values That Contribute to a Preference for a Corporate Brand.” *International Journal of Information Management Data Insights*, vol. 1, no. 2, 2021, p. 100024., doi:10.1016/j.jjimei.2021.100024.