Estimating vehicle MSRP from Power Performance Ratio

The Principal Components Ed Brown; Daphne Lin; Lihn Tran; Lisa Wu

Introduction

With technology advancements, automobile manufacturers are keen to introduce new vehicle models that provide better performance and driving experience to increase brand loyalty (retain customers) as well as attract new customers and increase their market share. Price is a key consideration to consumers when presented with factors to switch brands, as a consumer reports survey[1] in 2010 identified that price was an important factor to 67% of consumers surveyed.

While car manufacturers may get expert opinions on pricing a new car model, they can also benefit from data-based insights to reduce the uncertainty in their pricing strategy. Principal Components Consulting Group has been contracted by a big three global automotive company to evaluate how key features (specially, power performance) in new automobile design will influence their car pricing strategy. We use the Manufacturer's Suggested Retail Price (MSRP) to represent the car price. Once a target MSRP is estimated from power performance ratio, the manufacturer may then refine vehicle trim and model options to achieve a target profit margin for their new car product.

We will focus our study to explain the car price in terms of engine power performance ratio, using observations of 157 vehicle models from many automobile manufacturers across Europe, US and Asia. We apply a set of regression models to estimate the car price in relation to engine performance ratio and perform statistical analysis to evaluate the uncertainty of our estimates.

Data and Methodology

This is a car sales data set which is taken from Analytixlabs by way of Kaggle[2]. The data set was assembled in 2013 and contains characteristics of 157 different vehicle models from thirty (30) different vehicle manufacturers. This dataset represents the major global vehicle manufacturers across the three continents (Europe, US and Asia), and does not include some regional manufacturers that had a major presence in their local markets. Each row of data represents a single vehicle model. We encountered three samples from the dataset which were missing the values required for model development and these were dropped from the sample.

We developed the regression model and created the report creation using 154 observations after filtering the data set. An alternative approach was to use a 30% training and 70% testing data split. However, given our sample size, the alternative approach would result in a too small training dataset which may not produce consistent model estimates. Therefore, we apply the EDA work and a set of regression models to the entire data set.

Our outcome variable is price and our measured variable is the power performance ratio. It is important to understand what power performance ratio represents and why it has explanatory power of price. Power performance ratio is a continuous variable based on an engineering equation which represents the peak horsepower of an engine multiplied by the angular velocity or RPM of the engine at peak horsepower. This variable allows us to explain price in three ways. First, vehicles with high horsepower command a higher

[1] https://www.consumerreports.org/cro/news/2010/05/survey-car-buyers-most-influenced-by-quality-and-fuel-economy/index.htm

[2] https://www.kaggle.com/datasets/gagandeep16/carsales

price point due to the amount of engineering invested in engine development. Second, these high horsepower engines are predominantly found in luxury and sport vehicles. Third, higher angular velocities (RPM) in power performance ratio are also indicative of luxury or sports vehicles due to the engineering investment both in terms of design and materials to develop high RPM engines.

With the above causal relationship theory, we then conducted exploratory analysis to understand the distribution of our outcome variable (price) and the measured variable (power performance ratio), as well as the correlation between two variables. The histogram of price showed that it was right skewed; therefore, we decided to transform price by utilizing the natural log of price to bring it close to a normal distribution and correct the linearity issue for model development. Similarly, right skewness was observed in our primary explanatory variable (power performance ratio) and again we used the natural log to transform this variable into a more normal distribution. We also performed data transformation of other variables as needed to make sure the distribution shape is not too skewed.

We also analyzed the correlation between two variables (in natural log form) and observed that they are highly correlated and linear. We expect that when power performance increases, car value increases and price increases. In addition to the power performance ratio, we are also interested in whether this relationship shows differences by region (the headquarter location of the manufacturer) and observed some differences, as expected.

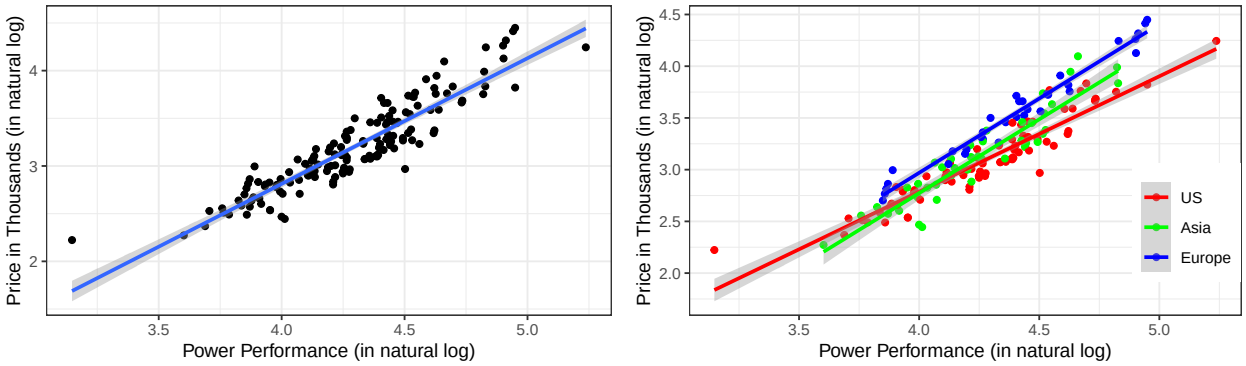

Figure 1: Car Price vs. Power Performance With and Without Manufacturer Region

We then fit a regression of the form:

$$\widehat{\ln(Price)} = \beta_0 + \beta_1 \cdot \ln(PowerPerformaceRatio) + \mathbf{Z}\gamma$$

β_1 represents the amount the price of the vehicle which will be adjusted up or down based on the power performance ratio of the engine designed for the vehicle. Since we use log transformation of the variables, this means increase power performance ratio by 1% will increase price byβ_1%. $\mathbf{Z}$ is a row vector of additional covariates (for example, region and car weight), and γ is a column vector of coefficients.

Results

Table 1 below shows the results of three representative regressions. For all three models, the coefficient on power performance ratio was highly statistically significant (with p-value less than 0.001). The coefficient estimate ranges from 1.152 to 1.318. This means that for a 1% change in power performance ratio, we estimate approximately 1.152% -1.318% change in price in the same direction. For example, in Model 1, a hypothetical $30,000 car will increase price by approximately $395 if the power performance ratio increases slightly from 60 to 60.6 (1% increase). The estimated price increase would be ~$378 by Model 2, and ~$345 by Model 3.

Model 2 adds the Region variable and reflects US in the base model. Each of these regions will have different engineering practices which may be reflected both in engine quality and the price (premium or discount) that manufacturers believe they can set for MSRP. Manufacturers from Asia and Europe have a positive effect on car price, in comparison to US manufacturers, which could be due to higher perceived engineering quality and precision. The coefficient of the Region variable is statistically significant (with p-value less than 0.001). To provide some context, for a hypothetical $30K car, the presence of a Asian manufacturer has a positive effect of adding $1,980 (beta of 0.066 multiplied by $30K) to car price, holding all other variables constant.

In Model 3, we want to consider the effect of vehicle physical features on MSRP, which we operationalize by using the curb_weight variable (representing the car/passenger capacity). curb weight had a positive correlation to MSRP indicating that the higher priced vehicles were larger than lower priced vehicles. The coefficient of the curb weight variable is statistically significant (with p-value less than 0.001).

Across all models, R2 and adjusted R2 are very high (from 0.84 to 0.91) which show that our measured variable (power performance ratio) explains price very well. In addition, we used Pearson correlation test to measure the practice significance which is 0.92, large practical significance as well.

Table 1: Estimated Car Price Linear Regression Models

	Output Variable: Price in Thousands of Dollars (in natural log)		
	(1)	(2)	(3)
Power Performance Ratio (in natural log)	1.318***	1.259***	1.152***
	(0.062)	(0.053)	(0.064)
Asian Manufacturer		0.066*	0.070*
		(0.030)	(0.028)
European Manufacturer		0.291***	0.314***
		(0.024)	(0.025)
Weight			0.075*
			(0.032)
Constant	−2.458***	−2.286***	−2.087***
	(0.262)	(0.232)	(0.233)
Observations	154	154	154
R^2	0.844	0.905	0.911
Adjusted R^2	0.843	0.904	0.908
Residual Std. Error	0.180 (df = 152)	0.141 (df = 150)	0.138 (df = 149)
F Statistic	825.440*** (df = 1; 152)	478.810*** (df = 3; 150)	380.641*** (df = 4; 149)
Note:	*p<0.05; **p<0.01; ***p<0.001 HC_1 robust standard errors in parentheses. American Vehicles are the reference level.		

Limitations

Independent and identically distributed (i.i.d) data is an essential assumption for our linear regression model to produce consistent estimates. Because there are different brands and models made by the same manu-

facturer based on market segmentation, there is a possibility of manufacturer and geographical clustering. Competition could also affect the outcome of the model as companies may adjust their price in response to competitors' prices. We partially account for geographical clustering by creating a dummy variable to indicate region. However, our model could not account for brand clustering and strategic interaction (competition among manufacturers). The i.i.d. limitations may cause biases and inconsistency in our model estimates.

Consistent linear regression estimates also require a normal distribution of residuals, which our diagnostic Normal Q-Q plot shows that there was no visual evidence of heavy tailed distributions in any diagnostic plot. Variables are automatically dropped to avoid perfect collinearity and the Variance Inflation Factor(VIF) test does not detect multicollinearity as well.

In addition, since we ran the linear regression model on the entire dataset, instead of the 30% training and 70% test data split, we took a conservative approach to assess all five classical linear model (CLM) assumptions which include i.i.d, no perfect colinearity and normality of errors, as assessed above, as well as linear conditional expectation and homoskedastic conditional variance. We examined the residuals vs. fitted value scatterplot and noted the line (mean of errors) is generally close to zero, which meets the linear conditional expectation. The same scatterplot also shows relative constant variance of errors and meets the homoskedastic conditional variance assumptions. We confirmed homoskedasticity by performing the Breusch-Pagan test (p-value of 0.07959 which fails to reject the null hypothesis of homoskedasticity).

In terms of structural limitations, several omitted variables may bias our model estimates. Two examples of omitted variables in our mode include fuel type and brand effect. Vehicle models require different fuel types (e.g., regular, mid, premium). Engines that require premium gas could have a positive effect on the power performance, compared to those that only require regular gas. The main effect is likely that the coefficient of power performance ratio is inflated and biased away from zero, making our hypothesis test overconfident.

Another omitted variable is the brand effect. For example, prestigious brands have higher perceived value and have a positive effect on price, while lesser known brands tend to demand lower prices. More prestigious brands such as BMW, may have more R&D capabilities and incentives to enhance engine technology, which has a positive effect on power performance, which could inflate the power performance coefficient and result in an omitted variable bias away from zero. This may also make our hypothesis test overconfident.

We evaluated reverse causality of our variables and don't expect price will reversely impact engine power performance, as the sequence goes from the engineering design process to the pricing strategy, and not the other way around. We also don't expect any outcome variables on the right hand side in our models.

Another limitation of our dataset is the timeframe in which the data was collected. The dataset only includes models launched in 2011-2012. With the speed of innovation and technology integration into cars, our regression model could be outdated as it does not consider new technology and features that could affect price.

Conclusion

This study estimated car price in terms of power performance ratio. For a small 1% change in power performance ratio, a hypothetical $30,000 car will increase price by approximately $345-$395. The power performance ratio in our data ranges from 23 to 188, so for a large percentage increase, we could expect a meaningful increase in car price. For car manufacturers who intent to invest in engineering R&D to improve power performance, this study provides the estimate of price return that they can expect.

In future research, it may be valuable to include new car models and brands after 2012 to incorporate the more recent technology advancements in price effect estimation. For manufacturers who are interested in competing in certain regions, it will be important to identify new datasets that include regional manufacturers (not present in this dataset) to evaluate power performance effect on price.