

# 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 fearful of being behind the curve and keen to invest in technology research and development (R&D) initiatives to introduce vehicle models that provide better performance and driving experience to increase brand loyalty and attract new customers. Technology R&D spent is a vital part of the total R&D cost for the automobile industry (\$76 billion based on the 2000 McKinsey report<sup>1</sup>). Manufacturers must gauge how these performance enhancements can generate economic value and influence the vehicle price consumers are willing to pay. A Consumer Reports survey in 2010<sup>2</sup> identified price as a critical consideration for 67% of consumers when presented with factors to switch brands. Therefore, pricing strategy is critical to the marketing and financial success of launching vehicle models.

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

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

## Data and Methodology

We utilized a car sales data set sourced from Analytixlabs by Kaggle<sup>3</sup>. The data set was assembled in 2013 and contained the characteristics of 157 different vehicle models from thirty (30) vehicle manufacturers. This continuous dataset includes the critical variables needed for our study. This dataset represents the major global vehicle manufacturers across the three continents (Europe, the US, and Asia). It does not include some regional manufacturers that may have a significant presence in their local markets. We performed the exploration data analysis (EDA) and developed the regression model on the entire 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 and may not produce consistent model estimates. Therefore, we chose the entire dataset for our EDA and modeling work and evaluated the critical assumptions in the limitations section.

Each row of data represents a single vehicle model. We reviewed the data quality of the critical variables. We noted that three samples from the dataset were missing the values required for model development, and we dropped these from the sample. We performed data transformation of the critical variables and added several new variables, as discussed below.

---

<sup>1</sup><https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/building-an-r-and-d-strategy-for-modern-times>

<sup>2</sup><https://www.consumerreports.org/cro/news/2010/05/survey-car-buyers-most-influenced-by-quality-and-fuel-economy/index.htm>

<sup>3</sup><https://www.kaggle.com/datasets/gagandeep16/carsales>

In our study, price is the outcome variable, and power performance ratio is the measured variable. It is essential to understand what the power performance ratio represents and why it has explanatory power of price. Power performance ratio is a continuous variable based on an engineering equation that 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: *Vehicles with high horsepower command a higher price point due to the amount of engineering invested in engine development.* These high horsepower engines are differentiating features, especially for high-end vehicles. \*Higher angular velocities (RPM) in power performance ratio indicate higher monetary value due to the engineering investment in design and materials to develop high RPM engines.

With the above causal relationship theory, we then conducted an exploratory analysis to understand the distribution of our outcome variable (price) and the measured variable (power performance ratio) and the correlation between the two variables. The price histogram 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, we observed right skewness in our primary measured variable (power performance ratio), and we used the natural log to transform this variable into a more normal distribution. We added a new indicator variable, “region,” to identify the location of the manufacturer headquarters to account for the geographical effect. We also performed data transformation of other variables as needed to ensure the data distribution shape is not too skewed.

We 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 and price increase. Given our background knowledge, we are also interested in whether this price vs. power performance ratio relationship shows many regional differences to detect engineering practice differences in each region.

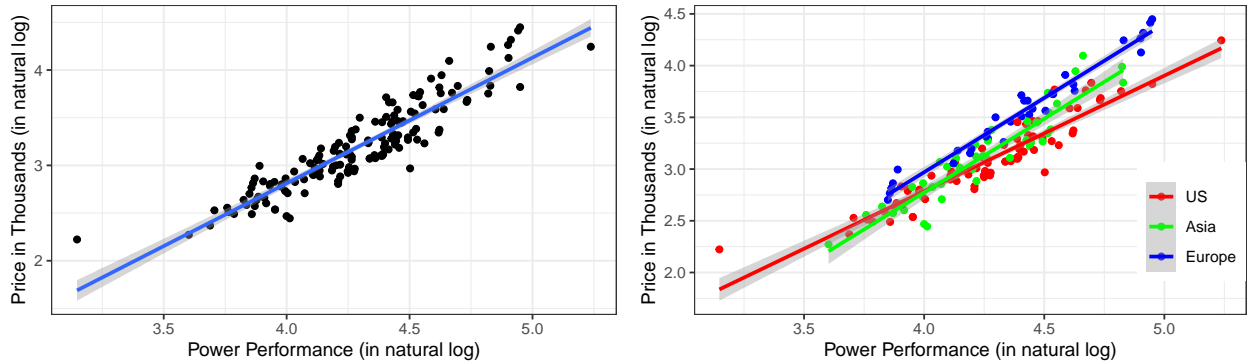


Figure 1: Car Price vs. Power Performance With and Without Manufacturer Region

We then fit a regression of the form:

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

$\beta_1$  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 increasing power performance ratio by 1% will increase price by  $\beta_1\%$ .  $\mathbf{Z}$  is a row vector of additional covariates (for example, region and car weight), and  $\gamma$  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. Therefore, a 1% change in power performance ratio will result in approximately 1.152% -1.318% change in price in the same direction. For example, in Model 1, a hypothetical

\$30,000 car will increase its 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 the US in the base model. Each region has different engineering practices, reflected 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 prices compared 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 0.05). To provide some context, for a hypothetical \$30K car, Asian manufacturers have a positive effect on price by adding \$1,980 (beta of 0.066 multiplied by \$30K) to price, holding all other variables constant

In Model 3, we want to consider the effect of physical vehicle features on MSRP, which we operationalize by using the curb\_weight variable (representing the car/passenger capacity). Curb weight positively correlated to MSRP, indicating that the higher-priced vehicles were larger than lower-priced ones. 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 shows that our measured variable (power performance ratio) explains price very well. In addition, we used the Pearson correlation test to measure the practice significance, which is 0.92, indicating a 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*** (0.062)	1.259*** (0.053)	1.152*** (0.064)
Asian Manufacturer		0.066* (0.030)	0.070* (0.028)
European Manufacturer		0.291*** (0.024)	0.314*** (0.025)
Weight			0.075* (0.032)
Constant	-2.458*** (0.262)	-2.286*** (0.232)	-2.087*** (0.233)
Observations	154	154	154
R <sup>2</sup>	0.844	0.905	0.911
Adjusted R <sup>2</sup>	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<sub>1</sub> robust standard errors in parentheses. American Vehicles are the reference level.

## Limitations

Independent and identically distributed (i.i.d) data is a crucial assumption for linear regressions to produce consistent estimates. Since the same manufacturer makes different brands and models for different market segments, manufacturer clustering is possible. The competition also affects data independence, as companies may adjust their price in response to competitors' prices in the same geographical areas. We partially account for geographical clustering by creating the "Region" indicator variable. However, our model could not account for manufacturer 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. Our diagnostic Normal Q-Q plot shows no visual evidence of heavy-tailed distributions in any diagnostic plot. The Variance Inflation Factor(VIF) test does not detect multicollinearity either.

We ran the linear regression model on the entire dataset; instead of the 30% training and 70% test data split. Therefore, we took a conservative approach to assess all five classical linear model (CLM) assumptions, including 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 that the line (mean of errors) is generally close to zero, which meets the linear conditional expectation. The same scatterplot also shows a relatively 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).

Several omitted variables may post structural limitations and bias our model estimates. Two examples of omitted variables are fuel type and brand effect. Vehicle models require different fuel types (economy/regular/premium). Engines that require premium gas could positively affect the power performance compared to those that only require economy/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 positively affect 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 positively affects power performance and could inflate the power performance coefficient and result in an omitted variable bias away from zero. These factors may make our hypothesis test overconfident.

We evaluated the reverse causality of our variables and do not expect that price will reversely impact engine power performance, as the engineering design process leads to the pricing strategy, and not the other way around. We also do not expect any outcome variables on the right-hand side in our models.

Another limitation of our dataset is the time the data was collected. The dataset only includes models launched in 2011-2012. With the speed of technology innovation and integration, our regression model could be outdated as it does not consider new technology, power performance data, and price effect.

## 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 its price by approximately \$345-\$395. The power performance ratio in our data ranges from 23 to 188, so we can expect a meaningful increase in car price for a significant percentage increase. For car manufacturers who intend to invest in engineering R&D to improve power performance, this study provides the expected price return estimate.

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