Power and Size : An Exploratory Analysis in Product Investment Paths

2023-12-12

Introduction

Current vehicle market trends indicate a customer preference away from smaller passenger vehicles, to larger ($>$ 600lb) SUVs and trucks. EPA analysis reveals a 45% market share of SUVs in the US in 2021 vs 20% in 2013. Due to this recent industry trend, the team was tasked with answering the question: Will an increase in vehicle power (horsepower) increase vehicle sales?

We built an initial model comparing horsepower and vehicle sales in thousands between car models. In addition to horsepower, other specifications likely play a factor in the vehicle sales. This study aims to reveal the relationship between horsepower and vehicle sales as well as any interactions with other vehicle specs to determine if certain product changes can result in better sales. In particular, we considered wheelbase (length between front and rear tires), as one of the first key design decisions for any car, and price, a major factor that can affect sales performance. As such, we built two additional models factoring in wheelbase and price.

Data and Methodology

The data in this study comes from Analytixlabs, an Indian data science institution made available to the public through Kaggle. The data set contains observational data about various models of cars. Each row corresponds to a different model with various measurable attributes for each model. We first split our data into a training set, consisting of 100 of 157 data points, and a validation set consisting of the other 57 data points. We performed all of our exploratory data analysis on the training set, and used the validation set to generate our statistical analysis and reports.

We found five data points across our training and validation sets that included N/A values across our variables of interest, and since they were an insignificant amount compared to our dataset, we chose to remove the data points. Overall, we built three models adhering to the principles of the classical linear model, the assumptions of which we discuss in the limitations section of this report.The three models are described chronologically as follows: Model 1: We were initially interested in explaining how the power of a car correlates with the sales of the vehicle, so our initial model included only a prediction of sales based on horsepower:

$$Sales\ in\ Thousands = \beta_0 + \beta_1 * Horsepower$$

where β_1 represents the immediate increase in sales in thousands if horsepower was increased by one unit. Model 2: However, we also were interested in two other factors that we thought could affect sales: The wheelbase of a vehicle and its price. Both variables were investigated and added along with horsepower into a second model:

$$Sales\ in\ Thousands = \beta_0 + \beta_1 * Horsepower + \beta_2 * Wheelbase + \beta_3 * Price\ in\ Thousands$$

Where β_1 represents the immediate increase in sales in thousands if horsepower was increased by one unit, β_2 represents the immediate increase in sales in thousands if wheelbase was increased by one inch, and β_3 represents the immediate increase in sales in thousands if the price were increased by \$1000. Model 3: Finally,

after seeing a non-linear underlying distribution of errors in our second model, we applied a log transformation to our outcome variable, sales:

$$log(Sales\ in\ Thousands) = \beta_0 + \beta_1 * Horsepower + \beta_2 * Wheelbase + \beta_3 * Price\ in\ Thousands$$

Where β_1 represents the log-transformed increase in sales in thousands if horsepower was increased by one unit, β_2 represents the log-transformed increase in sales in thousands if wheelbase was increased by one inch, and β_3 represents the log-transformed increase in sales in thousands if the price were increased by $1000.

We explore the main potential contributors to the sales of our vehicles - Power and Size. These will be operationalized through the use of two common industry metrics: Horsepower for power and wheelbase for size. Another metric that cannot be ignored is the price of the vehicle, as two major determining factors in consumer purchasing decisions are price and performance.

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```

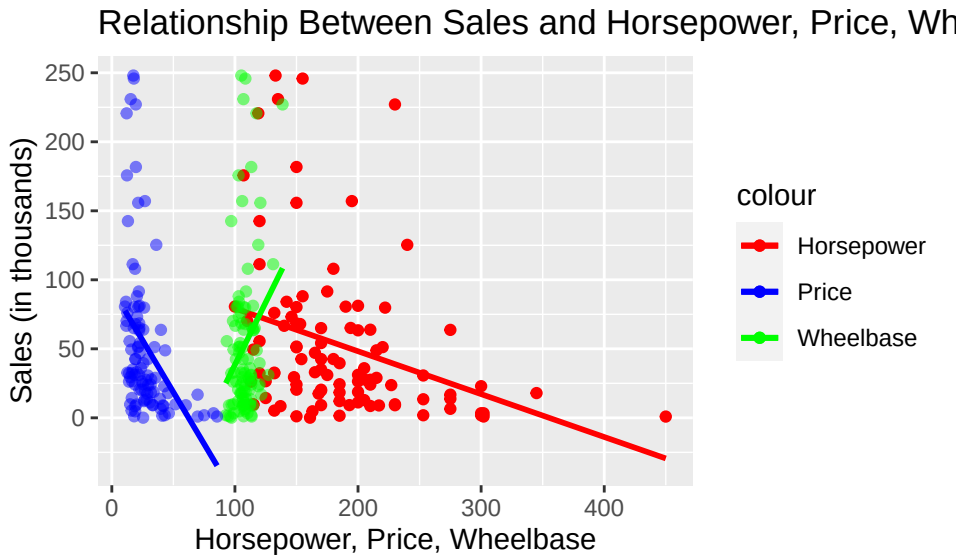

Figure 1: inital exploration of the variables for the sales model. Horsepower and Price seem to be negatively correalted while wheelbase is positively correlated

Exploratory plots suggest inital correlations between sales and horsepower, price, and wheelbase as observed in 1 . We fit the following model form to guide our model:

$$log(Sales\ in\ Thousands) = \beta_0 + \beta_1 * Horsepower + \beta_2 * Wheelbase + \beta_3 * Price\ in\ Thousands$$

Modeling

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Tue, Dec 12, 2023 - 10:00:03 PM

The three models were tested against the validation data set in order to determine the validity and usefulness of the model. With a RMSE of 58.3 and 55.8 respectively: models one and two were very clearly determined to be not useful with an over 100-fold error to constant ratio. While significantly improved with an RMSE of 1.06, model 3 will most likely not precise in being able to detemrine the exact number of sales for a given vehicle as the constants are in the tenths range.

```
## `geom_smooth()` using formula = 'y ~ x'
```

Table 1: Estimated Regressions

	Output Variables:		
	Sales in Thousands		Log(Sales in Thousands)
	(1)	(2)	(3)
Horsepower	−0.31***	−0.09	−0.0000
	(0.09)	(0.16)	(0.003)
Wheelbase		2.03***	0.05***
		(0.69)	(0.01)
Price in Thousands		−1.24**	−0.05***
		(0.62)	(0.01)
Constant	110.34***	−115.40	−0.83
	(18.26)	(72.05)	(1.51)
Observations	100	100	100
R^2	0.10	0.23	0.40
Residual Std. Error	54.85 (df = 98)	51.42 (df = 96)	1.08 (df = 96)

Note: *p<0.1; **p<0.05; ***p<0.01

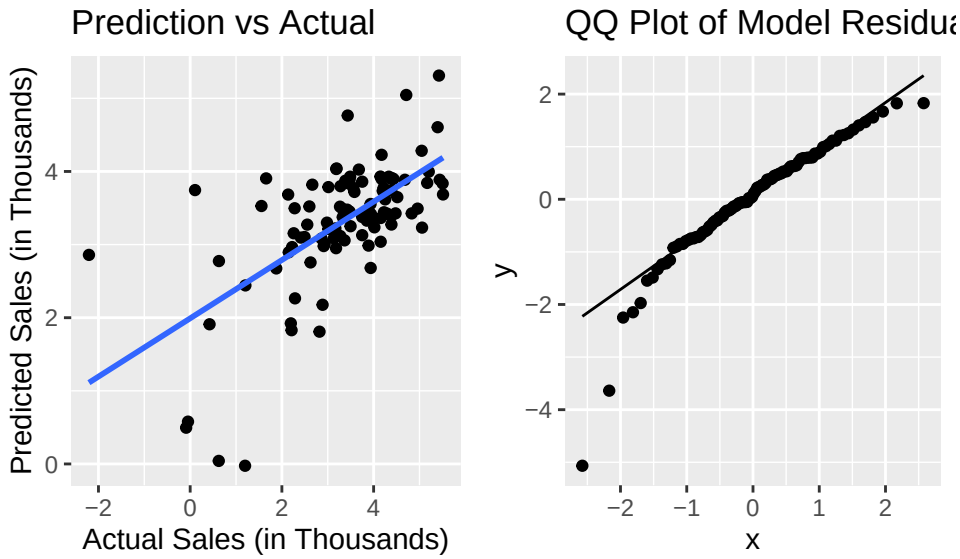

Figure 2: predicted vs actual on validation data set

While this exploatroy analysis reveals there may be a potential benefit to increasing the number of long wheelbase SKUs as observed in Figure 2. More detailed study and analysis is required to make accurate predictions of sales as can be observed in the error normality plot.

Limitations

Consistency in the linear regression estimate requires the sample to be independent and identically distributed. Because the dataset included multiple vehicle models from the same manufacturer, clustering by manufacturer may be a possible risk. Individuals may be loyal to car brands rather than models, so the popularity of the manufacturer would likely influence the sales of their vehicles. However, we did an exploratory analysis of sales by manufacturer and saw that the manufacturer itself did not have as much of an influence on sales compared to the number of unique vehicles manufactured. Additionally, car design is heavily influenced by seasonal trends, possibly resulting in temporal clustering. Overall though, we believe that overall the independence and identical distribution assumptions are not violated enough to invalidate our study. Consistency in the linear regression estimate also requires the sample to have an underlying linear distribution. In the first two models we designed, we actually saw a non-linear distribution in our error terms. We decided to mitigate this risk by log-transforming the outcome variable, sales in thousands. After this transformation, the error terms had a relatively linear distribution, so this assumption was met in our third model. We also note that we did not have perfect collinearity, as no variables were dropped from our model, and there was no near-perfect collinearity either.

As for structural limitation with our research design, one aspect that is worth calling out as a potential limiting factor is the bias caused by potential omitted variables. One potential omitted variable is economic conditions. As economic conditions and sales is expected to be positively correlated (better economy, more people purchase high ticket items such as cars) and prices of vehicles will likely to increase with increasing market demand - the omitted variable of the economic conditions would be positive. This would be a bias toward zero with the potential causing an understimate for the impact of price on the total sales.

It is also possible that there is a case of reverse causality here due to the price. From the data source, it is unclear whether the listed price is the price upon release, or the latest price. If it is the latter, it is possible that the lack or surplus of sales drives the price up or down accordingly.

Overall, we designed an explanatory model that shows a statistically significant relationship between several vehicle characteristics and sales, but we did not design a causal experiment, so our results cannot be interpreted as such.

Conclusions and Next Steps

This study explored what the correct investment path is for the future model years. Although the current model is unable to accurately predict the sales in relation to the increase in price and wheelbase length per model. This exploratory analysis revealed that a change to vehicle power may not influence the sales, while longer wheelbase models perform better for sales. This also supported our suspicion that higher priced models will sell less units.

In future research, more detailed and a wider spread of data across the industry may be taken to improve the dataset. This can include critical parameters such as model year (time) and annual GDP (economic conditions). Some more exploratory work may be required with consumer surveys to collect additional parameters, such as tech features, or vehicle height in order to address additional omitted variables. With these additional datasets, a more accurate model may be developed to predict sales as a function of vehicle parameters.