Gasoline Price Model

To consumers, perhaps the most frequent price-sensitive purchase is gasoline. Differences of a penny or two can draw consumers from one provider to another. I wondered if I could find data to create a simplistic, first-look model of gasoline prices over the past decade. I used R for this fun exercise; a condensed version of the code can be found at [GitHub](https://github.com/HendelData/gas_price_model).

There are times when gas prices seem to increase without explanation, but some quick research shows that gas prices are most affected by the cost of crude oil, production costs and federal and state taxes. Two other potential influences I discovered in my research are the number of barrels imported from OPEC nations and the difference in refining costs associated with summer and winter blends of gasoline; summer blend is more expensive to produce. Many articles I read indicated that gasoline prices varied by region, so I downloaded data by Petroleum Administration for Defense Districts (PADD) regions. There are five [PADD regions](https://www.eia.gov/todayinenergy/detail.php?id=4890) in the US: East Coast, Midwest, Gulf Coast, Rocky Mountain Region and West Coast (including Alaska and Hawaii).

I located the monthly data I needed from the [US Energy Information Association](https://www.eia.gov/) website. I could not locate any data on the actual production costs of gasoline by month or by PADD. As a proxy for the missing data, I included a dichotomous variable representing the summer blend production months (May through September). I believe this variable reflects the largest changes in production costs over the course of a year, so it is a good proxy for the missing data.

After combining the data sets, I had a single data set by month and by PADD containing data from January 2004 through December 2021. Because the data contains prices of gasoline over time, I used consumer price index (CPI) data from the [US Census Bureau](https://www2.census.gov/programs-surveys/demo/tables/p60/279/annual-index-value_annual-percent-change.xls) to convert all prices to real 2004 dollars.

I completed some univariate correlation analyses by PADD to verify that the variables I had selected were highly correlated with the price of gasoline and to be sure variables I planned to include were not highly correlated with each other. The cost of crude oil was highly correlated with the price of gasoline (a Pearson’s correlation coefficient [PCC] in the range of 0.95 with very small confidence intervals); federal excise tax on gasoline was not correlated with the price of gasoline as the tax has remained constant at 18.4¢ per gallon over the time period of interest; state tax and OPEC imports were not correlated with the price of gasoline as the PCC confidence intervals included zero with p values that were non-significant; the summer variable had low but significant PCC values; the year had significant PCC values. The year appears to be an important proxy for shock events, as without the year variable in the model, it performed worse in 2020 and 2021, years affected by the COVID-19 pandemic. Other shocks include unexpected refinery shutdowns due to mechanical issues or natural events. In general, for each PADD, the model underestimated prices.

The variable for cost of crude oil, the year and the proxy for summer all remained in the model based on the correlation analyses. I divided the data into a training data set (years 2004 through 2010) and a validation data set (years 2011 through 2021). After constructing the model on the training data by PADD, I applied the coefficient estimates to the validation data. Results were acceptable for all PADDS, with the East Coast, Midwest and Gulf Coast performing the best. As expected, an ANOVA showed that the cost of crude oil explained much of the variance in the price of gasoline over time: 87% for PADD1 and PADD2, 89% for PADD3, 65% for PADD4 and only 48% for PADD5. I believe that the low explanatory power for the Rocky Mountain Region and West Coast may be related to the higher transportation costs associated with those two regions due to geographical challenges, as described in this [report](https://www.energy.ca.gov/sites/default/files/2019-05/PADD_5_transportation_fuels_markets.pdf).

In general, the fit of the models was acceptable, with the PADD1, PADD2 and PADD3 models having the best fit, as expected from the ANOVA results above. The adjusted r-squared values for the models were: 0.9154, 0.9139, 0.9360, 0.8202 and 0.8201 for PADDs one through five respectively and the sum of the residuals was zero in all cases.

I produced graphs comparing the actual price to the predicted price for the validation data set, as shown below.

The variance of the coefficients across models for the cost of crude oil is relatively small. The values show that for each increase of $1 in the cost of crude oil, the price of gasoline can be expected to increase, on average across the PADDs, around 3.13¢. The coefficients for the summer variable show the most variance across the five models, with switching to summer blend costing an additional 10.12¢ in PADD1 (East Coast) to 28.68¢ in PADD4 (Rocky Mountain Region).

#R #datascientist #regression #ggplot #facets #linearmodel