Predicting Wholesale Power Prices in N.California using Linear Regression

This analysis uses multiple regression to predict hourly wholesale power prices at the main hub for Northern California, relating to PG&E (the principal utility). The specific context is the pertinent question of whether *"If the charging of electric vehicles were to be done at an hour other than the onset of the Off-Peak period (typically 9pm in the Winter and 10pm in the Summer), significant savings could be accomplished". This is an important question, as if EV chargers could be offered deeper savings to charge in the middle of the night, the power load might thereby be significantly smoothed out, which would also reduce the thermal stress on the vulnerable residential transformers that typically serve clusters of 5-10 houses.*

The independent variables may be divided into two groups. First, sources of generation, such as generation from Solar photovoltaic, Wind, small and large hydro, nuclear, thermal plants, and imports from other states. Load is another relevant variable, but since it is effectively the sum of the above variables, it introduces multicollinearity. Most of these variables are also available as projected values, from the so- called Day-Ahead-Market ("DAM"). The rationale for these variables is that collectively they constitute a supply curve, with the variable cost typically having a characteristic value for each generating type. For example, the variable cost of Wind and Solar is in each case very low. However, nuclear is more expensive, and thermal plants (usually natural gas-fired) are even more expensive. So, for example, hours in which a large amount of the generation is from wind and solar might reasonably be expected to have lower prices. Conversely, hours of high demand, such as Peak hours in the Summer months, may require even the most inefficient and expensive Thermal plants to run in order to meet the air-conditioning related high demand peaks.

Second, temporal variables such as the hour of the day, the specific weekday, and the month of the year were also included. Common sense suggests that the off-peak night hours with very little demand might evidence lower prices. Conversely, the Peak hours might be expected to have higher prices.

Exploratory data analysis revealed that most of the expectations of the data were confirmed, with however the significant surprise that solar energy is becoming so abundant that the midmorning hours often had the lowest prices of the day, being lower even than the lowest night prices. Another interesting result was that the standard deviation of hourly prices tended to vary linearly with the average price level for a given hour. This is contrary to the standard assumption for OLS models.

The method chosen to predict the prices was multiple linear regression. This was deemed appropriate as the variable to be predicted (power price) is continuous, as are each of the generation variables cited above. While Decision Trees are often an excellent method (and also the related Random Forests), their accuracy can be constrained by the fact that the number of different predicted values is limited to the number of distinct nodes. With continuous independent and independent variables, multiple regression was deemed most appropriate.

The subject hourly data for prices for the Day-Ahead-Market were downloaded from CAISO, the California Independent System Operator, in monthly intervals, and concatenated to make a continuous two-year time series. This is therefore a fairly large sample of more than 17,000 hours, and t-tests might therefore be expected to yield statistically significant results (as indeed was the case). The temporal variables were readily extracted from this time series, after first configuring the index into a datetime object.

The subject data for the hourly generation variables was scraped from daily reports, with expert scraping assistance provided by my mentor, Jeff Hevrin. Kudos to Jeff on this, as my Github search had revealed unsuccessful attempts to do something similar by others in the past.

The "statsmodels" package was used to do the OLS regressions, as it provides t-values for the independent variables. Using these t-values, a so-called parsimonious model can be developed by progressively dropping variables which are not statistically significant, in a step-wise manner. In some cases the sign of the coefficients of the remaining variables changed, and also in a manner that was more intuitive.

Principal Findings

The first initiative was to try to predict prices for all 24 hours of the day. This is challenging task, as prices for the Peak hours (eg from 6pm until 10pm) were highly volatile. An initial OLS model for all 24 hours of the day was therefore fitted to the time series for 2018, and yielded highly significant t-values for all variables except "Thermal" and "Large Hydro". These non-significant variables were sequentially dropped.

This model had an adjusted R2 of 49%, which seemed somewhat encouraging for a first model that covered all of the hours of the day – including the most volatile peak hours.

A purely temporal OLS model was also fitted, which excluded the generating variables and used only the hour of the day, the day of the week, and the month. This model had an R2 of 32%.

A third OLS model was then developed, that exclusively focused on the eight "Off-Peak" hours of interest. This model had of course only one-third the number of observations, but still more than 2,800 which suggests a potentially statistically powerful model (ie not a model with low t-values, in part due to high standard errors on account of a limited number of observations).

The R2 increased to 76%, an improvement that had been expected as the most volatile Peak period hours of the day were excluded. The only variable that did not have a statistically highly significant t-value was "solarPV", which makes perfect sense as the sun does not shine at night in Northern California. When this variable was dropped, all of the remaining variables had t-values that were statistically significant at the 1% level, and the R2 had declined modestly to 74%.

One important question in this context is the extent to which the relatively high R2 for the designated “Off-Peak” hours could have been due to “overfitting” of the model. To investigate this possibility a Train/Test Split approach was adopted using *scikit.learn*. Interestingly the R2 for this approach (using all of the available variables) was 79%, which is actually higher than the 76% for the comparable first model, which effectively had a 100% Training set. Of course, the R2 itself is a stochastic variable that changes somewhat with each random Train/Test split, so the observed results seemed quite reasonable from this perspective.

The major point of the foregoing is that the wholesale power prices tend to follow systematic patterns, with a strong temporal element. The hour of the day, and the day of the week, are each highly important variables that systematically effect prices in predictable ways.

Also, as derived in the related Jupyter Notebook, the prices at the hours of three and four in the morning are on average 53% lower than the late evening hours that mark the onset of the Off-Peak period that has a separate tariff for retail power prices. This has an important implication. Utilities can afford to offer much lower prices to EV chargers if the chargers are willing to delay their charging until 3 in the morning. Since most EV charging is at home, and as a matter of practice for a period of on average about two hours, this approach seems eminently possible. In fact, in Spain, EV chargers are offered an especially low “*Super Off Peak*” power price if they delay their charging until 1am. Norther California might therefore consider something analogous, but starting at 3am, when an offer of a generous price discount can be afforded by the electric utility.