Capstone Project Statistical Inference

Problem Definition: The project addresses the following question. *"Fact or Fiction: The intermittent nature of Renewable sources of energy, such as Wind and Solar, is such that the requirement to back them up imposes inefficiencies on existing fossil fuel plants, as measured by higher Heat Rates that negate most of the purported emission-reductions benefits from such Renewables*". The heat rate of a power plant is the amount of heat (Btu) from the combustion of fuel that is required to generate a standard kilowatt hour of power. This is important as the CO2 emissions are 100% directly related to the volume of fuel combusted – more fuel means more CO2!

Statistical Inference: the question is whether we can identify variables for which data is available that have a statistically significant relationship with the observed heat rate

Principal Findings: For the ERCOT electrical grid in Texas, 87% of the observed variability in the monthly heat-rates of the combined-cycle gas turbine power plants that constitute 93% of the gas-fired electrical power generation in Texas can be explained by a multivariate linear regression model that uses just six variables. These variables are, listed in order of their decreasing statistical significance, respectively:

1. The aggregate amount of electrical power generated by all gas-fueled power plants in ERCOT that month,
2. Time (since over time there has been a trend of adding new, and more efficient units),
3. Heating Requirements (specifically the number of Heating Degree Days that month),
4. The aggregate amount of Wind power generated by all wind turbine power plants in ERCOT that month,
5. The aggregate amount of nuclear power generated by all nuclear power plants in ERCOT that month, and
6. The month of the year (the colder Winter months having lower heat rates, the hotter Summer months having higher heat rates, and the lower-load months of March and April being subject to important seasonal maintenance of the nuclear plants.

The first four variables were each observed to be statistically significant at the 1% level (the likelihood of such a strong relationship being observed by chance being less than one percent), with the fifth variable (*nuclear power generation*) being significant at the 5% level. Of the twelve months of the year, eight were statistically significant at the 1% level. Recurring seasonal factors are clearly important in effecting the efficiency of power plants and need to be separately accounted for lest they act as a confounding variables.

While the amount of Wind Energy generated was observed to have a positive and highly statistically-significant relationship with the so-called Heat Rate of ERCOT's combined-cycle gas fired power plants, the relatively modest magnitude of this relationship implies that, when the added inefficiencies measured by the higher Heat Rates are factored in***, the imputed carbon footprint of an additional unit of wind generation is still approximately 80% lower than the average for the combined cycle gas-fired plants*** (and approximately [90]% lower than for ERCOT's coal-fired power plants). Importantly, these figures do not factor in the so-called "Peaker" single-cycle gas turbines, which although they constitute less than 10% of the gas-fired power plant capacity may be disproportionately used to back-up Wind generation on account of the speed with which they can be brought on line. This is a topic for potential future work. The point of this work is that at ERCOT, Renewables have been much more successful in reducing CO2 emissions than in Germany, for example.

Problem Significance: The question of whether Renewable forms of electrical power generation such as Wind and Solar truly convey reductions in aggregate CO2 emissions is highly controversial, with extreme views on either side. On the one hand, Green Groups fiercely claim that there is almost zero offsetting effect in terms of Wind generation making existing plants less efficient, and raising their heat rates. For an opposite view, see for example a recent piece by ASU's Professor Peter Rez entitled "*Why solar and wind won't make much difference to carbon dioxide emissions*". Available at https://blog.oup.com/2017/10/solar-wind-energy-carbon-dioxide-emissions/ . Just as cars are most fuel efficient when operated at a constant speed, so are power plants. Just a 20% drop in wind speed can cause a wind turbine to reduce its output by 50% or more, and so existing fossil fuel plants need to be able to ramp up and down their output within minutes. Power plants are usually most efficient when operated at full load, but the requirement to be able to increase output quickly means they may be obliged to operate at part load with a lower level of efficiency. So we have a possible so-called *“Fallacy of Composition”*: wind power itself emits no CO2 emissions, but the CO2 emissions of the supporting power plants may be materially increased.

Statistical Methodology: the model began with a multivariate linear regressions model from the Python STATSMODELS package. This package was chosen because it includes the “t-statistics” and other information commonly available from the R programming language (a successor to “S”, essentially a statistical programming language). The first model included thirteen potential variables, namely:

month, being the month of the year

ng, being the Aggregate electrical generation from power plants powered by natural gas

coal, being the Aggregate electrical generation from power plants powered by coal

nuclear, being the Aggregate electrical generation from power plants powered by nuclear

wind, being the Aggregate electrical generation from power plants powered by wind

wind2, a non-linear variable, the square of the above item

windpct, being the percent contribution of wind to total generation

load , being the average hourly capacity for the system as a whole for that month

dheat , being the number of heating days

dcool , being the number of cooling days (“*Heating and cooling degree-days are indicators of how much energy a typical household or building will use for space heating or cooling. The more heating degree-days you have, the more energy it will take to heat the inside of the home or building. The more cooling degree-days you have, the more energy it will take to cool the inside of the home or building”.)*

txgasprice , being the average natural gas price for gas supplied for power generation in Texas that month

usgasprice, being the national average natural gas price for gas supplied for power generation in that month

time, being the number of months since the start of the data set.

Commentary on Model Evolution: In the first model, with all variables included, apart from the single month of November, only three variables were statistically significant, namely wind, wind2 (being wind squared), and time. The R2 was 89%. Variables were then dropped stepwise in the following order, one at a time, and the model re-estimated (a standard statistical procedure for which stepwise packages are available in R). Specifically:

- After the first round, ***dcol*** was dropped as it was then the least significant remaining variable.
- After the second round, ***windpct*** was dropped as it was then the least significant remaining variable.
- After the third round, ***txgasprice*** was dropped as it was then the least significant remaining variable.
- After the fourth round, ***coal*** was dropped as it was then the least significant remaining variable.
- After the fifth round, ***usgasprice*** was dropped as it was then the least significant remaining variable.
- After the sixth round, ***load*** was dropped as it was then the least significant remaining variable.

At this point each of the remaining variables was statistically significant, so no further drops were made. Although six variables had been dropped, the R2 had declined by just 2%, from 89% to 87%.

Of particular note in this context is the negative sign of the coefficient for the variable ***wind2***, being the variable ***wind*** raised to the second power. There has been commentary from Europe and Australia that suggests that as wind generation hits a particular threshold of say, 20% of total power generation, then the inefficiencies imposed on the fossil fuel plants become increasingly marked. If this were indeed the case at ERCOT, then this variable ***wind2*** should be of positive sign, and statistically significant. Instead, we observe the opposite, a negative coefficient which is statistically significant. How can this be? One possible explanation is that as Wind generation has grown, investments to de-bottleneck electric power transmission lines have been made, so that curtailments have actually declined in relative importance over time. In other words, when wind generation attains a certain critical mass, more investments are made in transmission to integrate it. Another possibility is that as the number of wind generation sites has grown, generation has become more diversified and volatility has become averaged out (this is a readily testable thesis). As a third alternative, as Wind generation has become more important it may have become more worthwhile to invest in the development of sophisticated models that cam more accurately predict wind generation levels within different time horizons (eg of minutes, hours, and days). This area is a suitable topic for future study.