Project 2 - Statistical Computing

Modelling Peak Electricity Demand in Great Britain

An R-markdown template called `Report_project02.rmd` and a code template `code.R` have been provided. The `Report_project02.rmd` file outlines the required structure of your report and includes brief instructions for each section. You must upload a PDF of your Project report (generated from the `Report_project02.rmd` file), as well as the sources `Report_project02.rmd` and `code.R` files. Submission should be done through Gradescope. Please ensure that you tag all groups members on Gradescope, and also add all group member names in the report (see template).

Background

NESO (National Electricity System Operator) is responsible for planning and operating the electricity system in Great Britain (GB). NESO ensures that electricity supply meets demand at all times and supports long-term system planning, such as advising on the location of future generation and transmission infrastructure (https://www.neso.energy/).

Effective long-term planning requires reliable estimates of future electricity demand. Because new infrastructure is expensive and time-consuming to build, NESO needs to quantify uncertainty in peak demand many years ahead to avoid both over- and under-building.

Uncertainty in long-term electricity demand is influenced by various factors, including weather and climate, economic trends (e.g., productivity), energy efficiency measures, population changes, and technological shifts such as increased adoption of electric vehicles.

Your task is to develop a linear regression model using historical demand data to generate realistic simulations of future daily peak demand. This model will inform NESO's planning decisions and support security of supply analysis.

NESO is particularly interested in periods with a high risk of electricity shortfalls. In Great Britain, demand is significantly higher during winter due to heating and lighting needs, and shortfalls are unlikely in summer. Since shortages typically occur during extreme demand events, NESO is especially focused on ensuring that the model performs well in the upper tail of the demand distribution. The model will be used to estimate statistics such as the maximum annual demand and high quantiles (e.g., the 95th percentile of daily peak demand). Given the limited electricity storage currently available, NESO is not concerned with the clustering of high-demand days over time.

Data

You are provided with two datasets: `SCS_demand_modelling.csv` contains most of the data required for the analysis and focuses on winter periods when demand is highest and `SCS_hourly_temp.csv` contains hourly temperature data, which may be useful for modeling the temperature–demand relationship. You may assume that daily peak demand occurs at 6 p.m.

The variables in `SCS_demand_modelling.csv` are:

- Date
- wind - estimated capacity factor of wind generation at 6pm based on wind speeds on the given date, with installed wind generation as at 1 January 2015. The capacity factor is the wind generation divided by

the installed wind capacity. The wind generation is estimated using a physical model of installed wind generators across GB combined with historical wind speed data. The wind speeds used are generated from the NASA MERRA1 reanalysis dataset.

- solar S - estimated capacity factor of solar generation based on solar output on the given date at 6pm, with installed solar generation as at 1 January 2015. The capacity factor is the solar generation divided by the installed solar capacity. The solar generation is estimated using a physical model of installed solar generators across GB combined with historical solar output data. The solar output time series is generated from CMSAF SARAH satellite images.
- demand gross (MW) - the electricity demand recorded for 6pm on the given date. Demand is for all of GB as reported by NESO. This is measured as supply from all major power stations minus exports minus pumping for hydro minus transformer load. Alternatively it is end-user consumption minus embedded generation. It has estimates of embedded renewable added back on (these are wind farms and solar farms whose output is not recorded by NESO).
- temp - British population-weighted average temperature from MERRA1 at 6pm on given date.
- wdayindex - day of week, beginning from Sunday (0).
- monthindex - month of year, beginning with January (0).
- year
- TO - average temperature from 3pm-6pm on given date.
- TE - average of TO at 6pm and TE at 6pm on the previous day.
- start year - the year at the start of the winter period.
- DSN - the number of days since the 1st of November.

Goals

NESO have asked you to help develop a new model based on linear regression that they can use to generate realistic traces of daily peak demand for use in their security of supply analysis.

They want you to test and fit different models using the data sets provided and to recommend a model for their use. With that in mind, your regression model will have `demand_gross` as the response variable and the explanatory variables should be derived from the provided datasets. As a basic model to help estimate future demand patterns, consider

$$M_0 : Y_i = \beta_0 + \beta_1 Wind_i + \beta_2 Solar_i + \beta_3 temp_i + \beta_4 wdayindex_i + \beta_5 monthindex_i + \varepsilon_i, \quad \varepsilon_i \sim \text{Normal}(0, \sigma^2)$$

where Y_i is the demand gross at time i, $\beta_j, j = 1, \ldots, 5$ are the regression coefficients and ε_i is the error term at time i.

You should compare the performance of your model(s) against the baseline model M_0. We have covered a number of models and modeling approaches in the lectures and workshops, and you should explore multiple modeling strategies, but your report should focus on describing and justifying your final chosen model. When choosing a model to recommend to NESO, consider the following questions:

1. M_0 does not incorporate historic data to estimate future demand patterns, but the past demand is a good indication of the future. To account for trends in population, the economy, climate etc, think how you can reescale historic data using a 'year effect' statistic from for historic data.
2. How well does your model fit the historic data?
3. Is the prediction accuracy the same across all months? or is the model better at predicting specific months?

4. How could the maximum annual demand have varied in the 2013-14 winter season if different weather conditions had occurred? For this you can investigate variation in the maximum annual demand in 2013/14 by inputting the weather conditions from previous winters (i.e. test weather in 1991/92, 1992/93, ...).
5. How does peak daily demand depend on temperature? NESO currently use a variable known as TE (described below and in your dataset) - is there a better alternative?
6. What would be the limitations of your model?

You may use few or as many of the provided variables, and you may transform and manipulate these variables to generate additional covariates, but you must not use external datasets.

To assess how well your model can predict demand at different months and compare different models, you should construct a cross-validation scheme that groups the data by month and computes the prediction scores for each month. You can do the same to evaluate, e.g., how well the model predict for weekdays versus weekends.

It is important that any conclusions you draw from your model are well supported and sound and that you understand limitations of the model and the data. You are asked to build a linear regression models and NESO explicitly do not want a `blackbox model`. You should be able to explain and justify your modeling choices and your model's predictions.

Report structure

Your completed assignment must follow the section structure in the `Report_project02.rmd` template (you may add subsections, but not remove main sections). Please remove the instructions for each section in the final report. All your code should be well documented and placed in `code.R`, which will appear in the Appendix of the `Report_project02.rmd` report (see template). Your code will be assessed for clarity and adaptability (e.g., changing input data without altering core functions).

Your report must contain no more than 3000 words (not including code, references, tables, figures and their captions). Please state the word count under the title of your report. The total report length (excluding Appendix) should not exceed 15 pages.

All code and figures should be accompanied by text that provides an overview and context to what is being done or presented. This means that models and results should be presented mathematically and with graphs or tables as necessary, and not using computer code and output. It is advisable that you make use of the `ggplot2`, `tidyverse` and `kable` R-packages for constructing figures and tables in your report. Do not include raw output or R commands in the main report body.

Your report must include all of your work, but make sure that you are only retaining required components, e.g. remove unused code and figures (if a figure is not explicitly discussed in the text it should not be in the final document). Overall, your project will be partially assessed on the organization presentation of the document, so it should be as polished and streamlined as possible. Try to be as concise as possible while creating your write-up.

Marking scheme

There is no single correct analysis for this Project. The marks will be allocated on a combination of statistical approach and justification, interpretation of results in context, presentation, code reproducibility and efficiency, as well as how close you get to some particular model answers to the given challenges and questions.

- 80–100% The analysis is sound so that conclusions are well-supported statistically. Interpretation goes beyond simply interpreting plots and tables. The project should demonstrate a clear overview of the work while being as clear and concise as possible and have no statistical errors. The code is efficient

and carefully commented, so that someone reviewing the code can easily tell exactly what it is for, what it is doing and how it is doing it without having read this sheet. The work is to a publishable standard.

- 70-79% Analysis is sound so that conclusions are well-supported statistically. Interpretation is reasonably mature. The project should demonstrate a clear overview of the work, without getting lost in details, and be free of all but minor statistical errors. Clear and efficient code.
- 60 – 69% Some flaws in the analysis or presentation (or minor flaws in both), but statistically sound with a clear assessment of the results and conclusion. A good grasp of the statistics and context, so that interpretation is reasonable.
- 50 - 59% Should contain some sound statistics insights demonstrating understanding of statistical modelling and its application. Reasonable presentation and organisation.
- 40 – 49% Major flaws in analysis and presentation, but demonstrating some understanding of statistics, and a reasonable attempt to present the results.
- Below 40% - Flawed analysis demonstrating little or no understanding of statistics, and/or incomprehensible or very badly organised presentation.

Additional resources

- NESO documentation on how they estimate historic year effects (known as the Average Cold Spell peak): https://www.emrdeliverybody.com/Capacity%20Markets%20Document%20Library/NGESO%20ACS%20Methodology%202022.pdf
- Details from NESO on how they develop subjective views of the future energy system: https://www.neso.energy/publications/future-energy-scenarios-fes
- NESO's annual electricity capacity report - this is one of the uses of this analysis: https://nationalenergyso-emr.my.salesforce.com/sfc/p/#8d000002dUGC/a/J70000004CYD/cv3SY3Z5cLuiRsHHJuK5FZcNebxJDmgEeAqjo9ot1oo