Nuclear

Annison. M, Evans-Tovey. I, Li. J, McHugh. O, Satik. E, Wang. J

AIM:

To develop a multiple linear regression model to predict the costs of constructing nuclear reactors in France, identifying key predictors and assessing the impact of various factors such as reactor construction time, joint planning, and reactor number.

BACKGROUND:

In 1973, there was an oil embargo imposed by the Middle East, which ceased the shipping of oil to countries supporting Israel during the Yom Kippur War. Whilst France was not a direct target, it lead to a major increase in oil price, and caused France to reassess their reliance on Middle Eastern oil, which is what inspired the mass construction of these power stations.

INTRODUCTION:

This project examines the factors influencing the cost of constructing nuclear reactors in France, using a multiple linear regression model to predict costs. By accounting for inflation and assessing the relationships between covariates and the dependent variable, we identify the most significant predictors. Our results highlight the importance of factors such as the length of the construction process, reactor number, reactor type, and joint planning with another reactor in determining the overall cost.

METHOD:

What Data Points were removed and why?

We flagged 5 data points Chooz B (16,457 and 19,545), Civaux (47,630 and 57,729), Golfech (14,726) due to their unusually high construction costs which were significantly higher than the upper bound threshold of 12,180. Upon reviewing the QQ plots and analysing the distribution, it became clear that certain data points, Chooz B (16,457 and 19,545) and Golfech (14,726), would still be reasonable points to include given they are below 20,000. We identified the anomalous data points to be Civaux (47,630 and 57,729). It is important to note that the power stations built in Civaux were the only stations in the data to be built after the Chernobyl disaster, so this is likely to be the reason for the outliers.

How will this be likely to affect the data?

Removing these data points will allow a better model fit when predicting the costs for future nuclear reactors. It will improve the robustness of the model since it will not be

skewed by the unusually high values. Removing these outliers will lower the overall variance, which will result in narrower confidence intervals.

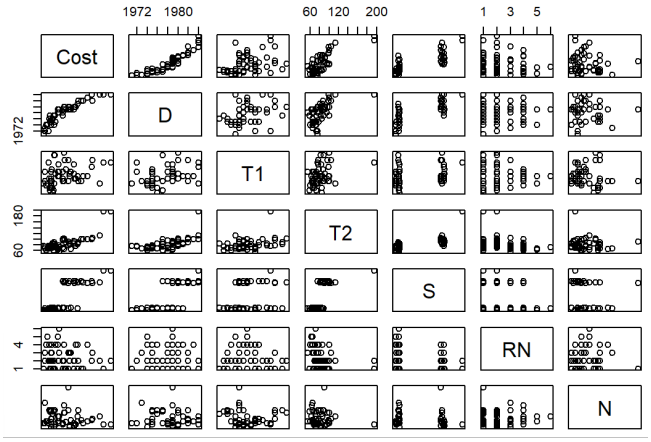
Why was this data anomalous?

The anomalous data were the two most recent nuclear plants built. The Chernobyl disaster began in April 1986, this could explain the higher costs for the two data points that we removed since the application process for these two anomalies started in 1987 and 1990. There could be more health and safety precautions that cause an increase in cost.

Exploratory Analysis

First we looked at a pairwise scatterplot of variables to see if there are any associations between them - potential multicollinearity in Date and T2 (length of construction process). Will compare models with and without date accounted for through inflation.

Observe an association between Cost and T2 (and date)



Modelling process

We briefly explored the location variable, looking into where in France it is located e.g. east/west, coastal etc. and then the population for each commune. From this there was no obvious association so we decided no further analysis for this variable was necessary.

Fitted all variables except date (not a good fit) and location, and removed variables at each step by looking at the significance level of an F-test and statistics such as AIC and adjusted R^2. At the final model removing any more variables (that are all significant/ ly different from 0 at 5% level) would lead to the AIC increasing and R^2 decreasing.

The residual plots were showing that the residuals were not randomly spread around 0 - looked to be a square function. Taking the square root of the response (cost) has made the plot look much better. Log function was considered but although the residual plot appeared better the scale-location plot was much worse - suggesting homoscedasticity present.

RESULTS:

Model Formula:

$$\sqrt{\{Cost\}} \sim T2 + RN + CP + JP + Intercept$$

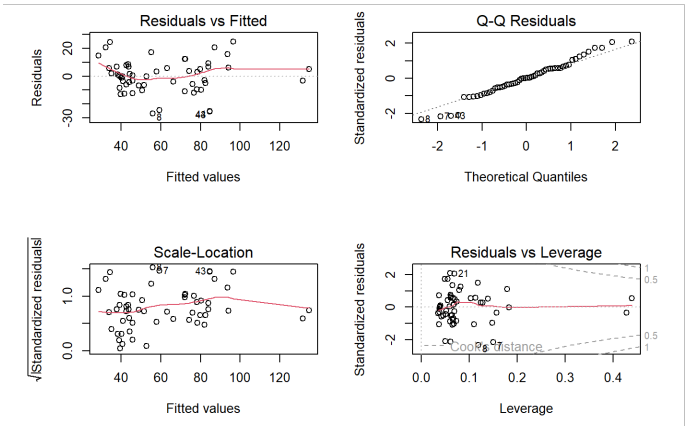
- **Strongest association:** T2 (construction time).
- **No significant points in Cook's distance.**
- **AIC value:** 446.7, **R^2 value:** 0.7882.
-

Comparing Inflation:

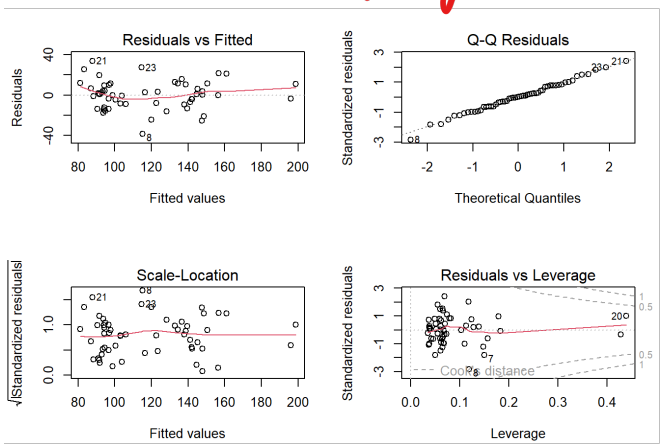
When inflation was accounted for in the model, residual plots remained unchanged. However, the QQ plot and scale-location plot improved. The association with RN and CP became milder in this model.

- **AIC value:** 465.1 (increase), **R^2 value:** 0.7894 (very slight increase).
- Multicollinearity and limited data on currency units made it difficult to fully assess the impact of inflation on costs.

no date



has date (inflation)



DISCUSSION:

Limitations:

- Small dataset.
- Limited number of variables.
- Lack of clarity regarding payment methods, currency units, and cost units.

CONCLUSION:

Key findings:

- **Outlier Detection:** Identified 2 key outliers, Civaux (47,630 and 57,729).
    - In the oral presentation, explain the process of identifying these outliers, highlighting the initial flagging of 5 data points, followed by excluding 3 after reviewing QQ plots and distribution analysis.
- **Variable Transformation:** Applied square root transformation to "Cost" for better model fit.
- **Data Exploration:** Removed non-numerical variables (location) and used pairwise scatterplots to identify associations. Boxplots were used to explore the effects of CP and JP on costs.
    - Briefly discuss the influence of CP and JP on cost.
- **Model Selection:** The final model selection was based on AIC, with model 6 being the best fit.

Overall, this project highlights the importance of key factors like construction time (T2) and reactor number (RN) in predicting nuclear reactor construction costs. The process of removing outliers, applying variable transformations, and refining the model demonstrates

the critical role of data cleaning and analysis in improving predictive accuracy. Despite the limitations of the dataset, the final model provides valuable insights into the cost drivers of nuclear reactors, laying the groundwork for further research and model improvements. Future work could benefit from more detailed data, including cost breakdowns and inflation-adjusted figures, to refine these predictions further.