Statistical Data Science

Example project presentation

Stuti Malik

Table of Contents

Project overview and objective

A file containing data from an experiment on a chemical reactor contains the following variables:

- X1: the reaction temperature, coded so that X1 = (Temperature - 200)/50, where Temperature is in degrees Celsius
- X2: the pH of the reaction, coded so that X2 = pH - 5
- X3: the pressure under which the reaction is run, coded so that X3 = (Pres- sure - 175)/25, where Pressure is measured in kPa
- Y: the yield of the reaction as a percentage

Task

The aim of this project is to recommend settings of temperature, pH and pressure which will maximise the yield of the reaction

Discussion of data science or statistical techniques

Initial process for model selection

Applying several short-listed linear models in R to analyse the data in order to be able to make good predictions of the yield. Points to consider:

- Is a multiple linear regression model adequate, or should a second (or higher) order polynomial model be preferred
- More generally, which terms should be included in the model, i.e. what should the linear predictor be?
- Does the model seem to be adequate? Do any of the required assumptions seem to be contradicted? If so, how can the model be improved?
- How can the model be used to estimate the expected yield at different set- tings of temperature, pH and pressure?

Process of selecting a multivariate statistical model

Model selection process

A process involved in creating a multivariate regression model:

- How to find a good-fit model to make the predictions
- Consider the statistic values and plots from the model which illustrate the adequacy of the model
- How will the analysis change if, for example, several engineers agree that the yield can be improved by using pH $= 6$, but not sure about the best settings for the temperature and pressure

Applying the model

Answering the question

At what settings would you recommend the reaction be run in order to maximise the yield?

- Summarise the steps involved in using the selected model to solve the optimisation problem (finding the stationary point to find the maximum yield)

Further discussions

- Statistical modelling and data science techniques, applied in agricultural and environmental research
- Data integration and data fusion techniques from various data sets, meta-analysis, considering the existing research, and selecting the right tools for analysis
- Research methods and important considerations
- Challenges for agriculture