STA 325: Case Study

DUE: Oct 22 11:59pm (on Canvas)

For the case study, you will develop a predictive model for a real-world research problem on turbulence, which has important applications in astrophysics, climatology, and engineering. In the workshop (held in class on Oct 6), Prof. Mak (on behalf of an ongoing collaboration with Duke Civil & Environmental Engineering) will present a research problem on predicting distributions of particle clusters in turbulence. You will then work in groups of at most five to build a machine learning model that not only is well-reasoned out and gives good predictive performance, but also investigates key scientific questions and provides interpretable insights. Each group will present their findings in a 10-minute recorded presentation, and submit a short report on Canvas.

The dataset consists of $n = 89$ simulations, each conducted at a different set of input parameters. There are three parameters (predictors): Reynolds number Re, gravitational acceleration Fr, and particle characteristic St; details on such parameters are provided in presentation slides. The raw response variable (obtained from numerical simulation and cluster analysis) is a probability distribution for particle cluster volumes. Since probability distributions are hard to work with, your collaborators instead summarized these distributions by their first four *raw* moments (i.e., $\mathbb{E}[X]$, $\mathbb{E}[X^2]$, $\mathbb{E}[X^3]$, $\mathbb{E}[X^4]$), which are easily computable. This data is provided in `data-train.csv`.

Your job as a statistical collaborator is to build a machine learning model that can predict the following summary statistics of the particle cluster distribution at a new input setting:

- *Mean*: $\mu = \mathbb{E}(X)$. This measures a distribution's centrality.
- *Standard deviation*: $\sigma = \sqrt{\mathbb{E}[(X-\mu)^2]}$. This measures a distribution's variability around its mean.
- *Skewness*: $\gamma = \mathbb{E}[(X-\mu)^3]/\sigma^3$. This measures a distribution's degree of asymmetry.
- *Kurtosis*: $\kappa = \mathbb{E}[(X-\mu)^4]/\sigma^4$. This measures how "heavy" a distribution's tails are.

These four statistics are more meaningful than raw moments when summarizing a distribution. To compute them, you will need to compute *central* moments (e.g., the second central moment $\mathbb{E}[(X-\mu)^2]$) from the provided *raw* moments (e.g., the second raw moment $\mathbb{E}(X^2)$). The function `raw2central` in R may be useful here.

Your machine learning model should be developed with the following two goals in mind:

- *Prediction*: For a new parameter setting of (Re, Fr, St), predict its particle cluster volume distribution in terms of its four summary statistics (mean, variance, skewness and kurtosis).
- *Inference*: Investigate and interpret how each parameter affects the probability distribution for particle cluster volumes.

For model building and justification, you may consider exploring the following points:

- *Predictive modeling*:
  - Take a look at the ranges and histograms of input and output variables. Do any of the variables require transformations? If so, what transformations are appropriate?
  - Try fitting a linear regression model – does it fit well? If not, try fitting a more complex (nonlinear) model. Does the data give any evidence for nonlinearity?
  - Does there appear to be interaction effects? If so, which variables interact (e.g., x_1 and x_2), and how do they interact (e.g., x_1x_2, or more complex interactions)?
- *Scientific inference*:
  - How does each of the three parameters (Re, Fr, St) affect the distribution of particle cluster volumes? Do these effects appear linear or nonlinear? Try to interpret these effects using what the three parameters mean physically.
  - Are the effects identified above similar over all response variables (i.e., over the distribution mean, variance, skewness and kurtosis), or are there effects that differ between, say, the mean and variance? Try to interpret any differences in effects using the three parameters mean physically.
  - Try to interpret any interaction effects using what the three parameters mean physically.

The above points should only be used as a guideline for your analysis. You are strongly encouraged to explore beyond these points, in order to provide comprehensive and well-supported answers to research objectives.

This assignment will be worth 15% of your grade, and will be evaluated on the following:

- *Report*: A short report describing your findings on the scientific problem. This should be no more than 4 pages (including figures and tables), and should have four sections:
  - An *introduction* outlining key research objectives and how your model achieves such objectives.
  - A *methodology* section describing your statistical model, how your model is fit from data, and justifying why your model is appropriate given the problem or dataset.

  - A *results* section discussing your predictive results (don't forget uncertainty!), as well as insights on the scientific problem. You should also submit your predictions on the hold-out set in `data-test.csv`, in the form of a `.csv` file.
  - A *conclusion* section summarizing key findings of your study.

  Your report will be evaluated on writing flow and organization, logical presentation and insightful interpretation of results, and how well it addresses research objectives. Bonus points will be awarded to the group with best predictive performance on the hold-out set.

- *Presentation*: A 10-minute recorded presentation summarizing your findings and recommendations. This can be recorded on Zoom. Your presentation will be evaluated on presentation clarity and flow, as well as the criteria above.