STA 325: Data Expeditions

For Data Expeditions, 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 first workshop (October 11), Reza Momenifar and Jonathan Holt (two Ph.D. students from Civil & Environmental Engineering) will present a research problem on predicting distributions of particle clusters in turbulence. You will then work in groups of four to build a statistical model which not only gives good predictive performance, but also investigates key scientific questions. In the second workshop (November 1), each group will present their findings in a 10-minute presentation, and submit a short report on Sakai.

The dataset consists of $n = 89$ simulations, each conducted at a different parameter setting. There are three parameters (predictors): Reynolds number Re, gravitational acceleration Fr, and particle characteristic St (details in presentation slides). The original response variable (obtained from numerical simulation and cluster analysis) is a probability distribution for particle cluster volumes. However, since probability distributions are hard to work with, we will instead summarize it by its first four raw moments (i.e., $\mathbb{E}[X]$, $\mathbb{E}[X^2]$, $\mathbb{E}[X^3]$, $\mathbb{E}[X^4]$), thus creating four new response variables. This data is provided in `data-train.csv`.

Your job as a data analyst is to build a statistical model which achieves two goals:

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

Some points which may be worth exploring:

- *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., $x_1 x_2$, 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 central moments (i.e., over all response variables), or are there effects which differ between, say, the mean and the variance? Try to interpret the latter 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 10% of your grade, and will be evaluated on the following:

- *Report* (November 1, 11:59pm on Sakai): A short report describing your findings on the scientific problem. This should be no more than 5 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* (November 1, in lab): A 10-minute presentation on the same topics. Your presentation will be evaluated on presentation poise, as well as the criteria above.