Statistics of Big Data, Fall 2021

Warmup homework exercise

Due date: 25 October 2021 before class

This exercise relies on the 2009 Nature paper[1] describing Google Flu Trends and the 2014 Science paper[2] discussing the major errors this model made starting in 2011. Both papers are available from the class home page.

Important comment: The goal here is to think like statisticians, and offer answers that are concise and mathematically and statistically sound and accurate. As much as possible use (correct) formulas and concrete statistical terms like "correlation", "bias", "variance", "stationarity" in their correct meaning in explaining your answers.

1. Read the Nature paper, and identify two flaws in the statistical methodology, and explain how each of these aspects could have been better addressed. Your answers must be specific, accurate and concise (and as mathematical as possible), rather than vague statements. The goal here is not to propose completely different approaches and ideas, but to concentrate on how they could have implemented their chosen approach better. These are a few points in the paper where you can find answers:
   (a) The internal four-fold cross validation in fitting each one of the $5 \cdot 10^6 \cdot 9$ models : is it necessary? Could the same goal have been accomplished with a simpler approach computationally and conceptually? **Hint:** make sure you think about the basic relationship between regression and correlation.
   (b) The methodology for combining the 36 Z scores into one score: is there an obvious better / more powerful approach? Possibly combined with the solution of the previous item.
   (c) Given that they want to predict the actual percentage in the future, would the use of other evaluation measures in the final step (selecting the number of terms) be more appropriate? Which and why?
2. Beyond general disparaging comments, the Science paper seems to attribute the failure of GFT to two main flaws:
   (a) Failure to take advantage of CDC data in modeling and failure to compare to baseline based on CDC data auto-regression.
   (b) "Algorithm dynamics" which seems to relate to how Google is collecting the data and how users are using the search engine. Because these are changing over time, the model no longer predicts well and is consistently overestimating.

   Explain each point briefly in concise and mathematical terms. Especially for the second point ("dynamics") explain which more general phenomenon is being described here that applies to all data that are collected and analyzed over time.
3. Of the points in question 1, and the two points in question 2, which one do you think is the most responsible for the extreme deterioration of GFT performance over time? Explain clearly and briefly.

[1] http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html
[2] https://www.science.org/doi/full/10.1126/science.1248506