Final Project – published 2022 April 17

Throughout the semester, I have emphasized that the goal of this course is to prepare you for a program of exploration and learning as a necessary component of successful statistical practice in data science. In the second half of the semester I have intentionally chosen texts that you could begin reading in class and complete after the semester. This final project is intended as a final push out of the nest.

You may start your final assignment based on either *In All Likelihood* or *Introduction to Empirical Bayes*. Each starting place has it's own instructions.

In All Likelihood

Begin by reading Chapter 4, working through the examples as usual.

Using R, prepare answers to exercises 4.25, 4.39, and 4.27.

Now consider the data in the spreadsheet file **Illinois_rain_1960-1964.xlxs** which reports amounts of precipitation during storms in Illinois from 1960 to 1964. These data were gathered in a study of the natural variability of rainfall. The rainfall from summer storms was measured by a network of rain gauges in southern Illinois for the years 1960-1964 (Changnon and Huff, 1967). The average amount of rainfall (in inches) from each storm, by year, is contained in the spreadsheet.

Use the data to identify the distribution of rainfall produced by the storms in southern Illinois. Estimate the parameters of the distribution using MLE. Prepare a discussion of your estimation, including how confident you are about your identification of the distribution and the accuracy of your parameter estimates.

Using this distribution, identify wet years and dry years. Are the wet years wet because there were more storms, because individual storms produced more rain, or for both of these reasons?

To what extent do you believe the results of your analysis are generalizable? What do you think the next steps would be after the analysis? An article by Floyd Huff, one of the authors of the 1967 report is included.

Introduction to Empirical Bayes

Continue reading *Introduction to Empirical Bayes*, chapters 5, 6, 11, 12, and 13.

Now, read Chapter 6 in *Computer Age Statistical Inference* which takes you to the next step in using empirical Bayes. Chapter 6 contains four examples – insurance claims, species discovery, Shakespeare's vocabulary, and lymph node counts. In each example, the result of the empirical Bayes analysis is given.

Using R, reproduce the analyses in Chapter 6 describing how you are proceeding and providing commentary. Note that at the end of the chapter, the authors noted that empirical Bayes is often used to support analyses of false-discovery rates which is discussed in Chapter 15.

Final project submission and rubric

The final project is due on 2022 May 12 at 5:00 pm Boston time.

As you work, keep in mind that the project is not intended to be the end of your work on the topics we discussed during 677, but an commencement of the work you have ahead of you as a practicing data scientist.

For this reason, the assignment instructions are open ended. Your job is to engage with the material. Turn in the results of your work which should include working R code, a description of what you have done, and a plan for further work. You will run into problems as you do this work. Write about the problems you encountered. How did you identify the problem? What did you do to overcome them?

Assessment of submitted projects

Assessment of submitted projects will focus on six criteria:

Organization: The work is well-organized and is presented in a form that is easy to read and understand. Your submission should include running text, data analysis, graphical presentations, R code, and mathematical notation.

Planning: You should describe how you distilled the project description into strategy to complete the assignment. Describe what you did as completed the project.

Execution: Use only as much code as you need. Make sure that the code you submit will run. I will be downloading it from your github site.

Clarity: Your code should be easy to read and understand. The code, associated comments, documentation, and presentation should form a seamless whole. Be sure to include a PDF containing your report in case there are any problems running the code

Curiosity: Throughout the year, the important role of curiosity in professional development has been emphasized in the practicum and other classes. In this project, curiosity is essential.

Describe what you learned: The bottom line for the work I have asked you to do is leaning. What did you learn? What will you do next? What will you do differently as you move forward?

Submitted Assignments

Submit your github address after you make your final push. Make sure that your name and the title of the assignment are in the file name and at the top of the assignment. Please push a PDF of the document your code produces to your github just before you submit. The assignment must be your own work but your may consult other students as long as you cite their help and describe what you learned from them.