Statistics 679/992 Assignment 2, Spring 2018

Due Tuesday, February 13, at 11:59pm

Purpose

The purpose of this assignment is to examine a data set with similar structure to the one we have been using as a course example, and to use the tools developed in the course to analyze this data in order to: (1) provide numerical and graphical summaries of the data; (2) fit a random intercept model using both **lmer()** from the *lme4* package in R and with **Stan**; (3) compare estimates and predictions from both models; (4) assess model goodness-of-fit and robustness; (5) make summaries about the relationship between socio-economic status and mathematics exam scores while accounting for school effects; (6) communicate findings effectively.

Background

The data in the file *hw02.csv* contains three variables: Score, which is a score on a mathematics exam; SES, which is an index of socioeconomic status (larger values correspond to a higher status), and School, which is an identifying number of a school. Each case is from a single student. These students are sampled from a larger population of students, and you may make interpretations as if this sample is random. You may also think of the schools as sampled from a larger population of schools.

What to Turn In

I expect that you will create an R Markdown document that contains your data analysis and written responses. You should process this document into a single electronic report (which may be either PDF or HTML) which you will upload through the course Canvas website. When you process your document, please **do not** include any code you write to carry out the analysis unless explicitly requested. Do include graphs, well formatted tables, and well crafted written responses as needed for each question.

Problems

1. Calculate the mean, median, standard deviation, minimum, and maximum of the mathematics exam scores. Produce a graph that displays the distribution of scores. Comment on any notable features of this distribution.
2. Repeat problem 1 for the socioeconomic variable SES.
3. Create a graph that shows the relationship between SES and exam score.
4. For each school, calculate the mean SES value, the mean exam score, and the number of students sampled from the school. How many total students and schools are in the data set?
5. Create a graph with a point for each school that shows the relationship between the mean SES and exam score values. Comment on any notable features.
6. Create a graph that displays the distribution of sample sizes among schools.
7. Identify the labels of these five schools:
   - **School A (small/low)**: Among all schools with the minimum sample size, the school with the smallest average SES value.
   - **School B (small/high)**: Among all schools with the minimum sample size, the school with the largest average SES value.
   - **School C (large/low)**: Among all schools with the maximum sample size, the school with the lowest average SES.
   - **School D (large/high)**: Among all schools with the maximum sample size, the school with the highest average SES.

- **School E (typical)**: Among all schools with the median sample size, the school with the median mean SES score.

8. Fit a random intercept model using **lmer()** from the *lme4* package. Report the estimated parameter values and associated standard errors in a table. Write all of the distributional assumptions among the parameters and data.
9. Fit a random intercept model using **Stan**. Use normal(0,100) prior distributions for the slope and intercept and Half-Cauchy(0,5) prior distributions for the individual level and school level standard deviations. Report the estimated parameter values and associated standard errors in a table. Write all of the distributional assumptions among the parameters and data.
10. Compare 95% confidence intervals/credible regions for the slope and intercept of each model.
11. Run another Stan analysis with the same model as problem 9, but a different random seed. How do the 95% credible regions of the slope and intercept compare with those computed in the Stan analysis in problem 10?
12. Run another Stan analysis, but change the prior distributions for the slope and intercept to be Cauchy(0,5). How do the 95% credible regions of the slope and intercept compare with those computed in the Stan analysis in problem 10?
13. Create a graph that compares the estimated school effects from the lmer analysis with the posterior means of the school effect distributions from the Stan analysis. Comment on what the graph says about the similarity of the two analyses.
14. Refer to the schools identified in problem 7. For each school, find 95% confidence intervals/credible regions for: (1) the mean mathematics exam score for all students in the school; and (2) the mean mathematics exam score for all students in the school with an SES value of 7. Compare these intervals from the lmer and Stan analyses and briefly comment on similarities and differences (among the schools and between the two analyses).
15. Refer to the schools identified in problem 7. For each school, find a 95% prediction interval for the mathematics exam score of a single student with an SES value of 7. Comment on differences and similarities between the two analyses and among the five schools.
16. Create residual plots for each analysis. Refer to these residual plots and comment on the suitability of the models as a framework for making inferences about the relationships between SES and mathematics exam scores for students in this population. If you were to pursue an alternative model, what might you consider doing?