RateMyProfessors Scraping with Hierarchical Models in JAGS

Part 1 of this project is done in Python with the part1.py script, dynamically web scraping RateMyProfessors to obtain data on Acadia University, Carleton University, Memorial University of Newfoundland, Mount Allison University, and Mount Saint Vincent University. For each school, the data is then structured (professor name, rating, difficulty, "would-take-again" percentage, and department) and saved to a .csv file in data/. We are now ready to apply hierarchical models to this data using JAGS in R. The resulting HDR/distribution plots from these R scripts are saved to the plots/ folder. (As an aside: essentially the only reason there is multithreading in that script is because I went insane with the wait times during the prototyping phase. No point in removing it now.)

In utils.R, we set up a few constants and utility functions common across Parts 2–4. In part2.R, we build a simple hierarchical model to estimate the modal rating of professors both throughout the entire population and for each school. In a normal distribution, the mode and mean are the same; deciding to model our data normally, we simply use JAGS to model means. Pretty much all six resulting distributions have modes centered around 3.6, with some but not too much variation in this.

In part3.R, we build a more complex hierarchical model to estimate the modal rating of professors, also including a level of department below the school level. This is arguably more realistic, as some programs are far more difficult than others. Given this refinement, we now see now more variation than in Part 2, even at the school level; for instance, MSVU has a much lower mode now, closer to 3.4. We opt to only make HDR plots for each level 1 distribution here (just as in Part 2), as there are far too many departments at each school.

In part4.R, we cut out the school level, keeping just the department level. Refraining from creating a plot for every single department as in Part 3, we plot only the population-level HDR and distribution. Our mode generated by JAGS is now closer to 3.7 than 3.6, somewhat higher than in Parts 2 and 4. Based off of real-world knowledge, this model is likely better than Part 2 but worse than Part 3—there is likely more variation in difficulty (and thus in rating) across departments than there is across school, but different student attitudes at different schools likely still plays a significant factor, so we should include both.