Visually Analyzing and Running Multilevel Data in R and BUGS

Jimmy Wong[1,*], Dr. Beth Chance[1]

1. California Polytechnic State University, San Luis Obispo
*Contact author: jwong100@calpoly.edu

Keywords: multilevel models, bugs, lme4, ggplot2, animation

Education data collected for research on students' performance and attitudes in a randomization-based introductory curriculum were analyzed using a multilevel modelling approach. Our data consisted of two sets of observational units that were students and instructors, where we had multiple students nested in each instructor. The intention of applying a multilevel modelling approach originated from the nature of the data, our desire to use variables collected on students and instructors in the same regression model, and our intention to account for any instructor-to-instructor variability in the response variable while running a regression model.

R was the primary statistical software used in running our multilevel models, producing graphical displays, and generating animations/gifs. During the first stages of our analysis, the main packages that we used in the order of importance to our study were **lme4**, **ggplot2**, **animation,** and **gridExtra**. Prior to running the actual multilevel models, we conducted exploratory analysis on the data by using functions from **ggplot2, gridExtra,** and **animation** packages**.** We primarily created multi-dimensional graphs to observe how variables associated and interacted with each other. Examples of such graphs were facetted boxplots conditional on two variables, scatterplots of students and instructors' variables weighted by class size with regression lines imposed, spaghetti plots color coded by an instructor-level variable, mosaic plots, and overlaid marginal and conditional histograms. The codes required to create these graphs extended to more than just using `ggplot` and `geom_point`, for example. There were many tweaks made to our graphs to have them appear more intriguing and original. After generating our graphs, **gridExtra** assisted us in combining multiple graphs to a single cohesive graph when we wanted to compare the idea of marginal versus conditional. These combined graphs were then saved to jpeg files using `ggsave` so we can access them directly in the future without having to run any code in *R*. Shifting our focus from graphs to animations, the **animation** package assisted us in further presenting some of our ideas such as showing the process of creating spaghetti plots and the considerable variability in the many variables among instructors in a more interesting, but yet, educational way by compiling individual graphs to unified gif files with the functions of `saveGIF` and `saveHTML`.

After our exploratory analysis was done, **lme4** assisted us in running multilevel linear and logistic models. The functions of `lmer` and `glmer` in **lme4** were used for quantitative and categorical response variables, respectively. The regression outputs provided us estimates of the coefficients and the variances components as well as giving us information on the significance of the predictor variables. However, we decided to obtain another set of estimates with a Bayesian approach. This idea lead us to the two *R* packages of **arm** and **R2WinBUGS.** We ran our multilevel models in *BUGS* by communicating from *R* to *BUGS* mainly with the function `bugs`. Therefore, our final results were a combination of frequentist and Bayesian methods and comparisons were made between the estimates from these two methods.

In summary, we used a multilevel modelling approach to analyze education data that had variables measured on both students and instructors. Our exploratory analysis consisted of generating graphs and animations to visually analyze the associations of different variables from our data set by using functions from the *R* packages of **ggplot2, gridExtra,** and **animation**. The assistance of **lme4, arm,** and **R2WinBUGS** in running our multilevel models allowed us to combine results from a frequentist and Bayesian method. As future work, we may consider compiling an R package that generalizes our code for graphs to allow users to input their variables of interest and also creating a Shiny application that encompasses as much stages of our analysis as possible to allow users to modify input parameters and to educate multilevel modelling to students or anyone who may be interested in multilevel models.