Homework 1

140.656 Multi-level models

BEFORE YOU GET STARTED:

Here are the rules of conduct for Homework 1:

* We encourage you to work in groups on completing the required data analysis for the homework assignment, i.e. Parts A – E.
* Post questions about Stata/R commands! You should make a reasonable attempt at figuring out the commands (Stata help is great and so is a Google search); however, do not spend HOUR(S) on a single command/figure. Please ask for help!
* Parts F and G of the assignment may be completed in teams of up to 4. Note: if you are submitting as a team, all team members will receive the same score for the assignment.

**Submitting your assignment:**

Check course plus for the deadline. **If you are submitting as a team, a single team member should submit the document to the Course Plus drop box.**

Your Homework1 submission should consist of:

* The one-page abstract from Part F
  + Document should be single-spaced, 10 point Times font, one inch margins.
  + Code does not need to accompany results for the main abstract, but code SHOULD accompany short answer questions. Since we grade for correctness, we will need to see your code.
  + A title and author list are at the top of the document (may be included in a header above the one inch margin)
  + Word docs are preferred, but PDFs are ok. (R users can knit to word docs in R studio, too!)
* At most 2 tables/figures that are referenced in the abstract
* Answers to Question 1 – 3 from Part G
* Students should try their best to adhere to formatting guidelines for reporting statistics in manuscripts. You may find this reference helpful: https://dx.doi.org/10.21315%2Fmjms2016.23.5.1
* If you are submitting as a team, then you should also include the following contribution section on a separate page at the end of your Homework document:
  + A statement about who contributed what to the data analysis (e.g. Initials1, Initials2, Initials3 met as a working group to complete the data analysis in Parts A – E)
  + A statement about who contributed what to the abstract writing (e.g. Initials1 drafted the Objective, Study Design and Conclusion section. All three authors contributed equally to the writing of the methods and results section. Initials2 edited the final abstract.)
  + A statement about who contributed what to the short answer questions (e.g. All three authors met as a working group to compile answers to the short answer questions.)

No late submissions will be accepted.

**Description of the Data:**

The data are from the High School and Beyond (HS&B) study conducted within the National Education Longitudinal Studies (NELS) program of the National Center for Education Statistics (NCES). The NELS was established to study the educational, vocational, and personal development of young people beginning with their elementary or high school years, and following them over time as they begin to take on adult roles and responsibilities. Thus far, the NELS program consists of five major studies: the National Longitudinal Study of the High School Class of 1972 (NLS-72), High School and Beyond (HS&B), the National Education Longitudinal Study of 1988(NELS:88), the Education Longitudinal Study of 2002 (ELS:2002), and the High School Longitudinal Study of 2009 (HSLS:09).

The HS&B survey included two cohorts: the 1980 senior class, and the 1980 sophomore class. Both cohorts were surveyed every two years through 1986, and the 1980 sophomore class was also surveyed again in 1992.

We have available data from 7042 students within 156 schools.

The study variables include:

Level 1: student

mathach: a measure of mathematics achievement

minority: dummy variable for student being non-white

female: dummy variable for student being female

ses: socioeconomic status (SES) based on parental education, occupation and income (z-score)

Level 2: school

schoolid: school identified

sector: dummy variable for a school being Catholic

pracad: proportion of students in the academic track

disclaim: scale measuring disciplinary climate

himinty: dummy variable for more than 40% minority enrollment

newid: rescaled school identifier, counts 1 to 156 (we created this for you)

**Objectives:**

The goals of the analysis are

1. to understand the variation in average mathematics achievement (MA) scores across schools
2. to determine if the composition of the school (students’ ethnicity, gender and SES) explains the observed variation in average MA scores across schools
3. to determine if school-level characteristics account for variation in the average MA scores across schools

**Part A: Exploratory Data analysis**

1. Descriptive statistics
   1. Calculate the minimum, median, mean and maximum number of students per school.
   2. Calculate descriptive statistics for the student level variables: MA score, ethnicity, gender and SES.
   3. Calculate descriptive statistics for the school level variables: school type (sector: Catholic vs. not), proportion of students in the academic track, measure of disciplinary climate and minority enrollment (indicator of more than 40% minority).
   4. Calculate descriptive statistics for the school-level average MA score, proportion of minority students, proportion of female students and average SES. Here the goal is to quantify variation in the composition of the schools; i.e. how do the students within a school vary across schools.
2. Assess the variation in MA scores within and across schools by creating a figure that displays both the within-school and between-school variation in MA score.

**Part B: Ratio of between to within-school variation (ICC)**

1. The two-stage normal-normal model
   1. Write out the appropriate two-stage normal-normal model to estimate an overall population average MA score assuming that the schools in our sample represent a random sample of all schools. Be sure to define all appropriate distributional assumptions.
   2. Fit the model and save the estimated random intercepts
   3. From the model fit, estimate the percentage of the total variation in average MA scores that is attributable to differences between schools; the intra-class correlation coefficient.
   4. Perform an analysis to explore whether the estimated random intercepts are approximately normally distributed.

**Part C: Compositional Effect**

1. Review the following definition of compositional effects (extracted from http://www.paho.org/english/dd/ais/be\_v24n3-multilevel.htm): **“**When inter-group (or inter-context) differences in an outcome (for example, disease rates) are attributable to differences in group composition (that is, in the characteristics of the individuals of which the groups are comprised) they are said to result from compositional effects.”
2. Extend the two-stage normal-normal model to include variables describing the composition of the students within each school: proportion of minority students, proportion of female students, and average SES.
   1. Create a summary figure assessing the form and strength of the relationship between the school average MA score and the student composition variables listed above.
   2. Write out the model, defining all variables and distributional assumptions
   3. Fit the model and save the estimated random intercepts.
   4. Compare the estimated variance of the school-specific intercepts to the value obtained from Part B.
   5. Does the student composition of the schools help explain variation we observe in the average MA score across schools?
   6. What, if any, of the student characteristics are statistically associated with the school average MA score?
   7. Explore the normality assumption of the random intercepts.

**Part D: Adjustment for additional school-specific characteristics.**

1. Extend the model you fit in Part C to include the school-level characteristics school type (sector: Catholic vs. not), proportion of students in the academic track, measure of disciplinary climate and minority enrollment (indicator of more than 40% minority).
   1. Create a summary figure assessing the form and strength of the relationship between school average MA score and the school-level characteristics listed above.
   2. Write out the model, defining all variables and distributional assumptions
   3. Fit the model and save the estimated random intercepts.
   4. Describe any changes to the estimated random intercept variance; i.e. after adjusting for the composition of the students within the school, do any of the school-level factors explain variation we observe in the school average MA scores?
   5. What, if any, school-specific characteristics explain variation in the school average MA scores?
   6. Assess the normality assumption of the random intercepts.

**Part E: Summary figure**

Create a figure displaying the reduction in the estimated variance for the true school mean MA scores. Hint: you can use the estimated random intercepts from the three models.

**Part F: Abstract**

Write a BRIEF abstract summarizing your findings. You MUST adhere to the following guidelines:

* 1. Your abstract should be NO LONGER than ONE page (standard 1 inch margins, single spaced). We will only read the first page if you submit an abstract longer than one page. Do NOT cut and paste Stata output within the one page abstract!
  2. In ADDITION to the ONE page abstract, you may include up to 3 tables/figures.
  3. Your abstract should include:
  4. An **objective** or description of the goal of the analysis
  5. A **study design** section describing the dataset and key variables of interest
  6. A **methods** section describing your statistical analysis, including enough detail so that your reader could replicate your analysis
  7. A **results** sectionthat includes a) descriptive statistics for study population, b) a summary of your key findings relating to the presence of a compositional effect, c) a summary of your key findings relating to school-level characteristics explaining variation across schools and d) identification of which student composition and school-level characteristics were associated with school mean MA scores.
  8. A **conclusion**.

**Part G: Short answer questions**

**Fit the following models in Stata or R:**

import delimited “hsb\_data.csv”, clear

bys schoolid: egen prop\_female = mean(female)

bys schoolid: egen mean\_ses = mean(ses)

bys schoolid: egen prop\_minority = mean(minority)

\* Model 1:

mixed mathach prop\_female mean\_ses prop\_minority || schoolid:

\* Model 2:

mixed mathach prop\_female mean\_ses prop\_minority sector pracad disclim himinty || schoolid:

df <- read\_csv("C:/Users/Elizabeth/Dropbox/MLM2019/Labs/Lab0/hsb\_data.csv")

df <- df %>%

group\_by(newid) %>%

mutate(prop\_female = mean(female),

mean\_ses = mean(ses),

prop\_minority = mean(minority))

# Fit Model 1

model1 = lmer(mathach~prop\_female+mean\_ses+prop\_minority+(1|newid),data=df)

# Fit Model 2

model1 = lmer(mathach~prop\_female+mean\_ses+prop\_minority+

sector+pracad+disclim+himinty+(1|newid),data=df)

1. Interpret the estimated regression coefficient for the proportion of minority students variable from Model 1. Be sure to include the estimate and 95% confidence interval.
2. Interpret the estimated regression coefficient for sector (catholic vs. not) from Model 2. Be sure to present the estimate and 95% confidence interval.
3. Suppose I fit the following model:

**STATA:** mixed mathach female ses minority || schoolid:

**R:** lmer(mathach~female+ses+minority+(1|newid),data=df)

Under what conditions would the estimated random intercept variance obtained from this model be the same as the results of Model 1?