In Module 12, we will review checking assumptions in multilevel models.

Note that when using R, you can usually do any one task in multiple ways. This is especially true of checking assumptions because it essentially entails exploring the data. There are many ways to explore your data/check assumptions beyond those in this data demo.

About the data:

The data for Module 10 are a subsample from the 1982 High School Beyond data collected by the National Center for Educational Statistics, used as example data throughout Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*: SAGE Publications.

* 7,185 students clustered within 160 schools
* *Minority* (1 = Minority, 0 = White)
* *Female* (1 = Female, 0 = Male)
* *SES* (Z-score from parental edu., income, and occupation)
* *Math Achievement* (mean = 12.75, SD = 6.88)
* *Size* (Number of students in the school)
* *Sector* (1 = Catholic school, 0 = public school)
* *SES\_mean –* the school level aggregate of SES
* *N\_Break —* the number of students in each school in our sample
* *CWCses —* the school mean (centered within cluster) value of SES for each student (their distance from their school’s mean on SES)

1. Load the data and libraries we will use for this module: lme4, lmerTest, dplyr, and ggplot2.
2. Create a subset of 30 schools dataset to make plotting easier for this demo. For your data, you should consider all observations.
3. Consider how related SES is to the outcome.
   1. Create a scatterplot of SES and mathach for the whole sample versus the subset, what are some of the limitations of this plot given the data we have?
   2. Create a correlation matrix for SES, SES\_mean, and Mathach
4. Let’s take a look at the relationship between SES and Mathach to identify any potential outlying schools, nonlinear relationships at the school level, or other funny business. Create a scatterplot for each school. From now on we will only work with the 30 school subsample.
5. Create a scatterplot with fitted regression lines by school. What sort of information does this give you? What are the limitations?
6. Run a model with CWCses at level 1 and meanses at level 2 (estimate the variances and the crosslevel interaction)
7. Make a histogram and QQplot of the residuals at level 1, do they look normal?
8. Check that the predictors are uncorrelated with the level 1 residuals by creating a scatterplot and calculating the bivariate correlation for each.
9. Create a plot to see if the level 1 error variances are equal and independent of sex.
10. Create an aggregate of the outcome at level 2 and ses and save to a new dataset
11. Create a scatterplot of the level 2 relationship from the new dataset. Add the regression equation and line of best fit to the scatterplot
    1. How is this regression equation related to the full MLM we ran earlier?
12. Output the level 2 residuals.
13. Using the intercept and slope residuals, create a histogram, qq-plot, and boxplot for each.
    1. \*side note—having a L2 dataset is nice because you can get simple information like how many clusters and easily run other descriptives for L2 of the data that you might want to report in a paper.
    2. Are there any L2 units you might want to follow up on, from these plots?
    3. Should we be concerned with normality?
14. Merge the L2 residuals with the L2 dataset from earlier
15. Check that the L1 and L2 residuals are unrelated
16. Are predictors at one level related to error at another level? Plot L1 residuals by Mean SES and L2 residuals by CWCses
17. Create a scatterplot of the intercept residual and mean\_ses
18. Create a scatterplot of the slope residual and mean\_ses