Instructions for the homework with Virginia Education data

Minimum requirements:

Model the pass percentage in high school mathematics in Virginia schools using a multilevel linear model with schools as individual units and divisions as groups. Consider only high schools and the highest grade in the school only. Also, consider only the 2015-16 pass percentage here. Optionally, you may consider other groupings – gender, ethnicity etc.

The predictor variables at the individual level should necessarily include the accreditation score for mathematics for that high school and the ethnicity of students in that school. In addition to these two predictors, come up with ONE extra predictor variable and justify why you chose this variable.

At the group level, the predictor variables are the average teacher salaries for public schools in that division and the truancy (raw numbers are given – normalize this by the total number of students in that division). Again, you need to come up with ONE extra group level predictor variable.

In your presentation, start with a couple of slides of descriptive data analysis using figures. Show the others what interesting facets of the data did you see and how it influenced your modeling decisions – these may help justify why you would expect some predictor to be important or not etc. In addition, explain any data cleaning steps you may have to undertake here –missing data imputations, for example, or considering a composite variable because that is what makes sense etc.

Build the model next and explain what the model means in the notations we have discussed, and more importantly, in words.

Fit the model, explain the model fit estimates etc. When doing this, do not directly start with the full model. Instead, start with baseline models like the complete and the no-pooling models. Consider also the model with just indicator variables for divisions, and the simple “two-step” model we discussed in class. Doing this tells you if multilevel modeling is really going to give you information. Discuss this point also with reference to the variances in the data and the structure of the groupings etc.

Do model diagnostics – both using numerical summaries and graphical methods such as showing posterior predictive distributions, cross-validation etc.

Throughout these steps, use only a certain number of divisions (randomly chosen preferably) and keep aside a small percentage of counties to do predictive tests. Also note that in order to do cross-validation, you will have to set aside a certain fraction of cases as test data and work on only a subset of the full dataset as the training data. Try and come up with heuristics on when the model becomes bad – in predicting out-of-sample data points within existing groups and also in predicting completely new groups. Do some “stress testing” to figure out how robust the model is.

Optionally, you can look at more complicated models with additional levels of grouping, with more variables etc.