EXERCISE ON FIXED EFFECTS, RANDOM EFFECTS, CLUSTERED STANDARD ERRORS

In the directory containing this exercise is a dataset on California schools, in .csv, .xlsx, and .RData formats. There is also a .docx file with brief descriptions of the variables. In this exercise, you consider what explains variation across school districts in test scores, Some of the variables — expenditures per student, computers per student, and student-teacher ratio — might be subject to policy choice, so accurately estimating them could be important.

You can do this exercise in R using R's built-in regression estimation functions together with a function in the "sandwich" R package, plus the "nlme" package. The sandwich package you probably have to install with **install.packages**(), but the nlme package is one of those distributed with R itself, so you only need to load it with **library**(nlme). The nlme package takes a likelihood approach and thus delivers a probability model of the data (instead of just estimates with standard errors). However, it applies maximum likelihood, rather than allowing posterior simulation. There is a package, "brms", that allows that, but it is a huge package that takes a long time to install and to learn, so you're not asked to use it in the exercise.

You are welcome to use any software you like in doing the exercise. Much of it, and possibly all of it, could be done with Stata.

(1) Estimate a linear regression of the average test score (`testscr`) on student-teacher ratio, computers per student, and expenditures per student. Determine whether the three variables have expanatory power by an F-Test of the hypothesis that all three have zero coefficients and via the Bayesian information criterion (BIC). The latter can be computed from an F-statistic: The BIC rejects the restriction when the F-statistic exceeds the log of the sample size.

(2) Do the same thing with a regression that adds the demographic variables: Average income, subsidized meals, calWorks per cent, and English learners percent. Again check whether the three "policy variables have explanatory power using an F test and BIC. Here you may need to extract the covariance matrix of coefficients from the **lm**() output to construct the F or chi-squared statistic.

(3) Repeat the previous estimations and tests in models that add county fixed effects. In R using **lm**(), this is accomplished by just adding "county" to the list of right-hand side variables. (county is a "factor" in the R dataframe, so R automatically converts it into the appropriate array of dummy variables when including it in a regression.)

(4) The districts vary greatly in size. Average scores might have more sampling variation in small districts. Plot the squared residuals from the estimated model in 2 against the total enrollment variable. Estimate a linear regression of these squared residuals on `1/enrl_tot`. Use the inverse of these predicted values as the **weights** argument

Date: February 26, 2021.

in `lm()` (or otherwise estimate the corresponding weighted regression estimates) in the question 2 regression.

(5) For at least two of the above regression models, calculate standard errors clustered by county. This is done very easily with the `vcovCL()` function from the sandwich package — so easily that if you're doing it this way you might want to see how much difference it makes in all of the above regressions.

(6) Estimate a random effects model, with county effects. In R, use the `lme()` function from the nlme package to estimate the 7-variable regression, with random effects by county. You do this by giving `lme` the argument `random = ~1 | county`. Also use the argument `method="ML"`, so that the estimation is by maximum likelihood.

(7) Compare the random effects 7-variable model to the fixed effects model. In R, you can do this by re-estimating the fixed-effect model with the `gls()`function from the nlme package, again being sure to use `method="ML"` argument. The `summary()` function applied to either random effects or fixed effects models computed this way deliver both log likelihood and BIC values, so the models can be compared both by a frequentist chi-squared test based on the log likelihood and via the BIC.

(8) Finally, for the random effects model, use a regression of its squared residuals on 1/(total enrollment) to generate weights for a weighted random effects estimation; see if this improves likelihood and/or changes important estimates. [Note: I think that in the nlme estimation functions the "weights" arguments are variance scales — the inverse of the weights used in `lm()`. So you would use a `weights= ~w` argument to `lme()` if you used `weights=1/w` in `lm()`].

(9) Be ready to discuss: Does the evidence favor an important effect from the "controllable" variables? The sizes and signs of the estimated effects, not just the significance levels of tests, should inform your views on this.