3 Analysis of Experimental Data: Project STAR

In the late 1980s the Tennessee state legislature funded a four-year experiment to evaluate the effect of small class sizes on learning, project STAR (Student Teacher Achievement Ratio). The experiment compared three different class arrangements for children in kindergarten through third grade:

- a regular-size class (22-25 pupils) with a single teacher – the control group
- a small class (13-17 pupils) with a single teacher
- a regular-size class with a single teacher and a teacher's aide

Participating schools were picked at random from the universe of public schools in Tennessee. Each participating school had at least one class of each type. Within schools both pupils and teachers were randomly assigned to one of the three types of classes. Each year the children were given standardized tests (the SAT) and these are the outcome measures we look at here.

In this exercise we ask you only to analyze the results for children in kindergarten and to compare the children in the small class with those in the regular class without aide. The dataset `STAR.dta` can be downloaded from OLAT.

The variables included in the dataset are:

- tscorek (Y): the outcome variable, test score of the child
- sck (D): whether kid in small class (1) or regular class (0)
- schidkn (X): code for particular school
- girl (W_1): whether the kid is a girl (1) or boy (0)
- freelunk (W_2): whether the kid gets a free lunch (poor household) (1) or not (0)
- totexpk_m (W_3): months of teaching experience of teacher

1. Summary statistics
   (a) Are there more girls or boys?
   (b) What't the average test score? What's the median? What are the average test scores in treatment and in control group?
   (c) How many years of experience does the most experienced teacher have?
2. OLS Regression
   (a) Run a regression of the test score on the treatment indicator, whether the kid was in a small or regular class (we are going to ignore the other treatment – being in a small class with aide). Interpret the coefficient, i.e. what does the number tell us. Assess whether the effect is small or large in economic terms.
   Hint: Use the sample standard deviation of testscores.
   (b) Now include school fixed effects (use the `xi` command in Stata or the `plm` package in R). Which estimate of the class size effect do you prefer and why?
3. Standard Errors

(a) A researcher argues that because D is randomly assigned it must be independent of the error term, U and hence we do not have to worry about using robust standard errors. Do you think this is correct? Which standard error do you prefer to use in this application and why?

(b) Now include school fixed effects and adjust your standard errors for clustering within schools (use the `cluster` command in Stata or `vcovHC` in R). Explain why allowing for correlated errors within schools might be important and comment on your results relative to the specification in 2) above.

4. Testing whether randomization worked (using Ws)

   (a) First, test whether girl, freelunk or totexpk_m are correlated with D by running separate regressions. What should you be looking for in the results from these regressions?

   (b) Test whether any of girl, freelunk or totexpk_m can "explain" treatment status. Interpret the test result.

   (c) Include girl, freelunk and totexpk_m into the regression you ran in 2(a) above. What would you expect to happen to the estimate of the treatment effect compared to 2(a) above? Is this what happens? Is the parameter estimate on totexpk_m causal? Is the freelunk estimate causal? And the estimate on girl?

5. Heterogeneity in Treatment Effects

   (a) Generate the interaction of the treatment indicator, D or sck, with girl, freelunk and totexpk_m. Call these interactions `sck_girl, sck_freelunk, sck_totexpk_m`.

   Estimate the regression including these interactions (including girl, totexpk_m and freelunk separately and school fixed effects as well). Interpret the coefficients on sck, sck_girl, sck_freelunk and sck_totexpk_m. For what groups does the treatment effect seem to be largest?

   Use an F-test for the joint null hypotheses that the treatment effect is the same for everyone.

   (b) Compute an estimate of the average treatment effect (ATE) by averaging the individual treatment effects. You can do this by typing in STATA:

```
generate te = _b[sck] + _b[sck_girl]*girl + _b[sck_freelunk]*freelunk +
_b[sck_totexpk_m]*totexpk_m
```

   after the estimation and then use (again) the `summarize` command (the command showing summary statistics). How does this estimate of ATE compare to the estimate of ATE we got when we did not include any interactions (i.e. the result found in the answer to questions 3) and 4) above?