Homework 12: Teacher, Leave Those Kids Alone! (They're the Control Group)

36-402, Spring 2025

Due at 6:00 pm on Thursday, 17 April 2025

Agenda: Causal inference will continue until morale improves.

The Tennessee STAR project was a randomized experiment which sought to determine whether children learn more in classrooms with fewer students. Students within participating schools were randomly assigned to either (i) small (< 18 student) classrooms, (ii) to ordinary-sized classrooms, or (iii) to ordinary-size classrooms where the teacher had an aide. We will focus on the contrast between (i) small classrooms and (ii) ordinary-size-with-no-aide classrooms (or "regular classrooms" for short).

The study began in kindergarten and continued through third grade. Students initially assigned to the small-class condition for the most part stayed in it (there were a few unavoidable exceptions for administrative reasons); students assigned to the two large-class conditions were re-randomized in the second year of the study, and thereafter changed only minimally. New students entering the schools in the study were randomized into the three conditions. Teachers were also randomized as to which kind of classroom they got. Learning was assessed (in the initial phase of the project) through annual standardized tests of reading and math.

A standard version of the data set is available as `STAR` in the `AER` package, which you may need to install. See `help(STAR)` for the definitions of the variables named below.

General: Whenever you are asked estimate something (a coefficient, a treatment effect, etc.), you should also give standard errors for your estimate. Those standard errors should be bootstrapped, *unless* you explain why, for a particular situation, R's default calculations of standard errors should be reliable. Unless explicitly called for, do not report R's p-values, or any significance stars.

1. *Causality? Reverse causality?*

   (a) (3) Linearly regress `readk` and `mathk` on `stark`. Report the coeffcients and standard errors. Explain why a non-parametric regression would be redundant here.

   (b) (2) Linearly regress `read3` and `math3` on `stark`. Report the coefficients as above.

(c) (5) Explain how a randomized treatment received in kindergarten can predict test scores three years later.

(d) (2) Linearly regress `readk` and `mathk` on `star3`. Report the coefficients as above.

(e) (5) Explain how a treatment received in the third grade can predict test scores in kindergarten, three years *earlier*.

(f) (5) To find the causal effect of the `stark` on `readk` and `mathk`, should we control for `star3`? (Explain.)

(g) (5) To find the causal effect of the `star3` on `read3` and `math3`, should we control for `stark`? (Again, explain.)

2. (10) For each year from kindergarten through third grade, estimate the *difference* in expected reading and math scores when students are assigned to a small classroom versus an ordinary classroom. (That is, estimate the average treatment effect of a small versus an ordinary classroom.) You may present your results either as a table or graphically; make sure it's easy to read and compare across years and subjects.

   Explain how you obtained your estimates, and why that procedure is, for this data, a valid way of estimating the desired average treatment effects. If you have to control or adjust for any covariates to get the average treatment effects, explain which ones you used and why.

3. *Heterogeneity of effects* There is considerable interest in knowing whether the effects of smaller classes are different for different groups of students.

   (a) (8) Estimate the average treatment effects of small versus ordinary-size-without-an-aid classrooms on kindergarten reading and math scores, for Caucasian and for African-American students[1].

   (b) (2) Explain why, to get such estimates from linear regression, the right models would be of the form `lm(readk~stark*ethnicity)`, and why `lm(readk~stark+ethnicity)` would be uninformative.

4. *Observational inference in an experimental study* Students whose families are sufficiently poor qualify for free lunches at school. This is recorded in the variables `lunchk` through `lunch3`. We want to know whether being above or below this threshold level of poverty has a causal effect on student's scores.

   (a) (2) Estimate the mean scores for reading and for math in each grade for students who do and do not qualify for free lunches (in that grade).

   (b) (5) If we want to find the effect of `lunchk` on kindergarten reading and math scores, does it make sense to control for `stark`? Explain.

[1] These are the names used at the time in the records. The other four racial/ethnic groups in the study had so few students that their results are very uninformative; do not report them.

(c) (10) Consider the following variables: `gender`, `ethnicity`, `schoolk`, `experiencek`, `tethnicityk`, `systemk`, `schoolidk`, `lunch1`. When estimating the effect of `lunchk` on kindergarten test scores, which of these should be controlled for, which of them should not be controlled for, and which of them do you not have enough information to say? If you answer "not enough information" for any variables, what more would you have to know? (Be more specific than "I would need the complete causal graph".)

(d) (5) If we want to find the effect of `lunchk` on first-grade reading and math scores, under what assumptions should we control for `readk` and `mathk`? Under what assumptions should we not control for them?

5. (1) *Timing* How long, roughly, did you spend on this assignment? How much of that time was spent on math, on coding/debugging, and on writing?

Presentation Rubric (15): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. All parts of all problems are answered with actual coherent sentences, and raw computer code or output are only shown when explicitly asked for. Text from the homework assignment, including this rubric, is included only when relevant, not blindly copied.

(In Gradescope, assign *all* pages to this rubric.)

Code Rubric (15): The code is logically organized and easy to read. No redundant code; no needlessly repetitive code; no unused code. Variables and functions have descriptive and appropriate names. (Loop or array indices, arguments, etc., can have short, conventional names such as `i`, `x`, `df`, etc.) All functions have comments defining their purpose, their inputs, their outputs, and any dependencies on other code you wrote. Vectorization is used wherever appropriate. Allowed packages: `knitr`, `tidyverse`, `dplyr`, `ggplot2`, and those explicitly mentioned in the textbook or the assignment for implementing particular methods. (Any other packages require prior permission from the professor, which must be renewed for each assignment; record the date on which you got permission in your comments.) Code from the textbook and class examples is used wherever possible and appropriate. In particular, it should be used for tasks like bootstrapping, calibration plots, and cross-validation (*unless* the package implementing a model includes its own cross-validation functions). All plots and tables are generated by code included in the R Markdown file. Numerical results (etc.) appearing in text are neither hand-copied nor spat out with `cat()`, `print()`, `sprintf()` etc., but instead properly formatted through in-line code.

(Do not assign any pages to this rubric; instead, submit your Rmd file to the "HW *k* R Markdown File" assignment on Gradescope, for the appropriate *k*.)