Regression II
Vanderbilt University
Prof. Sean P. Corcoran

Due Thursday October 5, 2023
113 total points

Problem Set 3

Instructions: Answer the following questions in a Stata do-file. Submit your problem set as a PDF via email to `sean.corcoran@vanderbilt.edu`. Use your last name and problem set number as the filename. Working together is encouraged, but all submitted work should be that of the individual student.

Question 1. Return to the NELS-88 data used in Problem Set 2 to estimate the academic benefits to attending a Catholic high school. This time—rather than exact or nearest neighbor matching based on covariates—we will use propensity scores. (**30 points**)

(a) For this analysis, you will want to use the "continuous" family income variable you created in Problem Set 2 and the dummy variables for parents' highest level of education. In addition, create z-scores for the 8th and 12th grade reading and math scores. (**3 points**)

(b) Use `teffects psmatch` or the `psmatch2` commands to estimate a propensity score model where the "treatment" is attending a Catholic high school. The goal will ultimately be (in part c) to estimate the ATT on 12th grade test scores, high school graduation, and post-secondary enrollment, using nearest neighbor matching (based on the propensity scores). This will be an iterative process where you experiment with a propensity score model, check for balance, and then make adjustments to your propensity score model and matching rules as needed. I recommend using the `quietly` prefix with `teffects psmatch`, or omitting the `outcome()` option if you use the `psmatch2` command. (You do not want treatment effect estimates to guide your specification search). Here are a few tips/requirements: (**12 points**)

- Choose predictor variables that are likely to be associated with the "treatment" that you would ultimately like to see balanced between your treated and untreated group. You can use any variables in the dataset that you deem appropriate.
- After fitting the propensity score model, check for balance on your predictor variables using `teffects summarize, box` and `density`. (Or use `pstest` after `psmatch2`).
- You may be able to improve balance by changing your propensity model specification—e.g., omitting or including variables (depending on how predictive or theoretically important they are), entering variables as continuous or categorical, adding interactions or nonlinear terms (e.g., quadratic)—and/or by tinkering with the number of nearest neighbors or caliper. Use your own judgment when deciding on "good enough" balance.

- You do not need to provide results for all of the iterations you attempt. Just include a short written explanation justifying your choice, and provide balance tests/figures for your final specification. (You can refer to other specifications you tried in your write-up).

- Lastly, once you have settled on a propensity score model, check for overlap in the distribution of propensity scores between your treated and untreated group. (E.g., `teffects overlap` after `teffects psmatch`, or using code provided in class). Include your graphical results with your output. Presuming you have good overlap, you can proceed to the next step. If there is poor overlap, you should revisit your propensity score model.

(c) Once you are satisfied with your propensity score model, calculate the ATT for the four outcome variables: 12th grade reading and math z-scores, high school graduation, and post-secondary enrollment. Interpret the results in words. What assumption(s) are required for these estimates to be considered causal? (**5 points**)

(d) Rather than nearest neighbor matching, propensity scores can also be used in weighting estimators. Using the same propensity score model you settled on in part (b), estimate the ATT and ATE using inverse probability weighting (`teffects ipw`). How do your results differ from those in part (c), if at all? (**5 points**)

(e) Imbens (2015) proposed an algorithm for estimating propensity score models. The algorithm iteratively adds linear and quadratic terms (and interactions), keeping terms that improve the predictive fit of the model. The user-written command `psestimate` executes this algorithm and outputs the resulting propensity scores to your dataset. (You will need to install it using `ssc install pseestimate`). The syntax of the command is:

```
psestimate treat, totry(varnames) genpscore(newvarname)
```

where *treat* is the treatment variable, *varnames* is a list of predictor variables you would like to try in the algorithm, and *newvarname* will be the new variable containing the propensity score. Choose a set of variables for the algorithm to try. (Note if you choose a long list it make take quite awhile for the algorithm to run. If it takes an exceedingly long time, start with a shorter list). What specification did the algorithm end up with? (**5 points**)

Note: the resulting propensity scores from part (e) could then be used for matching, stratification, or weighting.

Question 2. This problem will use the panel dataset of Texas elementary schools used in class (*texas_elementary_panel_2004_2007.dta*) to estimate the effects of student mobility on school average performance on standardized tests. (**37 points**)

```
use https://github.com/spcorcor18/LPO-8852/raw/main/data/Texas_elementary_panel_2004_2007.dta
```

(a) The variable *cpemallp* is defined as the percentage of students in a school who were enrolled less than 83% of the school year (i.e., were not present 6 or more weeks at that school). Rename this variable *mobility*, report the overall mean and standard deviation for this variable, and produce a kernel density plot for this variable (use the `kdensity` command). Describe what this distribution looks like. (**3 points**)

(b) Declare this dataset to be a panel using `xtset`. Use the same cross-sectional unit and time dimension variables used in class. Use `xtsum` to get a set of descriptive statistics for *mobility*. Does it appear that school mobility is primarily a between-school phenomenon, or something that varies more within schools over time? Explain how you know, and explain in words how the standard deviations (overall, within, and between) are calculated. (**4 points**)

(c) Estimate a simple OLS regression of the average TAKS exam passing rate (*avgpassing*) on mobility (refer to the lecture notes and in-class example for the *avgpassing* variable). How are these variables related? Report your results and interpret your coefficient estimate in words. Is the coefficient statistically significant? Practically significant? (**4 points**)

(d) Should the coefficient estimated in part (c) be interpreted as the causal effect of student mobility on school performance? Briefly explain why or why not, with reference to potential outcomes. (**2 points**)

(e) Add the following explanatory variables to your regression in (c): percent black, white, Hispanic, Asian or Pacific Islander (API), Limited English Proficient (LEP), and economically disadvantaged. Also include year effects and a dummy variable for charter schools (*charter*, which may need to be encoded as numeric). How does the inclusion of these covariates affect your estimated coefficient on *mobility*? Is it still statistically significant? Does the change make sense to you (explain)? Finally, provide a written interpretation of the estimated coefficients for the three year dummies (2005, 2006 and 2007). (**4 points**)

(f) Should the coefficient estimated in (e) be interpreted as the causal effect of student mobility on school performance? Briefly explain why or why not. How might a regression model with school fixed effects improve upon the model in (e)? (**3 points**)

(g) Estimate the regression in (e) with school fixed effects. (Do this using both `xtreg` and `areg`). How does this approach affect the estimated coefficient on *mobility*? Is it still statistically significant? Does the change make sense to you? Provide an intuitive explanation of the finding. Were any explanatory variables dropped from the model (or are there any that should be dropped that didn't)? (**5 points**)

(h) What statistical assumptions must hold in order to interpret the coefficient estimate in (g) as causal? Are they likely to hold here? (**4 points**)

(i) For parts (i)-(j), keep only four variables—*campus, year, avgpassing* and *mobility*—and drop any cases where *avgpassing* or *mobility* are missing. Create a scatterplot showing the relationship between *avgpassing* and *mobility* and calculate the sample mean for these two variables. (**4 points**)

(j) Use `xtdata` to transform your data using the fixed effects (within) transformation. Create another scatterplot showing the relationship between *avgpassing* and *mobility* and calculate the sample mean for these two variables. How do these compare with part (i), and what is the basic difference between these two? (**4 points**)

Question 3. This problem will examine teacher effects on students' math and reading achievement using student-level data from a large urban school district. You will use methods that are closely related to those used in practice for estimating teacher "value-added." You can find the necessary data on Github under the name *LUSD4_5.dta*. All students in this database are in grades 4 and 5, and the test results are from 2005 and 2006. (**26 points**)

```
use https://github.com/spcorcor18/LPO-8852/raw/main/data/LUSD4_5.dta
```

Note, unlike Question 2, the regressions in this problem are not designed to estimate the causal effect of any particular input or intervention. Rather, we will be estimating fixed effects for individual teachers.

(a) First provide some descriptive information about the contents of this panel database. How many student observations are there in each grade and year? How many students appear in *both* grades 4 and 5 in this data? How many unique schools are in the data? How many unique teachers? The variable *school* is a unique school identifier, and *teacher* is the unique teacher identifier. Be clear in your Stata code how you answered these questions. (**3 points**)

(b) Estimate four separate regressions: by grade (4 and 5) and by subject (math and reading). The dependent variable will be either the standardized math score (*mathz*) or standardized reading score (*readz*). Both are z-scores with a mean of zero and standard deviation of 1 (standardized for the grade, subject, and year). Use the following explanatory variables: age, female, LEP, special ed, immigrant, economically disadvantaged, black, Hispanic, Asian, and a year effect (i.e., a dummy variable for 2006). At this point, do not include any other fixed effects. Provide a brief interpretation of your regression results. (**5 points**)

(c) Now estimate the same regressions as in part (b), but add as an additional control the lagged math score (in the math regressions) and the lagged reading score (in the reading regressions). These variables are already in the dataset as *mathz_1* and *readz_1*. How do the results change, and how should our interpretation of these results change, given the inclusion of lagged (prior grade) achievement? (**5 points**)

(d) Next, estimate the regressions in part (c) (with the lagged score), but this time use `xtreg` and include a fixed effect for the classroom teacher. (Instead of using `xtset`, you could include the options `fe` and `i(teacher)` in the `xtreg` command. This is equivalent to `xtset` without officially setting the panel variables). How should our interpretation of the coefficients change, if at all, given the inclusion of teacher fixed effects? (**5 points**)

(e) Teacher fixed effects—systematic variation in achievement after controlling for prior student achievement and other student characteristics—are often referred to as the teacher's "value added." How much of the variance in achievement is attributable to the teacher effect? (This is reported as the "rho" in the `xtreg` output). (**3 points**)

(f) Save the estimated teacher fixed effects using `predict`, as shown in class. Keep one observation per teacher (you can use `duplicates drop` to do this) and create a histogram of the estimated teacher fixed effects. What is the standard deviation of these teacher fixed effects? What is the difference between a teacher at the 75th percentile of the teacher effect distribution and a teacher at the 25th percentile? (**5 points**)

Question 4. This problem will use the same student-level data from a large urban school district to estimate the impact of having a same-race teacher on achievement. (That is, how a student performs when they share the same race/ethnicity as their teacher, relative to when they don't.) For a study that tackles this very question see Dee (2004). (**20 points**)

```
use https://github.com/spcorcor18/LPO-8852/raw/main/data/LUSD4_5.dta
```

(a) Create a variable called *same_race* that equals zero unless the student and teacher share the same race/ethnicity, in which case *same_race* should be coded as one. Use the white, black, Hispanic, and Asian categories, but not the "other" race category. In what percent of cases (i.e., student-year observations) are students assigned to a teacher of their same race/ethnicity? How does this rate of same race exposure vary by student race/ethnicity? (**4 points**)

(b) Estimate two regressions where the dependent variables are the math and reading z-scores, respectively, and *same_race* is the explanatory variable. Explain why the estimated coefficient on *same_race* should not be interpreted as causal. (**4 points**)

(c) Briefly explain how a regression model with *student fixed effects* might improve upon the regressions in part (b). What problem might this solve? (**2 points**)

(d) Use `xtset` to designate student as the panel variable, and year as the time dimension. Estimate the same regressions as in Question (#3) part (d) (with student covariates and lagged score), and use `xtreg, fe` to include student fixed effects. Also include *same_race* among your explanatory variables. Do **not** run the model separately by grade; you need multiple observations per student for this model to make sense. Describe what you find for the "same race" coefficient. Is it statistically significant? Practically significant? Can one make a strong claim for causal inference in this case? Explain why or why not. (**6 points**)

(e) Are there any explanatory variables that are dropped in the models in (d)? Are there any explanatory variables that should be dropped that weren't? What does the latter indicate to you? (**2 points**)

(f) Finally, use the command `xttrans` to describe the frequency of changes in exposure to a same-race teacher over time. Interpret the results of this command. (**2 points**)