Regression II
Vanderbilt University
Prof. Sean P. Corcoran

Due September 23, 2024
30 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 `tebalance 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., `teoverlap` 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`* `[`*`x1`*`], totry(`*`x2`*`) genpscore(`*`newvarname`*`)`

where *treat* is the treatment variable, *x1* is a list of covariates that you definitely want to include in your model, and *x2* is a list of additional variables you would like to try in the algorithm. *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.