Regression II
Vanderbilt University
Prof. Sean P. Corcoran

Due September 22, 2022
62 total points

Problem Set 2

Instructions: Answer the following questions in a Stata do-file, and submit your resulting log file via email to `sean.corcoran@vanderbilt.edu`, preferably as a .pdf or .txt file. Use your last name and problem set number as the filename (e.g., *Swift PS2.pdf*). The resulting log should include the questions below (commented), your commands, output, and written responses. Edit this file as appropriate, with any requested interpretations of your output. Graphical output can be submitted separately, preferably as a PDF file. Working together is encouraged, but all submitted work should be that of the individual student.

This problem set will use the National Education Longitudinal Study (NELS-88) data and matching methods to estimate the academic benefits, if any, to attending a Catholic high school. The variable definitions in this dataset should be self-explanatory, but if you have any questions, just ask. Some of these items provide flexibility on analytic decisions. Use your best judgment, and provide your rationale in your write-up.

You can read the data into Stata directly using this syntax:

```
use https://stats.idre.ucla.edu/stat/stata/examples/methods_matter/chapter12/catholic, clear
```

1. Provide some basic descriptive information about the students in this dataset. How many observations are there? What proportion attended a Catholic high school? What proportion graduated high school on time? What proportion entered post-secondary education after high school? What are the overall means and standard deviations for 12th grade math and reading scores, respectively? (**5 points**)

2. Now provide descriptive statistics (means, standard deviations) *separately* for Catholic and all other students. How do these populations of students differ, if at all? Include all outcomes and background variables that you think are relevant for a comparison of academic achievement between Catholic and all other students. (**5 points**)

    - Some background variables are categorical (e.g., *faminc, mothed8, mhowfar*). For the family income variable, create an alternative version in dollars, assigning values equal to the midpoints of each range of income (for example, $4000 for $3,000-$4,999). For a later step, you should also create a second categorical measure of family income in which 1=≤$19,999, 2=$20,000 to $34,999, and 3=$35,000 to $74,900. For parents' education and other categorical variables you can use Stata's factor variable notation throughout, or create separate dummies indicating, say, the parents' highest level of education (<HS grad, HS grad, some college, college+). You may wish to simplify some categorical variables; for example, for each parents' report of the highest level of education expected for the child, you could reduce the six categories to two (e.g., college or higher).

- After this step, standardize your reading and math scores to be mean zero, standard deviation one (z-scores). Do this for both 8th and 12th grade scores, and use these variables in later steps.

3. Estimate simple bivariate OLS regressions relating each of the following outcomes to Catholic high school attendance: math z-score, reading z-score, on-time high school graduation, enrolled in post-secondary education. Provide a brief explanation of what you find. Does there appear to be a Catholic high school effect? If so, is the difference statistically significant? Practically significant? Is the slope coefficient here a plausible estimate of the average treatment effect on the treated (ATT)? Why or why not? (**5 points**)

4. Provide a brief rationale for why one might prefer matching or a re-weighting approach to a regression model. What conditions must be satisfied for a treatment effect estimate based on matching or re-weighting to be convincing as a causal estimate? (**5 points**)

5. Use `teffects nnmatch` to estimate the ATE and ATT (ATET in Stata) of Catholic school attendance by exact matching on family income, using the 3-category version you created above. Use the same four outcomes from part (3). Summarize what you find. (This analysis is comparable to the subclassification example from Table 12.1 in Murnane & Willett—see the lecture notes. If you want to compare your estimates to theirs, do this step again with *math12*, the math score on its original scale). Note you can use Stata's factor variable notation in this command. (**5 points**)

6. Re-estimate the exact matching ATT from part (5) using the math z-score as the outcome. Following that, use the `tebalance summarize` command to check for balance on the following variables: your 3-category family income measure, your family income measure in dollars, and 8th grade math and reading z-scores. (You *can* conduct balance checks on variables that were not included in your original exact matching algorithm). How do the Catholic and public school students in the matched sample compare with respect to their distribution of these variables? In light of your findings here, how comfortable are you interpreting the estimates in part (5) as causal? (**5 points**)

7. In this part, you will develop a propensity score model to later estimate the ATT of Catholic school attendance. The first step of a propensity score analysis is to determine which confounding variables should be included in the estimation of the propensity score. You should be able to defend your choices based on theory and your understanding of the likely factors predicting selection into Catholic school *and* subsequent academic outcomes. Ultimately, however, the aim of a propensity score analysis is to create balance in the treated and untreated groups. (**15 points**)

   Your first task will be to settle on a propensity score model, iterating on included covariates and model specifications and checking balance until you are reasonably satisfied with the balance you have attained. More specifically:

- Use `psmatch2` or `teffects psmatch` to estimate your propensity score model. If you use `psmatch2`, do not request an ATE/ATT estimate in this step! Your specification of the propensity score model should not be influenced by what treatment effects they produce. If you use `teffects psmatch`, include the `quietly` prefix to suppress the output.
- Personally, I prefer `psmatch2` for formulating a propensity score model, since the balance and overlap diagnostics are easier. While you may use `psmatch2` at this step, you should use `teffects` later in estimating the treatment effect to ensure proper calculation of standard errors.
- At each new iteration of a propensity score model, use `pstest` (after `psmatch2`) or `tebalance summarize`, `box`, and `density` (after `teffects`) to check your balance.
- You may be able to achieve better balance by changing your model specification—e.g., including or omitting variables (depending on how relevant they are); entering variables as continuous, categorical, or as indicator measures; including interactions or nonlinear terms (e.g. a quadratic), etc.—and tinkering with the number of nearest neighbors or caliper. Experiment with different adjustments before deciding on a final approach.
- You do not need to provide all of your iterative work leading to your final model specification, just an explanation and the final code and balance statistics for the model you decide upon. (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.
- Note: `teffects overlap` does not show the distribution of propensity scores for the matched sample alone. To do this, use the work-around shown in the Stata matching commands handout.

8. Using your final propensity score model from part (7), answer the following:

   (a) Interpret your propensity score model. What types of students are more or less likely to be "treated" (i.e., attend Catholic high school)? You do not need to interpret the specific probit or logit coefficients since these do not have a natural interpretation. Hint: if you used `teffects psmatch`, you will not see the coefficient estimates from the propensity score model. In this case I recommend estimating your model in `psmatch2` or using `logit` or `probit` directly. (**5 points**)

   (b) Calculate a ATT estimate for each of the four outcome variables (math and reading test scores, high school graduation, and post-secondary enrollment). Provide a written interpretation of your treatment effect estimates. How do these differ from those you estimated using earlier methods (regression and exact matching)? (**5 points**)

(c) Keeping in mind that an untreated observation may be matched multiple times, how many *unique* students not enrolled in Catholic school were used as matches in your analysis? (**2 points**)

9. Using the same propensity score model from part (7), estimate the ATE and ATET using inverse probability weighting (`teffects ipw`). How do your results differ from those in part (8)? How many students not enrolled in Catholic school were used in this analysis? (**5 points**)