Regression II
Vanderbilt University
Prof. Sean P. Corcoran

Problem Set 3
88 total points

Problem Set 3

Instructions: Answer the following questions and submit your results via email to `sean.corcoran@vanderbilt.edu`. Use your name and problem set number as the filename. The Stata log for Q1 should include the instructions, your commands, and output. Graphical output may be submitted separately, preferably as a PDF file. Working together is encouraged, but all submitted work should be that of the individual student.

Question 1. This problem will use regression difference-in-differences to estimate the impact of a breakfast in the classroom (BIC) program on school meals program participation in New York City. BIC was not implemented under random assignment; rather, schools voluntarily adopted the program. We do, however, have data for these and other schools before and after adoption. (See Corcoran, Elbel, & Schwartz 2015 for details). (**51 points**)

(a) In Stata, open the panel dataset called *NYCbkfastlunch.dta* from Github:

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

This file consists of school-level data in which the rows are elementary or middle schools observed in year t (t=2005 to 2012). The outcome variables of interest are *bkfast_part* and *lunch_part*, which are average daily participation rates in the school breakfast and lunch programs. Provide some descriptive statistics for these two variables. On what scale are they measured? (**2 points**)

(b) Stata has a set of `xt` commands that make working with panel data easier. Use `xtset` *`panelvar timevar`* to declare the data as a panel. Which variable is the cross-sectional unit (*panelvar*) and which is the time dimension (*timevar*)? Is this a balanced panel? (Use `xtdescribe` to inspect the panel balance). How many schools are observed in all 8 years? (**2 points**)

(c) This dataset contains three types of schools: (1) schools that adopted BIC <u>in 2010</u> *(bic2010==1)*, (2) schools that <u>never</u> adopted BIC *(bicever==0)*, and (3) schools that adopted BIC in a year other than 2010 *(bic2010==0 & bicever==1)*. For parts (c)-(i) we will only work with types (1) and (2). Think of (1) as the treated group and (2) as the untreated group. We are excluding type (3) for now so that the "pre" and "post" periods are clearly defined.

Estimate a difference-in-differences regression that compares mean breakfast participation for the treated and untreated groups in two time periods: before 2010 and 2010-2012. (In other words, do a simple pre-post comparison for the two groups). Do the same for lunch participation. Interpret your results. Did the adoption of BIC have an impact on breakfast or lunch program participation? Is the effect statistically significant? Practically significant? What assumption(s) must be satisfied for this difference-in-differences to be considered a causal effect? (**6 points**)

(d) Compare the mean characteristics of treated and untreated schools. Look at the following: total enrollment, % ELL, % special education, % Asian, % black, % Hispanic, % female, % free lunch eligible (*free1*), % reduced price lunch eligible (*redu1*). How do schools that adopted BIC compare to those that didn't? (**4 points**)

(e) Now estimate the same regression models in part (c), but include the school covariates listed above. How do your estimates of the "BIC effect" change, if at all? (And how are these covariates related to meal participation?) (**4 points**)

(f) Repeat parts (c) and (e), but include a linear time trend in the regression. Center your time variable to be equal to 0 in 2010. How does this affect your impact estimates for BIC, if at all? What assumption(s) must be satisfied for this difference-in-differences to be considered a causal effect? (**4 points**)

(g) Repeat part (f) but use year dummies in the regression model in place of the linear time trend. How does this affect your impact estimates for BIC, if at all? Explain why one of the post-2010 year effects is not estimable. (**4 points**)

(h) Next, estimate a two-way fixed effects version of the models in (g). One way to do this is to include separate dummy variables for every school. A preferable approach is to use `xtreg` with the `fe` option. (Be sure you have used the `xtset` command in part b). How do your estimates compare to those in (g)? (**5 points**)

(i) One way to test the common trends assumption (for the pre-period) is to fit an "event study" regression, which estimates a treatment-comparison group difference in every year. To do this, estimate the model described in (h), but instead of the usual difference-in-differences variables, include an interaction of *bic2010* and the year dummies. Use the year prior to treatment as the reference year. How should you interpret the *bic2010*year* interaction effects? Do they provide any evidence that BIC schools were on a different trajectory prior to 2010? (**5 points**)

(j) Repeat part (i) but use the `eventdd` command described in class to obtain the event study regression results and graph. (You may have to install this using `ssc install eventdd`). This command also requires installation of another user-written command called `matsort`. (**5 points**).

(k) Some schools adopted BIC in years other than 2010. (Do a crosstabulation of *year* and *bicpost* to see this). Using the full dataset (school types 1-3), re-estimate your difference-in-differences model with school fixed effects (as in part h) and include a $BIC \times post$ interaction for BIC schools in years following their adoption. This *bicpost* variable has already been created for you. Try the models without and with covariates, and include year effects since the "post" period varies by school. How do your results compare to the earlier ones? (**5 points**)

(l) Repeat the event study regression and graph for the full sample used in part (j), using `eventdd`. (**5 points**)

Question 2. For these questions, refer to the recent article by Cellini and Turner (2019), "Gainfully Employed? Assessing the Employment and Earnings of For-Profit College Students Using Administrative Data." You can find the article here: `http://jhr.uwpress.org/content/54/2/342.abstract`. **(37 points)**

(a) Cellini and Turner use a generalized difference-in-differences regression model to estimate the causal effect of attending a for-profit certificate program on labor market outcomes. What specific outcome variables do they examine, and what dataset(s) do they use? **(4 points)**

(b) How is the "treatment" variable defined here and what are the possible "pre" and "post" years? How many potential pre and post years are there? **(4 points)**

(c) The authors use three different groups of "untreated" individuals as comparison groups. What were they, and what was their rationale for looking at each? Which comparison group is their "preferred" one, and why? **(4 points)**

(d) Equation (1) on page 350 shows their regression specification. Carefully explain what each term represents, and how the causal effect of attending a for-profit certificate program is being identified. Why is there not a main effect for "For-Profit" in the model? **(5 points)**

(e) Carefully explain the main assumption necessary to interpret the difference-in-differences estimate here as a causal effect. What evidence do the authors provide that this assumption holds for their three different comparison groups? **(5 points)**

(f) The paper's main results are reported in Table 3. Carefully interpret the coefficients reported in Panel B. What additional evidence does Figure 2 provide? **(10 points)**

(g) Finally, Figure 4 shows the distribution of earnings effects *by school* for public and for-profit institutions. Cellini and Turner describe these as the result of "single-difference" regressions. Briefly explain what they mean by this, and why these should not be interpreted as the causal effects of attending specific institutions. **(5 points)**