Regression II
Vanderbilt University
Prof. Sean P. Corcoran

Due Friday September 27, 2024
84 total points

Problem Set 4

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 (e.g., *Walz_PS4.pdf*). Working together is encouraged, but all submitted work should be that of the individual student.

Question 1. 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 rates on school average performance on standardized tests. (**38 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. (**3 points**)

(e) Add the following covariates 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 statistically significant? Does the change in the estimated coefficient make sense to you (explain)? Finally, provide a written interpretation of the estimated coefficients for the three year effects (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. (Show this using both `xtreg` and `areg`). Ensure the standard errors allow for clustering at the school level. How does this approach affect the estimated coefficient on *mobility*, if at all? Is it statistically significant? Does the change in the estimated coefficient 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 have been dropped that weren't)? If so, why? (**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? Explain your answer. (**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 2. 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 1, 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. Note also that the panel structure is classrooms (students nested within teacher) rather than cross-sectional units observed at multiple points in time.

(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 student's 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 can 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 only 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 estimated 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 3. 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.) One of the original studies that examined this question is 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 equal to 1. In doing this, 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 (#2) 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 have been 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**)