

Econometrics (Econ 3220) - Spring 2020
Problem Set 4

This assignment is due on 5 May. A PDF with your solutions must be uploaded to NYU Classes before 10:25am (the start of lecture) that day.

Question 1: You're interested in studying the effect of salary bonuses (*Bonus*, continuous, coded 0 to 100, representing a percentage of an employee's base salary) on performance (*Score*, continuous, coded 0 to 100, representing an employee's score on their annual performance review). But, you recognize that these concepts are endogenous. Bonuses are likely to increase productivity, but productive employees are the ones likely to have earned larger bonuses. So you decide to estimate the effect of *Bonus* on *Score* with 2SLS. As instruments, you choose *Salaried* (dummy variable, coded 1 if the employee is paid a fixed amount per year and 0 if they are paid by hour) and *Experience* (number of years that the employee has worked in the industry).

- (a) Write the equations for this 2SLS model.
- (b) Next, assume that your instruments have passed all of the relevant tests. Is your second stage estimate biased or unbiased? Why? Is your estimate consistent or inconsistent? Why? (this question is asking for an explanation for each, not a formal mathematical proof)
- (c) You realize that you have omitted *Educ* (years of education), which might be an important second-stage exogenous predictor of performance score. Write a new set of equations for your 2SLS model, but including education.
- (d) Suppose that you estimated the model from (a) and then found that $\text{corr}(Educ, Score) > 0$ and $\text{corr}(Educ, \widehat{Bonus}) > 0$. What would this imply about your second stage estimates in (a)?
- (e) Now suppose that you estimated the model from (c) and found that the t -statistics on your first stage instruments were no longer significant. Why might this have happened? Write the equation of the test you could run on your first stage model to see if you can still use these variables as instruments.
- (f) Looking back at parts (a) through (e), what does this imply about

the difficulty of estimating 2SLS relative to a single-stage OLS model and the care that must be taken? Your answer should be no more than one paragraph.

Question 2: A global network university has campuses in Country A, Country B, and Country C. You are a professor who is interested in student achievement in Econometrics, a course that is offered on all campuses. An event begins in Country A that you suspect will eventually affect all countries. You have the foresight to start collecting data on *Achievement* (0-100, continuous, representing scores on a standardized test) at the start of the semester (time period 1), before any of the campuses are affected. The campus in Country A is affected first (time period 2) and the campus in Country B is affected second (time period 3). You stop collecting data in time period 3, so you never observe the campus in Country C ever being affected. In each time period, you measure this same *Achievement* variable.

(a) Design a diff-in-diff model to measure the effect of the event on *Achievement*. Note that you have three time periods and three groups. For now, assume that Country C is not affected by the event at all and that the equation has a constant.

(b) Why do we need two separate dummy variables for the campuses? Why do we need two separate dummy variables for the time periods?

(c) What coefficient or coefficients represent the treatment effect in your model? If we had data on this, how would we interpret the coefficient(s)?

(d) You start analyzing your data, but then realize that in country B, Econometrics has multiple sections, some taught by Prof. X and some taught by Prof. Y. You suspect that Prof. X will adapt especially well to the situation, better than Prof. Y. Thankfully, you have a variable in your dataset for which professor was teaching each student. How would you need to modify your model to account for differences in *Achievement* between students learning from one professor vs. the other? Write an equation (one equation only) for this model. What assumption are testing when you estimate this model instead of the one from part (a)? How would you know if the assumption is violated?

Question 3: Refer back to Question 2 from Problem Set 3 (the question about compliance). The rule implemented by the university community has been in place longer than you hoped. You're bored, so

you decide to repeat the study a second time, collecting data from the same group who responded to the first study. Consider the original study to be time period 1 and the second study to be time period 2.

(a) Assuming that values of *Student*, *Child1*, *Child2*, and *ChildGT2* are the same in your second time period, could you implement a first differences model here? Why or why not?

(b) As suggested in some of the answers given in Problem Set 3, it is likely that there is a difference between subjective compliance (whether a person considers themselves to be compliant) and objective compliance (whether a person is meeting a set of standards imposed by the researcher). It could be that people think that they are complying, when they are not really doing so. This means that our dependent variable *Comply* has measurement error. Assuming that this measurement error is the same for all groups, what does this imply about our ability to infer a relationship between our independent variables and our dependent variable?

(c) Now suppose that you repeat a survey several times with the same panel (meaning the same group of respondents) and intend to run a unit Fixed Effects model. Come up with three variables you might be able to include in a model predicting compliance and write your new model equation in fixed effects form.

Question 4: You have panel data on 2000 individuals in country *X*. This country offers many programs in adult education, so schooling changes for many people over the panel. You are asked to estimate the equation

$$\ln(wage_{jt}) = \alpha_0 + \alpha_1 S_{jt} + \alpha_2 A_{jt} + v_{jt}$$

, where $wage_{jt}$ denotes the average hourly wage of person j in period t , S_{jt} is the level of schooling for person j in period t , and A_{jt} is the age of person j in period t .

(a) You will use a fixed-effect model to avoid omitted variable bias, because ability is unobserved. How does such a model avoid the omitted variable problem? Can you estimate the coefficient on age?

(b) However, you are now informed that there is considerable measurement error in S_{jt} . Assuming classical measurement error, why

is this a bigger problem with fixed-effect estimation than with *OLS* estimation?

(c) You are presented with a variable Z_{jt} that you want to use as an IV. What conditions have to hold for it to be a good instrument? Assuming that it is a good instrument, how would you use it to get a consistent estimate of α_1 ?