14.74 Foundations of Development Policy
Spring 2009

14.74 Problem Set #2

Due Wednesday, Mar. 4, 2008 by 5 pm

Keep your answers short and to the point. Please submit homework in the format of the solutions. There should be 2 parts 1) The write up your answers including *relevant* parts of the log file, 2) The *final, working* version of your do file (one comprehensive program).

1. **Estimating the Returns to Schooling: the INPRES School Construction Program in Indonesia**

   *This problem will lead you through the instrumental variables techniques introduced in lecture to estimate the returns to schooling (in particular, the effect of **years of education** on **log hourly wages**). This exercise uses the Stata data set named supa.dta, which is a subset of the data used by Professor Duflo in her paper listed in the syllabus, "Schooling and labor market consequences of school construction in Indonesia: Evidence from an unusual policy experiment." Because this data set is 2.5Mb while the Stata default memory setting is 1Mb, begin your Stata session by typing "set mem 20m"; you will also need to type the command "set matsize 800"). Spend some time understanding the notation and data well, and the methods are explained step by step.*

   a. Recall the wage-setting equation from the education model discussed in class:

   $\ln w_i = a_i + bS_i$

   Interpret the parameter b in this equation. Why might the coefficient on an OLS regression of log wages on years of schooling not be a good estimate of b?

   In the remainder of this problem, you will be asked to use a number of instrumental variable techniques to try to get an unbiased estimate of b. We want to estimate the effect of years of schooling on log wages, but years of schooling is an endogenous variable (we call this the "endogenous regressor"). The idea behind instrumental variables is to use other variables, called instruments, which generate exogenous variation in years of schooling. In this problem, we can exploit the exogenous variation in years of schooling generated by the INPRES program.

   b. Wald Estimate

   i. First, calculate the difference-in-differences (DD) estimate of the effect of INPRES on **years of schooling**; this is the **first stage**. Use cross-cohort (young versus old) and cross-regional (high-intensity versus low-intensity) variation. Interpret the estimate. Under what assumption is this an unbiased estimate of the effect of INPRES on education?

   ii. Next, calculate the DD estimate of the effect of INPRES on **log hourly wages**, using the same sources of variation as in (i); this is the **reduced form**. Interpret the estimate. Under what assumption is this an unbiased estimate of

the effect of INPRES on wages? How might we test this assumption, given the available data?

iii. The Wald estimate for b in the wage-setting equation is the DD estimate from the reduced form divided by the DD estimate from the first stage. What is the additional assumption for unbiasedness of the Wald estimate? How might we test this assumption, given the available data?

iv. Calculate and interpret the Wald estimate of the returns to schooling.

c. **Indirect least squares (ILS)**

In (b), we used only some of the available variation in the INPRES program (in particular, we used only 2 age and 2 regional groups). Now we will use all of the regional variation. Before, you were calculating means for high- and low-intensity regions, young vs. old; in this question, you are calculating means for the young and old, region-by-region.

First, generate a new data set that contains the mean schooling of the young, mean schooling of the old, mean wage of the young, mean wage of the old, mean number of children in 1971, and mean program intensity, *by region*. (Hint: Your first line of Stata code should say something like `generate educ_yng = yeduc if young == 1`. Then, use the *collapse* command .)

Next, generate a variable called `educ_dif` which equals young mean education minus old mean education. Likewise, generate a variable called `wage_dif` which equals young mean log wage minus old mean log wage.

i. The **first stage** is: $S_{Yj} - S_{Oj} = \alpha P_j + v_j$ (where *S denotes years of schooling; P denotes program intensity; Y and O index young and old; and j indexes the region*). The equation is written as though the only right-hand-side variable is program intensity; however, the v_j term can be understood to include the error term, the constant term, and any control variables (i.e., other right-hand-side variables). Note that the left-hand side of the equation is just `educ_dif`. Estimate α by regressing `educ_dif` on program intensity, including the number of children in 1971 as a control variable; report the estimate.

ii. The **reduced form** is given by equation (2): $y_{Yj} - y_{Oj} = \gamma P_j + \varepsilon_j$ (where *y denotes log hourly wage*). Note that the left-hand side of the equation is just `wage_dif`. Estimate γ by regressing `wage_dif` on program intensity, including the number of children in 1971 as a control variable; report the estimate.

iii. Recall the wage-setting equation $y_i = a_i + bS_i$, where y is log hourly wage. Taking means by region and cohort, we have that $y_{Yj} = a_{Yj} + bS_{Yj}$ and $y_{Oj} = a_{Oj} + bS_{Oj}$. Subtracting the latter from the former, we get $y_{Yj} - y_{Oj} = a_{Yj} - a_{Oj} + b(S_{Yj} - S_{Oj})$. Use this equation, along with the first-stage and reduced-form equations, to argue algebraically that $b = \gamma/\alpha$.

iv. Calculate the indirect least squares estimate of b in this way, report and interpret this estimate. How does it compare to the Wald estimate? Which do you prefer? Explain.

v. Indirect Least Squares uses the program as an instrument to generate exogenous variation in years of schooling. Essentially, it predicts years of schooling based on the exposure to the program; it then estimates returns to schooling using only the variation in predicted schooling. Now let's use a

second technique to estimate the return to schooling using Indirect Least Squares. Run the first stage regression from part (i) above and use the *predict* command to save predicted values of the variable `educ_dif` as a new variable that you can call `edif_pred`. Next, run a regression of `wage_dif` on the predicted values `edif_pred`, again including the control for number of children in 1971. Report your new estimate of b and check to make sure the coefficient on the predicted values `edif_pred` is indeed the same as what you calculated in part (iv)?

d. **Two-stage least squares (2SLS)**

In (c), we did not use fully the variation by cohort; instead, we aggregated cohorts into two groups, young and old. Now we will use all of the cohort variation as well. In (c) you modified the data set; for this part, **return to using the original data set**.

Note that we could repeat the indirect least squares exercise cohort-by-cohort; that is, for each cohort k (= year of birth), we could run the regressions:
$S_{kj} - S_{50j} = \alpha_k P_j + v_{kj}$ and
$y_{kj} - y_{50j} = \gamma_k P_j + \varepsilon_{kj}$,
where 50 is the oldest cohort (year of birth = 1950), and j indexes the region.

i. You are **not** being asked to run the above regressions. However, if you were to run those regressions, how would you interpret the estimates for α_k and γ_k?

2SLS combines these different ways of estimating b in an optimal fashion (in the sense of minimizing the variance of the estimator). Like all instrumental variable estimators, 2SLS uses a set of variables, called instruments, to generate exogenous variation in the endogenous regressor. In our case, years of schooling is the endogenous regressor, so we use "exposure to the INPRES program" as instruments that will generate exogenous variation in years of schooling.

This data set has individuals who were age 2 to 24 in 1974; those who were age 2 to 12 in 1974 (i.e., `yob` between 1962 and 1972) are young enough to have been exposed to the program. Hence, we have 11 instruments available (1 program * 11 cohorts).

ii. For each cohort exposed to the program, generate a dummy variable equal to 1 if the individual was born in that year and 0 otherwise (e.g., `generate d62 = (yob==62)` , etc., up to `d72`). Now, for each exposed cohort, generate a variable that equals the cohort dummy multiplied by program intensity (e.g., `generate z62 = d62*prog_int` , etc., up to `z72`). These eleven variables (`z62` through `z72`) are the 11 instruments. **State clearly** the requirements for these to be good instruments? **Explain** briefly why you might think that these variables satisfy those two requirements.

For control variables in our regressions, we will use number of children in 1971, year-of-birth dummies, and year-of-birth dummies interacted with number of children in 1971 ("interacted with" means "multiplied by"). Year-of-birth dummies mean dummy variables, one for each possible year of birth, which take the value 1 if the individual is born in that year, and 0 otherwise. As you might have noticed, these are a lot of control variables (1 variable for number of children + 23 year-of-birth dummies + 23 interaction terms = 47 control variables!). Fortunately, Stata makes it easy to handle dummies and interaction terms based on an existing variable (in this case, year of birth `yob`). Within a Stata command, you can use the formulation `i.variable` to refer to a set of dummy variables based on the values taken by the variable named `variable`; you can use the formulation `i.variable1*variable2` to refer to the the set of dummy variables based on `variable`, the variable `variable2`, ***and*** the interaction terms between the two. When using the `i.` formulation in a command, you must precede the command with `xi:` . Thus, to regress the variable `y` on `x`, number of children in 1971, year-of-birth dummies, and year-of-birth dummies interacted with number of children, you would use the command

```
xi: regress y x i.YOB*ch71
```

iii. To calculate the 2SLS estimate of b,

1. First, use the instruments and control variables to predict years of education (i.e., run an OLS regression of years of education on the instruments and control variables, and form the predicted values of education from this regression).
2. Regress log hourly wages on this predicted value of years of education and the control variables. The coefficient on the predicted years of education is the 2SLS estimate of b.
   Compute and interpret the 2SLS estimate of b.
3. Can we trust this 2SLS estimate if the INPRES program improves school quality? Why or why not?

2. **Essay Question**

Read the paper "Peer Effects, Pupil-Teacher Ratios, and Teacher Incentives: Evidence from a Randomized Evaluation in Kenya" by Duflo, Dupas, and Kremer, which has been posted to the course website under "Problem Set 2".

- Succinctly but clearly describe the experimental design. What were the different treatments offered?
- Outline the specific questions/hypotheses that the paper tests. To guide yourself, look through Figures 4, 5, 6, and Tables 3, 4, 5, 6, 7, 8, and identify the hypothesis that the reported results are testing. State the hypotheses as questions. It is okay to bullet your response to this section. For example, I would begin with these questions from Table 3:
  - Does reducing class size via the Extra Teacher Program (ETP) improve teacher attendance/effort? (Table 3)
  - Does tracking students improve teacher attendance/effort? (Table 3)
  - Do contract teachers in tracking schools have higher attendance/effort than non-contract teachers? (Table 3)

    - etc.

Then, after reading the paper on tracking, read the paper "Many Children Left Behind: Textbooks and Test Scores in Kenya" by Glewwe, Kremer, and Moulin, which has also been posted to the course website under Problem Set 2.

- Briefly summarize the experimental design.
- Think about how you could modify the experimental design to answer more questions than those addressed directly in the paper. Keep the experiment related to textbook use, but think of other questions you may have about textbooks, or other programs that might interact with the effectiveness of textbooks. For example: what types of textbooks are effective? who benefits from textbooks? how should textbooks be distributed? what secondary programs may increase the efficacy of textbooks? etc. Be specific about what additional hypotheses you would like to test, beyond those covered in the paper.
- What control and treatment groups would you need to test your new hypotheses? What data would you like to collect? What comparisons would you make in your data?