

Faculty of Economics & Politics  
Part 2A Paper 3: Microeconometrics and Panel Data  
Class 1: Instrumental Variables

Please note: All data are located in J:/STUDENTS/MW217/MICROECONOM.  
All data files have the `.raw` extension. Variable descriptions etc are in  
accompanying `.des` files.

1. Consider the simple regression model

$$y = \beta_0 + \beta_1 x + u$$

and let  $z$  be a binary instrumental variable for  $x$ . Show that the IV estimator  $\beta$ , can be written as

$$\hat{\beta}_1 = (\bar{y}_1 - \bar{y}_0) / (\bar{x}_1 - \bar{x}_0)$$

where  $\bar{y}_0$  and  $\bar{x}_0$  are the sample averages for  $z = 0$ , and  $\bar{y}_1$  and  $\bar{x}_1$  are the sample averages for  $z = 1$ .

2. Suppose you want to test whether girls who attend a girls' high school do better in math than girls who attend coed schools. You have a random sample of senior high school girls from a state in the United States, and *score* is the score on a standardized math test. Let *girlhs* be a dummy variable indicating whether a student attends a girls, high school.
  - (i) What other factors would you control for in the equation? (You should be able to reasonably collect data on these factors.)
  - (ii) Write an equation relating *score* to *girlhs* and the other factors you listed in part (i).
  - (iii) Suppose that parental support and motivation are unmeasured factors in the error term in part (ii). Are these likely to be correlated with *girlhs*? Explain.
  - (iv) Discuss the assumptions needed for the number of girls' high schools within a twenty-mile radius of a girl's home to be a valid instrumental variable (IV) for *girlhs*.

### 3. The Returns to Schooling I

Use the data in WAGE2.RAW for this exercise.

- (i) In seeking to estimate the returns to education based upon a simple bivariate regression model,  $\log(\text{wage}) = \beta_0 + \beta_1 \text{educ} + u$ , a number of analysts have used family variables such as *sibs* (*number of siblings*), as an instrument for *educ*. Using the WAGE2.RAW data the IV estimate of the return to education is .122. To convince yourself that using *sibs* as an IV for *educ* is not the same as just plugging *sibs* in for *educ* and running an OLS regression, run the regression of  $\log(\text{wage})$  on *sibs* and explain your findings.
- (ii) The variable *brthord* is birth order (*brthord* is one for a first-born child, two for a second-born child, and so on). Explain why *educ* and *brthord* might be negatively correlated. Regress *educ* on *brthord* to determine whether there is a statistically significant negative correlation.
- (iii) Now use *brthord* as an IV for *educ*. Report and interpret the results.
- (iv) Now, suppose that we include number of siblings as an explanatory variable in the wage equation; this controls for family background, to some extent:

$$\log(\text{wage}) = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{sibs} + u.$$

Suppose that we want to use *brthord* as an IV for *educ*, assuming that *sibs* is exogenous. The reduced form for *educ* is

$$\text{educ} = \pi_0 + \pi_1 \text{sibs} + \pi_2 \text{brthord} + v.$$

State and test the identification assumption.

- (v) Estimate the equation from part (iv) using *brthord* as an IV for *educ* (and *sibs* as its own IV). Comment on the standard errors for  $\hat{\beta}_{\text{educ}}$  and  $\hat{\beta}_{\text{sibs}}$ .
- (vi) Using the fitted values from part (iv),  $\hat{\text{educ}}$ , compute the correlation between  $\hat{\text{educ}}$  and *sibs*. Use this result to explain your findings from part (v).

#### 4. *The Determinants of Fertility*

The data in FERTIL2.RAW includes, for women in Botswana during 1988, information on number of children, years of education, age, and religious and economic status variables.

- (i) Estimate the model

$$children = \beta_0 + \beta_1 educ + \beta_2 age + \beta_3 age^2 + u$$

by OLS and interpret the estimates. In particular, holding age fixed, what is the estimated effect of another year of education on fertility? If 100 women receive another year of education, how many fewer children are they expected to have?

- (ii) *Frsthalf* is a dummy variable equal to one if the woman was born during the first six months of the year. Assuming that *frsthalf* is uncorrelated with the error term from part (i), show that *frsthalf* is a reasonable IV candidate for *educ*. (Hint: You need to do a regression.)
- (iii) Estimate the model from part (i) by using *frsthalf* as an IV for *educ*. Compare the estimated effect of education with the OLS estimate from part (i).
- (iv) Add the binary variables *electric*, *tv*, and *bicycle* to the model and assume these are exogenous. Estimate the equation by OLS and 2SLS and compare the estimated coefficients on *educ*. Interpret the coefficient on *tv* and explain why television ownership has a negative effect on fertility.

#### 5. *The Returns to Schooling II*

Use the data in CARD.RAW for this exercise. The data is comprised of observations on 3010 men taken from the US National Longitudinal Survey of Young Men<sup>1</sup>. This is a panel dataset, tracking a group of individuals since 1966 when they were aged between 14-24. Here we use data from 1976.

---

<sup>1</sup>Note: it is likely that you will encounter a problem using all 3010 observations in MicroFit. If this is so you should simply discard the last 10 observations by setting the sample 1 3000.

## Background

Does the observed positive correlation between education and higher earnings reflect a causal effect of schooling; or do individuals with greater ability, and therefore greater earnings potential, self-select into more education? This is the question we address below.

We write a standard human capital earnings function as

$$w_i = \alpha + \beta_1 S_i + \beta_2 E_i + \beta_3 E_i^2 + \varepsilon_i, \quad (1)$$

where  $w_i$  denotes the log of individual earnings,  $S_i$  denotes years of schooling and  $E_i$  years of experience. Note that this variable is often measured as  $age_i - S_i - 6$ , assuming a start school date of age 6. We then augment this equation with other variables such as regional dummies.

There are a number of reasons why  $S_i$  is likely to be correlated with  $\varepsilon_i$ . We have examined 2 of these in lectures:

- if ability is unobserved, and ability is correlated with earnings, then more able students select more education, the OLS estimator of  $\beta_2$  will exhibit upward bias

(Note that if there is a problem of endogeneity with  $S_i$  then this will also generate similar problems for the experience variables.)

- if there is measurement error in the schooling variable then this may generate *downward* bias in the OLS estimator of  $\beta_2$ .

- (i) Augment (1) with three dummy variables: whether the individual was black (*black*), lived in a metropolitan area (*smsa*), and lived in the south (*south*). Estimate this model and comment on your results.
- (ii) If schooling is endogenous we will need to find instruments for  $S_i$ ,  $E_i$ , and  $E_i^2$ . For  $S_i$  use the dummy variable *nearc4*. Why might this provide a IV for  $S_i$ . What instruments are available for  $E_i$  and  $E_i^2$ ? By locating the necessary instruments estimate the parameters in (1) using the IV estimator. Interpret your results.
- (iii) The equation estimated in Example 15.4 (see the Wooldridge text) can be written as

$$w_i = \alpha + \beta_1 S_i + \beta_2 E_i + \beta_3 E_i^2 + \dots + \varepsilon_i,$$

where the other explanatory variables are: a black dummy variable, dummy variables for living in a SMSA and living in the south, a full set of regional dummy variables, and a SMSA dummy for where the man was living in 1966. In order for IV to be consistent, the IV for *educ*, *nearc4*, must be uncorrelated with *u*. Could *nearc4* be correlated with things in the error term, such as unobserved ability? Explain.

- (iv) For a subsample of the men in the data set, an IQ score is available. Regress IQ on *nearc4* to check whether average IQ scores vary by whether the man grew up near a four-year college. What do you conclude?
- (iv) Now regress IQ on *nearc4*, *smsa66*, and the 1966 regional dummy variables *reg662*, ..., *reg669*. Are IQ and *nearc4* related after the geographic dummy variables have been partialled out? Reconcile this with your findings from part (ii).
- (vi) From parts (ii) and (iii), what do you conclude about the importance of controlling for *smsa66* and the 1966 regional dummies in the  $\log(\textit{wage})$  equation?

Melvyn Weeks  
Lent 2008