

Problem Set 3: Instrumental Variables

1. (Wooldridge 15.1) Consider a simple model to estimate the effect of personal computer (PC) ownership on college grade point average for graduating seniors at a large public university:

$$GPA = \beta_0 + \beta_1 PC + u,$$

where PC is a binary variable indicating PC ownership.

- (a) Why might PC ownership be correlated with u ?
 - (b) Explain why PC is likely to be related to parents' annual income. Does this mean parental income is a good instrument for PC ? Why or why not?
 - (c) Suppose that, four years ago, the university gave grants to buy computers to roughly one-half of the incoming students, and the students who received grants were randomly chosen. Carefully explain how you would use this information to construct an instrumental variable for PC .
2. (Wooldridge 15.7) The following is a simple model to measure the effect of a school choice program on standardized test performance [see Rouse (1998) for motivation]:

$$score = \beta_0 + \beta_1 choice + \beta_2 faminc + u_1,$$

where $score$ is the score on a statewide test, $choice$ is a binary variable indicating whether a student attended a choice school in the last year, and $faminc$ is family income. The instrument for $choice$ is $grant$, the dollar amount granted to students to use for tuition at choice schools. The grant amount differed by family income level, which is why we control for $faminc$ in the equation.

- (a) Even with $faminc$ in the equation, why might $choice$ be correlated with u_1 ?
 - (b) If within each income class, the grant amounts were assigned randomly, is $grant$ uncorrelated with u_1 ?
 - (c) Write the reduced form equation for $choice$ (that is, the equation of $choice$ on all exogenous variables). What is needed for $grant$ to be partially correlated with $choice$?
 - (d) Write the reduced form equation for $score$. Explain why this is useful. (Hint: How do you interpret the coefficient on $grant$?)
3. Use the data file `Smoke.dat` for this exercise.

- (a) A model to estimate the effects of smoking on annual income (possibly through lost work days due to illness, or productivity effects), is

$$\log(income) = \beta_0 + \beta_1 cigs + \beta_2 educ + \beta_3 age + \beta_4 age^2 + u_1$$

where $cigs$ is number of cigarettes smoked per day, on average. How do you interpret β_1 ?

- (b) Under what assumption is the income equation from part (a) identified?
- (c) Estimate the income equation by OLS and discuss the estimate of β_1 .
- (d) Estimate the following reduced form for $cigs$:

$$cigs = \gamma_0 + \gamma_1 educ + \gamma_2 age + \gamma_3 age^2 + \gamma_4 \log(cigprice) + \gamma_5 restaurn + u_2$$

where $cigprice$ is the price of a pack of cigarettes (in cents), and $restaurn$ is a binary variable equal to unity if the person lives in a state with restaurant smoking restrictions. Are $\log(cigprice)$ and $restaurn$ significant in the reduced form?

- (e) Estimate the income equation by 2SLS. Compare the resulting estimate of β_1 with the OLS estimate.
 - (f) Do you think that cigarette prices and restaurant smoking restrictions are exogenous in the income equation?
4. Use the data file `mus06data.dta`. for this exercise. We analyze medical expenditures of individuals 65 years and older who qualify for health care under the U.S. Medicare program. The equation to be estimated has the dependent variable *ldrugexp*, the log of total out-of-pocket expenditures on prescribed medications. The regressors are an indicator for whether the individual holds either employer or union-sponsored health insurance (*hi_empunion*), number of chronic conditions (*totchr*), and four sociodemographic variables: age in years (*age*), indicators for whether female (*female*) and whether black or Hispanic (*blhisp*), and the natural logarithm of annual household income in thousand of dollars (*linc*).
- (a) Why is the health insurance variable *hi_empunion* likely to be endogenous?
 - (b) According to the OLS estimates, what is the impact of being insured through a union on the level of medical expenditures?
 - (c) Provide arguments to defend that the ratio of an individual's social security income to the individual's income from all sources (*ssiratio*) would be a good instrument for *hi_empunion* (Hint: High values in *ssiratio* indicate a significant income constraint.)
 - (d) Estimate the model by simple IV using *ssiratio* as instrument. Control for heteroskedasticity errors and provide output that additionally reports results from the first-stage regression. Is the coefficient of *hi_empunion* plausible? Explain.
 - (e) The variable *multlc* indicates whether the firm is a large operator with multiple locations. This variable is intended to capture whether the individual has access to supplementary insurance through the employer. Use both *ssiratio* and *multlc* as instruments for *hi_empunion*. Again control for heteroskedasticity. Give the IV estimates obtained in part (e) and the 2SLS estimates of the coefficients and their standard errors in a unique table. Comment the results of the table.
 - (f) Perform a Hausman test of endogeneity (after 2SLS estimation). What do you conclude?
 - (g) Test the validity of overidentifying instruments in the previous over-identified model.

References

- ROUSE, C. (1998): "Private school vouchers and student achievement: An evaluation of the Milwaukee parental choice program," *The Quarterly Journal of Economics*, 113(2), 553–602.