International School of Economics at TSU

Econometrics 2

Lasha Chochua

Problem Set 8

Instructions: You are encouraged to solve the problems before the recitation. Additionally, you are encouraged to work in groups. It is **not mandatory** to submit solutions unless stated otherwise. However, if you would like to share your solution, I would be happy to review it.

Problem 1: Consider a model for the health of an individual:

$$health = \beta_0 + \beta_1 age + \beta_2 weight + \beta_3 height + \beta_4 male + \beta_5 work + \beta_6 exercise + u_1 \quad (1)$$

where *health* is some quantitative measure of the person's *health*; *age*, *weight*, *height*, and *male* are self-explanatory; *work* is weekly hours worked; and *exercise* is the hours of exercise per week.

a. Why might you be concerned about *exercise* being correlated with the error term u_1?

Solution

Unobserved factors that tend to make an individual healthier also tend to make that person exercise more. For example, if *health* is a cardiovascular measure, people with a history of heart problems are probably less likely to exercise. Unobserved factors such as prior health or family history are contained in u_1, and so we are worried about correlation between *exercise* and u_1. Self-selection into exercising predicts that the benefits of exercising will be, on average, overestimated. Ideally, the amount of exercise could be randomized across a sample of people, but this can be difficult.

b. Suppose you can collect data on two additional variables, *disthome* and *distwork*, the distances from home and from work to the nearest health club or gym. Discuss whether these are likely to be uncorrelated with u_1.

Solution

If people do *not* systematically choose the location of their homes and jobs relative to health clubs based on unobserved health characteristics, then it is reasonable to believe that

disthome and *distwork* are uncorrelated with u_1. But the location of health clubs is not necessarily exogenous. Clubs may tend to be built near neighborhoods where residents have higher income and wealth, on average, and these factors can certainly affect overall health. It may make sense to choose residents from neighborhoods with very similar characteristics but where one neighborhood is located near a health club.

c. Now assume that *disthome* and *distwork* are in fact uncorrelated with u_1, as are all variables in equation (1) with the exception of *exercise*. Write down the reduced form for *exercise*, and state the conditions under which the parameters of equation (1) are identified.

Solution

The reduced form for *exercise* is

$$exercise = \pi_0 + \pi_1 age + \pi_2 weight + \pi_3 height + \pi_4 male + \pi_5 work + \pi_6 disthome + \pi_7 distwork + u_1$$

For identification we need at least one of π_6 and π_7 to be different from zero. This assumption can fail if the amount that people exercise is not systematically related to distances to the nearest health club.

Problem 2: Consider the following model to estimate the effects of several variables, including cigarette smoking, on the weight of newborns:

$$\log(bwght) = \beta_0 + \beta_1 male + \beta_2 parity + \beta_3 \log(faminc) + \beta_4 packs + u, \tag{2}$$

where *male* is a binary indicator equal to one if the child is male, *parity* is the birth order of this child, *faminc* is family income, and *packs* is the average number of packs of cigarettes smoked per day during pregnancy.

a. Why might you expect *packs* to be correlated with u?

Solution

There may be unobserved health factors correlated with smoking behavior that affect infant birth weight. For example, women who smoke during pregnancy may, on average, drink more coffee or alcohol, or eat less nutritious meals.

b. Suppose that you have data on average cigarette price in each woman's state of residence. Discuss whether this information is likely to satisfy the properties of a good instrumental variable for *packs*.

Solution

Basic economics says that *packs* should be negatively correlated with cigarette price, although the correlation might be small (especially because price is aggregated at the state level). At first glance it seems that cigarette price should be exogenous in equation (5.54), but we must be a little careful. One component of cigarette price is the state tax on cigarettes. States that have lower taxes on cigarettes may also have lower quality of health care, on average. Quality of health care is in u, and so maybe cigarette price fails the exogeneity requirement for an IV.

c. Use the data in BWGHT.RAW to estimate equation (2). First, use OLS. Then, use 2SLS, where *cigprice* is an instrument for *packs*. Discuss any important differences in the OLS and 2SLS estimates.

Solution

See Jupyter Notebook for regression results.

The difference between OLS and IV in the estimated effect of *packs* on *bwght* is huge. With the OLS estimate, one more pack of cigarettes is estimated to reduce *bwght* by about 8.4%, and is statistically significant. The IV estimate has the opposite sign, is huge in magnitude, and is not statistically significant. The sign and size of the smoking effect are not realistic.

d. Estimate the reduced form for *packs*. What do you conclude about identification of equation (2) using *cigprice* as an instrument for *packs*? What bearing does this conclusion have on your answer from part c?

Solution

See Jupyter Notebook for regression results.

The reduced form estimates show that *cigprice* does not significantly affect *packs*. In fact, the coefficient on *cigprice* does not have the sign we expect. Thus, *cigprice* fails as an IV for *packs* because *cigprice* is not partially correlated with *packs* with a sensible sign for the correlation. This is separate from the problem that *cigprice* may not truly be exogenous in

the
birth weight equation.

Problem 3 Problem 3 A researcher is interested in estimating the variance of the error term in Equation (1) in the slides.

a. Suppose she uses the estimator from the second-stage regression of TSLS:

$$\hat{\sigma}_a^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0^{TSLS} - \hat{\beta}_1^{TSLS}\hat{X}_i\right)^2,$$

where $\hat{X}_i$ is the fitted value from the first-stage regression. Is this estimator consistent? (For the purposes of this question, suppose that the sample is very large and the TSLS estimators are essentially identical to β_0 and β_1.)

Solution

Estimator:

$$\hat{\sigma}_a^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0^{TSLS} - \hat{\beta}_1^{TSLS}\hat{X}_i\right)^2$$

Note that:

- $\hat{X}_i$ is the **fitted value** from the first-stage regression.
- Define $\hat{u}_i = Y_i - \hat{\beta}_0^{TSLS} - \hat{\beta}_1^{TSLS}X_i$ — the second-stage residual.
- Then:

$$\hat{\sigma}_a^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(\hat{u}_i - \hat{\beta}_1^{TSLS}(\hat{X}_i - X_i)\right)^2$$

Assuming $\hat{\beta}_1^{TSLS} \to_p \beta_1$, write:

$$\hat{\sigma}_a^2 \approx \frac{1}{n}\sum_{i=1}^{n}\left(u_i - \beta_1(\hat{X}_i - X_i)\right)^2$$

Now expand the square:

$$= \frac{1}{n}\sum_{i=1}^{n} u_i^2 + \beta_1^2(\hat{X}_i - X_i)^2 - 2\beta_1 u_i(\hat{X}_i - X_i)$$

- First term $\to \sigma_u^2$
- Second and third terms converge to non-zero quantities due to approximation error from using $\hat{X}_i$ instead of X_i.

Conclusion:

$\hat{\sigma}_a^2$ is **not consistent** because it includes extra variation from $\hat{X}_i \neq X_i$.

b. Is

$$\hat{\sigma}_b^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0^{TSLS} - \hat{\beta}_1^{TSLS} X_i \right)^2$$

consistent?

Solution

The estimator

$$\hat{\sigma}_b^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0^{TSLS} - \hat{\beta}_1^{TSLS} X_i \right)^2$$

is consistent.

Using the same logic as in (a), we can write:

$$\hat{\sigma}_b^2 \approx \frac{1}{n} \sum_{i=1}^{n} u_i^2,$$

and this estimator converges in probability to σ_u^2.