Problem Set: Instrumental Variables

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

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

where `PC` *is a binary variable indicating* `PC` *ownership.*

(i) Why might `PC` *ownership be correlated with* u*?*

(ii) Explain why `PC` *is likely to be related to parents' annual income. Does this mean parental income is a good instrumental variable (IV) for* `PC`*? Why or why not?*

(iii) 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`*.*

Problem 2. *Evans and Schwab (1995) studied the effects of attending a Catholic high school on the probability of attending college. For concreteness, let* `college` *be a binary variable equal to unity if a student attends college, and zero otherwise. Let* `CathHS` *be a binary variable equal to one if the student attends a Catholic high school. A linear probability model is*

$$college = \beta_0 + \beta_1 CathHS + other\ factors + u,$$

where the other factors include gender, race, family income, and parental education.

(i) Why might `CathHS` *be correlated with* u*?*

(ii) Evans and Schwab have data on a standardized test score taken when each student was a sophomore. What can be done with this variable to improve the ceteris paribus estimate of attending a Catholic high school?

(iii) Let `CathRel` *be a binary variable equal to one if the student is Catholic. Discuss the two requirements needed for this to be a valid Instrumental Variable (IV) for* `CathHS` *in the preceding equation. Which of these can be tested?*

Problem 3. *Use the data in WAGE2 (Description) for this exercise.*

(i) *Estimate the following using OLS. What could be a problem?*

$$\log(wage) = \beta_0 + \beta_1 educ + u.$$

(ii) *Use the variable* **`sibs`** *(number of siblings) as an instrument for* **`educ`**. *Compare the results.*

(iii) *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* $\log(wage)$ *on* **`sibs`** *and explain your findings.*

(iv) *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.*

(v) *Use* **`brthord`** *as an IV for* **`educ`**. *Report and interpret the results.*

(vi) *Now, suppose that we include number of siblings as an explanatory variable in the wage equation; this controls for family background, to some extent:*

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

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

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

State and test the identification assumption.

(vii) *Estimate the equation from part (vi) using* **`brthord`** *as an IV for* **`educ`** *(and* **`sibs`** *as its own IV). Comment on the standard errors for* $\hat{\beta}_{educ}$ *and* $\hat{\beta}_{sibs}$.

(viii) *Using the fitted values from part (vi),* $\widehat{\texttt{educ}}$, *compute the correlation between* $\widehat{\texttt{educ}}$ *and* **`sibs`**. *Use this result to explain your findings from part 7.*

Problem 4. *Use the data in 401KUBS (Description) for this exercise. The equation of interest is a linear probability model:*

$$pira = \beta_0 + \beta_1 p401k + \beta_2 inc + \beta_3 inc^2 + \beta_4 age + \beta_5 age^2 + u.$$

The goal is to test whether there is a tradeoff between participating in a 401(k) plan and having an individual retirement account (IRA). Therefore, we want to estimate β_1.

(i) *Estimate the equation by OLS and discuss the estimated effect of* **p401k**.

(ii) *For the purposes of estimating the ceteris paribus tradeoff between participation in two different types of retirement savings plans, what might be a problem with ordinary least squares (OLS)?*

(iii) *The variable* **e401k** *is a binary variable equal to one if a worker is eligible to participate in a 401(k) plan. Explain what is required for* **e401k** *to be a valid Instrumental Variable (IV) for* **p401k**. *Do these assumptions seem reasonable?*

(iv) *Estimate the reduced form for* **p401k** *and verify that* **e401k** *has significant partial correlation with* **p401k**. *Since the reduced form is also a linear probability model, use a heteroskedasticity-robust standard error.*

$$p401k = \delta_0 + \delta_1 e401k + \delta_2 inc + \delta_3 inc^2 + \delta_4 age + \delta_5 age^2 + e.$$

(v) *Now, estimate the structural equation by IV and compare the estimate of β_1 with the OLS estimate. Again, you should obtain heteroskedasticity-robust standard errors.*