Problem Set II: Instrumental Variables

1. You are interested in the causal effect of cigarette smoking during pregnancy on the weight of newborns. Use the data in BWGHT.dta to estimate this effect. You specify the equation

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

where *bwght* is the birth weight (in ounces, 1 oz $\approx$ 28 gram), *male* is a binary indicator equal to one if the child is male, *parity* is the birth order of this child (between 1 and 6), $faminc$ is family income (in 1000 US$), and *packs* is the average number of packs of cigarettes smoked per day during pregnancy.

   (a) Explain why including the regressor $\log(faminc)$ may be important to correctly estimate the causal effect of *packs*.

   (b) Estimate the equation by OLS. What is the estimated effect? Is it plausible? Is it quantitatively relevant and statistically significant?

   (c) Why might you expect *packs* to be correlated with u? What does this imply for the OLS results?

   (d) Why could the cigarette price, *cigprice*, in the home state of the mother be a sensible instrument for *packs*?

   (e) Estimate the equation by 2SLS. What is the estimated effect? Is it plausible? Is it quantitatively relevant and statistically significant?

   (f) Discuss any important differences in the OLS and 2SLS estimates.

   (g) Estimate the reduced form (first stage regression) for packs and compute the relevant first-stage F statistic. What do you conclude about identification of the equation above using *cigprice* as an instrument for *packs*? What does this conclusion imply for the 2SLS estimate?

2. You want to estimate the return to schooling. Use the data in CARD.dta for this problem (from Card, D. (1995), "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in Aspects of Labour Market Behavior: Essays in

Honour of John Vanderkamp, ed. L. N. Christophides, E. K. Grant, and R. Swidinsky. Toronto: University of Toronto Press, 201-222).

(a) Use OLS to regress *lwage* (log of the hourly wage in cents, 1976) on *educ* (years of schooling, 1976), *exper* (proxy for experience in years), $exper^2$, *black* (=1 if black), *south* (=1 if in south 1976), *smsa* (=1 if in Standard Metropolitan Statistical Area, 1976), *reg*661 through *reg*668 (regional dummies, 1966), and *smsa*66 (=1 if in Standard Metropolitan Statistical Area, 1966). Compare your results with Table 2, Column (2) in Card (1995).

(b) Estimate a first-stage regression for *educ* containing all explanatory variables from part a and the dummy variable *nearc*4 (=1 if residence near 4 year college, 1966). Do *educ* and *nearc*4 have a practically and statistically significant partial correlation? Is this plausible? Use the relevant first-stage F statistic to check intrumental relevance. (See also Table 3, Column (1) in Card (1993).)

(c) Estimate the *lwage* equation by IV, using *nearc*4 as an instrument for *educ*. Interpret the estimated return to education. Is it quantitatively relevant? Compare the 95 percent confidence interval for the return of education with that obtained from part a. (See also Table 3, Column (5) in Card (1993).)

(d) Now use *nearc*2 along with *nearc*4 as instruments for *educ*. First estimate the reduced form for *educ*, and comment on whether *nearc*2 or *nearc*4 is more strongly related to *educ*. How do the 2SLS estimates compare with the earlier estimates? Based on the first-stage regressions, would you prefer to use *nearc*4 alone or together with *nearc*2 as instruments?

(e) For a subset of the men in the sample, IQ score is a available. Regress *iq* on *nearc*4. Is IQ score unrelated with *nearc*4? Try to explain.

(f) Now regress *iq* on *nearc*4 along with *smsa*66, and *reg*661 through *reg*668. Are *iq* and *nearc*4 partially correlated? What do you conclude about the importance of controlling for the 1966 location and regional dummies in the *lwage* equation when using *nearc*4 as an IV for *educ*?