PS5 Part 2

Chunyu Qu

Oct 25, 2021

5.3

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

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

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?

Answer

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, and all of these are significant explanatory variables of wage.

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.

Answer

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 (5.54). First, use OLS. Then, use 2SLS, where cigprice is an instrument for packs. Discuss any important differences in the OLS and 2SLS estimates.

Answer

```
# Read the data first
BWGHT = read.csv("D:/Google Drive/Fordham/2019 Spring/AE/MITDataCSV/BWGHT.csv", header = TRUE)

# Run the OLS
lm_53ols = lm(lbwght ~ male + parity + lfaminc + packs, data=BWGHT)
summary(lm_53ols)
```

```
##
```

```
## Call:
## lm(formula = lbwght ~ male + parity + lfaminc + packs, data = BWGHT)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.63729 -0.08845  0.02034  0.12271  0.84409
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.675618   0.021881 213.681  < 2e-16 ***
## male         0.026241   0.010089   2.601  0.00940 **
## parity       0.014729   0.005665   2.600  0.00942 **
## lfaminc      0.018050   0.005584   3.233  0.00126 **
## packs       -0.083728   0.017121  -4.890 1.12e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1876 on 1383 degrees of freedom
## Multiple R-squared:  0.03504,    Adjusted R-squared:  0.03225
## F-statistic: 12.55 on 4 and 1383 DF,  p-value: 4.905e-10
```

```
# Run 2SLS with cigprice as the IV
lm_53iv = ivreg(lbwght ~ male + parity + lfaminc + packs | male + parity + lfaminc + cigprice, data=BWG
summary(lm_53iv)
```

```
##
## Call:
## ivreg(formula = lbwght ~ male + parity + lfaminc + packs | male +
##     parity + lfaminc + cigprice, data = BWGHT)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.19538 -0.06910  0.07829  0.19077  0.89686
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.467861   0.258829  17.262   <2e-16 ***
## male         0.029821   0.017779   1.677   0.0937 .
## parity      -0.001239   0.021932  -0.056   0.9550
## lfaminc      0.063646   0.057013   1.116   0.2645
## packs        0.797106   1.086275   0.734   0.4632
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3202 on 1383 degrees of freedom
## Multiple R-Squared: -1.812,  Adjusted R-squared: -1.82
## Wald test: 2.391 on 4 and 1383 DF,  p-value: 0.04896
```

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 (5.54) using cigprice as an instrument for packs? What bearing does this conclusion have on your answer from part c?

Answer

```
# Run the 1st stage reduced form for packs
pack_rf = lm(packs ~ male + parity + lfaminc + cigprice, data=BWGHT)
summary(pack_rf)
```

```
##
## Call:
## lm(formula = packs ~ male + parity + lfaminc + cigprice, data = BWGHT)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36386 -0.11365 -0.08285 -0.04761  2.36602
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1374075  0.1040005   1.321   0.1866
## male        -0.0047261  0.0158539  -0.298   0.7657
## parity       0.0181491  0.0088802   2.044   0.0412 *
## lfaminc     -0.0526374  0.0086991  -6.051 1.85e-09 ***
## cigprice     0.0007770  0.0007763   1.001   0.3171
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2945 on 1383 degrees of freedom
## Multiple R-squared:  0.03045,    Adjusted R-squared:  0.02765
## F-statistic: 10.86 on 4 and 1383 DF,  p-value: 1.137e-08
```

```
# Generate estimated packs
packs_hat = fitted.values(pack_rf)

# Run the 2nd stage regression
lm_2nd = lm(lbwght ~ male + parity + lfaminc + packs_hat, data=BWGHT)
summary(lm_2nd)
```

```
##
## Call:
## lm(formula = lbwght ~ male + parity + lfaminc + packs_hat, data = BWGHT)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.62699 -0.08716  0.02033  0.12158  0.84799
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.467861   0.152848  29.231  < 2e-16 ***
## male         0.029821   0.010499   2.840  0.00457 **
## parity      -0.001239   0.012952  -0.096  0.92380
## lfaminc      0.063646   0.033668   1.890  0.05891 .
```

```
## packs_hat    0.797106   0.641484   1.243  0.21423
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1891 on 1383 degrees of freedom
## Multiple R-squared:  0.01945,    Adjusted R-squared:  0.01661
## F-statistic: 6.857 on 4 and 1383 DF,  p-value: 1.829e-05
```

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.

5.7

Consider model (5.45) where v has zero mean and is uncorrelated with $x_1, ..., x_K$ and q. The unobservable q is thought to be correlated with at least some of the xj. Assume without loss of generality that $E[q] = 0$. You have a single indicator of q, written as $q_1 = \delta_1 q + a_1, \delta_1 \neq 0$, where a_1 has zero mean and is uncorrelated with each of x_j, q. and v. In addition $z_1, z_2, ..., z_m$ is a set of variables that are (1) redundant in the structural equation (5.45) and (2) uncorrelated with a_1.

a. Suggest an IV method for consistently estimating the b_j. Be sure to discuss what is needed for identification.

Answer

Take q_1 as a proxy for q, then we express q by q_1 as

$$q = (1/\delta_1)q_1 - (1/\delta_1)a_1$$

Plug this into (5.45)

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_K x_K + \eta_1 q_1 + v - \eta_1 a_1$$

where $\eta_1 = 1/\delta_1$. Now, because the z_h are redundant in (5.45), they are uncorrelated with the structural error, v (by definition of redundancy). Further, we have assumed that the z_h are uncorrelated with a_1. Since each x_j is also uncorrelated with $v - \eta_1 a_1$ we can estimate (5.56) by 2SLS using instruments $(1, x_1, ...x_K, z_1, ..., z_M)$ to get consistent of the β_j and η_1.

Given all of the zero correlation assumptions, what we need for identification is that at least one of the z_h appears in the reduced form for q_1. More formally, in the linear projection

$$q_1 = \pi_0 + \pi_1 x_1 + ... + \pi_K x_K + \pi_{K+1} z_1 + ...\pi_{K+M} z_M + r_1$$

at least one of $\pi_{K+1}...\pi_{K+M}$ must be different from zero.

b. If equation (5.45) is a log(wage) equation, q is ability, q1 is IQ or some other test score, and $z_1, ..., z_M$ are family background variables, such as parents' education and number of siblings, describe the economic assumptions needed for consistency of the the IV procedure in part a.

Answer

We need family background variables to be redundant in the log(wage) equation once ability (and other factors, such as educ and exper), have been controlled for. The idea here is that family background may influence ability but should have no partial effect on log(wage) once ability has been accounted for. For the rank condition to hold, we need family background variables to be correlated with the indicator, q_1 say IQ, once the x_j have been netted out. This is likely to be true if we think that family background and ability are (partially) correlated.

c. Carry out this procedure using the data in NLS80.RAW. Include among the explanatory variables exper, tenure, educ, married, south, urban, and black. First use IQ as q1 and then KWW. Include in the z_h the variables meduc, feduc, and sibs. Discuss the results.

Answer

```
# Read the data first
NLS80 = read.csv("D:/Google Drive/Fordham/2019 Spring/AE/MITDataCSV/NLS80.csv", header = TRUE)

# Use IQ as q1
lm_57iq = ivreg(lwage ~ exper + tenure + educ + married + south + urban + black + iq | exper + tenure +
summary(lm_57iq)
```

```
##
## Call:
## ivreg(formula = lwage ~ exper + tenure + educ + married + south +
##     urban + black + iq | exper + tenure + educ + married + south +
##     urban + black + meduc + feduc + sibs, data = NLS80)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -2.134057 -0.217364  0.005651  0.231091  1.402071
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.471615   0.468913   9.536  < 2e-16 ***
## exper        0.016219   0.004008   4.047 5.76e-05 ***
## tenure       0.007675   0.003096   2.479   0.0134 *
## educ         0.016181   0.026198   0.618   0.5370
## married      0.190101   0.046759   4.066 5.33e-05 ***
## south       -0.047992   0.036742  -1.306   0.1919
## urban        0.186938   0.032799   5.700 1.76e-08 ***
## black        0.040027   0.113868   0.352   0.7253
## iq           0.015437   0.007708   2.003   0.0456 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3878 on 713 degrees of freedom
## Multiple R-Squared: 0.1546,  Adjusted R-squared: 0.1451
## Wald test: 25.81 on 8 and 713 DF,  p-value: < 2.2e-16
```

```
# KWW IQ as q1
lm_57kww = ivreg(lwage ~ exper + tenure + educ + married + south + urban + black + kww | exper + tenure
summary(lm_57kww)
```

```
##
## Call:
## ivreg(formula = lwage ~ exper + tenure + educ + married + south +
##     urban + black + kww | exper + tenure + educ + married + south +
##     urban + black + meduc + feduc + sibs, data = NLS80)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
```

```
## -2.319371 -0.238609  0.003009  0.252612  1.496516
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.217818   0.162759  32.059  < 2e-16 ***
## exper        0.006868   0.006747   1.018 0.309044
## tenure       0.005115   0.003774   1.355 0.175765
## educ         0.026081   0.025505   1.023 0.306857
## married      0.160527   0.052976   3.030 0.002532 **
## south       -0.091887   0.032215  -2.852 0.004466 **
## urban        0.148400   0.041160   3.605 0.000333 ***
## black       -0.042445   0.089370  -0.475 0.634974
## kww          0.024944   0.015058   1.657 0.098045 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3874 on 713 degrees of freedom
## Multiple R-Squared: 0.1563,  Adjusted R-squared: 0.1468
## Wald test:  25.7 on 8 and 713 DF,  p-value: < 2.2e-16
```

Even though there are 935 men in the sample, only 722 are used for the estimation because data are missing on meduc and feduc. The return to education is estimated to be small and insignificant whether IQ or KWW used is used as the indicator. This could be because family background variables do not satisfy the appropriate redundancy condition, or they might be correlated with a1. (In both first-stage regressions, the F statistic for joint significance of meduc,feduc and sibs have p-values below .002, so it seems the family background variables have some partial correlation with the ability indicators.)