Econometrics HW5

Tianyu Gao

2022-11-10

1

a)

First stage:

$$\text{educ}_i = \pi_0 + \pi_1 \, \text{huseduc}_i + \pi_2 \, \text{motheduc}_i + \pi_3 \, f\,\text{atheduc}_i + \nu_i$$

b)

If we denote the function as follows:

$$\text{educ}_i = \pi_0 + \pi_1 \, \text{exper}_i + \pi_2 \, \text{exper}_i^2 + \pi_3 \, \text{motheduc}_i + \pi_4 \, f\,\text{atheduc}_i + \pi_5 \, \text{huseduc}_i + u_i$$

then we can re-write the hypothesis testing in the question into following form:

$$\begin{gathered} H_0 : \pi_3 = \pi_4 = \pi_5 = 0 \\ H_1 : \exists i \in [3,5], \ \pi_i \neq 0 \end{gathered}$$

$$F = \frac{\frac{R^2_{\text{unrestricted}} - R^2_{\text{restricted}}}{q}}{\frac{1 - R^2_{\text{unrestricted}}}{n - k_{\text{unrestricted}} - 1}} = \frac{\frac{.4286 - .0049}{3}}{\frac{1 - .4286}{428 - 5 - 1}} = 104.346$$

```
((.4286 - .0049) / 3) / ((1 - .4286) / (428 - 5 - 1))
```

```
## [1] 104.306
```

```
qf(.95, 3, 422)
```

```
## [1] 2.626046
```

$$\begin{gathered} \because F_{3,422}(.95) = 2.62 < 104.306 \\ \therefore H_0 \text{ is rejected at } 95\% \text{ CI} \end{gathered}$$

Therefore we are sure about the instrument are relative.

However, I don’t think the exogenousity holds, because the education of parents determines the wealth of the family, which would attract other high end firms hire their daughters so that the family can be the clients. For example, the investment banks tend to hire some “VIP” employees so that the employees’ family would choose the investment bank to manage their assets.

Also, according to the model in question 1, the education level of husbands would be correlated with the education level of the wives. In this way, the education level of fathers would also be correlated with the education level of the mothers. So there is a trend of multicollinearity among the instrumental variables.

2

In the following answer, the subscript is omitted, indicating that all the variables are in vectors of composed of individual observations.

a)

i)

$$\text{Childern} = -4.13 - .09\text{educ} + .33\text{age} - .002\text{age}^2 + \text{u}$$

On average, when the women receive one more year of education, holding other variables constant, the women would have have $.09$ less children.

ii)

$$100 * .09 = 9$$

Thus, 9 fewer children are expected to have in the group.

b)

Regression 2:

$$\text{Children} = -.07\text{educ} + .34\text{age} - .002\text{age}^2 - .30\text{electric} - .25\text{tv} + .3\text{bicycle} - 4.39 + \text{u}$$

I prefer model 2. This is because the financial conditions tends to have effect on the birth of the children. In poor countries, the parents would like to have more children, due to traditional religious values (Paul Harrison, 1990), traditional masculinity, patriarchy and dependency toward the children(Adamson, 1986). Meanwhile, in places of lower degree of development, the women are less educated. Mathematically,

$$\text{E}(\text{u}|\text{educ}) \neq 0$$

leading to a omitted variable bias.

By adding the $\text{tv}, \text{electric}, \text{bicycle}$ as control variables, we can control the omitted variable bias caused by the financial conditions, here they work as proxy variables.

c)

Robust is a modification for ordinary least square. It usually uses weighted least square the make the errors in the model a constant, mitigating the heteroskedasticity of model.

We use this regression method because the effect of willingness of giving birth to children toward the number of children tends to change during the life span of women. When the observation is younger, she will have fewer children no matter what her willingness of giving birth to children is. As the age goes up, some people willing to have children will give birth to more children while some people not willing to give birth will not. This caused a heteroskedasticity in the model.

d)

When a woman have more children in Botswana, the probability of having more education is lower because she needs to take care of her kid, which costs a lot of time and energy.

e)

According the the result of the regression 3, $\mathrm{firsthalf}$ is relevant to educ, which assures its relevance. The willingness to give birth to children tend to be uniformly distributed considering the birth month, which ensures its exogeneity. Moreover, girls born in the first half of year can receive education of several more months, because the girls born in the first half of year can attend school one year before those born in the last half, and they drop of school in the same year.

3

In answer to this question, $\mathrm{u_i}$ denotes the error in the model.

a)

Zip code is not a good instrument. Families better off cluster in the US, and for kids in the families better off, they have more opportunity, and their income would be higher. Also, the kids from families better off tend to receive more education because of the optimism of the future. They tend to believe they can repay their student loan, and some parents would sponsor their kids for education.

In this way,

$$\mathrm{corr}(\mathrm{zip\ code}, \mathrm{u_i}) \neq 0$$

which doesn’t satisfy the requirement of exogeneity.

b)

The IQ and work-related test score are not good instruments. Obviously IQ can effect the way people solve problems, and work-related test scores represents the ability to deal with affairs in the job, these two indicators would have effect on salaries.

In other words,

$$\mathrm{corr}(\mathrm{IQ}, \mathrm{u_i}) \neq 0$$
$$\mathrm{corr}(\mathrm{work_score}, \mathrm{u_i}) \neq 0$$

which don’t satisfy the requirement of exogeneity.

c)

The education year of parents is not a good instrument, because the education of parents determines the wealth of the family, which would attract other high end firms hire their daughters so that the family can be the clients. For example, the investment banks tend to hire some “VIP” employees so that the employees’ family would choose the investment bank to manage their assets.

That is to say,

$$\mathrm{corr}(\mathrm{parents}'\ \mathrm{education}, \mathrm{u_i}) \neq 0$$

which doesn’t satisfy the requirement of exogeneity.

d)

The number of siblings is a good instrument, and this is because the number of siblings leads to the distribution of the resource in the family. In most cases, the more siblings a person have, the less resources he would share in a family (including the resources for education), leading to a relevance between education and siblings.

Equivalently,

$$\mathrm{corr}(\mathrm{sibling}, \mathrm{educ}) \neq 0$$

which satisfies the relevance

While the number of siblings doesn’t seem to correlate with the personal earning, or

$$\mathrm{corr}(\mathrm{earning}, \mathrm{sibling}) = 0$$

which satisfies the exogeneity.

4

```
load('NewsStudy.RData')
```

```
summary(dta)
```

```
##     resid               Female          watchnat        ReadNews
##  Length:507         Min.   :0.0000   Min.   :1.000   Min.   :1.000
##  Class :character   1st Qu.:0.0000   1st Qu.:6.000   1st Qu.:5.000
##  Mode  :character   Median :1.0000   Median :7.000   Median :6.000
##                     Mean   :0.5422   Mean   :6.102   Mean   :5.501
##                     3rd Qu.:1.0000   3rd Qu.:7.000   3rd Qu.:7.000
##                     Max.   :1.0000   Max.   :7.000   Max.   :7.000
##                     NA's   :9        NA's   :5       NA's   :2
##    pnintst        Education          income          Voted
##  Min.   :1.000   Min.   : 1.000   Min.   : 1.000   Min.   :0.0000
##  1st Qu.:3.000   1st Qu.: 8.000   1st Qu.: 4.000   1st Qu.:1.0000
##  Median :3.000   Median :10.000   Median : 5.000   Median :1.0000
##  Mean   :3.275   Mean   : 9.143   Mean   : 5.901   Mean   :0.8337
##  3rd Qu.:4.000   3rd Qu.:11.000   3rd Qu.: 8.000   3rd Qu.:1.0000
##  Max.   :4.000   Max.   :13.000   Max.   :11.000   Max.   :1.0000
##  NA's   :1       NA's   :2        NA's   :63       NA's   :8
##      prop         infopro      WatchProgram       learnpro
##  Min.   :1.00   Min.   :-3.0   Min.   :0.0000   Min.   :-2.000
##  1st Qu.:1.00   1st Qu.: 1.0   1st Qu.:0.0000   1st Qu.: 2.000
##  Median :1.00   Median : 2.0   Median :0.0000   Median : 2.000
##  Mean   :1.11   Mean   : 1.7   Mean   :0.2525   Mean   : 2.312
##  3rd Qu.:1.00   3rd Qu.: 2.0   3rd Qu.:1.0000   3rd Qu.: 3.000
##  Max.   :2.00   Max.   : 4.0   Max.   :1.0000   Max.   : 4.000
##  NA's   :98                                     NA's   :379
##  TreatmentGroup     prop_dv       InformationLevel   prop_vote
##  Min.   :0.0000   Min.   :0.0000   Min.   :1.000    Min.   :0.00
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:3.000    1st Qu.:1.00
##  Median :1.0000   Median :0.0000   Median :3.000    Median :1.00
##  Mean   :0.5108   Mean   :0.1933   Mean   :3.233    Mean   :0.89
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:4.000    3rd Qu.:1.00
##  Max.   :1.0000   Max.   :1.0000   Max.   :4.000    Max.   :1.00
##                                    NA's   :9        NA's   :98
##      white           partyid      PoliticalInterest
##  Min.   :0.0000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.0000   1st Qu.:1.000   1st Qu.:3.000
##  Median :1.0000   Median :3.000   Median :3.000
##  Mean   :0.7613   Mean   :3.506   Mean   :3.275
##  3rd Qu.:1.0000   3rd Qu.:6.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :7.000   Max.   :4.000
##                   NA's   :15      NA's   :1
```

```
unique(dta$Voted, rm.na =TRUE)
```

```
## [1]  1  0 NA
```

```
unique(dta$InformationLevel)
```

```
## [1]  4  3  2  1 NA
```

a)

```
modela = lm(InformationLevel ~ WatchProgram, data = dta)
summary(modela)
```

```
##
## Call:
## lm(formula = InformationLevel ~ WatchProgram, data = dta)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4531 -0.4531 -0.1568  0.8432  0.8432
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.15676    0.04270  73.927  < 2e-16 ***
## WatchProgram  0.29637    0.08423   3.519 0.000474 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8214 on 496 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.02435,    Adjusted R-squared:  0.02239
## F-statistic: 12.38 on 1 and 496 DF,  p-value: 0.0004736
```

```
library(lmtest)
library(sandwich)
library(ivreg)
```

```
coeftest(modela, vcov = vcovHC(modela, type = 'HC1'))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  3.156757   0.044738  70.560 < 2.2e-16 ***
## WatchProgram 0.296368   0.076051   3.897 0.0001108 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Therefore, for people who watch program, their information level is .30 higher than the people who don’t holding other constant.

However, this result could be biased, because some people might be attracted by the information they know and they watch more report on the TV.

b)

```
modelb = lm(InformationLevel~PoliticalInterest + WatchProgram + ReadNews + Education, da
ta = dta)
summary(modelb)
```

```
##
## Call:
## lm(formula = InformationLevel ~ PoliticalInterest + WatchProgram +
##     ReadNews + Education, data = dta)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5258 -0.5223  0.2404  0.4777  1.9283
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.6942656  0.1640090  10.330  < 2e-16 ***
## PoliticalInterest 0.2650756  0.0460088   5.761 1.48e-08 ***
## WatchProgram      0.2329059  0.0769573   3.026  0.00261 **
## ReadNews          0.1087893  0.0182718   5.954 5.01e-09 ***
## Education         0.0008844  0.0124248   0.071  0.94328
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7421 on 488 degrees of freedom
##   (14 observations deleted due to missingness)
## Multiple R-squared:  0.2053, Adjusted R-squared:  0.1987
## F-statistic: 31.51 on 4 and 488 DF,  p-value: < 2.2e-16
```

```
coeftest(modelb, vcov = vcovHC(modelb, type = 'HC1'))
```

```
##
## t test of coefficients:
##
##                     Estimate Std. Error t value  Pr(>|t|)
## (Intercept)       1.69426560 0.17823418  9.5058 < 2.2e-16 ***
## PoliticalInterest 0.26507563 0.05046391  5.2528 2.242e-07 ***
## WatchProgram      0.23290588 0.07301845  3.1897  0.001516 **
## ReadNews          0.10878926 0.02108053  5.1606 3.587e-07 ***
## Education         0.00088438 0.01258454  0.0703  0.944003
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

PoliticalInterest settles the concern I mentioned above, and the ReadNews controls other methods interviewees could know about the information.

c)

On one hand, people assigned to watch the program will more likely to watch the program, which contributes to the relevance of the instrument variable.

On the other hand, if we randomly assign people to watch the program regardless of the information level, this variable is exogenous.

```
modelc = ivreg(InformationLevel~ WatchProgram|TreatmentGroup, data = dta)
summary(modelc)
```

```
##
## Call:
## ivreg(formula = InformationLevel ~ WatchProgram | TreatmentGroup,
##     data = dta)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4055 -0.4055 -0.1732  0.8268  0.8268
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.17325    0.05842  54.321   <2e-16 ***
## WatchProgram  0.23221    0.17642   1.316    0.189
##
## Diagnostic tests:
##                  df1 df2 statistic p-value
## Weak instruments   1 496   146.646  <2e-16 ***
## Wu-Hausman         1 495     0.171   0.679
## Sargan             0  NA        NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8219 on 496 degrees of freedom
## Multiple R-Squared: 0.02321, Adjusted R-squared: 0.02124
## Wald test: 1.732 on 1 and 496 DF,  p-value: 0.1887
```

According to the result above, the instrument is strongly relevant to the endogenous variable, and the $p - \text{value}$ of the Wu-Hausman test is large, indicating that there is no endogeneity of the instrument. However, the endogenous variable itself is not significant.

d)

```
modeld = ivreg(InformationLevel~ WatchProgram + PoliticalInterest + ReadNews + Education
|PoliticalInterest + ReadNews + Education + TreatmentGroup, data = dta)
summary(modeld)
```

```
##
## Call:
## ivreg(formula = InformationLevel ~ WatchProgram + PoliticalInterest +
##     ReadNews + Education | PoliticalInterest + ReadNews + Education +
##     TreatmentGroup, data = dta)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5086 -0.5071  0.2001  0.4929  1.9362
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.6887506  0.1646510  10.257  < 2e-16 ***
## WatchProgram      0.2913412  0.1613956   1.805   0.0717 .
## PoliticalInterest 0.2640640  0.0461014   5.728 1.78e-08 ***
## ReadNews          0.1078766  0.0184163   5.858 8.64e-09 ***
## Education         0.0007689  0.0124353   0.062   0.9507
##
## Diagnostic tests:
##                   df1 df2 statistic p-value
## Weak instruments    1 488    143.82  <2e-16 ***
## Wu-Hausman          1 487      0.17   0.681
## Sargan              0  NA        NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7426 on 488 degrees of freedom
## Multiple R-Squared: 0.2043,  Adjusted R-squared: 0.1978
## Wald test:    30 on 4 and 488 DF,  p-value: < 2.2e-16
```

The results are similar, but the degree of freedom is different. This is because some observations containing null value are removed from the model.

e)

The result does not show me that watching TV program increases the information level. Now I defeated the endogenous problem, but the variable of interest is not significant.