ECO375H1S

Short Written Assignment

Professor: Martin Burda

Teaching Assistant: Remi Daviet

Student: Isaac Aktam

998986575

Note: all of the code can be found in the appendix. Code seen in the below solutions is added for the convenience of the reader.

a) Regress weeksm1 (weeks worked in 1979, a measure of labor supply) on the indicator variable morekids (=1 for 2+ children in family) using OLS. On average, do women with more than two children work less than women with two children? How much less?  
Answer:

> ols\_reg\_a <- lm(weeksm1 ~ morekids)

> ols\_reg\_a

Call:

lm(formula = weeksm1 ~ morekids)

Coefficients:

(Intercept) morekids

21.068 -5.387

> summary(ols\_reg\_a)

Call:

lm(formula = weeksm1 ~ morekids)

Residuals:

Min 1Q Median 3Q Max

-21.07 -21.07 -13.68 24.93 36.32

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 21.06843 0.05466 385.4 <2e-16 \*\*\*

morekids -5.38700 0.08861 -60.8 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.71 on 254652 degrees of freedom

Multiple R-squared: 0.01431, Adjusted R-squared: 0.0143

F-statistic: 3696 on 1 and 254652 DF, p-value: < 2.2e-16

As can be seen from the code above, the coefficient for morekids is -5.387. Therefore, women, who on average have more than two kids, work approximately 5 weeks less than women with two children or less.

b) Briefly explain why the OLS regression estimated in (a) is inappropriate for estimating the causal effect of fertility (morekids) on labor supply (weeksm1).

Answer:

When people have a choice, their desire to have more kids may correlate with an error term. This means that an error contains variables that correlate with morekids. Ergo, morekids is an endogenous variable and is subject to selection bias. For example, morekids may be correlated with

* Cultural believes where the use of contraceptives is not allowed or people are unaware of contraceptives or can’t afford them
* Male member of the family earns a sufficient amount of income to support a family with many children
* Government provides financial support to families as an incentive to give birth to more children

On the other hand, morekids may be an invalid explanatory variable as we are not concerned with women who have 2 children or less. Therefore we are showing a selection bias since we are dealing only with a subset of female demographic For example, a woman with 1 child or not may work less or not work at all if

* she has a sufficient financial capital that she can rely upon
* she has sickness that affects her work performance and thus causes her to work less
* company may ask her go on long vacation because she works so hard every year

Therefore, there are many other factors, apart from woman’s fertility, that affect the labor supply.

(c) The data set contains the variable samesex, which is equal to 1 if the first two children are

of the same sex (boy-boy or girl-girl) and equal to 0 otherwise. Are couples whose first two children are of the same sex more likely to have a third child (as measured by morekids)? Is the effect large? Is it statistically significant?

Answer:

> ols\_reg\_c <- lm(morekids ~ samesex)

> ols\_reg\_c

Call:

lm(formula = morekids ~ samesex)

Coefficients:

(Intercept) samesex

0.34642 0.06753

> summary(ols\_reg\_c)

Call:

lm(formula = morekids ~ samesex)

Residuals:

Min 1Q Median 3Q Max

-0.4139 -0.4139 -0.3464 0.5860 0.6536

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.346425 0.001365 253.79 <2e-16 \*\*\*

samesex 0.067525 0.001920 35.17 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4844 on 254652 degrees of freedom

Multiple R-squared: 0.004835, Adjusted R-squared: 0.004831

F-statistic: 1237 on 1 and 254652 DF, p-value: < 2.2e-16

From the R-code we see that the coefficient of samesex is 0.067525. Since the coefficient is positive, we expect that the third child will be of the same sex as the previous two children with a degree of belief (or, probability) of 6.7525% compared to families whose previous two children are not of the same sex. Therefore, the degree of difference is roughly 7%. The true value of the difference can be calculated as e0.067525 - 0 = 1.069. T-value for samesex is 35.17 > 1.96. The p – value for the same the samesex coefficient is <2e-16 \*\*\* < 0.05. Therefore, the effect is statistically significant such that samesex does not equal to zero.

(d) Briefly explain why samesex is a valid instrument for the IV regression of weeksm1 on

morekids.

Answer:

> cor(morekids,samesex)

[1] 0.06953403

> ols\_reg\_c <- lm(morekids ~ samesex)

> ols\_reg\_c\_res = resid(ols\_reg\_c)

> cor(ols\_reg\_c\_res,samesex)

[1] -3.134887e-17

To test the validity of the instrumental variable samesex, one may need to satisfy both of the following conditions:

* Instrument relevance, calculate the correlation between samesex and morekids
* Instrument validity, calculate the samesex and the OLS residuals

From the above, we see the desired result of instrument’s relevance, cor(morekids, samesex) = 0.06953403 ≠ 0. Second condition of exogeniety is also satisfied, cor(ols\_reg\_c\_res, samesex) = -3.134887e-17 ≈ 0. Therefore, samesex is a valid instrumental variable.

But, does this result apply to the whole population? No, since we are only given a sample. Therefore, we proceed with an intuitive explanation. Samesex appears to be a random instrumental variable independent of other control variables in the model. Also, it appears to be independent of the error term. Additionally, parents have no control over the gender of their children and thus there is no selection bias involved.

We can also test the validity of the samesex using the t-test from the R-code. As we can see, the p-value for samesex is <2e-16 \*\*\*<0.05. The test is significant of 0.05 level and coefficient samesex is not 0. Therefore, it is valid to include it in the model.

> ols\_reg\_c <- lm(morekids ~ samesex)

> summary(ols\_reg\_c)

Call:

lm(formula = morekids ~ samesex)

Residuals:

Min 1Q Median 3Q Max

-0.4139 -0.4139 -0.3464 0.5860 0.6536

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.346425 0.001365 253.79 <2e-16 \*\*\*

samesex 0.067525 0.001920 35.17 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4844 on 254652 degrees of freedom

Multiple R-squared: 0.004835, Adjusted R-squared: 0.004831

F-statistic: 1237 on 1 and 254652 DF, p-value: < 2.2e-16

(e) Estimate the IV regression of weeksm1 on morekids using samesex as an instrument. How large is the fertility effect on labor supply?

Answer:

> iv\_reg = ivreg(weeksm1~ morekids | samesex)

> summary(iv\_reg)

Call:

ivreg(formula = weeksm1 ~ morekids | samesex)

Residuals:

Min 1Q Median 3Q Max

-21.42 -21.42 -13.42 24.89 36.89

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 21.421 0.487 43.988 < 2e-16 \*\*\*

morekids -6.314 1.275 -4.953 7.3e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.71 on 254652 degrees of freedom

Multiple R-Squared: 0.01388, Adjusted R-squared: 0.01388

Wald test: 24.54 on 1 and 254652 DF, p-value: 7.296e-07

Answer:

From the above 2SLS regression, the coefficient of morekids is -6.314. This suggests, that on average, having more than two kids reduces the female labor supply by the amount of 6.314 weeks.

(f) Do the results change when you include the variables agem1, black, hispan, and othrace in the labor supply regression?

Answer:

> iv\_reg\_f <- ivreg(weeksm1 ~ morekids + agem1 + black + hispan + othrace | samesex + agem1 + black + hispan + othrace)

> iv\_reg\_f

Call:

ivreg(formula = weeksm1 ~ morekids + agem1 + black + hispan + othrace | samesex + agem1 + black + hispan + othrace)

Coefficients:

(Intercept) morekids agem1 black hispan othrace

-4.7919 -5.8211 0.8316 11.6233 0.4042 2.1310

> summary(iv\_reg\_f)

Call:

ivreg(formula = weeksm1 ~ morekids + agem1 + black + hispan +

othrace | samesex + agem1 + black + hispan + othrace)

Residuals:

Min 1Q Median 3Q Max

-36.34 -17.66 -10.99 22.72 45.15

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -4.79189 0.40657 -11.786 <2e-16 \*\*\*

morekids -5.82105 1.24631 -4.671 3e-06 \*\*\*

agem1 0.83160 0.02289 36.336 <2e-16 \*\*\*

black 11.62327 0.22893 50.772 <2e-16 \*\*\*

hispan 0.40418 0.25986 1.555 0.12

othrace 2.13096 0.20586 10.352 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.38 on 254648 degrees of freedom

Multiple R-Squared: 0.04368, Adjusted R-squared: 0.04366

Wald test: 1335 on 5 and 254648 DF, p-value: < 2.2e-16

From the updated 2SLS, we see that once we include the demographic variables (or hidden control variables), the coefficient of morekids changes by a slight amount from -6.314 to -5.8211. One of the reasons why change is small is because samesex is an exogenous variable and thus independent of added demographic control variables. From the updated 2SLS, women work 5.8 weeks less compared to 5.38 weeks (OLS). Therefore, in both cases of ivreg, women with more than two children work less for more than 5.3 weeks.

Lastly, we can test whether to use the IV estimator or OLS estimator. To do so, we use the Hausman Test for IV:

> # Hausman test for endogeneity of regressors

> cf\_diff = coef(iv\_reg) - coef(ols\_reg\_a)

> vc\_diff = vcov(iv\_reg) - vcov(ols\_reg\_a)

> samesex\_diff = as.vector(t(cf\_diff) %\*% solve(vc\_diff) %\*% cf\_diff)

> pchisq(samesex\_diff, df = 1, lower.tail = FALSE)

[1] 0.4661212

As we can see, at the significance level of 5%, 0.4661212 > 0.05. Therefore, we do not reject the null hypothesis H0: cov(y,u) = 0 which leads to d = = 0. It suffices to use OLS estimator in our case.

Appendix:

# Question a

> ols\_reg\_a <- lm(weeksm1 ~ morekids)

> ols\_reg\_a

Call:

lm(formula = weeksm1 ~ morekids)

Coefficients:

(Intercept) morekids

21.068 -5.387

> summary(ols\_reg\_a)

Call:

lm(formula = weeksm1 ~ morekids)

Residuals:

Min 1Q Median 3Q Max

-21.07 -21.07 -13.68 24.93 36.32

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 21.06843 0.05466 385.4 <2e-16 \*\*\*

morekids -5.38700 0.08861 -60.8 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.71 on 254652 degrees of freedom

Multiple R-squared: 0.01431, Adjusted R-squared: 0.0143

F-statistic: 3696 on 1 and 254652 DF, p-value: < 2.2e-16>

# Question c

ols\_reg\_c <- lm(morekids ~ samesex)

> ols\_reg\_c

Call:

lm(formula = morekids ~ samesex)

Coefficients:

(Intercept) samesex

0.34642 0.06753

> summary(ols\_reg\_c)

Call:

lm(formula = morekids ~ samesex)

Residuals:

Min 1Q Median 3Q Max

-0.4139 -0.4139 -0.3464 0.5860 0.6536

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.346425 0.001365 253.79 <2e-16 \*\*\*

samesex 0.067525 0.001920 35.17 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4844 on 254652 degrees of freedom

Multiple R-squared: 0.004835, Adjusted R-squared: 0.004831

F-statistic: 1237 on 1 and 254652 DF, p-value: < 2.2e-16

# Question d

> cor(morekids,samesex)

[1] 0.06953403

> ols\_reg\_c <- lm(morekids ~ samesex)

> ols\_reg\_c\_res = resid(ols\_reg\_c)

> cor(ols\_reg\_c\_res,samesex)

[1] -3.134887e-17

> iv\_reg = ivreg(weeksm1~ morekids | samesex)

> summary(iv\_reg)

Call:

ivreg(formula = weeksm1 ~ morekids | samesex)

Residuals:

Min 1Q Median 3Q Max

-21.42 -21.42 -13.42 24.89 36.89

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 21.421 0.487 43.988 < 2e-16 \*\*\*

morekids -6.314 1.275 -4.953 7.3e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.71 on 254652 degrees of freedom

Multiple R-Squared: 0.01388, Adjusted R-squared: 0.01388

Wald test: 24.54 on 1 and 254652 DF, p-value: 7.296e-07

#Question e

> iv\_reg\_e = ivreg(weeksm1~ morekids | samesex)

> summary(iv\_reg\_e)

Call:

ivreg(formula = weeksm1 ~ morekids | samesex)

Residuals:

Min 1Q Median 3Q Max

-21.42 -21.42 -13.42 24.89 36.89

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 21.421 0.487 43.988 < 2e-16 \*\*\*

morekids -6.314 1.275 -4.953 7.3e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.71 on 254652 degrees of freedom

Multiple R-Squared: 0.01388, Adjusted R-squared: 0.01388

Wald test: 24.54 on 1 and 254652 DF, p-value: 7.296e-07

# Question f

> iv\_reg\_f <- ivreg(weeksm1 ~ morekids + agem1 + black + hispan + othrace | samesex + agem1 + black + hispan + othrace)

> iv\_reg\_f

Call:

ivreg(formula = weeksm1 ~ morekids + agem1 + black + hispan + othrace | samesex + agem1 + black + hispan + othrace)

Coefficients:

(Intercept) morekids agem1 black hispan othrace

-4.7919 -5.8211 0.8316 11.6233 0.4042 2.1310

> summary(iv\_reg\_f)

Call:

ivreg(formula = weeksm1 ~ morekids + agem1 + black + hispan +

othrace | samesex + agem1 + black + hispan + othrace)

Residuals:

Min 1Q Median 3Q Max

-36.34 -17.66 -10.99 22.72 45.15

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -4.79189 0.40657 -11.786 <2e-16 \*\*\*

morekids -5.82105 1.24631 -4.671 3e-06 \*\*\*

agem1 0.83160 0.02289 36.336 <2e-16 \*\*\*

black 11.62327 0.22893 50.772 <2e-16 \*\*\*

hispan 0.40418 0.25986 1.555 0.12

othrace 2.13096 0.20586 10.352 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.38 on 254648 degrees of freedom

Multiple R-Squared: 0.04368, Adjusted R-squared: 0.04366

Wald test: 1335 on 5 and 254648 DF, p-value: < 2.2e-16

> # Hausman test for endogeneity of regressors

> cf\_diff = coef(iv\_reg) - coef(ols\_reg\_a)

> vc\_diff = vcov(iv\_reg) - vcov(ols\_reg\_a)

> samesex\_diff = as.vector(t(cf\_diff) %\*% solve(vc\_diff) %\*% cf\_diff)

> pchisq(samesex\_diff, df = 1, lower.tail = FALSE)

[1] 0.4661212