em hw4

2020-15709 서승환

1.

(a) 대우명제 $E(u_i|X_i)=0 \Rightarrow E(u_iX_i)=0$을 증명하면 충분하다.

$E(u_iX_i)$ ~~$=E(E(u_i|X_i)X_i)=E(E(u_iX_i|X_i))$~~

$=E(E(u_iX_i|X_i))$ (iterated expectation)

$=E(X_iE(u_i|X_i))=E(0)=0$이므로

$E(u_iX_i)\neq 0 \Rightarrow E(u_i|X_i)\neq 0$

(b) Exogeneity Condition : $Corr(Z_i,u_i)=0$

Relevance Condition : $Corr(Z_i,X_i)\neq 0$

$Corr(X,Y)=cov(X,Y)/\sqrt{Var(X)Var(Y)}$이므로

$Cov(Z_i,u_i)=0$ (Exo) / $cov(Z_i,X_i)\neq 0$ (Relevance)로 하는 조건을 찾으면 충분하다.

(i) Exogeneity

$Cov(Z_i,u_i)=E(Z_iu_i)+E(Z_i)\overset{=0}{E(u_i)}$

$=E(u_i(X_i^2+v_i))=E(X_i^2u_i)+E(\overset{=0}{u_i}v_i)$

$=E(X_i^2u_i)=0$에서 외생성이 만족된다.

(ii) Relevance

$Cov(Z_i,X_i)=E(Z_iX_i)-E(Z_i)\overset{=0}{E(X_i)}$

$=E((X_i^2+v_i)X_i)=E(X_i^3)+E(X_iv_i)$.

이때 $\int_{-\infty}^{\infty}x^3f(x)dx=0$ when $f(x)$는 우함수 $\Rightarrow E(X_i^3)=0$

$\therefore E(X_iv_i)\neq 0$에서 Revelance가 성립한다.

(c) $cov(Y_i,Z_i)=cov(X_i^2+v_i,\beta_1X_i+u_i)$

$=\beta_1cov(Z_i,X_i)+cov(Z_i,u_i)$

$=\beta_1cov(Z_i,X_i)$

$\therefore \beta_1=\dfrac{cov(Y_i,Z_i)}{cov(X_i,Z_i)}$

$\hat{\beta}_1=\dfrac{\widehat{cov}(Y_i,Z_i)}{\widehat{cov}(X_i,Z_i)} \xrightarrow{P} \dfrac{cov(Y_i,Z_i)}{cov(X_i,Z_i)}$

(d) Relevance가 만족된다. ~~항을 수 있다.~~

이 경우에도 ~~변수가~~ $\hat{\beta}_1$에 대한 (c)의 결론은 변하지 않는다.

2.

(a) $Y_i=\{Y_i(1)-Y_i(0)\}X_i+Y_i(0)$

$=E[Y_i(0)|X_i=1]+E[Y_i(1)-Y_i(0)|X_i=1]X_i$

$+Y_i(0)+\{\{Y_i(1)-Y_i(0)\}-E[Y_i(1)-Y_i(0)|X_i=1]\}X_i$

$-E[Y_i(0)|X_i=1]$

$\therefore \beta_0=E(Y_i(0)|X_i=1)$

$u_i=Y_i(0)-E(Y_i(0)|X_i=1)+\{Y_i(1)-Y_i(0)\}X_i$

$-E[Y_i(1)-Y_i(0)|X_i=1]X_i$

(b) $(X_i=0)$

$E(u_i|X_i=0)=E(Y_i(0)|X_i=0)-E(Y_i(0)|X_i\in 1)=0$

이는 $Y_i(0)$이 $X_i=0$이든 $X_i=1$이든 conditional expectation이 서로 같아야 함을, 즉 $Y_i(0)\perp X_i$를 의미한다. 한편 아래의 $X_i=1$ case에서 $E(u_i|X_i=1)$은 항등적으로 0이고, 이는 $Y_i(1)\perp X_i$를 조건으로 부과하지 '않아도' 0이 된다는 뜻이다.

$(X_i=1)$

$E(u_i|X_i=1)=E(Y_i(0)|X_i=1)-E(Y_i(0)|X_i=1)$

$+\{Y_i(1)-Y_i(0)\}-\{Y_i(1)-Y_i(0)\}=0.$

Econometrics HW4 Q3

Na SeungChan

Data Importing

```
raw_data <- readxl::read_xlsx('./fertility.xlsx')
q32_data <- raw_data %>%
  mutate(twoboys = ifelse(boy1st == 1&boy2nd == 1, 1, 0), twogirls = ifelse(boy1st == 0&boy2nd == 0, 1, 0))
```

이때, Q3.2에서 요구하는 'twoboys' 변수와 'twogirls' 변수를 생성하기 위해 data wrangling을 하였다.

Q3.1

반응변수 Y : 어머니의 노동공급 weeksm1 독립변수 X : 2명 넘는 아이 갖기 morekids (내생성 문제 존재)
X를 포착하기 위한 도구변수 Z : samesex(첫 두 아이가 같은 성별이면 1, 그렇지 않으면 0) control variables : agem1(출산 연령), black, hispan, othrace (인종 dummy)

우선 OLS estimation result를 계산한다. 이는 weeksm1을 morekids 변수만을 사용해 분석하고 내생성 문제를 고려하지 않은 분석이 된다. 단, 이 경우에도 age와 race는 control variables로 고려하였다. variance-robust 추정량을 얻기 위해 일반적인 lm() function이 아닌 lm_robust() function을 사용하였다.

```
lm_q11 <- lm_robust(weeksm1 ~ morekids + agem1 + black + hispan + othrace, data = raw_data)
summary(lm_q11)
```

```
##
## Call:
## lm_robust(formula = weeksm1 ~ morekids + agem1 + black + hispan +
##     othrace, data = raw_data)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper     DF
## (Intercept)  -4.8345    0.36735 -13.161 1.525e-39  -5.5545  -4.1145 254648
## morekids     -6.2304    0.08624 -72.246 0.000e+00  -6.3994  -6.0614 254648
## agem1         0.8379    0.01212  69.144 0.000e+00   0.8141   0.8616 254648
## black        11.6642    0.19553  59.653 0.000e+00  11.2810  12.0475 254648
## hispan        0.4661    0.18071   2.579 9.902e-03   0.1119   0.8203 254648
## othrace       2.1421    0.20828  10.285 8.354e-25   1.7339   2.5504 254648
##
## Multiple R-squared:  0.04376 ,   Adjusted R-squared:  0.04374
## F-statistic:  2545 on 5 and 254648 DF,  p-value: < 2.2e-16
```

```
lm_q11$p.value
```

```
##  (Intercept)     morekids        agem1        black      hispan      othrace
## 1.524745e-39 0.000000e+00 0.000000e+00 0.000000e+00 9.901915e-03 8.354171e-25
```

다음으로 IV estimator를 사용한다. IV estimation을 위해 AER packages의 ivreg() 함수를 사용하였다.

```
lm_q12 <- ivreg(weeksm1 ~ morekids + agem1 + black + hispan + othrace | samesex + agem1 + black + hispan + othra
summary(lm_q12)
```

```
##
## Call:
## ivreg(formula = weeksm1 ~ morekids + agem1 + black + hispan +
##     othrace | samesex + agem1 + black + hispan + othrace, data = raw_data)
##
## 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
```

앞서의 분석에서는 내생성 문제가 확실히 크게 존재했음을 t value를 비교하여 파악할 수 있다. IV Estimator를 활용한 경우에도 인과관계가 존재한다고 결론지을 수 있지만, p-value가 3e-06 수준으로 앞서의 극단적 수치에 비해서는 크게 낮아졌다.

Q3.2

Data importing part에서 두 변수를 추가한 q32_data dataset을 사용하여 분석을 진행한다.

반응변수 Y : 어머니의 노동공급 weeksm1 독립변수 X : 2명 넘는 아이 갖기 morekids (내생성 문제 존재)
X를 포착하기 위한 도구변수 Z : twoboys, twogirls (morekids = constant + b1twoboys + b2*twogirls + errors)
control variables : agem1(출산 연령), black, hispan, othrace (인종 dummy)

```
lm_q21 <- ivreg(weeksm1 ~ morekids + agem1 + black + hispan + othrace | twoboys + twogirls + agem1 + black + hisp
summary(lm_q21)
```

```
##
## Call:
## ivreg(formula = weeksm1 ~ morekids + agem1 + black + hispan +
##     othrace | twoboys + twogirls + agem1 + black + hispan + othrace,
##     data = q32_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -36.07 -17.89 -11.02  22.81  44.84
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.75132    0.40571 -11.711  < 2e-16 ***
```

```
## morekids    -5.43131    1.21859  -4.457 8.31e-06 ***
## agem1        0.82561    0.02253  36.641  < 2e-16 ***
## black       11.58427    0.22745  50.931  < 2e-16 ***
## hispan       0.34524    0.25685   1.344    0.179
## othrace      2.12033    0.20576  10.305  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.39 on 254648 degrees of freedom
## Multiple R-Squared: 0.04345, Adjusted R-squared: 0.04343
## Wald test:  1335 on 5 and 254648 DF,  p-value: < 2.2e-16
```

여전히 유의하다. 이와 같은 약간의 차이가 발생하는 것은 두 개의 모수 `b1`, `b2`를 모두 추정하게 되어 각 추정의 정확성이 낮아진 것으로 보인다.