Assignment 4

Foodies with hoodies

Contents

```r
# load packages
if(!require(pacman)){install.packages("pacman")}

p_load(devtools,tidyverse,dplyr,ggplot2,latex2exp,
       sampleSelection, quantreg, plm, nlme, knitr,car, ivreg,stargazer)
```

1 Question 1

Judge	Jones		Smith	
Sentences	Prison	Other	Prison	Other
Cases	70%	30%	40%	60%
Future arrests	40%	60%	20%	50%

1.1 (i)

We can treat the following problem as follows: Y_i is the outcome of whether an individual is arrested later. The instrument variable Z_i is which judge they are assigned to in the first case and D_i is the treatment whether the individual is sentenced to prison or not in the first case. Then the Wald estimator can be calculated by:

$$\delta_{Wald} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{Pr(D_i = 1|Z_i = 1) - Pr(D_i = 1|Z_i = 0)} \tag{1}$$

1.2 (ii)

1.3 (iii)

2 Question 2

2.1 (i)

From what is given, we have $MDE = 0.1$, the power $p = 0.7$, the proportion of students in control group is $p = 0.5$. The variance of the binomial variable is $\sigma^2 = p(1-p) = 0.25$ To get the number of students the teacher should include in the experiment, we use the following formula:

$$
\begin{aligned}
n &= \left(\frac{t_{1-\alpha/2} - t_{1-q}}{MDE}\right)^2 \frac{\sigma^2}{p(1-p)} \\
&= \left(\frac{1.960 + 0.524}{0.1}\right)^2 \frac{0.25}{0.5(1-0.5)} \\
&\approx 617
\end{aligned} \tag{2}
$$

Thus, the teacher should include at least 617 students in the experiment.

2.2 (ii)

This will change the proportion of students in treatment to $p = 0.5 \times 20\% = 0.1$, using the formula in Equation (2), the number of students required to participate in the experiment is:

$$
\begin{aligned}
n &= \left(\frac{1.960 + 0.524}{0.1}\right)^2 \frac{0.25}{0.1(1-0.1)} \\
&\approx 1713
\end{aligned} \tag{3}
$$

Thus, the number of students required to participate in the experiment increases by $6856 - 2468 = 4388$ students.

3 Question 3

3.1 (i)

```
# Load data
dfData = read.csv("AngristEvans80.csv")
attach(dfData)

# Fraction of girls among the first born child
count_girl1 = table(dfData$SEXK)
fraction_girl1= count_girl1[[2]]/(count_girl1[[1]]+count_girl1[[2]])

# Fraction of girls among the second born child
```

```r
count_girl2 = table(dfData$SEX2ND)
fraction_girl2= count_girl2[[2]]/(count_girl2[[1]]+count_girl2[[2]])

cat("Fraction of girls among the first born child is:
 ↪  ",fraction_girl1,"\n", "Fraction of girls among the second born
 ↪  child is: ",fraction_girl2)
```

```
Fraction of girls among the first born child is:  0.4876463
 Fraction of girls among the second born child is:  0.4884266
```

```r
#Regress gender of second child on gender of first child
lm_second_first = lm(SEX2ND~SEXK, data = dfData)
summary(lm_second_first)
```

```
Call:
lm(formula = SEX2ND ~ SEXK, data = dfData)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4908 -0.4862 -0.4862  0.5092  0.5138

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.4861744  0.0008672 560.626   <2e-16 ***
SEXK        0.0046185  0.0012418   3.719    2e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4999 on 648470 degrees of freedom
Multiple R-squared:  2.133e-05, Adjusted R-squared:  1.979e-05
F-statistic: 13.83 on 1 and 648470 DF,  p-value: 2e-04
```

3.2 (ii)

```r
# First stage regression
lm_first_stage = lm(CHILD3 ~ SAMESEX, data= dfData)
summary(lm_first_stage)
```

```
Call:
lm(formula = CHILD3 ~ SAMESEX, data = dfData)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4093 -0.4093 -0.3552  0.5907  0.6448

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.3552366  0.0008544  415.79   <2e-16 ***
SAMESEX     0.0540534  0.0012051   44.85   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4852 on 648470 degrees of freedom
Multiple R-squared:  0.003093,	Adjusted R-squared:  0.003091
F-statistic:  2012 on 1 and 648470 DF,  p-value: < 2.2e-16
```

Is the instrumental variable sufficiently strong? => yes

```r
# Regress number of children on whether the first two children have the
↪  same gender
lm_total = lm(KIDCOUNT ~ SAMESEX, data= dfData)
summary(lm_total)
```

```
Call:
lm(formula = KIDCOUNT ~ SAMESEX, data = dfData)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5752 -0.5752 -0.5040  0.4248  9.4960

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.504033   0.001458 1716.93   <2e-16 ***
SAMESEX     0.071200   0.002057   34.61   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8283 on 648470 degrees of freedom
Multiple R-squared:  0.001844,	Adjusted R-squared:  0.001842
F-statistic:  1198 on 1 and 648470 DF,  p-value: < 2.2e-16
```

3.3 (iii)

In this study, the treatment group includes those who have a third child and the control group includes those who have two children or less. The variables that affect decision for mothers to be assigned into treatment or control group is $Z = SAMESEX$, indicating whether the first two child are of the same sex or not.

The always takers are those who have a third child regardless of whether the first two children is of the same sex or not.

```
df_always = dfData[dfData$CHILD3 == 1 & dfData$SAMESEX == 0,]
cat("The share of always takers is: ", nrow(df_always)/nrow(dfData))
```

```
The share of always takers is:  0.1766969
```

The compliers are those who only have a third child if the first two kids are of the same sex.

```
df_compliers1 = dfData[dfData$CHILD3 == 1 & dfData$SAMESEX == 1,]
df_compliers0 = dfData[dfData$CHILD3 == 0 & dfData$SAMESEX == 0,]
cat("The share of compliers is: ",
↪   (nrow(df_compliers1)+nrow(df_compliers0))/nrow(dfData))
```

```
The share of compliers is:  0.5264159
```

The never takers are those who will never have the third child regardless of whether the first two chidren are of the same sex or not.

```
df_never = dfData[dfData$CHILD3 == 0 & dfData$SAMESEX == 1,]
cat("The share of never takers is: ", nrow(df_never)/nrow(dfData))
```

```
The share of never takers is:  0.2968871
```

Lastly, the defiers are those who will have a third child if the first two kids are of different sexes and will not have a third child if the first two kids are of the same sex. We cannot observe this as they are divided among the always taker and never taker's group.

3.4 (iv)

```r
iv_reg_hour <- ivreg(HOURSM ~ CHILD3 | SAMESEX, data = dfData)
iv_reg_income <- ivreg(INCOME1M ~ CHILD3 | SAMESEX, data = dfData)

stargazer(iv_reg_hour, iv_reg_income,
          type="text", report="vc*stp",
          keep.stat=c("n","adj.rsq"),
          title = "...") #remember to change title
```

```
...
=========================================
                  Dependent variable:
             ----------------------------
                 HOURSM        INCOME1M
                  (1)             (2)
-----------------------------------------
CHILD3         -3.585***     -786.830***
                (0.864)        (244.939)
              t = -4.150     t = -3.212
             p = 0.00004      p = 0.002

Constant       20.304***    3,825.464***
                (0.331)         (93.899)
              t = 61.311     t = 40.740
              p = 0.000       p = 0.000

-----------------------------------------
Observations    648,472        648,472
Adjusted R2      0.007          0.008
=========================================
Note:         *p<0.1; **p<0.05; ***p<0.01
```

3.5 (v)

```r
# Subgroup 1: Always taker
hour1=mean(df_always$HOURSM)
income1=mean(df_always$INCOME1M)
```

```
cat("The mean working hour of always takers is: ", hour1, ", the mean
↪  income of always takers is: ",income1)
```

```
The mean working hour of always takers is:  17.04711 , the mean income of always takers is:
```

```
# Subgroup 2: never takers
hour2=mean(df_never$HOURSM)
income2=mean(df_never$INCOME1M)
cat("The mean working hour of never takers is: ", hour2, ", the mean
↪  income of never takers is: ",income2)
```

```
The mean working hour of never takers is:  20.20379 , the mean income of never takers is:  39
```

```
# Subgroup 3: complier 1
hour3=mean(df_compliers1$HOURSM)
income3=mean(df_compliers1$INCOME1M)
cat("The mean working hour of complier in treatment group is: ", hour3,
↪  ", the mean income of this group is: ",income3)
```

```
The mean working hour of complier in treatment group is:  16.8629 , the mean income of this g
```

```
# Subgroup 4: complier 0
hour4=mean(df_compliers0$HOURSM)
income4=mean(df_compliers0$INCOME1M)
cat("The mean working hour of complier in control group is: ", hour4, ",
↪  the mean income of this group is: ",income4)
```

```
The mean working hour of complier in control group is:  20.12279 , the mean income of this gr
```

To-dos: USE these means to say something about the preference of having a third child

3.6 (vi)

3.7 (vii)

First, we stratify the sample by gender of the first child:

```r
df_first_girl = dfData[dfData$SEXK == 1,]
df_first_boy = dfData[dfData$SEXK == 0,]
```

(But they ask to use the first stage result?) I try to to it manually below:

```r
# First stage regression
lm_1st_girl = lm(CHILD3 ~ SAMESEX, data=df_first_girl)
lm_1st_boy = lm(CHILD3 ~ SAMESEX, data=df_first_boy)
stargazer(lm_1st_girl, lm_1st_boy,
          type="text", report="vc*stp",
          keep.stat=c("n","adj.rsq"),
          title = "...") #remember to change title
```

```
...
=========================================
                 Dependent variable:
             ----------------------------
                       CHILD3
                  (1)            (2)
-----------------------------------------
SAMESEX         0.063***      0.046***
                (0.002)        (0.002)
               t = 36.381    t = 27.351
               p = 0.000      p = 0.000

Constant        0.355***      0.356***
                (0.001)        (0.001)
              t = 293.159    t = 294.887
               p = 0.000      p = 0.000

-----------------------------------------
Observations    316,225        332,247
Adjusted R2      0.004          0.002
=========================================
Note:         *p<0.1; **p<0.05; ***p<0.01
```

If the first child is a girl and first two children are of the same sex, one is more likely to have a third child.

Then, we perform instrumental variable regressions:

```r
iv_reg_girl_hour <- ivreg(HOURSM ~ CHILD3 | SAMESEX, data =
 ↪  df_first_girl)
iv_reg_girl_income <- ivreg(INCOME1M ~ CHILD3 | SAMESEX, data =
 ↪  df_first_girl)
iv_reg_boy_hour <- ivreg(HOURSM ~ CHILD3 | SAMESEX, data = df_first_boy)
iv_reg_boy_income <- ivreg(INCOME1M ~ CHILD3 | SAMESEX, data =
 ↪  df_first_boy)
stargazer(iv_reg_girl_hour,
 ↪  iv_reg_girl_income,iv_reg_boy_hour,iv_reg_boy_income,
          type="text", report="vc*stp",
          keep.stat=c("n","adj.rsq"),
          title = "...") #remember to change title
```

```
...
==============================================================
                           Dependent variable:
             -------------------------------------------------
               HOURSM     INCOME1M      HOURSM      INCOME1M
                (1)         (2)           (3)         (4)
--------------------------------------------------------------
CHILD3         -1.104     -343.922    -6.695***  -1,320.535***
               (1.064)    (303.245)    (1.425)     (400.868)
             t = -1.037  t = -1.134  t = -4.698   t = -3.294
             p = 0.300   p = 0.257   p = 0.00001  p = 0.001

Constant     19.409***  3,676.891***  21.424***  4,006.674***
              (0.412)    (117.357)    (0.541)     (152.305)
             t = 47.115  t = 31.331  t = 39.567   t = 26.307
             p = 0.000   p = 0.000    p = 0.000   p = 0.000

--------------------------------------------------------------
Observations  316,225     316,225      332,247     332,247
Adjusted R2    0.004       0.005       -0.001       0.007
==============================================================
Note:                              *p<0.1; **p<0.05; ***p<0.01
```

Here we see that if the first child is a girl, having a third child does not significantly influence the hour and income.