

# Assignment 4

Group name: Foodies with hoodies

## Contents

<b>1</b>	<b>Question 1</b>	<b>2</b>
1.1	(i) . . . . .	2
1.2	(ii) . . . . .	2
1.3	(iii) . . . . .	3
<b>2</b>	<b>Question 2</b>	<b>3</b>
2.1	(i) . . . . .	3
2.2	(ii) . . . . .	3
<b>3</b>	<b>Question 3</b>	<b>4</b>
3.1	(i) . . . . .	4
3.2	(ii) . . . . .	5
3.3	(iii) . . . . .	7
3.4	(iv) . . . . .	8
3.5	(v) . . . . .	9
3.6	(vi) . . . . .	10
3.7	(vii) . . . . .	10

```
# 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 Sentences	Jones		Smith	
	Prison	Other	Prison	Other
Cases	70%	30%	40%	60%
Future arrests (of those who were sentenced to prison in the first case)	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. Assuming monotonicity, the Wald estimator can be calculated by:

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

If we suppose that  $Z_i = 1$  for Judge Jones, and  $Z_i = 0$  for Judge Smith, we can calculate the Wald estimator in the following way:

$$\delta_{Wald} = \frac{0.740 - 0.420}{70 - 40} = \frac{2}{3} \quad (2)$$

### 1.2 (ii)

In this case, the interpretation of the estimated effect of 0.667 is that the treatment difference between the groups leads to 66.7% increase in the chance that someone who has been sentenced to prison by Judge Jones will be arrested and sent to prison again compared to the ones sentenced by Judge Smith. However, we only examine the ones who have been involved in a case and sentenced to prison, so this only applies to the part of the population

who have been arrested at least once (70% in the case of Judge Jones and 40% in the case of Judge Smith).

### 1.3 (iii)

In this case, the always takers of the group are the ones who would be caught and arrested later no matter if they would have been sentenced to prison or not in their first case. That means that the portion of the population who are arrested in the future and sent to prison, which is 20% in the case of Judge Smith and 40% in the case of Judge Jones. We do not need to know how many of these were sent to prison in their first case and how many were not, because they are the ones who would be sent to prison one way or another.

## 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:

$$\begin{aligned} n &= \left( \frac{t_{1-\alpha/2} - t_{1-q}}{MDE} \right)^2 \frac{\sigma^2}{p(1 - p)} \\ &= \left( \frac{1.960 + 0.524}{0.1} \right)^2 \frac{0.25}{0.5(1 - 0.5)} \\ &\approx 617 \end{aligned} \tag{3}$$

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 (3), the number of students required to participate in the experiment is:

$$n = \left( \frac{1.960 + 0.524}{0.1} \right)^2 \frac{0.25}{0.1(1 - 0.1)} \approx 1713 \quad (4)$$

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
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
```

```
#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

```

To briefly comment on the results, the output shows a relatively similar proportion of genders among first and second babies as well. This is relatively expected since we did not manipulate the composition of the data and the natural trend is that there are slightly more male than female children born in the population. Furthermore the regression results imply that the gender of the first child is a strong predictor of the gender of the second child, namely if the first child is a girl, there is a stronger probability that the second child will also be a girl. This outcome is slightly more interesting, showing that the outcome of the gender of the first and second children are not necessarily independent.

### 3.2 (ii)

```

# 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

The instrumental variable is sufficiently strong, because the variable *SAMESEX* has a strongly significant explanatory power over whether the third child is born. This means that the instrumental variable (IV) can be used in a 2SLS model provided that it is not correlated with the second stage dependent variable.

```
# 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

The first model already implied that the *SAMESEX* is a strong instrument for whether a third child is born. Continuing along this logic, we also find that the IV is a strong instrument for the

total number of children, implying that the “treatment” of having same sex children for the first two kids significantly increases the number of children and in particular the chance of a third child being born.

### 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 children 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)

```
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
```

Conducting the IV regression with estimating the *CHILD3* variable with *SAMESEX*, we find that having a third child significantly reduces (1) the mother's weekly hours of work (*HOURSM*) by an approximate 3.585 hours and that in model (2) significantly reduces the mother's annual income as well (*INCOME1M*), by an estimated 786.8 \$ per year.



### 3.5 (v)

```
# Subgroup 1: Always taker
hour1=mean(df_always$HOURLSM)
income1=mean(df_always$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$HOURLSM)
income2=mean(df_never$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: 3

```
# Subgroup 3: complier 1
hour3=mean(df_compliers1$HOURLSM)
income3=mean(df_compliers1$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$HOURLSM)
income4=mean(df_compliers0$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 g

We can use these means to confirm our results from Section 3.4, namely that having a third child will reduce the annual income and weekly hours of the mother. We can also see the magnitude differences being relatively similar to the estimated effects. Compliers also change along with their decision, compliers in the control group (only 2 children with different genders) have similar group means to the never takers (always 2 children regardless of the gender being the same), while compliers in the treatment group (first two children same gender, hence had a third child) resemble to the always takers group. Therefore, this implies that it is not necessarily the treatment but rather simply the preference to have a third child or not will be important for the dependent variables in the model (and this might also imply that the treatment variable is a good instrument due to seemingly no correlation with the dependent variables).

### 3.6 (vi)

This selection does not concur with a potential bias in the OLS regression when regressing the labour market outcome variables on having a third child. It seems that the treatment groups in having a third child or not do not differ too much, implying that even though we may find a good instrument on the *CHILD3* variable, since the two groups without third child and the two groups with third child have relatively similar outcome means respectively, simply an OLS model with  $HOURS_M \sim CHILD3$  or  $INCOME1M \sim CHILD3$  would not have an underlying heterogeneity bias or bias due to omitting *SAMESEX*.

### 3.7 (vii)

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

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

Then we conduct a simple OLS regression for the would-be first stage, similarly to Section 3.2 on the subset of observations where the first child is female and the other subset where the first child is male.

```
# 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 = "Subset first stage OLS on (1) first child girl and
↪ (2) first child boy")
```

Subset first stage OLS on (1) first child girl and (2) first child boy

```
=====
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
```

Interestingly, although we find that the treatment is significant in predicting the probability of a third child for both subgroups, 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. This implies that the parents with two girls as first children have a stronger preference for a third child, while parents with two boys as first children, albeit still more likely to have a third child, have a weaker preference. Subsequently, we perform instrumental variable regressions:

```
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:				
	HOURLSM	INCOME1M	HOURLSM	INCOME1M
	(1)	(2)	(3)	(4)
CHILD3	-1.104 (1.064) t = -1.037 p = 0.300	-343.922 (303.245) t = -1.134 p = 0.257	-6.695*** (1.425) t = -4.698 p = 0.00001	-1,320.535*** (400.868) t = -3.294 p = 0.001
Constant	19.409*** (0.412) t = 47.115 p = 0.000	3,676.891*** (117.357) t = 31.331 p = 0.000	21.424*** (0.541) t = 39.567 p = 0.000	4,006.674*** (152.305) t = 26.307 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 find very interesting results. We observe that if the first child is a girl, having a third child does not significantly influence the hour and income, while if the first child is a boy, having a third child will significantly reduce both working hours and annual income, even more than previously in Section 3.4, with an estimated 6.695 hours of decrease in weekly working hours and an estimated 1320.535 \$ decrease in annual income. Therefore these results hint at strong heterogeneity in the effect of having a third child on labour statistics in terms of the gender of the children.