Bookmarks

Module 12: Endogeneity, Instrumental Variables, and Experimental Design > Module 12: Homework > 2SLS Estimates: Questions 1 - 9

2SLS Estimates: Questions 1 - 9

Bookmark this page

In this problem, we are going to replicate part of the results of Joshua Angrist and William Evans' article *"Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size."* Here is the abstract of the study:

> Research on the labor-supply consequences of childbearing is complicated by the endogeneity of fertility. This study uses parental preferences for a mixed sibling-sex composition to construct instrumental variables (IV) estimates of the effect of childbearing on labor supply. IV estimates for women are significant but smaller than ordinary least-squares estimates. The IV are also smaller for more educated women and show no impact of family size on husbands' labor supply. A comparison of estimates using sibling-sex composition and twins instruments implies that the impact of a third child disappears when the child reaches age 13. (JEL J13, J22)

The purpose of this exercise is to study how fertility affects female labor supply. In order to do this, we are going to compare female labor supply in households with two children versus households with three children. Since fertility decisions are endogenous, we are going to use two sets of instruments: whether there is a multiple pregnancy in the second pregnancy and sex composition of the first two children. This latter instrument was the one proposed by Angrist & Evans (1998). Intuitively, parents are more likely to have a third child when the first two have the same sex. Assuming that whether the first two children have the same sex is random, we can use this variable as an instrument for the number of children in the household.

We have provided you with the data set census80.csv that corresponds to an extract of the 1980 US Census. It has been restricted to the set of families with two or three children and with mother's age between 21 and 35 years. The data set contains the following variables:

- **workedm**: whether the mother works.
- **weeksm**: number of weeks the mother works.
- **whitem**: mother is White.
- **blackm**: mother is Black.
- **hispm**: mother is Hispanic.
- **othracem**: mother is of other race.
- **sex1st**: sex of the first child (0 corresponds to male and 1 to female).
- **sex2nd**: sex of the second child (0 corresponds to male and 1 to female).
- **ageq2nd**: age in quarters of the second child.
- **ageq3rd**: age in quarters of the third child.
- **numberkids**: number of children in the household.

Load the data into R, follow our instructions, and answer the following questions.

Question 1

1.0/1.0 point (graded)

Use the command **summary** to summarize the variables in the data. Using your output, fill in the following information:

Please round all answers to the second decimal place, i.e. if the answer is 6.6728, round to 6.67 and if it is 6.6788, round to 6.68

a. Fraction of mothers that work:

Endogeneity and Instrumental Variables
Finger Exercises due Dec 14, 2016 05:00 IST

Experimental Design
Finger Exercises due Dec 14, 2016 05:00 IST

Module 12: Homework
Homework due Dec 12, 2016 05:00 IST

Exit Survey

0.57 ✓ **Answer:** 0.57

0.57

b. 3rd quartile of weeks worked:

48.00 ✓ **Answer:** 48.00

48.00

c. Proportion of Hispanic mothers:

0.03 ✓ **Answer:** 0.03

0.03

d. Median age of the second child in quarters:

19.00 ✓ **Answer:** 19.00

19.00

Explanation

The following code in R:

```
#Preliminaries
#---------------------------------------------------------
library(car)
library(rdd)
setwd("/Users/raz/Dropbox/14.31 edX Building the Course/Problem Sets/PSET 12")

# I. DiD Estimations
#---------------------------------------------------------
rm(list = ls())
census80 <- read.csv('census80.csv')
sumamry(census80)
```

Produces the following output:

```
   workedm           weeksm          whitem           blackm           hispm
Min.   :0.0000   Min.   : 0.00   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
1st Qu.:0.0000   1st Qu.: 0.00   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.00000
Median :1.0000   Median :12.00   Median :1.0000   Median :0.0000   Median :0.00000
Mean   :0.5716   Mean   :20.82   Mean   :0.8314   Mean   :0.1125   Mean   :0.02725
3rd Qu.:1.0000   3rd Qu.:48.00   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.00000
Max.   :1.0000   Max.   :52.00   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000

   othracem           sex1st           sex2nd          ageq2nd         ageq3rd
Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   : 0.00   Min.   : 0.00
1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 9.00   1st Qu.: 5.00
Median :0.00000   Median :0.0000   Median :0.0000   Median :19.00   Median :13.00
Mean   :0.02886   Mean   :0.4871   Mean   :0.4881   Mean   :21.75   Mean   :16.59
3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:33.00   3rd Qu.:26.00
Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :71.00   Max.   :67.00
                                                                    NA's   :305132
  numberkids
Min.   :2.000
1st Qu.:2.000
Median :2.000
Mean   :2.286
3rd Qu.:3.000
Max.   :3.000
```

Submit You have used 1 of 2 attempts

Question 2

1.0/1.0 point (graded)

Use the variable **ageq2nd** and the variable **ageq3rd** to construct an indicator variable on whether there was a multiple pregnancy during the mother's second pregnancy. What is the proportion of households with a multiple pregnancy in the second pregnancy?

Please round your answer to the fourth decimal place, i.e. if your answer is 0.12435, please round to 0.1244, and if it is 0.12433, please round to 0.1243

0.0073 ✓ **Answer:** 0.0073

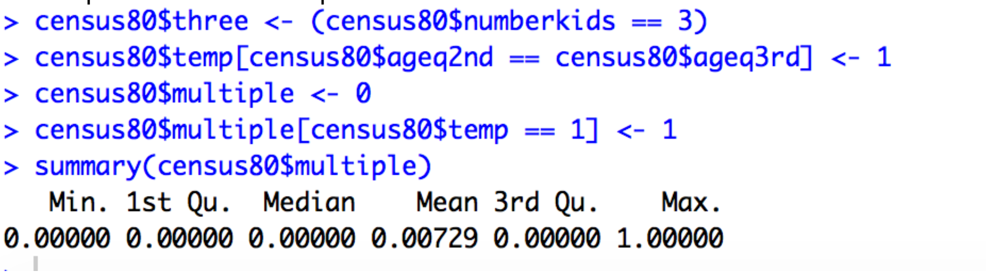
0.0073

Explanation

We can construct this indicator by running the following code:

```
census80$temp[census80$ageq2nd == census80$ageq3rd] <- 1
census80$multiple <- 0
census80$multiple[census80$temp == 1] <- 1
summary(census80$multiple)
```

which produces this output:



```
> census80$three <- (census80$numberkids == 3)
> census80$temp[census80$ageq2nd == census80$ageq3rd] <- 1
> census80$multiple <- 0
> census80$multiple[census80$temp == 1] <- 1
> summary(census80$multiple)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
0.00000 0.00000 0.00000 0.00729 0.00000 1.00000 
```

Submit You have used 1 of 2 attempts

Question 3

1.0/1.0 point (graded)

Use the variables **sex1st** and **sex2nd** to construct an indicator variable on whether the first and the second born children have the same sex. What is the proportion of households in which the first two children have the same sex?

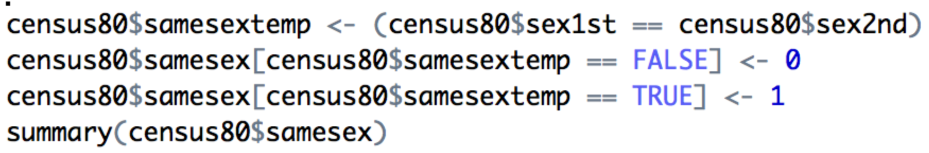
Please provide your answer to the fourth decimal place, i.e. exactly how it appears in the output

0.5019 ✓ **Answer:** 0.5019

0.5019

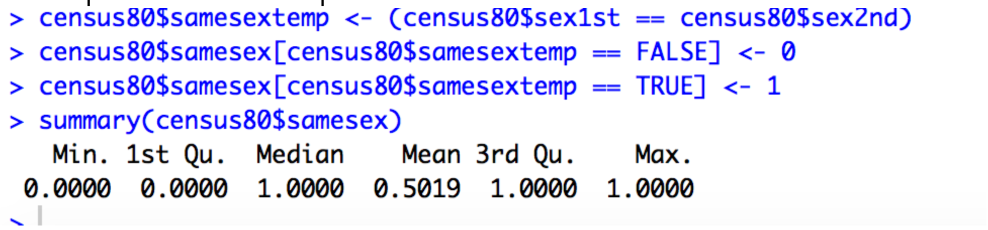
Explanation

We can construct this indicator by running the following code:



```
census80$samesextemp <- (census80$sex1st == census80$sex2nd)
census80$samesex[census80$samesextemp == FALSE] <- 0
census80$samesex[census80$samesextemp == TRUE] <- 1
summary(census80$samesex)
```

which produces this output:



```
> census80$samesextemp <- (census80$sex1st == census80$sex2nd)
> census80$samesex[census80$samesextemp == FALSE] <- 0
> census80$samesex[census80$samesextemp == TRUE] <- 1
> summary(census80$samesex)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  1.0000  0.5019  1.0000  1.0000 
> 
```

Submit You have used 1 of 2 attempts

Now, let's set up the model we want to estimate. In particular we are interested in estimating the following equation:

$$labor\ supply_h = \alpha_0 + \alpha_1 \mathbf{1}_{3\ children_h} + \alpha_2 black\ mother_h + \alpha_3\ hispanic\ mother_h + \alpha_4 other\ race_h + \varepsilon_h$$

(equation 1)

where $labor\ supply_h$ corresponds to a labor supply variable of the mother in household h, $\mathbf{1}_{3\ children_h}$ is an indicator on whether there are three children born in the households, and the other variables correspond to the race categories of the mother; finally, ε_h corresponds to an error term.

Question 4

1/1 point (graded)

Run this model through OLS using whether the mom works and the number of weeks she works as the dependent variables. According to your estimates, which of the following statements are correct? (Select all that apply)

☑ According to the OLS estimates, having a third child reduces the likelihood that the mother works by 8.39 percentage points.

□ According to the OLS estimates, having a third child reduces the likelihood that the mother works by 3.94 percentage points.

□ According to the OLS estimates, having a third child reduces the number of weeks a mother decides to work by 8.39 weeks.

☑ According to the OLS estimates, having a third child reduces the number of weeks a mother decides to work by 3.94 weeks.

✓

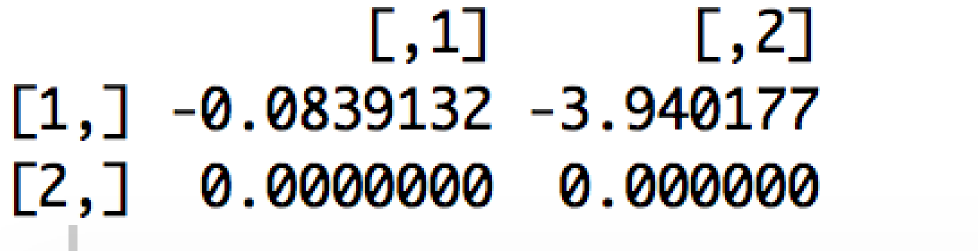
Explanation

If we run the following code in R:

```
#OLS models
#----------------------------------------------------------
ols1 <- lm(workedm ~ three + blackm + hispm + othracem, data = census80)
OLS[1, 1] <- ols1$coefficients[2]
pvalue <- summary(ols1)
OLS[2, 1] <- pvalue$coefficients[2, 4]

ols2 <- lm(weeksm ~ three + blackm + hispm + othracem, data = census80)
OLS[1, 2] <- ols2$coefficients[2]
pvalue <- summary(ols2)
OLS[2, 2] <- pvalue$coefficients[2, 4]
OLS
```

This is the output that we get:



```
           [,1]      [,2]
[1,] -0.0839132 -3.940177
[2,]  0.0000000  0.000000
```

The first column corresponds to the model where the dependent variable is whether the mother works or not. In the second column, the dependent variable is the number of weeks she worked.

Submit You have used 2 of 2 attempts

✓ Correct (1/1 point)

Since fertility is an endogenous variable, we want to use the multiple pregnancy and the same sex variables as instruments for having three children in the household. We are going to estimate the first-stage using each variable separately. Run a regression for each of these instruments using as a dependent variable the indicator of having three children and controlling for the race of the mother.

Question 5

1/1 point (graded)

According to your estimates, by having a multiple pregnancy during the second pregnancy, the likelihood of having a third child increases by how many percentage points?

Please round your answer to the second decimal place, i.e. if your answer is 51.2322, please round to 51.23, and if it is 51.2382, please round to 51.24

71.79 ✔ **Answer:** 71.79

71.79

Explanation

You should run a regression of the following model:

$$\mathbf{1}_{3\ children_h} = \beta_0 + \beta_1 multiple_h + \beta_2 black\ mother_h + \beta_3\ hispanic\ mother_h + \beta_4 other\ race_h + \nu_h$$

(equation 2)

The solution to Question 6 provides the correct code and output in R. The output should show that having a multiple pregnancy at the second pregnancy increases the likelihood of having a third child by 71.79 percentage points.

Submit You have used 2 of 2 attempts

✔ Correct (1/1 point)

Question 6

1/1 point (graded)

According to your estimates, when the first two children are of the same sex, the likelihood of having a third child increases by how many percentage points?

Please round your answer to the third decimal place, i.e. if your answer is 7.7283, round to 7.728 and if it is 7.7288, please round to 7.729.

4.902 ✓ **Answer:** 4.902

4.902

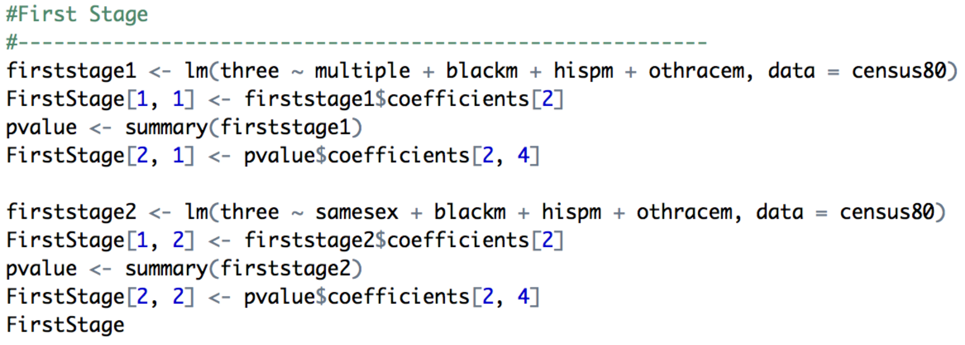
Explanation

You should run a regression of the following model:

$$\mathbf{1}_{3\ children_h} = \beta_0 + \beta_1 same\ sex_h + \beta_2 black\ mother_h + \beta_3\ hispanic\ mother_h + \beta_4 other\ race_h + \nu_h$$

(equation 3)

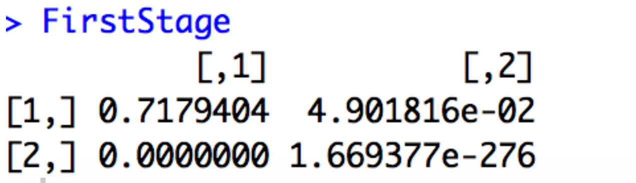
This code in R:



```
#First Stage
#-----------------------------------------------------------
firststage1 <- lm(three ~ multiple + blackm + hispm + othracem, data = census80)
FirstStage[1, 1] <- firststage1$coefficients[2]
pvalue <- summary(firststage1)
FirstStage[2, 1] <- pvalue$coefficients[2, 4]

firststage2 <- lm(three ~ samesex + blackm + hispm + othracem, data = census80)
FirstStage[1, 2] <- firststage2$coefficients[2]
pvalue <- summary(firststage2)
FirstStage[2, 2] <- pvalue$coefficients[2, 4]
FirstStage
```

Produces this output:



```
> FirstStage
          [,1]          [,2]
[1,] 0.7179404  4.901816e-02
[2,] 0.0000000 1.669377e-276
```

Thus, when first two children are of the same sex, the likelihood of a third child increases by 4.901816 percentage points.

Submit You have used 2 of 2 attempts

✓ Correct (1/1 point)

Question 7

1/1 point (graded)

Now, run the IV regression using whether the mother works as the dependent variable and multiple pregnancy as the instrument. According to this model, how does the likelihood that the mother works changes when a third child is born?

○ a. It increases by 6.412559 percentage points

● b. It decreases by 6.412559 percentage points ✓

○ c. It increases by 8.39132 percentage points

○ d. It decreases by 8.39132 percentage points

○ e. It increases by 9.8220536 percentage points

○ f. It decreases by 9.8220536 percentage points

Explanation

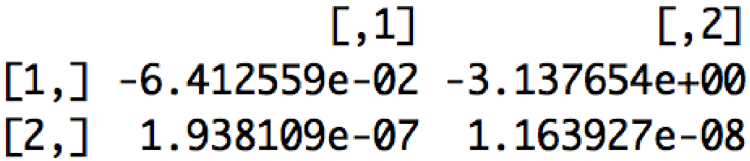
If we use multiple pregnancy at the second pregnancy variable as an instrument, then we find that having a third child decreases the likelihood that the mother works by 6.41 percentage points.

This code in R:

```
#IV model using multiple pregnancy
#-----------------------------------------------------------
iva1 <- ivreg(workedm ~ three + blackm + hispm + othracem |
                     blackm + hispm + othracem + multiple, data = census80)
IVa[1, 1] <- iva1$coefficients[2]
pvalue <- summary(iva1)
IVa[2, 1] <- pvalue$coefficients[2, 4]

iva2 <- ivreg(weeksm ~ three + blackm + hispm + othracem |
                blackm + hispm + othracem + multiple, data = census80)
IVa[1, 2] <- iva2$coefficients[2]
pvalue <- summary(iva2)
IVa[2, 2] <- pvalue$coefficients[2, 4]
IVa
```

Produces this output:



```
              [,1]          [,2]
[1,] -6.412559e-02 -3.137654e+00
[2,]  1.938109e-07  1.163927e-08
```

Submit You have used 1 of 2 attempts

✓ Correct (1/1 point)

Question 8

1/1 point (graded)

Now, run the IV regression using whether the mother works as the dependent variable and the same-sex variable as the instrument. According to this model, how does the likelihood that the mother works change when a third child is born?

○ a. It increases in 6.412559 percentage points

○ b. It decreases in 6.412559 percentage points

○ c. It increases in 8.39132 percentage points

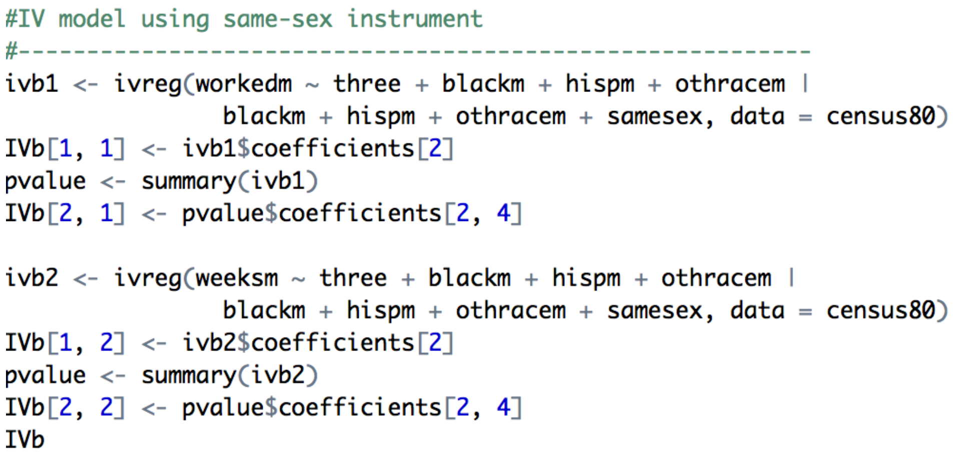
○ d. It decreases in 8.39132 percentage points

○ e. It increases in 9.8220536 percentage points

● f. It decreases in 9.8220536 percentage points ✓

Explanation

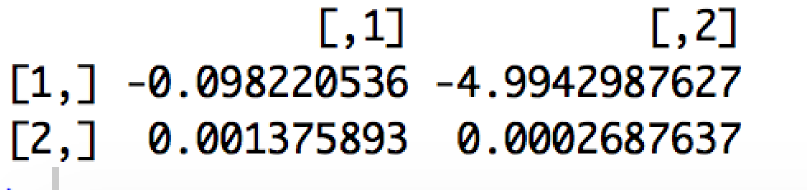
If we use the same-sex variable as the instrument, then we find that having a third child decreases the likelihood that the mother works by 9.82 percentage points. This code in R:



```
#IV model using same-sex instrument
#-----------------------------------------------------------
ivb1 <- ivreg(workedm ~ three + blackm + hispm + othracem |
                blackm + hispm + othracem + samesex, data = census80)
IVb[1, 1] <- ivb1$coefficients[2]
pvalue <- summary(ivb1)
IVb[2, 1] <- pvalue$coefficients[2, 4]

ivb2 <- ivreg(weeksm ~ three + blackm + hispm + othracem |
                blackm + hispm + othracem + samesex, data = census80)
IVb[1, 2] <- ivb2$coefficients[2]
pvalue <- summary(ivb2)
IVb[2, 2] <- pvalue$coefficients[2, 4]
IVb
```

Produces this output:



```
            [,1]          [,2]
[1,] -0.098220536 -4.9942987627
[2,]  0.001375893  0.0002687637
```

Submit You have used 1 of 2 attempts

✔ Correct (1/1 point)

Question 9

1/1 point (graded)

As you should see, the following relationship holds between the point estimates of the three strategies that we have used: $\hat{\alpha}_1^{IV-multiple} \leq \hat{\alpha}_1^{OSL} \leq \hat{\alpha}_1^{IV-same\ sex}$. Assuming a model of heterogeneous effects, what might explain these differences?

○ a. The instruments seem to be not valid since they show an opposite sign of the bias.

○ b. Women whose first two children are of the same sex are very different from women whose first two children are of different sex.

● c. IV estimates are local treatment effects. Thus, we are identifying the effect of fertility over women who have a third child when the relevant instrument changes. ✓

○ d. Women with a multiple pregnancy in the second pregnancy are very different than women with no-multiple pregnancy.

○ e. Fertility doesn't seem to be a relevant variable when women take labor supply decisions.

Explanation

As it was discussed in the lecture, under heterogeneous effects, IV estimates correspond to LATE (local average treatment effects). Thus, we are able to identify the average effect over the population that decides to have a third child when the instrument is switched on. This implies, that $\hat{\alpha}_1^{IV-multiple} = 0.06412559$ is the treatment effect on those that have a third child due to a multiple pregnancy. In general, for most of the population, having a multiple pregnancy would imply having a third child. On the other hand, $\hat{\alpha}_1^{IV-same\ sex} = 0.098220536$ corresponds to the treatment effect on those that decide to have a third child when the first two children have the same sex.

Submit You have used 1 of 2 attempts

✓ Correct (1/1 point)