Econometric Methods:
Solutions to Empirical Exercise 12.1
Chapter 12: Instrumental Variables Regression
Stock & Watson, 3^{rd} Edition

Zaeen de Souza [*] Deepti Goel[†]

Azim Premji University
12 March 2022

[*]

[†]Solution key prepared jointly by Zaeen and Deepti. R code and presentation in Rmarkdown by Zaeen.

Contents

Background: Empirical Exercise 11.1

These are the solutions to **E12.1** from **Chapter 12** of *Introduction to Econometrics (Updated Third edition)* by Stock & Watson. You should have the following on your computer in order to check answers/run the code and follow the questions in this assignment:

- An updated version of `R` and `Rstudio`.
- The following packages installed:
  - `ggplot2`
  - `readxl`
  - `fixest`
- The datasets called `fertility.xlsx`, `WeakInstrument.xlsx`
- The data description pdf to understand the variables being used.

Reading guide

All the code needed to complete the assignments is within this document. R code will be in a grey box and will look like this:

```
summary(iris)
```

And all R output i.e what R shows you once you run some code, will have `#` signs next to it, and will look like this:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.300   5.100   5.800   5.843   6.400   7.900
```

As far as possible these guides will show the **exact output** that comes from running code in `R`, and at times will use formatted tables made in latex. The results themselves, will be identical. Some things to note, that might make output look different across different computers:

- R reports things like p-values using scientific notation, but some computers report the numbers with many trailing zeros.
- If you have an old version of `R` or `Rstudio` it is highly recommended that you update it using the following code:

```
# Use this to update R from within RStudio
install.packages("installr")
library(installr)
# This last command, will open up a download prompt; choose yes/no accordingly.
updateR()
```

For updating `Rstudio`, un-install your version of `RStudio`, and download a fresh version from the `RStudio` website.

Loading the data and libraries

The following code sets the working directory to the folder where you have downloaded the data, loads the libraries needed fo the assignment and loads the excel dataset.

```
# Loading excel files
library(fixest)
library(readxl)
library(ggplot2)
library(dplyr)

# Setting working directory - this is unique to your computer
setwd("~Zaeen de Souza/Chapter 12 Instrument Variables, IV Weak Instruments")

# Loading the data as 'pset_data'
pset_data1 <- read_excel("fertility.xlsx")
pset_data2 <- read_excel("WeakInstrument.xlsx")
```

E12.1 Problem Context

How does fertility affect labor supply? That is, how much does a woman's labor supply fall when she has an additional child? In this exercise you will estimate this effect using data for married women from the 1980 U.S. Census.4 The data are available on the textbook website, www .pearsonglobaleditions.com/Stock_Watson, in the file Fertility and described in the file Fertility_Description. The data set contains information on married women aged 21–35 with two or more children.

Exercise E12.1

a. Regress weeks worked on the indicator variable morekids, using OLS. On average, do women with more than two children work less than women with two children? How much less?

The results are in table 1. We see that on average, mothers who had > 2 children, worked 5.39 weeks less than mothers who had ≤ 2 children.

```
model_1 <- feols(weeksm1 ~ morekids, data = pset_data1, vcov = "HC1")
```

Table 1: E.12.1 a

	Weeks Worked (1)
> 2 Children	-5.39***
	(0.087)
(Intercept)	21.1***
	(0.056)
R^2	0.014
Observations	254,654
F-test	3,696.022

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

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

The estimate of β_1, i,e $\widehat{\beta}_1$ is a biased estimate of the effect of having more kids on mother's work time for at least two reasons. First, there are omitted variables that are correlated with fertility decisions and weeks worked. Woman's education is one such variable: more educated mothers choose to have fewer children and may also choose to work more (or less). Second, there could be reverse causality. Women who work more may choose to have fewer children. For both these reasons $\widehat{\beta}_1$ is a biased estimate.

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? Is the effect large? Is it statistically significant?

On average, mothers whose first two children were of the same sex, were 6.8 percentage points more likely to have > 2 children compared to mothers with first two children of different gender. The result is significant at 1 percent level. The coefficient value of .068 is large as mothers with first two children of different sex have only a 34.6 percent chance of having more than two kids (0.068 is 20% of 0.346).

```
model_2 <- feols(morekids ~ samesex, data = pset_data1, vcov = "HC1")
```

Table 2: E.12.1 c

	$>$ 2 Children
	(1)
Same Sex	0.068***
	(0.002)
(Intercept)	0.346***
	(0.001)
R^2	0.005
Observations	254,654
F-test	1,237.219

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

d. Explain why samesex is a valid instrument for the instrumental variable regression of weeksworked on morekids.

For an instrument to be valid it need to satisfy two conditions.

1. Relevance: i.e $Cov(Z_i, X_i) \neq 0$
2. Exogeneity: i.e $Cov(Z_i, \varepsilon_i) = 0$

We can see from table 2 that the first condition is met—mothers whose two first children are of the same sex, are significantly more likely to have $>$ 2 children. The second condition cannot be tested directly, however it is plausible, in this setting (i.e the US) that the sex of the two firstborn children, only affects woman's labour supply via its effect on number of children.

e. Is samesex a weak instrument?

No, it is not. As can be seen in table 2, the first-stage F-statistic is 1,237.219 (table 2). As a general "rule of thumb", the F-stat of the first stage should be > 10 (Chapter 12, page 492, S&W). Here, the F-statistic is substantially larger.

f. Estimate the regression of weeksworked on morekids, using samesex as an instrument. How large is the fertility effect on labor supply?

We will estimate the following regression:

$$X_i = \alpha_1 + \beta_1 Z_i + \varepsilon_i \tag{1}$$

$$Y_i = \alpha_2 + \beta_1 \widehat{X}_i + \varepsilon_i \tag{2}$$

Where X is the endogenous variable that takes the value 1 if mother i had more than two children, Z is the IV for whether or not mother i has two first born children of the same sex, and Y is her labour supply measured in weeks worked in 1979.

The IV results are in table 3. Column 1 shows the OLS estimate whereas column 3, shows the IV estimate. The IV estimate is larger than the OLS estimate—for mothers whose fertility decisions were driven by exogenous variation in the sex of their first two children, having > 2 children reduced their labour supply by 6.3 weeks.

```
# syntax: Y ~ controls | endogenous var ~ IV
# we have no controls, so we just add a constant
IV <- feols(weeksm1 ~ 1 | morekids ~ samesex, pset_data1, vcov = "HC1")
```

Table 3: E.12.1 f

	Weeks Worked OLS (1)	>2 Children IV First-Stage (2)	Weeks Worked IV (3)
>2 Children	-5.39***		-6.31***
	(0.087)		(1.27)
Same Sex		0.068***	
		(0.002)	
(Intercept)	21.1***	0.346***	21.4***
	(0.056)	(0.001)	(0.487)
R^2	0.014	0.005	0.014
Observations	254,654	254,654	254,654
F-test	3,696.022	1,237.219	24.198

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

g. Do the results change when you include the variables agem1, black, hispan, and othrace in the labor supply regression (treating these variable as exogenous)? Explain why or why not.

The results of the 2*SLS* are in table 4. The estimates do change slightly, however, the change is not large. These results suggest that the covariates are not substantially correlated with our IV. This is also evident from the first-stage results of the 2SLS, which are in column (3) table 4—the coefficient is unchanged. The covariates can be used to be make the IV seem plausibly more exogenous.

```
IV2 <- feols(weeksm1 ~ agem1 + black +
                       hispan + othrace |
             morekids ~ samesex,
             pset_data1, vcov = "HC1")
```

Table 4: E.12.1 g

	>2 Children First-Stage (IV) (1)	Weeks Worked IV (2)	>2 Children First-Stage (2SLS) (3)	Weeks Worked 2SLS (4)
Same Sex	0.068***		0.068***	
	(0.002)		(0.002)	
>2 Children		-6.31***		-5.82***
		(1.27)		(1.25)
Mothers Age (Years)			0.015***	0.832***
			(0.0003)	(0.023)
Black			0.101***	11.6***
			(0.004)	(0.232)
Hispanic			0.151***	0.404
			(0.004)	(0.261)
Other Race			0.028***	2.13***
			(0.005)	(0.211)
(Intercept)	0.346***	21.4***	-0.140***	-4.79***
	(0.001)	(0.487)	(0.008)	(0.390)
R^2	0.005	0.014	0.024	0.044
Observations	254,654	254,654	254,654	254,654
F-test	1,237.219	24.198	1,262.087	1,309.956

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*