Recitation 6: Instrumental Variables and Experiments

Matthew Davis

July 6, 2023

```r
# Practice Question 1
fert <- read_dta('data/fertility.dta') %>%
  rename(weeksworked = weeksm1) # to match textbook question
```

Practice Question 1: Stock-Watson Empirical Exercise 12.1

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. The data set contains information on married women aged 21–35 with two or more children.

1a: Regress *weeksworked* 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?

```r
mod.1a <- feols(weeksworked ~ morekids, fert, vcov = 'hetero')
etable(mod.1a)
```

```
##                              mod.1a
## Dependent Var.:         weeksworked
##
## Constant          21.07*** (0.0561)
## morekids         -5.387*** (0.0872)
## ________________ __________________
## S.E. type        Heteroskedas.-rob.
## Observations                254,654
## R2                          0.01431
## Adj. R2                     0.01430
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coefficient is -5.387 which indicates women with more than two children work 5.4 fewer weeks than women with two or fewer children

1b: Explain why the OLS regression estimated in (a) is inappropriate for estimating the causal effect of fertility (*morekids*) on labor supply (*weeksworked*).

Both fertility and weeks worked are choice variables. A woman with a positive labor supply regression error (a woman who works more than average) may also be a woman who is less likely to have an additional child.

This would imply that *morekids* is positively correlated with the regression error, so that the OLS estimator of its coefficient is positively biased.

1c: 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?

Since our outcome variable is binary and we are using OLS, we are estimating a linear probability model

```r
mod.1c <- feols(morekids ~ samesex, fert, vcov = 'hetero')
etable(mod.1c, fitstat = 'F')
```

```
##                               mod.1c
## Dependent Var.:             morekids
##
## Constant        0.3464*** (0.0013)
## samesex         0.0675*** (0.0019)
## _______________ __________________
## S.E. type       Heteroskedas.-rob.
## F-test                     1,237.2
## ---
## Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

This suggests that couples with same-sex children are 6.8% more likely to have an additional child. The ‘effect’ is significant at the 0.1% level.

1d: Explain why *samesex* is a valid instrument for the IV regression of *weeksworked* on *morekids*

Samesex is random and is unrelated to any of the other variables in the model including the error term in the labor supply equation. Thus, the instrument is exogenous.

From 1c, the first stage F-statistic is large (F = 1237) so the instrument is highly relevant. Together, these satisfy the two conditions for a valid instrument.

1e: Is *samesex* a weak instrument?

No, the F-statistic of 1237 is much higher than the (arbitrary) conventional threshold of $F = 10$. THus, we consider it a very strong instrument assuming it is exogenous.

1f: Estimate the IV regression of *weeksworked* on *morekids*, using *samesex* as an instrument. How large is the fertility effect on labor supply?

This is asking for a regression with one instrument for one endogenous variable and zero exogenous variables:

```r
mod.1e <- feols(weeksworked ~ 1 | morekids ~ samesex, fert, vcov = 'hetero')
etable(mod.1a, mod.1e)
```

```
##                              mod.1a            mod.1e
## Dependent Var.:        weeksworked       weeksworked
##
## Constant         21.07*** (0.0561) 21.42*** (0.4873)
## morekids        -5.387*** (0.0872) -6.314*** (1.275)
## _______________ __________________ _________________
## S.E. type       Heteroskedas.-rob. Heteroskeda.-rob.
## Observations               254,654           254,654
## R2                         0.01431           0.01388
## Adj. R2                    0.01430           0.01388
## ---
## Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The first model is the OLS regression from part 1a and the second model is the IV regression asked for here. The estimated coefficient is -6.31, indicating that women with more than two children are estimated to work 6.3 fewer weeks than women with two or fewer children.

1g: Do the results change when you include the variables *agem1*, *black*, *hispan*, and *othrace* in the labor supply regression (treating these variables as exogenous)? Explain why or why not.

Here’s a command that estimates the relevant OLS and 2SLS models with and without the suggested control variables

```
mods.ols <- feols(weeksworked ~ morekids + sw0(agem1 + black + hispan + othrace),
                  fert)
mods.iv <- feols(weeksworked ~ sw(1, agem1 + black + hispan + othrace) |
                   morekids ~ samesex,
                 fert)
# The fitstat argument here adds the first-stage F-statistic for IV models
etable(mods.ols, mods.iv, vcov = 'hetero', order = c('!Constant'), fitstat = 'ivf')
```

```
##                                      mods.ols.1         mods.ols.2
## Dependent Var.:                     weeksworked        weeksworked
##
## morekids                     -5.387*** (0.0872) -6.230*** (0.0862)
## agem1                                           0.8379*** (0.0121)
## black                                            11.66*** (0.1955)
## hispan                                           0.4661** (0.1807)
## othrace                                          2.142*** (0.2083)
## Constant                      21.07*** (0.0561) -4.835*** (0.3673)
## ____________________________ __________________ __________________
## S.E. type                    Heteroskedas.-rob. Heteroskedas.-rob.
## F-test (1st stage), morekids                 --                 --
##
##                                      mods.iv.1          mods.iv.2
## Dependent Var.:                    weeksworked        weeksworked
##
## morekids                     -6.314*** (1.275)  -5.821*** (1.246)
## agem1                                          0.8316*** (0.0226)
## black                                           11.62*** (0.2318)
## hispan                                            0.4042 (0.2608)
## othrace                                        2.131*** (0.2110)
## Constant                    21.42*** (0.4873) -4.792*** (0.3898)
## ---------------------------- ----------------- ------------------
## S.E. type                   Heteroskeda.-rob. Heteroskedas.-rob.
## F-test (1st stage), morekids          1,237.2            1,279.8
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Setting the order argument to '!Constant' will order the coefficient for constant last in the output. Without the exclamation mark, it would order it first.)

Comparing the IV models, the estimates for the coefficient on morekids do not change substantially, suggesting that *samesex* is largely unrelated to the control variables.

Notice that we can set the standard errors used for all models being displayed just by putting it in the etable function: vcov = 'hetero' or 'HC1' for heteroskedasticity-robust standard errors. For the default standard errors which assume homoskedasticity, we'd just set vcov = 'iid' or 'regular' or blank if we did not indicate a preference when defining a model.

Extra R examples useful for the pset

The formula for estimating IV moels in fixest is intuitive: all presumed-exogenous variables are to the left of the "|" while endogenous variables are immediately to their right with instruments for those endogenous variables separated by a tilde. In short, the general formula is

"y ~ w1.exog + w2.exog | x1.exog + x2.exog ~ z1.instrument + z2.instrument = z3.instrument"

Recall our reduced-form/first-stage regression of *morekids* on *samesex* above. Suppose we were to use another variable in the dataset like *boy1st* as a second instrument for this endogenous variable. Here, we have two instruments for one endogenous variable and a set of control variables assumed to be exogenous:

```r
mod.example <- feols(weeksworked ~ agem1 + black + hispan + othrace | morekids ~ samesex + boy1st,
                     data = fert)
```

And in the case where we have no exogenous variables, we simply leave a 1 in place of the exogenous variables (referring to the constant regressor, i.e., the intercept).

```r
mod.endog <- feols(weeksworked ~ 1 | morekids ~ samesex + boy1st,
                   data = fert)
```

And of course, we can estimate these in one command using the stepwise/cumulative stepwise functions sw0/sw/csw0/csw as in this example:

```r
mod.both <- feols(weeksworked ~ sw(1,
                                   agem1 + black + hispan + othrace) | morekids ~ samesex + boy1st,
                  data = fert)
etable(mod.both, fitstat = ~ ivf + F)
```

```
##                                       mod.both.1         mod.both.2
## Dependent Var.:                      weeksworked        weeksworked
##
## Constant                    21.41*** (0.4812) -4.788*** (0.4061)
## morekids                    -6.290*** (1.259)  -5.788*** (1.231)
## agem1                                          0.8311*** (0.0227)
## black                                           11.62*** (0.2281)
## hispan                                            0.3991 (0.2582)
## othrace                                         2.130*** (0.2058)
## ____________________________ _________________ __________________
## S.E. type                                  IID                IID
## F-test (1st stage), morekids            633.65             655.88
## F-test                                  24.599            1,310.0
## ---
## Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Notice how we included both the first-stage and the second-stage F-statistics in the output

It will be useful in the problem set to refer to a regression’s predictions or residuals in the process of manually deriving the two-stage estimator. When doing so, it will be convenient to save these as separate variables in the original dataset. This way we can refer to them as we would any other variable:

```
mod.rf <- feols(morekids ~ samesex + boy1st, fert)
fert$rf.residuals <- mod.rf$residuals
fert$morekids.predictions <- mod.rf$fitted.values
```

Finally, it may be useful to refer to some test statistic (e.g., distributed t, F, or Chi2) and look up its associated p-value:

```
# F statistic
mod.f <- fitstat(mod.rf, 'F')
# Some associated values: F-stat, p-value, degrees of freedom
mod.f$f
```

```
## $stat
## [1] 633.6531
##
## $p
## [1] 3.09354e-275
##
## $df1
## [1] 2
##
## $df2
## [1] 254651
```

```
# p-value of a Chi2 distributed statistic
# (setting stat and degrees of freedom randomly to 3 and 1)
chisq.stat <- 3
chisq.dof <- 1
chisq.p <- 1-pchisq(chisq.stat, chisq.dof)
chisq.p
```

```
## [1] 0.08326452
```