# Lab 6

Ray Caraher

2025-05-06

# Setup

# Setting up our script

Before we get into any real coding, let's make sure that the preamble for our code looks good. Here is how I set it up:

```
## Load packages
library(haven)
library(fixest)
library(tidyverse)
## Set options

options(scipen = 999)

## Clear environment

rm(list = ls())

## Set directories

base_directory <- '/Users/rcaraher/Library/CloudStorage/OneDrive-UniversityofMassachusetts/Academic/Teachi
data_directory <- file.path(base_directory, 'Data')
results_directory <- file.path(base_directory, 'Results')
```

# Instrumental Variable Estimate in T

# Overview of IVs

Instrumental variables is another route to "causal" inference

- In DiD approach, only looking at changes in treatment status that are exogenous (caused by policy changes, lottery, natural experiment, etc.)
- In IV approach, only looking at variation in outcome that is correlated with variation in an exogenous variable (the instrument)

# Identifying IVs

We want to estimate the following regression:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

However, we have reason to believe that $X$ is not exogenous.

Examples:

1. Effect of schooling ($X$) on wages ($Y$): Some unobservable omitted variable (e.g., ability) is correlated with both $X$ and $Y$
2. Effect of insurance ($X$) on health ($Y$): Selection in that healthier individuals ($Y$) may be more likely to get insurance ($X$)
3. Effect of policing ($X$) on crime ($Y$): Reverse causality as police ($X$) are often deployed to areas with high crime rates ($Y$)

# Identifying IVs

In these cases, $X$ will not have a valid causal interpretation.

What can we do?

One solution is to identify an **instrument** $Z$ which is correlated with $X$ and correlated with $Y$ *only through its correlation with* $X$

In other words, the instrument was be **relevant and excusable**

# Conditions for a Valid Instrument

If the model is

$$Y = \beta_0 + \beta_1 X + \epsilon$$

,

then a valid instrument $Z$ must be:

1. Relevant: $Cov(Z, X) \neq 0 \rightarrow Z$ must be correlated with $X$
2. Excludable: $Cov(Z, \epsilon) = 0 \rightarrow Z$ must *not* be correlated with the error term

# Estimation via 2SLS: Stage 1

To estimate $Y = \beta X + \epsilon$ with instrument $Z$, use 2SLS:

Stage 1: Regress $X$ on $Z$ (get predicted $\hat{X}$):

$$X = \pi_0 + \pi_1 Z + \eta$$

This is called the **first stage**

# Notes about the First Stage - Weak Instruments

If $\pi_1$ is close to zero, $Z$ is a weak instrument:

- $\hat{X}$ contains little exogenous variation
- Stop here: 2SLS estimates become biased and unreliable (can be worse than OLS)

Can use statistical tests to look for weak instruments, most common being the **F-statistic**

# Estimation via 2SLS: Stage 2

Stage 2: Regress $Y$ on $\hat{X}$:

$$Y = \beta_0 + \beta_1 \hat{X} + \mu$$

# Why use 2SLS?

- OLS is biased when $X$ is endogenous
- IV isolates exogenous variation in $X$
- 2SLS is consistent (though often less efficient than OLS)

# 2SLS in R

# Estimating IVs in R

- Estimating 2SLS in R is a simple extension from our normal regression tools

# The setting

What is the effect of fertility on labor supply?

Important empirical question for many reasons:

- Having children could push women out of the labor force for some time, having career implications
- Having children may lead to increased premiums for fathers in the labor market
- More educated/higher class families may be differentially affected by fertility

# The setting

Want to estimate the following:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where $Y$ is a labor market outcome (hours worked, employment status, wages, etc.) and $X$ is fertility (number of children, having any children, etc.)

However, naive estimates of $X$ on $Y$ may not be credible if we aim to estimate a causal effect for many reasons:

1. Families may time when to have children based on labor market factors (reverse causality)
2. Families may have children when they expect their labor market outcomes to improve (OVB)
3. Families that are less resource/time constrained may have more children (selection)

# The setting

Angrist and Evans (1998) propose an instrumental variable to overcome these biases: **sex composition of current children**

Intuition:

Relevance:

- Families have a strong desire to have mixed-sex children and will increase fertility to do so
  - If you have two girls, will likely have a third child in an attempt to have a son
  - But if you have one boy and one girl already, less likely to have a third child

Exculdable:

- Initial sex of children is *randomly assigned*

Therefore, having two same-sex children is (arguably) a valid **instrument for fertility**

# Getting data

Let's read in the data and take a look.

```
ae_pums <- read_dta(file.path(data_directory, "angrist_evans_data.dta"))

glimpse(ae_pums)
```

```
## Rows: 402,014
## Columns: 11
## $ twins_1         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ mom_weeks_worked <dbl> 0, 52, 30, 0, 0, 0, 22, 26, 40, 0, 52, 0, 52, 0, 52
## $ kidcount        <dbl> 2, 2, 2, 2, 2, 3, 2, 3, 2, 2, 2, 2, 2, 2, 4, 3, 2,
## $ mom_worked      <dbl> 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1,
## $ twins_2         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ moreths         <dbl> 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0,
## $ whitem          <dbl> 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
## $ blackm          <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ morekids        <dbl> 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0,
## $ hispm           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ samesex         <dbl> 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0,
```

# Looking at data

Let's read in the data and take a look at some descriptives:

```
desc_tab <- ae_pums %>%
  summarise(mean_kids = mean(kidcount, na.rm = T),
            samesex = mean(samesex, na.rm = T))
desc_tab
```

```
## # A tibble: 1 x 2
##   mean_kids samesex
##       <dbl>   <dbl>
## 1      2.56   0.505
```

# Looking at data

Let's read in the data and take a look at some descriptives

```
ae_pums %>%
  count(kidcount)
```

```
## # A tibble: 11 x 2
##    kidcount      n
##       <dbl>  <int>
##  1        2 239150
##  2        3 117203
##  3        4  33577
##  4        5   9046
##  5        6   2160
##  6        7    607
##  7        8    205
##  8        9     39
##  9       10     18
## 10       11      7
## 11       12      2
```

# Looking at data

Let's read in the data and take a look at some descriptives

```
ae_pums %>%
  group_by(samesex) %>%
  summarise(morekids = mean(morekids, na.rm = T))
```

```
## # A tibble: 2 x 2
##   samesex morekids
##     <dbl>    <dbl>
## 1       0    0.375
## 2       1    0.435
```

# Estimating OLS

Let's first estimate using OLS the effect of having at least three kids on the likelihood mom worked:

```
ae_pums <- ae_pums %>%
  mutate(mt2kids = case_when(kidcount > 2 ~ 1,
                             is.na(kidcount) ~ NA_real_,
                             TRUE ~ 0))

mworked_ols <- feols(mom_worked ~ mt2kids + whitem + blackm + hispm + moreths,
                     data = ae_pums)
summary(mworked_ols)
```

```
## OLS estimation, Dep. Var.: mom_worked
## Observations: 402,014
## Standard-errors: IID
##              Estimate Std. Error   t value                Pr(>|t|)
## (Intercept)  0.599837   0.004663 128.63257              < 2.2e-16 ***
## mt2kids     -0.121478   0.001587 -76.52607              < 2.2e-16 ***
## whitem      -0.017575   0.004642  -3.78571 0.0001532915018691329 ***
## blackm       0.107326   0.005067  21.17960              < 2.2e-16 ***
## hispm       -0.049719   0.006368  -7.80765 0.0000000000000058403 ***
## moreths      0.059676   0.001712  34.86374              < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.48968   Adj. R2: 0.024116
```

This is pretty close to the estimated effect in AE (1998) of -0.176 (see table 5, row 1 in the NBER working paper draft)

# Estimating OLS

Let's now estimate using OLS the effect of having at least three kids on the number of weeks mom worked:

```
mweeks_ols <- feols(mom_weeks_worked ~ mt2kids + whitem + blackm + hispm + moreths,
                    data = ae_pums)
summary(mweeks_ols)
```

```
## OLS estimation, Dep. Var.: mom_weeks_worked
## Observations: 402,014
## Standard-errors: IID
##             Estimate Std. Error   t value              Pr(>|t|)
## (Intercept) 23.44623   0.208961 112.20389             < 2.2e-16 ***
## mt2kids     -6.11931   0.071133 -86.02610             < 2.2e-16 ***
## whitem      -1.70262   0.208033  -8.18435 0.00000000000000027457 ***
## blackm       5.26005   0.227075  23.16443             < 2.2e-16 ***
## hispm       -3.01258   0.285356 -10.55726             < 2.2e-16 ***
## moreths      2.35799   0.076702  30.74229             < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 21.9   Adj. R2: 0.030178
```

# Estimating 2SLS

We have already discussed why a naive regression of fertility on labor market outcomes may be endogenous. Now, let's instrument using $Z$ as an indicator for if the family had 2 kids of the same sex at birth.

We will first implement it by-hand, then using `feols()`

# Estimating the First Stage

Let's estimate the first stage.

When we use covariates, its important to include them on the RHS of the first-stage regression as well!

```
fs <- feols(mt2kids ~ samesex + whitem + blackm + hispm + moreths,
            data = ae_pums)
summary(fs)
```

```
## OLS estimation, Dep. Var.: mt2kids
## Observations: 402,014
## Standard-errors: IID
##              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)  0.435078   0.004630  93.9641 < 2.2e-16 ***
## samesex      0.059407   0.001532  38.7794 < 2.2e-16 ***
## whitem      -0.052779   0.004603 -11.4656 < 2.2e-16 ***
## blackm       0.067258   0.005024  13.3866 < 2.2e-16 ***
## hispm        0.096116   0.006313  15.2241 < 2.2e-16 ***
## moreths     -0.096315   0.001691 -56.9677 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.485619   Adj. R2: 0.021451
```

# Estimating the First Stage

The regression is significant and positive, suggesting that having 2 kids of the same sex significant increases your probability of having a third.

Now let's generate the predicted value of $X$ using the variation in `samesex`

```
ae_pums <- ae_pums %>%
  mutate(X_hat = predict(fs, type = "response"))
```

# Estimating the second stage

Now, we use the predicted values from the first stage to estimate the effect of the **exogenous** part of fertility on labor market outcomes:

```
mworked_s2 <- feols(mom_worked ~ X_hat + whitem + blackm + hispm + moreths,
                    data = ae_pums)
summary(mworked_s2)
```

```
## OLS estimation, Dep. Var.: mom_worked
## Observations: 402,014
## Standard-errors: IID
##              Estimate Std. Error  t value           Pr(>|t|)
## (Intercept)  0.603134   0.013028 46.29646          < 2.2e-16 ***
## X_hat       -0.128571   0.026190 -4.90908 0.0000009154029933 ***
## whitem      -0.017947   0.004873 -3.68286 0.0002306660443972 ***
## blackm       0.107805   0.005401 19.95974          < 2.2e-16 ***
## hispm       -0.049035   0.006892 -7.11519 0.0000000000011195 ***
## moreths      0.058992   0.003052 19.32714          < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.493219   Adj. R2: 0.00996
```

```
mweeks_s2 <- feols(mom_weeks_worked ~ X_hat + whitem + blackm + hispm + moreths,
                    data = ae_pums)
summary(mweeks_s2)
```

```
## OLS estimation, Dep. Var.: mom_weeks_worked
## Observations: 402,014
## Standard-errors: IID
##             Estimate Std. Error  t value              Pr(>|t|)
## (Intercept) 23.45461   0.584881 40.10148             < 2.2e-16 ***
## X_hat       -6.13733   1.175832 -5.21956 0.0000001794336473852 ***
## whitem      -1.70356   0.218786 -7.78644 0.0000000000000069086 ***
## blackm       5.26127   0.242485 21.69727             < 2.2e-16 ***
## hispm       -3.01084   0.309403 -9.73114             < 2.2e-16 ***
```

# Estimating 2SLS using a built-in routine

The `feols()` function we have been working with has built-in ability to estimate 2SLS.

Simply add another | after the fixed-effects using `X ~ Z` formula syntax.

The control variables on the immediate RHS of the formula will automatically be included.

Since we have no fixed-effects here, we can just use a `0`.

Our results perfectly match those computations we did by-hand.

# Estimating 2SLS using a built-in routine

```
mworked_tsls <- feols(mom_worked ~ whitem + blackm + hispm + moreths |
                0 |
                mt2kids ~ samesex,
              data = ae_pums)
summary(mworked_tsls)
```

```
## TSLS estimation - Dep. Var.: mom_worked
##                   Endo.    : mt2kids
##                   Instr.   : samesex
## Second stage: Dep. Var.: mom_worked
## Observations: 402,014
## Standard-errors: IID
##              Estimate Std. Error  t value            Pr(>|t|)
## (Intercept)  0.603134   0.012934 46.62990           < 2.2e-16 ***
## fit_mt2kids -0.128571   0.026003 -4.94444 0.00000076395140220 ***
## whitem      -0.017947   0.004838 -3.70938 0.00020779515550137 ***
## blackm       0.107805   0.005362 20.10349           < 2.2e-16 ***
## hispm       -0.049035   0.006842 -7.16643 0.00000000000077108 ***
## moreths      0.058992   0.003030 19.46634           < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.489692   Adj. R2: 0.024068
## F-test (1st stage), mt2kids: stat = 1,503.8     , p < 2.2e-16 , on 1 and 402,008 DoF.
##                  Wu-Hausman: stat =     0.074685, p = 0.784633, on 1 and 402,007 DoF.
```

# Estimating 2SLS using a built-in routine

```r
mweeks_tsls <- feols(mom_weeks_worked ~ whitem + blackm + hispm + moreths |
                0 |
                mt2kids ~ samesex,
              data = ae_pums)
summary(mweeks_tsls)
```

```
## TSLS estimation - Dep. Var.: mom_weeks_worked
##                   Endo.    : mt2kids
##                   Instr.   : samesex
## Second stage: Dep. Var.: mom_weeks_worked
## Observations: 402,014
## Standard-errors: IID
##             Estimate Std. Error  t value              Pr(>|t|)
## (Intercept) 23.45461   0.579591 40.46754             < 2.2e-16 ***
## fit_mt2kids -6.13733   1.165196 -5.26721 0.0000001385845302174 ***
## whitem      -1.70356   0.216807 -7.85752 0.0000000000000039279 ***
## blackm       5.26127   0.240292 21.89533             < 2.2e-16 ***
## hispm       -3.01084   0.306604 -9.81997             < 2.2e-16 ***
## moreths      2.35626   0.135795 17.35160             < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 21.9   Adj. R2: 0.030178
## F-test (1st stage), mt2kids: stat = 1,503.8     , p < 2.2e-16 , on 1 and 402,008 DoF.
##                  Wu-Hausman: stat =     2.401e-4, p = 0.987636, on 1 and 402,007 DoF.
```

## 2SLS Diagnostics

One benefit of using the built-in 2SLS models is that they automatically compute test-statistics such as the F-statistic.

Recall the intuition of the F-statistic: *Do the explanatory variables in this regression explain a meaningful amount of variation in the outcome?*

1. Estimate a "restricted model" without the instrument $Z$ and an "unrestricted model" with the instrument $Z$
2. Compare the residual sum of squares (RSS)
3. If the RSS drops substantially, the F-stat is large.

If the F-stat is large, this implies that it explains a lot of the variation in $Y$.

In other words, it is not a **weak instrument**

A ballpark F-stat of 10 is usually considered the rule-of-thumb

# 2SLS Diagnostics: Problems with the F-stat

However, the F-stat rule-of-thumb is really only valid with certain assumptions.

In the presence of clustered, heteroskedastic, or a number of other error cases we likely need a much larger F-stat than 10.

The `ivDiag` package has some useful methods for more thorough 2SLS diagnostics

# Advanced 2SLS Diagnostics

Let's load in the package (don't forget to install if the first time using it)

```
#install.packages("ivDiag")
library(ivDiag)
```

```
## ## Tutorial: https://yiqingxu.org/packages/ivDiag/
```

# Advanced 2SLS Diagnostics

The `ivDiag()` function requires us to supply its arguments as strings.

Let's first use it to look at some F-stats.

```
mworked_ivDiag <- ivDiag(data = ae_pums,
                         Y = "mom_worked",
                         D = "mt2kids",
                         Z = "samesex",
                         controls = c("whitem", "blackm", "hispm", "moreths"),
                         bootstrap = FALSE,
                         run.AR = FALSE)
```

# Advanced 2SLS Diagnostics

In addition to returning the OLS, 2SLS, first stage, and reduced form estimates, the `ivDiag()` function will return a range of F-stat tests.

```
mworked_ivDiag$est_ols
```

```
##              Coef     SE        t CI 2.5% CI 97.5% p.value
## Analytic -0.1215 0.0016 -76.1443 -0.1246  -0.1184       0
```

```
mworked_ivDiag$est_2sls
```

```
##              Coef    SE       t CI 2.5% CI 97.5% p.value
## Analytic -0.1286 0.026 -4.9445 -0.1795  -0.0776       0
```

```
mworked_ivDiag$F_stat
```

```
##  F.standard    F.robust   F.cluster F.effective
##    1503.843    1504.630          NA    1504.630
```

# The Anderson-Rubin test

If we are estimating an IV with a weak instrument, the standard error on the 2SLS estimate will not be valid.

Anderson-Rubin (1949) suggest a test statistic which is robust the weak instrument concern.

The Anderson-Rubin test statistic works backwards:

For a set of possible $\beta_0$ values, it computes if $Y - \beta_0 X$ regressed on $Z$ is statistically significant. Then, the set of $\beta_0$ such that $Z$ is not significant is the AR Confidence interval.

We can implement this by setting the `run.AR` argument in `ivDiag()` to `TRUE`

# The Anderson-Rubin test

```
mworked_ivDiag_v2 <- ivDiag(data = ae_pums,
                            Y = "mom_worked",
                            D = "mt2kids",
                            Z = "samesex",
                            controls = c("whitem", "blackm", "hispm", "moreths"),
                            bootstrap = FALSE,
                            run.AR = TRUE)
```

```
## AR Test Inversion...
```

```
## Parallelising on 11 cores
```

```
mworked_ivDiag_v2$AR$ci.print
```

```
## [1] "[-0.1795, -0.0776]"
```