Matching and IV

Antonella Bandiera

March 7, 2023

Last class: Selection on Observables

Last class we examined the first non-experimental estimators

- We said that SOO requires very strong, sometimes implausible assumptions
- But if there is nothing else, you can:

1. Regression adjustment (control with additional independent variables)
2. Matching (Pairing treated and untreated based on observables)

Last class: Selection on Observables

Last class we examined the first non-experimental estimators

- We said that SOO requires very strong, sometimes implausible assumptions
- But if there is nothing else, you can:

1. Regression adjustment (control with additional independent variables)
2. Matching (Pairing treated and untreated based on observables)

- As we mentioned, it is hard to believe SOO designs. Why?

Last class: Selection on Observables

Last class we examined the first non-experimental estimators

- We said that SOO requires very strong, sometimes implausible assumptions
- But if there is nothing else, you can:

1. Regression adjustment (control with additional independent variables)
2. Matching (Pairing treated and untreated based on observables)

- As we mentioned, it is hard to believe SOO designs. Why?
- Let's examine one Matching example

Matching

- Matching is the process of controlling for potential confounders between a treatment and an outcome by constructing comparison groups that are similar according to a set of matching variables.
- Imagine that we are interested in getting the effect of a job-training program on your chances of getting a good job.
- We notice that, while the pool of unemployed people eligible was about 50% male/50% female, the program just happened to be advertised heavily to men
- So the people actually in the program were 80% male/20% female, what is the problem?

Matching

- Matching is the process of controlling for potential confounders between a treatment and an outcome by constructing comparison groups that are similar according to a set of matching variables.
- Imagine that we are interested in getting the effect of a job-training program on your chances of getting a good job.
- We notice that, while the pool of unemployed people eligible was about 50% male/50% female, the program just happened to be advertised heavily to men
- So the people actually in the program were 80% male/20% female, what is the problem?

Outcomes $\leftarrow$ Gender $\rightarrow$ JobTrainingProgram

Matching

- The matching approach would look at all the untreated people and would construct a “control group” that was also 80% male/20% female, to compare to the already 80-20 treated group.

Matching

- The matching approach would look at all the untreated people and would construct a “control group” that was also 80% male/20% female, to compare to the already 80-20 treated group.
- This gets rid of Gender → JobTrainingProgram
- And leaves us with JobTrainingProgram → Outcomes

Matching

- What are we trying to do?
- Matching methods create a set of weights for each observation, perhaps calling that weight w
- Those weights are designed to make the treatment and control groups comparable.
- When we want to estimate the effect of treatment, we would calculate a weighted mean of the outcomes for the the treatment and control groups, and compare those.
- A weighted mean multiplies each observation's value by its weight, adds them up, and then divides by the sum of the weights
- The weighted mean of Y,

$$\frac{\sum wY}{\sum w}$$

Weights

- But where do the weights come from?
- Many different matching processes
- But they all do so using a set of “matching variables,” and using those matching variables to construct a set of weights so as to close any back doors that those matching variables are on.
- The idea is to create a set of weights such that there’s no longer any variation between the treated and control groups in the matching variables.

Weights: A simple example

Let's say we have a treated group, each of whom has received job training, consisting of 80 men and 20 women. Of the 80 men, 60 end up with a job and 20 without. Of the women, 12 end up with a job and 8 without.

Now let's look at the control group, which consists of 500 men and 500 women. Of the men, 350 end up with a job and 150 without. Of the women, 275 end up with a job and 225 without.

If we look at the raw comparison, we get that 72% of those with job training end up with jobs, while in the control group 62.5% end up with jobs. That's a treatment effect of 9.5 percentage points.

Weights: A simple example

Let's say we have a treated group, each of whom has received job training, consisting of 80 men and 20 women. Of the 80 men, 60 end up with a job and 20 without. Of the women, 12 end up with a job and 8 without.

Now let's look at the control group, which consists of 500 men and 500 women. Of the men, 350 end up with a job and 150 without. Of the women, 275 end up with a job and 225 without.

If we look at the raw comparison, we get that 72% of those with job training end up with jobs, while in the control group 62.5% end up with jobs. That's a treatment effect of 9.5 percentage points.

What happens if we use weights? The treatment effect might be biased due to differences in the gender composition of the control group and treatment group. How?

Weights: A simple example

There are many different matching methods, each of which might create different weights, but one method might create weights like this:

- Give a weight of 1 to everyone who is treated
- Give a weight of 80/500=.16 to all untreated men
- Give a weight of 20/500=.04 to all untreated women

Weights: A simple example

There are many different matching methods, each of which might create different weights, but one method might create weights like this:

- Give a weight of 1 to everyone who is treated
- Give a weight of 80/500=.16 to all untreated men
- Give a weight of 20/500=.04 to all untreated women

Let's see if we've eliminated the variation in gender between the groups. The treated group will still be 80% men - giving all the treated people equal weights won't change anything on that side. How about the untreated people?

If we calculate the proportion male, using a value of 1 for men and 0 for women, the weighted mean for them is $(.16 \times 500 + 0.4 \times 0)/(.16 \times 500 + 0.4 \times 500) = 80\%$, or the same as the treatment group.

Weights: A simple example

Now that we've balanced gender between the treated and control groups, what's the treatment effect?

It's still 72% of people who end up with jobs in the treated group - again, nothing changes there.

But in the untreated group it's
$(.16 \times 350 + 0.4 \times 275)/(.16 \times 500 + 0.4 \times 500) = 67\%$

The treatment effect is $.72 - .67 = 5p.p$

A single matching variable

- It is uncommon to match on a single variable
  1. What will our matching criteria be?
  2. Are we selecting matches or constructing a matched weighted sample?
  3. If we're selecting matches, how many?

1. What will our matching criteria be?

When performing matching, we are trying to select observations in the control group that are similar to those in the treated group

- What is similar?

1. What will our matching criteria be?

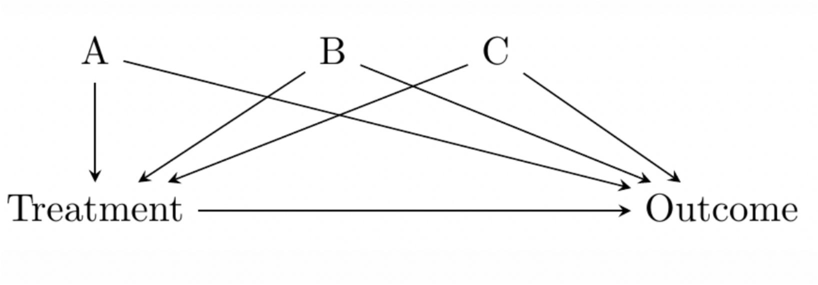
When performing matching, we are trying to select observations in the control group that are similar to those in the treated group

- What is similar?

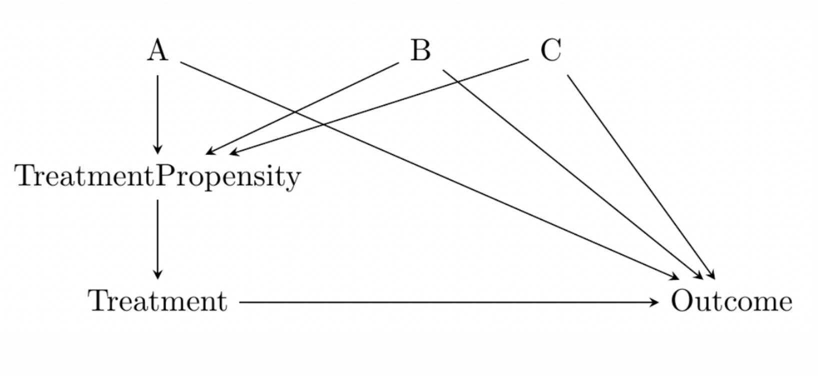
Two Approaches

1. Multivariate Distance Matching (observations are similar if they have similar values of the matching variables)
2. Propensity Score Matching (observations are similar if they were equally likely to be treated)

Distance Matching



Propensity Score



2. Are we selecting matches or constructing a matched weighted sample?

- Both of these represent different ways of matching the treatment and control groups. Neither is the default approach.
- The process of selecting matches means that we're picking control observations to either be in or out of the matched control sample.
- In the process of selected a matched weighted sample, instead of simply being in or out, each control will receive a different weight depending on how close it is to a treated observation, or how handy it will be in making the matched control group look like the treated group.
- Each has its pros and cons (selecting matches is intuitive, but matched weighted sample is less sensitive)

3. If we're selecting matches, how many?

Matching algorithms

- Pair matching (one-to-one matching without replacement)
- Nearest-neighbor matching (with replacement)
- Radius matching
- Kernel matching
- Caliper matching

Multiple Matching Variables

- Going from one variable to many just boils down to how we do we collapse all of our matching variables to one single variable?
- We will use example data from Broockman (2013)
- *Intrinsic movitations* of American politicians (stuff they'll do even if there's no obvious reward to it)
- One example of such a motivation is that Black politicians in America may be especially interested in supporting the Black American community.
- In 2010 Broockman sent a whole bunch of emails to state legislators (politicians), simply asking for information on unemployment benefits.

Multiple Matching Variables

- Answering the email could make a material improvement in the person's life, not just advance a policy issue.
- Each email was sent by the fictional "Tyrone Washington," which in the United States is a name that strongly suggests it belongs to a Black man.
- Broockman varies whether the letter-writer claims to live in the legislator's district or far away.

Broockman 2013

Question 1: Do Black legislators respond less often to out-of-district emails from Black people than in-district emails from Black people? Yes!

Question 2: Do non-Black legislators respond less to out-of-district emails from Black people than in-district emails from Black people? Also yes!

Question 3: is the in-district/out-of-district difference smaller for Black legislators than non-Black ones? If so, that implies that Black legislators have additional intrinsic motivation to help the Black emailer, evidence in favor of the intrinsic-motivation hypothesis and the legislators-help-those-like-themselves hypothesis

Broockman 2013

- Where does the matching come in? It's in that last step where we compare the in/out-of-district gap between Black and non-Black legislator: Those groups tend to be elected in very different kinds of places.
- In the original study, Broockman used median household income in the district, the percentage of the district's population that is Black, and whether the legislator is a Democrat as matching variables

Distance Matching: Mahalanobis Distance

First, take each matching variable and divide its value by its standard deviation.

Now we can calculate distance. For a given treated observation A and a given control observation B, the Mahalanobis distance is the sum of the squares of all the differences between A and B.

Then, after you've taken the sum, you take the square root.

- When you are using many variables, you divide out the whole covariance matrix

$$d(x_1, x_2) = \sqrt{(x_1 - x_2)' S^{-1} (x_1 - x_2)}$$

Broockman and one-to-one nearest neighbor matching

- Three variables: income, black percent, leg democrat
- We will match districts with black legislators and those without

```
library(Matching)
library(tidyverse)
library(causaldata)
library(dplyr)
br <- causaldata::black_politicians
```

Broockman and one-to-one nearest neighbor matching

```r
# Outcome
Y <- br %>%
    pull(responded)
# Treatment
D <- br %>%
    pull(leg_black)
# Matching variables
X <- br %>%
    dplyr::select(medianhhincom, blackpercent,
                  leg_democrat) %>%
    as.matrix()

# Weight = 2, oddly, denotes Mahalanobis distance
M <- Match(Y, D, X, Weight = 2, caliper = 1)
```

Broockman and one-to-one nearest neighbor matching

```r
# See treatment effect estimate
summary(M)
```

```
##
## Estimate...  -0.0073462
## AI SE......  0.072683
## T-stat.....  -0.10107
## p.val......  0.91949
##
## Original number of observations..............  5593
## Original number of treated obs...............  364
## Matched number of observations...............  363
## Matched number of observations  (unweighted).  405
##
## Caliper (SDs)..................................
## Number of obs dropped by 'exact' or 'caliper'  1
```

Broockman and one-to-one nearest neighbor matching

```
# Get matched data for use elsewhere. Note that this appro
# duplicate each observation for each time it was matched
matched_treated <- tibble(id = M$index.treated,
                          weight = M$weights)
matched_control <- tibble(id = M$index.control,
                          weight = M$weights)
matched_sets <- bind_rows(matched_treated,
                          matched_control)
# Simplify to one row per observation
matched_sets <- matched_sets %>%
                    group_by(id) %>%
                    summarize(weight = sum(weight))
# And bring back to data
matched_br <- br %>%
    mutate(id = row_number()) %>%
    left_join(matched_sets, by = 'id')
```

Broockman and one-to-one nearest neighbor matching

```
# To be used like this!
lm(responded~leg_black, data = matched_br,
   weights = weight)
```

```
##
## Call:
## lm(formula = responded ~ leg_black, data = matched_br, w
##
## Coefficients:
## (Intercept)    leg_black
##    0.398531    -0.007346
```

Broockman using propensity score

```r
library(Matching)
library(tidyverse)
library(causaldata)
library(dplyr)
br <- causaldata::black_politicians

# We can estimate our own propensity score
m <- glm(leg_black ~ medianhhincom + blackpercent +
           leg_democrat,
         data = br, family = binomial(link = 'logit'))
# Get predicted values
br <- br %>%
    dplyr::mutate(ps = predict(m, type = 'response'))
# "Trim" control observations outside of
# treated propensity score range
# (we'll discuss this later in Common Support)
```

Broockman using propensity score

```
# "Trim" control observations outside of
# treated propensity score range

minps <- br %>%
    filter(leg_black == 1) %>%
    pull(ps) %>%
    min(na.rm = TRUE)
maxps <- br %>%
    filter(leg_black == 1) %>%
    pull(ps) %>%
    max(na.rm = TRUE)
br <- br %>%
    filter(ps >= minps & ps <= maxps)
```

Broockman using propensity score

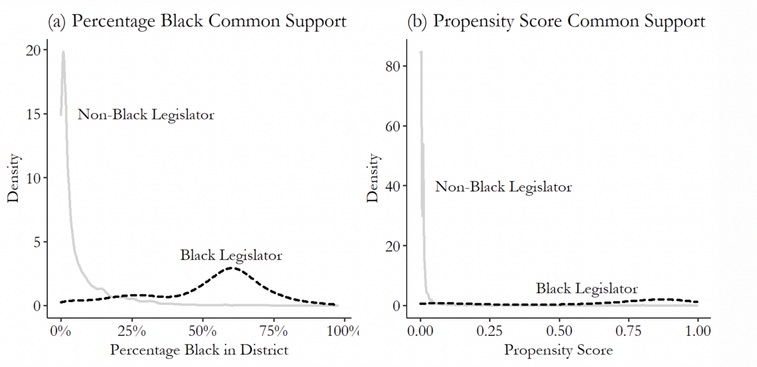
```r
# Create IPW weights
br <- br %>%
    mutate(ipw = case_when(
    leg_black == 1 ~ 1/ps,
    leg_black == 0 ~ 1/(1-ps)))

# And use to weight regressions
summary(lm(responded ~ leg_black, data = br, weights = ipw)
```

```
##
## Call:
## lm(formula = responded ~ leg_black, data = br, weights =
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -15.3666  -0.4056  -0.4039   0.5987  12.2721
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

Main Assumption

- **COMMON SUPPORT**: simply to look at the distribution of a variable for the treated group against the distribution of that same variable for the untreated group



Leaving SOO behind

SOO requires very demanding assumptions:

- I know the X that moves observations to select into treatment
- $(Y_i(1), Y_i(0)) \perp\!\!\!\perp D_i | X_i$
- Need to isolate all the unwanted variation in D_i

We move to selection on unobservables:

- I know something about what moves observations to select into treatment
- $(Y_i(1), Y_i(0)) \perp\!\!\!\perp Z_i$ and $Cov(Z_i, D_i) \neq 0$
- But for this we need to find some “good” variation in D_i

What is good variation and where does it come from?

We turn to natural experiments:

- Rather than needing to observe everything
- We observe some (quasi) random variation in D_i
- Natural experiments have naturally occurring random variation
  - Can come from actual nature (rainfall, hurricane, temperature, etc.)
  - But also many are policy-induced

What is good variation and where does it come from?

We turn to natural experiments:

- Rather than needing to observe everything
- We observe some (quasi) random variation in D_i
- Natural experiments have naturally occurring random variation
  - Can come from actual nature (rainfall, hurricane, temperature, etc.)
  - But also many are policy-induced

Some new vocabulary for the rest of the class:

- **Observational data**: Data generated from non-experimental settings
- **Identification strategy**: approach to using observational data to estimate causal effects -**Identifying assumptions**: Assumptions required for the identification strategy to causally estimate impacts

How to isolate good variation

Consider the following model:

$$Y_i = \alpha + \tau D_i + \beta X_i + \varepsilon_i$$

where

Y_i is the outcome of interest D_i is the treatment X_i is a set of covariates, where $Cov(X_i, \varepsilon_i) = 0$ ε_i is the error

What condition do we need to recover the causal effect of D_i?

How to isolate good variation

Consider the following model:

$$Y_i = \alpha + \tau D_i + \beta X_i + \varepsilon_i$$

where

Y_i is the outcome of interest D_i is the treatment X_i is a set of covariates, where $Cov(X_i, \varepsilon_i) = 0$ ε_i is the error

What condition do we need to recover the causal effect of D_i?

We need $E[\varepsilon_i | D_i] = 0 \leftrightarrow Cov(D_i, \varepsilon_i) = 0$

What happens when $E[\varepsilon|D_i] \neq 0$?

It means that D_i is endogenous:

$\rightarrow$ we cannot get an unbiased estimate of τ_{ATE}

- This can result from:
  - Ommitted variable bias
  - Reverse causality

Isolate exogenous variation

$$Y_i = \alpha + \tau D_i + \beta X_i + \varepsilon_i$$

Suppose we can separate D_i in two parts

$$D_i = B_i \varepsilon_i + C_i$$

with $Cov(C_i, \varepsilon_i) = 0$

Then we can write

$$Y_i = \alpha + \tau(B_i \varepsilon_i + C_i) + \beta X_i + \varepsilon_i$$

$$Y_i = \underbrace{\alpha + \tau C_i + \beta X_i + (1 + \tau B_i)\varepsilon_i}_{\text{rearrange}}$$

Isolate exogenous variation

$$Y_i = \alpha + \tau D_i + \beta X_i + \varepsilon_i$$

Suppose we can separate D_i in two parts

$$D_i = B_i \varepsilon_i + C_i$$

with $Cov(C_i, \varepsilon_i) = 0$

Then we can write

$$Y_i = \alpha + \tau(B_i \varepsilon_i + C_i) + \beta X_i + \varepsilon_i$$

$$Y_i = \underbrace{\alpha + \tau C_i + \beta X_i + (1 + \tau B_i)\varepsilon_i}_{\text{rearrange}}$$

- If we observed C_i we could recover τ
- But we can't

Instrumental variables

We want an instrumental variable:

Intuitively, an IV generates variation in C_i but it is uncorrelated with ε_i.

Z_i is a valid instrument for D_i when the following are satisfied:

Instrumental variables

<u>We want an instrumental variable:</u>

Intuitively, an IV generates variation in C_i but it is uncorrelated with ε_i.

Z_i is a valid instrument for D_i when the following are satisfied:

1. **First stage**: $Cov(Z_i, D_i) \neq 0$ - Z_i an D_i are related - Without this, you are capturing nothing - We can test this

Instrumental variables

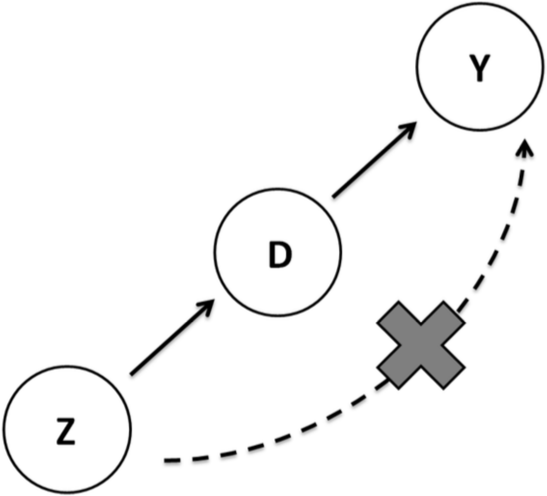
We want an instrumental variable:

Intuitively, an IV generates variation in C_i but it is uncorrelated with ε_i.

Z_i is a valid instrument for D_i when the following are satisfied:

1. **First stage**: $Cov(Z_i, D_i) \neq 0$ - Z_i an D_i are related - Without this, you are capturing nothing - We can test this
2. **Exclusion restriction** $Cov(Z_i, \varepsilon_i) = 0$ - Z_i and ε_i are **not** related - Z_i only affects Y_i through D_i - Fundamentally untestable

The exclusion restriction



Estimation of IV

Without covariates:

$$\hat{\tau}_{IV} = Cov(Z_i, Y_i)/Cov(Z_i, D_i)$$

Two stage least squares (2SLS)

The classic way to perform IV is via 2SLS:

1. **First stage**

Regress endogenous D_i on all exogenous variables including Z_i

$$D_i = \alpha + \gamma Z_i + \beta X_i + \eta_i$$

- Store the predicted values of D_i, $\hat{D}_i$
- This is our test of the first assumption!

The first stage

$$D_i = \alpha + \gamma Z_i + \beta X_i + \eta_i$$

This equation

- Estimates the effect of our instrument on treatment

The first stage

$$D_i = \alpha + \gamma Z_i + \beta X_i + \eta_i$$

This equation

- Estimates the effect of our instrument on treatment
- **This is isolating the “good variation”** in D_i
  - By assumption, Z_i is unrelated to Y_i except through D_i
  - Thus, a regression of D_i only keeps the good part
  - When we extract $\hat{D}_i$ we have eliminated the bad (endogenous) variation
- Lets us test whether $Cov(Z_i, D_i \neq 0)$:
  - The proper way to do this is with an F test
  - Rule of thumb: need F-stat > 20
- Should have an intuitive sign

Two stage least squares (2SLS)

The classic way to perform IV is via 2SLS:

1. **First stage**

Regress endogenous D_i on all exogenous variables including Z_i

$$D_i = \alpha + \gamma Z_i + \beta X_i + \eta_i$$

- Store the predicted values of D_i, $\hat{D}_i$
- This is our test of the first assumption!

2. **Second stage**

Regress outcome Y_i on predicted $\hat{D}_i$ and other Xs:

$$Y_i = \alpha + \tau \hat{D}_i + \gamma X_i + \varepsilon$$

- $\hat{\tau}$ in this equation is our IV estimate
- The standard error will be wrong, so always use canned routines

The second stage

$$Y_i = \alpha + \tau \hat{D}_i + \gamma X_i + \varepsilon$$

The second stage

$$Y_i = \alpha + \tau \hat{D}_i + \gamma X_i + \varepsilon$$

This equation:

- Estimates the effect of our treatment on our outcomes
- This only uses **good variation** in D_i
    - By assumption, Z_i is unrelated to Y_i except through D_i
    - So a regression of Z_i on $\hat{D}_i$ only uses the good variation (exogenous)
- Central assumption of IV:
    - Z_i is as good as randomly assigned

The reduced form

Good IV studies will also report the reduced form

- This is a regression of Y_i on your instrument Z_i

$$Y_i = \alpha + \theta Z_i + \pi X_i + \eta_i$$

- This does **not** recover $\hat{\tau}_{IV}$
- Instead this tells us how your outcome varies with the instrument
- This needs to have a causal interpretation

Estimation

We have estimated three total coefficients

1. First stage $\hat{\gamma}$ is the effect of our instrument on treatment
2. Reduced form: $\hat{\theta}$ is the effect of our instrument on our outcome
3. Second stage: $\hat{\tau}_{IV}$ is the effect of treatment on outcome

Estimation

We have estimated three total coefficients

1. First stage $\hat{\gamma}$ is the effect of our instrument on treatment
2. Reduced form: $\hat{\theta}$ is the effect of our instrument on our outcome
3. Second stage: $\hat{\tau}_{IV}$ is the effect of treatment on outcome

We can just compute:

$$\hat{\tau}_{IV} = \frac{\hat{\theta}}{\hat{\gamma}}$$

Estimation

We have estimated three total coefficients

1. First stage $\hat{\gamma}$ is the effect of our instrument on treatment
2. Reduced form: $\hat{\theta}$ is the effect of our instrument on our outcome
3. Second stage: $\hat{\tau}_{IV}$ is the effect of treatment on outcome

We can just compute:

$$\hat{\tau}_{IV} = \frac{\hat{\theta}}{\hat{\gamma}}$$

- This should be intuitive
  - We are just scaling the reduced form by the first stage
  - Our IV estimate is just the effect of instrument on our outcome, weighted by how much the instrument moves the treatment
  - Standard errors are still tricky (use canned routine)

The exclusion restriction is the key to any IV

You should always ask: *What is the exclusion restriction in this analysis saying?*

- Do we believe this, why or why not?

Misc

When doing IV

- Your standard errors will be bigger than OLS standard errors
  - This should be intuitive: IV is throwing out variation
  - You need a canned routine to correctly estimate them
- Always include the same covariates X_i in the first stage, second stage, and reduced form
- The exclusion restriction is fundamentally untestable