Conceptual Causality: IV

Section 2

Sima Biondi

Spring 2025

Gov 51: Data Analysis and Politics

Overview

1. Instrumental variables (IV)
2. R crash course
3. Back to IV

Roadmap

Last week: the **fundamental problem of causal inference**, and introduction to one estimation strategy (DiD)

Roadmap

Last week: the **fundamental problem of causal inference**, and introduction to one estimation strategy (DiD)

$\rightarrow$ Potential outcomes framework and notation

Roadmap

Last week: the **fundamental problem of causal inference**, and introduction to one estimation strategy (DiD)

- $\rightarrow$ Potential outcomes framework and notation
- $\rightarrow$ Estimand vs. estimator vs. estimate

Roadmap

Last week: the **fundamental problem of causal inference**, and introduction to one estimation strategy (DiD)

- $\rightarrow$ Potential outcomes framework and notation
- $\rightarrow$ Estimand vs. estimator vs. estimate
- $\rightarrow$ Parallel trends assumption

Roadmap

Last week: the **fundamental problem of causal inference**, and introduction to one estimation strategy (DiD)

- $\rightarrow$ Potential outcomes framework and notation
- $\rightarrow$ Estimand vs. estimator vs. estimate
- $\rightarrow$ Parallel trends assumption

Roadmap

Last week: the **fundamental problem of causal inference**, and introduction to one estimation strategy (DiD)

- → Potential outcomes framework and notation
- → Estimand vs. estimator vs. estimate
- → Parallel trends assumption

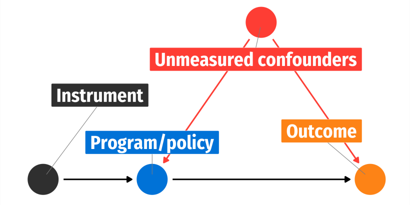
In this section, we continue to examine ways to estimate causal quantities via instrumental variables (IV)

Overview

1. Instrumental variables (IV)
2. R crash course
3. Back to IV

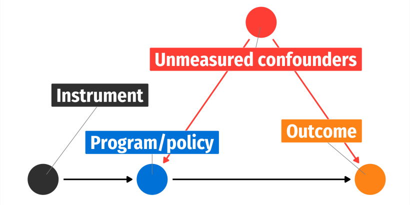
Intro to IV

- Explanatory variables are oftentimes correlated with our errors (also known as endogenity)



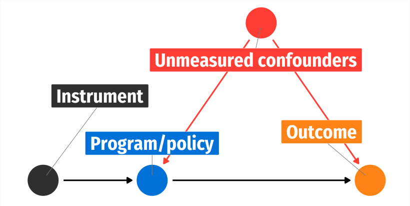
Intro to IV

- Explanatory variables are oftentimes correlated with our errors (also known as endogenity)
- Examples: conflict and economic growth, government information and inequality, efficacy of canvassing



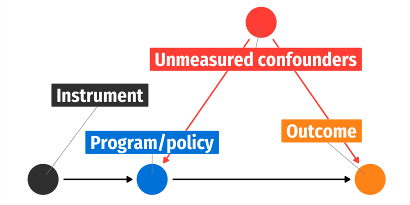
Intro to IV

- Explanatory variables are oftentimes correlated with our errors (also known as endogenity)
- Examples: conflict and economic growth, government information and inequality, efficacy of canvassing
- OLS and controls are insufficient to account for this bias



Intro to IV

- Explanatory variables are oftentimes correlated with our errors (also known as endogenity)
- Examples: conflict and economic growth, government information and inequality, efficacy of canvassing
- OLS and controls are insufficient to account for this bias
    - *Check-in*: why?



IV assumptions

- For an IV to be identified, it must:

IV assumptions

- For an IV to be identified, it must:
    1. Be assigned as-if random

IV assumptions

- For an IV to be identified, it must:
  1. Be assigned as-if random
  2. Affect treatment assignment

IV assumptions

- For an IV to be identified, it must:
    1. Be assigned as-if random
    2. Affect treatment assignment
    3. Only affect outcome through treatment (exclusion restriction)

IV assumptions

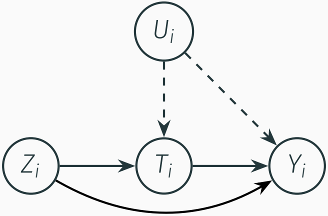
- For an IV to be identified, it must:
    1. Be assigned as-if random
    2. Affect treatment assignment
    3. Only affect outcome through treatment (exclusion restriction)
- IF we meet these assumptions $\rightarrow$ consistent estimate of the local Average Treatment Effect (LATE)

IV assumptions

- For an IV to be identified, it must:
  1. Be assigned as-if random
  2. Affect treatment assignment
  3. Only affect outcome through treatment (exclusion restriction)
- IF we meet these assumptions $\rightarrow$ consistent estimate of the local Average Treatment Effect (LATE)
  - Check-in: the LATE is an estimate of the causal quantity for the compliers, why?

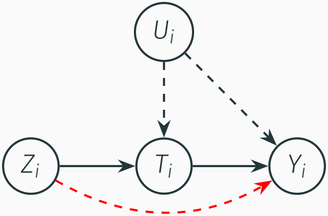
IV assumptions: a DAG

Does this DAG fit our necessary assumptions for a IV strategy?



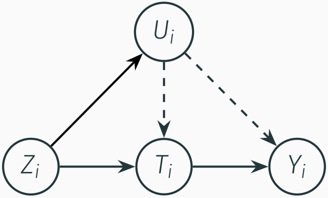
IV assumptions: a DAG

Does this DAG fit our necessary assumptions for a IV strategy?



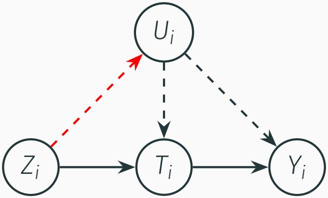
No, it violates the exclusion restriction

What about this DAG? Does it fit our necessary assumptions for a IV strategy?



IV assumptions: a DAG

What about this DAG? Does it fit our necessary assumptions for a IV strategy?



No, because this also violates the exclusion restriction!

The assumptions strike back

Why does the LATE only estimate the causal effect on the compliers?

The assumptions strike back

Why does the LATE only estimate the causal effect on the compliers?

- The exclusion restriction means that always and never takers always get the same treatment

The assumptions strike back

Why does the LATE only estimate the causal effect on the compliers?

- The exclusion restriction means that always and never takers always get the same treatment
- If treatment is static $\rightarrow$ outcomes are consistent

The assumptions strike back

Why does the LATE only estimate the causal effect on the compliers?

- The exclusion restriction means that always and never takers always get the same treatment
- If treatment is static $\rightarrow$ outcomes are consistent
- Monotonicity

The assumptions strike back

Why does the LATE only estimate the causal effect on the compliers?

- The exclusion restriction means that always and never takers always get the same treatment
- If treatment is static $\rightarrow$ outcomes are consistent
- Monotonicity
- Is this a useful estimate? Why or why not?

IV example: voter turnout

Question: What is the effect of voter turnout on Democratic vote share?

IV example: voter turnout

Question: What is the effect of voter turnout on Democratic vote share?

- What's are potential confounders?

IV example: voter turnout

Question: What is the effect of voter turnout on Democratic vote share?

- What's are potential confounders?

Question: What is the effect of voter turnout on Democratic vote share?

- What's are potential confounders? Strategic voters who only turn out because they think Democrats can win

IV example: voter turnout

Question: What is the effect of voter turnout on Democratic vote share?

- What's are potential confounders? Strategic voters who only turn out because they think Democrats can win
- Using an IV approach our estimates tell us the effect of strategic voter turnout when Democrats are favored or unfavored

IV example: voter turnout

Question: What is the effect of voter turnout on Democratic vote share?

- What's are potential confounders? Strategic voters who only turn out because they think Democrats can win
- Using an IV approach our estimates tell us the effect of strategic voter turnout when Democrats are favored or unfavored
- Changing the question: captures measure of Democratic strength

IV example: voter turnout

Question: What is the effect of voter turnout on Democratic vote share?

- What's are potential confounders? Strategic voters who only turn out because they think Democrats can win
- Using an IV approach our estimates tell us the effect of strategic voter turnout when Democrats are favored or unfavored
- Changing the question: captures measure of Democratic strength

IV example: voter turnout

Question: What is the effect of voter turnout on Democratic vote share?

- What's are potential confounders? Strategic voters who only turn out because they think Democrats can win
- Using an IV approach our estimates tell us the effect of strategic voter turnout when Democrats are favored or unfavored
- Changing the question: captures measure of Democratic strength

How do we achieve identification?

IV example: voter turnout

$\rightarrow$ Need an IV to capture variation in voter turnout that is independent from strategic voting

IV example: voter turnout

$\rightarrow$ Need an IV to capture variation in voter turnout that is independent from strategic voting

Let's think back to the IV assumptions:

1. Randomization: treatment is assigned as-if random

IV example: voter turnout

$\rightarrow$ Need an IV to capture variation in voter turnout that is independent from strategic voting

Let's think back to the IV assumptions:

1. Randomization: treatment is assigned as-if random
2. First-stage: IV affects treatment assignment

IV example: voter turnout

$\rightarrow$ Need an IV to capture variation in voter turnout that is independent from strategic voting

Let's think back to the IV assumptions:

1. Randomization: treatment is assigned as-if random
2. First-stage: IV affects treatment assignment
3. Exclusion restriction: IV only affects outcome through treatment

IV example: voter turnout

$\rightarrow$ Need an IV to capture variation in voter turnout that is independent from strategic voting

Let's think back to the IV assumptions:

1. Randomization: treatment is assigned as-if random
2. First-stage: IV affects treatment assignment
3. Exclusion restriction: IV only affects outcome through treatment
4. Monotonicity: no defiers

IV example: voter turnout

$\rightarrow$ Need an IV to capture variation in voter turnout that is independent from strategic voting

Let's think back to the IV assumptions:

1. Randomization:

IV example: voter turnout

$\rightarrow$ Need an IV to capture variation in voter turnout that is independent from strategic voting

Let's think back to the IV assumptions:

1. Randomization:

IV example: voter turnout

$\rightarrow$ Need an IV to capture variation in voter turnout that is independent from strategic voting

Let's think back to the IV assumptions:

1. Randomization: rain is assigned as-if random to districts
2. First-stage:

IV example: voter turnout

$\rightarrow$ Need an IV to capture variation in voter turnout that is independent from strategic voting

Let's think back to the IV assumptions:

1. Randomization: rain is assigned as-if random to districts
2. First-stage:

IV example: voter turnout

$\rightarrow$ Need an IV to capture variation in voter turnout that is independent from strategic voting

Let's think back to the IV assumptions:

1. Randomization: rain is assigned as-if random to districts
2. First-stage: Rain depresses voter turnout
3. Exclusion restriction:

IV example: voter turnout

$\rightarrow$ Need an IV to capture variation in voter turnout that is independent from strategic voting

Let's think back to the IV assumptions:

1. Randomization: rain is assigned as-if random to districts
2. First-stage: Rain depresses voter turnout
3. Exclusion restriction:

IV example: voter turnout

$\rightarrow$ Need an IV to capture variation in voter turnout that is independent from strategic voting

Let's think back to the IV assumptions:

1. Randomization: rain is assigned as-if random to districts
2. First-stage: Rain depresses voter turnout
3. Exclusion restriction: Rain only affects Dem vote share through voter turnout
4. Monotonicity:

IV example: voter turnout

$\rightarrow$ Need an IV to capture variation in voter turnout that is independent from strategic voting

Let's think back to the IV assumptions:

1. Randomization: rain is assigned as-if random to districts
2. First-stage: Rain depresses voter turnout
3. Exclusion restriction: Rain only affects Dem vote share through voter turnout
4. Monotonicity:

IV example: voter turnout

$\rightarrow$ Need an IV to capture variation in voter turnout that is independent from strategic voting

Let's think back to the IV assumptions:

1. Randomization: rain is assigned as-if random to districts
2. First-stage: Rain depresses voter turnout
3. Exclusion restriction: Rain only affects Dem vote share through voter turnout
4. Monotonicity: rain doesn't increase turnout

IV example: voter turnout

$\rightarrow$ Need an IV to capture variation in voter turnout that is independent from strategic voting

Let's think back to the IV assumptions:

1. Randomization: rain is assigned as-if random to districts
2. First-stage: Rain depresses voter turnout
3. Exclusion restriction: Rain only affects Dem vote share through voter turnout
4. Monotonicity: rain doesn't increase turnout

$\Rightarrow$ Candidate strength in any given election is independent (does not affect the variation) in turnout caused by rain

Overview

Using R scripts

- R Scripts are different than R Markdown files

Using R scripts

- R Scripts are different than R Markdown files
- R Markdown files are often used to generate documents, BUT

Using R scripts

- R Scripts are different than R Markdown files
- R Markdown files are often used to generate documents, BUT
    - Not all coding requires a generation of a PDF

Using R scripts

- R Scripts are different than R Markdown files
- R Markdown files are often used to generate documents, BUT
    - Not all coding requires a generation of a PDF
    - May actually hinder our code

Using R scripts

- R Scripts are different than R Markdown files
- R Markdown files are often used to generate documents, BUT
    - Not all coding requires a generation of a PDF
    - May actually hinder our code
- If you want to run a line of code in a script, you don't need to click Run!

Using R scripts

- R Scripts are different than R Markdown files
- R Markdown files are often used to generate documents, BUT
  - Not all coding requires a generation of a PDF
  - May actually hinder our code
- If you want to run a line of code in a script, you don't need to click Run!
  - Mac: Click on the line of interest, CMD + Return

Using R scripts

- R Scripts are different than R Markdown files
- R Markdown files are often used to generate documents, BUT
    - Not all coding requires a generation of a PDF
    - May actually hinder our code
- If you want to run a line of code in a script, you don't need to click Run!
    - Mac: Click on the line of interest, CMD + Return
    - Windows: Click on the line of interest, Ctrl + Enter

Using R scripts

- R Scripts are different than R Markdown files
- R Markdown files are often used to generate documents, BUT
  - Not all coding requires a generation of a PDF
  - May actually hinder our code
- If you want to run a line of code in a script, you don't need to click Run!
  - Mac: Click on the line of interest, CMD + Return
  - Windows: Click on the line of interest, Ctrl + Enter
- Scripts allow us to save our code, AND run the functions in our console

Working directories

- You must tell R where your files are, or what your ”working directory” is.

Working directories

- You must tell R where your files are, or what your ”working directory” is.
- `getwd()` and `setwd()` respectively ”get” your working directory and ”set” your working directory.

Working directories

- You must tell R where your files are, or what your ”working directory” is.
- `getwd()` and `setwd()` respectively ”get” your working directory and ”set” your working directory.
- If you get an error along the lines of ”Cannot establish connection....,” it is because you are loading in data that is not in your working directory.

Setting up a working directory

- You can either set your working directory through code:

```
setwd("/Users/simabiondi/GitHub/teaching/gov51")
```

Setting up a working directory

- You can either set your working directory through code:

```
setwd("/Users/simabiondi/GitHub/teaching/gov51")
```

- OR click `Session` → `Set Working Directory` → `Choose Directory`.

Setting up a working directory

- You can either set your working directory through code:

```
setwd("/Users/simabiondi/GitHub/teaching/gov51")
```

- OR click `Session` → `Set Working Directory` → `Choose Directory`.
- OR use the `here` package:

```
here("teaching", "gov51", "mydataset.csv")
```

Setting up a working directory

- You can either set your working directory through code:

```
setwd("/Users/simabiondi/GitHub/teaching/gov51")
```

- OR click `Session` → `Set Working Directory` → `Choose Directory`.
- OR use the `here` package:

```
here("teaching", "gov51", "mydataset.csv")
```

- *Note*: R Projects set your working directory to the folder that it is in.

Loading data from a URL

- You can also load in data easily from the course website if you have an internet connection:

```
url <- "https://naijialiu.github.io/Gov_51/
    Causal/simulated_iv.csv"
df <- read.csv(url)
```

Why do we use base R?

Advantages of tidyverse: Efficient recall of variable names, consistent within the Tidyverse universe, some unique data wrangling functions

↪ BUT: Clunky in function creation, reliance on package functions

Pedagogical reason: **Reliance on package functions** - think a calculator before learning basic arithmetic

Grammar of base R

Some basic functions and their tidyverse equivalents:

```
# Creating new variable
df$newvar <- 1:10

df <- df |>
  mutate(newvar = 1:10)
```

For more information, check out the tidyverse guide to base R:
`https://dplyr.tidyverse.org/articles/base.html`

Grammar of base R

Some basic functions and their tidyverse equivalents:

```
# Creating a conditional new variable
df$newvar2 <- ifelse(df$newvar %% 2 == 0,
                     1,
                     0)

df <- df |>
  mutate(newvar2 = ifelse(df$newvar %% 2 == 0,
                          1,
                          0))
```

For more information, check out the tidyverse guide to base R:
`https://dplyr.tidyverse.org/articles/base.html`

A quick note about coding

I spend most of my time in section getting everyone up to speed on the concepts, but that doesn't mean that coding isn't important!

A quick note about coding

I spend most of my time in section getting everyone up to speed on the concepts, but that doesn't mean that coding isn't important!

What if I have a coding question?

A quick note about coding

I spend most of my time in section getting everyone up to speed on the concepts, but that doesn't mean that coding isn't important!

What if I have a coding question? $\rightarrow$ *Come to office hours*!

Overview

Now, let's estimate!

Recall: $\rightarrow$ Estimand: LATE = ITT effect on the outcome for compliers

Now, let's estimate!

Recall: $\rightarrow$ Estimand: LATE = ITT effect on the outcome for compliers

- Many estimators exist to estimate the LATE

Now, let's estimate!

Recall: $\rightarrow$ Estimand: LATE = ITT effect on the outcome for compliers

- Many estimators exist to estimate the LATE
- Two of the most popular: (1) Two Stage Least Squares (TSLS) and (2) Wald estimators

Now, let's estimate!

Recall: $\rightarrow$ Estimand: LATE = ITT effect on the outcome for compliers

- Many estimators exist to estimate the LATE
- Two of the most popular: (1) Two Stage Least Squares (TSLS) and (2) Wald estimators
- Logic of of TSLS

Now, let's estimate!

Recall: $\rightarrow$ Estimand: LATE = ITT effect on the outcome for compliers

- Many estimators exist to estimate the LATE
- Two of the most popular: (1) Two Stage Least Squares (TSLS) and (2) Wald estimators
- Logic of of TSLS
    1. Regress treatment (T_i) on the instrument (Z_i)

Now, let's estimate!

Recall: $\rightarrow$ Estimand: LATE = ITT effect on the outcome for compliers

- Many estimators exist to estimate the LATE
- Two of the most popular: (1) Two Stage Least Squares (TSLS) and (2) Wald estimators
- Logic of of TSLS
    1. Regress treatment (T_i) on the instrument (Z_i)
    2. Regress outcome of interest (Y_i) on the fitted values ($\hat{T}_i$) generated in the previous stage

Now, let's estimate!

Recall: $\rightarrow$ Estimand: LATE = ITT effect on the outcome for compliers

- Many estimators exist to estimate the LATE
- Two of the most popular: (1) Two Stage Least Squares (TSLS) and (2) Wald estimators
- Logic of of TSLS
    1. Regress treatment (T_i) on the instrument (Z_i)
    2. Regress outcome of interest (Y_i) on the fitted values ($\hat{T}_i$) generated in the previous stage
- IMPORTANT: we are estimating the LATE, not the ATE,

Now, let's estimate!

Recall: $\rightarrow$ Estimand: LATE = ITT effect on the outcome for compliers

- Many estimators exist to estimate the LATE
- Two of the most popular: (1) Two Stage Least Squares (TSLS) and (2) Wald estimators
- Logic of of TSLS
    1. Regress treatment (T_i) on the instrument (Z_i)
    2. Regress outcome of interest (Y_i) on the fitted values ($\hat{T}_i$) generated in the previous stage
- IMPORTANT: we are estimating the LATE, not the ATE,

Now, let's estimate!

Recall: $\rightarrow$ Estimand: LATE = ITT effect on the outcome for compliers

- Many estimators exist to estimate the LATE
- Two of the most popular: (1) Two Stage Least Squares (TSLS) and (2) Wald estimators
- Logic of of TSLS
  1. Regress treatment (T_i) on the instrument (Z_i)
  2. Regress outcome of interest (Y_i) on the fitted values ($\hat{T}_i$) generated in the previous stage
- IMPORTANT: we are estimating the LATE, not the ATE, *why?*

Now, let's estimate!

Recall: $\rightarrow$ Estimand: LATE = ITT effect on the outcome for compliers

- Many estimators exist to estimate the LATE
- Two of the most popular: (1) Two Stage Least Squares (TSLS) and (2) Wald estimators
- Logic of of TSLS
    1. Regress treatment (T_i) on the instrument (Z_i)
    2. Regress outcome of interest (Y_i) on the fitted values ($\hat{T}_i$) generated in the previous stage
- IMPORTANT: we are estimating the LATE, not the ATE, *why?*
    - $\rightarrow$ Can't estimate ATE because we don't know the proportions of compliers, always, and never takers

IV example: estimation using Wald estimator

The Wald Estimator estimates the LATE among compliers

$$\frac{\widehat{ITT}}{\widehat{Encouragement}} = \frac{\widehat{ITT_Y}}{\widehat{ITT_T}}$$

$$= \frac{E[Y_i(Z_i = 1] - E[Y_i(Z_i = 0)]}{E[T_i(Z_i = 1] - E[T_i(Z_i = 0)]}$$

What does this mean in English?

IV example: estimation using Wald estimator

The Wald Estimator estimates the LATE among compliers

$$\frac{\widehat{ITT}}{\widehat{Encouragement}} = \frac{\widehat{ITT_Y}}{\widehat{ITT_T}}$$

$$= \frac{E[Y_i(Z_i = 1] - E[Y_i(Z_i = 0)]}{E[T_i(Z_i = 1] - E[T_i(Z_i = 0)]}$$

What does this mean in English?

Let's go to R!

IV example: using Wald estimator

```
set.seed(02138)
mydf <- data.frame(draft = rbinom(20, 1, 0.5),
                   military = rbinom(20, 1, 0.3),
                   earning = rnorm(20, 10000, sd =
                       5000))
summary(mydf)
dim(mydf)
ITT <- mean(mydf$earning[mydf$draft == 1]) -
  mean(mydf$earning[mydf$draft == 0])
Encouragement <- mean(mydf$military[mydf$draft ==
    1]) -
  mean(mydf$military[mydf$draft == 0])
tauhat <- ITT/Encouragement
```

Estimate of the effect of military service on lifetime earnings (for compliers) is `r round(tauhat,2)`

IV example: wrap-up

- Endogeneity concerns are real! Our estimates of military service on lifetime earnings are clearly affected by confounding variables

IV example: wrap-up

- Endogeneity concerns are real! Our estimates of military service on lifetime earnings are clearly affected by confounding variables

IV example: wrap-up

- Endogeneity concerns are real! Our estimates of military service on lifetime earnings are clearly affected by confounding variables
    - -6261.08 vs. 583.49 is huge!!

IV example: wrap-up

- Endogeneity concerns are real! Our estimates of military service on lifetime earnings are clearly affected by confounding variables
    - -6261.08 vs. 583.49 is huge!!
- Instrumental variables are a useful strategy to achieve identification of an estimand

IV example: wrap-up

- Endogeneity concerns are real! Our estimates of military service on lifetime earnings are clearly affected by confounding variables
    - -6261.08 vs. 583.49 is huge!!
- Instrumental variables are a useful strategy to achieve identification of an estimand
- Finding a good instrument is difficult, as the assumptions are stringent

Next week

- Matching
- In the background: start brainstorming and talking to classmates