# Quant II

## Lab 1

Yinxuan Wang

2025-01-29

# Hi!

- Yinxuan Wang, $4^{th}$ year PhD.
- Fields: Methods, Comparative Politics
- Email: yinxuan.wang@nyu.edu
- Office: 420

# Hi!

- Yinxuan Wang, $4^{th}$ year PhD.
- Fields: Methods, Comparative Politics
- Email: yinxuan.wang@nyu.edu
- Office: 420

- What do you want to get out of Quant II?

# Logistics

- Lab: Thursday, 10 am - 12 pm EST, Room 212
- Lab materials will be posted on the lab's GitHub repo: https://github.com/yinxuanwang/quant2-labs-spring2026
- Office hours: by appointment

# Logistics

- Lab: Thursday, 10 am - 12 pm EST, Room 212
- Lab materials will be posted on the lab's GitHub repo: https://github.com/yinxuanwang/quant2-labs-spring2026
- Office hours: by appointment

- Homework due via email to Cyrus and me by the indicated deadline
- Deadline is *strict*
- Submit **PDF document** with **code used** embedded in the document

# Some purposes of lab

- Build intuition and motivation
- Review and extend
- Ask questions
- Learn how to do the analysis we are learning about (i.e., in R)

# Today's Lab

- Getting set up with RStudio and Quarto
- Potential outcomes and ATE
- DAG and Do-calculus

# Quarto

- Tool that combines R, LaTeX, and Markdown
  - 'Next-generation' of RMarkdown
  - Easy integration with other languages, e.g. Python
- Create **reproducible** documents
- Combine text, code, and analysis results
- Your homework should be prepared using Quarto or similar tools
- Code should be clean, well named, and properly formatted

# Some useful packages

**Here**:

- No file paths in code
- Works either with `.Rproj` or `.here` file

# Some useful packages

**Here**:

- No file paths in code
- Works either with `.Rproj` or `.here` file

**Pacman**

- R package for package management
- `pacman::p_load()` loads or installs automatically
- `pacman::p_install_version()` installs a specific version

# Some useful packages

**Here**:

- No file paths in code
- Works either with `.Rproj` or `.here` file

**Pacman**

- R package for package management
- `pacman::p_load()` loads or installs automatically
- `pacman::p_install_version()` installs a specific version

**Tables**

- `modelsummary`, `stargazer`: regression tables
- `kable`, `kableExtra`: easy LaTeX/HTML table styling

# Potential Outcomes Framework

- **Potential outcomes** formally encode counterfactuals
  - $Y_i(1)$: outcome that unit $i$ would have if treated;
  - $Y_i(0)$: outcome that unit $i$ would have if untreated

# Potential Outcomes Framework

- **Potential outcomes** formally encode counterfactuals
  - $Y_i(1)$: outcome that unit $i$ would have if treated;
  - $Y_i(0)$: outcome that unit $i$ would have if untreated
- **Unit-level treatment effect**: $\rho_i = Y_i(1) - Y_i(0)$
- **ATE**: $\rho = E[Y_i(1) - Y_i(0)]$

# Potential Outcomes Framework

- **Potential outcomes** formally encode counterfactuals
  - $Y_i(1)$: outcome that unit $i$ would have if treated;
  - $Y_i(0)$: outcome that unit $i$ would have if untreated
- **Unit-level treatment effect**: $\rho_i = Y_i(1) - Y_i(0)$
- **ATE**: $\rho = E[Y_i(1) - Y_i(0)]$
- Connect **observed outcomes** to potential outcomes
  - $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$
- Expected difference in means
  - $E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0]$
    $= E[Y_i(1) \mid D_i = 1] - E[Y_i(0) \mid Di = 0]$

# Difference-in-Means and ATE

$$E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = \text{ATT} + \text{Selection bias w.r.t. } Y_0$$

# Difference-in-Means and ATE

$$E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = \text{ATT} + \text{Selection bias w.r.t. } Y_0$$

$$E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = \text{ATC} + \text{Selection bias w.r.t. } Y_1$$

# Difference-in-Means and ATE

$$E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = \text{ATT} + \text{Selection bias w.r.t. } Y_0$$

$$E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = \text{ATC} + \text{Selection bias w.r.t. } Y_1$$

$$\begin{aligned} E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = & \underbrace{\rho}_{\text{Average treatment effect}} \\ & + \underbrace{E[Y_i(0) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0]}_{\text{Selection bias w.r.t. } Y_0} \\ & + \underbrace{(1-\pi)\Big(E[\rho \mid D_i = 1] - E[\rho \mid D_i = 0]\Big)}_{\text{Selection bias w.r.t. } \rho}. \end{aligned}$$

# Lab Activities

- [See the `lab01_exercise.qmd` in the Github]

# DAGs

- **DAG** stands for Directed Acyclic Graph

# DAGs

- **DAG** stands for Directed Acyclic Graph
  - *Directed*: No reverse causality or simultaneity;

# DAGs

- **DAG** stands for Directed Acyclic Graph
  - *Directed*: No reverse causality or simultaneity;
  - *Acyclic*: No cycles

# DAGs

- **DAG** stands for Directed Acyclic Graph
  - *Directed*: No reverse causality or simultaneity;
  - *Acyclic*: No cycles
- Representation of the data generating process (DGP)

# DAGs

- **DAG** stands for Directed Acyclic Graph
  - *Directed*: No reverse causality or simultaneity;
  - *Acyclic*: No cycles
- Representation of the data generating process (DGP)
  - Nodes ($X$,$D$,$Y$ etc.) are random variables

# DAGs

- **DAG** stands for Directed Acyclic Graph
  - *Directed*: No reverse causality or simultaneity;
  - *Acyclic*: No cycles
- Representation of the data generating process (DGP)
  - Nodes ($X$,$D$,$Y$ etc.) are random variables
  - Edges ($X \rightarrow Y$) denote a direct causal effect of $X$ on $Y$

# DAGs

- **DAG** stands for Directed Acyclic Graph
  - *Directed*: No reverse causality or simultaneity;
  - *Acyclic*: No cycles
- Representation of the data generating process (DGP)
  - Nodes ($X$,$D$,$Y$ etc.) are random variables
  - Edges ($X \rightarrow Y$) denote a direct causal effect of $X$ on $Y$
- Tools to help understand whether a research design can identify a causal relationship

# DAGs

- **DAG** stands for Directed Acyclic Graph
  - *Directed*: No reverse causality or simultaneity;
  - *Acyclic*: No cycles
- Representation of the data generating process (DGP)
  - Nodes ($X$,$D$,$Y$ etc.) are random variables
  - Edges ($X \rightarrow Y$) denote a direct causal effect of $X$ on $Y$
- Tools to help understand whether a research design can identify a causal relationship
  - No assumptions about the functional form or distribution.

# The Simulated Data, i

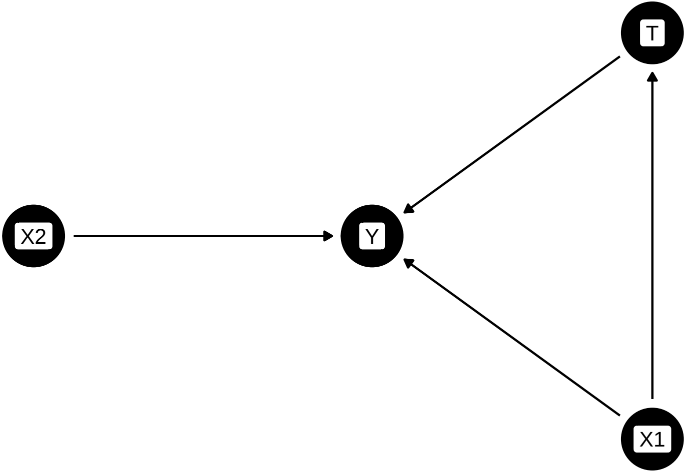
```r
set.seed(123)
N <- 1000

# 2 random covariates
x1 <- runif(N, 0, 1)
x2 <- rnorm(N, 0, 0.5)

# Some noise
e <- rnorm(N, 0, 1)

# Treatment effect
d <- rnorm(N, 1, 1)

# Potential outcomes
y0 <- x1 * 4 + x2 + e
y1 <- y0 + d
```

# The Simulated Data, ii

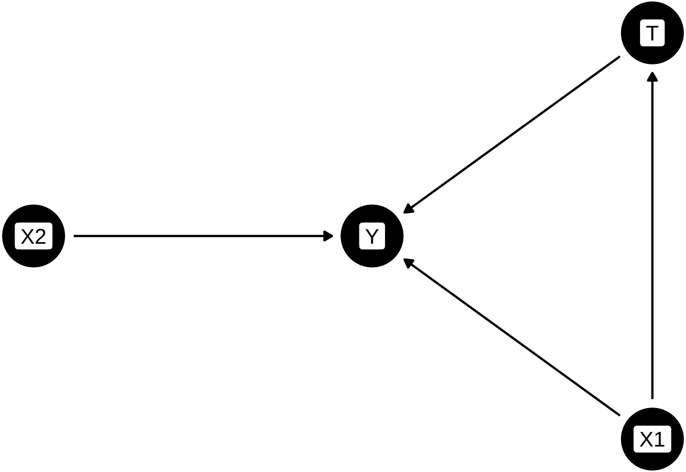
```
# Treatment assignment, confounded by x1
ts <- rbinom(n = N, size = 1, prob = x1)

# Observed outcome
y <- ts * y1 + (1-ts) * y0
```

# DAG

# DAG

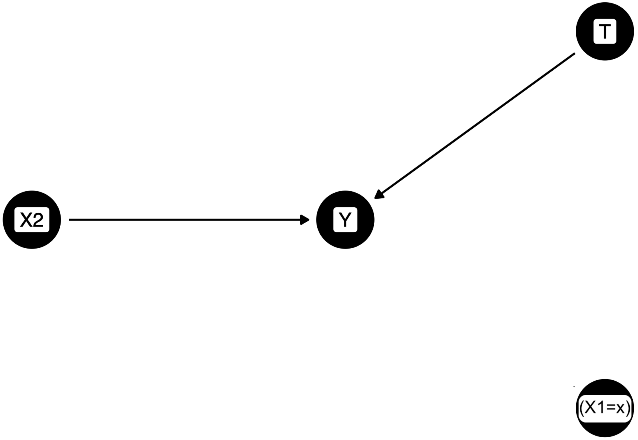


# DAG




Conditioning strategy to identify the effect of D on Y?

# DAG, ii



# Do-operator

Target of inference is based on an hypothetical intervention

$$\rho = E[Y_i \mid do(D_i = 1)] - E[Y_i \mid do(D_i = 0)]$$

# Do-operator

Target of inference is based on an hypothetical intervention

$$\rho = E[Y_i \mid do(D_i = 1)] - E[Y_i \mid do(D_i = 0)]$$

**Intervention distribution**: $P(Y \mid do(D = d))$

$$E[Y \mid do(D = d)] = \sum_y y\, P(Y = y \mid do(D = d))$$

$$(\text{or } \int y f(y \mid do(D = d)) dy)$$

# Do-operator

Target of inference is based on an hypothetical intervention

$$\rho = E[Y_i \mid do(D_i = 1)] - E[Y_i \mid do(D_i = 0)]$$

**Intervention distribution**: $P(Y \mid do(D = d))$

$$E[Y \mid do(D = d)] = \sum_y y\,P(Y = y \mid do(D = d))$$

$$(\text{or } \int y f(y \mid do(D = d)) dy)$$

Identifiability: Can we write the intervention distribution using only observed quantities?

- Trying to remove $do()$ operators

# Do-calculus, notations

[Cyrus Slide 2-23]

Let X, Y, Z and W be arbirary disjoint sets of nodes in a causal DAG, G

$G_{\overline{X}}$: Graph obtained by deleting from $G$ all arrows pointing to $X$

$G_{\underline{X}}$: Graph obtained by deleting from $G$ all arrows emerging from $X$

# Do-calculus, rules

[Cyrus Slide 2-24]

**Rule 1 (Insertion/deletion of observations)**

$$P(y \mid \mathrm{do}(x), z, w) = P(y \mid \mathrm{do}(x), w) \quad \text{if } Y \perp\!\!\!\perp Z \mid X, W \text{ in } G_{\overline{X}}$$

# Do-calculus, rules

[Cyrus Slide 2-24]

**Rule 1 (Insertion/deletion of observations)**

$$P(y \mid \mathrm{do}(x), z, w) = P(y \mid \mathrm{do}(x), w) \quad \text{if } Y \perp\!\!\!\perp Z \mid X, W \text{ in } G_{\overline{X}}$$

**Rule 2 (Action/observation exchange)**

$$P(y \mid \mathrm{do}(x), \mathrm{do}(z), w) = P(y \mid \mathrm{do}(x), z, w) \quad \text{if } Y \perp\!\!\!\perp Z \mid X, W \text{ in } G_{\overline{X}\,\underline{Z}}$$

# Do-calculus, rules

[Cyrus Slide 2-24]

**Rule 1 (Insertion/deletion of observations)**

$$P(y \mid \mathrm{do}(x), z, w) = P(y \mid \mathrm{do}(x), w) \quad \text{if } Y \perp\!\!\!\perp Z \mid X, W \text{ in } G_{\overline{X}}$$

**Rule 2 (Action/observation exchange)**

$$P(y \mid \mathrm{do}(x), \mathrm{do}(z), w) = P(y \mid \mathrm{do}(x), z, w) \quad \text{if } Y \perp\!\!\!\perp Z \mid X, W \text{ in } G_{\overline{X}\,\underline{Z}}$$

**Rule 3 (Insertion/deletion of actions)**

$$P(y \mid \mathrm{do}(x), \mathrm{do}(z), w) = P(y \mid \mathrm{do}(x), w) \quad \text{if } Y \perp\!\!\!\perp Z \mid X, W \text{ in } G_{\overline{X}\,\overline{Z(W)}}$$

$Z(W)$ is the set of Z-nodes that are not ancestors of any $W$-nodes in $G_{\overline{X}}$

# Do-calculus, example

Consider a discrete example

$$P(y \mid do(d)) = \sum_x P(y, x \mid do(d)) = \sum_x P(y \mid x, do(d)) P(x \mid do(d))$$

# Do-calculus, example

Consider a discrete example

$P(y \mid do(d)) = \sum_x P(y, x \mid do(d)) = \sum_x P(y \mid x, do(d)) P(x \mid do(d))$

$Y \perp\!\!\!\perp D|X$ in $G_{\underline{D}}$, so we can apply Rule 2

$P(Y \mid x, do(d)) = P(y \mid x, d)$

# Do-calculus, example

Consider a discrete example

$P(y \mid do(d)) = \sum_x P(y, x \mid do(d)) = \sum_x P(y \mid x, do(d)) P(x \mid do(d))$

$Y \perp\!\!\!\perp D | X$ in $G_{\underline{D}}$, so we can apply Rule 2

$P(Y \mid x, do(d)) = P(y \mid x, d)$

$X \perp\!\!\!\perp D$ in $G_{\overline{D}}$, so we can apply Rule 3

$P(x \mid do(d)) = P(x)$

# Do-calculus, example

Consider a discrete example

$P(y \mid do(d)) = \sum_x P(y, x \mid do(d)) = \sum_x P(y \mid x, do(d))P(x \mid do(d))$

$Y \perp\!\!\!\perp D|X$ in $G_{\underline{D}}$, so we can apply Rule 2

$P(Y \mid x, do(d)) = P(y \mid x, d)$

$X \perp\!\!\!\perp D$ in $G_{\overline{D}}$, so we can apply Rule 3

$P(x \mid do(d)) = P(x)$

Then, $P(y \mid do(d)) = \sum_x P(y \mid x, d)P(x)$