Research Design & Causal Inference

ECON 692 – Applied Economics Seminar

Andrew Hobbs

University of San Francisco

February 19, 2026

Today's Agenda

Time	Duration	Activity
4:35	5 min	Welcome & check-in
4:40	35 min	Potential outcomes & DAG theory
5:15	25 min	Practical DAG skills
5:40	10 min	Break
5:50	20 min	DAG software demo
6:10	30 min	DAG workshop
6:40	10 min	Break
6:50	55 min	Empirical strategy drafting
7:45	20 min	Share-outs
8:05	10 min	Wrap-up

Check-in

- How is your data coming together?
- Has anyone hit a wall with their data source?
- Has anyone started sketching your causal story?

Reminder

Data Report is due **tomorrow, Friday February 20** at 11:59pm.

Potential Outcomes

The Setup: Paid Family Leave

Research question: Does access to paid family leave increase female employment?

- Unit i: a woman in a particular state and year
- Treatment: $D_i = 1$ if her state has paid family leave; 0 otherwise
- Outcome: $Y_i = 1$ if employed

Potential outcomes:

- $Y_i(1)$: her employment *if* her state **has** PFL
- $Y_i(0)$: her employment *if* her state does **not** have PFL

We observe: $Y_i = D_i \cdot Y_i(1) + (1 - D_i) \cdot Y_i(0)$ – only **one** potential outcome per unit.

The Fundamental Problem

Individual treatment effect: $\tau_i = Y_i(1) - Y_i(0)$

This is **never observed –** we cannot put the same woman simultaneously in a PFL state and a non-PFL state.

So we estimate averages:

$\text{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)]$	Effect averaged over *everyone*
$\text{ATT} = \mathbb{E}[Y_i(1) - Y_i(0) \mid D_i = 1]$	Effect averaged over the *treated*

For policy evaluation, we usually want the **ATT**.

"Did PFL raise employment for women in states that adopted it?"

Selection Bias

The naive comparison mixes up the treatment effect with pre-existing differences:

$$\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] \;=\; \mathbb{E}[Y_i(1) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 0]$$

Add and subtract $\mathbb{E}[Y_i(0) \mid D_i = 1]$:

$$= \underbrace{\mathbb{E}[Y_i(1) - Y_i(0) \mid D_i = 1]}_{\text{ATT (what we want)}} + \underbrace{\mathbb{E}[Y_i(0) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 0]}_{\text{Selection bias}}$$

Selection bias for PFL: states that adopt PFL may have had higher female employment *even without* the policy – liberal, urban states pass PFL *and* have stronger female labor markets.

The naive comparison cannot tell us whether high employment in PFL states reflects τ or pre-existing advantages.

Where Does Bias Come From?

Potential outcomes **define** the target and **name** the problem.

But they don't tell us *where* the bias comes from or *what to condition on* to remove it.

Potential outcomes	**DAGs**
Define the estimand (ATE, ATT)	Map the data-generating process
Name the problem (selection bias)	Show *where* bias enters
What to estimate	*How* to identify it

Use potential outcomes to define your target. Use DAGs to design your identification strategy.

Directed Acyclic Graphs

DAG Building Blocks

A **Directed Acyclic Graph** is a map of your causal assumptions.

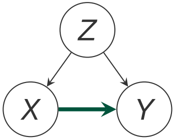
- **Nodes**: variables (observed or unobserved)
- **Directed edges**: $X \to Y$ means “X has a direct causal effect on Y, holding all other variables fixed”
- **Acyclic**: no feedback loops

Key insight

Every missing arrow is also an assumption: “these two things are not directly connected.” A DAG encodes your **substantive theory**, not just your statistical model.

Three Fundamental Structures

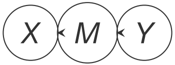
Fork (Confounder)



Z opens a backdoor path

Block it:
condition on Z

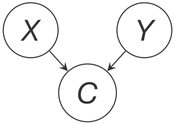
Chain (Mediator)



X affects Y through M

Caution:
conditioning on M
blocks this path

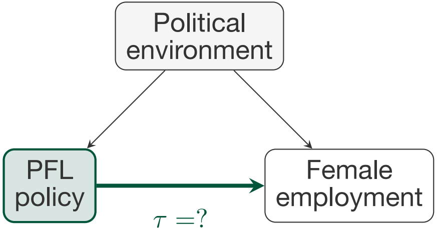
Collider



Both X and Y cause C

Danger:
conditioning on C
opens a spurious path

The Fork: Confounders

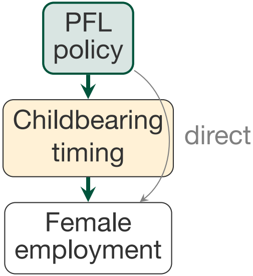


Backdoor path: PFL ← POL → EMP

- Liberal states adopt PFL earlier
- Liberal states also have stronger female labor markets for *other* reasons
- Naive comparison cannot separate τ from the POL effect

Block the backdoor: condition on political environment to isolate τ
In practice: state fixed effects, vote-share controls, ideology index

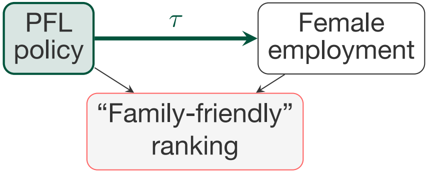
The Chain: Mediators



- PFL enables women to have children *and* remain employed – a core mechanism
- Both paths are genuine effects of PFL; controlling for childbearing blocks one and **underestimates** the total effect

Rule: do **not** condition on mediators unless you want the *direct* effect only, net of the mechanism.

The Collider: A Subtle Danger

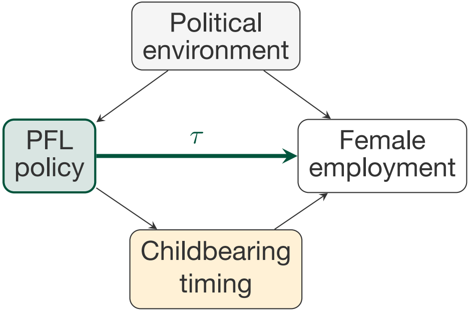


Both PFL adoption and high female employment feed into “family-friendly” rankings.

- Collider paths are **closed by default** – no bias
- Restricting your sample to highly-ranked states conditions on the collider – opening a spurious path

Rule: do not condition on variables caused by both X and Y – e.g., selecting on outcomes or post-treatment variables.

The PFL DAG



Backdoor path (bias):

- PFL $\leftarrow$ POL $\rightarrow$ EMP

Mediator (don't control):

- PFL $\rightarrow$ CHILD $\rightarrow$ EMP

Minimum adjustment set:

- **{Political environment}**

Controlling for POL closes the backdoor path without blocking the treatment effect.

The Backdoor Criterion

Backdoor Criterion (Pearl, 2009)

A set of variables $\mathbb{Z}$ identifies the causal effect of X on Y if:

1. No variable in $\mathbb{Z}$ is a **descendant** of X (no mediators or outcomes)
2. $\mathbb{Z}$ **blocks every path** between X and Y that begins with an arrow *into* X

Applied to PFL:

- Backdoor path: PFL $\leftarrow$ POL $\rightarrow$ EMP
- Adjustment set $\{$POL$\}$: blocks the path; POL is not a descendant of PFL ✓
- **Do not** include CHILD – descendant of PFL (violates condition 1)

Practical shortcut: `dagitty.net` finds the adjustment set automatically.

DAGs and Parallel Trends

Parallel trends requires: absent PFL, treated and control states would have followed the same employment trajectory.

When does this fail? When a confounder Z satisfies all three:

1. $Z \rightarrow$ PFL adoption *timing*
2. $Z \rightarrow$ EMP over *time* (affects trends, not just levels)
3. Z is not controlled for

In the PFL DAG: political environment drove both adoption timing and employment growth – early adopters will diverge from controls before the policy, which a pre-trends test detects.

Practical rule: check your DAG before running your event study. Every time-varying confounder is a candidate for a parallel trends violation – and a robustness check to add.

DAG Software

dagitty.net

Free, browser-based – no installation required.

What it does:

- Draw nodes and edges with point-and-click
- Specify treatment, outcome, observed variables
- Automatically finds: adjustment sets, testable implications, all open paths

Quick start:

1. Go to `dagitty.net`
2. *Model* → *New Model*
3. Click to add nodes; drag to connect
4. Right-click a node to mark it as treatment or outcome
5. Read the **Adjustment** panel

Use this during the workshop: enter your project DAG and let dagitty find your adjustment set.

ggdag in R

For publication-quality figures in your paper or final presentation.

```
library(ggdag)

pfl_dag <- dagify(EMP ~ PFL + POL + CHILD,
                  PFL ~ POL, CHILD ~ PFL,
                  exposure = "PFL", outcome = "EMP")

ggdag(pfl_dag, layout = "sugiyama") + theme_dag()
ggdag_adjustment_set(pfl_dag)
```

`install.packages("ggdag")` – also installs `dagitty` as a dependency.

DAG Workshop

Draw Your DAG

Step 1 – Individual (15 minutes)

For your own research project, draw a DAG that includes:

- Your **treatment** and **outcome**, clearly labeled
- At least **two confounders** (forks – things that affect both treatment and outcome)
- At least **one mediator**, if applicable (a mechanism through which treatment operates)
- Mark what you plan to **condition on** – and why

Use paper, a whiteboard, or `dagitty.net`.

Step 2 – Groups of 3 (15 minutes)

Share your DAG. For each person, the group asks:

- What is the main backdoor path?
- Does the adjustment set satisfy the backdoor criterion?
- Is anything being controlled that shouldn't be (a mediator or collider)?

Empirical Strategy Drafting

Drafting Sprint

55 minutes. This starts your **Empirical Strategy Draft** – due **March 6**, a separate submission from the Data Report.

Write **2–3 paragraphs** covering:

1. **Identification strategy** – what variation are you exploiting? Why is it plausibly exogenous? Name the method and the specific source of variation.
2. **Causal diagram** – what does your DAG imply about what you need to control for? Be explicit: which confounders will you include and why? Which variables will you *not* control for, and why not?
3. **Main threat and response** – what is the single biggest threat to your identification? How will you address or test it (robustness check, placebo test, pre-trends plot)?

Goal: a rough draft to build on over the next two weeks. I'll be circulating.

Share-outs

3–4 students, 5 minutes each.

Cover:

1. **Research question** – one sentence
2. **Key finding from your DAG** – what was the most important backdoor path you identified? How are you addressing it?
3. **Main identification threat** – and your plan for testing or addressing it

Listeners: one piece of feedback each.

Wrap-Up & Next Week

Due this week:

- **Data Report – due tomorrow, Friday February 20**
- Weekly progress report (Friday)

Next week (Week 5):

- Data cleaning and merging for reproducibility
- Hands-on implementation with your own data

Before next class:

- Submit your Data Report
- Make sure your data loads cleanly in R or Python
- Push your current code to GitHub
- Revisit your DAG – refine it based on today's feedback