Statistical Modeling and Causal Inference with R

Week 4: Causal Graphs

Manuel Bosancianu

Max Schaub

September 28, 2020

Hertie School

Today's focus

- ✓ Causal graphs: components and terminology
- ✓ Rules for independence: *d-separation*
- ✓ Examples!

Last week

- Linear regression as the most "used and abused" method in statistical inference
- OLS ('ordinary least squares") as the estimation method for β: the treatment effect (NATE)
- Mechanics: finding the set of values which minimize the sum of squared residuals.

$$Minimize : \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \underbrace{\hat{\beta}_1}_{NATE} X_i)^2 \tag{1}$$

Last week

$$\underbrace{E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{NATE} = \underbrace{\kappa}_{ATE} + \underbrace{E[u_i|D_i = 1] - E[u_i|D_i = 0]}_{Selection\ bias} \tag{2}$$

Selection bias is a threat to recovering the ATE.

Regression can incorporate additional variables, in the quest for removing selection bias.

Limitations: some confounders are difficult to measure, or are not present in the data.

Causal graphs

Why bother with graphs?

Controlling for confounders can eliminate *selection bias*, but how do we <u>choose</u> the confounders?

Should we control for everything we can measure?

Causal graphs:

- ✓ provide an concise account of the DGP
- ✓ allow us to decide which controls to include
- ✓ offer a framework to discuss all aspects of causal inference (experimental and observational)

Building blocks

Variables are the nodes (or *vertices*) of the graph.

Links between nodes are called edges (or *arcs*).

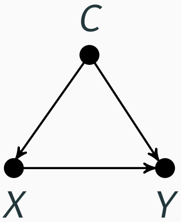
Implications

$$X \longrightarrow Y$$

- time flows from left to right
- presence of an edge means a direct causal effect from X to Y
- absence of an edge suggests the *lack* of a direct causal effect

Directed edge means X is parent and Y is child.

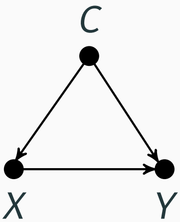
More complexity



Nodes that are only parents are called exogenous (or *root* nodes).

Nodes that are both parent and child are called endogenous.

DAGs

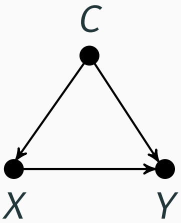


We focus on a special type of causal model: the directed acyclic graph.

- directed: the edges all have directions
- acyclic: variable cannot cause itself, through another variable (or variables)

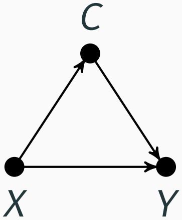
Typology of nodes

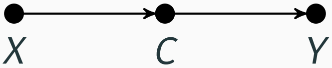
A few foundational configurations in a DAG.



C is a confounder: causal impact on both treatment assignment and outcome.

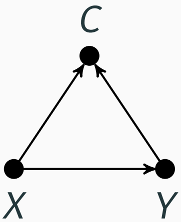
Typology of nodes





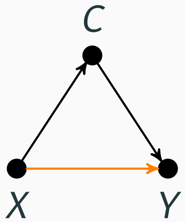
C is a mediator: carries part or all of the effect of X on Y.

Typology of nodes

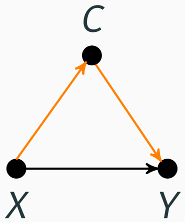


C is a collider: a common child node of at least two parent nodes.

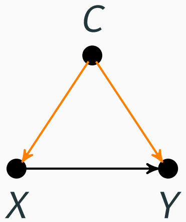
Paths



(a) Direct path



(b) Front-door path

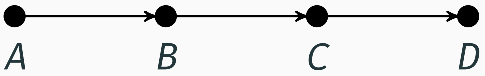


(c) Back-door path

Whether a causal path is open or closed depends on:

1. whether or not we control for nodes on the path in our analysis
2. what type of nodes we control for (colliders, confounders, or mediators)

Paths



With longer paths, we have ancestors and descendants.

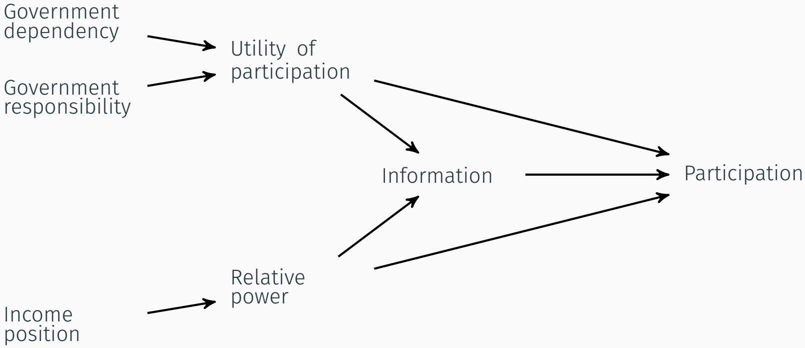
A is an ancestor for B, C, and D. The latter are all descendants for A.

The path from A to D is causal. The path from P to S is non-causal.

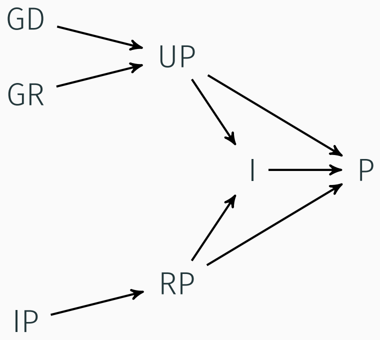
Example #1

Relative power theory

Proposed by Goodin and Dryzek (1980) as an explanation for why poorer people participate less in politics.



Stylized



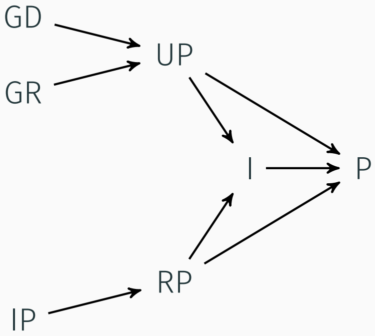
How many paths from RP to P?

- ✓ RP $\rightarrow$ P
- ✓ RP $\rightarrow$ I $\rightarrow$ P
- ✓ RP $\rightarrow$ I $\leftarrow$ UP $\rightarrow$ P

How many causal/ non-causal?

- ✓ 1 & 2 causal
- ✓ 3 non-causal

Stylized



How many exogenous nodes?

3: GD, GR, IP

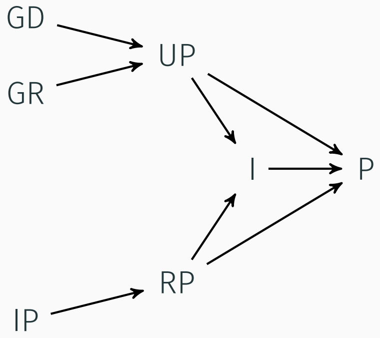
How many endogenous?

4: UP, RP, I, P

How many back-door paths (from RP to P)?

In this case, none.

Stylized



How many colliders?

2: I, UP

How many confounders (if we want the effect of I on P)?

2: UP, RP

How many mediators (if we want the effect of IP on P)?

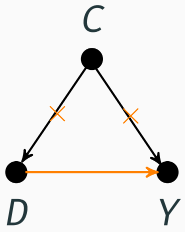
2: RP, I

d-separation

Linkage to regression

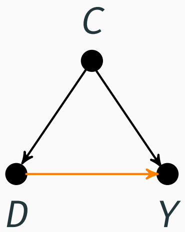
β is unbiased if $E[u_i|D_i] = 0$ ("0 conditional mean assumption").

Reformulating: treatment assignment should be unaffected by selection bias.



Treatment assignment should be as good as random.

Closing open back-door paths



$D \rightarrow Y$ is what we're trying to capture.

$D \leftarrow C \rightarrow Y$ is a spurious source of association between D and Y.

Back-door criterion

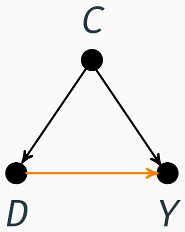
To compute $NATE = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0]$, we need to isolate the direct path: $D \rightarrow Y$.

Open backdoor paths create spurious correlations between $\boldsymbol{D}$ and $\boldsymbol{Y}$.

2 strategies to close them:

- ✓ If path contains confounder, condition on confounder
- ✓ if path contains collider it is already closed, so do not condition on collider

Canonical configurations: confounders

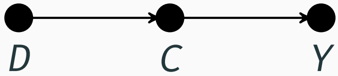


$D \leftarrow C \rightarrow Y$ is an open back-door path (C is not a collider).

Solution: condition on C to close the path.

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 C_i + \epsilon_i \tag{3}$$

Canonical configurations: mediators



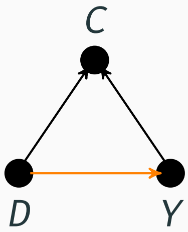
$D \rightarrow C \rightarrow Y$: causal path between treatment and outcome.

Solution: do not condition on C, to keep the path open.

Rules:

- ✓ close all non-causal paths linking D to Y
- ✓ do not close any causal path between D and Y

Canonical configurations: colliders



$D \rightarrow C \leftarrow Y$: noncausal path between treatment and outcome.

Solution: do not condition on C, as path is closed already.

Don't condition on a descendant of a collider, either (Pearl, Glymour, & Jewell, 2016, p. 44-45).

Strategy for d-separation

A few simple steps (Cunningham, 2021, p. 73):

1. write down all paths between D and Y
2. identify open/closed back-door paths (any confounders or colliders?)
3. find conditioning strategy that closes all open back-doors

Last step is not always possible.

When all non-causal paths between D and Y are blocked, D and Y are d-separated.

Example #2

Google gender pay discrimination

Google accused of 'extreme' gender pay discrimination by US labor department

Allegations of possible employment violations emerge at court hearing as part of lawsuit to compel company, a federal contractor, to provide compensation data

Sam Levin *in San Francisco*

@SamTLevin Email

Fri 7 Apr 2017 23.48 BST

3,301

▲ Google is a federal contractor, which means it is required to allow the DoL to inspect and copy records and information about its compliance with equal opportunity laws. Photograph: Thomas Trutschel/Photothek via Getty Images

Google has discriminated against its female employees, according to the US Department of Labor (DoL), which said it had evidence of "systemic compensation disparities".

As part of an ongoing DoL investigation, the government has collected information that suggests the internet search giant is violating federal employment laws with its salaries for women, agency officials said.

Google gender pay discrimination

Lisa Barnett Sween, one of Google's attorneys, testified in opening remarks that the DoL's request constituted a "fishing expedition that has absolutely no relevance to the compliance review". She said the request was an unconstitutional violation of the company's fourth amendment right to protection from unreasonable searches.

Marc Pilotin, a DoL attorney, said: "For some reason or another, Google wants to hide the pay-related information."

In a statement to the Guardian, Google said: "We vehemently disagree with [Wipper's] claim. Every year, we do a comprehensive and robust analysis of pay across genders and we have found no gender pay gap. Other than making an unfounded statement which we heard for the first time in court, the DoL hasn't provided any data, or shared its methodology."

The company has recently claimed that it has closed its gender pay gap globally and provides equal pay across races in the US.

As a federal contractor Google has to share data relevant to equal opportunity law compliance, if asked.

The company initially resisted the DoL request.

Google performs internal analysis

At Google, Employee-Led Effort Finds Men Are Paid More Than Women

Zitation exportieren

The New York Times

September 8, 2017 Friday 00:00 EST

Section: TECHNOLOGY

Length: 1465 words

Byline: DAISUKE WAKABAYASHI

Highlight: A spreadsheet created by employees to share salary information shows **pay** for women is falling short of what **men** make at various levels.

The spreadsheet covers levels one through six of **Google**'s job hierarchy, from entry-level data center workers at level one to managers and experienced engineers at level six. It does not include company executives and top-level engineers, who receive a wider range of salaries.

At five of the six job levels, women are **paid** less than **men**. At level three, the entry level for technical positions, women make 4 percent less than **men** at $124,000 in salary and bonus. But it widens to 6 percent by the time employees reach midcareer status, around level five, with women earning, on average, $11,000 less than **men**.

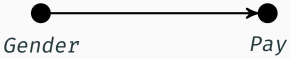
Google said the spreadsheet's information does not take into account a number of factors, like where employees are based, whether they are in higher-**paying** technical positions, and job performance.

Based upon its own analysis from January, **Google** said female employees make 99.7 cents for every dollar a **man** makes, accounting for factors like location, tenure, job role, level and performance.

Google said its analysis includes salary, bonus and equity compensation for 95 percent of employees between levels one and nine — three levels beyond what was reflected in the data shared with The Times — while excluding vice presidents and above. **Google** did not provide a breakdown of how it arrived at that calculation.

Eileen Naughton, **Google**'s vice president of people operations, said the gender **pay** disparity reflected in the internal spreadsheet is "not a representative sample" for other, **more** complex reasons. For example, a person in a nontechnical role may be at the same job level as an engineer, but will be **paid** significantly less because "there is a premium **paid** in all markets for highly technical talent."

Who is right?

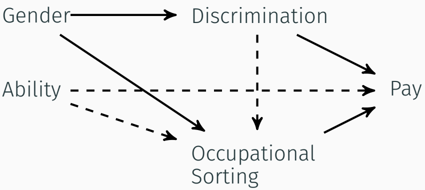


The DoL's investigation is presumably based on more than one spreadsheet, so it's possible that there might be more evidence of discrimination.

Other factors are certainly associated with gender and pay.

Fair for Google to control for some of these in their analysis, but which ones?

Let's start with a DAG...

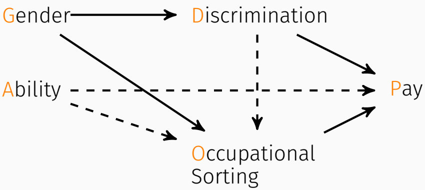


Example taken from Cunningham (2021, p. 73)

Absence of a directed edge in DAG means no causal effect: gender has no direct effect on pay (equal performance).

Ability is an unobserved factor (we don't have it in the salary data).

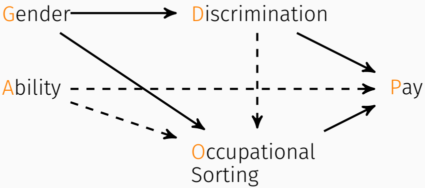
Let's start with a DAG...



At stake is estimating the effect of D on P.

1. $D \rightarrow P$
2. $D \leftarrow G \rightarrow O \rightarrow P$
3. $D \leftarrow G \rightarrow O \leftarrow A \rightarrow P$
4. $D \rightarrow O \rightarrow P$
5. $D \rightarrow O \leftarrow A \rightarrow P$

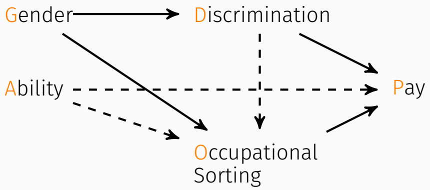
Let's start with a DAG...



Back-door path 2 is open, but 4 is closed (collider O). Paths 3 and 5 are closed, due to a collider (O).

Notice what happens when we control for occupation: paths 3 and 5 actually become *open*!

Google's approach: control for occupation



1. $D \rightarrow P$
2. $D \leftarrow G \rightarrow O \rightarrow P$
3. $D \leftarrow G \rightarrow O \leftarrow A \rightarrow P$
4. $D \rightarrow O \rightarrow P$
5. $D \rightarrow O \leftarrow A \rightarrow P$

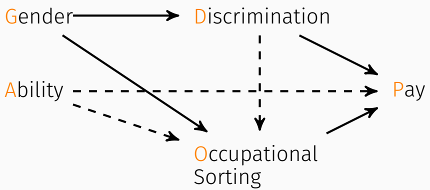
Could we control just for gender?

1. $D \rightarrow P$
2. $D \leftarrow G \rightarrow O \rightarrow P$
3. $D \leftarrow G \rightarrow O \leftarrow A \rightarrow P$
4. $D \rightarrow O \rightarrow P$
5. $D \rightarrow O \leftarrow A \rightarrow P$

Paths 3 and 5 are closed, and controlling for gender would close path 2 as well.

On the face of it, it's fine to leave path 4 as is.

Let's start with a DAG...



But the NATE would be a mix of 2 dynamics, only one of which is under the company's control.

Controlling for occupational sorting, though, opens up a closed back-door path, biasing the estimate!

The aftermath

Google Finds It's Underpaying Many Men as It Addresses Wage Equity

Zitation exportieren

The New York Times

March 4, 2019 Monday 10:12 EST

Section: TECHNOLOGY

Length: 1180 words

Byline: Daisuke Wakabayashi

Highlight: After a recent study, the company gave raises to thousands of men after determining they were earning less than women in similar jobs.

Gender inequality is a radioactive topic at Google. The Labor Department is investigating whether the company systematically underpays women. It has been sued by former employees who claim they were paid less than men with the same qualifications. And last fall, thousands of Google employees protested the way the company handles sexual harassment claims against top executives.

Critics said the results of the pay study could give a false impression. Company officials acknowledged that it did not address whether women were hired at a lower pay grade than men with similar qualifications.

Google seems to be advancing a "flawed and incomplete sense of equality" by making sure men and women receive similar salaries for similar work, said Joelle Emerson, chief executive of Paradigm, a consulting company that advises companies on strategies for increasing diversity. That is not the same as addressing "equity," she said, which would involve examining the structural hurdles that women face as engineers.

Google has denied paying women less, and the company agreed that compensation among similar job titles was not by itself a complete measure of equity. A more difficult issue to solve — one that critics say Google often mismanages for women — is a human resources concept called leveling. Are employees assigned to the appropriate pay grade for their qualifications?

The company said it was now trying to address the issue.

"Because leveling, performance ratings and promotion impact pay, this year we are undertaking a comprehensive review of these processes to make sure the outcomes are fair and equitable for all employees," Lauren Barbato, Google's lead analyst for pay equity, people analytics, wrote in a blog post made public on Monday.

Conclusion

Benefits of DAGs

DAGs are great to help you make your assumptions clear to your audience.

They also let you understand whether an effect can be causally identified or not, if an assumed model about the world is true.

The NATE is identified if all open back-door paths are closed.

Benefits of DAGs

Conditioning on confounders closes open back-door paths.

DAGs help you choose the right conditioning strategy for your analysis.

Valuable throughout, but especially at early stages of research design:

- ✓ to understand what needs to be measured
- ✓ to determine whether effect is causally identified

Thank you for the kind attention!

References

Cunningham, S. (2021). *Causal Inference: The Mixtape*. New Haven, CT: Yale University Press.

Goodin, R., & Dryzek, J. (1980). Rational Participation: The Politics of Relative Power. *British Journal of Political Science*, *10*(3), 273–292.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal Inference in Statistics: A Primer*. Chichester, UK: Wiley.