Lecture 2 - Selection on Observables, DAGs, & Bad Controls

Francis J. DiTraglia

University of Oxford

Treatment Effects: The Basics

A New Twist on the Disease Example[1]

	D	Y	Y_0	Y_1	X
Aiden	0	1	1	1	Young
Bella	0	1	1	1	Young
Caden	0	1	1	1	Young
Dakota	1	1	1	1	Young
Ethel	0	0	0	1	Old
Floyd	0	0	0	0	Old
Gladys	0	0	0	0	Old
Herbert	1	1	0	1	Old
Irma	1	0	0	0	Old
Julius	1	0	0	0	Old

Warmup Exercise: Calculate

1. ATE
2. $\mathbb{E}(Y|D=1) - \mathbb{E}(Y|D=0)$
3. TOT
4. Selection Bias

[1] Different people / potential outcomes from last time: no allergic!

```r
library(tidyverse)

people <- c("Aiden", "Bella", "Carter", "Dakota", "Ethel", "Floyd",
            "Gladys", "Herbert", "Irma", "Julius")

x <- c("young", "young", "young", "young", "old", "old",
       "old", "old", "old", "old")

y0 <- c(1, 1, 1, 1, 0, 0, 0, 0, 0, 0)
y1 <- c(1, 1, 1, 1, 1, 0, 0, 1, 0, 0)
d <- c(0, 0, 0, 1, 0, 0, 0, 1, 1, 1)
y <- (1 - d) * y0 + d * y1

tbl <- tibble(name = people, d, y, y0, y1, x)
rm(y0, y1, d, y, x, people)
```

```r
# ATE
ATE <- tbl |>
  summarize(mean(y1 - y0)) |>
  pull()

ATE
```

```
## [1] 0.2
```

```r
# E(Y|D=1) and E(Y|D=0)
means <- tbl |>
  group_by(d) |>
  summarize(y_mean = mean(y))

means
```

```
## # A tibble: 2 x 2
##         d y_mean
##     <dbl>  <dbl>
## 1       0    0.5
## 2       1    0.5
```

```r
# Naive difference of means
naive <- means |>
  pull(y_mean) |>
  diff()

naive
```

```
## [1] 0
```

```r
# TOT
TOT <- tbl |>
  filter(d == 1) |>
  summarize(mean(y1 - y0)) |>
  pull()

TOT
```

```
## [1] 0.25
```

```r
# Selection Bias
SB <- tbl |>
  group_by(d) |>
  summarize(y0_mean = mean(y0)) |>
  pull(y0_mean) |>
  diff()

SB
```

```
## [1] -0.25
```

Solution

```r
# Everything we've calculated
c(ATE = ATE, naive = naive, TOT = TOT, SB = SB)
```

```
##   ATE naive   TOT    SB
##  0.20  0.00  0.25 -0.25
```

- This revised version of the disease example *still* features selection into treatment.
- As a sanity check, notice that our results satisfy the "Fundamental Decomposition"

$$\underbrace{\mathbb{E}(Y|D=1)-\mathbb{E}(Y|D=0)}_{\text{Observed Difference of Means}}=\underbrace{\mathbb{E}(Y_1-Y_0|D=1)}_{\text{TOT}}+\underbrace{[\mathbb{E}(Y_0|D=1)-\mathbb{E}(Y_0|D=0)]}_{\text{Selection Bias}}$$

Conditional Average Treatment Effects (CATEs)

	D	Y	Y_0	Y_1	X
Aiden	0	1	1	1	Young
Bella	0	1	1	1	Young
Caden	0	1	1	1	Young
Dakota	1	1	1	1	Young
Ethel	0	0	0	1	Old
Floyd	0	0	0	0	Old
Gladys	0	0	0	0	Old
Herbert	1	1	0	1	Old
Irma	1	0	0	0	Old
Julius	1	0	0	0	Old

Intuition

How do treatment effects vary with observed characteristics X?

Definition

$\text{CATE}(x) \equiv \mathbb{E}(Y_1 - Y_0 | X = x)$

Exercise

1. Compute CATE(Young)
2. Compute CATE(Old)
3. Relate these to the *overall* ATE.

Solution: No treatment effect for Young; positive effect for Old.

```r
# Conditional ATEs
tbl |>
  group_by(x) |>
  summarize(CATE = mean(y1 - y0))
```

```
## # A tibble: 2 x 2
##   x      CATE
##   <chr> <dbl>
## 1 old   0.333
## 2 young 0
```

But how can we relate the CATEs to the overall ATE of 0.2?

Recall: Properties of Conditional Expectation $\mathbb{E}(W|X=x)$

Definition

$$\mathbb{E}(W|X=x) \equiv \sum_{\text{all } w} w \cdot \mathbb{P}(W=w|X=x)$$

Linearity

$$\mathbb{E}(cW|X=x) = c\mathbb{E}(W|X=x)$$

$$\mathbb{E}(W+Z|X=x) = \mathbb{E}(W|X=x) + \mathbb{E}(Z|X=x)$$

The Law of Iterated Expectations[2]

In Words

The overall average is the sum of the group averages weighted by relative group size.

In Mathematics

$$\mathbb{E}(W) = \mathbb{E}_X[\mathbb{E}(W|X)] \equiv \sum_{\text{all } x} \mathbb{E}(W|X = x)\mathbb{P}(X = x)$$

Example

$$\mathbb{E}(Y_1 - Y_0) = \mathbb{E}(Y_1 - Y_0|X = \text{Young})\mathbb{P}(\text{Young}) + \mathbb{E}(Y_1 - Y_0|X = \text{Old})\mathbb{P}(\text{Old})$$

[2] See this note for a proof and more discussion.

The Law of Iterated Expectations

```r
group_stats <- tbl |>
  group_by(x) |>
  summarize(CATE_x = mean(y1 - y0), count = n()) |>
  mutate(p_x = count / sum(count))

group_stats
```

```
## # A tibble: 2 x 4
##   x     CATE_x count   p_x
##   <chr>  <dbl> <int> <dbl>
## 1 old    0.333     6   0.6
## 2 young  0         4   0.4
```

The Law of Iterated Expectations

```r
# E[E(Y1 - Y0 | X)]
group_stats |>
  summarize(sum(CATE_x * p_x)) |>
  pull()
```

```
## [1] 0.2
```

```r
# E(Y1 - Y0)
tbl |>
  summarize(mean(y1 - y0)) |>
  pull()
```

```
## [1] 0.2
```

Wait, what is this lecture supposed to be about again?

	D	Y	Y_0	Y_1	X
Aiden	0	1	1	1	Young
Bella	0	1	1	1	Young
Caden	0	1	1	1	Young
Dakota	1	1	1	1	Young
Ethel	0	0	0	1	Old
Floyd	0	0	0	0	Old
Gladys	0	0	0	0	Old
Herbert	1	1	0	1	Old
Irma	1	0	0	0	Old
Julius	1	0	0	0	Old

Disease Example

Selection into treatment: naive comparison of means doesn't give ATE.

Iterated Expectations

If we learn the CATEs, we can average them to get the ATE.

Idea

Maybe if we **adjust for age**, we can address the selection problem.

Selection-on-observables

A pair of assumptions that shows us when this idea will work out.

Propensity Score: Who is more likely to be treated?

	D	Y	Y_0	Y_1	X
Aiden	0	1	1	1	Young
Bella	0	1	1	1	Young
Caden	0	1	1	1	Young
Dakota	1	1	1	1	Young
Ethel	0	0	0	1	Old
Floyd	0	0	0	0	Old
Gladys	0	0	0	0	Old
Herbert	1	1	0	1	Old
Irma	1	0	0	0	Old
Julius	1	0	0	0	Old

Propensity Score $p(x)$

- $p(x) \equiv \mathbb{P}(D = 1|X = x)$
- Share treated by age group.

Exercise

Calculate $p(\text{Young})$ and $p(\text{Old})$

Propensity Score: Who is more likely to be treated?

	D	Y	Y_0	Y_1	X
Aiden	0	1	1	1	Young
Bella	0	1	1	1	Young
Caden	0	1	1	1	Young
Dakota	1	1	1	1	Young
Ethel	0	0	0	1	Old
Floyd	0	0	0	0	Old
Gladys	0	0	0	0	Old
Herbert	1	1	0	1	Old
Irma	1	0	0	0	Old
Julius	1	0	0	0	Old

Propensity Score $p(x)$

- $p(x) \equiv \mathbb{P}(D = 1 | X = x)$
- Share treated by age group.

Exercise

Calculate $p(\text{Young})$ and $p(\text{Old})$

Solution

$p(\text{Young}) = 1/4, \quad p(\text{Old}) = 1/2$

Old people are more likely to take treatment and more likely to die with or without it!
Age *confounds* the relationship between D and Y.

Wishful Thinking

Wouldn't it be great if $\text{CATE}(x) = \mathbb{E}(Y|D=1, X=x) - \mathbb{E}(Y|D=0, X=x)$?

	D	Y	Y_0	Y_1	X
Aiden	0	1	1	1	Young
Bella	0	1	1	1	Young
Caden	0	1	1	1	Young
Dakota	1	1	1	1	Young
Ethel	0	0	0	1	Old
Floyd	0	0	0	0	Old
Gladys	0	0	0	0	Old
Herbert	1	1	0	1	Old
Irma	1	0	0	0	Old
Julius	1	0	0	0	Old

Stratify by Age

- Perhaps *within* age groups there is no selection problem.
- If so, learn the CATE for each group.

Exercise

Check if this claim holds in our example.

Stratifying by age works in this example

$$\text{CATE}(x) = \mathbb{E}(Y|D = 1, X = x) - \mathbb{E}(Y|D = 0, X = x)$$

```r
tbl |>
  group_by(x) |>
  summarize(CATE = mean(y1-y0)) |>
  knitr::kable(digits = 2)
```

x	CATE
old	0.33
young	0.00

```r
tbl |>
  group_by(x, d) |>
  summarize(y_mean = mean(y)) |>
  knitr::kable(digits = 2)
```

x	d	y_mean
old	0	0.00
old	1	0.33
young	0	1.00
young	1	1.00

Final Step

$$\text{ATE} = \text{CATE}(\text{Young})\mathbb{P}(\text{Young}) + \text{CATE}(\text{Old})\mathbb{P}(\text{Old}) = 2/5 \times 0 + 3/5 \times 1/3 = 0.2$$

This worked because our example satisfies two key assumptions.

Definition: Conditional Independence

- $W \perp\!\!\!\perp Z | R \iff \mathbb{P}(W, Z|R) = \mathbb{P}(W|R) \cdot \mathbb{P}(Z|R)$.
- See chapter 2 of the lecture notes and this video for more details.

Assumption 1 – Selection on Observables:[3] $D \perp\!\!\!\perp (Y_0, Y_1)|\boldsymbol{X}$

- Implies that people with the same observed characteristics have the same potential outcomes, on average, regardless of whether they were *actually* treated or not.
- See my blog post for more discussion of this assumption.

Assumption 2 – Overlap: $0 < p(\boldsymbol{x}) < 1$ for all values of $\boldsymbol{x}$.

- Recall that $p(\boldsymbol{x}) \equiv \mathbb{P}(D = 1|\boldsymbol{X} = \boldsymbol{x})$.
- Among people with given characteristics $\boldsymbol{x}$, some but not all are treated.

[3] This can be weakened to $\mathbb{E}(Y_d|D, \boldsymbol{X}) = \mathbb{E}(Y_d|\boldsymbol{X})$ for $d = 0, 1$, i.e. *mean* independence.

The approach we used above is called "Regression Adjustment"

Intuition

- Form **strata** based on common value $\boldsymbol{x}$ of covariates.
- Within each stratum, compute the average outcome among treated and untreated.
- Subtract these to estimate CATE($\boldsymbol{x}$), the stratum-specific ATE.
- Average the stratum-specific ATEs, weighting by the fraction of people in each.

Main Result[4]

Under the selection on observables and overlap assumptions:

$$\text{CATE}(\boldsymbol{x}) \equiv \mathbb{E}(Y_1 - Y_0 | \boldsymbol{X} = \boldsymbol{x}) = \mathbb{E}(Y|D=1, \boldsymbol{X} = \boldsymbol{x}) - \mathbb{E}(Y|D=0, \boldsymbol{X} = \boldsymbol{x}).$$

By iterated expectations, $\text{ATE} = \mathbb{E}[\text{CATE}(\boldsymbol{X})]$ so we can learn the ATE.

[4] See my video for the proof: https://expl.ai/BJWTFKG

Alternative Approach: Propensity Score Weighting

Intuition

- Disease example: older people are more likely to be treated and more likely die regardless of whether they are treated.
- *Too few* young people among the treated and *too few* old people among the untreated relative to what we'd have in a randomized experiment.
- To compensate: **upweight** treated young people untreated old people when computing average outcomes for the treated and untreated groups.

Main Result[5]

Under the selection on observables and overlap assumptions:

$$\text{ATE} = \mathbb{E}\left[w_1(\boldsymbol{X}) \cdot Y\right] - \mathbb{E}\left[w_0(\boldsymbol{X}) \cdot Y\right], \quad w_1(\boldsymbol{X}) = \frac{D}{p(\boldsymbol{X})}, \quad w_0(\boldsymbol{X}) = \frac{1-D}{1-p(\boldsymbol{X})}$$

[5]See my video for the proof: https://expl.ai/BASRRGX.

Propensity Score Weighting in Our Example

```r
psw <- tbl |>
  group_by(x) |>
  mutate(pscore = mean(d)) |>
  ungroup() |>
  mutate(weight1 = d / pscore,
         weight0 = (1 - d) / (1 - pscore))
```

Propensity Score Weighting in Our Example

```
psw |> select(-y0, -y1)
```

```
## # A tibble: 10 x 7
##    name        d     y x      pscore weight1 weight0
##    <chr>   <dbl> <dbl> <chr>   <dbl>   <dbl>   <dbl>
##  1 Aiden       0     1 young    0.25       0    1.33
##  2 Bella       0     1 young    0.25       0    1.33
##  3 Carter      0     1 young    0.25       0    1.33
##  4 Dakota      1     1 young    0.25       4    0
##  5 Ethel       0     0 old      0.5        0    2
##  6 Floyd       0     0 old      0.5        0    2
##  7 Gladys      0     0 old      0.5        0    2
##  8 Herbert     1     1 old      0.5        2    0
##  9 Irma        1     0 old      0.5        2    0
## 10 Julius      1     0 old      0.5        2    0
```

Propensity Score Weighting in Our Example

```r
psw |> summarize(sum(weight1), sum(weight0))
```

```
## # A tibble: 1 x 2
##   `sum(weight1)` `sum(weight0)`
##            <dbl>          <dbl>
## 1             10             10
```

```r
psw |>
  summarize(mean(weight1 * y) - mean(weight0 * y)) |>
  pull()
```

```
## [1] 0.2
```

```r
ATE
```

```
## [1] 0.2
```

How can we evaluate the assumptions?

Overlap

- Since D and $\boldsymbol{X}$ are observed, we can check this directly.
- The more characteristics we put into $\boldsymbol{X}$, the harder it becomes to satisfy overlap.

Selection on Observables

- Without outside data or extra assumptions, there's no way to check this.
- Else equal, the more characteristics we put into $\boldsymbol{X}$, the more plausible this becomes.

Bad Controls

- More is **not always better**. Some characteristics definitely **shouldn't** go into $\boldsymbol{X}$.
- This is what we'll discuss for the rest of the lecture!

The Birthweight Paradox[6]

The analyses in Yerushalmy's paper indicated that, among low birthweight infants of less than 2500g, maternal smoking was associated with lower infant morality. The results have been replicated in a number of studies and populations, and these seemingly paradoxical associations are now often referred to as the 'birthweight paradox'

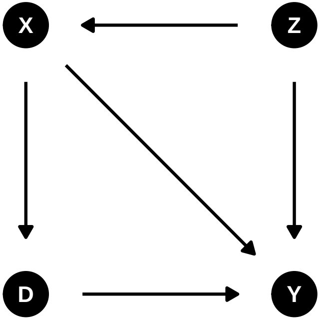
- $D = 1$ mother smokes while pregnant
- $Y = 1$ infant dies
- $X = 1$ low birthweight

Should we adjust for birthweight when studying the causal effect of maternal smoking on infant mortality?

[6] Quote from VanderWeele (2014).

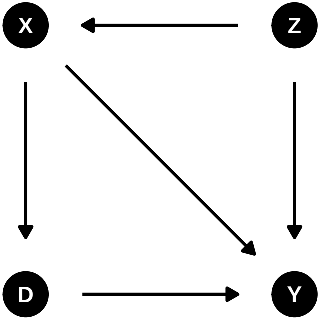
Graph: set of **nodes** connected by **edges**.

- Two nodes are **adjacent** if connected by an edge.
- Edges can be **directed** (figure) or **undirected**.
- Directed edge points from **parent** to **child**.
- **Directed graph** has only directed edges.
- **Path**: sequence of connected vertices.
- **Directed Path**: a path that "obeys one-way signs"
- Directed path points from **ancestor** to **descendant**.
- **Cycle**: directed path that returns to starting node.
- **Acyclic Graph**: a graph without any cycles.



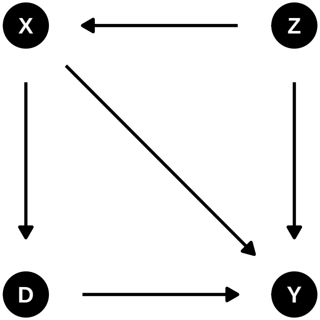
Exercise

1. Is this graph directed?
2. Is this graph acyclic?
3. Are Z and D adjacent?
4. List all paths between D and Y.
5. List all *directed* paths from D to Y.



Exercise

1. Is this graph directed?
2. Is this graph acyclic?
3. Are Z and D adjacent?
4. List all paths between D and Y.
5. List all *directed* paths from D to Y.



Solution

1. Yes: all edges in the graph are directed.
2. Yes: there is no directed path that takes you back to the node where you started.
3. Z and D are not adjacent: there is no edge between them.
4. There are three: $(D \to Y)$, $(D \leftarrow X \to Y)$, and $(D \leftarrow X \leftarrow Z \to Y)$.
5. There is only one: $(D \to Y)$.

Graphical Causal Models with DAGs

Graphical Causal Model

Directed edges encode assumptions about the “flow” of causation (edge) or lack thereof (no edge).

Potential Cause

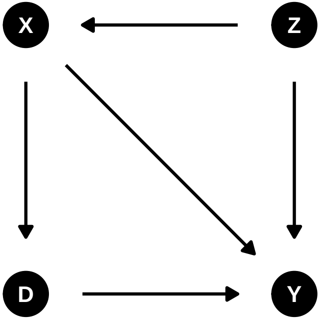
If D is an ancestor of Y, it is a **potential cause** of Y.

Direct Cause

If D is a parent of Y, it is a **direct cause** of Y.

Back Door Criterion

Can we learn $(D \rightarrow Y)$ using selection on observables? If so, what covariates should we adjust for?



“Draw Your Assumptions” – Birthweight Example

Birthweight Paradox

- Y mortality
- X birthweight
- D maternal smoking
- U unobserved: e.g. malnutrition / birth defect

Should we condition on X?

Can’t adjust for U: unobserved. Should we adjust for birthweight when studying (smoking $\rightarrow$ mortality) effect?

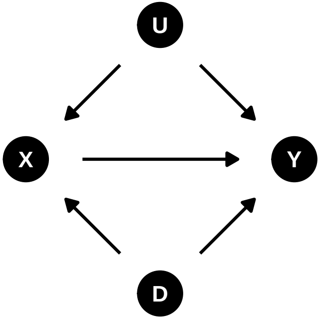


Figure 1: A possible model for the birthweight example.

Causal and Non-causal Paths

Causal Path

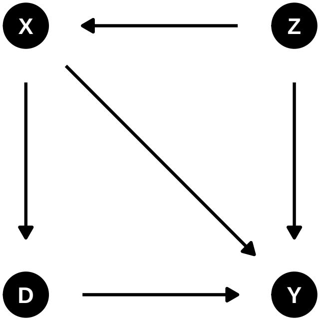
Directed path between treatment and outcome; always starts with an edge pointing *out* of treatment.

Backdoor Path

Noncausal path path between treatment and outcome; always starts with an edge pointing *into* treatment.

Exercise

1. List all causal paths from D to Y.
2. List all backdoor paths between D and Y.



Causal and Non-causal Paths

Causal Path

Directed path between treatment and outcome; always starts with an edge pointing *out* of treatment.

Backdoor Path

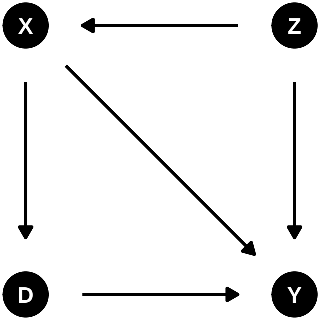
Noncausal path path between treatment and outcome; always starts with an edge pointing *into* treatment.

Exercise

1. List all causal paths from D to Y.
2. List all backdoor paths between D and Y.

Solution

1. $(D \rightarrow Y)$
2. $(D \leftarrow X \rightarrow Y)$, and $(D \leftarrow X \leftarrow Z \rightarrow Y)$.



Graph Surgery

Observational Distribution: $\mathbb{P}(Y|D=d)$

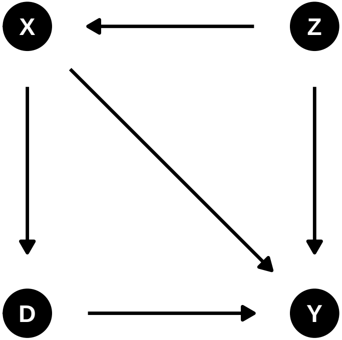
- *Actual* distribution of Y among people observed to have $D=d$.
- DAG shows the observational distribution and how it arises from our causal model.

Interventional Distribution: $\mathbb{P}(Y|\text{do}(D=d))$

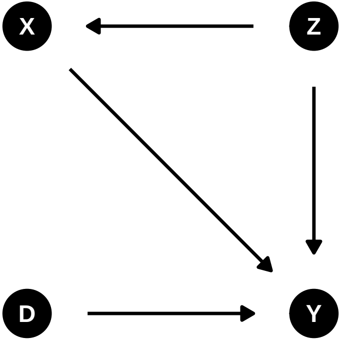
- Distribution of Y that we *would obtain* if we *intervened* and set $D=d$ for everyone.
- Obtain from DAG by removing edges pointing into D.
- Causal effect of interest is the path from D to Y in this "modified" graph.
- $\text{ATE} = \mathbb{E}(Y_1 - Y_0) = \mathbb{E}(Y|\text{do}(D=1)) - \mathbb{E}(Y|\text{do}(D=0))$
- This is what an experiment does: removes all causes of treatment!

Graph Surgery: Delete Edges Pointing Into D

Observational Distribution



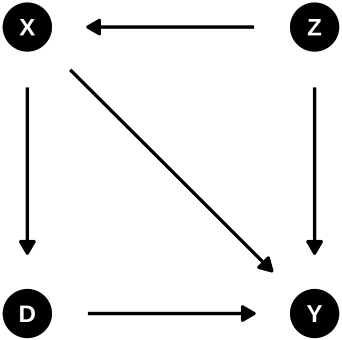
Interventional Distribution: do(D)



Interventional DAG has *no backdoor paths*. To use the observational distribution for causal inference, we will attempt to "block" the backdoor paths by conditioning.

Exercise: Draw the DAG for the do(X) Interventional Distribution

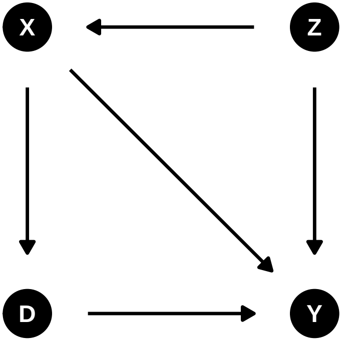
Observational Distribution



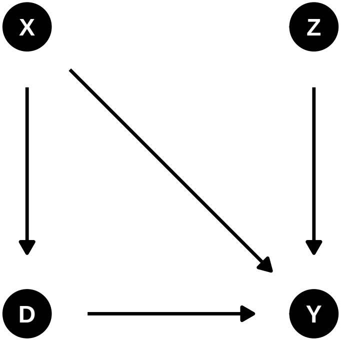
Interventional Distribution: do(X)

Exercise: Draw the DAG for the do(X) Interventional Distribution

Observational Distribution



Interventional Distribution: do(X)



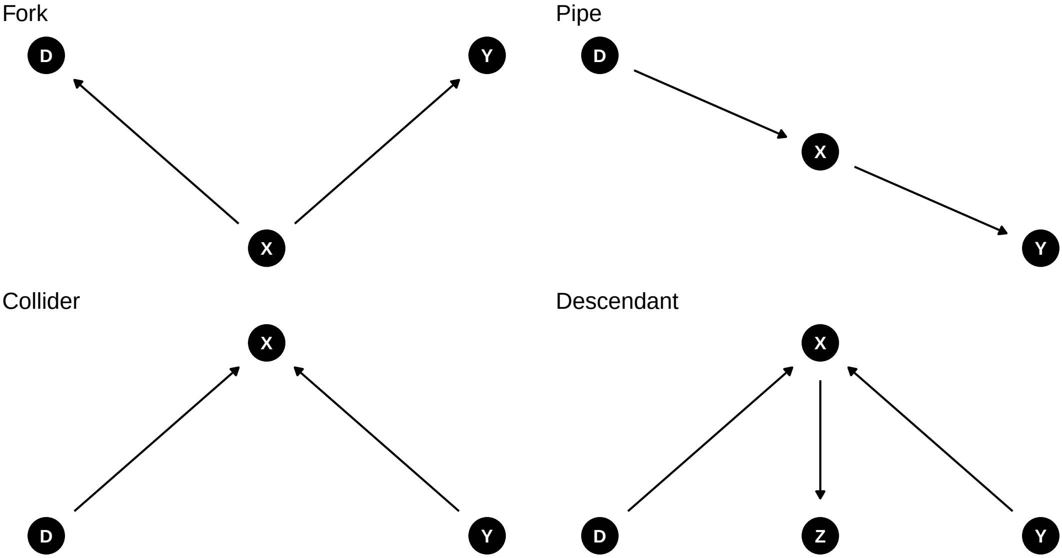


Figure 2: The Four Basic DAGs

Fork = Common Cause / Confounder

Confounder = Good Control

- D and Y are dependent: **open** path between them.
- But D doesn't cause Y: X causes D and Y.
- Conditioning on X **blocks the path** from D to Y.

Example

D is shoe size, Y is reading ability, X is age.

Fork Rule

If X is a common cause of D and Y and there is only one path between D and Y, then $D \perp\!\!\!\perp Y | X$.

“Condition on things that cause both D and Y.”

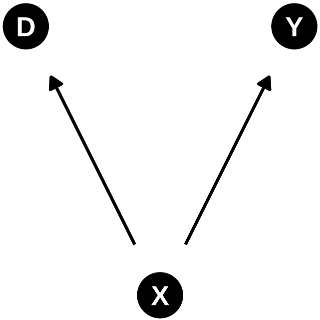


Figure 3: X is a confounder. Good control for $D \to Y$.

Pipe = Mediator

Mediator = Bad Control

- D and Y are dependent: **open** path between them.
- D causes Y through its causal effect on X.
- Conditioning on X **blocks the path** from D to Y.

Example

D is SAT coaching, X is SAT score, Y is college acceptance

Pipe Rule

If there is only one directed path from D to Y and X intercepts that path, then $D \perp\!\!\!\perp Y | X$.

"Don't condition on an intermediate outcome."

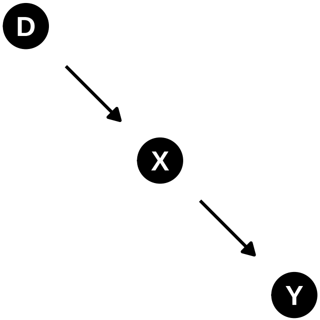


Figure 4: X is a mediator. Bad control for $D \to Y$.

Collider = Common Effect

Common Effect = Bad Control

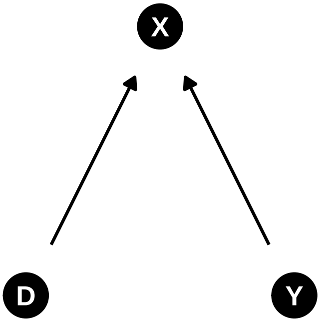
- D and Y are independent: **blocked** path between them.
- D and Y both cause X, but neither causes the other.
- Conditioning on X **unblocks** the path between D and Y.

Example

D, Y indep. coins; X = bell rings if at least one HEADS.

Collider Rule

If there is only one path between D and Y and X is their common effect, then $D \perp\!\!\!\perp Y$ but $D \not\perp\!\!\!\perp Y | X$.



Why are brilliant researchers lousy teachers?

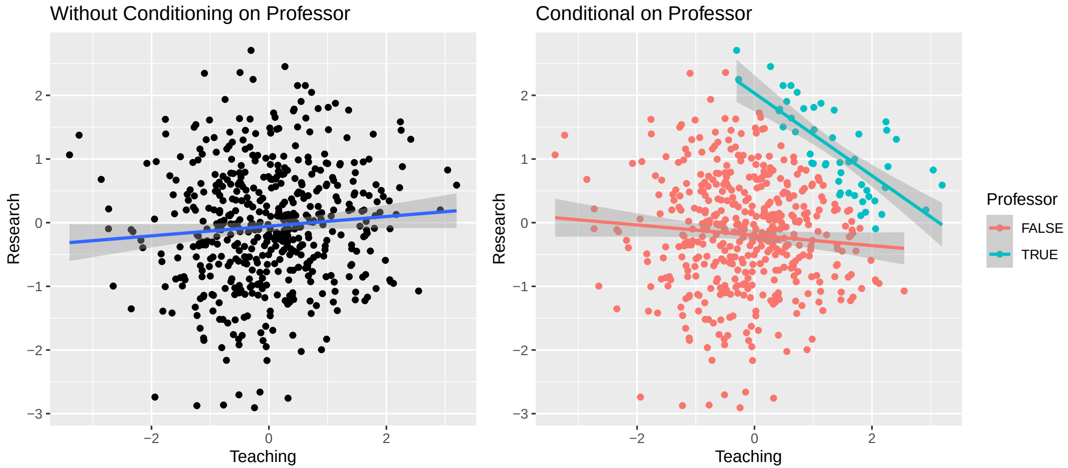


Figure 5: Teaching and Research are independent $N(0, 1)$. Professor is a collider: TRUE if the sum of Research and Teaching is in the top 10th percentile of all observations.

The Descendant

Descendant Rule

Conditioning on a descendant Z of X has the effect of *partially conditioning* on X itself.

Collider Corollary

In the figure, $D \perp\!\!\!\perp Y$ but $D \not\!\perp\!\!\!\perp Y | W$.

Discussion

- What this means depends on the situation.
- In the figure X is a collider.
- Could also have X as the middle node in pipe/fork.
- Pipe/fork: adjust for $W \Rightarrow$ **partially block** D, Y path.

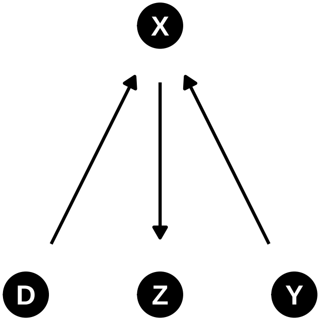


Figure 6: Z is a descendant of the collider X. Bad control for $D \to Y$

Exercise: Find all examples of the four basic DAGS.

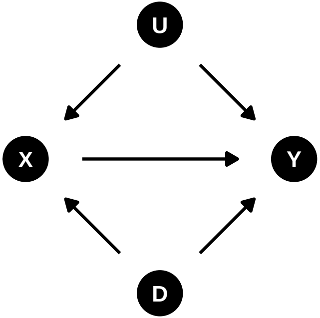


Figure 7: Birthweight DAG

Exercise: Find all examples of the four basic DAGS.

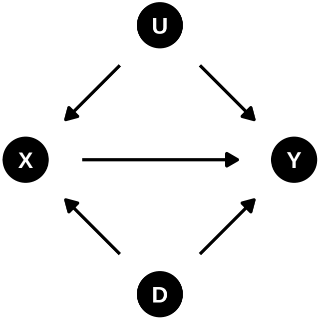


Figure 7: Birthweight DAG

Solution

1. **Forks**: $X \leftarrow U \rightarrow Y$ and $X \leftarrow D \rightarrow Y$
2. **Pipe**: $D \rightarrow X \rightarrow Y$
3. **Colliders**: $D \rightarrow X \leftarrow U$ and $D \rightarrow Y \leftarrow U$.
4. **Descendant**: Y is a descendant of the collider $D \rightarrow X \leftarrow U$.

Blocking and Opening Paths in the Four Basic DAGs

Fork

$D \leftarrow X \rightarrow Y$ is an **open** path; conditioning on the **confounder** X **blocks** the path.

Pipe

$D \rightarrow X \rightarrow Y$ is an **open** path; conditioning on the **mediator** X **blocks** the path.

Collider

$D \rightarrow X \leftarrow Y$ is a **blocked** path; conditioning on the **collider** X **opens** the path.

Descendant

Conditioning on the descendant of a **confounder** / **mediator** partially blocks the open path. Conditioning on the descendant of a **collider** partially opens the blocked path.

Backdoor Criterion

Use what we know about the four basic DAGs to **block** all backdoor paths between D and Y in our "big" DAG. Obtain interventional distribution from observational data.

The Backdoor Criterion

Recall: Backdoor Path

Noncausal path between D and Y; starts with edge pointing **into** D.

Blocked Path

A set of nodes X **blocks** a path p if and only if p contains: (1) a **pipe** or **fork** whose middle node is in X or (2) a **collider** that is *not* in X and has no descendants in X.

Backdoor Criterion

A set of nodes X satisfies the back-door criterion relative to (D, Y) if no node in X is a descendant of D and X blocks every back-door path between D and Y.

A Less Formal Statement of the Back-door Criterion

1. List all the paths that connect treatment and outcome.
2. Check which of them *open*. A path is *open* unless it contains a collider.
3. Check which of them are *back-door paths*: contain an arrow pointing at D.
4. If there are no open back-door paths, you're done. If not, look for nodes you can condition on to **block** remaining open back-door paths without opening new ones.

Of course we can only condition on *observed* variables!

Backdoor Theorem = Selection on observables!

Backdoor Theorem

If X satisfies the back-door criterion relative to (D, Y), then

$$\mathbb{P}(Y = y | \mathrm{do}(D = d)) = \sum_z \mathbb{P}(Y = y | D = d, X = x) \cdot \mathbb{P}(X = x)$$

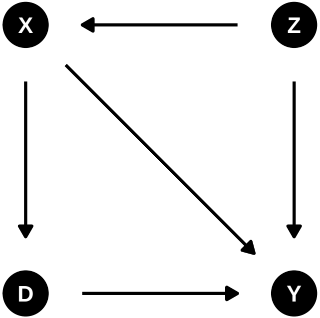
Counterfactual Interpretation

If X satisfies the back-door criterion relative to (D, Y), then $Y_d \perp\!\!\!\perp D | X$ for all d.

Translating to Potential Outcomes

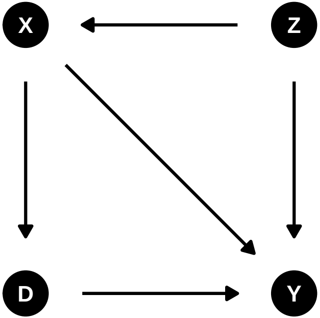
- The "counterfactuals" Y_D are our potential outcomes from earlier in this lecture.
- Back-door criterion implies selection on observables assumption for D given X.
- The formula above is nothing more than **regression adjustment**.

Exercise: What to adjust for to learn the effect of each intervention?



1. The effect of D on Y.
2. The effect of X on Y.
3. The effect of Z on Y?

Exercise: What to adjust for to learn the effect of each intervention?



1. The effect of D on Y.
2. The effect of X on Y.
3. The effect of Z on Y?

Solution

1. There are two backdoor paths. In $(D \leftarrow X \rightarrow Y)$, the middle node in a fork is X. In $(D \leftarrow X \leftarrow Z \rightarrow Y)$ the middle node in a pipe is X. Adjusting for X blocks both.
2. The only backdoor path is $(X \leftarrow Z \rightarrow Y)$, a fork with Z as its middle node. Adjusting for Z blocks this path.
3. There are no arrows pointing into Z, hence no backdoor paths. We don't have to adjust for anything.

(Possible) Solution to Birthweight Paradox

Among low birthweight infants... maternal smoking was associated with lower infant mortality.

Notation

Y mortality, X birthweight, D maternal smoking, and U unobserved: e.g. malnutrition / birth defect

Birthweight is a bad control!

- Can't adjust for U because it's unobserved.
- No arrows pointing into D so no backdoor paths.
- X is a collider: conditioning on it creates spurious dependence between D and U.

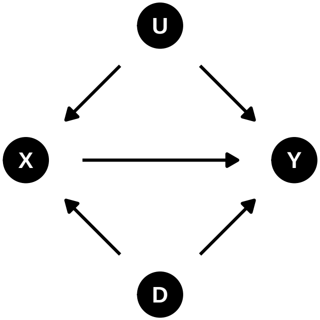


Figure 8: If we believe this model, X is a bad control.

Low birthweight infants whose mothers did *not* smoke must have an unfavorable value of U, making it appear as though smoking has health benefits.