Econ 4v98/5v98

Causal Inference and Research Design

Scott Cunningham

Baylor University

Spring 2014

Introduction and stats review

- Review syllabus and expectations
- Introductory material and statistics review

Three Types of Empirical Research

- Descriptive analysis

  Establish facts about economic reality that need to be explained by theoretical reasoning and yield new insights about economic phenomena

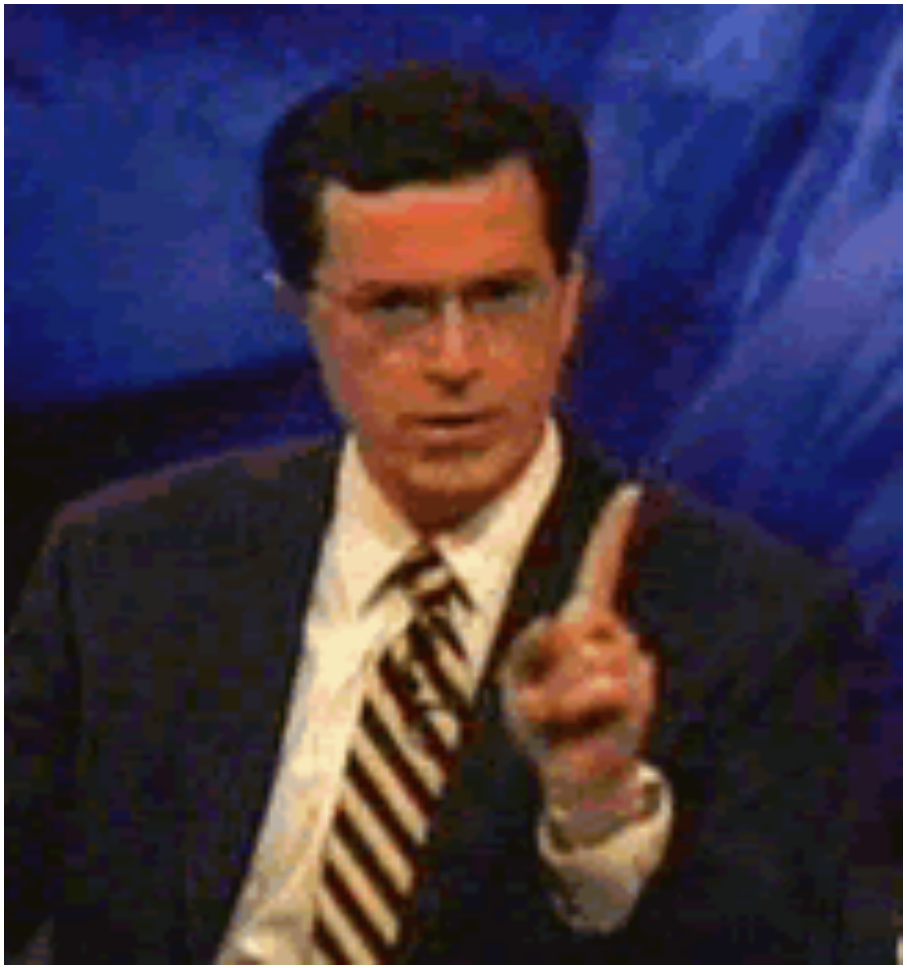
- Non-causal prediction

  Using known quantitative information to predict either future events or other relevant associations

- Causal prediction (or "causal inference")

  Seeks to determine the effects of particular interventions and policies, or to estimate the behavioral relationships suggested by economic theory

What do we mean by "Scientific Methodology"?



- Scientific methodology is the epistemological foundation of our scientific knowledge
- Science does not collect evidence in order to "prove" what people already believe or want others to believe.
- Science accepts unexpected and even undesirable answers.
- Science is process oriented, not outcome oriented.

Causality statements over time

- Potential outcomes and counterfactuals
  - John Stuart Mill: "If a person eats of a particular dish, and dies in consequence, that is, would not have died if he had not eaten it, people would be apt to say that eating of that dish was the source of his death."
  - Roland Fisher: "If we say, 'This boy has grown tall because he has been well fed,' ... we are suggesting that he might quite probably have been worse fed, and that in this case he would have been shorter."
- Experimentation as a source of knowledge
  - Stock and Watson: "A causal effect is defined to be the effect of a given action or treatment, as measured in an ideal, randomized controlled experiment. In such an experiment, the only systematic reason for differences in outcomes between the treatment and control groups is the treatment itself."
- Physical randomization
  - William Sealy Gossett: "If now the plots had been randomly placed..."

Why randomization?

- How to recognize causal questions
  - "What randomized experiment would I have run if I were a dictator with infinite resources?"
  - But what if randomization may be unethical, expensive, or simply infeasible?
- Observational ("non-experimental") studies vs. Randomized experiment
  - In a controlled experiment, study participants are randomly selected to receive some intervention ("treatment") or some "placebo", usually by the researcher designing the experiment
  - In an observational study, study participants self-select themselves to receive the treatment
  - So what? Both have a treatment group and a control so why is this distinction important?
- Example: What is the causal effect of smoking on lung cancer?
  - What randomized experiment would you run if you were a dictator with infinite resources, and discuss any practical limitations?
  - Describe the differences between those in the treatment group and control group prior to the experiment beginning under randomization?
- Describe an alternative observational study. What do we gain and what do we lose when we forego randomization? How important is that?

What is a "counterfactual"?

- Causality is based on comparisons between reality and hypothetical "what-if" scenarios, or "counterfactual"
- Counterfactuals in popular culture
  - It's a Wonderful Life: George Bailey
  - A Christmas Carol by Charles Dickens and Ebinezer Scrooge
- Example: Amy is a lifelong smoker recently diagnosed with lung cancer. Did smoking cause her lung cancer? It did *if and only if* Amy would not have gotten lung cancer had she not smoked.
- But how can we know if something causes something else if causality requires comparing two states of the world, but we only have one?

Well-defined causal states

- **Causal claim**: Uggen and Manza (2002) claim that Al Gore would have won the 2000 presidential election had felons and ex-felons been allowed to vote.
  - Restrictions on votes $\rightarrow$ who votes $\rightarrow$ who wins election
- **But...** if in this hypothetical world, (ex-)felons could vote, might anything else relevant to outcomes changed?
  - Would Gore and Bush have run on different policies?
  - Would that different campaign have affected other votes?
  - Would someone other than Gore and/or Bush have run?
- Causal question are implicitly **ceteris paribus** statements – "holding everything except for the intervention and the potential outcomes constant"

> *"When a ceteris paribus assumption is relied on to rule out other contrasts that are nearly certain to occur at the same time, the posited causal states are open to the charge that they are too metaphysical to justify the pursuit of causal analysis." - Morgan and Winship (p. 33)*

- We call causal states "well-defined" if the reliance on the *ceteris paribus* assumption is plausible – but never forget that this assumption is critical and one you have to seriously contemplate yourself

Brief review of summation and expectation

Summation Operator

$$\sum_{i=1}^{n} x_i \equiv x_1 + x_2 + \ldots + x_n \tag{1}$$

Properties of the Summation operator

1. For any constant c:

$$\sum_{i=1}^{n} c = nc \tag{2}$$

2.

$$\sum_{i=1}^{n} cx_i = c\sum_{i=1}^{n} x_i \tag{3}$$

3. For any constant a and b:

$$\sum_{i=1}^{n}(ax_i + by_i) = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} y_i \tag{4}$$

Sum of the quotients does **not** equal the quotient of the sums:

$$\sum_{i=1}^{n} \frac{x_i}{y_i} \neq \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} y_i} \tag{5}$$

Sums of squared does **not** equal the sum squared:

$$\sum_{i=1}^{n} x_i^2 \neq \left(\sum_{i=1}^{n} x_i \right)^2 \tag{6}$$

The **average** (mean) can be computed as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n} \tag{7}$$

and the sum of the deviations from the mean:

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0 \tag{8}$$

One result that will be very useful throughout the semester is:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2 \tag{9}$$

and more generally,

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i = \sum_{i=1}^{n} x_i y_i - n(\bar{x}\bar{y}) \tag{10}$$

Expected Value

The **expected value** of a random variable, also called the expectation and sometimes the *population mean* is simply the weighted average of the possible values that variable can take, with the weights being given by the probability of each value "happening" in the population.

- Suppose the variable X can take values $x_1, x_2, \ldots, x_k$ with probability $f(x_1), f(x_2), \ldots, f(x_k)$, respectively.

$$E(X) = x_1 f(x_1) + x_2 f(x_2) + \ldots + x_k f(x_k) = \sum_{j=1}^{k} x_j f(x_j) \quad (11)$$

For Example: if X takes on values -1, 0, and 2 with probabilities 0.3, 0.3, and 0.4, respectively,

$$E(X) = (-1)(0.3) + (0)(0.3) + (2)(0.4) = 0.5 \tag{12}$$

In fact you could take the expectation of a function of that variable, say X^2. Note that X^2 take only the values 0, 1 and 4 with probabilities 0.3, 0.3, and 0.4,

$$E(X) = (-1)^2(0.3)+(0)^2(0.3)+(2)^2(0.4) = 0.3+0+1.6 = 1.9 \tag{13}$$

Properties of Expected Values

1. For any constant c, $E(c) = c$
2. For any constants a and b,
   $E(aX + b) = E(aX) + E(b) = aE(X) + b$
3. If we have many constants, $a_1, a_2, \ldots, a_n$, and many random variables, $X_1, X_2, \ldots, X_n$, then

$$E(a_1X_1 + a_2X_2 + \ldots + a_nX_n) = a_1E(X_1) + a_2E(X_2) + \ldots + a_nE(X_n) \tag{14}$$

Using summation notation,

$$E\left(\sum_{i=1}^{n} a_iX_i\right) = \sum_{i=1}^{n} a_iE(X_i) \tag{15}$$

In the special case in which each $a_i = 1$:

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) \tag{16}$$

Rules of the Expectation Operator, E(.)

Expectation operators, $E[x]$, calculate the population mean

Let a and b be a constant number (e.g., 2, 0.94) and $\mathbf{X}$, $\mathbf{Y}$ random variables that take on different values (e.g., a column of student grades for the semester)

Expectation operator follow algebraic rules:

$$
\begin{aligned}
E[a] &= a && (17)\\
E[a + b\mathbf{X}] &= a + bE[\mathbf{X}] && (18)\\
E[\mathbf{X} + \mathbf{Y}] &= E[\mathbf{X}] + E[\mathbf{Y}] && (19)\\
E[b\mathbf{X}]^2 &= b^2 E[\mathbf{X}]^2 && (20)\\
E[\mathbf{X} + \mathbf{Y}]^2 &= E[(\mathbf{X} + \mathbf{Y})(\mathbf{X} + \mathbf{Y})]\\
&= E[(\mathbf{X}^2 + 2\mathbf{XY} + \mathbf{Y}^2)]\\
&= E[(\mathbf{X}^2] + 2E[\mathbf{XY}] + E[\mathbf{Y}^2)] && (21)
\end{aligned}
$$

Potential Outcomes Notation

Group	Cancer (Y^1) under smoking	Cancer (Y^0) under no smoking
Treatment ($D = 1$)	Observable as Y	Unobservable Counterfactual
Control ($D = 0$)	Unobservable Counterfactual	Observable as Y

- **Potential outcomes (Y^1, Y^1)**: Y^1 and Y^0 are the "potential" cancer outcomes under a lifetime of smoking (Y^1) or not (Y^0)
- **Intervention (D)**: Smoking (D) is the intervention and $D \in \{0, 1\}$ corresponding to "control" ($D = 0$) or "treatment" ($D = 1$)

$$Y = Y^1 \text{ if } D = 1 \tag{22}$$
$$Y = Y^0 \text{ if } D = 0 \tag{23}$$

- **Observed outcomes (Y)**: Combining (1) and (2) with the following switching equation:

$$Y = DY^1 + (1 - D)Y^0 \tag{24}$$

- **Counterfactuals**: Y^1 for the $D = 0$ group and Y^0 for the $D = 1$ groups are unobserved but exist in principle – Amy *could've* not smoked
- We need to review some statistics before we can continue with definitions...

Important Definitions

- **Definition 1: Causal effect**. The causal effect of an intervention for every individual, i, is:

$$\delta_i = Y_i^1 - Y_i^0 \tag{25}$$

- **Definition 2: Fundamental problem of causal inference**: It is impossible to observe both Y_i^1 and Y_i^0 for the same individual. Therefore, individual causal effects are unknowable.
- **But . . .**, it is possible using *groups* of units if *assumptions* about treatment selection are credible

Average Treatment Effect Definitions

- **Definition 3: Average Treatment Effect (ATE)**. Aggregated causal effects over all individuals.

$$\begin{aligned} E[\delta] &= E[Y^1 - Y^0] \\ &= E[Y^1] - E[Y^0] \end{aligned} \tag{26}$$

- **Definition 4: Average Treatment Effect on the Treated (ATT)**. Average treatment effect for those who typically take the treatment

$$\begin{aligned} E[\delta|D=1] &= E[Y^1 - Y^0|D=1] \\ &= E[Y^1|D=1] - E[Y^0|D=1] \end{aligned} \tag{27}$$

- **Definition 5: Average Treatment Effect on the Untreated (ATU)**. Average treatment effect for those who typically don't take the treatment

$$\begin{aligned} E[\delta|D=0] &= E[Y^1 - Y^0|D=0] \\ &= E[Y^1|D=0] - E[Y^0|D=0] \end{aligned} \tag{28}$$

- **Naive ATE (NATE)**: Difference in mean outcome for the self-selected treatment from self-selected control groups. (Difference along the *observable* diagonal)

$$E_{NAIVE}[\delta] = E_N[y_i|d_i=1] - E_N[y_i|d_i=0] \tag{29}$$

Biased Naive ATE

- Usually, whenever people calculate NATE, they think they are calculating ATE. But compare ATE with NATE more closely:

$$\begin{aligned} ATE &= E[Y^1] - E[Y^0] \\ NATE &= E_N[y_i|d_i = 1] - E_N[y_i|d_i = 0] \end{aligned}$$

- All individuals in the population contribute to calculating ATE, whereas each sampled individual is used once to estimate $E_N[y_i|d_i = 1]$ or $E_N[y_i|d_i = 0]$.

Average Treatment Effects

Group	Units smoke Y^1	Units don't smoke Y^0	**Estimated Treatment Effects**
Treatment ($D=1$)	**(Observed)** $E[Y^1\|D=1]$	(Unobserved counterfactual) $E[Y^0\|D=1]$	$ATT = E[Y^1\|D=1] - E[Y^0\|D=1]$
Control ($D=0$)	(Unobserved counterfactual) $E[Y^1\|D=0]$	**(Observed)** $E[Y^0\|D=0]$	$ATU = E[Y^1\|D=0] - E[Y^0\|D=0]$
Averages	$E[Y^1]$	$E[Y^0]$	$ATE = E[\delta] = E[Y^1] - E[Y^0]$

- Keep in mind: potential outcomes are not the same as treatment assignment.
- We can calculate average treatment effects by averaging the columns and subtracting by rows
- But if a fraction of the population, π, receives treatment (e.g., 60%), then weight averages by π
  - $E[Y^1] = \pi E[Y^1|D=1] + (1-\pi)E[Y^1|D=0]$
  - $E[Y^0] = \pi E[Y^0|D=1] + (1-\pi)E[Y^0|D=0]$
- Conditional average treatment effects are calculated by differencing Y^1 and Y^0 by row
  - $ATT = E[Y^1|D=1] - E[Y^0|D=1]$, **Average treatment effect on treated** (e.g., typical smokers)
  - $ATU = E[Y^1|D=0] - E[Y^0|D=0]$, **Average treatment effect on untreated** (e.g., typical non-smokers)
- Average treatment effect (ATE) is calculated by summing (vertically) ATT and ATU, weighted by the percent the population in treatment (π) and control ($1-\pi$)
  - $ATE = E[\delta] = E[Y^1] - E[Y^0] = \pi E[Y^1|D=1] + (1-\pi)E[Y^1|D=0] + \pi E[Y^0|D=1] + (1-\pi)E[Y^0|D=0]$
- Notice: the three causal "definitions" – ATE, ATT and ATU – contain counterfactual information and therefore by definition **cannot be calculated with data**

Naive Average Treatment Effects (NATE)

Group	Y^1	Y^0	**Naive ATE**
Treatment ($D = 1$)	**Observable** as Y	–	
Control ($D = 0$)	–	**Observable** as Y	
			NATE $= E_N[y_i \mid d_i = 1] - E_N[y_i \mid d_i = 0]$

- Notice: diagonal subtraction, or just the differences in the observed groups' mean outcomes (e.g., cancer)
- Naive ATE isn't like any of the three we discussed (ATE, ATU, or ATT)
- Then what is it?

Decomposition of NATE

- Assume that a fixed portion of the population, π, always selects into treatment, and the rest of the population, $1 - \pi$, choose control.
- Assume that π is unchanging ("fixed") in the population and known simply by adding up and dividing by population
- Assume we "sample" the population using a large survey

Decomposition of NATE (cont.)

- ATE has five elements:

$$\begin{aligned} E[\delta] &= E[Y^1] - E[Y^0] \\ &= \{\pi E[Y^1|D=1] + (1-\pi)E[Y^1|D=0]\} \\ &\quad -\{\pi E[Y^0|D=1] + (1-\pi)E[Y^0|D=0]\} \end{aligned}$$

- Three of the five objects in equation (30) can be calculated from our survey:

$$\begin{aligned} E_N[d_i] &\rightarrow \pi \\ E_N[y_i|d_i=1] &\rightarrow E[Y^1|D=1] \\ E_N[y_i|d_i=0] &\rightarrow E[Y^0|D=0] \end{aligned}$$

- But not $E[Y^1|D=0]$ or $E[Y^0|D=1]$ as they are both **counterfactual**

Naive estimator, headache example

- You and I have headaches which we can rank on a scale of 1 to 10, with 1 being the mildest and 10 being a severe migraine
- I took an aspirin and record my headache as a **6**; you don't take an aspirin record yours as a **5**
- Assume had I not taken an aspirin, my headache would've been a 10; had you not taken an aspirin, your headache would've been a 3

Group	$E[Y^1\|.]$	$E[Y^0\|.]$	Treatment effect
Me ($D=1$)	**6**	10	$ATT = E[\delta\|D=1] =$ **-4**
You ($D=0$)	3	**5**	$ATU = E[\delta\|D=0] =$ **-2**
	$E[Y^1] = 4.5$	$E[Y^0] = 7.5$	$ATE = E[Y^1] - E[Y^0] =$ **-3**
			$NATE = E_N[Y\|D=1] - E_N[Y\|D=0] =$ **+1??**

- Aspirins clearly reduced headache severity for every person, ATT (-4), ATU (-2), ATE (-3)
- So how is it even possible that NATE says aspirin's caused a *positive change* in headache severity of +1?

Decomposition of NATE

- Step 1. Definition of ATE (equation 30)

$$\begin{aligned} E[\delta] &= E[Y^1] - E[Y^0] \\ &= \{\pi E[Y^1|D=1] + (1-\pi)E[Y^1|D=0]\} \\ &\quad -\{\pi E[Y^0|D=1] + (1-\pi)E[Y^0|D=0]\} \end{aligned}$$

- Step 2: Simplify

$$\begin{aligned} E[Y^1|D=1] &= a \\ E[Y^1|D=0] &= b \\ E[Y^0|D=1] &= c \\ E[Y^0|D=0] &= d \\ E[\delta] &= e \end{aligned} \tag{30}$$

- Step 3: Rewrite ATE (eq. 30') using the above variables

$$e = \{\pi a + (1-\pi)b\} - \{\pi c + (1-\pi)d\}$$

Decomposition of NATE (cont.)

- Step 4: Rearrange ATE (eq. 30') as follows:

$$
\begin{aligned}
e &= \{\pi a + (1-\pi)b\} - \{\pi c + (1-\pi)d\} \\
e &= \pi a + b - \pi b - \pi c - d + \pi d \\
e &= \pi a + b - \pi b - \pi c - d + \pi d + (\mathbf{a} - \mathbf{a}) + (\mathbf{c} - \mathbf{c}) + (\mathbf{d} - \mathbf{d}) \\
&= e - \pi a - b + \pi b + \pi c + d - \pi d - \mathbf{a} + \mathbf{a} - \mathbf{c} + \mathbf{c} - \mathbf{d} + \mathbf{d} \\
\mathbf{a} - \mathbf{d} &= e - \pi a - b + \pi b + \pi c + d - \pi d + \mathbf{a} - \mathbf{c} + \mathbf{c} - \mathbf{d} \\
\mathbf{a} - \mathbf{d} &= e + (\mathbf{c} - \mathbf{d}) + \mathbf{a} - \pi a - b + \pi b - \mathbf{c} + \pi c + d - \pi d \\
\mathbf{a} - \mathbf{d} &= e + (\mathbf{c} - \mathbf{d}) + (1-\pi)a - (1-\pi)b + (1-\pi)d - (1-\pi)c \\
\mathbf{a} - \mathbf{d} &= e + (\mathbf{c} - \mathbf{d}) + (1-\pi)(a-c) - (1-\pi)(b-d)
\end{aligned}
$$

- Step 5: Substitute original values back

$$
\begin{aligned}
E[Y^1|D=1] - E[Y^0|D=0] &= E[\delta] \\
&\quad + (E[Y^0|D=1] - E[Y^0|D=0]) \\
&\quad + (1-\pi)(\{E[Y^1|D=1] - E[Y^0|D=1]\} \\
&\quad - (1-\pi)\{E^1|D=0] - E[Y^0|D=0]\})
\end{aligned}
$$

- Step 6: Substitute equations 26, 27, 28

$$
\begin{aligned}
E[Y^1|D=1] - E[Y^0|D=0] &= ATE \\
&\quad + (E[Y^0|D=1] - E[Y^0|D=0]) \\
&\quad + (1-\pi)(ATT - ATU)
\end{aligned}
$$

Decomposition of NATE (cont.)

$$\begin{aligned} NATE &= ATE \\ &\quad +(E[Y^0|D=1] - E[Y^0|D=0]) \\ &\quad +(1-\pi)(ATT - ATU) \end{aligned} \tag{31}$$

- NATE is equal to the ATE + the second line + the third line
  - Second line is the average "baseline" differences between the two groups in a world where neither group receives treatment
  - Third line is difference between the ATT and the ATU weighted by the share of population in control
- The difference in means will only measure ATE if the second line equals zero and the third line equals zero, otherwise NATE is a "biased estimator of the ATE"

Ignorable treatment assignment

- What if we were to **randomly** assign units to receive some treatment – say, using a coin flip or some other randomizing device
- What could we assume about the differences in the averages of the two groups if treatment was random?

Ignorable treatment assignment

- What if we were to **randomly** assign units to receive some treatment – say, using a coin flip or some other randomizing device
- What could we assume about the differences in the averages of the two groups if treatment was random?
- Can you see why the following conditions should hold under randomization? Explain them in your own words.

$$E[Y^0|D=1] = E[Y^0|D=0] \quad (32)$$

$$E[Y^1|D=1] = E[Y^1|D=0] \quad (33)$$

Ignorable treatment assignment

- What if we were to **randomly** assign units to receive some treatment – say, using a coin flip or some other randomizing device
- What could we assume about the differences in the averages of the two groups if treatment was random?
- Can you see why the following conditions should hold under randomization? Explain them in your own words.

$$E[Y^0|D=1] = E[Y^0|D=0] \quad (34)$$
$$E[Y^1|D=1] = E[Y^1|D=0] \quad (35)$$

- Recalculate NATE if equations (32) and (33) hold and explain the results

Ignorable treatment assignment

- Substitute eq. (32) into (31)

$$\begin{aligned} NATE &= ATE \\ &\quad +(E[Y^0|D=1]-E[Y^0|D=0]) \\ &\quad +(1-\pi)(ATT-ATU) \\ NATE &= ATE+(1-\pi)(ATT-ATU) \end{aligned}$$

- Decompose ATT-ATU using eq. (27) and (28) and rearrange

$$\begin{aligned} ATT-ATU &= E[Y^1|D=1]-E[Y^0|D=1] \\ &\quad -E[Y^1|D=0]+E[Y^0|D=0] \\ &= E[Y^1|D=1]-E[Y^1|D=0] \\ &\quad +E[Y^0|D=0]-E[Y^0|D=1] \end{aligned}$$

- Since top row is zero by equation (32) and bottom row is zero by equation (33), $NATE = ATE$ under randomized treatment assignment

Ignorable treatment assignment

- "Ignorable treatment assignment", $(Y^0, Y^1) \perp\!\!\!\perp D$, is when units are assigned to receive some treatment independent of their potential outcomes
  - Ignorable treatment yields equations 32-33
- What do the equalities in equation 32-33 mean exactly?
  - Equation 32 means units in treatment are on average the same as those who aren't if neither got treated
  - Equation 33 means units in treatment are on average the same as those who aren't if both had been treated
- But since each comparison contains counterfactuals, how do we know they hold under randomization? In other words, why is randomization so crucial?

Ignorable treatment assignment

- Randomization is not, in fact, what I want you to learn from this exercise – equations 32-33 are what I mainly want you to understand and notice
- Randomization is special because randomization gives us gives us credible reasons, *a priori*, to believe equations 32-33 will hold.
- But as we will see, the rest of the class is mainly about identifying causal effects when randomization is not available to us. We start with this ideal situation, though, so that the causal effect parameters are clearly defined, and so that we know precisely what problems we face when going to data to try and estimate them
- Exercise: What if equation 32 can be credibly assumed, but you're not sure about equation 32? What does NATE identify?

Stable Unit-Treatment Value Assumption (SUTVA)

- Counterfactual tradition maintains an important but largely overlooked condition called SUTVA
- SUTVA stands for "stable unit-treatment value assumption", and might be better understood reading the acronym backwards than forwards
  1. **A**: It's an **a***ssumption*
  2. **TV**: that the **t***reatment-value* (or "treatment effect" or "causal effect")
  3. **S**: is **s***table*
  4. **U**: for all **u***nits*, or participants
- SUTVA is the *a priori* assumption that the value of Y for unit i will be the same when exposed to treatment D no matter what mechanism is used to assign the treatment D to unit i and no matter what treatments the other i units receive. It also assumes that the treatment is the same for all units.

SUTVA (cont)

- Causal effect definitions in this tradition always require SUTVA
  1. Every person receives the same dosage
  2. Potential outcomes of individuals must be unaffected by potential changes in the treatment exposures of other individuals
- Example of SUTVA violation: Externalities from treatment
  - When I took my aspirin, my headache was recorded as a 6. What if everyone in my family was given aspirin at the same time as me, and in so doing, I had recorded my headache as a 3? Why might more of us taking aspirin change the efficacy of the treatment?
  - SUTVA requires that treatment effects be invariant to how many are in treatment or control – treatment effects are "stable", because they are based on "ceteris paribus" assumptions holding
- Another form of SUTVA violation: Heterogeneity in dosage
  - I get regular aspirin, but someone else gets extra strength. SUTVA is violated because our treatments are not the same.
  - **Heterogeneity in dosage**, in other words, violates SUTVA. But heterogenous treatment effects are fine

History of graphical causal modeling in science

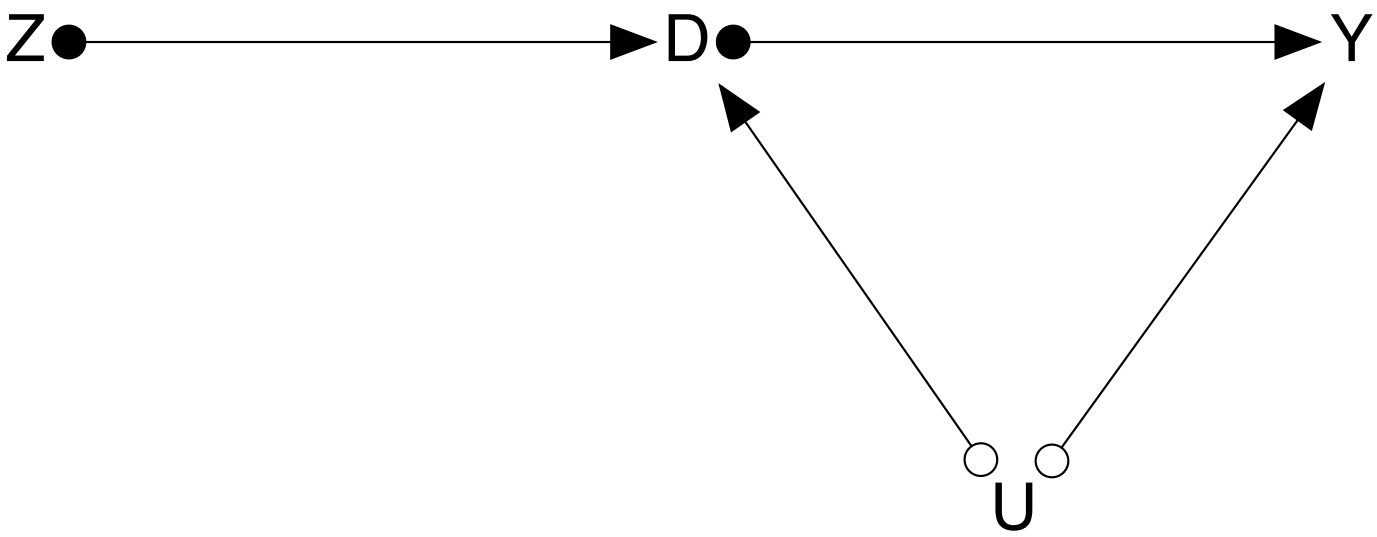
- Path diagrams were developed by Sewell Wright, early 20th century geneticist, for causal inference
- Sewell Wright's father, Phillip Wright, used them to prove the existence of the "instrumental variable" estimator (see Stock and Trebbi)
- Directed acyclic graphs (DAG) are not the same thing as path diagrams (Freedman ch. 6), though
- Judea Pearl and colleagues in Artificial Intelligence at UCLA developed DAG to help build functioning robots
- Both methods use figures with arrows to illustrate causality, but there are important and subtle differences between path analysis and DAGs

Three methods for causal inference

- Conditioning to block back door paths (e.g., matching and regression)
- Instrumental variables, regression discontinuity and randomized experiments
- Condition variables that allows for estimation by a mechanism

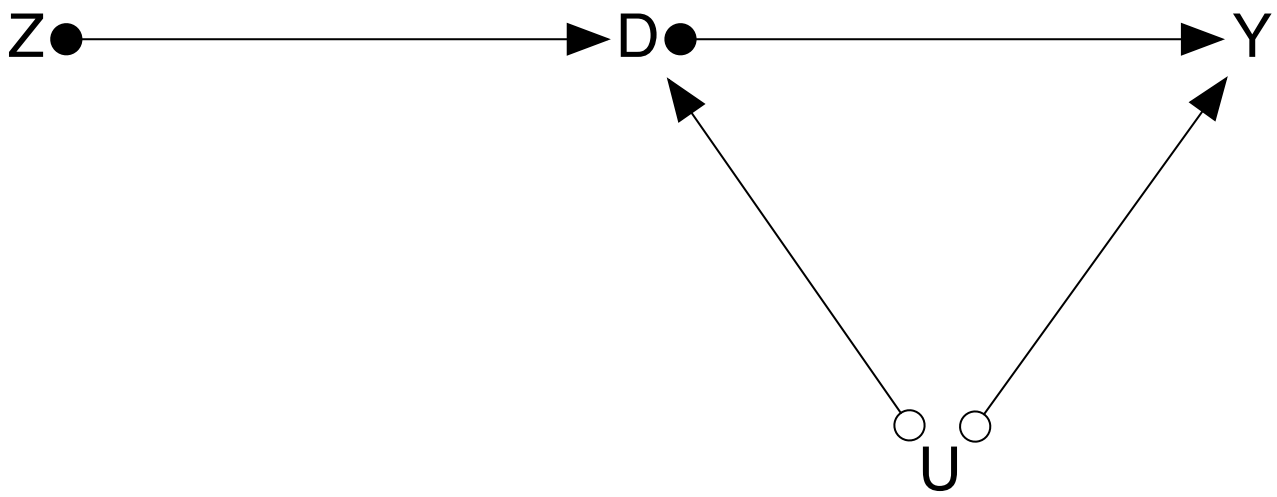
What is a DAG?

A **Directed Acyclic Graph (DAG)** is a set of nodes (letters and/or solid dots) and directed edges (lines with one arrow) with no directed cycles



- Nodes represent variables (e.g., Y is wage, D is education, U are unobserved determinants of salary, Z is parental education)
  - Solid circles are observed (by the researcher)
  - Hollow circles are unobserved (by the researcher)
- Arrows represent "direct" causal effects, and "direct" means that the variable isn't mediated by other variables in the graph
- A **causal DAG** must include:
  - All direct causal effects among the variables in the graph
  - All common causes (even if unmeasured) of any pair of variables in the graph

Basic DAG Concepts



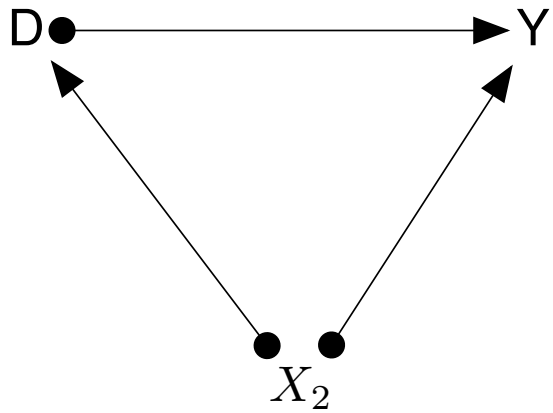
- U is a **parent** of D and Y
- D and Y are **descendants** of Z
- There is a **directed path** from Z to Y
- There are two paths from Z to U, but **no directed path**
- X is a collider of the path $Z \rightarrow D \leftarrow U$
- X is a non-collider of the path $Z \rightarrow D \rightarrow Y$

Three types of directed edges

1. Chain of mediation: $A \rightarrow C \rightarrow B$
   - A's causal effect on B is "mediated" by C
   - A and B are indirectly correlated – "unconditional association"
2. Fork of mutual dependence: $A \leftarrow C \rightarrow B$
   - A and B are both caused by C
   - A and B are indirectly correlated (unconditional association)
3. Inverted fork of mutual causation: $A \rightarrow C \leftarrow B$
   - A and B both cause C
   - C is a "collider" along non-directed path $A \rightarrow C \leftarrow B$
     - ⋆ If a variable receives two arrows, it is a "collider"
     - ⋆ NOTE: A and B's causal effects "collide" at C
   - A and B are uncorrelated, but if we condition on the collider, C, it will create spurious correlations between A and B

Bias #1: Confounding and backdoor paths

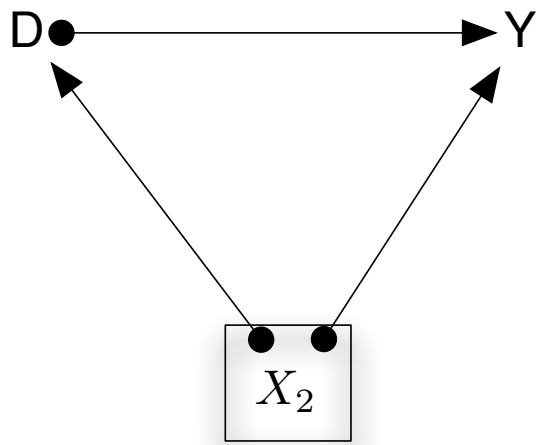
1. **Confounding** is when the causal variable, D, and the outcome, Y, are correlated through a fork of mutual dependence
   - Ex: Assume college increases earnings, labor market discrimination lowers minorities' earning (with or without college), and minorities attend college less than whites. Let D be college degree; Y earnings; and X_2 the worker's race-ethnicity.



- Race creates a **backdoor path** from D to Y through $D \leftarrow X_2 \rightarrow Y$
  - Confounders create backdoor paths between causal variables and outcomes of interest
  - Backdoor path: $D \leftarrow X_2 \rightarrow Y$

Solution #1: Blocking backdoor paths

- Question: How do we isolate the effect of college on earnings under confounding?
  - **Answer**: "Block" the backdoor path by **conditioning** on race (X_2)
- How-to-block-a-backdoor-path
  1. Calculate the correlation between D and Y for every strata of X_2
  2. Correctly combine $E[Y|D=1, Whites] - E[Y|D=0, Whites]$ and $E[Y|D=1, Blacks] - E[Y|D=0, Blacks]$
  3. Combination of the correlations for each strata of X_2 yields the unbiased estimate of ATE and can be done using multivariate regression, stratification/subclassification, and/or matching
- Visual representation of conditioning to close a backdoor path using DAGs:



Solution #1: Blocking backdoor paths (cont.)

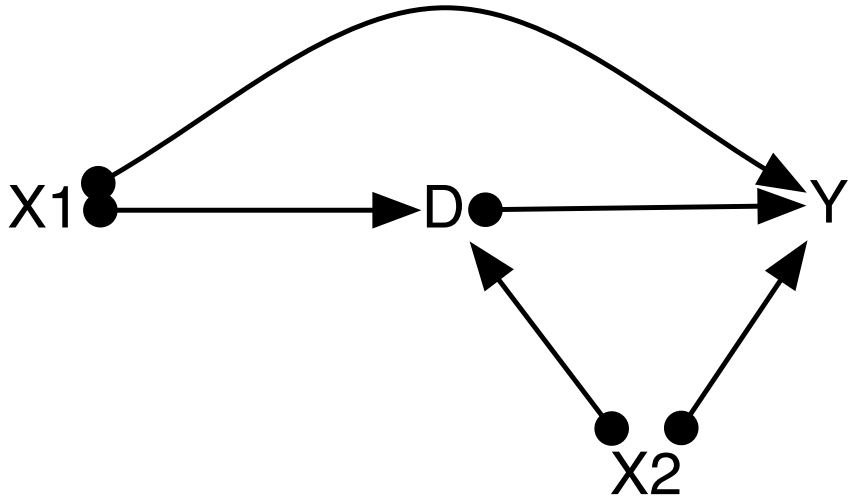
- A path is blocked if and only if:
  - the path contains a non-collider that has been conditioned on,
  - or the path contains a collider that has not been conditioned on and has no descendants that have been conditioned on.
- Words to live by with regards to conditioning:
  - Conditioning on a non-collider **blocks** a backdoor path
  - Conditioning on a collider opens a path (i.e., spurious correlation)
  - Not conditioning on a collider (or its descendants) blocks a path

Solution #1: "Backdoor criterion"

- Assume the directed edge, $D \rightarrow Y$, and backdoor paths connecting D and Y (e.g., forks of mutual dependence)
- The **backdoor criterion** specifies the necessary and sufficient conditions for identifying $D \rightarrow Y$
  - Conditioning on a set of variables, Z, will identify $D \rightarrow Y$ *if and only if* all backdoor paths from D to Y are blocked after conditioning on Z
- Z blocks all backdoor paths if and only if each backdoor path
  1. contains a chain of mediation $D \rightarrow Z \rightarrow Y$ **or**
  2. contains a fork of mutual dependence $D \leftarrow Z \rightarrow Y$, **or**
  3. contains an inverted fork of mutual causation $D \rightarrow C \leftarrow Y$, where C and all of C's descendants are **not** in Z.

Solution #1: Examples

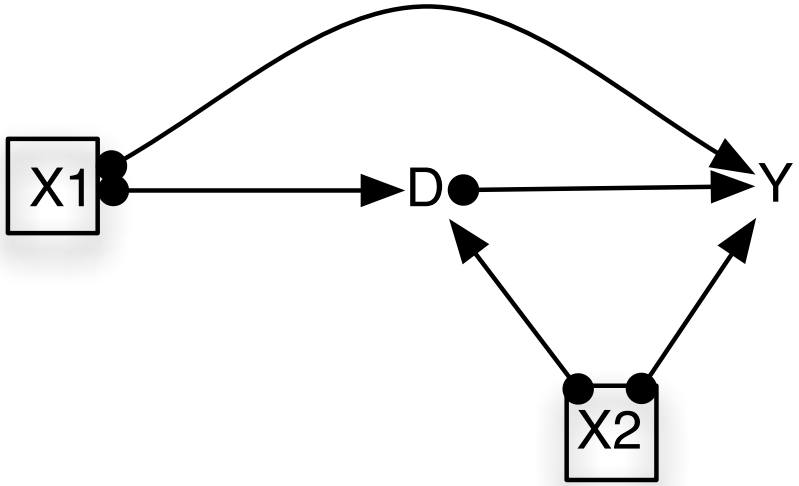
- **Matching on all common causes is sufficient**: There are two backdoor paths from D to Y.



1. Backdoor path 1: $D \leftarrow X1 \rightarrow Y$ (OPEN)
2. Backdoor path 2: $D \leftarrow X2 \rightarrow Y$ (OPEN)

Solution #1: Examples

- **Matching on all common causes is sufficient**: There are two backdoor paths from D to Y.

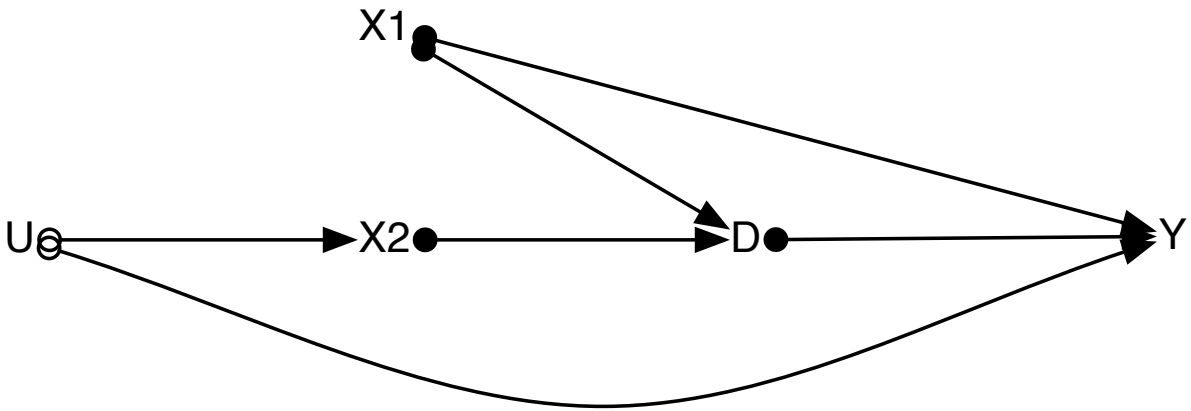


1. Backdoor path 1: $D \leftarrow \boxed{X1} \rightarrow Y$ (CLOSED)
2. Backdoor path 2: $D \leftarrow \boxed{X2} \rightarrow Y$ (CLOSED)

Conditioning on $X1$ and $X2$ blocks **both** backdoor paths and therefore meets the backdoor criterion.

Solution #1: Examples

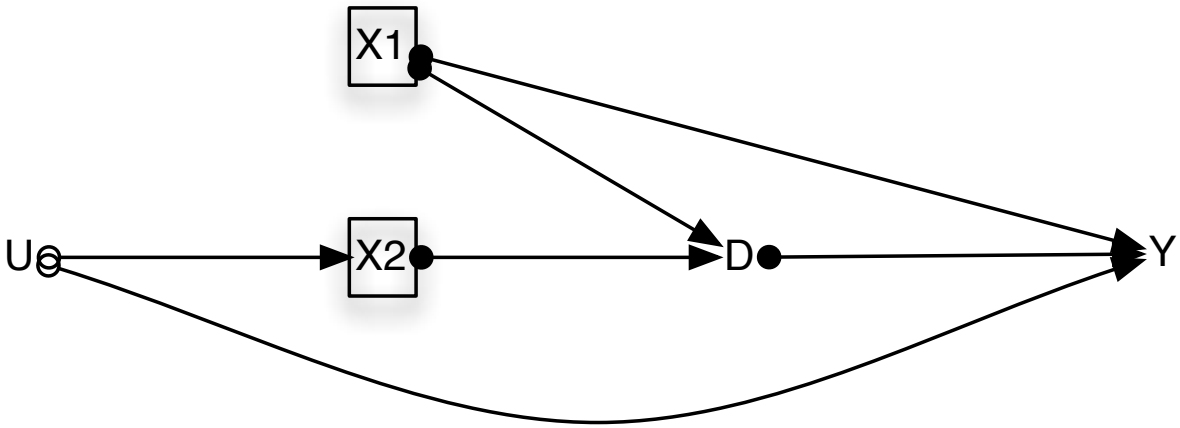
- **Matching may work even if not all common causes are observed**: U and $X1$ are common causes.



1. Backdoor path 1: $D \leftarrow X1 \rightarrow Y$ (OPEN)
2. Backdoor path 2: $D \leftarrow X2 \leftarrow U \rightarrow Y$ (OPEN)

Solution #1: Examples

- **Matching may work even if not all common causes are observed**: U and $X1$ are common causes.

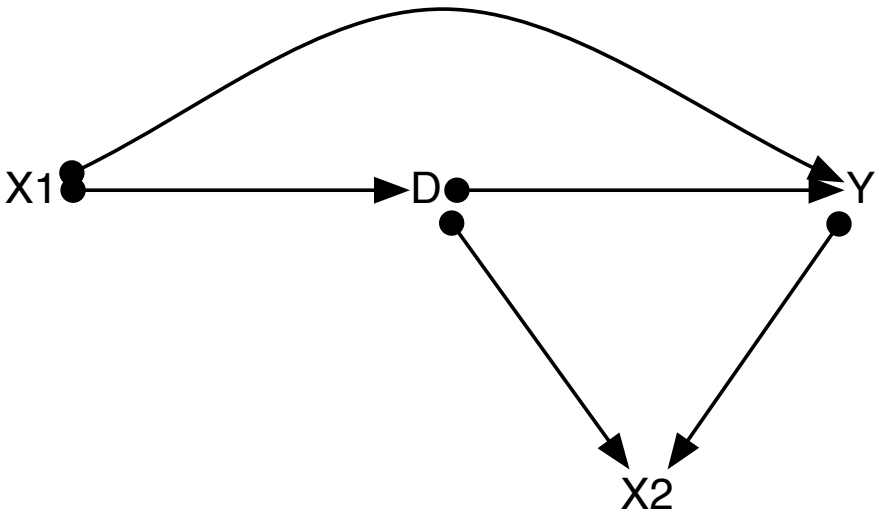


1. Backdoor path 1: $D \leftarrow \boxed{X1} \rightarrow Y$ (CLOSED)
2. Backdoor path 2: $D \leftarrow \boxed{X2} \leftarrow U \rightarrow Y$ (CLOSED)

Conditioning on $X1$ and $X2$ is sufficient to identify $D \rightarrow Y$

Bias #2: Conditioning on a collider

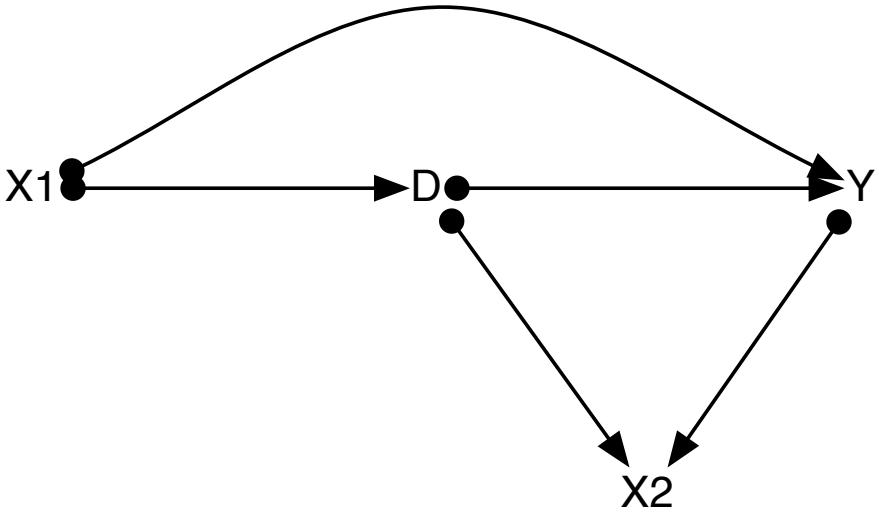
- **Matching on an outcome may create bias**: There is only one backdoor path from D to Y.



1. Backdoor path: $D \leftarrow X1 \rightarrow Y$ (OPEN)
2. Collider Path: $D \rightarrow X2 \leftarrow Y$ (OPEN or CLOSED?)

Bias #2: Conditioning on a collider

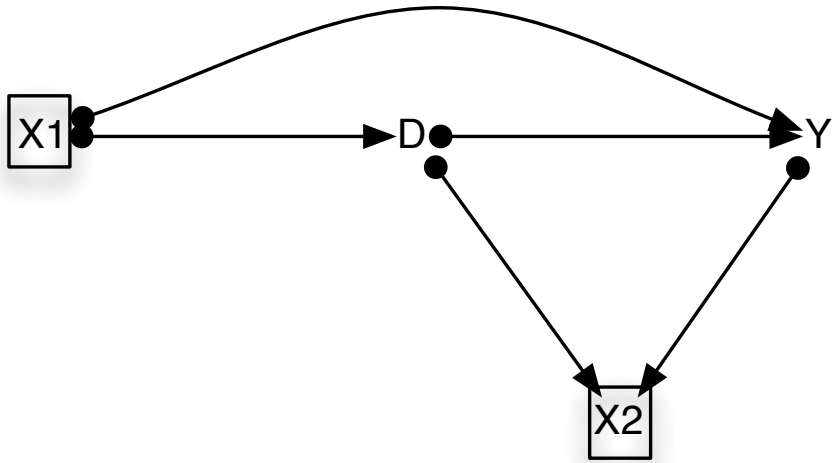
- **Matching on an outcome may create bias**: There is only one backdoor path from D to Y.



1. Backdoor path: $D \leftarrow X1 \rightarrow Y$ (OPEN)
2. Collider Path: $D \rightarrow X2 \leftarrow Y$ (CLOSED)

Bias #2: Conditioning on a collider

- **Matching on an outcome may create bias**: There is only one backdoor path from D to Y.



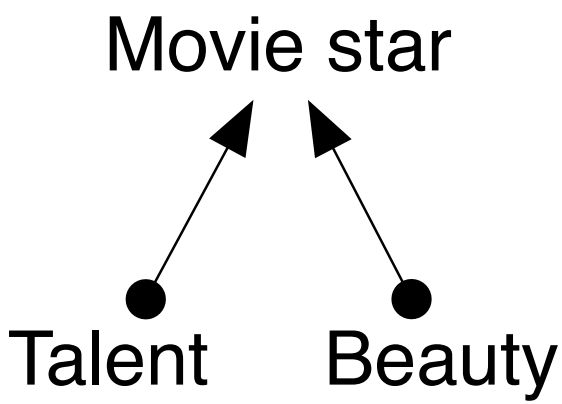
1. Backdoor path: $D \leftarrow \boxed{X1} \rightarrow Y$ (CLOSED)
2. Collider Path: $D \rightarrow \boxed{X2} \leftarrow Y$ (OPEN)

- Conditioning on $X1$ blocks the backdoor path, but conditioning on $X2$ opens the collider path. So what, right?

Bias #2: Conditioning on a collider

Important: Since unconditioned colliders block back-door paths, what exactly does conditioning on a collider do? Let's illustrate with a fun example and some made-up data

- CNN.com headline: Megan Fox voted worst – but sexiest – actress of 2009
  `http://marquee.blogs.cnn.com/2009/12/30/megan-fox-voted-worst-but-sexiest-actress-of-2009/`
- Is this negative correlation causal? Or this is what happens when we condition on a collider?
- Assume talent and beauty are independent, but each causes someone to become a movie star. What's the correlation between talent and beauty for a sample of movie stars compared to the population as a whole (stars and non-stars)?



Bias #2: Collider do-file (STATA)

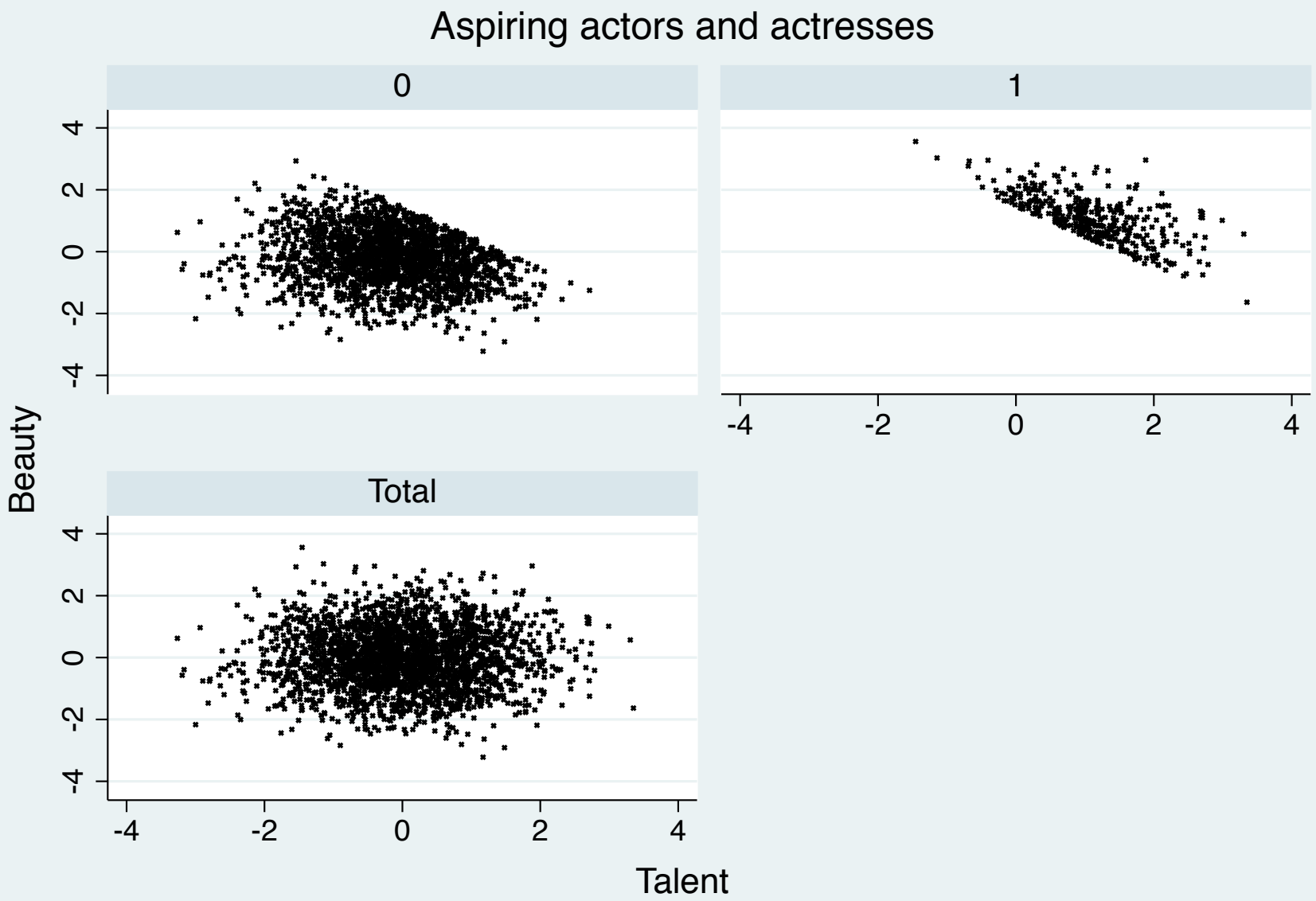
```
clear all
set seed 3444

* 2500 independent draws from standard normal distribution
set obs 2500
generate beauty=rnormal()
generate talent=rnormal()

* Creating the collider variable (star)
gen score=(beauty+talent)
egen c85=pctile(score), p(85)
gen star=(score>=c85)
label variable star "Movie star"

* Conditioning on the top 15%
twoway (scatter beauty talent, mcolor(black) msize(small)
msymbol(smx)), ytitle(Beauty) xtitle(Talent) subtitle(Aspiring actors
and actresses)) by(star, total)
```

Bias #2: Scatterplots



Summary

- We are working with two separate sets of conceptual tools to clarify what is meant by "causality" – the potential outcomes model and the DAG
- These are complementary tools describing the same idea – causality is thought of as comparing the same unit under two potential outcomes (treatment vs control)
- Decomposition of NATE reveals correlations contain additive selection bias and treatment heterogeneity bias in addition to true causal effects
- Randomized treatment assignment removes both forms of bias and allows NATE to identify ATE
- DAGs reveal how these additive selection problems can affect simple differences in means, as well as showcase how we might identify the causal effects when there are problems like confounding
- Conditioning on a collider introduces spurious correlation and

Fundamental problem of causal inference is missing data: we do not observe workers under both potential outcomes We have to find someone else to substitute for John's unobserved counterfactual state *What's different about John's counterfactual?*

Finding a credible counterfactual substitute is the crux of all sound causal inference

The best counterfactual substitutes are exchangeable – it's all about the comparison group!

Randomization is an important part of this because randomization ensures that the many other factors that also determine wages in the population have been equally distributed between the potential outcomes

Example: Does Information Matter?

- We want to estimate the effect on voting behavior of paying attention to an election campaign
- Survey research consistently finds a largely ignorant voting public (e.g., Berelson, Lazasfeld, and McPhee 1954; Campbell, Converse, Miller and Stokes 1960; Zaller 1992)
- There's the `fact` of public ignorance and there's the `meaning` of public ignorance – although the fact of public ignorance is largely accepted, the meaning has been challenged (Sniderman 1993).
- Can voters use information such as polls, interest group endorsements and partisan labels to vote like their better informed compatriots (e.g., Lupia 2004; McKelvey and Ordeshoot 1985a,b, 1986)

Fundamental Problem of Causal Inference

- Fundamental problem: we aren't observing all of the potential outcomes or counterfactuals. Frame the question using potential outcomes notation.
- Let Y_{i1} denote voter i's vote intention when voter i learns during the campaign
  - Call this the `treatment` group
- Let Y_{i0} denote voter i's vote intention when voter i does not learn during the campaign
  - Call this the `control` group

Fundamental Problem of Causal Inference

Let T_i be a `treatment` indicator: 1 when i is in the `treatment` group and 0 in the `control` group.

- The observed outcome for observation i is:
  - $Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0}$.
- If $T_i = 1$, ...
  - $Y_i = 1 Y_{i1} + (1 - 1) Y_{i0}$
  - $Y_i = Y_{i1}$
- If $T_i = 0$, ...
  - $Y_i = 0 Y_{i1} + (1 - 0) Y_{i0}$
  - $Y_i = Y_{i0}$
- And the treatment effect for i is ...
  - $Y_{i1} - Y_{i0}$
- Note that the causal effect of learning is the difference in the observed outcome to the `counterfactual` (i.e. the outcome when the voter did not learn).

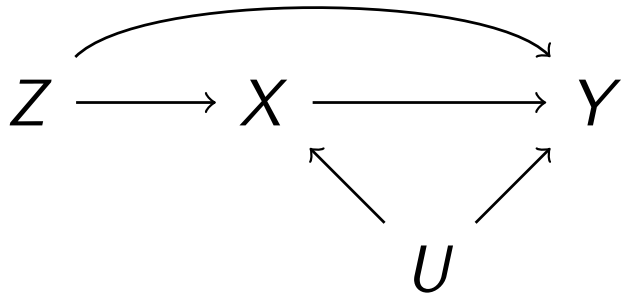
Experimental Data

- If assignment to treatment is randomized, the inference problem is simple because the two groups are from the same population.
  - $\{Y_{i1}, Y_{i0} \perp T_i\}$
  - $\perp$ means that the two potential outcomes, Y_{i1} and Y_{i0} are independent of the treatment itself
  - This is not the same thing as saying that the treatment does not affect the outcome
- Observations in the treatment and control groups are not exactly the same, but they are comparable – they are exchangeable
- Hence, for $j = 0, 1$ we have exchangeability:

$$E(Y_{ij}|T_i = 1) = E(Y_{ij}|T_i = 0) = E(Y_i|T_i = j)$$

What to match on: a brief introduction to DAGs

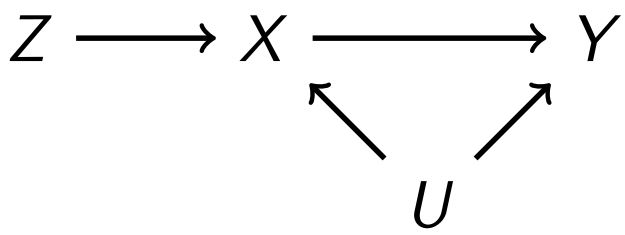
A **Directed Acyclic Graph** (DAG) is a set of nodes (vertices) and directed edges (arrows) with no directed circles



- Nodes represent variables
- Arrows represent direct causal effects ("direct" means not mediated by other variables in the graph)
- A **causal DAG** must include
  - All direct causal effects among the variables in the graph
  - All common causes (even if unmeasured) of any pair of variables in the graph

Some DAG concepts

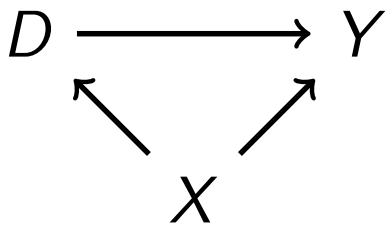
In the DAG:



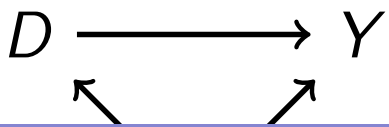
- U is a **parent** of X and Y
- X and Y are **descendants** of Z
- There is a **directed path** from Z to Y
- There are two **paths** from Z to U (but no directed path)
- X is a **collider** of the path $Z \rightarrow X \leftarrow U$
- X is a **noncollider** of the path $Z \rightarrow X \leftarrow Y$

Confounding

- Confounding arises when the treatment and the outcomes have common causes



  The association between D and Y does not only reflect the causal effect of D on Y
- Confounding creates **backdoor paths**, that is, paths starting with incoming arrows
- In the DAG, we can see a backdoor path from D to Y $(D \leftarrow X \rightarrow Y)$
- However, once we "block" the backdoor path by conditioning on the common cause X, the association between D and Y is only reflective of the effect of D on Y



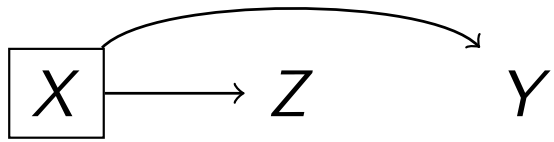
Blocked paths

A path is blocked if and only if:

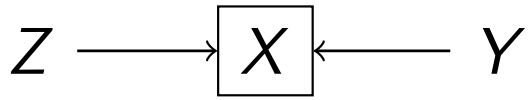
- It contains a noncollider that has been conditioned on
- Or it contains a collider that has not been conditioned on and has no descendents that have been conditioned on

Examples:

1. Conditioning on a noncollider blocks a path:



2. Conditioning on a collider opens a path:



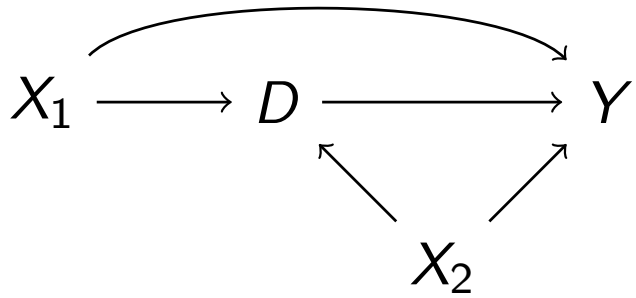
3. Not conditioning on a collider (or its descendents) leaves a path blocked:

Backdoor criterion

- Suppose that
  - D is a treatment
  - Y is an outcome
  - $X_1, ..., X_k$ is a set of covariates
- Is it enough to match on $X_1, ..., X_k$ in order to estimate the causal effect of D on Y?
- **?**'s (**?**) **backdoor criterion** provides sufficient conditions
- Backdoor criterion: $X_1, ..., X_k$ satisfies the backdoor criterion with respect to (D, Y) if:
  1. No element of $X_1, ..., X_k$ is a descendent of D
  2. All backdoor paths from D to Y are blocked by $X_1, ..., X_k$
- If $X_1, ..., X_k$ satisfies the backdoor criterion with respect to (D, Y), then matching on $X_1, ..., X_k$ identifies the causal effect of D on Y

Implications for practice I

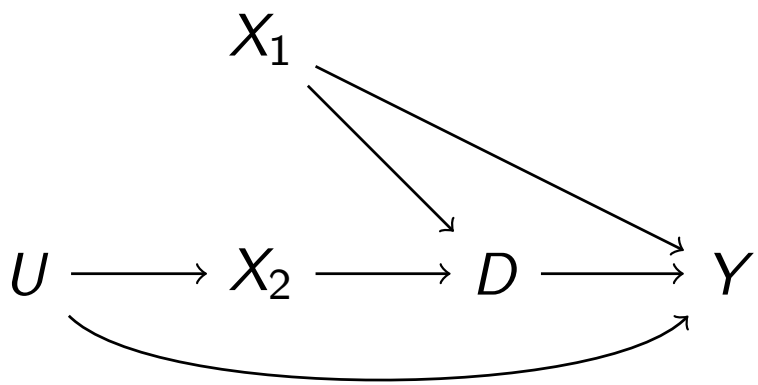
- **Matching on all common causes is sufficient**
  - There are two backdoor paths from D to Y



  - Conditioning on X_1 and X_2 blocks the backdoor paths

Implications for practice II

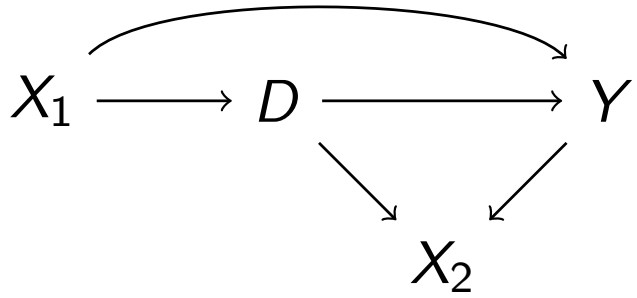
- **Matching may work even if not all common causes are observed**
  - U and X_1 are common causes



  - Conditioning on X_1 and X_2 is enough

Implications for practice III

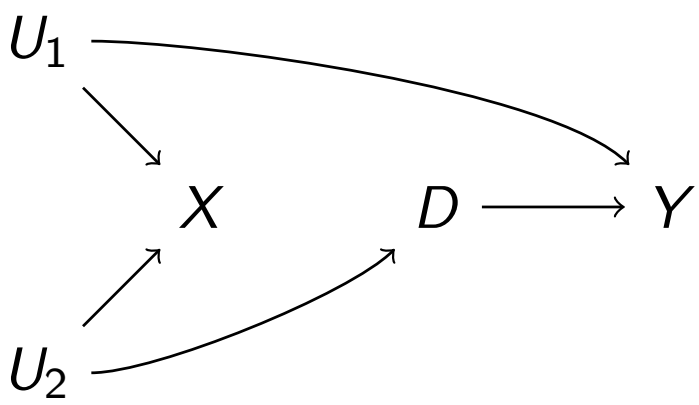
- **Matching on an outcome may create bias**
  - There is only one backdoor path from D to Y



  - Conditioning on X_1 blocks the backdoor path
  - Conditioning on X_2 would open a path!

Implications for practice IV

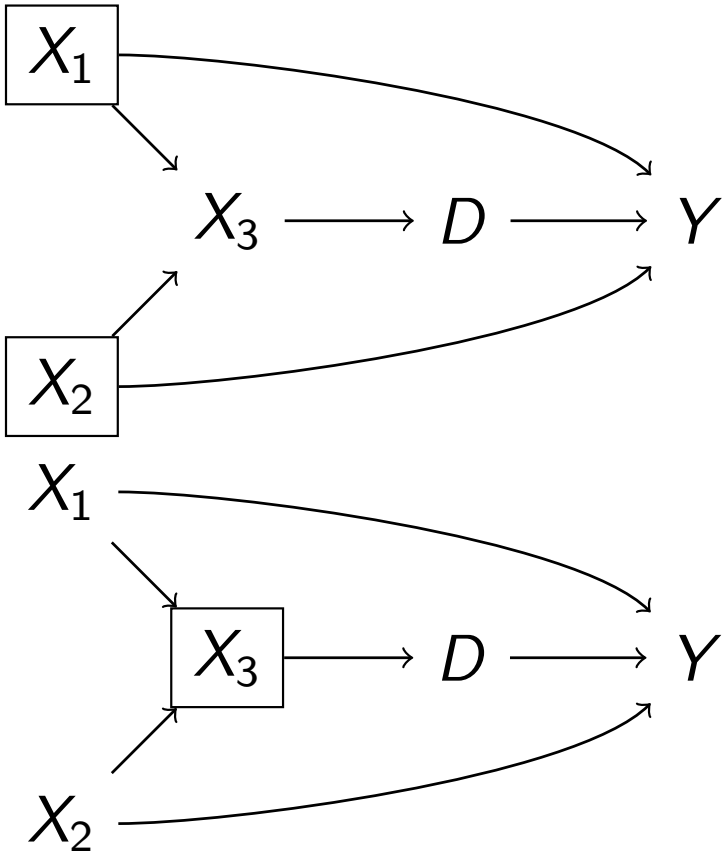
- **Matching on all pretreatment covariates is not always the answer**
  - There is one backdoor path and it is closed



  - No confounding; conditioning on X would open a path!

Implications for practice V

- **There may be more than one set of conditioning variables that satisfy the backdoor criterion**



Implications for practice VI

- Conditioning on the common causes, X_1 and X_2, is sufficient, as always
- But conditioning on X_3 only also blocks the backdoor paths

DAGs: final remarks

- The backdoor criterion provides a useful graphical test that can be used to select matching variables
- The backdoor criterion is only a set of sufficient conditions, but covers most of the interesting cases
  - There are general results (**?**)
- Applying the backdoor criterion requires knowledge of the DAG
  - There are results on the validity of matching under limited knowledge of the DAG (**?**)
  - Suppose that there is a set of observed pretreatment covariates and that we know if they are causes of the treatment and/or the outcomes
  - Suppose that there exists a subset of the observed covariates such that matching on them is enough to control for confounding
  - Then, matching on the subset of the observed covariates that are either a cause of the treatment or a cause of the outcome or both is also enough to control for confounding