Controls

Zahid Asghar

10/16/22

Endogeneity

- Last week was all about handling sampling variation and avoiding inference error
- This week we're all about endogeneity!
- Where it pops up and what we can do about it
- At least as a starter (we'll revisit this topic many times)

Endogeneity Recap

- We believe that our true model looks like this:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Where ε is *everything that determines* Y *other than* X
- If X is related to some of those things, we have endogeneity
- Estimating the above model by OLS, it will mistake the effect of those *other* things for the effect of X, and our estimate of $\hat{\beta}_1$ won't represent the true β_1 no matter how many observations we have

Endogeneity Recap

- For example, the model

$$IceCreamEating = \beta_0 + \beta_1 ShortsWearing + \varepsilon$$

- The true β_1 is probably 0. But since $Temperature$ is in ε and $Temperature$ is related to $ShortsWearing$, OLS will mistakenly assign the effect of $Temperature$ to the effect of $ShortsWearing$, making it look like there's a positive effect when there isn't one
- If $Temperature$ hangs around $ShortsWearing$, but OLS doesn't know about it, OLS will give $ShortsWearing$ all the credit for $Temperature$'s impact on $IceCreamEating$
- Here we're mistakenly finding a positive effect when the truth is 0, but it could be anything - negative effect when truth is 0, positive effect when the truth is a bigger/smaller positive effect, negative effect when truth is positive, etc. etc.

To the Rescue

- One way we can solve this problem is through the use of *control variables*
- What if $Temperature$ *weren't* in ε? Then we'd be fine! OLS would know how to separate out its effect from the $ShortsWearing$ effect. How do we take it out? Just put it in the model directly!

$$IceCreamEating = \beta_0 + \beta_1 ShortsWearing + \beta_2 Temperature + \varepsilon$$

- Now we have a *multivariate* regression model. Our estimate $\hat{\beta}_1$ will *not* be biased by $Temperature$ because we've controlled for it

(probably more accurate to say "covariates" or "variables to adjust for" than "control variables" and "adjust for" rather than "control for" but hey what are you gonna do, "control" is standard)

To the Rescue

- So the task of solving our endogeneity problems in estimating β_1 using $\hat{\beta}_1$ comes down to us *finding all the elements of ε that are related to X and adding them to the model*
- As we add them, they leave ε and hopefully we end up with a version of ε that is no longer related to X
- If $cov(X, \varepsilon) = 0$ then we have an unbiased estimate!
- (of course, we have no way of checking if that's true - it's based on what we think the data generating process looks like)

How?

- How does this actually work?
- Controlling for a variable works by *removing variation in X and Y that is explained by the control variable*
- So our estimate of $\hat{\beta}_1$ is based on *just the variation in X and Y that is unrelated to the control variable*
- Any accidentally-assigning-the-value-of-Temperature-to-ShortsWearing can't happen because we've removed the effect of $Temperature$ on $ShortsWearing$ as well as the effect of $Temperature$ on $IceCreamEating$
- We're asking at that point, *holding $Temperature$ constant*, i.e. *comparing two different days with the same $Temperature$*, how is $ShortsWearing$ related to $IceCreamEating$?
- We know we're comparing within the same $Temperature$ because we literally subtracted out all the $Temperature$ differences!

Example

The true effect is $\beta_1 = 3$. Notice Z is binary and is related to X and Y but isn't in the model!

```
tib <- tibble(Z = 1*(rnorm(1000) > 0)) %>%
  mutate(X = Z + rnorm(1000)) %>%
  mutate(Y = 2 + 3*X + 2*Z + rnorm(1000))
feols(Y~X, data = tib) %>%
  etable()
##                                  .
## Dependent Var.:                 Y
##
## (Intercept)     2.756*** (0.0461)
## X               3.421*** (0.0383)
## _______________ _________________
## S.E. type                     IID
## Observations                1,000
## R2                        0.88897
## Adj. R2                   0.88886
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example

To remove what part of X and Y is explained by Z, we can get the mean of X and Y by values of Z

```r
tib <- tib %>%
  group_by(Z) %>%
  mutate(Y_mean = mean(Y), X_mean = mean(X))
head(tib)
## # A tibble: 6 × 5
## # Groups:   Z [2]
##         Z       X      Y Y_mean  X_mean
##     <dbl>   <dbl>  <dbl>  <dbl>   <dbl>
## 1       1  0.855   5.98    6.91  0.967
## 2       1 -1.59   -0.226   6.91  0.967
## 3       1  0.806   6.08    6.91  0.967
## 4       1 -0.0302  2.94    6.91  0.967
## 5       0 -0.490  -0.997   1.95 -0.0121
## 6       0 -0.512   0.383   1.95 -0.0121
```

Example

Now, `Y_mean` and `X_mean` are the mean of Y and X for the values of Z, i.e. the part of Y and X *explained by* Z. So subtract those parts out to get *residuals* `Y_res` and `X_res`!

```r
tib <- tib %>%
  mutate(Y_res = Y - Y_mean, X_res = X - X_mean)
head(tib)
## # A tibble: 6 × 7
## # Groups:   Z [2]
##       Z       X      Y Y_mean  X_mean  Y_res  X_res
##   <dbl>   <dbl>  <dbl>  <dbl>   <dbl>  <dbl>  <dbl>
## 1     1  0.855   5.98    6.91  0.967  -0.928 -0.113
## 2     1 -1.59   -0.226   6.91  0.967  -7.14  -2.55
## 3     1  0.806   6.08    6.91  0.967  -0.827 -0.161
## 4     1 -0.0302  2.94    6.91  0.967  -3.98  -0.997
## 5     0 -0.490  -0.997   1.95 -0.0121 -2.95  -0.478
## 6     0 -0.512   0.383   1.95 -0.0121 -1.57  -0.499
```

Example

What do we get now?

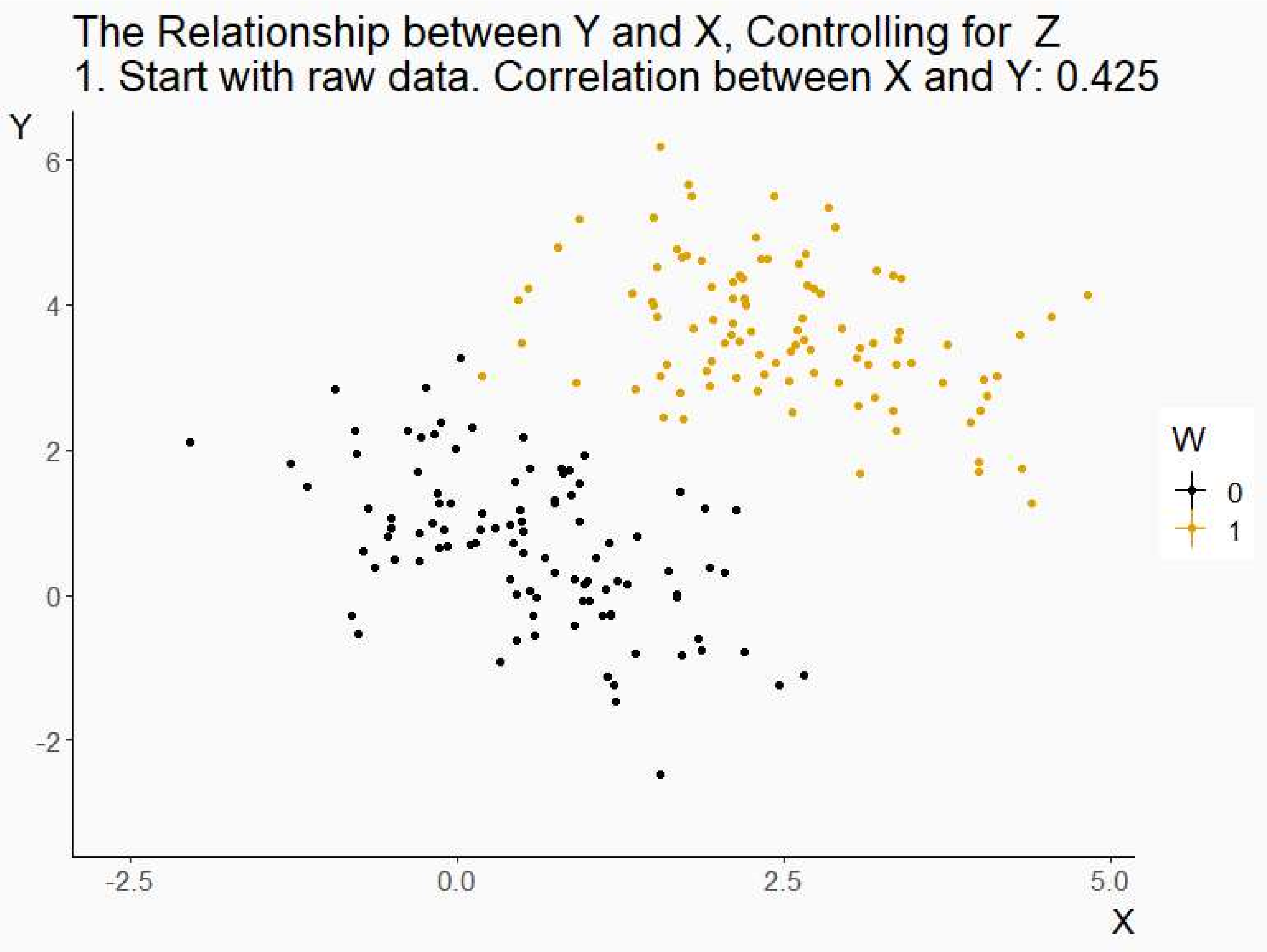
```
feols(Y_res ~ X_res, data = tib) %>%
  etable()
##                                   .
## Dependent Var.:               Y_res
##
## (Intercept)     6.13e-18 (0.0319)
## X_res           3.030*** (0.0319)
## _______________ _________________
## S.E. type                     IID
## Observations                1,000
## R2                        0.90060
## Adj. R2                   0.90050
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example

Compare this to actually including Z as a control:

```
feols(Y ~ X + Z, data = tib) %>%
  etable()
##                                   .
## Dependent Var.:                   Y
##
## (Intercept)      1.988*** (0.0441)
## X                3.030*** (0.0319)
## Z                1.993*** (0.0712)
## _______________ _________________
## S.E. type                       IID
## Observations                  1,000
## R2                          0.93782
## Adj. R2                     0.93769
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

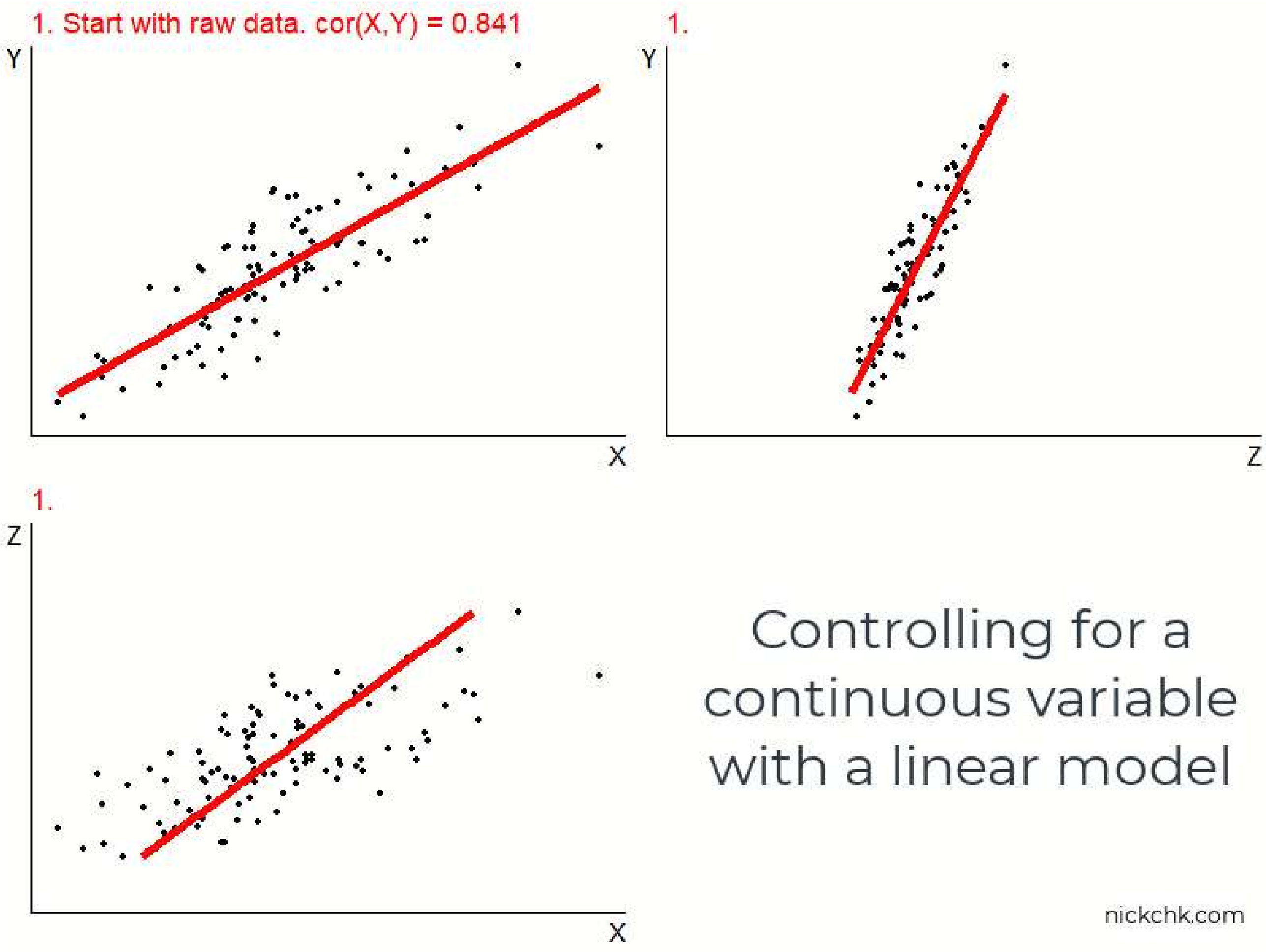
Graphically



Controlling

- We achieve all this just by adding the variable to the OLS equation!
- We can, of course, include more than one control, or controls that aren't binary
- Use OLS to predict X using all the controls, then take the residual (the part not explained by the controls)
- Use OLS to predict Y using all the controls, then take the residual (the part not explained by the controls)
- Now do OLS of just the Y residuals on just the X residuals

A Continuous Control



Controlling for a continuous variable with a linear model

nickchk.com

What do we get?

- We can remove some of the relationship between X and ε
- Potentially all of it, making $\hat{\beta}_1$ us an *unbiased* (i.e. correct on average, but sampling variation doesn't go away!) estimate of β_1
- Maybe we can also get some estimates of β_2, β_3 ... but be careful, they're subject to the same identification and endogeneity problems!
- Often in econometrics we focus on getting *one* parameter, $\hat{\beta}_1$, exactly right and don't focus on parameters we haven't put much effort into identifying

Concept Checks

- Selene is a huge bore at parties, but sometimes brings her girlfriend Donna who is super fun. If you regressed $PartyFunRating$ on $SeleneWasThere$ but not $DonnaWasThere$, what would the coefficient on $SeleneWasThere$ look like and why?
- Describe the steps necessary to estimate the effect of $Exports$ on $GrowthRate$ while controlling for $AmountofConflict$ (a continuous variable). There are three “explain/regress” steps and two “subtract” steps.
- If we estimate the same $\hat{\beta}_1$ with or without Z added as a control, does that mean we have no endogeneity problem? What *does* it mean exactly?

Have We Solved It?

- Including controls for every part of (what used to be) ε that is related to X clears up any endogeneity problem we had with X
- So… when we add a control, does that do it? How do we know?
- Inconveniently, the data alone will never tell us if we've solved endogeniety
- We can't just check X against the remaining ε because we never *see* ε - what we have left over after a regression is the real-world *residual*, not the true-model *error*

Causal Diagrams

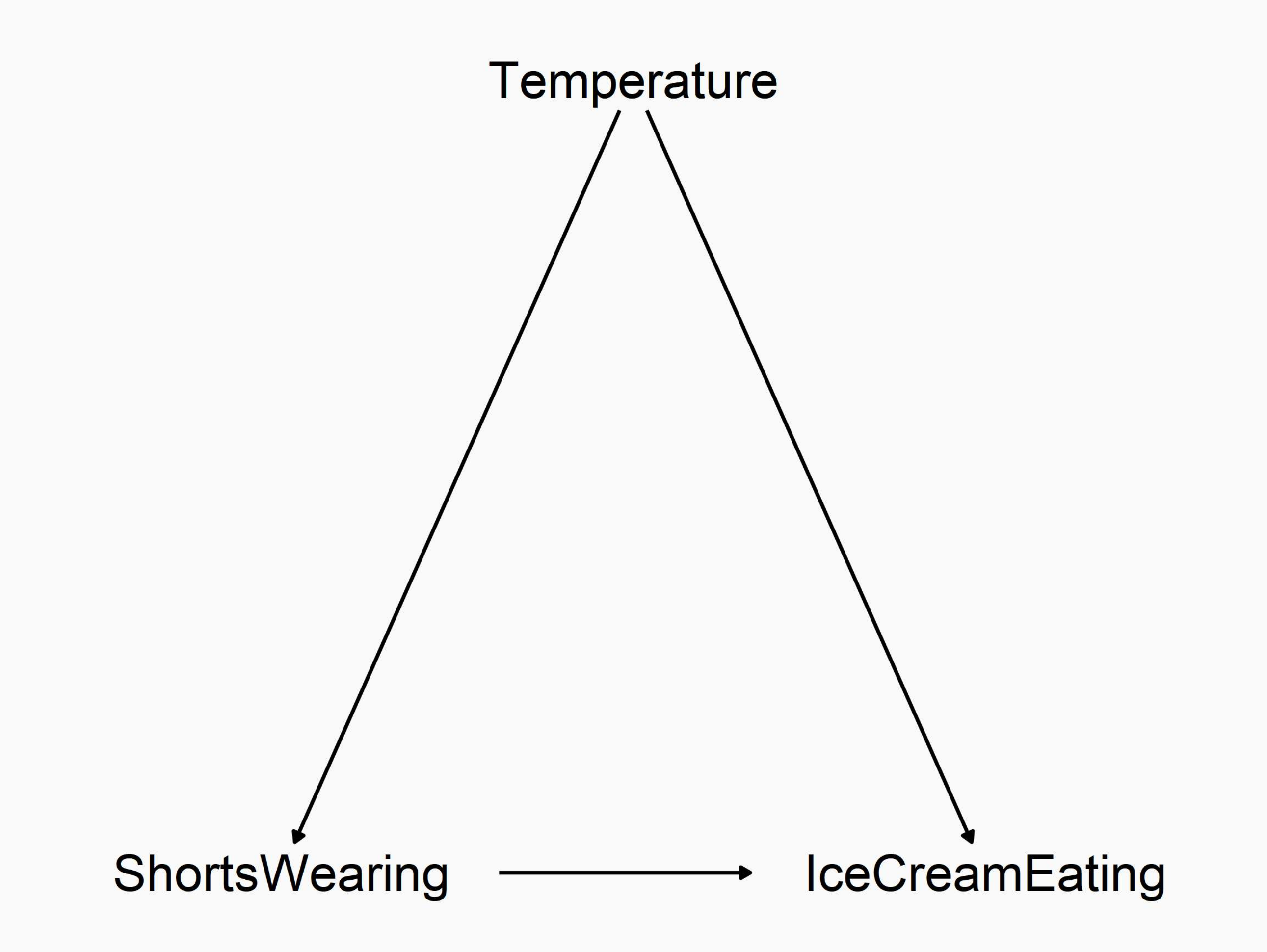
- "What do I have to control for to solve the endogeneity problem" is an important and difficult question!
- To answer it we need to think about the data-generating process
- One way to do that is to draw a *causal diagram*
- A causal diagram describes the variables responsible for generating data and how they cause each other
- Once we have written down our diagram, we'll know what we need to control for
- (hopefully we have data on everything we need to control for! Often we don't)

Drawing a Diagram

- Endogeneity is all about the *alternate reasons why* two variables might be related *other than the causal effect you want
- We can represent *all* the reasons two variables are related with a diagram
- Put down on paper how you think the world works, and where you think the data came from! This is economic modeling but with less math

1. List out all the variables relevant to the DGP (including the ones we can't measure or put our finger on!)
2. Draw arrows between them reflecting what causes what else
3. List all the paths from X to Y - these paths are reasons why X and Y are related!
4. Control for at least one variable on each path you *want to close* (isn't the effect you want)

Drawing a Diagram



Drawing a Diagram

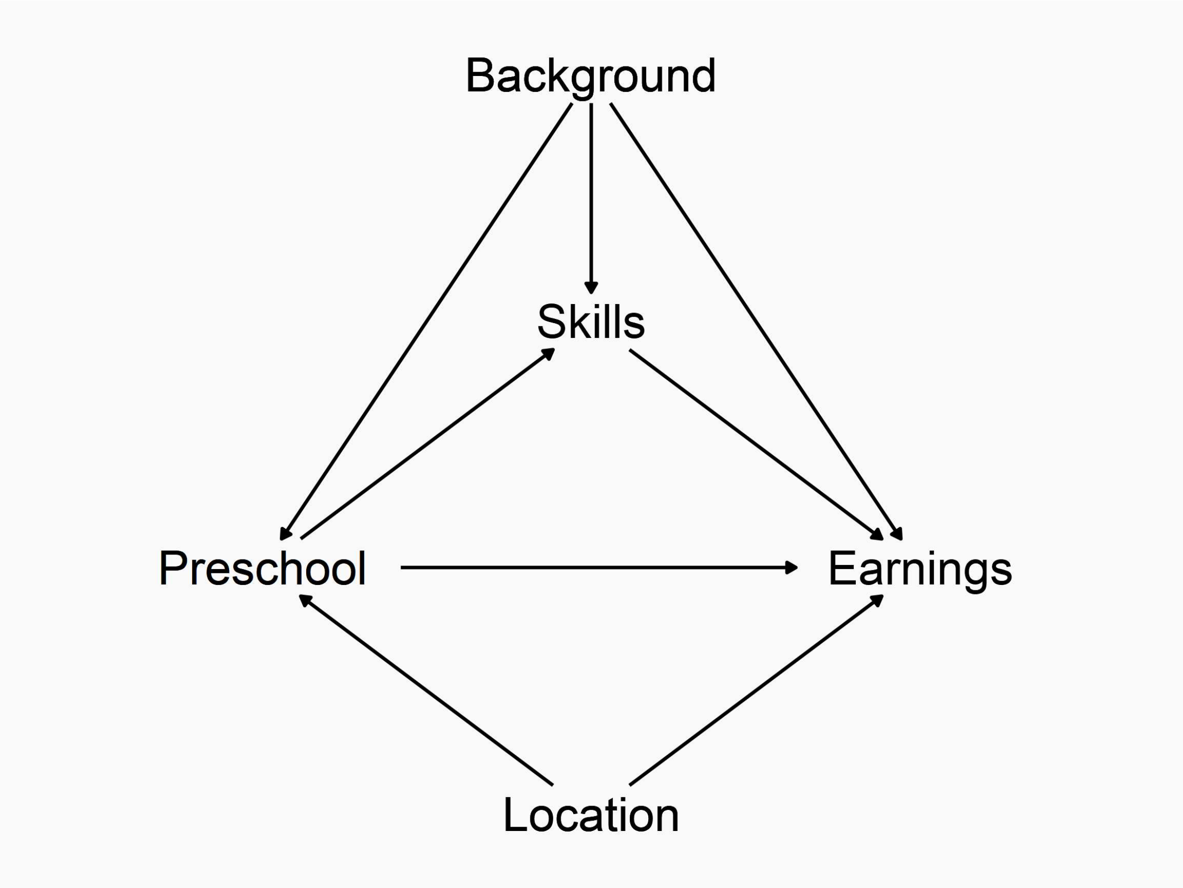
- We observe that, in the data, $ShortsWearing$ and $IceCreamEating$ are related. Why?
- Maybe, we theorize, that wearing shorts causes you to eat ice cream ($ShortsWearing \rightarrow IceCreamEating$)
- However, there's another explanation/path: $Temperature$ causes both ($ShortsWearing \leftarrow Temperature \rightarrow IceCreamEating$)
- We need to control for temperature to *close this path!*
- Once it's closed, the only path left is $ShortsWearing \rightarrow IceCreamEating$, so if we *do* see a relationship still in the data, we know we've identified the causal effect

Detailing Paths

- The goal is to list all the paths that go from the *cause* of our choice to the *outcome* variable (no loops)
- That way we know what we need to control for to close the paths!
- Control for any one variable on the path, and suddenly there's no variation from that variable any more - the causal chain is broken and the path is closed!
- A path counts no matter which direction the arrows point on it (the arrow direction matters but we'll get to that next time)
- If the path isn't part of what answers our research question, it's a *back door* we want to be closed

Preschool and Adult Earnings

Does going to preschool improve your earnings as an adult?



Paths

1. $Preschool \rightarrow Earnings$
2. $Preschool \rightarrow Skills \rightarrow Earnings$
3. $Preschool \leftarrow Location \rightarrow Earnings$
4. $Preschool \leftarrow Background \rightarrow Earnings$
5. $Preschool \leftarrow Background \rightarrow Skills \rightarrow Earnings$
6. $Preschool \rightarrow Skills \leftarrow Background \rightarrow Earnings$

Closing Paths

- We want the ways that $Preschool$ causes $Earnings$ - that's the first two, $Preschool \rightarrow Earnings$ and $Preschool \rightarrow Skills \rightarrow Earnings$
- The rest we want to close! They're *back doors*
- $Location$ is on $\#3$, so if we control for $Location$, 3 is closed
- $Background$ is on the rest, so if we control for $Background$, the rest are closed
- So if we estimate the below OLS equation, $\hat{\beta}_1$ will be unbiased!

$$Earnings = \beta_0 + \beta_1 Preschool + \beta_2 Location + \beta_3 Background + \varepsilon$$

And the Bad News…

- This assumes that *the model we drew was accurate*. Did we leave any important variables or arrows out? Think hard!
- What other variables might belong on this graph? Would they be on a path that gives an alternate explanation?
- Just because we *say* that's the model doesn't magically make it the *actual model!* It needs to be right! Use that economic theory and common sense to think about missing parts of the graph
- Also, *can* we control for those things? What would it mean to assign a single number for $Background$ to someone? Or if we're representing $Background$ with multiple variables - race, gender, parental income, etc., how do we know if we've fully covered it?

And the Bad News…

- Regardless, this is the kind of thinking we have to do to figure out how to identify things by controlling for variables
- There's no way to get around having to make these sorts of assumptions if we want to identify a causal effect
- Really! No way at all! Even experiments have assumptions
- The key is not avoiding assumptions, but making sure they're reasonable, and verifying those assumptions where you can

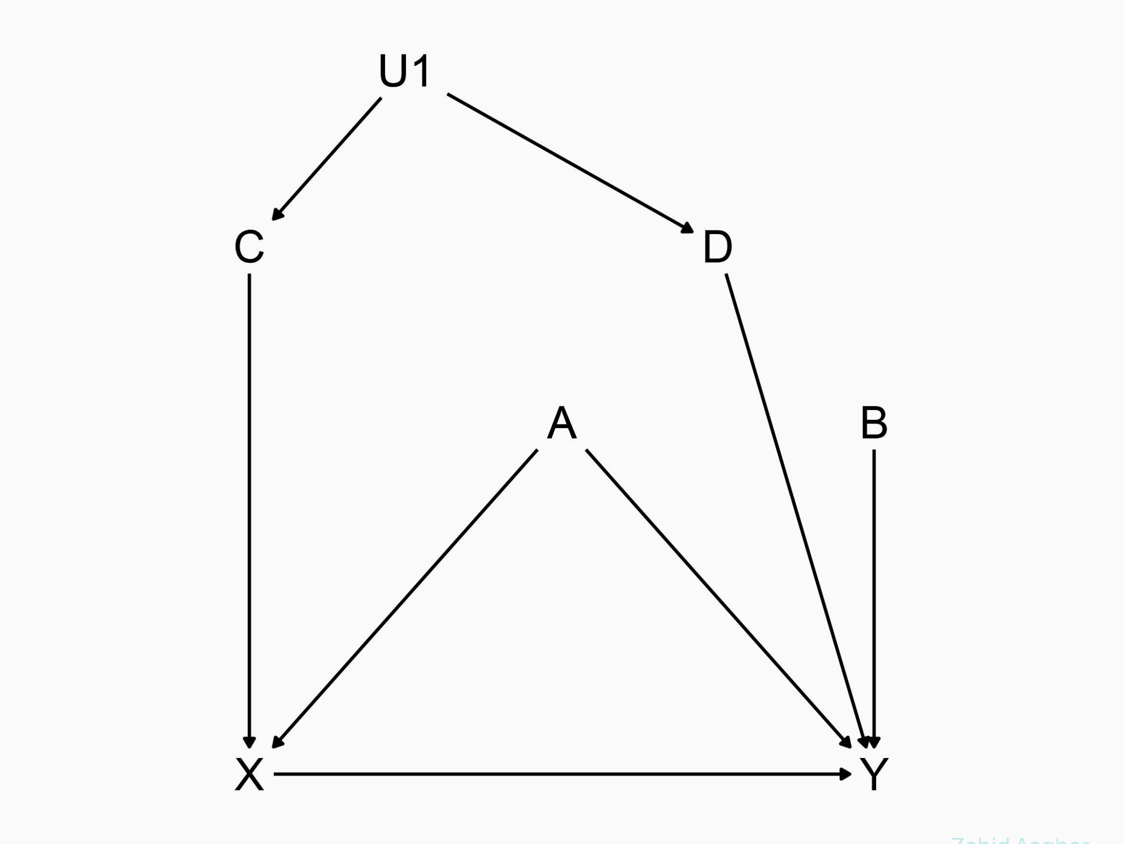
An Example

- Let's back off of those concerns a moment and generate the data ourselves so we know the truth!
- In the below data generating process, what is the true effect of X on Y?
- Let's figure out how to draw the causal diagram for this data generating process!
- (note: U1, U2, etc., often stand in as an unobserved common cause for two variables that are *correlated* but we think neither causes the other)

```r
tib2 <- data.frame(U1 = rnorm(1000), A = rnorm(1000), B = rnorm(1000)) %>%
  mutate(C = U1 + rnorm(1000), D = U1 + rnorm(1000)) %>%
  mutate(X = A + C + rnorm(1000)) %>%
  mutate(Y = 4*X + A + B + D + rnorm(1000))
m1 <- feols(Y~X, data = tib)
coef(m1)
## (Intercept)          X
##        2.76       3.42
```

The Diagram

- Here's the diagram we can draw from that information. What paths are there from X to Y?



The Paths

1. $X \to Y$
2. $X \leftarrow A \to Y$
3. $X \leftarrow C \leftarrow U_1 \to D \to Y$

What do we *need* to control for to close all the paths we don't want? Assume we can't observe (and so can't control for) U_1

The Adjusted Analysis

- Remember, the true β_1 was 4

```
##                               m1                 m2                 m3
## Dependent Var.:                Y                  Y                  Y
##
## (Intercept)     2.756*** (0.0461)  -0.0261 (0.0584)   0.0218 (0.0455)
## X               3.421*** (0.0383)  4.004*** (0.0600)  4.003*** (0.0290)
## A                                  1.044*** (0.0846) 0.9848*** (0.0549)
## C                                 0.4827*** (0.0739)
## D                                                    0.9827*** (0.0368)
## _______________ __________________ __________________ __________________
## S.E. type                      IID                IID                IID
## Observations                 1,000              1,000              1,000
## R2                         0.88897            0.96089            0.97624
## Adj. R2                    0.88886            0.96077            0.97617
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Concept Checks

- Why did we only need to control for $C\ or\ D$ in that last example?
- Draw a graph with five variables on it: X, Y, A, B, C. Then draw arrows at them completely at random (except to ensure there's no "loop" where you can follow an arrow path from arrow base to head and end up where you started). Then list every path from X to Y and say what you'd need to control for to identify the effect
- What would you need to control for to estimate the effect of "drinking a glass of wine a day" on "lifespan"? Draw a diagram.