Big Data and Economics

Causal Inference

Kyle Coombs
Bates College | ECON/DCS 368

Table of contents

Prologue

Prologue

- This week we are starting to think about causal inference
- Today, we're going to explore endogeneity a little bit
- We'll talk about how to solve it using *controls*
- As a warning: this approach is rarely the best approach to causal inference
- But it is a helpful starting point

Questions?

- Ask questions about course content, problem sets, etc.
- I am trying to build this step into future lectures

Endogeneity and omitted variable bias

Endogeneity vs. Exogeneity

- Last time I introduced **exogeneity** as a property of a variable in a model
- I suggested a new saying: Correlation plus **exogeneity** is causation.
- **Endogeneity** is the opposite of **exogeneity**
- 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.

Control variables

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 *multivariable* 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* $\boldsymbol{X}$ *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 $\boldsymbol{X}$

- If $\boldsymbol{cov(X, \varepsilon) = 0}$ 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

Example: coin value

- Let's say we have several piles of coins from a collector with different amounts of quarters and dimes
- The piles are labeled with amounts of money and the amounts of coins, but we don't know the value of the coins
- We could use regression to find out
- One thing we do know is that the collector always had at least as many dimes as quarters
- My friend Szymon Sacher suggested this example (he'll be presenting here in two weeks!)

Example: coin value

```
coins <- tibble(quarters = sample(0:10,1000,replace=TRUE),
            pennies=sample(0:10, 1000, replace=TRUE),
            nickels=sample(0:10, 1000, replace= TRUE)) %>%
  mutate(dimes = quarters + sample(0:10,1000,replace=TRUE)) %>%
  mutate(amount = 0.25*quarters + 0.10*dimes + 0.01*pennies + 0.05*nickels)

coins
```

```
## # A tibble: 1,000 × 5
##    quarters pennies nickels dimes amount
##       <int>   <int>   <int> <int>  <dbl>
##  1       10       5       2    13   3.95
##  2        1       7       0     4   0.72
##  3        6       3       4     7   2.43
##  4        1       6      10     4   1.21
##  5        8       5       0    16   3.65
##  6        0      10       1     6   0.75
##  7        3       4       1     4   1.24
##  8        6       9       1    13   2.94
##  9        4       5       9     9   2.4
## 10        4      10       6     9   2.3
## # ℹ 990 more rows
```

Straight-forward regression

If we just regress, we immediately see the values

```
allcoins <- feols(amount~ quarters + dimes + nickels+pennies, data = coins)
etable(allcoins,fitstat=~n,digits=2,se.below=TRUE)
```

```
##                      allcoins
## Dependent Var.:        amount
##
## Constant          -1.7e-14***
##                   (1.9e-15)
## quarters           0.25***
##                   (2.5e-16)
## dimes              0.10***
##                   (1.8e-16)
## nickels            0.05***
##                   (1.7e-16)
## pennies            0.01***
##                   (1.7e-16)
## _______________ ___________
## S.E. type                IID
## Observations           1,000
## ---
```

What if we remove quarters?

The coefficient on dimes changes a lot! Why?

```
noquarters <- feols(amount~ dimes + nickels+pennies, data = coins)
etable(allcoins,noquarters,fitstat=~n,digits=2,se.below=TRUE)
```

```
##                       allcoins noquar..
## Dependent Var.:         amount   amount
##
## Constant           -1.7e-14*** -0.01
##                    (1.9e-15)   (0.06)
## quarters            0.25***
##                    (2.5e-16)
## dimes               0.10***      0.23***
##                    (1.8e-16)   (0.004)
## nickels             0.05***      0.05***
##                    (1.7e-16)   (0.005)
## pennies             0.01***      0.01*
##                    (1.7e-16)   (0.005)
## _______________ ___________ ________
## S.E. type                IID      IID
## Observations           1,000    1,000
## ---
```

Endogeneity of quarters

- The number of dimes was a function of quarters
- When we dropped quarters, we omitted a variable that was related to dimes and the amount of money
- So the coefficient on dimes was biased

Concept check

- What happens if I drop nickels and pennies as well?

```
onlydimes <- feols(amount~ dimes, data = coins)
etable(allcoins,noquarters,onlydimes,fitstat=~n,digits=2,se.below=TRUE)
```

```
##                     allcoins noquar.. onlydi..
## Dependent Var.:       amount   amount   amount
##
## Constant         -1.7e-14*** -0.01     0.31***
##                  (1.9e-15)   (0.06)   (0.04)
## quarters          0.25***
##                  (2.5e-16)
## dimes             0.10***     0.23***  0.22***
##                  (1.8e-16)   (0.004)  (0.004)
## nickels           0.05***     0.05***
##                  (1.7e-16)   (0.005)
## pennies           0.01***     0.01*
##                  (1.7e-16)   (0.005)
## _______________ ___________ ________ ________
## S.E. type                IID      IID      IID
## Observations           1,000    1,000    1,000
## ---
```

See it directly

Let's drop nickels and pennies cause they're small, unrelated to quarters and dimes by construction, and we're just trying to illustrate a point

Let's subtract out the part of dimes that is related to quarters (on average)

```
coins <- coins %>%
  group_by(quarters) %>%
  mutate(amount_mean = mean(amount), dimes_mean = mean(dimes))
head(coins)
```

```
## # A tibble: 6 × 7
## # Groups:   quarters [5]
##   quarters pennies nickels dimes amount amount_mean dimes_mean
##      <int>   <int>   <int> <int>  <dbl>       <dbl>      <dbl>
## 1       10       5       2    13   3.95       4.28       14.9
## 2        1       7       0     4   0.72       1.13        5.56
## 3        6       3       4     7   2.43       2.91       11.0
## 4        1       6      10     4   1.21       1.13        5.56
## 5        8       5       0    16   3.65       3.63       13.4
## 6        0      10       1     6   0.75       0.806       5.28
```

Example: Residualize

Now, `amount_mean` and `dimes_mean` are the respective means for each amount of quarters, i.e. the part of `amount` and `dimes` *explained by* *`quarters`*. So subtract those parts out to get *residuals* `amount_res` and `dime_res`!

```
coins <- coins %>%
  mutate(amount_res = amount - amount_mean, dimes_res = dimes - dimes_mean)
head(coins)
```

```
## # A tibble: 6 × 9
## # Groups:   quarters [5]
##   quarters pennies nickels dimes amount amount_mean dimes_mean amount_res
##      <int>   <int>   <int> <int>  <dbl>       <dbl>      <dbl>      <dbl>
## 1       10       5       2    13   3.95       4.28       14.9     -0.334
## 2        1       7       0     4   0.72       1.13        5.56    -0.415
## 3        6       3       4     7   2.43       2.91       11.0     -0.479
## 4        1       6      10     4   1.21       1.13        5.56     0.0752
## 5        8       5       0    16   3.65       3.63       13.4      0.0176
## 6        0      10       1     6   0.75       0.806       5.28    -0.0565
## # ℹ 1 more variable: dimes_res <dbl>
```

Example: Regression residuals

What do we get now?

```
residuals <- feols(amount_res ~ dimes_res, data = coins)
etable(allcoins, noquarters,onlydimes,residuals,dict=c('dimes_res'='dimes'),fits
```

```
##                     allcoins noquar.. onlydi..  residuals
## Dependent Var.:       amount   amount   amount amount_res
##
## Constant         -1.7e-14*** -0.01     0.31***   -3.5e-17
##                  (1.9e-15)   (0.06)   (0.04)     (0.005)
## quarters           0.25***
##                  (2.5e-16)
## dimes              0.10***    0.23***  0.22***    0.10***
##                  (1.8e-16)   (0.004)  (0.004)    (0.002)
## nickels            0.05***    0.05***
##                  (1.7e-16)   (0.005)
## pennies            0.01***    0.01*
##                  (1.7e-16)   (0.005)
## _______________ ___________ ________ ________ __________
## S.E. type               IID      IID      IID        IID
## Observations          1,000    1,000    1,000      1,000
## ---
```

Example

Let's quickly check what happens when we exclude nickels and pennies

```
quartersdimes <- feols(amount ~ quarters + dimes, data = coins)
etable(allcoins,noquarters,onlydimes,residuals,quartersdimes,dict=c('dimes_res'=
```

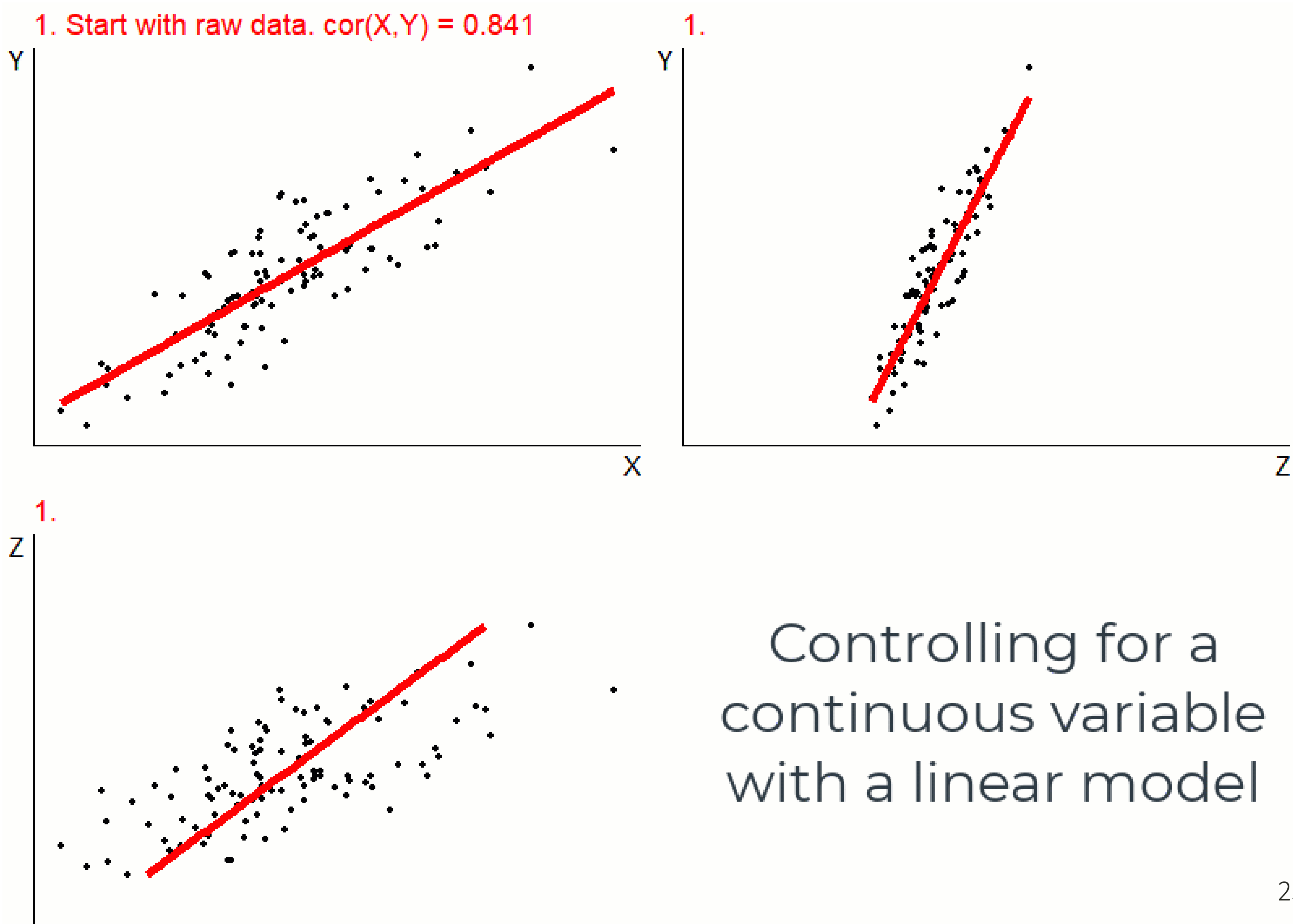
```
##                    allcoins noquar.. onlydi..  residuals quarte..
## Dependent Var.:      amount   amount   amount amount_res   amount
##
## Constant        -1.7e-14*** -0.01     0.31***   -3.5e-17  0.32***
##                 (1.9e-15)   (0.06)   (0.04)     (0.005)  (0.01)
## quarters          0.25***                                 0.25***
##                 (2.5e-16)                                (0.002)
## dimes             0.10***     0.23***  0.22***    0.10***  0.10***
##                 (1.8e-16)   (0.004)  (0.004)    (0.002)  (0.002)
## nickels           0.05***     0.05***
##                 (1.7e-16)   (0.005)
## pennies           0.01***     0.01*
##                 (1.7e-16)   (0.005)
## _______________ ___________ ________ ________ __________ ________
## S.E. type               IID      IID      IID        IID      IID
## Observations          1,000    1,000    1,000      1,000    1,000
## ---
```

Graphically with binary control Z

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



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

What if pennies were omitted?

- What if pennies were the omitted variable?

```
##                                    ..1              ..2
##                          penniesdimes       onlydimes
## Dependent Var.:                amount          amount
##
## Constant             1.4*** (0.06)   1.4*** (0.06)
## dimes              0.11*** (0.008) 0.11*** (0.005)
## pennies              0.007 (0.01)
## _______________ _______________ _______________
## S.E. type                     IID             IID
## Observations                1,000           1,000
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Huh? The bias is so much smaller!
- This is because the quarters explain a larger share of the variation in amount than the pennies do
- So when we remove quarters, dimes inherit a larger relationship with

Tough Concept Checks

- Describe the steps necessary to estimate the effect of $Education$ on $Income$ while controlling for $Segregation$ (a continuous variable). There are three "explain/regress" steps and two "subtract" steps.

- Selene is a huge bore at parties, but sometimes brings her girlfriend Donna who is super fun. If you regress $PartyFunRating$ on $SeleneWasThere$ but not $DonnaWasThere$, what would the coefficient on $SeleneWasThere$ look like and why?

- If we estimate the same $\hat{\beta}_1$ with or without some Z added as a control, does that mean we have no endogeneity problem? What *does* it mean exactly?

Go to: Menti.com: 9111 2607

Summary

- We can remove endogeneity by adding omitted variables into our regression model if:

  1. We know/correctly assume what they are
  2. We can measure them

- This works by removing the part X and Y that is related to the omitted variable, Z

- This is fairly common, but often inadequate approach to causal inference

- Sometimes it is the best we can do though!

Next week: Fixed Effects/Diff-in-diff