FIN 620
Emp. Methods in Finance

Lecture 3 – Causality

Professor Todd Gormley

Background readings for today

- Roberts-Whited
  - *Section 2*
- Angrist and Pischke
  - *Section 3.2*
- Wooldridge
  - *Sections 4.3 & 4.4*
- Greene
  - *Sections 5.8-5.9*

Outline for Today

- Quick review
- Motivate why we care about causality
- Describe three possible biases & some potential solutions
  - Omitted variable bias
  - Measurement error bias
  - Simultaneity bias
- Student presentations of "Classics #2"

Quick Review *[Part 1]*

- Why is adding irrelevant regressors a potential problem?

- Why is a larger sample helpful?

Quick Review *[Part 2]*

- Suppose, $\beta_1 < 0$ and $\beta_3 > 0$ … what is the sign of the effect of an increase in x_1 for the <u>average</u> firm in the below estimation?

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u$$

  - **Answer:** It is the sign of

$$\frac{dy}{dx_1}\Big|^{x_2=\bar{x}_2} = \beta_1 + \beta_3 \bar{x}_2$$

Quick Review *[Part 3]*

- How could we make the coefficients easier to interpret in the prior example?
  - Shift all the variables by subtracting out their sample mean before doing the estimation
  - It will allow the non-interacted coefficients to be interpreted as effect for average firm

Quick Review *[Part 4]*

- Consider the following estimate:

$$\ln(wage) = 0.32 - 0.11 female + 0.21 married$$
$$-0.30(female \times married) + 0.08 education$$

  - **Question:** How much lower are wages of married and unmarried females after controlling for education, and who is this relative to?

Outline for Today

- Quick review
- Motivate why we care about causality
- Describe three possible biases & some potential solutions
  - Omitted variable bias
  - Measurement error bias
  - Simultaneity bias
- Student presentations of "Classics #2"

Motivation

- As researchers, we are interested in making causal statements
  - Ex. #1 – what is the *effect* of a change in corporate taxes on firms' leverage choice?
  - Ex. #2 – what is the *effect* of giving a CEO more stock ownership in the firm on the CEO's desire to take on risky investments?
- I.e., we don't like to just say variables are 'associated' or 'correlated' with each other

What do we mean by causality?

- Recall from earlier lecture, that if our linear model is the following…

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

And we want to infer β_1 as the causal effect of x_1 on y, holding all else equal, then we need to make the following assumptions…

The basic assumptions

- *Assumption #1:* $E(u) = 0$
- *Assumption #2:* $E(u \mid x_1, \ldots, x_k) = E(u)$
  - In words, average of u (i.e., unexplained portion of y) does not depend on value of x
  - This is "conditional mean independence" (CMI)
- Generally speaking, you need the estimation error to be uncorrelated with all the x's

Tangent – CMI versus correlation

- CMI (which implies x and u are uncorrelated) is needed for unbiasedness *[which is again a finite sample property]*
- However, we only need to assume a zero correlation between x and u for consistency *[which is a large sample property]*
  - **This is why I will typically just refer to whether u and x are correlated in my test of whether we can make causal inferences**

Three main ways this will be violated

- Omitted variable bias
- Measurement error bias
- Simultaneity bias

Now, let's go through each in turn…

Omitted variable bias (OVB)

- Probably the most common concern you will hear researchers worry about
- **Basic idea** = the estimation error, u, contains another variable, e.g., z, that affects y **<u>and</u>** is correlated with an x
  - **Please note!** The omitted variable is only problematic if correlated with an x

OVB more formally, with one variable

- You estimate: $y = \beta_0 + \beta_1 x + u$
- But true model is: $y = \beta_0 + \beta_1 x + \beta_2 z + v$
- Then, $\hat{\beta}_1 = \beta_1 + \delta_{xz}\beta_2$, where δ_{xz} is the coefficient you'd get from regressing the omitted variable, *z*, on *x*; and

$$\delta_{xz} = \frac{\operatorname{cov}(x,z)}{\operatorname{var}(x)}$$

Interpreting the OVB formula

Bias

$$\hat{\beta}_1 = \beta_1 + \boxed{\frac{\operatorname{cov}(x,z)}{\operatorname{var}(x)}\beta_2}$$

Effect of *x* on *y*

Regression of *z* on *x*

Effect of *z* on *y*

- Easy to see, estimated coefficient is only unbiased if $cov(x, z) = 0$ [i.e., x and z are uncorrelated] **or** z has no effect on y [i.e., $\beta_2 = 0$]

Direction and magnitude of the bias

$$\hat{\beta}_1 = \beta_1 + \frac{\text{cov}(x,z)}{\text{var}(x)}\beta_2$$

- Direction of bias given by signs of β_2, $cov(x, z)$
  - E.g., If know z has positive effect on y [i.e., $\beta_2 > 0$] and x and z are positively correlated [$cov(x, z) > 0$], then the bias will be positive
- Magnitude of the bias will be given by magnitudes of β_2, $cov(x, z)/var(x)$

Example – One variable case

- Suppose we estimate: $\ln(wage) = \beta_0 + \beta_1 educ + w$
- But true model is:

$$\ln(wage) = \beta_0 + \beta_1 educ + \beta_2 ability + u$$

- What is likely bias on $\hat{\beta}_1$? Recall,

$$\hat{\beta}_1 = \beta_1 + \frac{\text{cov}(educ, ability)}{\text{var}(educ)} \beta_2$$

Example – **Answer**

- Ability & wages likely positively correlated, so $\beta_2 > 0$
- Ability & education likely positive correlated, so cov(*educ*, *ability*) > 0
- Thus, the bias is likely to positive! $\hat{\beta}_1$ is too big!

OVB – General Form

- Once move away from simple case of just one omitted variable, determining sign (and magnitude) of bias will be a <u>lot</u> harder
  - Let $\boldsymbol{\beta}$ be vector of coefficients on k included variables
  - Let γ be vector of coefficient on l excluded variables
  - Let $\mathbf{X}$ be matrix of observations of included variables
  - Let $\mathbf{Z}$ be matrix of observations of excluded variables

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \frac{E[\mathbf{X'Z}]}{E[\mathbf{X'X}]}\boldsymbol{\gamma}$$

OVB – General Form, Intuition

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \frac{E[\mathbf{X'Z}]}{E[\mathbf{X'X}]}\boldsymbol{\gamma}$$

Vector of regression coefficients

Vector of partial effects of excluded variables

- Same idea as before, but more complicated
- Frankly, this can be a real mess!
  [See Gormley and Matsa (2014) for example with just two included and two excluded variables]

Eliminating Omitted Variable Bias

- How we try to get rid of this bias will depend on the type of omitted variable
  - **Observable** omitted variable
  - **Unobservable** omitted variable

How can we deal with an observable omitted variable?

Observable omitted variables

- This is easy! Just add them as controls
  - E.g., if the omitted variable, z, in my simple case was 'leverage,' then add leverage to regression
- A functional form misspecification is a special case of an observable omitted variable

Let's now talk about this…

Functional form misspecification

- Assume true model is…

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + u$$

- However, we omit squared term, x_2^2
  - Just like any OVB, bias on $(\beta_0, \beta_1, \beta_2)$ will depend on β_3 and correlations among (x_1, x_2, x_2^2)
  - You get same type of problem if have incorrect functional form for y *[e.g., it should be ln(y) not y]*
- In some sense, this is minor problem… Why?

Tests for correction functional form

- You could add additional squared and cubed terms and look to see whether they make a difference and/or have non-zero coefficients
- This isn't as easy when the possible models are not nested…

Non-nested functional form issues…

- Two non-nested examples are:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

versus

$$y = \beta_0 + \beta_1 \ln(x_1) + \beta_2 \ln(x_2) + u$$

Let's use this example and see how we can try to figure out which is right

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

versus

$$y = \beta_0 + \beta_1 x_1 + \beta_2 z + u$$

Davidson-MacKinnon Test *[Part 1]*

- To test which is correct, you can try this…
  - Take fitted values, $\hat{y}$, from 1st model and add them as a control in 2nd model

$$y = \beta_0 + \beta_1 \ln(x_1) + \beta_2 \ln(x_2) + \theta_1 \hat{y} + u$$

  - Look at *t*-stat on θ_1; if significant rejects 2nd model!
  - Then, do reverse, and look at *t*-stat on θ_1 in

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \theta_1 \hat{\hat{y}} + u$$

    where $\hat{\hat{y}}$ is predicted value from 2nd model… if significant then 1st model is also rejected ☹

Davidson-MacKinnon Test *[Part 2]*

- Number of weaknesses to this test…
  - A clear winner may not emerge
    - Both might be rejected
    - Both might be accepted [If this happens, you can use the R^2 to choose which model is a better fit]
  - And rejecting one model does **NOT** imply that the other model is correct ☹

Bottom line advice on functional form

- Practically speaking, you hope that changes in functional form won't affect coefficients on key variables very much…
  - But, if it does… You need to think hard about why this is and what the correct form should be
  - The prior test might help with that…

Eliminating Omitted Variable Bias

- How we try to get rid of this bias will depend on the type of omitted variable
  - **Observable** omitted variable
  - **Unobservable** omitted variable

Unobservable are much harder to deal with, but one possibility is to find a proxy variable

Unobserved omitted variables

- Again, consider earlier estimation

$$\ln(wage) = \beta_0 + \beta_1 educ + \beta_2 ability + u$$

  - **Problem**: we don't observe & can't measure *ability*
  - What can we do? **Ans. =** Find a proxy variable that is correlated with the unobserved variable, E.g., IQ

Proxy variables *[Part 1]*

- Consider the following model…

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u$$

  where x_3^* is unobserved, but we have proxy x_3

- Then, suppose $x_3^* = \delta_0 + \delta_1 x_3 + v$
  - v is error associated with proxy's imperfect representation of unobservable x_3
  - Intercept just accounts for different scales *[e.g., ability has different average value than IQ]*

Proxy variables *[Part 2]*

- **<u>If</u>** we are only interested in β_1 or β_2, we can just replace x_3^* with x_3 and estimate

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

- But, for this to give us consistent estimates of β_1 and β_2 , we need to make some assumptions

  #1 – We've got the right model, and

  #2 – Other variables don't explain our unobserved variable after we've accounted for our proxy

Proxy variables – *Assumptions*

#1 – $E(u \mid x_1, x_2, x_3^*) = 0$; i.e., we have the right model and x_3 would be irrelevant if we could control for x_1, x_2, x_3^*, such that $E(u \mid x_3) = 0$

- This is a common assumption; not controversial

#2 – $E(v \mid x_1, x_2, x_3) = 0$; i.e., x_3 is a good proxy for x_3^* such that after controlling for x_3, x_3^* does not depend on x_1 or x_2

- I.e., $E(x_3^* \mid x_1, x_2, x_3) = E(x_3^* \mid x_3)$

Why the proxy works…

- Recall true model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u$
- Now plug-in for x_3^*, using $x_3^* = \delta_0 + \delta_1 x_3 + v$

$$y = \underbrace{\left(\beta_0 + \beta_3 \delta_0\right)}_{\alpha_0} + \beta_1 x_1 + \beta_2 x_2 + \underbrace{\left(\beta_3 \delta_1\right)}_{\alpha_1} x_3 + \underbrace{\left(u + \beta_3 v\right)}_{e}$$

  - Prior assumptions ensure that $E(e \mid x_1, x_2, x_3) = 0$ such that the estimates of $(\alpha_0, \beta_1, \beta_2, \alpha_1)$ are consistent
  - **Note:** β_0 and β_3 are **<u>not</u>** identified

Proxy assumptions are key *[Part 1]*

- Suppose assumption #2 is wrong such that

$$x_3^* = \delta_0 + \delta_1 x_3 + \underbrace{\gamma_1 x_1 + \gamma_2 x_2 + w}_{v}$$

where $E(w \mid x_1, x_2, x_3) = 0$

  - If above is true, $E(v \mid x_1, x_2, x_3) \neq 0$, and if you substitute into model of *y*, you'd get…

Proxy assumptions are key *[Part 2]*

- Plugging in for x_3^*, you'd get

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + e$$

where
$$\alpha_0 = \beta_0 + \beta_3 \delta_0$$
$$\alpha_1 = \beta_1 + \beta_3 \gamma_1$$
$$\alpha_2 = \beta_2 + \beta_3 \gamma_2$$
$$\alpha_3 = \beta_3 \delta_1$$

E.g., α_1 captures effect of x_1 on y, β_1, but also its correlation with unobserved variable

- We'd get consistent estimates of $(\alpha_0, \alpha_1, \alpha_2, \alpha_3)$ But that isn't what we want!

Proxy variables – *Example #1*

- Consider earlier wage estimation

$$\ln(wage) = \beta_0 + \beta_1 educ + \beta_2 ability + u$$

  - If we use IQ as proxy for unobserved *ability*, what assumption must we make? Is it plausible?
    - **Answer:** We assume $E(ability \mid educ, IQ) = E(ability \mid IQ)$, i.e., average ability does not change with education after accounting for IQ… Could be questionable assumption!

Proxy variables – *Example #2*

- Consider *Q*-theory of investment

$$investment = \beta_0 + \beta_1 Q + u$$

  - Can we estimate β_1 using a firm's market-to-book ratio (MTB) as proxy for Q? Why or why not?
    - **Answer:** Even if we believe this is the correct model (Assumption #1) or that Q only depends on MTB (Assumption #2), e.g., $Q=\delta_0+\delta_1 MTB$, we are still not getting estimate of β_1… see next slide for the math

Proxy variables – *Example #2 [Part 2]*

- Even if assumptions held, we'd only be getting consistent estimates of

$$investment = \alpha_0 + \alpha_1 MTB + e$$

where $\alpha_0 = \beta_0 + \beta_1\delta_0$

$\alpha_1 = \beta_1\delta_1$

  - While we can't get β_1, is there something we can get if we make assumptions about sign of δ_1?
  - **Answer:** Yes, the sign of β_1

Proxy variables – **Summary**

- If the coefficient on the unobserved variable isn't what we are interested in, then a proxy for it can be used to identify and remove OVB from the other parameters
  - Proxy can also be used to determine sign of coefficient on an unobserved variable

Random Coefficient Model

- So far, we've assumed that the effect of x on y (i.e., β) was the same for all observations
  - In reality, this is unlikely true; model might look more like $y_i = \alpha_i + \beta_i x_i + u_i$, where

$$\alpha_i = \alpha + c_i$$
$$\beta_i = \beta + d_i$$
$$E(c_i) = E(d_i) = 0$$

I.e., each observation's relationship between x and y is slightly different

  - α is the average intercept and β is what we call the "average partial effect" (APE)

Random Coefficient Model *[Part 2]*

- Regression would seem to be incorrectly specified, but if willing to make assumptions, we can identify the APE
  - Plug in for α_i and β_i

$$y_i = \alpha + \beta x_i + \left(c_i + d_i x_i + u_i\right)$$

If like, can think of the unobserved differential intercept and slopes as omitted variable

  - Identification requires

$$E\left(c_i + d_i x_i + u_i \mid x\right) = 0$$

What does this imply?

Random Coefficient Model *[Part 3]*

- This amounts to requiring

$$E(c_i \mid x) = E(c_i) = 0 \Rightarrow E(\alpha_i \mid x) = E(\alpha_i)$$
$$E(d_i \mid x) = E(d_i) = 0 \Rightarrow E(\beta_i \mid x) = E(\beta_i)$$

  - We must assume that the individual slopes and intercepts are mean independent (i.e., uncorrelated with the value of *x*) in order to estimate the APE
    - **I.e., knowing *x*, does not help us predict the individual's partial effect**

Random Coefficient Model *[Part 4]*

- Implications of APE
  - Be careful interpreting coefficients when you are implicitly arguing elsewhere in paper that effect of x varies across observations
    - **Keep in mind the assumption this requires**
    - **And describe results using something like… "we find that, *on average*, an increase in x causes a β change in y"**

Three main ways this will be violated

- Omitted variable bias
- Measurement error bias
- Simultaneity bias

Measurement error (ME) bias

- Estimation will have measurement error whenever we measure the variable of interest imprecisely
  - Ex. #1: Altman-z-score is noisy measure of default risk
  - Ex. #2: Avg. tax rate is noisy measure of marg. tax rate
- **Such measurement error can cause bias, and the bias can be quite complicated**

Measurement error *vs.* proxies

- Measurement error is like a proxy variable, but very different conceptually
  - Proxy is used for something that is entirely unobservable or unmeasureable (e.g., ability)
  - With measurement error, the variable we don't observe is well-defined and can be quantified… it's just that our measure of it contains error

ME of Dep. Variable *[Part 1]*

- Usually not a problem (in terms of bias); just causes our standard errors to be larger. E.g.,…
  - Let $y^* = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$
  - But we measure y^* with error $e = y - y^*$
  - Because we only observe y, we estimate

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + (u + e)$$

Note: we always assume E(e)=0; this is innocuous because if untrue, it only affects the bias on the constant

ME of Dep. Variable *[Part 2]*

- As long as E(e|x)=0, the OLS estimates are consistent and unbiased
  - I.e., as long as the measurement error of y is uncorrelated with the x's, we're okay
  - Only issue is that we get larger standard errors when e and u are uncorrelated *[which is what we typically assume]* because Var(u+e)>Var(u)

What are some common examples of ME?

ME of Dep. Variable *[Part 3]*

- Some common examples
  - **Market leverage** – typically use book value of debt because market value hard to observe
  - **Firm value** – again, hard to observe market value of debt, so we use book value
  - **CEO compensation** – value of options are approximated using Black-Scholes

Is assuming *e* and *x* are uncorrelated plausible?

ME of Dep. Variable *[Part 4]*

- **Answer** = Maybe… maybe not
  - Ex. – Firm leverage is measured with error; hard to observe market value of debt, so we use book value
    - But the measurement error is likely to be larger when firms are in distress… Market value of debt falls; book value does not
    - This error could be correlated with x's if it includes things like profitability (i.e., ME larger for low profit firms)
    - **This type of ME will cause inconsistent estimates**

ME of Independent Variable *[Part 1]*

- Let's assume the model is $y = \beta_0 + \beta_1 x^* + u$
- But we observe x^* with error, $e = x - x^*$
  - We assume that $E(y|x^*, x) = E(y|x^*)$ [i.e., x doesn't affect y after controlling for x^*; this is standard and uncontroversial because it is just stating that we have written the correct model]
- **What are some examples in CF?**

ME of Independent Variable *[Part 2]*

- There are lots of examples!
  - Average Q measures marginal Q with error
  - Altman-*z* score measures default prob. with error

Will this measurement error cause bias?

ME of Independent Variable *[Part 2]*

- Answer depends crucially on what we assume about the measurement error, e
- Literature focuses on two extreme assumptions

#1 – Measurement error, e, is uncorrelated with the observed measure, x

#2 – Measurement error, e, is uncorrelated with the unobserved measure, x^*

Assumption #1: e uncorrelated with x

- Substituting x^* with what we actually observe, $x^* = x - e$, into true model, we have

$$y = \beta_0 + \beta_1 x + u - \beta_1 e$$

  - Is there a bias?
    - **Answer** = No. x is uncorrelated with e by assumption, and x is uncorrelated with u by earlier assumptions
  - What happens to our standard errors?
    - **Answer** = They get larger; error variance is now $\sigma_u^2 + \beta_1^2 \sigma_e^2$

Assumption #2: *e* uncorrelated with *x**

- We are still estimating $y = \beta_0 + \beta_1 x + u - \beta_1 e$, but now, x is correlated with e
  - e uncorrelated with x^* guarantees e is correlated with x; $\text{cov}(x, e) = E(xe) = E(x^* e) + E(e^2) = \sigma_e^2$
  - I.e., an independent variable will be correlated with the error… we will get **biased** estimates!
- This is what people call the **Classical Error-in-Variables (CEV)** assumption

CEV with 1 variable = attenuation bias

- If you work out math, you can show that the estimate of β_1, $\hat{\beta}_1$, in prior example (which had just one independent variable) is…

$$p\lim(\hat{\beta}_1) = \beta_1 \boxed{\left(\frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_e^2} \right)} \longleftarrow$$

This scaling factors is always between 0 and 1

- The estimate is always biased towards zero; i.e., it is an **attenuation bias**
  - And, if variance of error, σ_e^2, is small, then attenuation bias won't be that bad

Measurement error… not so bad?

- Under current setup, measurement error doesn't seem so bad…
  - If error uncorrelated with observed x, no bias
  - If error uncorrelated with unobserved x^*, we get an attenuation bias… so at least the sign on our coefficient of interest is still correct
- Why is this misleading?

Nope, measurement error is bad news

- Truth is, measurement error is probably correlated a bit with both the observed x and unobserved x^*
  - I.e… some attenuation bias is likely
- **Moreover**, even in CEV case, if there is more than one independent variable, the bias gets horribly complicated…

ME with more than one variable

- If estimating $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$, and just one of the *x*'s is mismeasured, then…
  - **ALL** the β's will be biased if the mismeasured variable is correlated with any other *x* *[which presumably is true since it was included!]*
  - Sign and magnitude of biases will depend on all the correlations between *x*'s; **i.e., big mess!**
    - See Gormley and Matsa (2014) math for *Avg*E estimator to see how bad this can be

ME example

- Fazzari, Hubbard, and Petersen (1988) is classic example of a paper with ME problem
  - Regresses investment on Tobin's Q (it's measure of investment opportunities) and cash
  - Finds positive coefficient on cash; argues there must be financial constraints present
  - But Q is noisy measure; all coefficients are biased!
- Erickson and Whited (2000) argues the pos. coeff. disappears if you correct the ME

Three main ways this will be violated

- Omitted variable bias
- Measurement error bias
- Simultaneity bias

Simultaneity bias

- This will occur whenever any of the supposedly independent variables (i.e., the x's) can be affected by changes in the y variable; E.g.

$$y = \beta_0 + \beta_1 x + u$$

$$x = \delta_0 + \delta_1 y + v$$

  - I.e., changes in x affect y, and changes in y affect x; this is the simplest case of reverse causality
  - An estimate of $y = \beta_0 + \beta_1 x + u$ will be biased…

Simultaneity bias continued…

- To see why estimating $y = \beta_0 + \beta_1 x + u$ won't reveal the true β_1, solve for x

$$x = \delta_0 + \delta_1 y + v$$

$$x = \delta_0 + \delta_1 \left(\beta_0 + \beta_1 x + u \right) + v$$

$$x = \left(\frac{\delta_0 + \delta_1 \beta_0}{1 - \delta_1 \beta_1} \right) + \left(\frac{v}{1 - \delta_1 \beta_1} \right) + \left(\frac{\delta_1}{1 - \delta_1 \beta_1} \right) u$$

  - Easy to see that x is correlated with u! I.e., bias!

Simultaneity bias in other regressors

- Prior example is case of reverse causality; the variable of interest is also affected by y
- But, if y affects any x, their will be a bias; E.g.,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$
$$x_2 = \gamma_0 + \gamma_1 y + w$$

  - Easy to show that x_2 is correlated with u; and there will be a bias on all coefficients
  - This is why people use lagged x's

“Endogeneity” problem – *Tangent*

- In my opinion, the prior example is what it means to have an “endogeneity” problem or and “endogenous” variable
  - But, as I mentioned earlier, there is a lot of misusage of the word “endogeneity” in finance… So, it might be better just saying “simultaneity bias”

Simultaneity Bias – **Summary**

- If your x might also be affected by the y (i.e., reverse causality), you won't be able to make causal inferences using OLS
  - Instrumental variables or natural experiments will be helpful with this problem
- Also, you can't get causal estimates with OLS if controls are affected by the y

"Bad controls"

- Like simultaneity bias… this is when one *x* is affected by another *x*; e.g.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$
$$x_2 = \gamma_0 + \gamma_1 x_1 + v$$

  - Angrist-Pischke call this a "bad control," and it can introduce a subtle selection bias when working with natural experiments
    [we will come back to this in later lecture]

"Bad Controls" – TG's Pet Peeve

- **But just to preview it…** If you have an x that is truly exogenous (i.e., random) *[as you might have in natural experiment]*, do not put in controls, that are also affected by x!
  - Only add controls unaffected by x, or just regress your various y's on x, and x alone!

We will revisit this in later lecture…

Summary of Today *[Part 1]*

- We need conditional mean independence (CMI), to make causal statements
- CMI is violated whenever an independent variable, x, is correlated with the error, u
- Three main ways this can be violated
  - Omitted variable bias
  - Measurement error bias
  - Simultaneity bias

Summary of Today *[Part 2]*

- The biases can be very complex
  - If more than one omitted variable, or omitted variable is correlated with more than one regressor, sign of bias hard to determine
  - Measurement error of an independent variable can (and likely does) bias <u>all</u> coefficients in ways that are hard to determine
  - Simultaneity bias can also be complicated

Summary of Today *[Part 3]*

- To deal with these problems, there are some tools we can use
  - E.g., Proxy variables [discussed today]
  - We will talk about other tools later, e.g.
    - Instrumental variables
    - Natural experiments
    - Regression discontinuity

In First Half of Next Class

- Before getting to these other tools, will first discuss panel data & unobserved heterogeneity
  - Using fixed effects to deal with unobserved variables
    - What are the benefits? [There are many!]
    - What are the costs? [There are some…]
  - Fixed effects versus first differences
  - When can FE be used?
- Related readings: see syllabus

Assign papers for next week…

- Rajan and Zingales (AER 1998)
  - Financial development & growth
- Matsa (JF 2010)
  - Capital structure & union bargaining
- Ashwini and Matsa (JFE 2013)
  - Labor unemployment risk & corporate policy

Break Time

- Let’s take our 10-minute break
- We will do presentations when we get back