Empirical Methods

Topic 2b:

Sources of Bias

Sources of Bias

Latex Color = "LightGrey"

Bias and its Sources: Intro

- We're finally in a position to dig into the sources of bias...
  - ...more accurately the sources of inconsistency...
    - ★ ...that cause violations of

$$E(\epsilon_i|x_i) = 0$$

  - Recall that the violation of this assumption is called *endogeneity*.

Where does endogeneity come from?

The three most common reasons for endogeneity of a particular x_{ik} are:

1. Correlated unobservables
   - i.e., an unobserved factor determines both y_i and x_{ik}
2. Measurement error in x_{ik}
   - i.e., an explanatory variable is measured with error
     - ★ In which case, the *true* econometric model doesn't have any problems...
     - ★ ... but the one we *estimate* (with measurement error) does
3. "Simultaneous causality"
   - i.e., not only does x_{ik} cause y_i, but y_i also causes x_{ik}

$\rightarrow$ We'll cover each in turn

Correlated Unobservables

Latex Color = "LightGrey"

Correlated Unobservables I

- We already have a framework to understand the first reason for endogeneity, *Correlated unobservables*
  - It's given by our Omitted Variables formula:

$$
\begin{array}{rrcl}
\text{Truth:} & y & = & X\beta + \underbrace{\gamma q + \epsilon}_{\epsilon^*} \\
\text{You estimate:} & y & = & X\beta + \epsilon^* \\
\\
\Rightarrow & E(\hat{\beta}) & = & \beta + \gamma(X'X)^{-1}X'q \\
 & & = & \beta + \gamma\hat{\beta}_{q\_on\_X}
\end{array}
$$

Correlated Unobservables II

q is a "correlated unobservable":

- It impacts y_i,
  - (via γ)
- It is unobservable (to the econometrician)
  - And thus part of the (composite) error term, ϵ_i^*
    - ★ (We commonly attribute to the error term any "unobservables" influencing y_i)
- It is correlated with at least one of the x_{ik}
  - Else $\hat{\beta}_{q\_on\_X} = 0$

Correlated Unobservables III

The "classic" correlated unobservable measures the impact of education on wages

- Most people would agree that getting further education enhances skills, leading to professional opportunities that pay higher wages
  - (There is lots of empirical evidence to support this)
- The challenge is that it may be hard to measure all the factors that influence wages
  - In particular "ability"
    - ★ Short for "innate ability", a factor that differs across people in ways that make them more productive in the workplace
    - ★ (E.g. attention to detail, ability to identify and solve problems, working well with others, etc.)
    - ★ (It's hard to pin down exactly how to define ability...)

Correlated Unobservables IV

- The influence of ability on wages means that we will often have a correlated unobservable problem:

$$\text{Truth:} \quad y = \beta_1 + \beta_{educ} educ_i + \tilde{x}_i' \tilde{\beta} + \underbrace{\beta_{abil} ability_i + \epsilon_i}_{\epsilon_i^*}$$

$$\text{You estimate:} \quad y = \beta_1 + \beta_{educ} educ_i + \tilde{x}_i' \tilde{\beta} + \epsilon_i^*$$

where

  - As usual, we focus on a single covariate of interest ($educ_i$)...
    - ★ ...and lump all the other covariates (and their coefficients) into a composite $\tilde{x}_i$ ($\tilde{\beta}$)
  - *Even if* $E(\epsilon_i | x_i, ability) = 0$, we worry that $E(\epsilon_i^* | x_i) \neq 0$
    - ★ Because (the *conditional*) $Cov(educ_i, ability_i) \neq 0$
    - ★ (i.e. the covariation ... remaining after controlling for all the other $\tilde{x}_i$)

Correlated Unobservables V

- Applying the omitted variable bias formula to this setting yields:

$$E(\hat{\beta}_{educ}) = \beta_{educ} + \beta_{abil}\hat{\beta}_{abil\_on\_educ}$$

- Quick Quiz: What is the likely sign of any bias?*
  - Likely sign of β_{abil}?

    _____
  - Likely sign of $\hat{\beta}_{abil\_on\_educ}$?

    _____
  - Thus: _____

Correlated Unobservables VI

- I will use a real dataset to illustrate each of the three common sources of endogeneity bias,
  - And - more important - how to correct for each using Instrumental Variables!
- For this first, correlated unobservables/wage-ability example, we use data from Tom Mroz:
  - Mroz, T. (1987), "The Sensitivity of an Empirical Model of Married Womens Hours of Work to Economic and Statistical Assumptions," *Econometrica*, v55, 765-799.

See Stata Example in Class

  - Do these results seem reasonable to you?*
    - ★ Thus: _____
    - ★ (We'll discuss how to address this problem after introducing the other common sources of endogeneity)

Measurement Error

Latex Color = "LightGrey"

Measurement Error I

- We turn next to the second source of endogeneity
  - *Measurement Error*
- Suppose the true model is

$$y_i = \alpha + \beta x_i^* + \epsilon_i^*$$

  with $E(\epsilon_i^* | x_i^*) = 0$ (as in the CLRM)
- But there is measurement error in x_i : $x_i = x_i^* + \eta_i$:
  - With $\eta_i \sim (0, \sigma_\eta^2)$
  - The x_i *we* see, x_i, is equal to the true value, x_i^*, plus an error term, η_i

Measurement Error II

$$\boxed{\begin{aligned} y_i &= \alpha + \beta x_i^* + \epsilon_i^* \\ x_i &= x_i^* + \eta_i \end{aligned}}$$

- Let's make our life as easy as possible and assume the measurement error is uncorrelated with everything:

$$\begin{aligned} E(\eta_i | x_i^*) &= 0 \\ E(\eta_i | \epsilon_i^*) &= 0 \end{aligned}$$

  - This is called "classical measurement error"
    - ★ (Just be sure to check if this is reasonable in an application!)

Measurement Error III

- If we add and subtract $\beta\eta_i$ to our equation from the last slide we get

$$\begin{aligned} y_i &= \alpha + \beta x_i^* + \epsilon_i^* \\ &= \alpha + \beta x_i^* + \beta\eta_i + (\epsilon_{it}^* - \beta\eta_i) \\ &= \alpha + \beta x_i + \epsilon_{it} \end{aligned}$$

where $\epsilon_{it} = \epsilon_{it}^* - \beta\eta_{it}$ is a "composite error" of our true error, ϵ_i^*, and the measurement error, η_i

Measurement Error IV

- We now have a problem however:

$$\begin{aligned} Cov(x_i, \epsilon_i) &= Cov(x_i^* + \eta_i, \epsilon_i^* - \beta\eta_i) \\ &= Cov(x_i^*, \epsilon_i^*) + Cov(\eta_i, \epsilon_i^*) - \beta Cov(x_i^*, \eta_i) - \beta Cov(\eta_i, \eta_i) \\ &= -\beta V(\eta_i) \quad \text{as } Cov(\eta_i, \eta_i) = V(\eta_i) \\ &= -\beta\sigma_\eta^2 \end{aligned}$$

- And it's also the case that

$$\begin{aligned} V(x_i) &= V(x_i^* + \eta_i) \\ &= V(x_i^*) + V(\eta_i) + 2Cov(x_i^*, \eta_i) \\ &= \sigma_{x^*}^2 + \sigma_\eta^2 \end{aligned}$$

where $\sigma_{x^*}^2$ is the variance of the true variable, x_i^*

Measurement Error V

- We can apply our old stand-by equation to this setting (with only one x_i) and take $N \to \infty$ to show

$$\begin{aligned}
\hat{\beta} &= \beta + (X'X)^{-1}X'\epsilon \\
&= \beta + \frac{\frac{1}{N}X'\epsilon}{\frac{1}{N}X'X} \\
&\xrightarrow{p} \beta + \frac{-\beta\sigma^2_\eta}{\sigma^2_{x^*}+\sigma^2_\eta} \quad \text{as } \tfrac{1}{N}X'\epsilon \to Cov(x_i,\epsilon_i) \text{ and } \tfrac{1}{N}X'X \to V(x_i) \\
&\xrightarrow{p} \beta\left(1 - \frac{\sigma^2_\eta}{\sigma^2_{x^*}+\sigma^2_\eta}\right) \\
&\xrightarrow{p} \beta\left(\frac{\sigma^2_{x^*}}{\sigma^2_{x^*}+\sigma^2_\eta}\right) \\
&\xrightarrow{p} \lambda\beta
\end{aligned}$$

where $\lambda = \left(\frac{\sigma^2_{x^*}}{\sigma^2_{x^*}+\sigma^2_\eta}\right)$...

  - ...is the ratio of the variance of x_i^* $(\sigma^2_{x^*})$ to the total variation in x_i $(\sigma^2_{x^*} + \sigma^2_\eta)$

Measurement Error VI

$$\begin{aligned} \hat{\beta} &\xrightarrow{p} \beta \left(\frac{\sigma^2_{x^*}}{\sigma^2_{x^*} + \sigma^2_{\eta}} \right) \\ &\xrightarrow{p} \lambda \beta \end{aligned}$$

where $\lambda \equiv \left(\frac{\sigma^2_{x^*}}{\sigma^2_{x^*} + \sigma^2_{\eta}} \right)$

- Note:
  - There is no bias is there is no msmt error, $\sigma^2_{\eta} = 0$
  - If $\beta > 0$, then $\hat{\beta}$ has a negative bias
  - If $\beta < 0$, then $\hat{\beta}$ has a positive bias
- We call this *attenuation bias*: $\hat{\beta}_1$ is biased towards zero

Measurement Error VII

- For this second, measurement-error example, we use data from average health insurance coverage across the 50 US states (and the District of Columbia) in 2007

See Stata Example in Class

- Do these results seem reasonable to you?*
  - _____

Measurement Error VIII

- OK, that's the intuition when we have a single x_i
  - What if we have a richer model

$$y_{it} = \beta_1 + \beta_k x_{ik}^* + x_{i,-k}\beta_{-k} + \epsilon_i^*$$

where

- x_{ik}^* is the variable we care about and
  - ★ (And it has classical measurement error)
- $x_{i,-k}$ and β_{-k} are the remaining variables and parameters

Measurement Error IX

- Let's control for $x_{i,-k}$ using the "residual maker", M_{-k}:

$$\begin{aligned}
\hat{\beta}_k &= \beta_k + (X'M_{-k}X)^{-1}X'M_{-k}\epsilon \\
&= \beta_k + \frac{\frac{1}{N}X'M_{-k}\epsilon}{\frac{1}{N}X'M_{-k}X} \\
&\xrightarrow{p} \beta_k + \frac{-\beta_k\sigma^2_\eta}{\sigma^2_{x^*}+\sigma^2_\eta} \\
&\xrightarrow{p} \beta\left(\frac{\sigma^2_{x^*_{-k}}}{\sigma^2_{x^*_{-k}}+\sigma^2_\eta}\right)
\end{aligned}$$

where you can show (but we won't) by focusing on the term in red:

  - The numerator of the bias term *doesn't* change
    - ★ (As only x_{ik} has measurement error and this measurement error is uncorrelated with all the other x's)
  - The denominator term has *less* variation in x_i^* once we control for all the other x's using M_{-k}
    - ★ i.e., $\lim_{N\to\infty}(X_k^{*'}M_{-k}X_k^*) = \sigma^2_{x^*_{-k}} < \sigma^2_{x^*} = \lim_{N\to\infty}(X_k^{*'}X_k^*)$

- **Bottom line:** The problem is worse!

Simultaneous Equations

Latex Color = "LightGrey"

Simultaneous Equations

- We turn now to our third and final source of endogeneity
  - *Simultaneous Equations*
- In a system of simultaneous equations, not only does a change in x_{ik} cause a change in y_i...
  - ... but a change in y_i causes a change in x_{ik}!
    - ★ Or, more generally, y_i and x_{ik} are "jointly determined"
- The most common example of this is Supply and Demand

A Simple Example I

- Let me start with a simple demand equation:
- Suppose

$$q_i = \beta_0 + \gamma_1 p_i + \beta_1 v_i + \epsilon_i$$

  where q_i and p_i are quantity and price in market i
- Question: Is this a demand equation or a supply equation?
- Answer*:

A Simple Example II

$$q_i = \alpha_0 + \alpha_1 p_i + \alpha_2 v_i + \epsilon_i \tag{1}$$

- Suppose we were to estimate (1)...
  - Let's further suppose we were to specify exactly what v_i measures
    - For example, suppose it's a demand shifter like income
- How would we interpret the estimated coefficient on price, $\hat{\alpha}_1$?
  - Will $\hat{\alpha}_1$ estimate a demand elasticity?
- Answer*:

A Simple Example III

- Let's show the reasons for this...
- With an extended example using supply and demand:

$$\begin{aligned} \text{[Demand:]} \quad q_i &= \alpha_0 + \alpha_1 p_i + \alpha_2 inc_i + \epsilon_{i1} \\ \text{[Supply:]} \quad q_i &= \beta_0 + \beta_1 p_i + \beta_2 w_i + \epsilon_{i2} \end{aligned}$$

where

  - $(q_i, p_i) \equiv$ (quantity, price),
  - $inc_i \equiv$ average income, and
  - $w_i \equiv$ input price in market i
  - $E(\epsilon_i | inc_i, w_i) = 0, \quad V(\epsilon_i | inc_i, w_i) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ & \sigma_2^2 \end{bmatrix}$

Simultaneous Equations Definitions I

1. Endogenous Variables $\equiv$
   - Those variables jointly determined by the behavioral relationships under consideration
   - Must have one equation per endogenous variable to completely specify the system
   - (q_i, p_i) in our example
2. Exogenous Variables $\equiv$
   - Those variables determined outside the model but impact variables inside the model
   - (inc_i, w_i) in our example

Simultaneous Equations Definitions II

3. Structural Equations $\equiv$
   - Derived from economic theory, they summarize the behavioral relationships of interest
   - Typically relate endogenous variables as functions of each other and exogenous variables
   - In our example:

$$\begin{aligned} \text{[Demand:]} \quad q_i &= \alpha_0 + \alpha_1 p_i + \alpha_2 inc_i + \epsilon_{i1} \\ \text{[Supply:]} \quad q_i &= \beta_0 + \beta_1 p_i + \beta_2 w_i + \epsilon_{i2} \end{aligned}$$

Simultaneous Equations Definitions III

4. Reduced-Form Equations $\equiv$
   - Relate the endogenous variables to (*only*) the exogenous variables
   - In our example:

$$\begin{aligned} q_i &= \pi_{11} + \pi_{12} inc_i + \pi_{13} w_i + \nu_{i1} \\ p_i &= \pi_{21} + \pi_{22} inc_i + \pi_{23} w_i + \nu_{i2} \end{aligned}$$

   - Note: *Only* endog vars on LHS; *only* exog vars on RHS
     - ★ When you have multiple endogenous variables, the relationship between endogenous variables and exogenous variables, $f(endog_i|exog_i)$, is *all the data can (directly) tell you*.
     - ★ While we are often interested in the relationship between endogenous variables, these are not directly measurable from data
     - ★ We recover them using "Identification" (to come)

The Reduced Form

- How did we get to the reduced form?
  - By solving the structural equations for the endogenous variables.
  - In our example (Verify these outside class!):

$$\begin{aligned} q_i &= \pi_{11} + \pi_{12} inc_i + \pi_{13} w_i + \nu_{i1} \\ &= \frac{\alpha_1\beta_0 - \alpha_0\beta_1}{\alpha_1 - \beta_1} + \frac{-\alpha_2\beta_1}{\alpha_1 - \beta_1} inc_i + \frac{\alpha_1\beta_2}{\alpha_1 - \beta_1} w_i + \frac{-\beta_1\epsilon_{i1} + \alpha_1\epsilon_{i2}}{\alpha_1 - \beta_1} \end{aligned}$$

$$\begin{aligned} p_i &= \pi_{21} + \pi_{22} inc_i + \pi_{23} w_i + \nu_{i2} \\ &= \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{-\alpha_2}{\alpha_1 - \beta_1} inc_i + \frac{\beta_2}{\alpha_1 - \beta_1} w_i + \frac{-\epsilon_{i1} + \epsilon_{i2}}{\alpha_1 - \beta_1} \end{aligned}$$

A Few Questions I

1. Can we estimate a structural equation by OLS?, e.g.

$$q_i = \alpha_0 + \alpha_1 p_i + \alpha_2 inc_i + \epsilon_{i1}$$

2. Answer: No.
   - p_i is correlated with ϵ_{i1}
   - To see this, use the formula for p_i from the reduced form:

$$\begin{aligned} E(p_i \epsilon_{i1}) &= E\left[(\pi_{21} + \pi_{22} inc_i + \pi_{23} w_i + \nu_{i2})\epsilon_{i1}\right] \\ &= 0 + E(\nu_{i2}\epsilon_{i1}) \qquad \text{By (A2, Mean-zer} \\ &= E\left[\tfrac{1}{\alpha_1 - \beta_1}(-\epsilon_{i1} + \epsilon_{i2})\epsilon_{i1}\right] \\ &= \tfrac{1}{\alpha_1 - \beta_1}(-\sigma_1^2 + \sigma_{12}) \neq 0 \end{aligned}$$

     - ★ What is the sign of the bias?
     - ★ Answer:* _____

A Few Questions II

3. Can we estimate a reduced-form equation by OLS?, e.g.

$$q_i = \pi_{11} + \pi_{12} inc_i + \pi_{13} w_i + \nu_{i1}$$

4. Answer: Yes
   - Since $E(\epsilon_i | inc_i, w_i) = 0$, all the RHS variables are uncorrelated with ϵ_{i1} and ϵ_{i2}
     - ★ (As from two slides ago, $\nu_{i1} = \frac{-\beta_1 \epsilon_{i1} + \alpha_1 \epsilon_{i2}}{\alpha_1 - \beta_1}$)
     - ★ (Similarly ν_{i2} is a function of ϵ_{i1} and ϵ_{i2})
     - ★ (And thus $E(\nu_{i1} | inc_i, w_i) = 0$ and $E(\nu_{i2} | inc_i, w_i) = 0$...
     - ★ ... when $E(\epsilon_i | inc_i, w_i) = 0$)

Why Bother with Structural Equations? I

- You can hopefully see that it is much easier to estimate reduced-form equations than structural equations
- Question: Why bother with structural equations at all???
- Answer: We cannot answer many interesting economic questions using only a reduced-form analysis, e.g.
  - The price elasticity of demand
  - The strategic response of one firm's (advertising/price/product choice) to another's (advertising/price/product choice)
  - (For these - and many others - we *need* the parameters in the structural equations)

Why Bother with Structural Equations? II

- Bottom line: Identifying the parameters in structural equations is difficult but important
  - That's why you're here!
- Moving between the reduced form and structural parameters is the topic of *Identification*
  - (As defined by econometricians anyway...)

Identification I

- Identification is the process of recovering structural parameters from reduced-form parameters
- In our example:
  - We want to be able to solve for $(\alpha_0, \alpha_1, \alpha_2, \beta_1, \beta_2, \beta_3)$ from $(\pi_{11}, \pi_{12}, \pi_{13}, \pi_{21}, \pi_{22}, \pi_{23})$
    - ★ (Technically also for parameters in $f(\epsilon_{ij}|\Sigma)$ from $f(\nu_{ij}|\Omega)$, where $V(\nu_i) = \Omega$)
- Many texts cover this in detail
  - But I've concluded the math investment isn't worth it
  - What's the intuition???

Identification II

- In our example, you might think we have a chance at least:
  - We have six parameters we can estimate ($\hat{\pi}$'s)
  - And six parameters we want to recover from them (α's, β's)
- What is the general version of this?
  - Take one of our structural equations, e.g. Demand, to see:

$$q_i = \alpha_0 + \alpha_1 p_i + \alpha_2 inc_i + \epsilon_{i1}$$

Identification III

$$\text{[Demand:]} \qquad q_i = \alpha_0 + \alpha_1 p_i + \alpha_2 inc_i + \epsilon_{i1}$$

- Let $M =$ the number of endogenous variables in the overall system of equations
  - In our example, $M = 2$: (q_i, p_i)
- Let $K =$ the number of exogenous variables in the overall system of equations
  - In our example, $K = 3$: $(1, inc_i, w_i)$
    - ★ Where, recall, w_i enters the system of equations from Supply

Identification IV

$$\text{[Demand:]} \qquad q_i = \alpha_0 + \alpha_1 p_i + \alpha_2 inc_i + \epsilon_{i1}$$

- You can show (but we won't) that the equation relating the structural parameters for any single equation to reduced-form parameters can be written as a system of
  - K equations in $M + K$ unknowns*
    - ★ (To see this, see the next two - optional - slides)
- What happens when we have fewer equations than unknowns?
  - We're cooked
    - ★ (We're under-identified)
    - ★ There are infinitely many ways to solve for the structural parameters from the reduced-form parameters.
    - ★ Without restrictions, we can't recover the structure

Identification V*

- Aside: Why K equations in $M + K$ unknowns???
  - Need some extra notation to show.
  - Let $y_i' = [q_i \quad p_i]$, $x_i' = [1 \quad inc_i \quad w_i]$, and $\epsilon_i' = [\epsilon_{1i} \quad \epsilon_{2i}]$
    - ★ Also $\Gamma = \begin{bmatrix} 1 & 1 \\ -\alpha_1 & -\beta_1 \end{bmatrix}$ and $\beta = \begin{bmatrix} -\alpha_0 & \beta_0 \\ -\alpha_2 & 0 \\ 0 & -\beta_2 \end{bmatrix}$
  - We can then write a/our system of simultaneous equations as

$$\text{Structural equations:} \qquad y_i'\Gamma + x_i'\beta = \epsilon_i' \qquad \text{or}$$

$$y_i' = -x_i'\beta\Gamma^{-1} + \epsilon_i'\Gamma^{-1}$$

$$\text{Reduced form equations:} \qquad y_i' = x_i'\Pi + v_i'$$

Identification VI*

- Solving for the structural parameters in the j^{th} equation means solving

$$\begin{aligned} \Pi\Gamma_j + \beta_j &= 0 \\ \Leftrightarrow [\Pi \quad I_K] \begin{bmatrix} \Gamma_j \\ \beta_j \end{bmatrix} &= 0 \end{aligned} \tag{2}$$

  - If there are no restrictions, there are M+K elements in $\begin{bmatrix} \Gamma_j \\ \beta_j \end{bmatrix}$
  - And while there are KM terms in Π, $[\Pi \quad I_K]$ is of dimension $K \times (M + K)$
    - ★ As such, it has maximum rank K
    - ★ Thus there are only K "pieces of information" in the equations in (2)
  - Thus K equations in $M + K$ unknowns.

Identification: Restrictions I

- Question: Where can we find some restrictions?
- Answer: _____
  - The cost to Apple for microchips should not influence consumers' willingness-to-pay (WTP) for iPads
    - ★ i.e. w_i shouldn't enter the demand curve
  - Consumer income should not influence the cost of making an iPad
    - ★ i.e. inc_i shouldn't enter the supply curve
- How many do we need?
  - At least M

Identification: Restrictions II

- The most common types of restrictions are called exclusion restrictions
- These are either:
  1. One of the *endogenous* variables isn't in a structural equation
     - ★ In which case, there is one less unknown to solve for (Good!)
     - ★ (In our example, this isn't relevant...
     - ★ ...as each structural equation has all the endogenous variables)
  2. One of the *exogenous* variables isn't in a structural equation
     - ★ There is again one less unknown to solve for (Still Good!)
     - ★ In our example, w_i doesn't enter Demand

Identification: Restrictions III

- So where do we find our M restrictions?
  - One is free: we normalize the parameter on the dependent variable to 1
  - The $M-1$ others are usually one of the two types of excluded variables:
    - ★ Excluded endogenous variables (M^*)
    - ★ Excluded exogenous variables (K^*)
- For any equation, we divide each of the endogenous and exogenous variables into those that are included and excluded:
  - $M = \tilde{M} + M^* + 1$, where there are and $\tilde{M}$ included and M^* excluded (right-hand-side) endogenous variables
  - Similarly $K = \tilde{K} + K^*$, where $\tilde{K}$ (K^*) are the number of included (excluded) exogenous variables

Identification: Restrictions IV

- If the number of excluded variables has to exceed $M-1$, we get:

$$M^* + K^* \geq M - 1$$
$$\Rightarrow \boxed{K^* \geq \tilde{M}}$$

- In other words, the number of excluded exogenous variables must be at least as great as the number of the included (right-hand-side) endogenous variables in each equation
  - This is called the Order Condition for identification
- Also important: the Rank Condition for identification:
  - Our M restrictions cannot be linearly dependent.

Identification: Restrictions V

- Some simple intuition:
  1. If all our exclusions are excluded *endogenous* variables, what then?
     - ★ Easy! We don't have an endogeneity problem.
     - ★ All the structural parameters are just given by the reduced-form parameters for the j^{th} equation.

Identification: Restrictions VI

- Some simple intuition, cont:
  2. If all our exclusions are excluded *exogenous* variables, what then?
     - ★ The only way an excluded exogenous variables could be influencing the dependent variable in the excluded-endogenous equation (e.g. the input price in a demand equation) is through their influence on the included endogenous ones (e.g. price)
     - ★ This provide the variation in the data to recover the effect of the included endogenous variable
     - ★ We need to have (at least) one excluded exogenous variable for each included endogenous
     - ★ The excluded exogenous variables are what we will soon call our Instruments

Econometric v Economic Identification I

- Our discussion of identification to this point covers what I would call "econometric identification"
- There are instances when econometric models are identified "econometrically" but not "economically"*
  - An example: when we consider sample selection models, we might specify two equations, one relating wages to covariates, x_i, and another analyzing whether or not someone works at all:

$$wage_i = x_i'\beta + \epsilon_i$$
$$P(work_i) = f(x_i'\rho)$$

  - Where x_i includes covariates we think influence working decisions
    - ★ E.g., education, gender, marriage status, etc.
    - ★ ...and $f(\cdot)$ is the CDF of a Normal random variable

Econometric v Economic Identification II*

$$wage_i = x_i'\beta + \epsilon_i$$
$$P(work_i) = f(x_i'\rho)$$

- One can show (but we won't) that you can write the wage equation as:

$$wage_i = x_i'\beta + \sigma_{12}\lambda(x_i'\rho) + \epsilon_i^*$$

  - Where λ is a particular nonlinear function of $x_i'\rho$

- Note that the "selection equation" (the 2nd one)
  - ...*does not* have an excluded exogenous variable
    - ★ i.e., there are no *additional* variables (beyond x_i) that enters $\lambda(\cdot)$
  - If the $\lambda(\cdot)$ function was linear, it would be *obvious* that this model is unidentified
    - ★ (As $wage_i = x_i'\beta + \sigma_{12}(x_i'\rho) = x_t'(\beta + \sigma_{12}\rho)$)
    - ★ (And even if we had an estimate of $\hat{\rho}$, we couldn't separate β from σ_{12})

Econometric v Economic Identification III

$$wage_i = x_i'\beta + \sigma_{12}\lambda(x_i'\rho) + \epsilon_i^*$$

- When x_i enters both linearly and as part of the $\lambda(\cdot)$ function, the non-linearities in λ "technically" identify the model...
  - But I (and many) would argue that it really isn't identified.
    - ★ (Why not? Because your functional form has to be 100% correct...)
    - ★ (...and we're rarely so certain about functional forms)

Econometric v Economic Identification IV

- Many people (including me) don't like "econometric identification"...
- So you'll often hear questions in seminar about...
  - ..."what is it in an author's *data* that identifies an effect of interest?"
- It is the duty of the author to provide a (good!) answer for this
  - No matter how complicated the structural model
- Without it, you risk having your results discounted
- (The problem: sometimes it really is hard to know)

Simultaneous Equations Example I

- For this final, simultaneous equations example, we use annual data on the purchase of...
- The data include prices and quantities of _____ over 30 years.

See Stata Example in Class

Simultaneous Equations Example I

- A common problem in demand estimation is the presence of "unobserved quality", i.e. that the quality of a good in a market is observed by market participants, but not the econometrician.
  - What is likely to be the sign of the bias on the price coefficient in this case?*
    - ★ _____
- Let's estimate demand by regressing quantity on price. Do these results seem reasonable to you?*
  - _____

Sources of Bias: Example from Class I

And now some questions for you:

- Did any of you write an empirical Bachelor's thesis?
  - Or will any of you write an empirical Master's thesis?
- What was/is the underlying estimating equation(s)?
  - Were you worried about any sources of bias?
    - From which of our three types?
  - What do you think was the sign of any bias?

Figuring these things out can take some time!

Sources of Bias Conclusion

- The point: when doing econometrics, it's important to have in mind the structure of the full data generating process...
  - And how to fix it!
- That's coming next with

Instrumental Variables Estimation

Table of Contents