PS2030
Political Research and Analysis
Unit 2: Regression Models: Problems and Extensions
3. Endogenous Regressors

Spring 2025, Weeks 7-8

WW Posvar Hall 3600

Professor Steven Finkel

- Endogeneity and Instrumental Variables
  - OLS assumption of the independence of X and ε
  - This is also called the "exogeneity" assumption, that E(Xε=0)
  - Violation of this assumption indicates *endogeneity* in the model, or that X is an "endogenous regressor" in the equation predicting Y
  - Violation leads to all kinds of problems with OLS estimation, and severely inhibits the use of regression for making inferences about the ***causal effects*** of X on Y
  - Occurs when:
    - Relevant explanatory variables correlated with X have been omitted from the model
    - Y causes X ("reverse causality") in addition (or instead of) X causing Y
    - X contains random measurement errors
  - This week will discuss diagnostics and corrections, mainly via *instrumental variables* (IV) analysis or "*two-stage least squares*" (TSLS); we will discuss additional models for causal inference in Unit 4

Omitted Variables

- Already discussed to some extent in context of multiple regression, where we controlled for *observable* characteristics that may produce spurious association between X_1 and Y. Omit these factors, say X_2 and X_3, and they are folded into the Y equation's error term. If X_1 and X_2/X_3 are related, then X_1 and the error term would then be related
- True Specification: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$
- Your Specification $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i^*$

where $\varepsilon_i^* = \varepsilon_i + \beta_2 X_{2i} + \beta_3 X_{3i}$

X_1 is now related to ε^*

- OLS will yield *biased* (and *inconsistent*) estimates of β_1 to extent that X_1 is related to X_2/X_3. The estimated effect of β_1 will include the correlated effects of X_1 on Y through its relationships to X_2 and X_3
- OLS is greedy! It maximizes the effect of X_1 even though some of the effect properly belongs to ε (via the omitted variables)

- Can see this with "covariance algebra" solution to parameter estimation:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

multiply through by X_1 :

$$X_1 Y_i = X_1(\beta_0 + \beta_1 X_{1i} + \varepsilon_i)$$

$$X_1 Y_i = X_1\beta_0 + X_1\beta_1 X_1 + X_1\varepsilon_i$$

take expected covariances of each side of the equation:

$$\text{Cov}(X_1 Y_i) = \beta_1 \text{Var}(X_1) + \text{Cov}(X_1\varepsilon_i)$$

$$\beta_1 = \frac{\text{Cov}(X_1 Y_i) - \text{Cov}(X_1\varepsilon_i)}{\text{Var}(X_1)} \quad \text{and} \quad \beta_{1OLS} = \frac{\text{Cov}(X_1 Y_i)}{\text{Var}(X_1)}$$

- So the usual OLS formula is *only valid* when $\text{Cov}(X_1\varepsilon)=0$. If not, and there is a positive covariance between X_1 and the error term, OLS will *overestimate* X's true effect
- Problem: we don't observe $\text{Cov}(X_1\varepsilon)$ so can't estimate it directly!!
- This also shows that we can "solve" for regression coefficients in terms of the observed variances and covariances between variables in our sample -- but only if certain assumptions are satisfied

- We deal with this in many cases with multiple regression – we bring all observed covariates into the equation and then we try to defend assumption of no covariance between X and ε. Problem: what about *unobservables*, things you could not measure or do not necessarily know how to include ?
- Example: Civic education exposure on political knowledge
- We know that this "treatment" is subject to self-selection biases – people who select into the civic education treatment are different from others, even if we control for lots of observed factors like education, income, political interest, and the like. There may be *unobservable* differences between treated and untreated people on:
  - Personality
  - Motivation
  - Family and/or social network structure
  - All of which may be difficult, if not impossible to observe and bring into the analysis directly

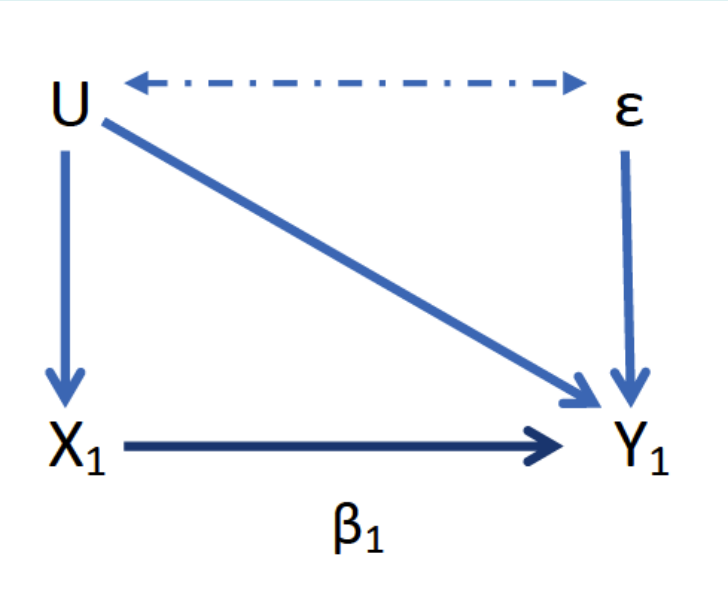
- So true specification:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_k X_{ki} + \beta_u U_i + \varepsilon_i$$

- Where U_i is the summary of all the unobservables that may be related to X_1 and also related to Y
- U_i is folded into the error term
- OLS will produce biased and inconsistent estimates of β_1 to the extent that X_1 is related to U, and U is related to Y
- Same problem: we don't observe U, so we can't control its effects directly
- Another way to look at it: we don't have enough information to estimate the true effects of both X and ε (which contains U) on Y
- Solutions: 1) Assume no covariation between X and ε, which is the usual OLS practice and which we know (in this case) is wrong!

  2) Add more information to the causal system!

Here is what this causal system looks like:

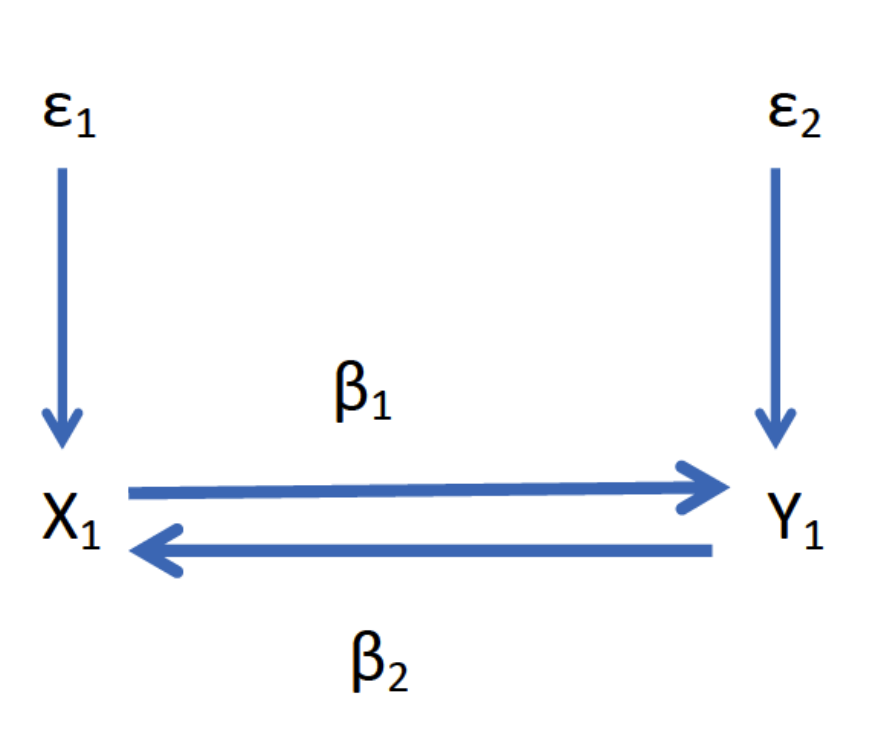


U causes both X_1 and Y_1
U is unobserved, so folded into ε
X is now related to ε and hence "endogenous"
OLS β_1 will give the true effect of X_1 *plus* some correlated effect from U
We need additional information to estimate the "true" effects of X, independent from U

Reciprocal Causality

- Similar problem of $E(X\varepsilon)\neq 0$ in cases where $Y \rightarrow X$ as well as the usual $X \rightarrow Y$ that we have been dealing with in the class
- We don't really think that factors *simultaneously* cause one another, i.e., at any given instant there is mutual causation, but we think that there is reciprocal (feedback) effects from one variable to another, and that the effects have or are taking place around the time period of observation
- Does political knowledge lead to exposure to civic education, or does exposure to civic education lead to political knowledge? In a cross-sectional study that takes place after the fact, it is difficult to disentangle these processes, so may have to assume possible reciprocal causal effects.
- This is extremely common problem in empirical political science research, in virtually all subfields!!
- Leads to error term being related to X, so biased and inconsistent estimates via OLS, *and* leads to general difficulties in estimation due to lack of information needed to "solve" for the model's parameters

A Cross-Sectional Example



$$Y_1 = \beta_1 X_1 + \varepsilon_2$$

$$X_1 = \beta_2 Y_1 + \varepsilon_1$$

Problems?

- Substitute the equation for Y_1 into the X_1 equation:

$$X_1 = \beta_2(\beta_1 X_1 + \varepsilon_2) + \varepsilon_1$$

$$X_1 = \beta_2\beta_1 X_1 + \beta_2\varepsilon_2 + \varepsilon_1$$

- So X_1 is a function of ε_2. But, in the equation predicting Y_1, OLS assumes that X_1 and ε_2 are unrelated. This is the $E(X\varepsilon)=0$ assumption for the OLS estimate of β_1 to be unbiased.
- **Therefore OLS cannot be used in context of reciprocal effects causal models**
- The same issue affects estimating the effect of Y_1 on X_1 (β_2). Y_1 is intrinsically related to ε_1, while OLS assumes it has no relationship, so an OLS estimation of β_2 would also be biased
- So both X and Y are "endogenous" variables, related to their equation's error terms

- Further problem: We can't manipulate the covariances and variances of the observed variables to solve for the unknown parameters.
- We call this a problem of **"underidentification"**

$$X_1 = \beta_2 Y_1 + \varepsilon_1$$

$$Y_1 = \beta_1 X_1 + \varepsilon_2$$

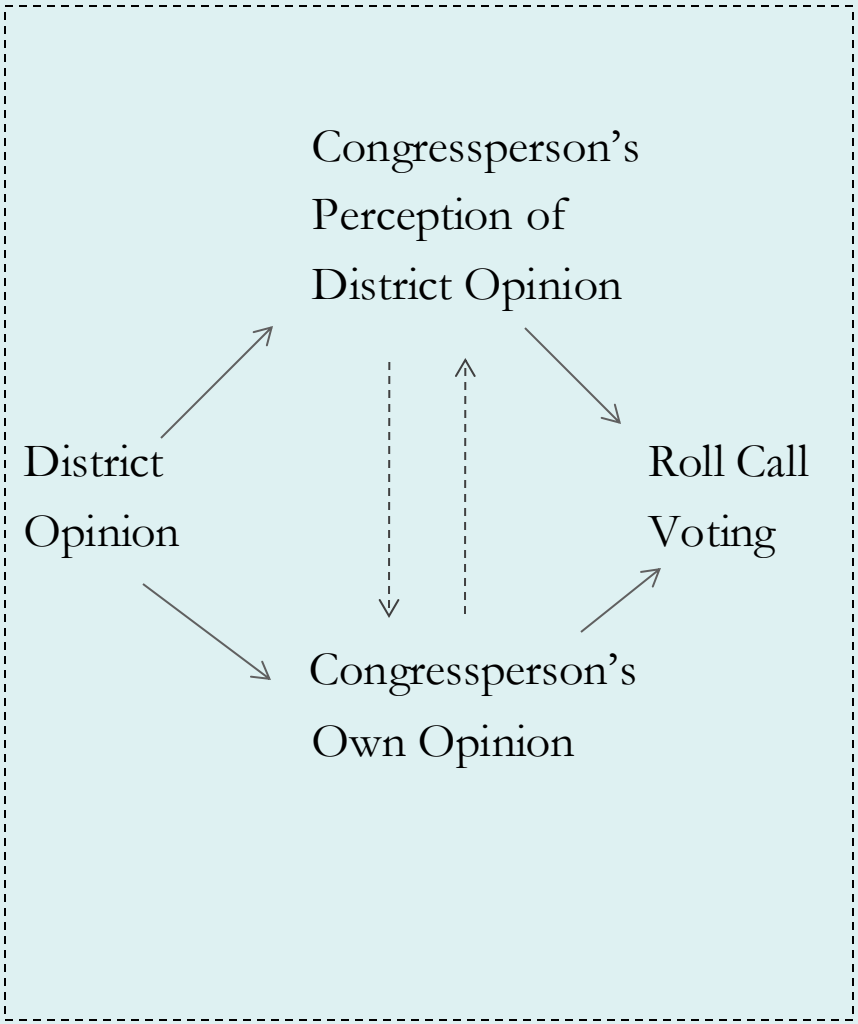
$$\beta_1 = \frac{\text{Cov}(X_1 Y_i) - \text{Cov}(X_1 \varepsilon_2)}{\text{Var}(X_1)} \text{ and } \beta_2 = \frac{\text{Cov}(X_1 Y_i) - \text{Cov}(Y_i \varepsilon_1)}{\text{Var}(Y_1)}$$

- Unknowns: 2 regression parameters and 2 error covariances ($X_1\varepsilon_2$ and $Y_1\varepsilon_1$). Knowns: the variance of X, the variance of Y, and their covariance. We cannot solve for the "unknowns" (uniquely) in terms of the known variances and covariances. We need more information!!!
- Note: this would be true even if $E(X_1\varepsilon_2)=E(Y_1\varepsilon_1)= 0$; the identification problem would still arise. We need to estimate the variance of the error terms ε in any regression model, so in the above example we actually have 6 total unknowns, and in the example in this bullet point we still have 4 total unknowns but only 3 knowns

Additional Terminology

- Many of these ideas have come into econometric/political science analysis via *structural equation models* (SEM), or *covariance structure analysis* (CSA). These models involve systems of equations with multiple dependent variables and possible reciprocal causality. Estimates are obtained by expressing the *unknown* model parameters in terms of the *known* variances and covariances of the observed variables, and solving for them (if the parameters are "identified").
- Basic distinctions
  - Models are either **"recursive"** (unidirectional causality), or **"non-recursive"** (reciprocal causality)
  - Variables are either **"endogenous"** (determined within the causal system), or **"exogenous"** (determined outside of the causal system, or "given")
  - Exogenous variables (by definition) satisfy the $E(X\varepsilon)=0$ condition
  - Until this week in the course, we have examined "recursive" models with one "endogenous" dependent variable and all "exogenous" independent variables
  - Now we are examining situations where the independent variables may also be "endogenous", and where the models may be "non-recursive" as well

An SEM Example: Miller and Stokes (1956) Model of Congressional Representation



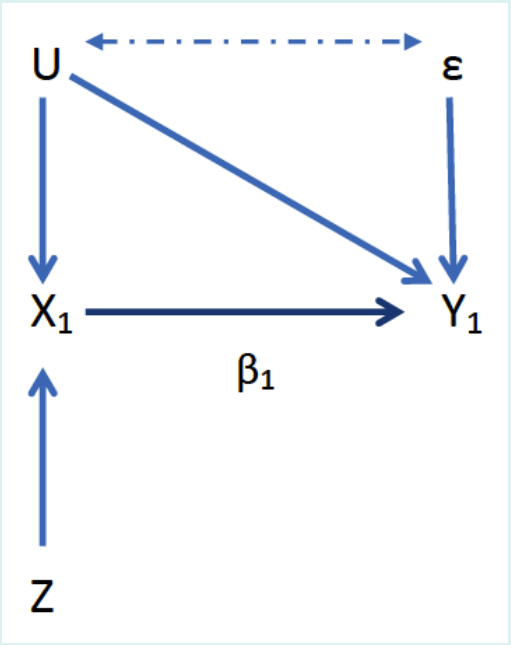
Illustrates features of SEMs:

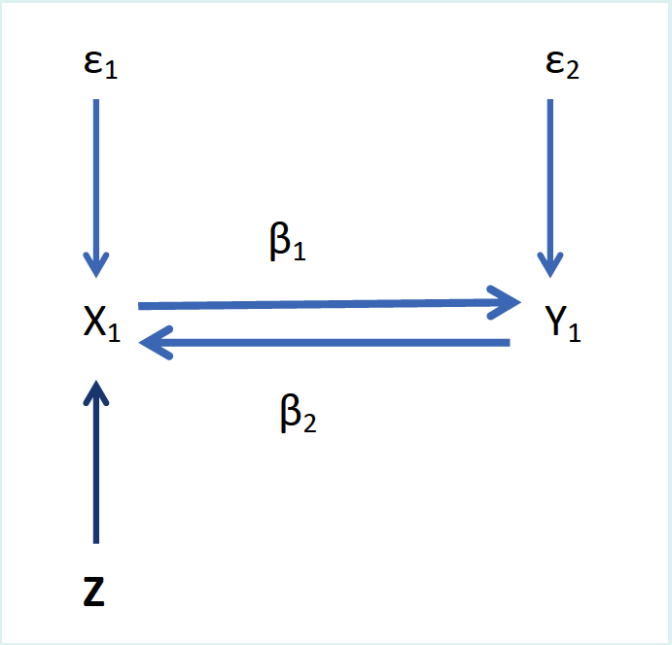
- *System of Equations*: here modeling three different DVs
- *Exogenous* versus *Endogenous* variables (as opposed to *Independent* versus *Dependent* variables)
- *Direct* versus *Indirect* Causal Effects
- *Recursive* versus *Non-Recursive* Models
- *Identification* of model parameters: is there enough information in the model to estimate all the coefficients of interest?

"Identifying" the Effects of Endogenous Variables

- We need more information to estimate the effects of variables that are endogenous in a given equation or (non-recursive) causal system
- We apply one simple rule: the "order condition" for identification:
  - If equation involves *p* endogenous variables, then there must be at least *(p-1)* **excluded exogenous** variables for the equation to be identified. Another way to put it: for each endogenous variable that is included as a predictor, there must be at least one excluded exogenous variable to identify the equation
  - That is, must have a Z that does *not* have an effect on a particular endogenous variable for every endogenous variable that does. This represents an "**exclusion" restriction** that allows identification of the given equation (i.e. the Z is "excluded" from the Y equation in question)
  - Z, being exogenous, is also unrelated to Y's error term ε
- Z is called an **"instrumental variable"** or an **"instrument"** for the endogenous X variable
- Hard to find, but *extremely* useful when (if) you do

Estimating Models with Instrumental Variables





To estimate β_1, we need an instrument Z in each case for X_1
Conditions that Z *MUST* Fulfill:

1) The "Exclusion Restriction": Z does not cause Y_1 except through X_1
2) The "Exogeneity Restriction": Z is unrelated to U and ε

- The Y_1 equation in either case is **"just-identified"**; it has one excluded exogenous variable and one included endogenous variable. (The X_1 equation on the right is not identified, so we cannot estimate β_2).

$$Y_1 = \beta_1 X_1 + \varepsilon_2$$

- Multiply through by the excluded exogenous variable Z:

$$ZY_1 = Z\beta_1 X_1 + Z\varepsilon_2$$

- Take expectations based on the the exogeneity assumption that $E(Z\varepsilon_2)=0$) and solve for β_1:

$$\beta_1 = \frac{Cov(Z_1 Y_1)}{Cov(Z_1 X_1)}$$

- Beautiful!! We take the covariance of the instrument with the *dependent* variable and divide by the covariance of the instrument with the endogenous *independent* variable to arrive at an unbiased effect of the endogenous independent variable. This is "**instrumental variables analysis**".

$$\hat{\beta}_{1IV} = \frac{\Sigma(Z_i - \bar{Z})(Y_i - \bar{Y})}{\Sigma(Z_i - \bar{Z})(X_i - \bar{X})}$$

and $$\sigma^2_{\hat{\beta}_{1IV}} = \frac{\sigma^2}{N\sigma_x^2 r_{XZ}^2}$$

So: variance (and standard error) of the IV estimator:

- Increases with smaller N, as in OLS
- Increases as explanatory power of equation decreases, as in OLS
- Increases as the covariation between X and the instrument Z decreases

- This means that we want instruments that are *strongly* related to X and that also satisfy the other IV assumptions
- This makes "good" instruments even harder to find

Logic of IV Analysis

- The logic of IV analysis is as follows. Given an endogenous regressor **X** in an equation with some dependent variable **Y**, we find some *exogenous* variable **Z** that produces change in Y through one mechanism only -- *the mediating effect of* X.
- Because Z is:
  - (a) Exogenous; and
  - (b) Has no direct effect on Y, then:
- Any changes in Y that may result from changes in Z ***must be attributable to X***, and must also be **unrelated to the problematic endogenous part of the X-Y relationship**.
- So the instrumental variable gives us *exogenously-induced* changes in X, and we then test whether these changes produce subsequent changes in Y.

- But lots can go wrong!!
  - First: IVs are notoriously difficult to find
  - In the reciprocal effects case, e.g.: we need an exogenous variable unrelated to each equations' error terms, that directly affects *one* of the endogenous variables in the reciprocal effects causal but is *excluded,* i.e., does not affect the other endogenous variable in question
  - Can we find an exogenous variable that affects one and only one of two reciprocally linked variables? Difficult.
  - This is one reason why *panel* or *longitudinal* data is very useful for causal inference; we can under some conditions use the *lags* of variables as instruments, as the assumptions of IV analysis are easier to defend, and, if lots of time periods of observation are available, lots of potential instruments exist also!!

- Also important: the exclusion restriction must hold in order to estimate this effect correctly. But this assumption cannot be definitively established empirically**, it must be justified theoretically**
- Let's say that Z also causes Y in the causal system in the figures. In that case, we could not derive an expression for (a hypothetical) β_1 in terms of the observed variances and covariances. As far as we can get is:

$$ZY_1 = Z\beta_1 X + Z\beta_3 Z + Z\varepsilon_2$$

$$Cov(ZY_1) = \beta_1 Cov(Z_1 X) + \beta_3 Var(Z)$$

$$\beta_1 = \frac{Cov(ZY_1) - \beta_3 Var(Z)}{Cov(Z_1 X)}$$

- And we can't estimate β_3 empirically to "prove" that it is zero because it is a partial regression coefficient controlling for X_1, which we cannot get because of the OLS violations
- So IV assumptions to some extent need to be based on theory and common sense, since the exclusion condition cannot definitively be tested

- Finally, often difficult to justify the exogeneity assumption of the instrument, i.e., that it is unrelated to the error term of the equation in question. This means it cannot be related to any other unmeasured or unobserved cause of the equation's dependent variable.
- We can test this under some conditions (as we will see below), but still must be justified in theory
- This is one reason why experimental analysis is so attractive: by randomizing assignment to treatment and control groups, one creates automatically an exogenous "instrument" that is uncorrelated with all error terms in the causal system, and which can be useful in later analyses (e.g. "intent to treat" models, "encouragement designs" etc.)
- This is also one reason why many recent IV analyses use **"natural experiments"**, as the IV assumptions can be justified more easily.

- Classic Example #1: The "return" on earnings from education. The kinds of people who seek education are likely to have unobservables that relate to earning power, over and above whatever education they get. So education and the error term in an earnings equation are likely to be correlated –education is "endogenous".
- Angrist and Krueger (1991). Uses "quarter of birth" as an "instrument" for education.
- Individuals born in the 4th quarter of a calendar year tend to stay in school longer than do individuals born in the 1st quarter of the year – they are either in earlier grades and need longer to complete the mandatory amount of school (in many states), and/or they turn 16 later and hence are legally compelled to stay in school longer than 1st quarter individuals
- So, if birth quarter can be assumed to be "randomly" determined (or "as if" randomly determined), it can be used as an instrument for educational attainment so long as birth quarter is not *directly* related to earnings (i.e., in ways unrelated to its effect on earnings through additional education)

- Classic Example II: Does military service affect earnings?
- Omitted variables/endogeneity problems: the kinds of individuals who select into military service may have different earning potentials than individuals who do not, and these differences may be unobserved
- Angrist (1990): uses Vietnam-era draft lottery number as an instrument for military service
- People with (randomly) low lottery numbers needed to serve, people with (randomly) high lottery numbers could avoid service; so this is an *exogenously-induced* change in military service. It is also unlikely that lottery number status had a *direct* impact on earnings – why should it?
- Finding: Military service is *negatively* correlated with earnings
- Applied in political science by Stoker and Erikson (2012 *APSR*); positive effects of having a low lottery number on antiwar political attitudes, Democratic party identification, etc. via "draft vulnerability"; though not confirmed in IV analyses of effects of actual military service

"Weak" and "Strong" Instruments

- Instrumental variable analysis depends not only on finding instruments to identify equations, but on finding *good* instruments. The opposite of a "good" instrument is a "**weak instrument**".
  - A weak instrument has little to no covariation between the instrument and the independent variables, i.e., the variable it is serving as an instrument. If this covariance is zero, the estimation process will break down altogether.

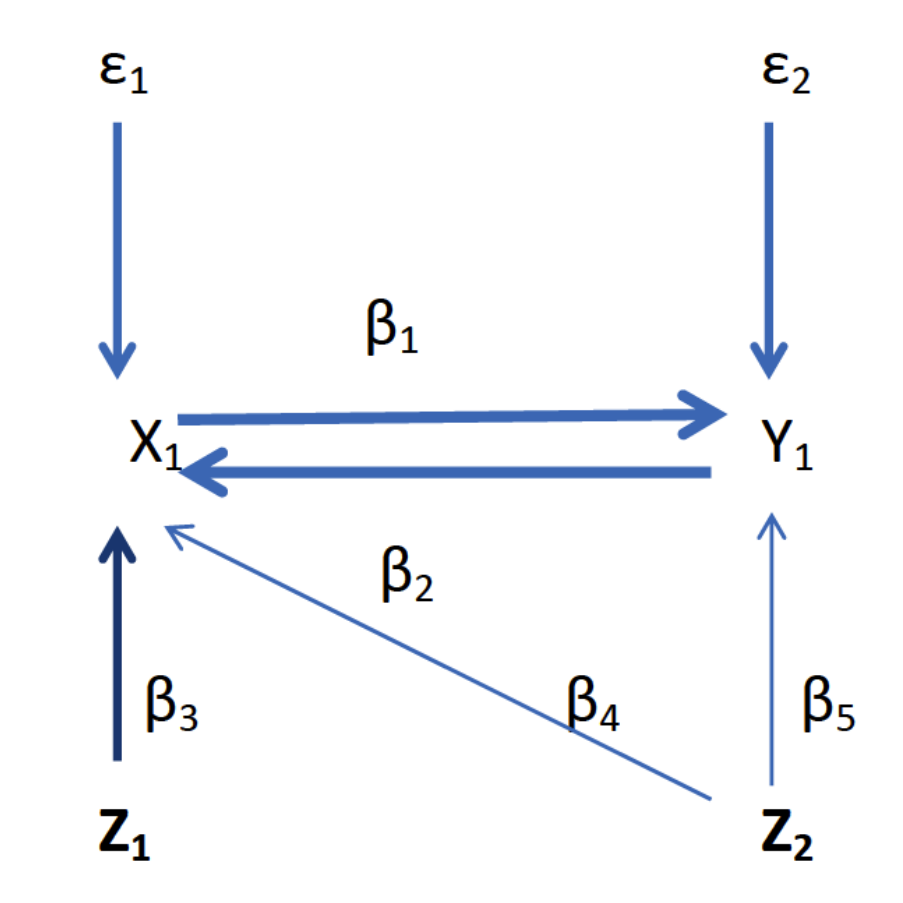
$$\sigma^2_{\hat{\beta}_{1IV}} = \frac{\sigma^2}{N\sigma_x^2 r_{XZ}^2}$$

  - As this covariance becomes smaller and smaller, the standard error for the β effect will become larger, hence increasing the chance of showing insignificant results. So *good* instruments are strongly related to the independent variable for which they are serving as instruments. (How good is "good"? See below)
  - Also can show (Woolridge p, 514-515), that a "weak" instrument will really exacerbate IV estimation problems if Z is related to the error term ε (i.e. if it is not 100% exogenous). In some cases, IV will then be worse than OLS in terms of inconsistency

Extending IV Analysis in "Just-Identified" Models

- When models are "just-identified" (with all equations having an equal number of excluded exogenous variables as endogenous independent variables), then it is straightforward to extend these procedures to handle more complex multivariate models
- Procedure is called **"Indirect Least Squares"**
- We express each endogenous variable in terms of only the *exogenous* variables in what are called the **"reduced form equations"**. We then estimate these reduced form equations (legitimately) using OLS, and we finally manipulate the reduced form regression coefficients to solve for the causal effects of interest
- IV analysis as we have been examining in the previous few slides is just a "reduced" version of Indirect Least Squares when there is only one instrument and one independent endogenous variable in an equation

Example of Indirect Least Squares



- Y_1 equation is just identified; Z_1 is the instrument for X_1
- But can't just use IV formula for β_1 since Y_1 is also caused by Z_2 and we need to estimate β_5 along with β_1
- Equation for X_1 is not identified (by the way); no excluded exogenous variables and 1 included endogenous variable (Y_1), but still useful for us

- We can express each endogenous variable in terms of the *exogenous* variables only. These are called the "reduced form" equations, and can be estimated legitimately with OLS (since exogenous variables are unrelated to their equation's error terms)

$$Y_1 = \beta_1 X_1 + \beta_5 Z_2 + \varepsilon_2$$

$$X_1 = \beta_2 Y_1 + \beta_3 Z_1 + \beta_4 Z_2 + \varepsilon_1$$

substitute for X_1 in the Y_1 equation:

$$Y_1 = \beta_1(\beta_2 Y_1 + \beta_3 Z_1 + \beta_4 Z_2 + \varepsilon_1) + \beta_5 Z_2 + \varepsilon_2$$

$$Y_1 = \beta_1\beta_2 Y_1 + \beta_1\beta_3 Z_1 + (\beta_1\beta_4 + \beta_5) Z_2 + \beta_1\varepsilon_1 + \varepsilon_2$$

$$Y_1(1 - \beta_1\beta_2) = \beta_1\beta_3 Z_1 + (\beta_1\beta_4 + \beta_5) Z_2 + \beta_1\varepsilon_1 + \varepsilon_2$$

$$Y_1 = \frac{\beta_1\beta_3}{1 - \beta_1\beta_2} Z_1 + \frac{\beta_1\beta_4 + \beta_5}{1 - \beta_1\beta_2} Z_2 + \frac{\beta_1\varepsilon_1 + \varepsilon_2}{1 - \beta_1\beta_2}$$

$$Y_1 = \frac{\beta_1\beta_3}{1-\beta_1\beta_2} Z_1 + \frac{\beta_1\beta_4 + \beta_5}{1-\beta_1\beta_2} Z_2 + \frac{\beta_1\varepsilon_1 + \varepsilon_2}{1-\beta_1\beta_2}$$

$$Y_1 = \pi_1 Z_1 + \pi_2 Z_2 + V_1$$

where

$$\pi_1 = \frac{\beta_1\beta_3}{1-\beta_1\beta_2} \text{ and } \pi_2 = \frac{\beta_1\beta_4 + \beta_5}{1-\beta_1\beta_2}$$

- So if we regress Y on both Z using OLS, we obtain the "reduced form" coefficients above
- Let's do the same for X – get the reduced form coefficients

$$Y_1 = \beta_1 X_1 + \beta_5 Z_2 + \varepsilon_2$$

$$X_1 = \beta_2 Y_1 + \beta_3 Z_1 + \beta_4 Z_2 + \varepsilon_1$$

substitute for Y_1 in the X_1 equation:

$$X_1 = \beta_2(\beta_1 X_1 + \beta_5 Z_2 + \varepsilon_2) + \beta_3 Z_1 + \beta_4 Z_2 + \varepsilon_1$$

$$X_1(1-\beta_2\beta_1) = (\beta_2\beta_5 + \beta_4)Z_2 + \beta_3 Z_1 + \beta_2\varepsilon_2 + \varepsilon_1$$

$$X_1 = \frac{\beta_3}{1-\beta_2\beta_1} Z_1 + \frac{(\beta_2\beta_5 + \beta_4)}{1-\beta_2\beta_1} Z_2 + \frac{\beta_2\varepsilon_2 + \varepsilon_1}{1-\beta_2\beta_1}$$

$$X_1 = \quad \pi_3 Z_1 + \quad \pi_4 Z_2 + \quad V_2$$

where

$$\pi_3 = \frac{\beta_3}{1-\beta_2\beta_1} \text{ and } \pi_4 = \frac{(\beta_2\beta_5 + \beta_4)}{1-\beta_2\beta_1}$$

- So if we regress X_1 on both Z_1 and Z_2 using OLS, we get the reduced form coefficients above

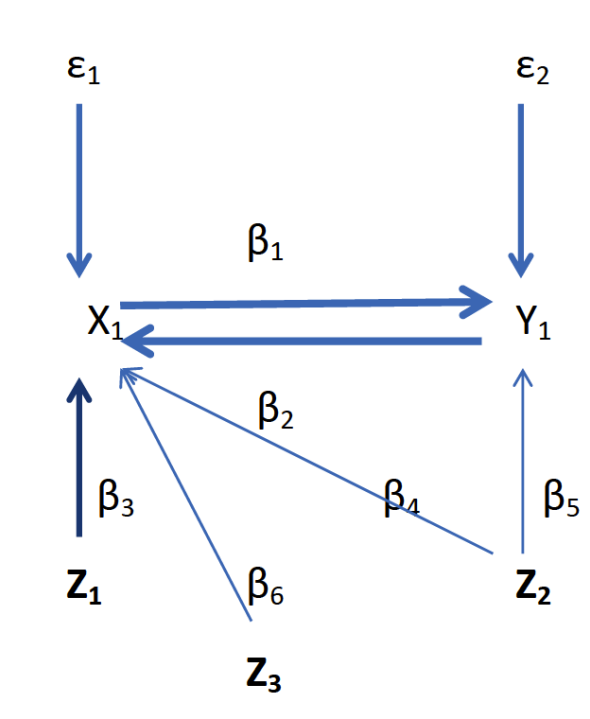
- Now, having obtained the OLS reduced form coefficients, we can manipulate them to solve for the parameters in the just-identified equation:

$$\beta_1 = \frac{\hat{\pi}_1}{\hat{\pi}_3} \text{ and } \beta_5 = \hat{\pi}_4 - \frac{\hat{\pi}_1 \hat{\pi}_2}{\hat{\pi}_3}$$

- So the effect of the endogenous variable X on Y is pretty much the same idea as in earlier IV – it is the reduced form regression coefficient of Y on the instrument Z_1 divided by the reduced form regression coefficient of X on the instrument Z_1
- Indirect Squares also gives you the other effects of other exogenous variables in a just-identified equation, in this case, β_5

Two-Stage Least Squares (TSLS)

- When equations are **over-identified**, more than one Indirect Least Squares solution for estimating model parameters is possible. In these instances, we make use of a procedure known as **Two Stage Least Squares**, which extends the logic of instrumental variables analysis discussed so far in a slightly different manner.



Here, the Y_1 equation is "overidentified" – we have two excluded exogenous variables (Z_1 and Z_3) and only one included endogenous variable

Both Z_1 and Z_3 could in principle be used as "instruments" for X_1

- Following the logic of Indirect Least Squares, we could obtain the reduced form estimates:

$$Y_1 = \pi_1 Z_1 + \pi_2 Z_2 + \pi_3 Z_3 + V_1$$
$$X_1 = \pi_4 Z_1 + \pi_5 Z_2 + \pi_6 Z_3 + V_2$$

- And we could arrive at **two** valid, consistent estimates of β_1:

$$\beta_1 = \frac{\hat{\pi}_1}{\hat{\pi}_4} \quad \text{and} \quad \beta_1 = \frac{\hat{\pi}_3}{\hat{\pi}_6}$$

- in other words, the reduced form effect of each instrument on the DV divided by the respective effect on the endogenous independent variable for which it is serving as an instrument
- How to get the "best " estimate? We could average them, but that is inefficient. Two Stage Least Squares (2SLS) is the way to combine all of the potential ILS estimates into a single "best" one

Steps in TSLS

- First stage: Regress X_1 on **all three** exogenous variables, **even those that may not affect it in its own equation. [In this example all three exogenous variables truly do affect X_1, but even if, say, Z_2 did not affect X_1 in the true causal model, we still use it in the first stage of TSLS for efficiency purposes]**
- This generates a predicted X_1 that is unrelated to the error term ε_2. Using **all** the exogenous variables generates the best possible prediction of X_1, and because this prediction is a linear combination of variables that are all unrelated to ε_2, the predicted variable will also be unrelated to ε_2.
- Call the predicted value of X_1 from the first stage $\hat{X}_1$
- Second stage: Regress Y_1 on Z_2 and $\hat{X}_1$ to generate estimates of β_5 and β_1. All independent variables in this equation are unrelated to the equation's error term ε_2, so OLS estimation is now possible.

- Two-Stage Least Squares, then, uses OLS estimation at two stages to generate estimates in over-identified models (and in just-identified models it would reduce to ILS)
- Our example, Step 1:

$$X_1 = \hat{\pi}_4 Z_1 + \hat{\pi}_5 Z_2 + \hat{\pi}_6 Z_3 + V_2$$
$$\hat{X}_1 = \hat{\pi}_4 Z_1 + \hat{\pi}_5 Z_2 + \hat{\pi}_6 Z_3$$

- Second step:

$$Y_1 = \hat{\beta}_1 \hat{X}_1 + \beta_5 Z_2 + \varepsilon_2$$

- TSLS or 2SLS now available in all canned software packages

  Stata: IVREGRESS R: IVREG

2SLS Notes

- Standard errors in the second stage will be biased using a manual two-step approach. STATA automatically adjusts the estimated standard errors in IVREGRESS, so just use those.
- Check model estimates for unrealistic results, which are common in IVREGRESS when the assumptions do not hold and/or when the instruments are poor.
- Problems exist as discussed above if there are weak instruments: large standard errors and unstable estimates. A formal test is suggested by Shea (1997): this is the F* for the effects of the instrumental variables (i.e., Z_1 and Z_3 in our case) in the first-stage equation. If the IVs taken together do not add a significant amount of explanatory power, over and above the other exogenous variables, you have problems.
- An informal rule of thumb is that you want an F*-statistic above 10.0 for the excluded instruments in the first stage of the process. You can obtain this with "estat firststage" post-estimation command

- The IV assumptions still must hold for procedure to be valid. Both the *exclusion* restriction and the *exogeneity* assumption must be justified in a given situation
- If you have an overidentified model, you can test the exogeneity assumption with the "Sargan test". The idea is to use the residuals from the second stage as dependent variables in an auxiliary regression, and, if the instruments are *truly* exogenous, they should not be related significantly to the residuals.
- Sargan: take the $\hat{\varepsilon}_2$ from our second stage Y_1 equation, and regress against all exogenous variables:

$$\hat{\varepsilon}_2 = \varpi_1 Z_1 + \varpi_2 Z_2 + \varpi_3 Z_3 + \zeta$$

- Under the null that the instruments are uncorrelated with the error term, $N*R^2$ from this equation ~ as chi-squared with *r* degrees of freedom (equal to the number of excess instruments)

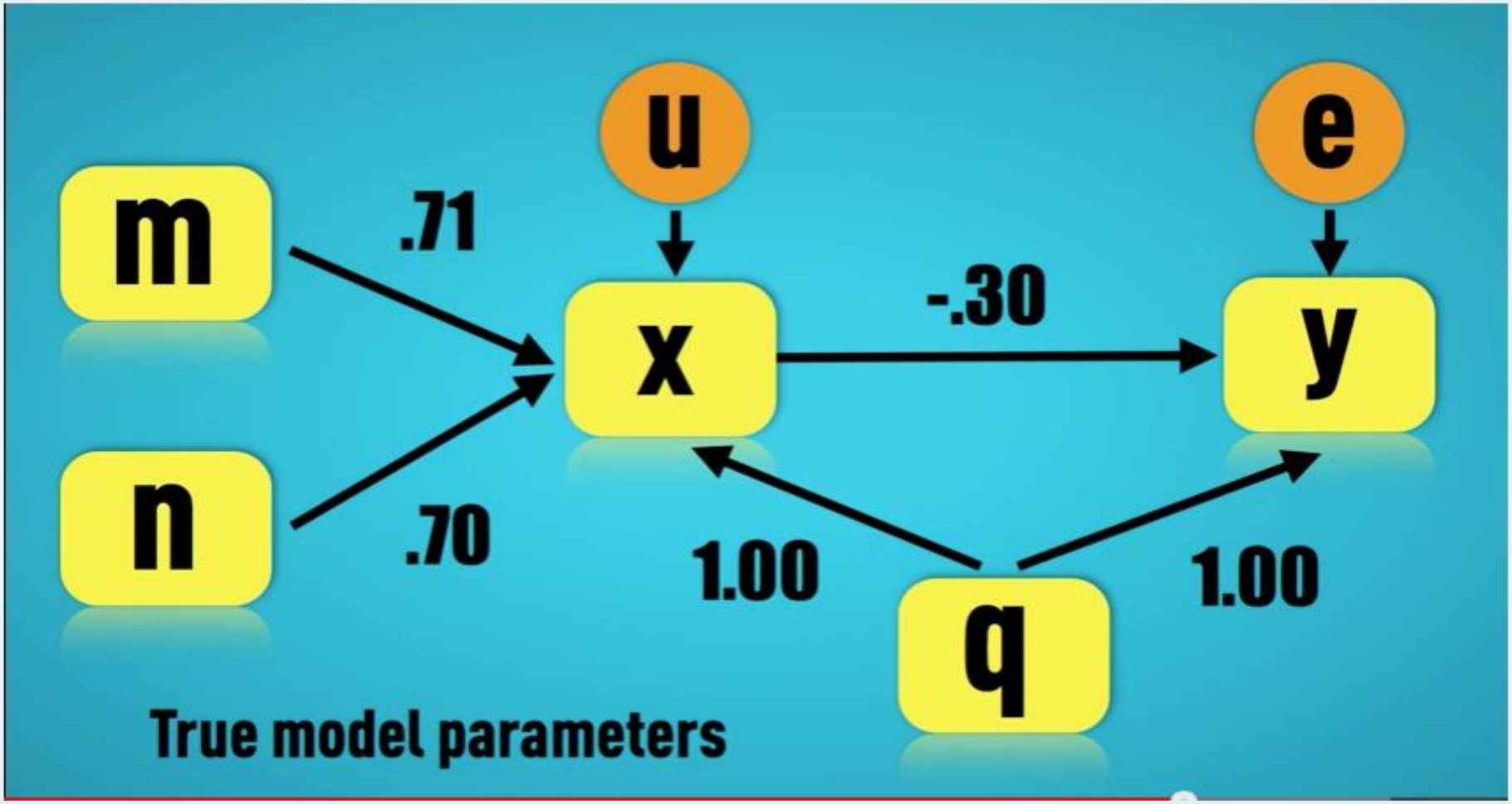
- Further notes on 2SLS:
- Can test for heteroskedasticity using "ivhettest", another add-on module, and then correct with robust standard errors
- Can test for whether 2SLS gives you significantly better estimates than OLS with the "**Durbin-Wu-Hausman**" test. The idea here is that if there is no endogeneity problems in the model, you should be using OLS because it is more efficient than 2SLS. So unless the coefficient really differ significantly, you are better off sticking with OLS. This test is a summary test of the difference in 2SLS and OLS estimates. You can obtain this with "estat endogenous" after running IVREGRESS
- Be careful running 2SLS and associated instrumental variable procedures: theory, theory, theory should be your guide!!!
- More advanced IV models, along with other ways of estimating models with omitted variable biases due to unobservables, will be discussed in unit 4

Alternative Method: "Two-Stage Residual Inclusion"

- A similar method for handling endogeneity is called "two stage residual inclusion" (TSRI), a type of "control function" regression
- There is also two stages in this procedure: in the the first stage, the endogenous variable is regressed on all exogenous variables, including the instrumental variable(s), and predicted values and a first-stage residual are generated
- What is this residual? It is the portion of the endogenous variable that **cannot** be explain by the exogenous variables in the system – in other words, it is a measure of the **endogenous** portion of the endogenous variable!
- Second stage: "*Control*" for the endogeneity by including the first stage residual along with all the other independent variables. Ingenious!
- Terza *et al.* "Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling" *Journal of Health Economics* 27 (2008): 531–543.

Example of 2SLS

John Antonakis "Endogeneity: An inconvenient truth" Podcast, Faculty of Business and Economics, University of Lausanne, Switzerland (also on Youtube)



- Generates 10000 cases on variables M, N, X, Q, Y according to the "true model" in previous slide ("antonakis.2slsdata.dta")
- Assume that you omit Q from consideration:
  - Omitted variable bias such that U and e are correlated
  - X is now "endogenous"
  - Regression of Y on X yields:

```
. reg y x

      Source |       SS       df       MS              Number of obs =   10000
-------------+------------------------------           F(  1,  9998) =   19.15
       Model |  32.146475     1   32.146475             Prob > F      =  0.0000
    Residual |  16781.5228  9998  1.67848798             R-squared     =  0.0019
-------------+------------------------------           Adj R-squared =  0.0018
       Total |  16813.6693  9999  1.68153508             Root MSE      =  1.2956

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .0327032   .0074728     4.38   0.000      .018055    .0473514
       _cons |  -.0098596   .0129559    -0.76   0.447     -.0352557    .0155366
------------------------------------------------------------------------------
```

- But fortunately, nature has provided us with 2 instruments: M and N!
- By construction, they are related to X, unrelated to U and e, and have no direct effect on Y!

```
Instrumental variables (2SLS) regression          Number of obs =   10000
                                                  Wald chi2(1)  =  416.27
                                                  Prob > chi2   =  0.0000
                                                  R-squared     =       .
                                                  Root MSE      =  1.4083
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |  -.2859296   .0140143   -20.40   0.000    -.3133971   -.2584621
       _cons |  -.0064556   .0140838    -0.46   0.647    -.0340594    .0211482
------------------------------------------------------------------------------
Instrumented:  x
Instruments:   m n
```

- True effect is recovered with 2SLS!

TRUE EFFECTS OF M AND N ON X ALSO RECOVERED

First-stage regressions

Number of obs	10000
F(2, 9997)	2529.00
Prob > F	0.0000
R-squared	0.3360
Adj R-squared	0.3358
Root MSE	1.4130

x	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
m	.7077401	.0139256	50.82	0.000	.6804432	.735037
n	.7051994	.0140834	50.07	0.000	.6775932	.7328057
_cons	.0031191	.0141303	0.22	0.825	-.024579	.0308173

```
. estat overid

  Tests of overidentifying restrictions:

  Sargan (score) chi2(1) =  .003151  (p = 0.9552)
```

```
. estat endogenous

  Tests of endogeneity
  Ho: variables are exogenous

  Durbin (score) chi2(1)          =   920.05  (p = 0.0000)
  Wu-Hausman F(1,9997)            =  1012.97  (p = 0.0000)
```

Final Endogeneity Issue: Measurement Error

- Indicators of variables may contain error, in that the value that is assigned to a given unit is not the "true" value for that variable for that unit. Errors in variables may be:
  - **Systematic**, in which case we may say that the observed indicator is always off from the true value in one direction or the other, or that some other variable is also systematically influencing indicator, aside from the true variable of interest. In that case, we say the measure is not a **valid** indicator.
  - **Random**, in which case the observed indicator is sometimes higher or lower than the true value depending on random factors in the measurement process, such as (among other things):
    - poor record keeping
    - individual coder decisions (e.g. the people at Freedom House deciding on a 2 versus a 3 for some country's civil liberties index).
    - ambiguous questions in surveys
    - mood or other transient factors in the interview or observation process
    - scaling of variables (e.g. where does "3.55" attitude go on a 1-2-3-4-5 scale? Most of the time to 4, but some of the time no doubt to 3).

- Random errors lead indicators to be *unreliable*
- **ALL SOCIAL SCIENCE MEASURES ARE UNRELIABLE TO SOME EXTENT!!!! SOME ARE ALSO INVALID, THOUGH THIS PROBLEM IS MUCH MORE DIFFICULT TO DETECT AND CORRECT.**
- We can correct for reliability problems with some of the same techniques we have considered thus far today; alternative solutions exist with techniques covered in courses on measurement and/or longitudinal analysis

Consequences of Measurement Error for OLS

- Different consequences for OLS if Y or X contains random error
  - In Y: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

    $\text{and } y_i^* = Y_i + u_i$

    $\text{then } y_i^* - u_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

    $\text{and } y_i^* = \beta_0 + \beta_1 X_i + (\varepsilon_i + u_i)$
  - so OLS overestimates error variance, lower R-squared and larger standard errors!
  - In X: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

    $\text{and } \mathrm{x}_i^* = X_i + v_i$

    $\text{then } \mathrm{Y}_i = \beta_0 + \beta_1(\mathrm{x}_i^* - v_i) + \varepsilon_i$

    $\text{and } \mathrm{Y}_i = \beta_0 + \beta_1 \mathrm{x}_i^* + (\varepsilon_i - \beta_1 v_i)$
  - So OLS overestimates error variance *and* E(Xε)≠0 again (since v and x^* are related). So OLS produces biased, inconsistent and inefficient estimates!!!

Measurement Error, Reliablity and the Attenuation of OLS Estimates

- Another way to look at this: the OLS estimate of β with the "fallible' measure x* is:

$$\beta_{OLS} = \frac{Cov(Yx^*)}{Var(x^*)}$$

- With a little substitution, we can arrive at:

$$\beta_{OLS} = \frac{Cov(\beta_{TRUE} X + \varepsilon, x^*)}{Var(x^*)} = \frac{Cov(\beta_{TRUE} X_i x^*) + Cov(x^*, \varepsilon)}{Var(x^*)} = \beta_{TRUE} \frac{Cov(x^* X)}{Var(x^*)}$$

- We can further show that the numerator is equal to Var(X), so:

$$(5) \qquad \beta_{OLS} = \beta_{TRUE} \frac{Var(X)}{Var(x^*)}$$

- **Conclusion: Unless there is no measurement error in *x*, the (bivariate) OLS estimate of β will be less than the true value, and will be *attenuated* by the factor** $\frac{Var(X)}{Var(x^*)}$
- We call this factor, which is the ratio of "true score variance to observed score variance," the **reliability** of x* (denoted as ϱ_{xx}). It is the proportion of the observed variance in x* that is composed of the latent true score and *not* the measurement error *v*. So the OLS bivariate β equals the true β multiplied by the reliability of *x*.

- Higher reliability means that the observed score is closely related to the true score and hence the attenuation of the OLS regression coefficient will be small; lower reliability means greater random noise in the indicator and consequently greater attenuation of the OLS regression coefficient in the bivariate case.

Notes on Reliability

- The direction of bias due to measurement error in explanatory variables is *always* downward in the bivariate case; in the multivariate case it may be downwards or upwards, depending on the amount of measurement error in particular variables and their intercorrelations.
- We can also take the measurement error equation for x*, square both sides and take expectations to yield:

$$Var(x^*) = Var(X) + Var(v)$$

- which expresses the variance in a fallible indicator as composed of two parts: the true score variance and the error variance. So the reliability of x is the proportion of its variance being "true score" variance – it is akin to R^2 in that we can say that the higher the reliability, the lower the error variance in an indicator and thus the greater amount of variance 'explained" by the latent true score X.

Correcting for Measurement Error: Instrumental Variables

$$Y_i = \beta_0 + \beta_1 x_i^* + (\varepsilon_i - \beta_1 v_i)$$

- Problem: x* is correlated with the error term, via its relationship with *v*
- Solution: Find an *exogenous* instrument Z for x* such that
  - the instrument affects x*, but
  - the instrument does not affect Y except via x* (the exclusion restriction)
- If you can find such a Z, then multiply and take expectations:

$$Cov(ZY) = Cov(Z\beta_0) + \beta_1 Cov(Zx^*) + Cov(Z\varepsilon) - \beta_1 Cov(Zv)$$

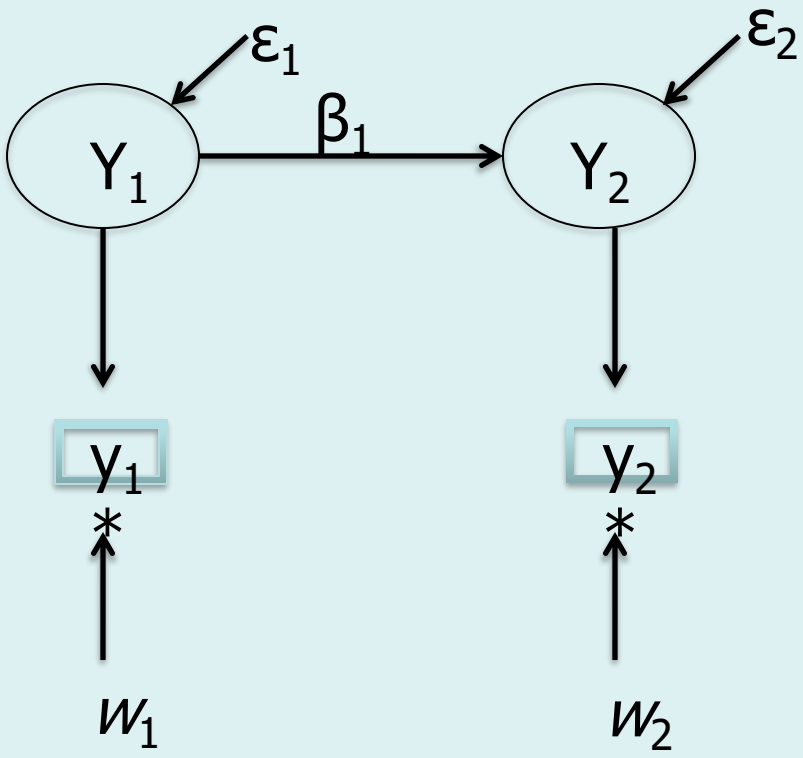
$$\beta = \frac{Cov(ZY)}{Cov(Zx^*)}$$

- This is exactly the same solution as in earlier IV models!

- So an unbiased estimate of β_1 can be obtained by dividing the covariance of the dependent variable with the instrument by the covariance of the independent variable with the instrument. In models with more than one instrument, we can use 2SLS procedures But: same problems as in earlier models holds here. Hard to find instruments that are relatively highly correlated with the error-filled independent variable x* **and** that have no direct effect on (or from) Y **and** are uncorrelated with the disturbance ε and the measurement error term *v*.
- One possibility: use a second x* indicator of X as an instrument for the first. This is the idea behind a procedure called "Model-Implied Instrumental Variables" (MIIV), a method that is growing in popularity for estimating models with measurement error.
  - See Bollen, K. A. *et al.* (2022). "An introduction to model implied instrumental variables using two stage least squares (MIIV-2SLS) in structural equation models (SEMs)" *Psychological Methods,* 27(5), 752–772.
- Depends on all IV assumptions made so far *and* that the measurement errors *v* for both indicators are unrelated, a possibly dubious assumption.

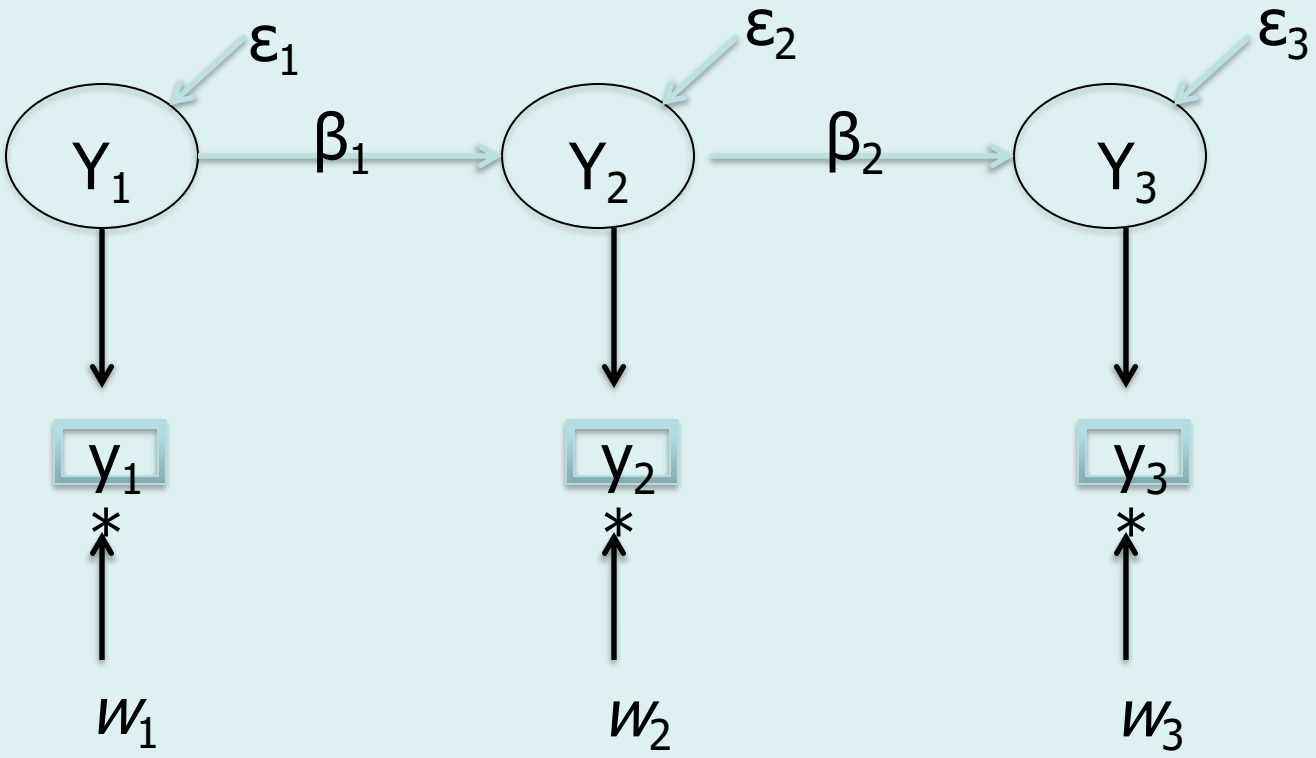
- The MIIV idea of handling measurement error by using *multiple indicators* of a latent construct is a good one that is also incorporated in Structural Equation Modeling (SEM) more generally.
- We solve the problems of measurement error with additional information. IVs and MIIVs do it one way, multiple indicator latent variable models with maximum likelihood estimation is another way, and longitudinal models of multiple and/or single indicator latent variables are yet another way
- The final slides show different kinds of longitudinal models that attempt to correct for measurement error. These can be estimated in structural equation programs like STATA, R (LAVAAN), AMOS, MPLUS, EQS, or LISREL
- The more waves of observation and the more indicators you have, the more flexibility you have in modeling both the measurement and structural processes
- These models are covered in courses on longitudinal analysis

Example: Two Wave "Y-Only" Model with Measurement Error

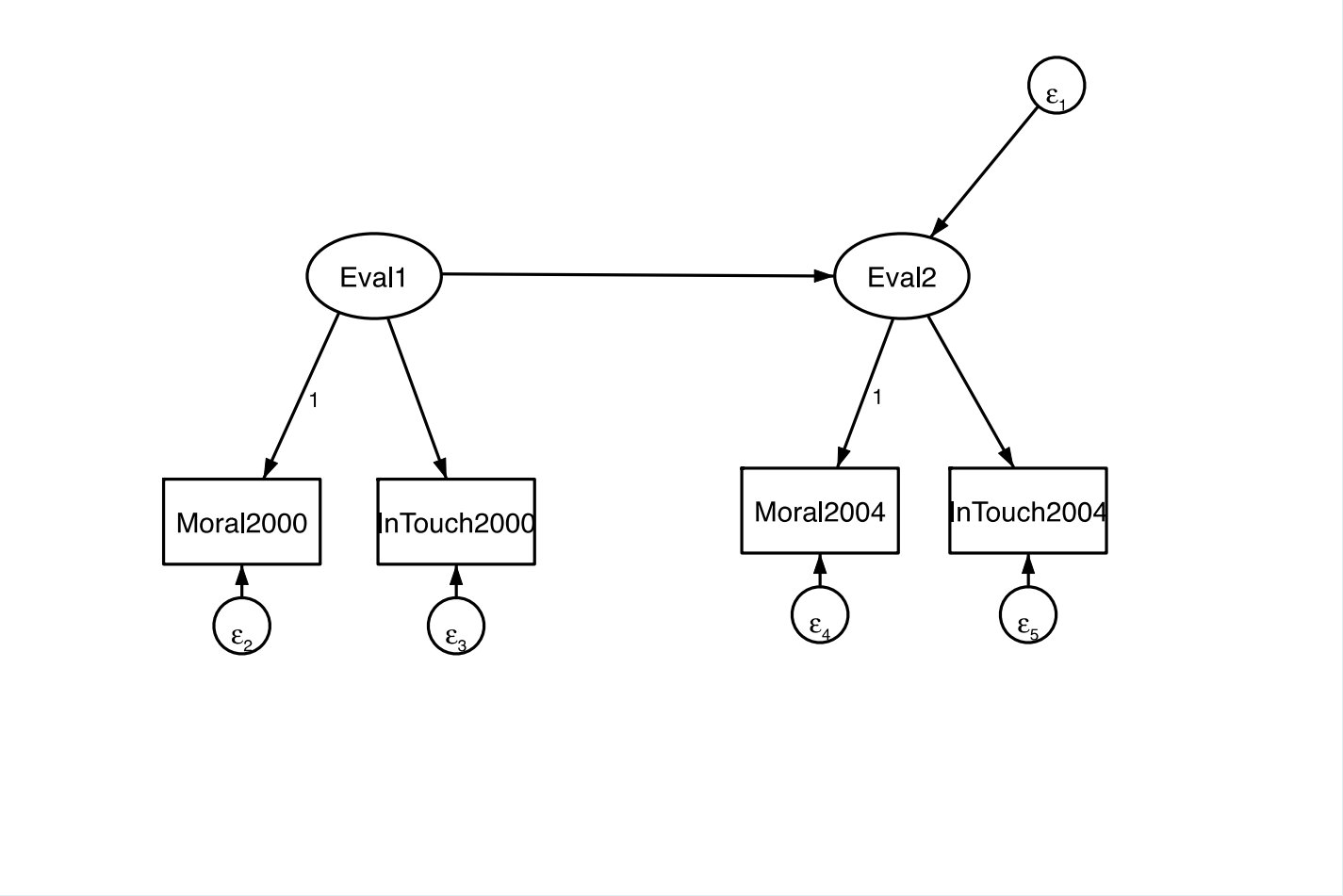


NOTE: SEM convention to use circles to represent "latent" variables, squares to represent "observed" variables

Three Wave, Single Indicator Model



Is this model identified? YES, but only with "equality constraints" on the measurement error parameters!



MEASUREMENT MODEL FIGURE 3:
Two Wave, Two Indicator Model with NES Panel Data 2000-2004