ECON 7130 - Microeconomics III

Spring 2016
Notes for Lecture #6

Today:

- Endogeneity
- Why IV?
- How to set up 2SLS
- Testing IV
- Examples of IV

Endogeneity

- Endogeneity
  - The major challenge in the social sciences
  - Due to lack of ability to use lab for many questions, hard to hold "all else equal" (control treatment)
  - As a result, much work biased since it suffers from endogeneity
- What is endogeneity?
  - When an independent variable is correlated with the error term ($corr(X, \varepsilon) \neq 0$)
  - It's a loop of causality between the dependent and independent variable (e.g., quantity demanded determines price, but price determines quantity demanded)
  - Results in bias in estimation of econometric model
  - Sources of endogeneity:
    1. Omitted Variable Bias
    2. Measurement Error
    3. Simultaneity
- Omitted Variables
  - Leads to OVB (omitted variable bias)
  - Fixes
    - ∗ Try to find proxies for omitted variables
    - ∗ Use panel and try to difference out OVB
    - ∗ Use Instrumental Variables
  - Selection Bias
    - ∗ A form of OVB - don't observe factor that causes people to selection into treatment
    - ∗ To fix - Heckman Selection Model
      - · Instrument for selection with Inverse Mills Ratio
      - · IMR related to prob selected into treatment
      - · Need excluded instrument that affects selection, but not outcome conditional on selection (often hard to find), else rely on functional form assumptions for identification
- Simultaneity
  - e.g., Classic supply and demand model

  - Economic outcomes determined jointly
  - To identify causality, need exogenous shifters (i.e., instruments)

- Instrumental Variables:
  - Instrumental variables (IV) regression is designed to control for endogeneity
  - IV regression is designed to correct the estimates in the main equation by finding exogenous shifters for the endogenous variable(s)
  - e.g. supply shifters to uncover demand curve

- What are instruments?
  - The excluded instruments (z_i) are variables used to explain a variable we suspect of being endogeneous and which are exogenous with respect to the main equation

$$cov(z_i, u_i) = 0 \tag{1}$$

$$cov(z_i, x_i) \neq 0 \tag{2}$$

  - (1) is the **exclusion restriction**, it determines if instrument is valid (note that u_i is the error term from the main equation)
  - (2) is the relevance criteria, determines if the instrument is informative to the independent variable x
  - These are tough criteria
  - Note that the term excluded instrument comes from the fact that this instrument is excluded from the main equation. Technically speaking, all exogenous variables are instruments (though here and elsewhere people are often loose with language and use the term "instruments" to specifically mean the excluded instruments).
  - Challenge: to find variables correlated with the endogenous variable but uncorrelated with the part of the error term that is due to the unobserved heterogeneity
  - Rule of thumb: a good instrument should correlate with the key independent variable, but should only affect the main equation dependent variable through this key independent variable

- How to estimate IV models:
  - A few ways:
    - 2SLS
    - 3SLS (e.g., simultaneous equations model)
    - GMM (see http://www.soderbom.net/lec2n_final.pdf)
    - LIML (limited info max likelihood)
  - Here, we'll focus on 2SLS - things generalize, but easiest to see intuition here (and really all are related and can be viewed as GMM estimators)
  - Need at least as many instruments as have endogenous variables
    - If number of instruments = number endogenous variables, the model is **just identified**
    - If number of instruments > number endogenous variables, the model is **overidentified**

- 2SLS model
  - Assume that we want to study y_i using x_{1i} and a number of controls x_{ji} where $j \in 2, 3, ..., k$
  - Assume x_{1i} to be endogenous
  - We have n instruments Z_{ni} useful for predicting x_{1i}

- The 2sls model becomes

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_j x_{ji} + \varepsilon_i$$
$$x_{1i} = \pi_0 + \pi_n z_{ni} + \pi_j x_{ji} + v_i$$

- Since $cov(z_{ni}, \varepsilon_i) = 0$ then it must be that $cov(\pi_0 + \pi_n z_{ni}, \varepsilon_i) = 0$
- In your instrumentation (regression against the endogenous variable), you also include the remaining explanatory variables → so $corr(v_i, x_{ij}) = 0$
- We call the instrumental variables z_{ni} the excluded instruments, because they do not appear in the main equation explaining y_i
- First run the regression against the endogenous variable (first stage) and calculate the predicted value, which we denote by $\hat{x_{1i}}$
- Then use the predicted $\hat{x}_{1i}$ rather than the observed x_{1i} in the main (regression) equation (second stage)

$$y_i = \beta_0 + \beta_1 \hat{x}_{1i} + \beta_j x_{ji} + \varepsilon_i$$

- Note that $corr(\hat{x}_{1i}, \varepsilon_i) = 0$ by construction
- We have removed the element of the endogenous variable that as correlated with the error term
- What is left is the effect of the predicted value of the endogenous variable on the outcome y_i, without bias from endogeneity
- Differences between OLS and IV
  - The OLS estimate is: $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
  - The IV estimate is: $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}$
- Interpreting IV
  - Generally, IV regression follows the same conventions as OLS.
  - Here the test statistics of the parameters can be shown to follow a standard normal distribution - the test statistics are hence referred to as z scores and compared to the standard normal distribution and not t-distributed as is the case in OLS
  - The R2 in IV is less useful since it in fact can be negative; the Residual Sum of Squares can be higher than the Total Sum of Squares in IV - do not interpret it.
- In Stata:
  - Can do 2SLS by hand - `reg x1 z x2 x3` → `predict x1hat` → `reg y x1hat x2 x3` (but in this case, need to correct standard errors in resulting in second stage)
  - Or, `ivregress` - default is 2SLS, but can also do with LIML or GMM (not sure when those advantageous - maybe for cluster/robust SE?)

- Things to check
  1. That the 'suspected' explanatory variable indeed is endogenous
  2. That you do not have weak instruments
  3. That instruments are valid (i.e., exogenous, an overidentication test)
- Testing endogeneity (Durbin-Wu-Hausman test)
  - May be able to see endogeneity by just looking at coefficients of IV vs OLS regressions - if OLS differ significantly, then biased
  - But there are more precise tests
  - To test:
    - Regress suspected endogenous variable on the excluded instruments and controls from the main regression equation: $x_{1i} = \pi_0 + \pi_n z_{ni} + \pi_j x_{ji} + v_i$

        - Get the residuals form this regression, v_i
        - Regress the variable of interest on the suspect variable, other controls, and the residuals from the regression above: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_j x_{ji} + \gamma v_i + \varepsilon_i$
        - Test the significance of the coefficient on the residuals from the first stage (F-test)
            - A small p-value (large F stat) implies that OLS is not consistent - i.e. that the suspect variable is endogenous
    - If no endogeneity, use OLS b/c more efficient (lower s.e.)
    - In Stata:
        - Assume x1 potential endogenous variable, z instrument
        - `reg x1 z x2 x3`
        - `predict x1_res, res`
        - `reg y x1 x2 x3 x1_res`
        - `test x1_res` (this is the DWH test)
    - Alternative: Hausman test
        - Run OLS, store results
        - Run IV, store resuls
        - Use `hausman` to test difference in coefficients

- Testing weak instruments
    - As pointed out by Bound, Jaeger, and Baker (1993; 1995), the "cure can be worse than the disease" when the excluded instruments are only weakly correlated with the endogenous variables.
    - IV estimates are biased in same direction as OLS, and Weak IV estimates may not be consistent.
        - See Chao and Swanson (2005) for a comparison of consistency results for related estimators.
    - With weak instruments, tests of significance have incorrect size, and confidence intervals are wrong.
    - Many tests for weak instruments have been proposed
    - No clear consensus on best approach for evaluation
    - Generally we consider the significance of the instruments in the first stage equation
    - Significance suggest instruments are note weak since the criteria is: $cov(z_i, x_i) \neq 0$
    - One formal test you might want to do is the Stock and Yogo test
        - Gives you some idea how strong your identification is (i.e., is the bias remaining in the IV estimate 20% of that from OLS? More?)
        - "Critical values" automatically calculated with `ivreg2` command. These are the values for an F-test of the first stage coefficients on the instruments

- Testing instruments valid (Sargan Test)
    - If the instruments are poor in the sense that they are not exogenous to the main equation, we obtain biased results (they are said not to be 'valid')
    - We can check this with a Sargan over identification test
    - This is an over identification test - meaning you need more excluded instruments than endogenous variables
    - It is only a valid test if at least one of the instruments is exogenous (it assumes that one is) - if all are endogenous, then problem
    - Follow the following steps:
        1. Estimate the 2SLS IV regression model - Extract residuals from the second stage

        2. Regress these residuals on all exogenous variables (i.e., both the included and excluded instruments) and extract R^2
        3. Calculate nR^2 which is χ^2 distributed
        4. Compare the value with the critical value in the chi-square table with degrees of freedom equal to # instruments less # endogenous variables
    - If the statistics (nR^2) exceeds the critical χ^2 value, we can conclude that the instruments are not exogenous and hence invalid.
    - They are correlated with the error term and hence have some explanatory power in the main equation.
    - Even when the overidentification test suggests that the instruments are valid, we should be very careful: The test assumes that one instrument is valid.
    - If all instruments do not fulfill the criteria $Cov(z_i, u_i) = 0$, then the test might suggest that the instruments are valid, even when they are not!!
    - Often hard have over identified model (because it's hard enough to come up with enough excluded instrumental variables to have a just identified model)
    - In Stata:
        * Use `ivreg2` (ins tread of `ivregress`) and it'll spit out Sargan test results automatically if over identified model
        * Lower Sargan stat (higher p-value) means that instruments exogenous

- Informal tests
    - Use raw correlations:
        * Endogenous variable and excluded instruments (want it to be larger)
        * Dependent variable and excluded instruments (want it to be zero)
    - Have a story
        * Why are instruments exogenous?
        * How do they drive the endogenous variable?
    - These are often the best you have if model is exactly identified

IV: Example 1, Angrist and Krueger, "Does Compulsory School Attendance Affect Schooling and Earnings?" (*QJE*, 1991):

- Question: What are the returns to education? I.e. has does additional schooling affect earnings later in life?

- Problems with answering this question:
    - There are omitted variables correlated with earnings and years of schooling
    - e.g. cognitive ability

- A natural, natural experiment: compulsory schooling laws
    - Quarter of birth correlated with years of schooling completed because of mandatory schooling, and institutional rules regarding ages start and can dropout of school
        * Most states require students to enter school in the calendar year in which they turn 6. Hence, those born late in the year are young for their grade.
        * Students are required to remain in school only until their 16th birthday.
        * So someone born early in the year may drop out in grade G while someone born late may not dropout until grade $G + 1$.

  - Quarter of birth random
- Data:
  - 1% sample of the 1970 and 1980 Census
- Basic Model:
  - 2SLS:
    1. $E_i = X_i\pi + \sum_c Y_{ic}\delta_c + \sum_c \sum_J Y_{ic}Q_{ij}\theta_{jc} + \varepsilon_i$
    2. $lnW_i = X_i\beta + \sum_c Y_{ic}\xi_c + \rho E_i + \mu_i$
  - Where, E_i are years of education for person i (which is endogenous because of correlation between omitted variable, education, and earnings)
  - Y_{ic} is a dummy variable for person i being born in year c
  - Q_{ij} is a dummy variable for person i being born in quarter j
  - X_i are a vector of demographic controls
  - W_i is the weekly wage of person i
- The coefficient of interest is ρ, the return to education
- Identification:
  - Difference-in-Differences: comparing differences in earnings across those born in different quarters and across states with difference compulsory schooling laws
  - 2SLS, IV approach - instrumenting for education with quarter of birth
  - Key assumption:
    - That compulsory schooling laws are binding for some people (if never bind, no help in identification)
    - That quarter of birth is exogenous to earnings
      - May not be true if rich/poor parents have different timing of births (they cite evidence not a concern at the time - maybe more so now that more studies out)
      - May not be true if students who are older are more mature and perform better - the authors note that this would bias downward the effects they find
- Results:
  - Returns to education between 6 and 7 percent (that is effect of one more year of school)
  - OLS and 2SLS results very similar
  - Sargan test mostly concludes valid instruments (over identified because multiple dummies for quarter of birth)
  - Never show first stage regression
  - Findings represent a Local Average Treatment Effect (LATE)
    - Valid only for an additional year of schooling by those affected by compulsory schooling laws - not everyone!

IV: Example 2, Knittel, Hughes, and Sperling, "Evidence of a Shift in the Short-Run Price Elasticity" , (*The Energy Journal*, 2008):

- Classic example of simultaneity in market equilibrium
- Question: What is the short run elasticity of demand for gasoline? Has it changed over time? (so want to trace out demand curve)

- Natural experiments used as instruments:
  - Venezuelan oil strike (December 2002 - March 2003)
  - The (Second) Iraq War (March - November 2003)
  - Hurricane Katrina (September 2005 - January 2006)
  - Claim that high prices in the 1975-1980 period result of supply forces (e.g. OPEC oil embargo)
- Data: EIA, aggregates on gasoline production and prices for 1975-2006
  - Mainly interested in comparing two periods with similar increases in gasoline prices: 1975-1980 and 2001-2006
- Basic Model:
  - Do OLS, but price may be a function of demand - they are determined simultaneously
  - Show how could have movement in supply and demand that results in higher price and less quantity
  - Simultaneous Equation models estimated via 2SLS:
    1. $lnP_{jt} = \beta_0 + \beta_1 lnY_{jt} + \beta_2 VZ + \beta_3 IQ + \beta_4 USA + \varepsilon_j + \varepsilon_t + \varepsilon_{jt}$
    2. $lnG_{jt} = \gamma_0 + \gamma_1 lnP_{jt} + \gamma_2 lnY_{jt} + u_j + u_t + u_{jt}$
  - Where, P_{jt} is average retail price of gasoline in month j of year t
  - G_{jt} are gallons of gasoline sold in month j of year t
  - Y_{jt} = real, per capital disposable income
  - VZ, IQ, and USA are dummies turned on if in the months of the production disruptions
  - Month and year dummies are included to account for seasonality and annual fluctuations in demand for gasoline
  - Not that log-log specification means that coefficients measure elasticities
- Identification:
  - Need supply shocks that affect price, but not be correlated with unobserved demand shocks for gasoline
  - DRAW how shifting supply traces out demand curve
  - Hard to find such instruments - so only do for 2001-2006 period, arguing earlier period have supply shocks (so OLS ok there)
    - Making sure that difference not do to demand shocks in later period - robustness test
- Results:
  - Short run elasticities are much smaller in the 2001-2006 period than the 1975-1980 period
  - Income elasticities are the same across periods