Applied Econometrics (MSc.)
Lecture 3
Instrumental Variables Estimation - Theory

Yonas Alem

Department of Economics
University of Gothenburg

December 4, 2014

Why IV estimation?

- So far, in OLS, we assumed independence. $E\{x_i\varepsilon_i\} = 0$.
- In other words, all the explanatory variables are exogenous.
- There are a number of cases in economics where this assumption is unrealistic.
- When a variable is endogenous, the error term will be correlated with the explanatory variable. Thus, OLS is no more unbiased and inconsistent.
- The linear model no longer corresponds to a conditional expectation or a best linear approximation
- Many reasons for contemporaneous correlation between the error term and one or more of the X variables.

- A regression equation may contain a lagged dependent variable as one explanatory variable.
- Common in Labor Economics (unemployment duration models), Development Economics (poverty and consumption dynamics), and Agricultural Economics (technology adoption).
- Let the regression equation be given by

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 y_{t-1} + \varepsilon_t \tag{1}$$

- suppose ε_t follows a MA(1) process given as

$$\varepsilon_t = \rho \varepsilon_{t-1} + \upsilon_t \tag{2}$$

- Rewriting the model as

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 y_{t-1} + \rho \varepsilon_{t-1} + \upsilon_t$$

- This implies

$$y_{t-1} = \beta_1 + \beta_2 x_{t-1} + \beta_3 y_{t-2} + \varepsilon_{t-1}$$

- The lagged dependent variable and the error term are correlated and OLS will be biased and inconsistent!
- The possible solution is an IV technique.

- measurement error in one or more of the explanatory variables leads to correlation with the error term
- suppose the relationship:

$$y_t = \beta_1 + \beta_2 w_t + \nu_t$$

- you can think of y_t as household savings, and w_t as disposable income
- where the error term has a zero mean and finite variance

- $E\{\nu_t|w_t\} = 0$ implying that $E\{y_t|w_t\} = \beta_1 + \beta_2 w_t$
- If there is measurement error, $x_t = w_t + u_t$.
- Even under a set of simplifying assumptions such as
  1. $u_i \backsim (0, \sigma_u^2)$,
  2. u_i is independent of ν_i, and
  3. The measurement error is independent of the underlying true value w_t
     - Estimating the equation $y_t = \beta_1 + \beta_2 x_t + \varepsilon_t$ using OLS will result in biased and inconsistent parameter estimates
- Note: $\varepsilon_t = \nu_t - \beta_2 u_t$

- One of the most common cause of endogeneity
- An omitted variable (which is captured by the error term) is correlated with one or more of the explanatory variables
- Also arises from unobservable omitted factors correlated with the explanatory variable(s)
- Consider an individual wage equation given by

$$y_i = x_{1i}'\beta_1 + x_{2i}\beta_2 + u_i\gamma + \nu_i$$

- where: y_i is log wage of an individual, x_{1i} is a vector of individual characteristics,and x_{2i} refers to years of schooling.
- Let the variable u_i represent ability
- Obviously, ability and years of schooling will be correlated $\Longrightarrow$ OLS will be biased and inconsistent
- Assuming for instance that $E(x_i, u_i) > 0$, $\Longrightarrow$ OLS overestimates the returns to schooling!
- We'll see an application on this in the next section

- A situation where not only x_i has an impact on y_i, but also y_i an impact on one or more of the x.
- Reverse causality arises when the two variables (x_i and y_i) are determined simultaneously.
- A number of examples in Macro-economics where one needs a system of equations to determine endogenous variables.
  - E.g: Demand and Supply.
- A classic example: the simple Keynesian consumption function - with a closed economy and no government.

- Assume a closed economy with no government
- Let the aggregate consumption function be given by $C_t = a_t + b\,Y_t + \varepsilon_t$
- Where $t = 1, ..., T$ years, and $0 < b < 1$
- Aggregate income will be determined by the identity. $Y_t = C_t + I_t + \varepsilon_t$
- Where I_t represents private investment - assumed exogenous.

- It is easy to show in the consumption function that $\text{Cov}(Y_t, \varepsilon_t) \neq 0$
- What will be the change in consumption for a unit of change in income?
- Estimating the equation using OLS will result in inconsistent estimates because consumption and income are endogenous.
- One needs to solve for the reduced form equations

$$Y = \frac{a}{1-b} + \frac{1}{1-b}I + \frac{1}{1-b}\varepsilon$$

$$C = \frac{a}{1-b} + \frac{b}{1-b}I + \frac{1}{1-b}\varepsilon$$

- Which can be written in more general form as

$$Y = \pi_1 + \pi_2 I + \upsilon_1$$

$$C = \pi_1 + \pi_2 I + \upsilon_2$$

- OLS can be used to estimate the equations separately. These coefficients (may be) used to estimate b depending on *identification*.

- More specifically, one needs to revert to other methods like, IV (instrumental variable), ILS (Indirect Least Squares), 2SLS(Two-stage least squares - a special case of the IV technique), LI/ML (Limited Information, Maximum Likelihood) methods, depending on identification. (We don't discuss these in detail in this lecture).

- Consider the linear wage model

$$y_i = x_{1i}'\beta_1 + x_{2i}\beta_2 + \varepsilon_i \tag{3}$$

- To make the conditional expectation (the best linear approximation) of y_i given x_{1i} and x_{2i}, we needed to impose

$$E\{\varepsilon_i x_{1i}\} = 0 \tag{4}$$

and

$$E\{\varepsilon_i x_{2i}\} = 0 \tag{5}$$

- If not, the model no longer corresponds to $E\{y_i|x_{1i}, x_{2i}\} \Longrightarrow$ OLS will be biased and inconsistent
- In the above wage equation, "ability" or "intelligence" (which is unobserved and hence included in ε_i) would be correlated with "education"

- In such a case, $E\{\varepsilon_i x_{2i}\} \neq 0$ and we say that x_{2i} is endogenous
- Other than education, many variables in the wage equation (union status, sickness, industry and marital status) are in fact potentially endogenous
- Married individuals earn on average 10% more wage than unmarried individuals in the US
  - But this is not reflecting the causal effect of being married, rather it reflects the difference in unobservable characteristics of married and unmarried people
- Under additional model identifying assumptions, we would be able to derive another estimator
- The conditions in (4) and (5) are called **moment conditions**
- These conditions would be sufficient to identify the unknown parameters in the model

- The K parameters in β_1 and β_2 should be such that the following K equalities hold:

$$E\{(y_i - x_{1i}'\beta_1 - x_{2i}\beta_2)x_{1i}\} = 0 \tag{6}$$

$$E\{(y_i - x_{1i}'\beta_1 - x_{2i}\beta_2)x_{2i}\} = 0 \tag{7}$$

- These conditions are imposed on the estimator when estimating OLS through the corresponding sample moments
- That is, the OLS estimator $b = (b_1', b_2)'$ for $\beta = (\beta_1', \beta_2)'$ is solved from

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - x_{1i}'b_1 - x_{2i}b_2)x_{1i} = 0 \tag{8}$$

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - x_{1i}'b_1 - x_{2i}b_2)x_{2i} = 0 \tag{9}$$

- These are the first-order conditions for the minimization of the least square criterion and the number of conditions = K (number of unknown parameters)
  - b_1 and b_2 can be solved uniquely from (8) and (9)
- When (5) is violated, (9) drops out and we can no longer solve for b_1 and $b_2 \Longrightarrow \beta_1$ and β_2 are no longer identified
- Identification requires at least one additional moment condition which is possible when we have what is called an **Instrumental Variable (IV)**
- An instrumental variable z_{2i} in this case is a variable such that: $E\{\varepsilon_i z_{2i}\} = 0$ **(the IV is exogenous)** and $E\{z_{2i} x_{2i}\} \neq 0$ **(the IV is relevant)**

- In this case we have

$$E\{(y_i - x_{1i}'\beta_1 - x_{2i}\beta_2)z_{2i}\} = 0 \tag{10}$$

- Such an IV would be referred to as "exogenous" and would be sufficient to the model's K parameters
- Condition (10) is known as the **exclusion restriction**
- The **IV estimator** $\hat{\beta}_{IV}$ can then be solved from

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - x_{1i}'\hat{\beta}_{1,IV} - x_{2i}\hat{\beta}_{2,IV})x_{1i} = 0 \tag{11}$$

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - x_{1i}'\hat{\beta}_{1,IV} - x_{2i}\hat{\beta}_{2,IV})z_{2i} = 0 \tag{12}$$

- Solving these equations analytically gives the IV estimator as follows.

$$\hat{\beta}_{IV} = (\sum_{i=1}^{N} z_i x_i')^{-1} (\sum_{i=1}^{N} z_i y_i) \tag{13}$$

- where $x_i' = (x_{1i}', x_{2i})$ and $z_i' = (x_{1i}', z_{2i})$
- Do you see what happens when $z_{2i} = x_{2i}$?
- Identification of the model and consistency of the IV estimator requires that the moment conditions uniquely identify the parameters of interest

- This is equivalent to saying that π_2 in the following equation is significantly different from zero

$$x_{2i} = x'_{1i}\pi_1 + z_{2i}\pi_2 + \upsilon_i \tag{14}$$

- z_{2i} should also not be a linear combination of the elements in $x_1\, i$
- If these conditions are satisfied, we say that the instrument is **relevant** (testable by $H_0 : \pi_2 = 0$)
- The IV estimator therefore would be implemented using a two-stage framework
  - Stage 1: Estimate (14) (the reduced form equation), and get the predicted values of x_{2i} (the endogenous variable)
  - Stage 2: Run OLS regression of the model using predicted values of from stage 1 instead of the endogenous varible (i.e., use $\hat{x}_{2i}$ in place of x_{2i})

$$\hat{\sigma}^2 = \frac{1}{N-K}\sum_{i=1}^{N}(y_i - x_i'\hat{\beta}_{IV})^2 \tag{15}$$

- Like OLS, one can compute standard errors robust to heteroskedasticity of unknown form

- Practical challenges with the IV estimator
  1. Finding an exogenous and relevant instrument
  2. High standard errors compared to OLS
- Note however that the moment conditions we stated earlier are identifying, **they cannot be tested statistically**
- They can however be tested if there are more conditions than actually needed for identification
- One can however test endogeneity of x_{2i} using a variant of the Hausman test called (Durbin-Wu-Hausman test) by comparing the OLS and IV estimators for β provided that the instrument z_{2i} is valid

- Durbin-Wu-Hausman test: steps
- Step 1: estimate a reduced-form equation explaining x_{2i} from x_{1i} and z_{2i} and save the residuals, say $\hat{\upsilon}_i$
- Step 2: add the residuals to the mode of interest and estimate an OLS model of

$$y_i = x_{1i}'\beta_1 + x_{2i}\beta_2 + \hat{\upsilon}_i\gamma + e_i \tag{16}$$

- One can test the endogeneity of x_{2i} by performing a standard t-test on $\gamma = 0$

- If more than one explanatory variable is endogenous, the dimension of x_{2i} is increased accordingly and the model becomes

$$y_i = x'_{1i}\beta_1 + x'_{2i}\beta_2 + \varepsilon_i \tag{17}$$

- To estimate this equation, we need an instrument for each element in x_{2i}
  - We need instrument for each element in x_{2i}, i.e., equal number of instruments with endogenous variables
- The IV estimator in this case would be

$$\hat{\beta}_{IV} = (\sum_{i=1}^{N} z_i x'_i)^{-1} (\sum_{i=1}^{N} z_i y_i) \tag{18}$$

where now $x'_i = (x'_{1i}, x'_{2i})$ and $z'_i = (x'_{1i}, z'_{2i})$

- The entire vector z_i is called the vector of instruments
- An exogenous variable doesn't need an instrument (or it serves as its own instrument)
- If $z_i = x_i$, the IV model reduces to an OLS model, where each variable is instrumented by itself

- In the **exactly identified** case, $(1/N)\sum_i \hat{\varepsilon}_i z_i = 0$ by construction $\Longrightarrow K = R$ and identifying restrictions are not testable
- But if the model is **overidentified** (i.e., if there are more instruments than endogenous variables), it would be possible to derive a test ststistic which has an asymptotic Chi-squared distribution with $R - K$ d.f
- The test is called an **overidentifying restrictions test** or **Sargan test**
- A simple way to compute the test statistic is by taking $N * R^2$ of an auxiliary regression of IV residuals $\hat{\varepsilon}_i$ upon the full set of instruments z_i

Weak Instruments

- The instrument may exhibit only weak correlation with the endogenous regressor(s)
  - The normal distribution provides a very poor approximation of the distribution of the true IV estimator (even if the sample size is large)
  - The standard IV estimator would therefore be biased, its standard errors are misleading and hypothesis test are unreliable
  - It is possible to investigate this using what is called the "reduced form" regression
- Consider the linear model.

$$y_i = x_{1i}'\beta_1 + x_{2i}\beta_2 + \varepsilon_i \tag{19}$$

and assume that $E\{x_{1i}\varepsilon_i\} = 0$, and additional instruments z_{2i} (for x_{2i}) satisfy $E\{z_{2i}\varepsilon_i\} = 0$

Weak Instruments Cont.

- The appropriate reduced form is given by

$$x_{2i} = x_{1i}'\pi_1 + z_{2i}'\pi_2 + \upsilon_i \tag{20}$$

- If $\pi_2 = 0$, the z_{2i}s are irrelevant and the IV estimator is inconsistent
- If π_2 is close to zero, the instruments are weak
- One can use the value of the F statistic for $\pi_2 = 0$ in this regression
- As a rule of thumb, if the $F-$statistic is greater than 10, we don't need to worry about weak instruments
- If the IVs are insignificant in the reduced form regression, do not trust the IV results!
- If you have more instruments, try by dropping the weakest ones