Graduate Econometrics
Lecture 5
Endogenous Regressors and the Instrumental Variables Estimator

Yonas Alem

Department of Economics
University of Gothenburg

December 2, 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

Causes of Endogeneity

2. Measurement error in the explanatory variables contd...

- $E\{\nu_t|w_t\} = 0$ 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

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

Causes of Endogeneity

3. Omitted variable bias contd...

- 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.

Causes of Endogeneity

4. Simultaneity and reverse causality contd...

- 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).

- One is interested in the best linear combination of $x_2, ..., x_K$ and a constant gives a good approximation of the dependent variable y
- Write an arbitrary linear combination:

$$\tilde{\beta}_1 + \tilde{\beta}_2 x_2 + ... + \tilde{\beta}_K x_K, \tag{3}$$

- Where $\tilde{\beta}_1, ..., \tilde{\beta}_K$ are constants to be chosen
- Index the observations by $i, i = 1, ..., N$
- The difference between the observed value of y and its linear approximation is given by

$$y_i - [\tilde{\beta}_1 + \tilde{\beta}_2 x_{i2} + ... + \tilde{\beta}_K x_{iK}]. \tag{4}$$

- Re-write [2] as:

$$y_i - x_i'\tilde{\beta}. \tag{5}$$

- The objective is to choose values for $\tilde{\beta}_1, ..., \tilde{\beta}_K$ such that these differences are small
- OLS: choose $\tilde{\beta}$ such that the sum of squared differences is as small as possible. i.e

$$S(\tilde{\beta}) \equiv \sum_{i=1}^{N} (y_i - x_i'\tilde{\beta})^2 \tag{6}$$

- FOC: Differentiate $S(\tilde{\beta})$ w.r.t the vector $\tilde{\beta}$, which gives the following K conditions (**Normal Equations**)

$$-2\sum_{i=1}^{N} x_i(y_i - x_i'\tilde{\beta}) = 0 \tag{7}$$

or

$$(\sum_{i=1}^{N} x_i x_i')\tilde{\beta} = \sum_{i=1}^{N} x_i y_i. \tag{8}$$

- The solution to the minimization problem, for b is given by

$$b = (\sum_{i=1}^{N} x_i x_i')^{-1} \sum_{i=1}^{N} x_i y_i \tag{9}$$

- One can confirm from the SOCs that b indeed corresponds to a minimum Thus,

$$\hat{y}_i = x_i' b \tag{10}$$

- Consider the linear wage model

$$y_i = x_{1i}'\beta_1 + x_{2i}\beta_2 + \varepsilon_i \tag{11}$$

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

$$E\{\varepsilon_i x_{1i}\} = 0 \tag{12}$$

and

$$E\{\varepsilon_i x_{2i}\} = 0 \tag{13}$$

- 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 (12) and (13) 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:

$$E\{(y_i - x_{1i}'\beta_1 - x_{2i}\beta_2)x_{1i}\} = 0 \tag{14}$$

$$E\{(y_i - x_{1i}'\beta_1 - x_{2i}\beta_2)x_{2i}\} = 0 \tag{15}$$

- 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

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - x_{1i}'b_1 - x_{2i}b_2)x_{1i} = 0 \tag{16}$$

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - x_{1i}'b_1 - x_{2i}b_2)x_{2i} = 0 \tag{17}$$

- 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 (16) and (17)
- When (13) is violated, (17) 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

$$E\{(y_i - x_{1i}'\beta_1 - x_{2i}\beta_2)z_{2i}\} = 0 \tag{18}$$

- Such an IV would be referred to as "exogenous" and would be sufficient to the model's K parameters
- Condition (18) is known as the **exclusion restriction**
- The **IV estimator** $\hat{\beta}_{IV}$ can then be solved from

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - x_{1i}'\hat{\beta}_{1,IV} - x_{2i}\hat{\beta}_{2,IV})x_{1i} = 0 \tag{19}$$

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - x_{1i}'\hat{\beta}_{1,IV} - x_{2i}\hat{\beta}_{2,IV})z_{2i} = 0 \tag{20}$$

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

$$\hat{\beta}_{IV} = (\sum_{i=1}^{N} z_i x_i')^{-1} (\sum_{i=1}^{N} z_i y_i) \tag{21}$$

- 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

$$x_{2i} = x'_{1i}\pi_1 + z_{2i}\pi_2 + \upsilon_i \tag{22}$$

- 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 (22) (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})

$$\hat{\sigma}^2 = \frac{1}{N-K}\sum_{i=1}^{N}(y_i - x_i'\hat{\beta}_{IV})^2 \tag{23}$$

- 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

$$y_i = x_{1i}'\beta_1 + x_{2i}\beta_2 + \hat{\upsilon}_i\gamma + e_i \tag{24}$$

- 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

$$y_i = x_{1i}'\beta_1 + x_{2i}'\beta_2 + \varepsilon_i \tag{25}$$

- 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

$$\hat{\beta}_{IV} = (\sum_{i=1}^{N} z_i x_i')^{-1} (\sum_{i=1}^{N} z_i y_i) \tag{26}$$

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

Returns to Schooling

- Does higher education result in higher wage?
- The answer seems obvious: people with higher education earn more
- Attempting to using the coefficient of education estimated by OLS would likely be misleading
- It is most likely the case that individuals with greater earning capacity have chosen more years of schooling
- If this is true, the OLS coefficient for education simply reflect differences in unobserved characteristics of working individuals
- Estimating the returns to education addressing the role of unobservables attracted a lot of attention in the past years

Returns to Schooling cont.

- Consider the human capital earnings equation

$$w_i = \beta_1 + \beta_2 S_i + \beta_3 E_i + \beta_4 E_i^2 + \varepsilon_i$$

Where:

- $w_i =$ log of individual earnings, $S =$years of schooling, E_i years of experience
- When information on experience is missing, it is replaced by $age - S_i - 6$
- Reformulate the wage equation as:

$$w_i = z_i'\beta + \gamma S_i + \varepsilon_i \tag{27}$$

where

- z_i includes all observable variables (except S_i)

Returns to Schooling cont.

- OLS estimation of (27) is consistent only if $E\{\varepsilon_i S_i\} = 0$
- But this won't be true and schooling will likely be correlated with ε_i! Why?
  - Ability bias (the most important reason) $\Longrightarrow$ an upward bias in γ, measurement error in education, and other unobserved factors
- In the above formulation there are no instruments available for schooling as all potential candidates are included in the wage equation, i.e.,
- The number of moment conditions in

$$E\{\varepsilon_i z_i\} = E\{(w_i - z_i'\beta - \gamma S_i) z_i\} = 0 \qquad (28)$$

Returns to Schooling cont.

- The moment condition is one short to identify β and γ
- But if there exists a variable in z_i say z_{2i} that affects schooling but not wages, this variable can be can be excluded from the wage equation so as to reduce the number of unknown parameters by 1
  - The model becomes **exactly (just) identified**
  - And the IV estimator of β and γ would be consistent
- IVs used the literature for education: Family background variables (e.g., the number of siblings, or parents' education), institutional factors of the schooling system, quarter of birth, the presence of a nearby college

Returns to Schooling cont.

- Let's use data from the US National Longitudinal Survey consisting 3010 men
- Individuals were followed from 1966 when they were aged 14-24, and interviewed in a number of consecutive years
- The labor market data used here covers 1976
- Note: if education is endogenous, so would be experience and its square
  - The OLS model would therefore suffer from endogeneity of three variables $\Longrightarrow$ we need at least three IVs!
  - Age and its square can be good instruments for experience and its square
  - For schooling, we can try "the presence of a nearby college"

Returns to Schooling - OLS Results

```
. reg  lwage76 ed76 exp76 exp762 black smsa76 south76
```

Source	SS	df	MS		
Model	172.16563	6	28.6942716	Number of obs	3010
Residual	420.476016	3003	.140018653	F(6, 3003)	204.93
				Prob > F	0.0000
				R-squared	0.2905
				Adj R-squared	0.2891
Total	592.641646	3009	.196956346	Root MSE	.37419

lwage76	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ed76	.074009	.0035054	21.11	0.000	.0671357	.0808823
exp76	.0835958	.0066478	12.57	0.000	.0705612	.0966305
exp762	-.0022409	.0003178	-7.05	0.000	-.0028641	-.0016177
black	-.1896315	.0176266	-10.76	0.000	-.2241929	-.1550702
smsa76	.161423	.0155733	10.37	0.000	.1308876	.1919583
south76	-.1248615	.0151182	-8.26	0.000	-.1545046	-.0952184
_cons	4.733664	.0676026	70.02	0.000	4.601112	4.866216

Returns to Schooling - Reduced form OLS results for schooling

```
. reg ed76 age76 age76sq black smsa76 south76 nearc4
```

Source	SS	df	MS		
Model	2555.48762	6	425.914603	Number of obs	= 3010
Residual	19006.5924	3003	6.32920161	F(6, 3003)	= 67.29
				Prob > F	= 0.0000
				R-squared	= 0.1185
				Adj R-squared	= 0.1168
Total	21562.0801	3009	7.16586243	Root MSE	= 2.5158

ed76	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age76	1.061441	.3013985	3.52	0.000	.4704727	1.65241
age76sq	-.0187598	.0052314	-3.59	0.000	-.0290173	-.0085024
black	-1.468367	.1154434	-12.72	0.000	-1.694723	-1.242011
smsa76	.8354027	.1092524	7.65	0.000	.6211856	1.04962
south76	-.4596997	.1024337	-4.49	0.000	-.6605469	-.2588524
nearc4	.347105	.1069972	3.24	0.001	.1373098	.5569002
_cons	-1.869524	4.298357	-0.43	0.664	-10.29755	6.558497

Returns to Schooling IV regression results

```
. ivregress 2sls lwage76 (ed76 exp76 exp762 = age76 age76sq nearc4) black smsa76 south76

Instrumental variables (2SLS) regression          Number of obs =    3010
                                                  Wald chi2(6)  =  890.41
                                                  Prob > chi2   =  0.0000
                                                  R-squared     =  0.1764
                                                  Root MSE      =   .4027
```

lwage76	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ed76	.1329473	.0513196	2.59	0.010	.0323627	.2335319
exp76	.0559614	.0259642	2.16	0.031	.0050725	.1068502
exp762	-.0007957	.0013387	-0.59	0.552	-.0034195	.0018282
black	-.1031403	.0772829	-1.33	0.182	-.254612	.0483314
smsa76	.1079848	.049682	2.17	0.030	.0106098	.2053598
south76	-.0981752	.028731	-3.42	0.001	-.154487	-.0418634
_cons	4.065667	.6077882	6.69	0.000	2.874424	5.25691

```
Instrumented:  ed76 exp76 exp762
Instruments:   black smsa76 south76 age76 age76sq nearc4
```

Returns to Schooling IV regression results cont.

- The estimated returns to schooling in the IV results is 13%
  - With high standard errors though (due to the fairly low correlation between the IV and endogenous regressors
  - This is indicated in the R^2 of the first stage regression (0.1185)
- There is no any goodness-of-fit measure in an IV estimator
- The OLS underestimates the true causal effects of schooling!
- This bias could also be due to measurement error in the schooling variable
- Finding a good instrument is always a challenge!

- 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.

$$y_i = x_{1i}'\beta_1 + x_{2i}\beta_2 + \varepsilon_i \tag{29}$$

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

$$x_{2i} = x_{1i}'\pi_1 + z_{2i}'\pi_2 + \upsilon_i \tag{30}$$

- 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

Heteroskedasticity & IV Estimation methods

Class work

- What is heteroskedasticity? What are its consequences for the OLS estimator?
- How do we test for multiplicative heteroskedasticity?
- State and discuss the options to deal with heteroskedasticity
- Elaborate how the Breusch-Pagan Test works
- Discuss how one checks for the White test of heteroskedasticity
- What are the sources of endogeneity in the linear model?
- Given the linear model $y_i - x_i'\tilde{\beta}$, derive the OLS estimator

- When does the OLS estimator $\hat{\beta}$ become BLUE?
- What is an instrumental variable?
- How does the IV estimator work?
- How do you test if a certain explanatory variable is indeed endogenous? Discuss the test in detail!
- How do you test whether your IV is exogenous?
- Discuss the problem of weak instruments and how to detect it