Endogeneity and Instrumental Variables (IV) Estimation

Hüseyin Taştan[1]

[1] Department of Economics
Ph.D. Program - Advanced Econometrics
Yildiz Technical University

December 3, 2013

Instrumental Variables (IV) and Two-Stage Least Square (2SLS)

OLS/GLS will be inconsistent under the following circumstances

- Ignoring an important variable(s) (Omitted Variable, Unobserved Heterogeneity)
- Measurement errors in X variables (Classical Errors in Variables)
- Simultaneity

Explanatory variables and the random disturbance term will be correlated:

$$\mathsf{E}(\boldsymbol{x}u) \neq \mathbf{0}$$

$\Longrightarrow$ ENDOGENEITY

Omitting a Relevant Variable

- What happens if we exclude an important variable?
- If a relevant variable is omitted this implies that its parameter is **not** 0 in the population regression function (PRF). This is called **underspecification** of the model.
- In this case OLS estimators will be **biased** and **inconsistent**.
- E.g. suppose that the PRF includes 2 independent variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

- Suppose that we omitted x_2 because, say, it is unobservable. Now the sample regression is

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$$

- Is the OLS estimator on x_1, $\tilde{\beta}_1$, still unbiased/consistent?

Omitting a Relevant Variable

- The impact of the omitted variable will be included in the error term:

$$y = \beta_0 + \beta_1 x_1 + \nu$$

- True model includes x_2. Thus the error term ν can be written as:

$$\nu = \beta_2 x_2 + u$$

- OLS estimator of β_1 in the model above is:

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1) y_i}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}$$

- To determine the magnitude and sign of the bias we will substitute y in the formula for $\tilde{\beta}_1$, re-arrange and take expectation.

Omitting a Relevant Variable

$$\begin{aligned}\tilde{\beta}_1 &= \frac{\sum(x_{i1}-\bar{x}_1)(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+u_i)}{\sum(x_{i1}-\bar{x}_1)^2}\\ &= \beta_1+\beta_2\frac{\sum(x_{i1}-\bar{x}_1)x_{i2}}{\sum(x_{i1}-\bar{x}_1)^2}+\frac{\sum(x_{i1}-\bar{x}_1)u_i}{\sum(x_{i1}-\bar{x}_1)^2}\end{aligned}$$

Taking (conditional) expectation we obtain

$$\begin{aligned}\mathsf{E}(\tilde{\beta}_1|x_1,x_2) &= \beta_1+\beta_2\frac{\sum(x_{i1}-\bar{x}_1)x_{i2}}{\sum(x_{i1}-\bar{x}_1)^2}+\frac{\sum(x_{i1}-\bar{x}_1)\overbrace{\mathsf{E}(u_i|x_1,x_2)}^{=0}}{\sum(x_{i1}-\bar{x}_1)^2}\\ &= \beta_1+\beta_2\left(\frac{\sum(x_{i1}-\bar{x}_1)x_{i2}}{\sum(x_{i1}-\bar{x}_1)^2}\right)\end{aligned}$$

Omitting a Relevant Variable

$$E(\tilde{\beta}_1|x_1,x_2)=\beta_1+\beta_2\left(\frac{\sum(x_{i1}-\bar{x}_1)x_{i2}}{\sum(x_{i1}-\bar{x}_1)^2}\right)$$

The expression in the parenthesis to the right of β_2 is just the regression of x_2 on x_1:

$$\tilde{x}_2=\tilde{\delta}_0+\tilde{\delta}_1 x_1$$

Thus

$$E(\tilde{\beta}_1)=\beta_1+\beta_2\tilde{\delta}_1$$

$$bias=E(\tilde{\beta}_1)-\beta_1=\beta_2\tilde{\delta}_1$$

This is called **omitted variable bias**.

Omitted Variable Bias

$$bias=E(\tilde{\beta}_1)-\beta_1=\beta_2\tilde{\delta}_1$$

- Iff $\tilde{\delta}_1=0$ or $\beta_2=0$ then bias is 0.
- The sign of bias depends on both β_2 and the correlation between omitted variable (x_2) and included variable (x_1).
- It is not possible to calculate this correlation if omitted variable cannot be observed.
- The following table summarizes possible cases:

Direction of Bias

	$Corr(x_1,x_2)>0$	$Corr(x_1,x_2)<0$
$\beta_2>0$	positive bias	negative bias
$\beta_2<0$	negative bias	positive bias

Omitted Variable Bias

$$bias=E(\tilde{\beta}_1)-\beta_1=\beta_2\tilde{\delta}_1$$

- The size of the bias is also important. It depends on both $\tilde{\delta}_1$ and β_2.
- A small bias relative to β_1 may not be a problem in practice.
- But in most cases we are not able to calculate the size of the bias.
- In some cases we may have an idea about the direction of bias. For example, suppose that in the wage equation true PRF contains both education and ability.
- Suppose also that ability is omitted because it cannot be observed, leading to omitted variable bias.
- In this case we can say that sign of the bias is + because it is reasonable to think that people with more ability tend to have higher levels of education and ability is positively related to wage.

Omitted Variable Bias

- The effect of omitted variable will be in u. Thus, the exogeneity assumption fails.

$$wage = \beta_0 + \beta_1 educ + \beta_2 ability + u$$

- Instead of this model we estimate

$$wage = \beta_0 + \beta_1 educ + \nu$$

$$\nu = \beta_2 ability + u$$

- Education will be correlated with the error term (ν).

$$\mathsf{E}(\nu|educ) \neq 0$$

- Education is endogenous. If we omit ability the effect of education on wages will be overestimated. Some part of the effect of education on wage comes from ability.

Using Proxy Variables for Unobserved Explanatory Variables

- Can we use a proxy variable for an omitted unobserved explanatory variable?
- We know that if the unobserved variable is an important, relevant variable then OLS estimators are biased and inconsistent.
- The question can be rephrased as follows: Can we solve or at least mitigate the omitted variable bias using proxy variables?
- A **Proxy variable** is something that is related to the unobserved variable that we would like to control for.
- Example: recall that in the wage equation we could not observe innate ability. Can we use intelligence quotient (IQ) as a proxy for ability?
- IQ does not have to be the same thing as ability, we know they are not. But what we need is for IQ to be correlated with ability.

Using Proxy Variables

- Consider the following model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u$$

y :log(wage), x_1:educ, x_2: exper, x_3^*:ability (unobserved)

- x_3^*: unobserved; x_3: proxy for unobserved variable
- Proxy variable must be related to the unobserved variable, represented by the following simple regression:

$$x_3^* = \delta_0 + \delta_3 x_3 + \nu_3$$

- We need the error term ν_3 because these variables are not exactly related.
- Typically, these variables are positively correlated so that $\delta_3 > 0$.
- If $\delta_3 = 0$ then x_3 cannot be a suitable proxy.

Using Proxy Variables

- How can we use x_3 to get an unbiased or at least consistent estimators?
- We can just pretend that x_3^* and x_3 are the same and run the regression of y on x_1, x_2, x_3. This is called **plug-in solution to the omitted variables problem**.
- How does this approach produce consistent estimators?
- To show this we need to make some assumptions about the error terms u and ν_3.
- The error term, u, is uncorrelated with x_1, x_2 and x_3^*. This is the standard assumption.
- In addition to this, u must be uncorrelated with x_3. Since x_3 is the proxy variable, it is irrelevant in the population model. It is x_3^* that affects y not x_3.

$$\mathsf{E}(u|x_1, x_2, x_3^*, x_3) = \mathsf{E}(u|x_1, x_2, x_3^*) = 0$$

Using Proxy Variables

- The error term ν_3 is uncorrelated with x_1, x_2 and x_3.
- This can be stated as follows:

$$\mathsf{E}(x_3^*|x_1, x_2, x_3) = \mathsf{E}(x_3^*|x_3) = \delta_0 + \delta_3 x_3$$

- This says that once x_3 is controlled for the expected value of x_3^* does not depend on x_1 and x_2.
- For example, in the wage equation where IF is the proxy variable for ability this condition becomes

$$\mathsf{E}(ability|educ, exper, IQ) = \mathsf{E}(ability|IQ) = \delta_0 + \delta_3 IQ$$

- This implies that the average level of ability only changes with IQ, not with $educ$ and $exper$. Is this a reasonable assumption?

Using Proxy Variables

- Plugging in $x_3^* = \delta_0 + \delta_3 x_3 + \nu_3$ into the model and rearranging we obtain

$$y = (\beta_0 + \beta_3\delta_0) + \beta_1 x_1 + \beta_2 x_2 + \beta_3\delta_3 x_3 + u + \beta_3\nu_3$$

- Let the composite error term be $e = u + \beta_3\nu_3$

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \alpha_3 x_3 + e$$

where $\alpha_0 = (\beta_0 + \beta_3\delta_0), \alpha_3 = \beta_3\delta_3$

- If the assumptions for the proxy variables are all satisfied then the composite error term e will be uncorrelated with the explanatory variables included in the model. Thus, OLS estimators of $\alpha_0, \beta_1, \beta_2, \alpha_3$ will be consistent.
- The coefficient on IQ, α_3, measures the impact of a one point change in IQ test score on wage.

Using Proxy Variables: Wage2.dta

- This data set contains information about monthly wages, education, experience, tenure, IQ scores, and several demographic characteristics for a sample of 935 working men in 1980.
- Adding IQ test scores we obtain the following results:

Model 1: OLS, using observations 1–935
Dependent variable: lwage

	Coefficient	Std. Error	t-ratio	p-value
const	5.17644	0.128001	40.4407	0.0000
educ	0.0544106	0.00692849	7.8532	0.0000
exper	0.0141458	0.00316510	4.4693	0.0000
tenure	0.0113951	0.00243938	4.6713	0.0000
married	0.199764	0.0388025	5.1482	0.0000
south	−0.0801695	0.0262529	−3.0537	0.0023
urban	0.181946	0.0267929	6.7908	0.0000
black	−0.143125	0.0394925	−3.6241	0.0003
IQ	0.00355910	0.000991808	3.5885	0.0004

Mean dependent var	6.779004	S.D. dependent var	0.421144
Sum squared resid	122.1203	S.E. of regression	0.363152
R^2	0.262809	Adjusted R^2	0.256441

Using Lagged Dependent Variables as Proxy Variables

- In some applications (eg, the wage example) we have at least a vague idea about which unobserved factor we want to control.
- In other applications, we suspect that one or more of the independent variables is correlated with an omitted variable, but we have no idea how to obtain a proxy for that omitted variable.
- In such cases, we can include the value of the dependent variable y from an earlier time period, y_{-1}.
- To do this we need the lagged value of the dependent variable. This provides a way of controlling historical factors that cause current differences in dependent variable.
- For example, some cities have had high crime rates in the past Many of the unobserved factors contribute to both high current and past crime rates. Slowly moving components in dependent variable (inertial effects) can be captured by the lagged value.

Using Lagged Dependent Variables as Proxy Variables

- Example: CRIME2.dta, 1987 crime data for 46 cities, information in 1982 also available
- The model without the lagged crime rate:

$$\widehat{\text{l\_crmrte87}} = \underset{(1.251)}{3.34} - \underset{(0.032)}{0.029}\,\text{unem87} + \underset{(0.173)}{0.203}\,\text{l\_lawexpc87}$$

$$n = 46 \quad R^2 = 0.057$$

- The model with lagged crime rate:

$$\widehat{\text{l\_crmrte87}} = \underset{(0.821)}{0.076} + \underset{(0.02)}{0.009}\,\text{unem87} - \underset{(0.109)}{0.140}\,\text{l\_lawexpc87} + \underset{(0.132)}{1.194}\,\text{l\_crmrte82}$$

$$n = 46 \quad R^2 = 0.680$$

- In the first model, crime rate decreases as unemployment increases. This is counterintuitive.
- After controlling for the crime rate in 1982 (5 years ago) coefficient on $unem$ is positive but insignificant.
- What is the elasticity of the current crime rate to the crime rate in the previous period?

Measurement Error in Explanatory Variable

- Measurement errors in y generally lead to increased in variance
- But measurement errors in x variables can lead to more serious problems (ie, inconsistency).
- To determine conditions under which OLS estimators become inconsistent let us consider the simple regression model:

$$y = \beta_0 + \beta_1 x_1^* + u$$

Suppose that the first 4 Gauss-Markov assumptions hold.

- Here, x_1^* is the unobserved actual value and x_1 is the observed value.
- Then, the measurement error is

$$e_1 = x_1 - x_1^*$$

- Assume that the expected value of the measurement error is zero: $\mathsf{E}(e_1) = 0$

Measurement Error in Explanatory Variable

- Assume that the error term u is uncorrelated with both x_1^* and x_1 so that:

$$\mathsf{E}(y|x_1^*, x_1) = \mathsf{E}(y|x_1^*)$$

- This means that after controlling for x_1^* we no longer need x_1 in the model.
- If we use x_1 instead of x_1^*, what are the properties of OLS estimators? Are they still consistent?
- This depends on the assumption we make about the measurement error.
- There are two possible assumptions: (1) measurement error is uncorrelated with x_1.
- (2) measurement error is uncorrelated with unobserved actual value, x_1^*.

(1) e_1 and x_1 are uncorrelated

- This assumption can be written as

$$\mathsf{Cov}(x_1, e_1) = 0$$

- Since $e_1 = x_1 - x_1^*$, it must be the case that e_1 and x_1^* are correlated.
- Under this assumption, substituting $x_1^* = x_1 - e_1$ in the model we obtain:

$$y = \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1)$$

- Expected value and variance of the composite error term:

$$\mathsf{E}(u - \beta_1 e_1) = 0, \quad \mathsf{Var}(u - \beta_1 e_1) = \sigma_u^2 + \beta_1^2 \sigma_{e_1}^2$$

- OLS estimators are consistent because the error term and x_1 are uncorrelated. But the variance will be higher.

(2) e_1 and x_1^* are uncorrelated (CEV Assumption)

- This is known as the "Classical Errors-in-Variables (CEV)". In the econometrics literature, when we talk about measurement error in explanatory variable we usually mean CEV.
- The CEV assumption can be written as:

$$\text{Cov}(x_1^*, e_1) = 0$$

- The observed value can be written as the sum of actual value and measurement error:

$$x_1 = x_1^* + e_1$$

- Obviously, if x_1^* and e_1 are uncorrelated, then, x_1 and e_1 must be correlated:

$$\text{Cov}(x_1, e_1) = \text{E}(x_1 e_1) = \text{E}(x_1^* e_1) + \text{E}(e_1^2) = 0 + \sigma_{e_1}^2 = \sigma_{e_1}^2$$

- Under CEV assumption, the covariance between x_1 and e_1 is equal to the variance of the measurement error.

(2) CEV Assumption: $\text{Cov}(x_1^*, e_1) = 0$

- Recall that the model was written as:

$$y = \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1)$$

- Since e_1 is included in the composite error term, its covariance with x_1 will create a problem.
- The covariance between composite error term and x_1 is

$$\text{Cov}(x_1, u - \beta_1 e_1) = -\beta_1 \text{Cov}(x_1, e_1) = -\beta_1 \sigma_{e_1}^2$$

- Because this covariance is not 0, OLS estimators will be biased and inconsistent under CEV assumption
- We can calculate the amount of inconsistency in OLS.

(2) CEV Assumption: $\text{Cov}(x_1^*, e_1) = 0$

- In the simple regression model, the probability limit of the OLS estimator of the slope parameter is:

$$\begin{aligned}
\text{plim}(\hat{\beta}_1) &= \beta_1 + \frac{\text{Cov}(x_1, u - \beta_1 e_1)}{\text{Var}(x_1)} \\
&= \beta_1 - \frac{\beta_1 \sigma_{e_1}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2} \\
&= \beta_1 \left(1 - \frac{\sigma_{e_1}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2} \right) \\
&= \beta_1 \left(\frac{\sigma_{x_1^*}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2} \right)
\end{aligned}$$

(2) CEV Assumption: $\text{Cov}(x_1^*, e_1) = 0$

- Probability limit of the OLS estimator:

$$\text{plim}(\hat{\beta}_1) = \beta_1 \underbrace{\left(\frac{\sigma_{x_1^*}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2} \right)}_{\leq 1} \neq \beta_1$$

- The term in the parenthesis will always be smaller than 1. If and only if $\sigma_{e_1}^2 = 0$ then it is 1.
- This means that: $\hat{\beta}_1$ is always closer to 0 than the true value β_1 is. This is called **attenuation bias**.
- If $\beta_1 > 0$ then $\hat{\beta}_1$ will approach a value smaller than the true value in the limit (underestimation). Otherwise, it will approach a bigger value (overestimation).

(2) CEV Assumption: Cov$(x_1^*, e_1) = 0$

- Probability limit of the OLS estimator:

$$\text{plim}(\hat{\beta}_1) = \beta_1 \underbrace{\left(\frac{\sigma^2_{x_1^*}}{\sigma^2_{x_1^*} + \sigma^2_{e_1}} \right)}_{\leq 1} \neq \beta_1$$

- If the variance of x_1^* is large as compared to the variance of e_1 then the ratio Var(x_1^*)/Var(x_1) will be close to 1. In this case the amount of inconsistency may not be large. But it is almost impossible to determine this.
- Things are more complicated when we add more explanatory variables.
- But we can say that measurement errors generally lead to inconsistency of all OLS estimators.

(2) CEV Assumption: Cov$(x_1^*, e_1) = 0$

- Consider the following model for the college success:

$$colGPA = \beta_0 + \beta_1 faminc^* + \beta_2 hsGPA + \beta_3 SAT + u$$

$faminc$: Family income, $hsGPA$: high school GPA, SAT: Scholastic Aptitude Test result

- $faminc^*$ is the actual family income. If a questionnaire method is used to collect data then the student will be asked to report family income.
- We can collect data on hsGPA and SAT scores from student records. But we cannot do this for family income levels.
- If the reported income is different from the actual income, and if the CEV assumption is valid (ie actual income and measurement error are uncorrelated) then, OLS estimator for β_1 will be biased and inconsistent.
- As a result, the impact of the family income on the college success will be underestimated (downward bias).

Simultaneous Equations Model (SEM) and Simultaneity

- The source of endogeneity in a SEM is simultaneity: explanatory and dependent variables are simultaneously determined.
- For example: supply and demand simultaneously determine price
- Relationship between crime and police force.
- Simultaneity: there is a feedback from dependent variable to error term
- OLS is inconsistent, IV or GMM should be used.

SEM Example: Demand-Supply Model for Labor

Labor supply:

$$h^s = \alpha_1 w + \beta_1 z_1 + u_1$$

- h^s: annual labor hours supplied by employees
- w: average hourly wage
- z_1: an observable variable that affects labor supply (supply-shifter, e.g., average wage level in another sector)

Labor demand:

$$h^d = \alpha_2 w + \beta_2 z_2 + u_2$$

- h^d: firms' demand for labor hours
- z_2: an observable variable that affects demand for labor (demand-shifter, e.g. prices of inputs)
- Both equations are structural, ie they can be derived from theory and have causal interpretation.

SEM Example: Demand-Supply Model for Labor

Market equilibrium:

$$h^s = h^d \equiv h$$

Ignoring observation subscripts SEM can be written as:

$$\begin{aligned} h &= \alpha_1 w + \beta_1 z_1 + u_1 \\ h &= \alpha_2 w + \beta_2 z_2 + u_2 \end{aligned}$$

- h, w: endogenous variables, $(\alpha_1 \neq \alpha_2)$
- z_1, z_2: exogenous variables
- u_1, u_2: structural error terms
- Which one is supply, which one is demand?

SEM: Crime rate and police force

$$crime\ rate = \alpha_1 police\ force + \beta_{10} + \beta_{11} income + u_1$$

$$police\ force = \alpha_2 crime\ rate + \beta_{20} + other\ factors + u_2$$

- $\alpha_1 < 0$, $\alpha_2 > 0$
- Crime and police force are simultaneously determined.
- We need the second equation to estimate the first.
- Simultaneity: error term is correlated with endogenous variables

OLS Simultaneity Bias

$$y_1 = \alpha_1 y_2 + \beta_1 z_1 + u_1 \qquad (1)$$

$$y_2 = \alpha_2 y_1 + \beta_2 z_2 + u_2 \qquad (2)$$

z_1, z_2: strictly exogenous (determined outside the model). Solving the model for y_2:

$$\begin{aligned} y_2 &= \alpha_2(\alpha_1 y_2 + \beta_1 z_1 + u_1) + \beta_2 z_2 + u_2 \\ (1 - \alpha_2\alpha_1) y_2 &= \alpha_2 \beta_1 z_1 + \beta_2 z_2 + \alpha_2 u_1 + u_2 \end{aligned}$$

OLS Simultaneity Bias

Assuming $\alpha_1\alpha_2 \neq 0$ the reduced form:

$$y_2 = \pi_{21} z_1 + \pi_{22} z_2 + v_2 \qquad (3)$$

$$\pi_{21} = \frac{\alpha_2 \beta_1}{1 - \alpha_2\alpha_1}, \quad \pi_{22} = \frac{\beta_2}{1 - \alpha_2\alpha_1}$$

$$v_2 = \frac{\alpha_2 u_1 + u_2}{1 - \alpha_2\alpha_1}$$

v_2: reduced form error (uncorrelated with exogenous variables, OLS is consistent)

OLS Simultaneity Bias

If we estimate (1) by OLS simultaneity bias will arise. α_1 and β_1 cannot be estimated consistently.

$$y_1 = \alpha_1 y_2 + \beta_1 z_1 + u_1$$

- z_1 and u_1 are uncorrelated.
- However, as can be seen in the reduced form, v_2 and u_1 are correlated. In this case, y_2 and u_1 will be correlated by definition.
- Also, if u_1 and u_2 are correlated, then v_2 and u_1 will always be correlated.
- If $\alpha_2 = 0$ and u_1, u_2 uncorrelated then y_2 and u_1 will be uncorrelated. But these are very strong assumptions.
- The bias caused by the correlation between y_2 and u_1 is called *OLS simultaneity bias*.
- It is very difficult to determine the sign and the magnitude of the bias.

Solving Endogeneity: Instrumental Variables

- Suppose that one of the $\boldsymbol{X}$ variables, say $\boldsymbol{X}_k$, is correlated with $\boldsymbol{u}$.
- In that case all OLS estimators $\hat{\boldsymbol{\beta}}_j$ will be inconsistent.
- Solution: Find an observable exogenous variable which is not in the model, $\boldsymbol{Z}_1$, and use that variable in the estimation. $\boldsymbol{Z}_1$ is called an instrument for $\boldsymbol{X}_k$.
- How to choose an instrument?
  - Cov$(Z_1, u) = 0$
  - $\boldsymbol{Z}_1$ and $\boldsymbol{X}_k$ should be correlated.

IV

Model: $y = \boldsymbol{x}^\top \boldsymbol{\beta} + u$. Let $k \times 1$ instrument vector be $\boldsymbol{z}$ (this may include explanatory variables). Instrumental variables should satisfy the following condition:

$$\mathsf{E}[\boldsymbol{z}u] = \boldsymbol{0}_k$$

From this population moment condition:

$$\mathsf{E}[\boldsymbol{z}\boldsymbol{x}^\top]\boldsymbol{\beta} = \mathsf{E}[\boldsymbol{Z}y]$$

$$\boldsymbol{\beta} = \left(\mathsf{E}\left[\boldsymbol{z}\boldsymbol{x}^\top\right]\right)^{-1} \mathsf{E}\left[\boldsymbol{z}y\right]$$

$$\hat{\boldsymbol{\beta}}_{IV} = (\boldsymbol{Z}^\top \boldsymbol{X})^{-1} \boldsymbol{Z}^\top \boldsymbol{y}$$

IV: How to choose instruments?

Example: Simple wage equation

$$wage = \beta_1 + \beta_2 educ + u$$

- Cov$(educ, u) = 0$??
- Can we estimate β_2 consistently using OLS?
- Educ may be correlated with unobserved factors in $\boldsymbol{u}$ (e.g., unobserved innate ability).
- Can we find an instrument for educ?
- Mother's, education level, father's education level, number of siblings, last four digits of citizen ID number (?), IQ level (?),...

Simple model

$$y = \beta_0 + \beta_1 x + u$$

- Instrument z: must be exogenous
- z must not have a partial effect on y, it should not be a part of the model.
- z must not be correlated with other factors affecting y
- z must be correlated with the endogenous x (positively or negatively)
- Must have $\text{Cov}(z, u) = 0 \implies$ cannot be tested ! (maintained assumption)
- Must have $\text{Cov}(z, x) \neq 0 \implies$ can be tested.

Simple Model

$$y = \beta_0 + \beta_1 x + u$$

$$\text{Cov}(z, u) = 0$$

$$\text{Cov}(z, x) \neq 0$$

How can we test the second condition?

$$x = \pi_0 + \pi_1 z + \nu$$

OLS estimator:

$$\pi_1 = \frac{\text{Cov}(z, x)}{\text{Var}(z)}$$

t-test: $H_0 : \pi_1 = 0$, $H_1 : \pi_1 \neq 0$ should be significant !

Simple Model: Example

$$score = \beta_0 + \beta_1 skipped + u$$

score: final exam grade
skipped: number of classes missed during the semester

- If there are omitted variables affecting the score then the OLS estimator of β_1 will be inconsistent. Can we find an IV for skipped?
- It should not have a direct affect on score, it should not be a part of the model
- Should be correlated with skipped but uncorrelated with student's ability (unobserved, included in u)
- E.g. distance: distance of the student's residence to the school
- We can assume that as the distance increases the probability of skipping classes will increase: positively correlated with skipped
- Is distance correlated with u?
- Student's income level is not included in the model, therefore

Simpel Model: Example (Wooldridge, Intro, pp.469-470)

Angrist (1990): the effect of being a veteran of the Vietnam war on lifetime earnings:

$$\log(earns) = \beta_0 + \beta_1 veteran + u$$

earns: lifetime earnings
veteran: dummy variable $= 1$ if Vietnam veteran

- Problem: sample selection: the decision to serve in the military may be correlated with individual characteristics that affect wage. $\text{Cov}(veteran, u) = 0$ may not hold. OLS is inconsistent
- draft lottery system provides a natural experiment environment (see Wooldridge, 2001, *Econometric Analysis of Cross Section and Panel Data*, ch.5, p.88).

Simpel Model: Example (Wooldridge, Intro, pp.469-470)

Angrist (1990): the effect of being a veteran of the Vietnam war on lifetime earnings:

$$\log(earns) = \beta_0 + \beta_1 veteran + u$$

earns: lifetime earnings
veteran: dummy variable $= 1$ if Vietnam veteran

- Vietnam Draft lottery: young men were given lottery numbers that determined whether they would be called to serve at the war. Numbers given were randomly assigned and should be uncorrelated with u
- Those with a low enough number had to serve in Vietnam. Veteran is correlated with the lottery number, can be an IV?
- Question: if some men who were assigned low draft lottery numbers obtained additional schooling to reduce the probability of being drafted, is lottery number a good instrument for veteran?

Identification in the simple model

$$y = \beta_0 + \beta_1 x + u$$
$$\text{Cov}(z, u) = 0$$
$$\text{Cov}(z, x) \neq 0$$

β_1 can be written in terms of population moments. The covariance between z and y can be written as follows:

$$\text{Cov}(z, y) = \beta_1 \underbrace{\text{Cov}(z, x)}_{\neq 0} + \underbrace{\text{Cov}(z, u)}_{=0}$$

This implies

$$\beta_1 = \frac{\text{Cov}(z, y)}{\text{Cov}(z, x)}$$

Sample analog:

$$\hat{\beta}_1 = \frac{\sum(z - \bar{z})(y - \bar{y})}{\sum(z - \bar{z})(x - \bar{x})}$$

Identification in the simple model

$$\beta_1 = \frac{\text{Cov}(z, y)}{\text{Cov}(z, x)}$$

Sample analog of the population moment condition:

$$\hat{\beta}_1 = \frac{\sum(z - \bar{z})(y - \bar{y})}{\sum(z - \bar{z})(x - \bar{x})}$$

IV estimator $\hat{\beta}_1$ is consistent: $plim(\hat{\beta}_1) = \beta_1$
If z and x are weakly correlated the standard errors of IV estimates can be high (Weak Instruments)

Weak Instruments

Probability limit of IV estimator:

$$plim\hat{\beta}_1 = \beta_1 + \frac{Corr(z, u)}{Corr(z, x)} \frac{\sigma_u}{\sigma_x}$$

Even if Corr(z, u) small, if Corr(z, x) is also small then IV may have large asymptotic bias. In some cases, using OLS would even be better. Probability limit of OLS estimator:

$$plim\tilde{\beta}_1 = \beta_1 + Corr(x, u)\frac{\sigma_u}{\sigma_x}$$

If

$$\frac{Corr(z, u)}{Corr(z, x)} < Corr(x, u)$$

then IV should be preferred.

Example: Birth weight and cigarette smoking

$$\log(bwght) = \beta_0 + \beta_1 packs + u$$

bwgth: birth weight of newly born babies,
packs: average cigarette smoked by mother during pregnancy (in packs)

- packs and u correlated?
- IV for packs: average cigarette price in the region
- Is cigarette price correlated with u?

Example: Birth weight and cigarette smoking, BWGHT.dta

```
. reg  packs cigprice

      Source |       SS       df       MS              Number of obs =    1388
-------------+------------------------------           F(  1,  1386) =    0.13
       Model |  .011648626     1  .011648626           Prob > F      =  0.7179
    Residual |  123.684481  1386  .089238442           R-squared     =  0.0001
-------------+------------------------------           Adj R-squared = -0.0006
       Total |  123.696129  1387  .089182501           Root MSE      =  .29873

------------------------------------------------------------------------------
       packs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cigprice |   .0002829    .000783     0.36   0.718    -.0012531    .0018188
       _cons |   .0674257   .1025384     0.66   0.511    -.1337215    .2685728
------------------------------------------------------------------------------
```

cigprice: should not be used as an IV

Example: Birthweight and cigarette smoking

IV regression with cigprice:

```
. ivregress 2sls  lbwght (packs =  cigprice)

Instrumental variables (2SLS) regression               Number of obs =    1388
                                                       Wald chi2(1)  =    0.12
                                                       Prob > chi2   =  0.7310
                                                       R-squared     =       .
                                                       Root MSE      =  .93818

------------------------------------------------------------------------------
      lbwght |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       packs |   2.988676   8.692619     0.34   0.731    -14.04854     20.0259
       _cons |   4.448136   .9075006     4.90   0.000     2.669468    6.226805
------------------------------------------------------------------------------
Instrumented:  packs
Instruments:   cigprice
```

R^2 is negative, coefficient estimate on packs does not have expected sign.

Identification in Two Equation System

$$supply:\ q = \alpha_1 p + \beta_1 z_1 + u_1$$
$$demand:\ q = \alpha_2 p + u_2$$

Notice that there are no demand-shifter in the demand equation. Suppose that we have a random sample (q, p, z_1). Which equation can be identified and consistently estimated?

- In this system, demand is identified not supply. Why?
- A variable that can shift the supply curve can be used to identify the demand equation. Since we only observe equilibrium values (q, p), we cannot know which equation they belong to.
- If one of the equations change (shift) while the other is fixed then we can identify the fixed equation.
- In other words, z_1 is an IV for demand.
- If we can find an observable demand shifter (for example, consumer income level) we can identify the supply equation.

More than one variable

Structural equation:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$$

- y_1: endogenous
- z_1: exogenous
- y_2: endogenous, can be correlated with u_1
- For example

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + u_1$$

- OLS estimators of all coefficients are inconsistent if educ is endogenous
- We need to find an exogenous variables that does not belong to the model. Let that variable be z_2:

$$\mathsf{E}(u_1) = 0, \quad \mathsf{Cov}(z_1, u_1) = 0, \quad \mathsf{Cov}(z_2, u_1) = 0$$

More than one variable

Structural equation:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$$

$$\mathsf{E}(u_1) = 0, \quad \mathsf{Cov}(z_1, u_1) = 0, \quad \mathsf{Cov}(z_2, u_1) = 0$$

Sample analogs:

$$\begin{aligned} \sum \hat{u}_1 &= 0 \\ \sum z_1 \hat{u}_1 &= 0 \\ \sum z_2 \hat{u}_1 &= 0 \end{aligned}$$

$\hat{u}_1 = y - \hat{\beta}_0 - \hat{\beta}_1 y_2 - \hat{\beta}_2 z_1$
IV estimators can be derived by solving the system above.

Identification

Structural Equation:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$$

Reduced Form:

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \nu_2$$

Note that endogenous variable is written in terms of exogenous variables.

$$\begin{aligned} \mathsf{E}(\nu_2) &= 0 \\ \mathsf{Cov}(z_1, \nu_2) &= 0 \\ \mathsf{Cov}(z_2, \nu_2) &= 0 \end{aligned}$$

For identification we must have $\pi_2 \neq 0$. z_2 is an IV for y_2. Note that the affect of z_1 is partialled out.

Example: Proximity to university (Wooldridge, Intro II.ed, pp.474-5)

Card (1995):

- IV for education: a dummy variable for whether someone grew up near a four-year college
- nearc4: $= 1$ for those who grew up near a 4-year college
- Dependent variable: logarithmic wages
- Explanatory variables: experience, black, SMSA, south, regional dummies, SMSA66
- SMSA: standard metropolitan statistical area (dummy)
- south: dummy variable =1 if located in the south

Example: Proximity to College and Wages

Reduced form for educ:

```
. regress  educ  nearc4 exper expersq black smsa south smsa66 reg662- reg669

      Source |       SS       df       MS              Number of obs =    3010
-------------+------------------------------           F( 15,  2994) =  182.13
       Model |  10287.6179    15  685.841194           Prob > F      =  0.0000
    Residual |  11274.4622  2994  3.76568542           R-squared     =  0.4771
-------------+------------------------------           Adj R-squared =  0.4745
       Total |  21562.0801  3009  7.16586243           Root MSE      =  1.9405

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      nearc4 |   .3198989   .0878638     3.64   0.000     .1476194    .4921785
       exper |  -.4125334   .0336996   -12.24   0.000    -.4786101   -.3464566
     expersq |   .0008686   .0016504     0.53   0.599    -.0023674    .0041046
       black |  -.9355287   .0937348    -9.98   0.000     -1.11932   -.7517377
        smsa |   .4021825   .1048112     3.84   0.000     .1966732    .6076918
       south |  -.0516126   .1354284    -0.38   0.703    -.3171548    .2139296
      smsa66 |   .0254805   .1057692     0.24   0.810    -.1819071    .2328682
      reg662 |  -.0786363   .1871154    -0.42   0.674    -.4455241    .2882514
      reg663 |   -.027939   .1833745    -0.15   0.879    -.3874918    .3316139
      reg664 |    .117182   .2172531     0.54   0.590    -.3087984    .5431624
      reg665 |  -.2726165   .2184204    -1.25   0.212    -.7008858    .1556528
      reg666 |  -.3028147   .2370712    -1.28   0.202    -.7676536    .1620242
      reg667 |  -.2168177   .2343879    -0.93   0.355    -.6763953    .2427598
      reg668 |   .5238914   .2674749     1.96   0.050    -.0005618    1.048344
      reg669 |    .210271   .2024568     1.04   0.299    -.1866975    .6072395
       _cons |   16.63825   .2406297    69.14   0.000     16.16644    17.11007
------------------------------------------------------------------------------
```

nearc4 is significant, can be an IV for educ (if uncorrelated with u)

Example: Proximity to College and Wages

```
. ivregress 2sls  lwage exper expersq black smsa south smsa66 reg662- reg669 (educ = nearc4)

Instrumental variables (2SLS) regression               Number of obs =    3010
                                                       Wald chi2(15) =  769.20
                                                       Prob > chi2   =  0.0000
                                                       R-squared     =  0.2382
                                                       Root MSE      =   .3873

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1315038   .0548174     2.40   0.016     .0240637     .238944
       exper |   .1082711   .0235956     4.59   0.000     .0620246    .1545176
     expersq |  -.0023349   .0003326    -7.02   0.000    -.0029868    -.001683
       black |  -.1467757   .0537564    -2.73   0.006    -.2521364   -.0414151
        smsa |   .1118083   .0315777     3.54   0.000     .0499171    .1736995
       south |  -.1446715    .027212    -5.32   0.000    -.1980061   -.0913369
      smsa66 |   .0185311   .0215511     0.86   0.390    -.0237082    .0607704
      reg662 |   .1007678   .0375854     2.68   0.007     .0271017    .1744339
      reg663 |   .1482588   .0367162     4.04   0.000     .0762964    .2202211
      reg664 |   .0498971   .0436234     1.14   0.253    -.0356032    .1353974
      reg665 |   .1462719   .0469387     3.12   0.002     .0542738    .2382701
      reg666 |   .1629029   .0517714     3.15   0.002     .0614328    .2643731
      reg667 |   .1345722   .0492708     2.73   0.006     .0380032    .2311413
      reg668 |   -.083077   .0591735    -1.40   0.160    -.1990548    .0329008
      reg669 |   .1078142   .0417024     2.59   0.010      .026079    .1895494
       _cons |   3.666151   .9223682     3.97   0.000     1.858342    5.473959
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   exper expersq black smsa south smsa66 reg662 reg663 reg664
               reg665 reg666 reg667 reg668 reg669 nearc4
```

Example: Proximity to College and Wages

For neat tables use `estout` STATA package. To install

```
. ssc install estout
```

website:

`http://repec.org/bocode/e/estout/`

Komutlar:

```
quietly regress lwage educ exper expersq black smsa south smsa66 reg662- reg669
estadd scalar Num_obs = e(N)
estadd scalar R2 = e(r2)
eststo

quietly ivregress 2sls  lwage exper expersq black smsa south smsa66 reg662- reg669 (educ = near
estadd scalar Num_obs = e(N)
estadd scalar R2 = e(r2)
eststo

esttab , compress b(3)  se(3) star(* 0.1 ** 0.05 *** 0.01) ///
      mtitles("OLS" "IV") stats(Num_obs R2, fmt(0 2)) ///
   title(CARD Data OLS and IV Estimation Results)
eststo clear
```

2SLS: Single Endogenous Variable

Structural Equation:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$$

Let z_2 and z_3 be two exogenous variables excluded from the model. If both of these are correlated with y_2 then they can be IVs. In this case we obtain two IV estimator both of which will in general be inefficient. But we can use linear combination of them as IV. Consider the reduced form:

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \nu_2$$

$\pi_2 \neq 0$ or $\pi_3 \neq 0$. If both are 0 then the structural model cannot be identified. This can be tested using an F-test.

2SLS: Single Endogenous Variable

Structural Equation:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$$

Reduced form equation (linear projection of y_2):

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \nu_2$$

IV-2SLS steps:

- Regress y_2 on z_1, z_2, z_3, obtain $\hat{y}_2$
- Regress y_1 on $\hat{y}_2$, z_1

Example: Wages and Education for working women, MROZ.dta (Wooldridge, Intro p.478)

IVs for education level of working women: mother's and father's education level. Reduced form estimation results:

```
. regress  educ exper expersq motheduc fatheduc if  inlf
      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  4,   423) =   28.36
       Model |  471.620998     4  117.90525           Prob > F      =  0.0000
    Residual |  1758.57526   423  4.15738833           R-squared     =  0.2115
-------------+------------------------------           Adj R-squared =  0.2040
       Total |  2230.19626   427  5.22294206           Root MSE      =   2.039
------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0452254   .0402507     1.12   0.262    -.0338909    .1243417
     expersq |  -.0010091   .0012033    -0.84   0.402    -.0033744    .0013562
    motheduc |    .157597   .0358941     4.39   0.000      .087044    .2281501
    fatheduc |   .1895484   .0337565     5.62   0.000     .1231971    .2558997
       _cons |    9.10264   .4265614    21.34   0.000     8.264196    9.941084
------------------------------------------------------------------------------
. test  motheduc fatheduc
 ( 1)  motheduc = 0
 ( 2)  fatheduc = 0
       F(  2,   423) =   55.40
            Prob > F =    0.0000
```

Example: MROZ.dta (Wooldridge, Intro p.478)

```
. ivregress 2sls  lwage exper expersq (educ =  motheduc fatheduc)
Instrumental variables (2SLS) regression               Number of obs =     428
                                                       Wald chi2(3)  =   24.65
                                                       Prob > chi2   =  0.0000
                                                       R-squared     =  0.1357
                                                       Root MSE      =  .67155
------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0613966   .0312895     1.96   0.050     .0000704    .1227228
       exper |   .0441704   .0133696     3.30   0.001     .0179665    .0703742
     expersq |   -.000899   .0003998    -2.25   0.025    -.0016826   -.0001154
       _cons |   .0481003    .398453     0.12   0.904    -.7328532    .8290538
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   exper expersq motheduc fatheduc
```

2SLS: More than one endogenous variables

Structural Equation:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 y_3 + \beta_3 z_1 + \beta_4 z_2 + \beta_5 z_3 + u_1$$

2SLS: We need at least two exogenous variables which are correlated with y_2 and y_3. Order condition: Number of exogenous variables (instruments) must be at least as large as the number of endogenous variables in the model.
Rank condition:$rank(\boldsymbol{Z}^\top \boldsymbol{X}) = k$

IV-2SLS

Number of instruments can be larger than the number of endogenous variables. Let the model be

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + u_i, \quad i = 1, 2, \dots, n$$

where $\boldsymbol{x}_i$ is $k \times 1$. Let $\boldsymbol{z}_i$ be $l \times 1$, $l \geq k$, vector of instrumental variables. Similarly, let $\boldsymbol{X}$ be $n \times k$ and $\boldsymbol{Z}$ $n \times l$ data matrices. The IV estimator is

$$\hat{\boldsymbol{\beta}}_{IV} = \left(\boldsymbol{X}^\top \boldsymbol{Z} (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top \boldsymbol{X} \right)^{-1} \boldsymbol{X}^\top \boldsymbol{Z} (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top \boldsymbol{y} \qquad (4)$$

- 2SLS
- 1. stage: Regress $\boldsymbol{X}$ on $\boldsymbol{Z}$, obtain $\hat{\boldsymbol{X}}$
- 2. stage: Regress $\boldsymbol{y}$ on $\hat{\boldsymbol{X}}$, obtain $\hat{\boldsymbol{\beta}}_{IV} \equiv \hat{\boldsymbol{\beta}}_{2SLS}$

Hausman Endogeneity Test

- IV performance can be worse than OLS if the instruments are weak.
- Is there a test that can help us to decide between 2SLS and OLS?
- Hausman test for endogeneity:
- $$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1$$
- If y_2 and u_1 are uncorrelated: OLS
- Hausman test: compares OLS and 2SLS estimates.
- If the difference is big: prefer 2SLS

Hausman Endogeneity Test

Steps:

- Estimate the reduced form, (ie, regress the endogenous variable on all exogenous variables and IVs), obtain the residuals: $\hat{\nu}$
- Add $\hat{\nu}$ to the structural equation, conduct a t-test (can use heteroskedasticity-robust se).
- If significant then the variable is endogenous, prefer 2SLS

Overidentifying Restrictions (OID) Test

If the number of instruments is larger than the number of endogenous variables, we can test whether some of them are uncorrelated with the structural error. Steps:

- Estimate the structural model by 2SLS and save the residuals: $\hat{u}$
- Regress $\hat{u}$ on all exogenous variables and obtain R^2.
- Under the null hypothesis of no correlation between u and instrumental variables

$$OID = nR^2 \sim \ \chi^2_q$$

$q = l - k > 0$

Large enough OID test statistic implies that some of the IVs are invalid.

STATA

```
ivregress post-estimation commands

. estat endogenous

  Tests of endogeneity
  Ho: variables are exogenous

  Durbin (score) chi2(1)          =  2.80707  (p = 0.0938)
  Wu-Hausman F(1,423)             =  2.79259  (p = 0.0954)

. estat overid

  Tests of overidentifying restrictions:

  Sargan (score) chi2(1) =  .378071  (p = 0.5386)
  Basmann chi2(1)        =  .373985  (p = 0.5408)
```