Instrumental Variables

Ercan Karadas

New York University
Department of Economics

Spring, 2018

Table of Contents

Cases Where the OLS Estimator Cannot Be Saved

Instrumental Variables Estimator

The Generalized Instrumental Variables Estimator (GIVE)

An Alternative Presentation of GIVE

Example: Returns to Schooling

Example: Institutions and economic development

Specification Tests

Table of Contents

Cases Where the OLS Estimator Cannot Be Saved

Instrumental Variables Estimator

The Generalized Instrumental Variables Estimator (GIVE)

An Alternative Presentation of GIVE

Example: Returns to Schooling

Example: Institutions and economic development

Specification Tests

Resources:

- Cameron-Trivedi, 4.8, 4.9
- Hayashi, Ch. 3.1-3.4
- Greene, Ch. 8.1
- Johnston and Dinardo, Ch.5
- Verbeek, Ch. 5

The linear regression model (Review)

- As a reference point consider the linear regression model:

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

- The goal: estimate the **conditional mean function**: $E[y|x]$
  - This gives the change in the conditional mean given an *exogenous* change in x
  - In another word, $E[y|x]$ tells us how much we should expect y to change in response to one unit *exogenous* change in x.
  - By "exogenous change" we mean keeping everything else constant.
- Until now, it was assumed that the error term ε and the explanatory variable x were contemporaneously uncorrelated: $E[x\varepsilon] = 0$. Then
  - The regression model describes a conditional expectation

$$E[y|x] = \beta_1 + \beta_2 x$$

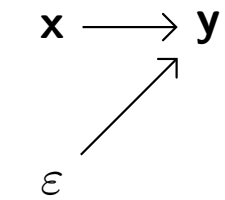
  - β_2 readily gives the expected change in y in response to one unit exogenous change in x:

$$\beta_2 = \frac{dy}{dx}$$

(When we have more than one explanatory variables $\beta_i = \frac{\partial y}{\partial x_i}$ and recall that the partial derivative assumes that all variables except x_i are kept constant)

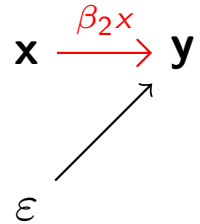
The linear regression model (Review)

- The only *expected* effect of x on y is a direct effect via the term $\beta_2 x$:

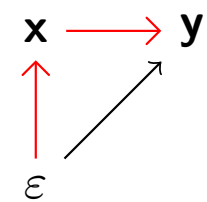


OLS

- If x and ε are correlated, x have two effects on y:

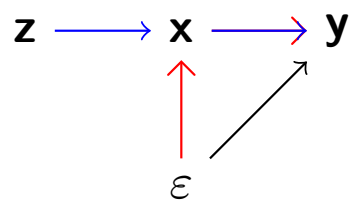


Direct Effect via $\beta_2 x$



Indirect Effect via u affecting x

- The goal of regression is to estimate *only the direct effect*: estimate β_2
- To do that we need an independent source of variation z that effects x but not correlated with ε (and y aside from the indirect route via x)



Instrumental Variables

- *z instruments* an independent source of variation in x: to get a consistent estimate of β_2 just focus on the blue route.

1. Endogeneity and omitted variable bias

- Consider a wage equation

$$y_i = \beta_1 + \beta_2 x_{2i} + \varepsilon_i$$

  where y_i is wage, x_i is years of schooling and importantly we now assume $E[x_2\varepsilon] \neq 0$.
- And suppose that $E[x_2\varepsilon] \neq 0$ because ε contains an omitted variable, say IQ, that is highly correlated with x.
- Suppose regardless the violation of the orthogonality assumption we applied OLS to get coefficient estimates

$$\hat{y}_i = b_1 + b_2 x_{2i}$$

- In order to see where can we go wrong, let's try to answer the following:
  - What is the (expected) contribution of one additional year of schooling on wage?
  - It's not b_2!
  - Because when we ask this question we mean keeping everything else constant, but we never observed the years of schooling in isolation
  - Instead b_2 contains the effect of IQ on wage as well and consequently will overestimate the true β_2.
  - In another word, we don't have enough information to differentiate what the conditional expectation corresponds to, i.e. example

$$E[\text{wage}|x = \text{college}, IQ = \text{low}] \quad \text{or} \quad E[\text{wage}|x = \text{college}, IQ = \text{high}]$$

IV as a solution to the Identification Problem

- What we have just discussed can be seen as an **identification problem** because we were unable to identify the true population parameter β_2 with the information we had
- Another classical example of this kind of identification problem arises when you regress quantity demanded on price because price is correlated with the error term. In order to identify the demand function we need an independent source of variation that effect the supply but not the price directly.
- For the details of this example see Example 8.4 in Greene.
- Later we will see that Keynesian consumption function example also suffer from the same type of identification problem.
- At this point some terminology: a variable x is called
  - an exogenous variable if $E[x\varepsilon] = 0$
  - an endogenous variable if $E[x\varepsilon] \neq 0$
- As we have seen OLS fails to provide a consistent estimator for an endogenous variable
- Therefore, this problem is also known as the **problem of endogeneity**
- Now, let's see when the problem identification (or endogeneity) arises

When can we expect $E[\mathbf{x}\varepsilon] \neq 0$?

- Sometimes there are statistical or economic reasons why we would not want to impose the condition

$$E[\varepsilon \mathbf{x}] = 0$$

- Some examples of such situations
  1) Endogeneity and Omitted Variable Bias (Omitted variable bias)
  2) Simultaneity and Reverse Causality
  3) Measurement error in the regressors
  4) Autocorrelation with a Lagged Dependent Variable
- In such cases, we can no longer argue that the OLS estimator is unbiased or consistent
- In these notes, I'll only mention the first two cases, the other two will be on the problem set

1. Endogeneity and omitted variable bias

- We will look at the omitted variable bias via a wage equation

$$y_i = \mathbf{x}_{1i}'\boldsymbol{\beta} + x_{2i}\beta_2 + \varepsilon_i$$

where

y_i : a person's log wage

$\mathbf{x}_{1i}$: a vector of individual characteristics (incl. an intercept term)

x_{2i} : years of schooling

- Omitted variable situation can arise in two ways
  - A relevant explanatory variable, say x_{2i}, is omitted from the model. This problem is simple to solve so this is not what we are interested in here
  - There are unobservable omitted factors in the model that happen to be correlated with one or more of the explanatory variables
- Suppose that the error term ε contains a variable, say a, that is correlated with the years of schooling x_2, i.e.

$$\varepsilon_i = a_i\gamma + v_i$$

where γ is a constant and v is an error term uncorrelated with a.

- The presence of an unobserved component in the equation that is potentially correlated with the observed regressors is also referred to as **unobserved heterogeneity**

1. Endogeneity and omitted variable bias

- Let's see why omitted variable bias lead to inconsistent estimates
- As we saw the *true model* for wage should be

$$y_i = \mathbf{x}_{1i}'\boldsymbol{\beta} + x_{2i}\beta_2 + a_i\gamma + v_i$$

where a is an unobserved variable reflecting ability, for simplicity say ability and the other variables as we defined before.

- Because a_i is unobserved, the econometrician simply estimates

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$$

where $\mathbf{x}_i' = (\mathbf{x}_{1i}', x_{2i})$, $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \beta_2)$ and ε_i is (mistakenly) considered to be uncorrelated with $\mathbf{x}_i$

- OLS estimator vector $\mathbf{b}$ is computed as usual

$$\mathbf{b} = \left(\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i'\right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i y_i$$

- Now substitute $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$, keeping in mind that $\varepsilon_i = a_i\gamma + v_i$

1. Endogeneity and omitted variable bias

- A little bit rearrangement yields

$$\mathbf{b} = \boldsymbol{\beta} + \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i a_i \gamma + \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i v_i$$

- Recall that $E[\mathbf{x}_i v_i] = 0$ and take plim on both sides to get

$$\text{plim } \mathbf{b} = \boldsymbol{\beta} + \mathbf{Q}_{xx}^{-1} E[\mathbf{x}_i a_i] \gamma \,.$$

  where we assumed the data matrix is well-behaved in the sense that

$$\frac{1}{N} \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i' \right) \xrightarrow{p} \mathbf{Q}_{xx} \text{ is finite and invertible}$$

- When $\gamma \neq 0$, consistency of the OLS estimator for $\boldsymbol{\beta}$ requires

$$E[\mathbf{x}_i a_i] = 0$$

  That is, the unobserved ability should be uncorrelated with schooling and the other explanatory variables in the model!

- But people with higher ability tend to end up with more schooling so this condition is very unlikely to hold in this example
- Therefore, the OLS estimator $\mathbf{b}$ is not a consistent estimator in this case

2. Simultaneity and reverse causality

- This happens if $\mathbf{x}_i$ not only has an impact on y_i, but at the same time y_i has an impact on one or more elements of $\mathbf{x}_i$, say x_{2i}.
- Example: The level of criminal activity in a city will be affected by the amount spent on law enforcement, while city officials may decide upon the budget for law enforcement partly by the expected level of criminal activity.
- For an analytical discussion consider a **Keynesian consumption function**:

$$C_i = \beta_1 + \beta_2 Y_i + \varepsilon_i$$

where

$$C_i : \text{per capita consumption}$$

$$Y_i : \text{per capita income}$$

and β_2 measures the *marginal propensity to consume*

- However, aggregate income is not exogenously given as it will be determined by the identity

$$Y_{2i} = C_i + I_i$$

where I_i denotes per capita investment

- We assume that investment is exogenous, which means that I_i and ε_i are uncorrelated, that is, $E[I_i \varepsilon_i] = 0$.

2. Simultaneity and reverse causality

$$\left.\begin{aligned} C_i &= \beta_1 + \beta_2 Y_i + \varepsilon_i \\ Y_i &= C_i + I_i \end{aligned}\right\} \quad \text{Structural Form}$$

- Since C_i influences Y_i through the equilibrium condition, we can no longer argue that Y_i and ε_i are uncorrelated.
- To see that suppose we increase C_i in the second expression by 1 unit and keep I_i constant, so Y_i will increase by 1 unit as well. Now go back to the first equation: if both C_i and Y_i increase by one unit and $0 < \beta_2 < 1$, then ε_i have to increase as well in order to satisfy the equality. Therefore, income Y_i and error term ε_i are correlated.
- Alternatively, this can be shown algebraically by deriving the **reduced form**, which describes C_i and Y_i as a function of exogenous variable(s) and error terms:

$$\left.\begin{aligned} Y_i &= \frac{\beta_1}{1-\beta_2} + \frac{1}{1-\beta_2} I_i + \frac{1}{1-\beta_2} \varepsilon_i \\ C_i &= \frac{\beta_1}{1-\beta_2} + \frac{\beta_2}{1-\beta_2} I_i + \frac{1}{1-\beta_2} \varepsilon_i \end{aligned}\right\} \quad \text{Reduced Form}$$

- The first equation shows that Y and ε are clearly correlated

2. Simultaneity and reverse causality

- In fact we can derive the exact expression for the correlation
- Note that from the first of these two equations

$$cov(Y_i, \varepsilon_i) = \frac{1}{1-\beta_2} cov(I_i, \varepsilon_i) + \frac{1}{1-\beta_2} Var(\varepsilon_i)$$
$$= \frac{\sigma^2}{1-\beta_2} \neq 0$$

- Accordingly, OLS is inconsistent for estimating the marginal propensity to consume β_2!
- Note, again, that the consumption function does not correspond to a conditional expectation, i.e. this specification does not allow us to compute

$$E[C_i | Y_i = y_i]$$

That is, we can't compute expected consumption level corresponding to some income level y_i, keeping everything else constant.

  - Because the correlation between Y and ε makes it impossible to *keep everything else constant*

Table of Contents

Cases Where the OLS Estimator Cannot Be Saved

Instrumental Variables Estimator

The Generalized Instrumental Variables Estimator (GIVE)

An Alternative Presentation of GIVE

Example: Returns to Schooling

Example: Institutions and economic development

Specification Tests

Origins of the method of IV

From **Angrist and Krueger (2001):**

- The canonical example, and earliest application, of instrumental variables involved attempts to estimate demand and supply curves.
- Economists such as P.G. Wright, Henry Schultz, Elmer Working and Ragnar Frisch were interested in estimating the elasticities of demand and supply for products ranging from herring to butter, usually with time series data.
- If the demand and supply curves shift over time, the observed data on quantities and prices reflect a set of equilibrium points on both curves. Consequently, an ordinary least squares regression of quantities on prices fails to identify-that is, trace out-either the supply or demand relationship.

Single Endogenous Regressor and a Single Instrument

- We have seen that when $E[\varepsilon_i x_{2i}] \neq 0$ the OLS method produces a biased and inconsistent estimator for the parameters in the model
- In these situations, we need an alternative estimation method
- We are going to start with a simple case: there is a $(K-1) \times 1$ vector of exogenous variables $\mathbf{x}_1$ and a single endogenous variable x_2
- Mathematically we have the following model

$$y_i = \mathbf{x}_{1i}'\boldsymbol{\beta}_1 + x_{2i}\beta_2 + \varepsilon_i$$

where the error term is related to the regressors as follows

$$E[\varepsilon_i \mathbf{x}_{1i}] = 0 \quad \text{and} \quad E[\varepsilon_i x_{2i}] \neq 0$$

- As we discussed when $E[\varepsilon_i x_{2i}] \neq 0$, we can't identify β_2 and can't obtain a consistent estimate of it
- At this point we need to impose additional assumptions to ensure that the model is identified

Single Endogenous Regressor and a Single Instrument

- We assume that there is a variable z_2 with the following two properties
  - **Exogeneity:** $E[\varepsilon_i z_{2i}] = 0$? (instrument uncorrelated to error term), and
  - **Relevance:** $cov(x_{2i}, z_{2i}) \neq 0$ (instrument correlated with endogenous regressor)
- This assumption together with $E[\varepsilon_i \mathbf{x}_{1i}] = 0$ provides us two orthogonality conditions that are called **theoretic moment conditions:**

$$\left.\begin{aligned} E\left[\varepsilon_i \mathbf{x}_{1i}\right] &= 0 \\ E\left[\varepsilon_i z_{2i}\right] &= 0 \end{aligned}\right\} \quad \text{Theoretical Moment Conditions}$$

- These moment conditions can written as

$$\left.\begin{aligned} E\left[\left(y_i - \mathbf{x}_{1i}'\boldsymbol{\beta}_1 - x_{2i}\beta_2\right)\mathbf{x}_{1i}\right] &= 0 \\ E\left[\left(y_i - \mathbf{x}_{1i}'\boldsymbol{\beta}_1 - x_{2i}\beta_2\right)z_{2i}\right] &= 0 \end{aligned}\right\} \quad \text{Theoretical Moment Conditions}$$

- Next, replace the expectations with their sample counterparts

$$\left.\begin{aligned} \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \mathbf{x}_{1i}'\boldsymbol{\beta}_1 - x_{2i}\beta_2\right)\mathbf{x}_{1i} &= 0 \\ \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \mathbf{x}_{1i}'\boldsymbol{\beta}_1 - x_{2i}\beta_2\right)z_{2i} &= 0 \end{aligned}\right\} \quad \text{Sample Moment Conditions}$$

Single Endogenous Regressor and a Single Instrument

- We have a system of K equations (the first equation contains $(K-1)$ equations and the second one contains a single equation) and K parameters to be estimated
- Solution to the system gives **the instrumental variables estimator $\mathbf{b}_{IV}$**

$$\mathbf{b}_{IV} = \left(\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^{N} \mathbf{z}_i y_i$$

  where $\mathbf{x}_i' = (\mathbf{x}_{1i}', x_{2i})$, $\mathbf{z}_i' = (\mathbf{x}_{1i}', z_{2i})$.
- Identification of the model and consistency of the IV estimator requires that the moment conditions uniquely identify the parameters of interest. This requires that the $K \times K$ matrix

$$\text{plim } \frac{1}{N} \sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i' = \mathbf{Q}_{zx}$$

  is finite and invertible. This means that the partial correlation between the instrument and the endogenous variable is nonzero.
- This requires that the coefficient π_2 in the *reduced form* equation

$$x_{2i} = \mathbf{x}_{1i}' \pi_1 + z_{2i} \pi_2 + v_i$$

  to be different from zero, which says that the endogenous regressor x_{2i} and the instrument z_{2i} have nonzero correlation, after netting out the effects of all other exogenous variables in the model.

Single Endogenous Regressor and a Single Instrument

- It can be shown that under these assumptions

$$\sqrt{N}(\boldsymbol{\beta} - \mathbf{b}) \xrightarrow{d} N\left(0, \sigma^2 \left(\mathbf{Q}_{xz}\mathbf{Q}_{zz}^{-1}\mathbf{Q}_{zx}\right)^{-1}\right)$$

  where the $K \times K$ matrix

$$\text{plim } \frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i\mathbf{z}_i' = \mathbf{Q}_{zz}$$

  is assumed to be finite and invertible.

- In finite sample, the covariance matrix can be estimated by

$$\text{Est.}V\left[\mathbf{b}_{IV}\right] = \hat{\sigma}^2 \left[\left(\sum_{i=1}^{N} \mathbf{x}_i\mathbf{z}_i'\right)^{-1}\left(\sum_{i=1}^{N} \mathbf{z}_i\mathbf{z}_i'\right)^{-1}\left(\sum_{i=1}^{N} \mathbf{z}_i\mathbf{x}_i'\right)\right]^{-1}$$

  where

$$\hat{\sigma}^2 = \frac{1}{N-K}\sum_{i=1}^{N}\left(y_i - \mathbf{x}_i'\mathbf{b}_{IV}\right)^2$$

  is a consistent estimator for σ^2 that is based on the residual sum of squares.

An alternative derivation of $\mathbf{b}_{IV}$

- There is an alternative way of obtaining $\mathbf{b}_{IV}$
- To show this method consider the simplest regression

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

where $E[\varepsilon_i x_i] \neq 0$, but we have another variable z satisfying both
  - $E[\varepsilon_i z_i] = 0$ (exogeneity)
  - $cov(x_i, z_i) \neq 0$ (relevance)
- Let us now take the covariance with z_i on both sides of to get

$$cov(y_i, z_i) = \beta_2 cov(x_i, z_i) + cov(\varepsilon_i, z_i)$$

- From the exogeneity condition $cov(\varepsilon_i, z_i) = 0$ so we can write

$$\beta_2 = \frac{cov(y_i, z_i)}{cov(x_i, z_i)}$$

- This is (theoretically) defining β_2. How to estimate it?
- Simply replace the population covariances by the sample covariances.

$$b_{2,IV} = \frac{\sum_{i=1}^{N}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{N}(z_i - \bar{z})(x_i - \bar{x})}$$

- Note that this reduces to OLS if $z_i = x_i$

Finding instruments

- Instruments need to be uncorrelated with the unobservables affecting y.
- E.g. we want to estimate a wage equation explaining earnings from schooling and other variables.
- Which factors affect schooling but not earnings directly? I.e. what affects schooling but not unobserved ability/intelligence that is determining wages?
  - Parents education?
  - Distance to school?
  - Quarter of birth?

Table of Contents

Cases Where the OLS Estimator Cannot Be Saved

Instrumental Variables Estimator

The Generalized Instrumental Variables Estimator (GIVE)

An Alternative Presentation of GIVE

Example: Returns to Schooling

Example: Institutions and economic development

Specification Tests

Moment conditions

- We will look at the general case: multiple endogenous regressors and arbitrary number of instruments
- Consider the linear model

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i$$

  where $\boldsymbol{\beta}$ is a K-dimensional vector of parameters.
- A crucial assumption is that the error term ε_i and the explanatory variables $\mathbf{x}_i$ are uncorrelated, i.e.

$$E[\varepsilon_i \mathbf{x}_i] = 0\,.$$

- This condition is absolutely essential to make OLS consistent. (Consistency of OLS also requires some additional regularity conditions.)

- Problems like
  - Measurement error in x_i
  - Endogeneity of $\mathbf{x}_i$ due to unobserved heterogeneity or reverse causality
- Lead to cases where

$$E[\varepsilon_i \mathbf{x}_i] \neq 0\,.$$

  (for one or more elements)
- In such a case OLS is necessarily inconsistent.
- Moreover, the model becomes unidentified unless we are willing to impose alternative assumptions.
- Identification is obtained if we can find an R-dimensional vector of (relevant) instruments $\mathbf{z}_i$ such that

$$E[\varepsilon_i \mathbf{z}_i] = 0 \quad (R \geq K).$$

How does that work?

- The conditions $E[\varepsilon_i \mathbf{z}_i] = 0$ are first written as

$$E[\varepsilon_i \mathbf{z}_i] = E[\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right) \mathbf{z}_i] = 0$$

- This is a set of R **moment conditions**.
- These R equations define (theoretically) what $\boldsymbol{\beta}$ is.
- We can exploit them to estimate $\boldsymbol{\beta}$.

How to estimate β?

- First, replace the expectations by sample averages.

$$\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \mathbf{x}_i'\beta\right)\mathbf{z}_i$$

- Next, choose an estimate for β that makes this as close as possible to zero.
- Why?
  - sample averages converge to population means if N becomes infinitely large, and
  - population mean is zero (only) for the true parameter values.

Different Cases: $R = K$

- $\mathbf{R} < \mathbf{K}$: we do not have enough instruments; there is an infinite number of values for $\mathbf{b}_{IV}$ such that

$$\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \mathbf{x}_i'\mathbf{b}_{IV}\right)\mathbf{z}_i = 0$$

  The model remains unidentified.

- $\mathbf{R} = \mathbf{K}$: there is (typically) one unique solution $\mathbf{b}_{IV}$ satisfying

$$\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \mathbf{x}_i'\mathbf{b}_{IV}\right)\mathbf{z}_i = 0$$

  The model is exactly identified and the solution is the instrumental variables estimator

$$\mathbf{b}_{IV} = \left(\sum_{i=1}^{N}\mathbf{z}_i\mathbf{x}_i'\right)^{-1}\sum_{i=1}^{N}\mathbf{z}_i y_i$$

$$\mathbf{b}_{IV} = \left(\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{Z}'\mathbf{y}$$

Different Cases: $R > K$

- $\mathbf{R} > \mathbf{K}$: The model is overidentified. There are more instruments than necessary for identification.
- Rather than choosing a subset of instruments, we can exploit them all by choosing $\boldsymbol{\beta}$ in such a way that the R sample moments

$$\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)\mathbf{z}_i = 0$$

  are close to zero as much as possible and this is done by minimizing a quadratic form in the sample moments, i.e.

$$\mathbf{Q}_N(\boldsymbol{\beta}) = \left[\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)\mathbf{z}_i\right]' W_N \left[\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)\mathbf{z}_i\right]$$

  where W_N is a $R \times R$ positive definite weighting matrix (that may be sample dependent), with $W_N \longrightarrow W$
- In matrix notation

$$\mathbf{Q}_N(\boldsymbol{\beta}) = \left[\frac{1}{N}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right]' W_N \left[\frac{1}{N}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right]$$

- First order conditions wrt $\boldsymbol{\beta}$ gives the resulting IV estimator as

$$\mathbf{b}_{IV} = \left(\mathbf{X}'\mathbf{Z}W_N\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}W_N\mathbf{Z}'\mathbf{y}$$

- Note that when $R = K$, the matrix $\mathbf{X}'\mathbf{Z}$ is square and by assumption invertible and the previous formula agrees with the one we gave for the case $R = K$. In this case, the weighting matrix is irrelevant
- The resulting estimator for β is consistent for any choice of weighting matrix.
- But the optimal weighting matrix yields the most efficient estimator for β by putting more weight on those sample moments that provide more accurate information on β
- The optimal weighting matrix is proportional to the inverse of the sample moments. When ε_i is $i.i.d.(0, \sigma^2)$ the optimal weighting matrix is given by

$$W_N^{opt} = \left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i\mathbf{z}_i'\right]^{-1} = \left[\frac{1}{N}\mathbf{Z}'\mathbf{Z}\right]^{-1}$$

- The resulting IV estimator is given as

$$\mathbf{b}_{GIVE} = \left(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$$

- This estimator is referred to as the **generalized instrumental variables estimator (GIVE)**

The GIVE / 2SLS estimator

- It is also known by the name the **two-stage least squares estimator (2SLS)** because the same estimator can be obtained in two-steps:
  - Estimate reduced forms (by OLS) that explain $\mathbf{x}_i$ from $\mathbf{z}_i$. Take the fitted values from these regressions. (These are interpreted as best linear approximations.)
  - Estimate the original model (by OLS) replacing the endogenous regressors by the fitted values from step 1.

  (!) It is a common mistake that the instruments themselves are included in the second stage. This is incorrect. One should include the fitted values from the reduced forms, which are linear combinations of all instruments
- The asymptotic distribution of $\mathbf{b}_{GIVE}$ is given by

$$\sqrt{N}(\mathbf{b}_{GIVE} - \beta) \xrightarrow{d} N\left(0, \sigma^2 \left(\mathbf{Q}_{xz}\mathbf{Q}_{zz}^{-1}\mathbf{Q}_{zx}\right)^{-1}\right)$$

  This is the same expression as we saw before except the dimensions of $\mathbf{Q}_{xz}$ and $\mathbf{Q}_{zz}$.
- Again in finite sample, the covariance matrix can be estimated by

$$\text{Est.}V\left[\mathbf{b}_{GIVE}\right] = \hat{\sigma}^2 \left(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}$$

  where $\hat{\sigma}^2 = \frac{1}{N-K}\sum_{i=1}^{N}\left(y_i - \mathbf{x}_i'\mathbf{b}_{GIVE}\right)^2$ is a consistent estimator for σ^2 that is based on the residual sum of squares

Keynesian Consumption Example Extended

- Recall our Keynesian Consumption example, but now suppose that there is also a government sector in the economy so the identity should change to $Y_i = C_i + I_i + G_i$ and we have the following structural form

$$\left.\begin{aligned} C_i &= \beta_1 + \beta_2 Y_i + \varepsilon_i \\ Y_i &= C_i + I_i + G_i \end{aligned}\right\} \quad \text{Structural Form}$$

- Both I_i and G_i are valid instruments and we only need one to identify the model, so we could go with either one but using *both* of them might be more efficient in which case we should form the matrices $\mathbf{X}$ and $\mathbf{Z}$ as follows

$$\mathbf{X} = [1 \quad Y], \qquad \mathbf{Z} = [1 \quad I \quad G]$$

  Now apply the formula $\mathbf{b}_{GIVE}$
- Alternatively, 2SLS can be applied as follows
  - **Step 1.** Regress endogenous variable Y_i on exogenous variables (non here) and instruments

$$Y_i = \pi_1 + \pi_2 I_i + \pi_3 G_i + v_i \qquad \text{(Reduced Form)}$$

    and obtain $\hat{Y}_i$
  - **Step 2.** Regress C_i on $\hat{Y}_i$

$$C_i = \beta_1 + \beta_2 \hat{Y}_i + \varepsilon_i$$

    to obtain the estimated $\boldsymbol{\beta}$ vector by OLS

$$\mathbf{b}_{2SLS} = \left(\hat{\mathbf{X}}'\hat{\mathbf{X}}\right)^{-1} \hat{\mathbf{X}}'\mathbf{C} = \mathbf{b}_{GIVE}$$

    where $\hat{\mathbf{X}} = [1 \quad \hat{Y}]_{N\times 2}$ and $\mathbf{C}$ is $N \times 1$ column vector.
- In the next section we will see how to express 2SLS in matrix notation

Some remarks

- Suppose we estimate a wage equation instrumenting
  - Experience, experience-squared and schooling, by
  - Age, age-squared and lived-near-college.
- Even though we choose instruments to match the endogenous regressors, for the resulting IV estimator it is irrelevant how we match instruments to individual regressors.
- All that matters is the space spanned by the instruments. (This result follows nicely form the 2SLS interpretation. Only the fitted values of the reduced form matter.)
- Instruments should be exogenous, i.e. uncorrelated with the equations error term.
- They should also be relevant, i.e. correlated with the regressors that they are supposed to be instrumenting.
- This means that in the reduced form, where we explain x_i from z_i, the instruments should be sufficiently important. (For example, lived-near-college should have a non-negligible impact upon schooling, conditional upon the other exogenous variables/ instruments.)
- Otherwise, we may have a *weak instruments* problem

Table of Contents

Cases Where the OLS Estimator Cannot Be Saved

Instrumental Variables Estimator

The Generalized Instrumental Variables Estimator (GIVE)

An Alternative Presentation of GIVE

Example: Returns to Schooling

Example: Institutions and economic development

Specification Tests

Obtaining IV Estimator within the GLS Framework

- In this section we will derive the IV estimator as a special case of the GLS estimator. Consider the linear model in matrix form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon, \qquad \text{Var}(\varepsilon) = \sigma^2\mathbf{I}$$

  where $\boldsymbol{\beta}$ is a K-dimensional vector of parameters.
- Suppose it is possible to find a data $(N \times R)$ matrix $\mathbf{Z}$ with the following properties
  - $\mathbf{Z}$ is correlated with $\mathbf{X}$ and
$$\frac{1}{N}(\mathbf{Z}'\mathbf{X}) \xrightarrow{p} \mathbf{Q}_{zx}$$
  - $\mathbf{Z}$ is uncorrelated with ε, that is
$$E\left[\mathbf{Z}'\varepsilon\right] = \mathbf{0}$$
- Premultiplying the linear model by $\mathbf{Z}'$ gives

$$\mathbf{Z}'\mathbf{y} = \mathbf{Z}'\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}'\varepsilon, \qquad \text{Var}(\mathbf{Z}'\varepsilon) = \sigma^2\left(\mathbf{Z}'\mathbf{Z}\right)\mathbf{I} \equiv \Omega$$

- Now we will proceed as in GLS framework: transform the model to get a homoscedastic variance-covarience matrix
- Decompose Ω

$$\Omega^{-1} = \mathbf{P}'\mathbf{P}$$

- Premultiply the linear model again by $\mathbf{P}$

$$\mathbf{P}\mathbf{Z}'\mathbf{y} = \mathbf{P}\mathbf{Z}'\mathbf{X}\boldsymbol{\beta} + \mathbf{P}\mathbf{Z}'\varepsilon, \qquad \text{Var}(\mathbf{P}\mathbf{Z}'\varepsilon) = \sigma^2\mathbf{I}$$

GLS Framework

- Applying OLS to this model we obtain GLS estimator vector

$$\begin{aligned}\mathbf{b}_{GLS} &= \left(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} \quad (=\mathbf{b}_{IV})\\ &= \left(\mathbf{X}'\mathbf{P}_z\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{P}_z\mathbf{y}\end{aligned}$$

  where $\mathbf{P}_z = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ is the projection of $\mathbf{X}$ on $\mathbf{Z}$.
- Note that GLS formulation shows that the IV estimator may be seen as a two-stage least square(2SLS) procedure:
  - **Step 1.** Regress each of the variables in $\mathbf{X}$ on $\mathbf{Z}$ to obtain a matrix of fitted values $\hat{\mathbf{X}}$,

$$\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} = \mathbf{P}_z\mathbf{X}$$

  - **Step 2.** Regress $\mathbf{y}$ on $\hat{\mathbf{X}}$ to obtain the estimated β vector

$$\begin{aligned}\mathbf{b}_{2SLS} &= \left(\hat{\mathbf{X}}'\hat{\mathbf{X}}\right)^{-1}\hat{\mathbf{X}}'\mathbf{y}\\ &= \left(\mathbf{X}'\mathbf{P}_z\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{P}_z\mathbf{y}\\ &= \mathbf{b}_{IV}\end{aligned}$$

- Since GLS estimator is BLUE, comparing $\mathbf{b}_{IV}$ of this section to the one in the previous section explains why we needed to take $W_N^{opt} = (\mathbf{Z}'\mathbf{Z})^{-1}$ to obtain the best combination of instruments.

GLS Framework

- The variance-covariance matrix is

$$\text{var}(\mathbf{b}_{IV}) = \sigma^2 \left(\mathbf{X}'\mathbf{P}_z\mathbf{X}\right)^{-1}$$

- and σ^2 can be estimated consistently by

$$\hat{\sigma}^2 = \frac{1}{N}\left(\mathbf{y} - \mathbf{X}\mathbf{b}_{IV}\right)'\left(\mathbf{y} - \mathbf{X}\mathbf{b}_{IV}\right)$$

- When $R = K$, that is, when $\mathbf{Z}$ and $\mathbf{X}$ has the same number of columns, $\mathbf{X}'\mathbf{Z}$ is $K \times K$ and nonsingular, therefore the $\mathbf{b}_{IV}$ reduces to

$$\mathbf{b}_{IV} = \left(\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{Z}'\mathbf{y} \quad \textit{with} \quad \text{var}(\mathbf{b}_{IV}) = \sigma^2 \left(\mathbf{Z}'\mathbf{X}\right)^{-1}\left(\mathbf{Z}'\mathbf{Z}\right)\left(\mathbf{X}'\mathbf{Z}\right)$$

GLS Framework

- When some of the $\mathbf{X}$ variables are used as instruments, we may partition $\mathbf{X}$ and $\mathbf{Z}$ as

$$\underset{N\times K}{\mathbf{X}} = \begin{bmatrix} \underset{N\times M}{\mathbf{X}_1} & \underset{N\times(K-M)}{\mathbf{X}_2} \end{bmatrix} \qquad \underset{N\times R}{\mathbf{Z}} = \begin{bmatrix} \underset{N\times M}{\mathbf{X}_1} & \underset{N\times(R-M)}{\mathbf{Z}_1} \end{bmatrix}$$

- It can be shown that $\hat{\mathbf{X}}$, the matrix of regressors in the second-stage regression, is then

$$\hat{\mathbf{X}} = \begin{bmatrix} \mathbf{X}_1 & \hat{\mathbf{X}}_2 \end{bmatrix}$$

where $\hat{\mathbf{X}}_2 = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}_2$.
That is, the variables in $\mathbf{X}_1$ serve as instruments for themselves, and the remaining second-stage regressors are the fitted values of $\mathbf{X}_2$, obtained from the regression of $\mathbf{X}_2$ on the *full* set of instruments.

Table of Contents

Cases Where the OLS Estimator Cannot Be Saved

Instrumental Variables Estimator

The Generalized Instrumental Variables Estimator (GIVE)

An Alternative Presentation of GIVE

Example: Returns to Schooling

Example: Institutions and economic development

Specification Tests

Estimating the returns to schooling: Motivation

- **Well-established empirical finding:** Better-educated individuals earn higher wages than their less-educated counterparts.
- **Question of Interest:** the higher earnings observed for better-educated workers are *caused* by their higher education, or whether individuals with greater earning capacity have chosen to acquire more schooling?
- Want: what is the effect on earnings of an exogenous increase in schooling?
- OLS estimates tend to be biased, because they reflect differences in unobserved characteristics of individuals that have attained different levels of schooling.
- This is referred to as ability bias. (Another cause of biased OLS estimates is measurement error in schooling.)
- This well-known example is based on Card 1995.

Estimating the returns to schooling

- Human capital earnings function (Mincer 1974):

$$w_i = \beta_1 + \beta_2 S_i + \beta_3 E_i + \beta_4 E_i^2 + \text{Other Control Var.} + \varepsilon_i$$

w_i : log of individual earnings

S_i : years of schooling

E_i : years of experience

- Years of experience is measured as $E_i = age_i - S_i - 6$
- "Other Control Variables" include race (`black`), and two geographical variables (`south` and `smsa`).
- As we discussed S_i causes an endogeneity problem in this model.
- To identify the model Card (1995) used a dummy variable for whether someone grew up near a four-year college (`nearc4`) as an instrumental variable for education S_i
- In IV analysis we will follow a three step process
  - **Step 1.** Do (and report) OLS. This provides a benchmark for what follows.
  - **Step 2.** Check instrument validity
  - **Step 3.** If instruments are valid, do use IV estimator

Step 1. OLS

```
# OLS Estimation
lm(lwage ~ educ + exp + exp2 + black + smsa + south)

Coefficients:
              Estimate Std. Error    t value Pr(>|t|)
(Intercept)  6.000e+00  4.542e-14  1.321e+14   <2e-16 ***
educ         1.000e+00  2.355e-15  4.246e+14   <2e-16 ***
exp          1.000e+00  4.467e-15  2.239e+14   <2e-16 ***
exp2         2.053e-16  2.136e-16  9.610e-01   0.3365
black        1.774e-14  1.184e-14  1.498e+00   0.1343
smsa         1.052e-14  1.046e-14  1.005e+00   0.3149
south       -1.838e-14  1.016e-14 -1.809e+00   0.0705 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.514e-13 on 3003 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:      1
F-statistic: 7.808e+28 on 6 and 3003 DF,  p-value: < 2.2e-16
```

Step 2. Instrument Validity

- In order to test the instrument validity, regress schooling (`educ`) on the instrument (`nearc4`) and the other exogenous variables (this is also called reduced form estimation)

```
lm(educ ~ nearc4 + exp + exp2 + black + smsa + south + nearc4)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.6591746  0.1763889  94.446  < 2e-16 ***
nearc4       0.3373208  0.0825004   4.089 4.45e-05 ***
exp         -0.4100081  0.0336939 -12.169  < 2e-16 ***
```

- In this regression, we only care whether the coefficient of `nearc4` is significant or not (that is why the rest of the output is not even reported)
- `nearc4` is highly significant, proceed to IV estimation

Step 3. IV Estimation

- For this example **X** and **Z** matrices as defined in the previous section should be formed as follows

$$\mathbf{X} = [educ \quad exp \quad exp2 \quad black \quad smsa \quad south]$$
$$\mathbf{Z} = [nearc4 \quad exp \quad exp2 \quad black \quad smsa \quad south]$$

- R commands to compute $\mathbf{b}_{IV}$ directly by using the formula (without any package)

```
ones <- rep(1,3010)
X <- cbind(ones, educ, exp, exp2, black, smsa, south)
Z <- cbind(ones, nearc4, exp, exp2, black, smsa, south)

b_IV <- solve(t(Z)%*%X)%*%t(Z)%*%lwage
b_IV =
ones   3.752781312
educ   0.132288842
exp    0.107497987
exp2  -0.002284072
black -0.130801893
smsa   0.131323663
south -0.104900534
```

Step 3. IV Estimation

- Card (1995) is a standard IV example and it's usually presented as in previous slide
- However, if you look at the definition of E_i it contains an endogenous variable, therefore E_i and E_i^2 should also be considered as endogenous variables as well
- Which means we need two additional instruments and we will use age and age^2 as two additional instruments
- In this case **X** and **Z** matrices should be revised but the rest of the code is the same:

```
ones <- rep(1,3010)
X <- cbind(ones, educ, exp, exp2, black, smsa, south)
Z <- cbind(ones, nearc4, age, age2, black, smsa, south)

b_IV <- solve(t(Z)%*%X)%*%t(Z)%*%lwage
b_IV =
ones   4.065667375
educ   0.132947268
exp    0.055961357
exp2  -0.000795658
black -0.103140265
smsa   0.107984806
south -0.098175164
```

Step 3. IV Estimation

- Once you have **X** and **Z** matrices, you can also compute variance-covariance matrix and t-statics
- Once you do these calculations you should get this table

Dependent variable: log(*wage*)

Variable	Estimate	Standard error	t-ratio
constant	4.0656	0.6085	6.682
schooling	0.1329	0.0514	2.588
exper	0.0560	0.0260	2.153
$exper^2$	−0.0008	0.0013	−0.594
black	−0.1031	0.0774	−1.333
smsa	0.1080	0.0050	2.171
south	−0.0982	0.0288	−3.413

Instruments: *age*, age^2 and *lived near college*
used for: *exper*, $exper^2$ and *schooling*

Table: Wage Equation, Estimated by IV

- R has a built-in command to run an IV regression
- For the last case, R command with package will be something like this:

```
install.packages("ivpack")
library(ivpack)
ivreg(lwage ~ educ + exp
+ exp2 + black + smsa + south , ~ nearc4 + age + age2
+ black + smsa + south)
```

Discussion

- Any IV estimate requires a choice of instruments that should be motivated. Always mention this choice.
- Reduced form explaining endogenous regressors from exogenous regressors and instruments, should show significant effect of the instruments. (If weak: weak instruments problem.)
- IV estimates are (much) less accurate than OLS (how much depends upon their correlation with the endogenous regressors).
- It is possible to use more instruments than required (overidentification).

Table of Contents

Cases Where the OLS Estimator Cannot Be Saved

Instrumental Variables Estimator

The Generalized Instrumental Variables Estimator (GIVE)

An Alternative Presentation of GIVE

Example: Returns to Schooling

Example: Institutions and economic development

Specification Tests

Institutions and economic development

- What is the causal impact of the quality of institutions (property rights enforcement, rule of law, ...) upon the economic development of a country (GDP per capita)?
- To answer that question Acemoglu, Johnson and Robinson (2001) employ the following specification

$$\ln GDP_i = \beta_1 + QI_i\beta_2 + \mathbf{x}_i\beta_3 + \varepsilon_i$$

  where $log(GDP_i)$ denotes the logarithm of GDP per capita in country i, QI_i is a measure of the quality of institutions, whereas $\mathbf{x}_i$ is a vector of other characteristics that are assumed to be exogenous, for example, related to climate or geography
- Institutional quality is potentially correlated with omitted variables, might be measured with error, and may partly be driven by GDP (reverse causality).
- This will bias OLS estimates and it is a challenge to find good instruments.
- Acemoglu et. all. (2001) argue that early settler mortality is a valid instrument.

Data

- Base sample: 64 countries
- Dependent variable: log GDP per capita
- Variable of interest: quality of institutions (QI)
- Other (control) variables: latitude, Africa and Asia dummies, and a measure of malaria risk (malfal94).
- Potential instrument: log of the mortality rate expected by the first European settles in the colonies (logem4),
- Alternative instrument: percentage of population from European descent in 1900 (euro1900).

Defending Instrument

Defending logem4 as an instrument:

- Settler mortality rates were a major determinant of settlements, which in turn were a major determinant of early institutions.
- Because early institutions are strongly correlated with current institutions, this instrument is relevant.
- When early mortality rates, conditional upon the controls in the model, have no impact on GDP per capita today, it is also exogenous.
- (Major concern: both may correlate through disease environment).

OLS Results

Dependent variable: log(*GDP*)		
Variable	(1)	(2)
constant	4.728	6.178
	(0.397)	(0.404)
QI	0.468	0.364
	(0.064)	(0.056)
latitude	1.577	0.234
	(0.710)	(0.625)
africa		−0.414
		(0.226)
asia		−0.457
		(0.221)
malfal94		−0.788
		(0.278)
R^2	0.575	0.740
Number of observations	64	62

Table: GDP Equation Estimated by OLS

Reduced Form (QI explained by exogenous variables)

Dependent variable: *QI*				
Variable	(1a)	(1b)	(2a)	(2b)
constant	8.529	7.853	7.872	5.861
	(0.812)	(0.831)	(0.963)	(0.962)
logem4 (instrument)	−0.510	−0.368	−0.328	−0.031
	(0.141)	(0.149)	(0.199)	(0.187)
euro1900 (instrument)	–	0.021	–	0.044
		(0.008)		(0.010)
latitude	2.002	0.200	1.888	−1.654
	(1.337)	(1.495)	(1.457)	(1.515)
africa			0.135	1.272
			(0.527)	(0.531)
asia			0.487	1.989
			(0.519)	(0.572)
malfal94			−0.774	−1.241
			(0.695)	(0.617)
R^2	0.296	0.367	0.322	0.493
F-test on instrument(s)	13.09	10.52	2.72	11.03
Number of observations	64	63	62	62

Table: Reduced Form, Estimated by OLS

IV Estimation

Dependent variable: log(*GDP*)				
Variable	(1a)	(1b)	(2a)	(2b)
constant	1.692	1.995	2.772	4.991
	(1.293)	(1.018)	(2.717)	(0.764)
QI (instrumented)	0.996	0.946	0.893	0.548
	(0.222)	(0.173)	(0.420)	(0.115)
latitude	−0.647	−0.597	−1.070	−0.220
	(1.335)	(1.186)	(1.425)	(0.723)
africa			−0.445	−0.425
			(0.365)	(0.247)
asia			−0.825	−0.585
			(0.455)	(0.250)
malfal94			−0.106	−0.550
			(0.691)	(0.328)
Instruments	*logem4*	*logem4 euro1900*	*logem4*	*logem4 euro1900*
Overidentifying restrictions test	–	0.069	–	1.928
(*p*-value)		(0.791)		(0.165)
Durbin–Wu–Hausman test	−4.33	−5.37	−2.14	−2.14
(*p*-value)	(0.000)	(0.000)	(0.037)	(0.037)
Number of observations	64	63	62	62

Table: GDP Equation, Estimated by IV

Discussion

- OLS results find a strong and statistically significant relationship between QI and GDP per capita.
- There are probably many omitted determinants of GDP per capita that will also correlate with institutions. We should not interpret the relation between QI and GDP as causal.
- Reduced forms show that settler mortality has significant impact on QI, but only when there are few controls in the model. With the second instrument (euro1900) added, their joint significance is okay.
- IV results show larger impact of QI than does OLS.

Table of Contents

Cases Where the OLS Estimator Cannot Be Saved

Instrumental Variables Estimator

The Generalized Instrumental Variables Estimator (GIVE)

An Alternative Presentation of GIVE

Example: Returns to Schooling

Example: Institutions and economic development

Specification Tests

Testing for endogeneity: Hausman Test

- We will consider the following model

$$y_i = \mathbf{x}_{1i}'\boldsymbol{\beta}_1 + x_{2i}\beta_2 + \varepsilon_i$$

- **Assume** that there is a valid instrument z_{2i} for the endogenous variable x_{2i}, i.e. $E[\varepsilon_i z_{2i}] = 0$.
- Under this assumption it is possible to test whether x_{2i} is endogenous, i.e. $E[\varepsilon_i x_{2i}] \neq 0$
- The basic idea is the following: When $E[\varepsilon_i z_{2i}] = 0$ we know that IV estimator is always consistent and under the null that $E[\varepsilon_i x_{2i}] = 0$ OLS estimator is consistent too and therefore they should differ by sampling error only. Under the alternative hypothesis, only the IV estimator is consistent (and OLS is inconsistent).
- Hausman based a test on the difference between the two estimators and for this particular model can be implemented as follows
  - Estimate the reduced form equation
$$x_{2i} = \mathbf{x}_{1i}'\pi_1 + z_{2i}\pi_2 + v_i$$
and save the residuals $\hat{v}_i$.
  - Add these residuals to the model and estimate
$$y_i = \mathbf{x}_{1i}'\boldsymbol{\beta}_1 + x_{2i}\beta_2 + \hat{v}_i\gamma + e_i$$
  - If $\gamma = 0$, x_{2i} is exogenous and this can be tested by performing a standard t-test on $\gamma = 0$ in the above regression

Testing Overidentification: Sargan Test

- It is not possible to test whether instruments are valid (exogenous) if they are needed to identify the model
- Here we are considering the GIVE model with K parameters and R instruments
- $R = K$: In exactly identified case, all instruments are needed to identify the model so we cannot test the instruments. Instrument validation should come from economic theory and introspection
- $R > K$: In the overidentified case, we can test the overidentifying restrictions.
- The idea is the following: since β is overidentified only K (linear combinations) of the R elements in the sample moments

$$\frac{1}{N}\sum_i e_i z_i$$

  are set equal to zero, where $e_i = y_i - \mathbf{x}_i'\mathbf{b}_{GIVE}$
- On the other hand, if the population moment conditions were true, one would expect the elements in this vector all to be sufficiently close to zero. This provides a basis for a test of the overidentification.

- Sargan showed that under the null hypothesis that all instruments are valid

$$\xi = N\mathbf{Q}_N(\mathbf{b}_{IV}) = \left[\sum_{i=1}^{N} e_i \mathbf{z}_i\right]' \left[\hat{\sigma}^2 \sum_{i=1}^{N} \mathbf{z}_i \mathbf{z}_i'\right]^{-1} \left[\sum_{i=1}^{N} e_i \mathbf{z}_i\right]' \sim \chi^2_{R-K}$$

  That is, ξ has an asymptotic Chi-squared distribution with $R - K$ degrees of freedom
- If the test rejects, the specification of the model is rejected in the sense that the sample evidence is inconsistent with the joint validity of all R moment conditions
- Without additional information it is not possible to determine which of the moments are incorrect, that is, which of the instruments are invalid

Testing for Weak Instruments

- In general, to figure out whether you have weak instruments, it is useful to examine the reduced-form regression and evaluate the explanatory power of the additional instruments that are not included in the equation of interest
- For example consider the model with one endogenous regressor x_{2i}

$$y_i = \mathbf{x}_{1i}'\boldsymbol{\beta}_1 + x_{2i}\beta_2 + \varepsilon_i$$

  where $E[\mathbf{x}_i\varepsilon_i] = 0$ and there is a valid instrument z_{2i}, $E[\varepsilon_i z_{2i}] = 0$
- Then the reduced form equation is given by

$$x_{2i} = \mathbf{x}_{1i}'\pi_1 + z_{2i}\pi_2 + v_i$$

- If $\pi_2 = 0$, the instrument z_{2i} is irrelevant and IV estimator is inconsistent
- If π is close to zero, the instruments are weak.
- The value of the F-statistic for $\pi_2 = 0$ is a measure for the information content contained in the instruments.
- As a simple rule-of-thumb based on theoretical analysis in Staiger and Stock (1997): F-statistic should exceed 10.
- Therefore, it is a good practice to compute and present the F-statistic of the reduced form in empirical work
- If the F-statistic for the significance of the instruments in the reduced form is too small, you should not put much confidence in the IV results