Instrumental Variables (IV)

Instrumental Variables (IV) is a method of estimation that is widely used in many economic applications when correlation between the explanatory variables and the error term is suspected

- for example, due to omitted variables, measurement error, or other sources of simultaneity bias

Context

$$y_i = x_i'\beta + u_i \qquad \text{for } i = 1, ..., N$$

The basic idea is that if we can replace the actual realized values of x_i (which are correlated with u_i) by predicted values of x_i that are

- related to the actual x_i

- but uncorrelated with u_i

then we can obtain a consistent estimator of β

Predicted values are formed by projecting x_i on a set of **instrumental variables** or instruments, which are required to have two important properties:

- related to the explanatory variable(s) x_i ('informative')

- uncorrelated with the errors u_i ('valid')

The first property ensures that the predicted values are related to x_i

The second property ensures that the predicted values are uncorrelated with u_i

The general problem in practice is finding instrumental variables that have *both* these properties

But assuming for the moment that we have good instruments available, we consider the method of Two Stage Least Squares (2SLS)

Note that in the context of multiple regression when some x_i variables are 'endogenous' (i.e. correlated with u_i) and other explanatory variables are 'exogenous' (i.e. uncorrelated with u_i), the instruments are required to have explanatory power for (each of) the endogenous x_i variable(s) *after conditioning* on all the remaining exogenous x_i variable(s)

Two Stage Least Squares (2SLS)

We first consider a model with a single endogenous explanatory variable (x_i) and a single instrument (z_i) - both assumed to have mean zero for simplicity

$$y_i = x_i\beta + u_i \qquad \text{for } i = 1, ..., N \qquad E(u_i) = 0, \quad E(x_i u_i) \neq 0$$

$$y = X\beta + u \qquad \text{(all vectors are } N \times 1)$$

First stage regression (linear projection)

$$x_i = z_i\pi + r_i \qquad \text{for } i = 1, ..., N$$

$$X = Z\pi + r \qquad \text{(all vectors are } N \times 1)$$

We require

$$E(z_i u_i) = 0$$

so that the instrumental variable z_i is **valid**

We require

$$\pi \neq 0 \qquad (\text{here} \leftrightarrow E(z_i x_i) \neq 0)$$

so that the instrumental variable z_i is **informative**

We estimate the first stage regression coefficient π using OLS

$$\widehat{\pi} = (Z'Z)^{-1}Z'X$$

And form the predicted values of $X = (x_1, ..., x_N)'$

$$\widehat{X} = Z\widehat{\pi} = Z(Z'Z)^{-1}Z'X$$

Second stage regression

$$y_i = \widehat{x}_i\beta + (u_i + (x_i - \widehat{x}_i)\beta) \qquad \text{for } i = 1, ..., N$$

$$y = \widehat{X}\beta + (u + (X - \widehat{X})\beta) \qquad (\text{all vectors are } N \times 1)$$

The second component of the error term is a source of finite sample bias but not inconsistency; also has to be noted when constructing standard errors

We estimate the second stage regression coefficient β using OLS

$$
\begin{aligned}
\widehat{\beta}_{2SLS} &= (\widehat{X}'\widehat{X})^{-1}\widehat{X}'y \\
&= [(Z(Z'Z)^{-1}Z'X)'(Z(Z'Z)^{-1}Z'X)]^{-1}(Z(Z'Z)^{-1}Z'X)'y \\
&= [(X'Z(Z'Z)^{-1}Z')(Z(Z'Z)^{-1}Z'X)]^{-1}(X'Z(Z'Z)^{-1}Z')y \\
&= [X'Z(Z'Z)^{-1}Z'Z(Z'Z)^{-1}Z'X]^{-1}X'Z(Z'Z)^{-1}Z'y \\
&= (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y
\end{aligned}
$$

using the symmetry of $(Z'Z)^{-1}$ s.t. $[(Z'Z)^{-1}]' = (Z'Z)^{-1}$

The 2SLS estimator can also be written as

$$\begin{aligned}\widehat{\beta}_{2SLS} &= (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y \\ &= [(Z(Z'Z)^{-1}Z'X)'X]^{-1}(Z(Z'Z)^{-1}Z'X)'y \\ &= (\widehat{X}'X)^{-1}\widehat{X}'y\end{aligned}$$

Notice that $\widehat{\beta}_{2SLS} = \widehat{\beta}_{OLS}$ in the special case where $z_i = x_i$ and hence $\widehat{x}_i = x_i$

This is very intuitive - if we project x_i on itself, we obtain perfect predictions and the second stage of 2SLS coincides with the standard OLS regression

The previous expressions for the 2SLS estimator remain valid when we have several explanatory variables (in the row vector x_i') and several instrumental variables (in the row vector z_i'), in place of the scalars x_i and z_i [although, as we will see, the requirements for $\widehat{\beta}_{2SLS}$ to be consistent become more subtle in this case]

Only in the (just-identified) special case where we have the same number of instruments as we have explanatory variables (i.e. where the row vectors x_i' and z_i' have the same number of columns), we can also express the 2SLS estimator as

$$\widehat{\beta}_{2SLS} = (Z'X)^{-1}Z'y$$

In the special case with one explanatory variable and one instrument, this follows from $\widehat{X} = Z\widehat{\pi}$ with $\widehat{\pi}$ being a scalar, using

$$\begin{aligned}
\widehat{\beta}_{2SLS} &= (\widehat{X}'X)^{-1}\widehat{X}'y = [(Z\widehat{\pi})'X]^{-1}(Z\widehat{\pi})'y \\
&= [\widehat{\pi}(Z'X)]^{-1}\widehat{\pi}(Z'y) \\
&= \frac{1}{\widehat{\pi}}(Z'X)^{-1}\widehat{\pi}(Z'y) \\
&= (Z'X)^{-1}Z'y
\end{aligned}$$

Proof of consistency

$$\begin{aligned}
\widehat{\beta}_{2SLS} &= (\widehat{X}'X)^{-1}\widehat{X}'y \\
&= (\widehat{X}'X)^{-1}\widehat{X}'(X\beta + u) \\
&= (\widehat{X}'X)^{-1}\widehat{X}'X\beta + (\widehat{X}'X)^{-1}\widehat{X}'u \\
&= \beta + \left(\frac{\widehat{X}'X}{N}\right)^{-1}\left(\frac{\widehat{X}'u}{N}\right)
\end{aligned}$$

Taking probability limits

$$p\lim_{N\to\infty}\widehat{\beta}_{2SLS} = \beta + p\lim_{N\to\infty}\left(\frac{\widehat{X}'X}{N}\right)^{-1} p\lim_{N\to\infty}\left(\frac{\widehat{X}'u}{N}\right)$$

We assume the data on (y_i, x_i, z_i) are independent over $i = 1, ..., N$

$$p \lim_{N\to\infty} \widehat{\beta}_{2SLS} = \beta + p \lim_{N\to\infty} \left(\frac{\widehat{X}'X}{N}\right)^{-1} p \lim_{N\to\infty} \left(\frac{\widehat{X}'u}{N}\right)$$

$p \lim_{N\to\infty} \left(\frac{\widehat{X}'u}{N}\right) = 0$ provided $\widehat{x}_i$ and u_i are uncorrelated, which is implied by $E(z_i u_i) = 0$

$p \lim_{N\to\infty} \left(\frac{\widehat{X}'X}{N}\right)$ exists and is non-singular, provided $\widehat{x}_i$ and x_i are correlated, which is implied by $\pi \neq 0$

Given these two properties of the instrumental variables z_i, we obtain

$$p \lim_{N\to\infty} \widehat{\beta}_{2SLS} = \beta$$

$\widehat{\beta}_{2SLS}$ is a consistent estimator of the parameter β

Alternative interpretation

The 2SLS estimator can also be interpreted as a **Generalized Method of Moments** (GMM) estimator

Write the model as

$$y_i - x_i'\beta = u_i(\beta) \qquad \text{with} \;\; E(u_i) = 0 \;\; \text{and} \;\; E(z_i u_i) = 0$$

GMM estimators choose β to minimize the weighted quadratic distance

$$u'ZW_N Z'u$$

or, equivalently, to minimize

$$\left(\frac{u'Z}{N}\right) W_N \left(\frac{Z'u}{N}\right) = \left(\frac{1}{N}\sum_{i=1}^{N} u_i z_i'\right) W_N \left(\frac{1}{N}\sum_{i=1}^{N} z_i u_i\right)$$

for some weight matrix W_N

Note that $\left(\frac{1}{N}\sum_{i=1}^{N} z_i u_i\right)$ is the sample analogue of $E(z_i u_i)$

Intuitively, GMM chooses the value of β that allows $u_i(\beta) = y_i - x_i'\beta$ to satisfy the population moment conditions $E(z_i u_i) = 0$ as closely as possible (in a weighted quadratic distance sense) in the sample

Different choices of the weight matrix W_N produce different GMM estimators, based on the same moment conditions $E(z_i u_i) = 0$

Consistency depends on the validity of the (population) moment conditions $E(z_i u_i) = 0$

Efficiency depends on the choice of the weight matrix

Setting $W_N = (Z'Z)^{-1}$ and minimizing

$$u'Z(Z'Z)^{-1}Z'u$$

yields the 2SLS estimator $\widehat{\beta}_{2SLS}$

$\widehat{\beta}_{2SLS}$ is thus *a* GMM estimator

For the linear model with independent observations and conditional homoskedasticity ($E(u_i^2|z_i) = \sigma^2$), this choice of the weight matrix $W_N = (Z'Z)^{-1}$ gives the asymptotically efficient estimator in this class

For 'just-identified' models, with the same number of instrumental variables in z_i as explanatory variables in x_i, the choice of the weight matrix is irrelevant, since the 2SLS estimator makes $u_i(\beta) = y_i - x_i'\beta$ and z_i orthogonal in the sample

This efficient GMM interpretation provides a more general rationale for 2SLS in the 'over-identified' case, where the number of instruments exceeds the number of explanatory variables, and the choice of the weight matrix matters for asymptotic efficiency

Aside - OLS is a just-identified GMM estimator, obtained by setting $z_i = x_i$

Large sample distribution theory

Asymptotic distribution theory for the 2SLS estimator (and for GMM estimators more generally) was developed by White (*Econometrica*, 1982) and Hansen (*Econometrica*, 1982)

Details can be found in many econometrics textbooks, e.g. Hayashi (2000), sections 3.5 and 3.8 (noting unconventional notation)

With independent observations, we obtain (asymptotic Normality)

$$\sqrt{N}\left(\widehat{\beta}_{2SLS} - \beta\right) \xrightarrow{D} N\left(0, V\right)$$

or $\widehat{\beta}_{2SLS} \overset{a}{\sim} N\left(\beta, avar(\widehat{\beta}_{2SLS})\right)$, with $avar(\widehat{\beta}_{2SLS}) = V/N$

In large samples, hypothesis tests about the true value of some or all of the elements of the parameter vector β can thus be conducted using familiar standard Normal or chi-squared test statistics, provided a consistent estimator of $avar(\widehat{\beta}_{2SLS})$ is available

Under the conditional homoskedasticity assumption that $E(u_i^2|z_i) = \sigma^2$, the asymptotic variance of $\widehat{\beta}_{2SLS}$ is given by

$$\begin{aligned} avar(\widehat{\beta}_{2SLS}) &= \sigma^2(X'Z(Z'Z)^{-1}Z'X)^{-1} \\ &= \sigma^2\left(\widehat{X}'\widehat{X}\right)^{-1} \end{aligned}$$

This can be estimated consistently using

$$\widehat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\widehat{u}_i^2$$

where $\widehat{u}_i = y_i - x_i'\widehat{\beta}_{2SLS}$

Notice that $\widehat{u}_i = y_i - x_i'\widehat{\beta}_{2SLS} \neq y_i - \widehat{x}_i'\widehat{\beta}_{2SLS}$, so that σ^2 is **not** estimated consistently using the residuals from the second stage regression (due to the additional $(x_i' - \widehat{x}_i')\beta$ component in the error term)

Calculating the 2SLS estimator explicitly using OLS at the second stage produces the correct parameter estimates but **not** the correct standard errors

Many software programs are now available that produce 2SLS estimates with the correct standard errors

The basic command that does this in Stata is ivregress

The syntax for specifying which explanatory variables are treated as endogenous, which explanatory variables are treated as exogenous, and which additional variables are used as instruments, is described in the Handout on *IV Estimation Using Stata*

Importantly, we can also obtain **heteroskedasticity-consistent** (White/ Huber) standard errors for the 2SLS estimator (and for GMM estimators more generally)

This is important because although the model with conditional homoskedasticity provides a useful theoretical benchmark, models with conditionally heteroskedastic errors are very common in applied research

- we do not expect shocks hitting rich and poor households to be draws from the same distribution

- we do not expect shocks hitting large and small firms to be draws from the same distribution

Provided large samples are available - so that efficiency is not a major concern - we can obtain reliable inference about parameters in linear models using estimators like 2SLS that are not efficient for models with conditionally heteroskedastic errors, but which allow heteroskedasticity-robust standard errors and related test statistics to be constructed

For the conditional heteroskedasticity case (i.e. $E(u_i^2|z_i) = \sigma_i^2$), the asymptotic variance of $\widehat{\beta}_{2SLS}$ can be estimated consistently using

$$\widehat{avar}(\widehat{\beta}_{2SLS}) = \left(\widehat{X}'\widehat{X}\right)^{-1}\left(\sum_{i=1}^{N}\widehat{u}_i^2\widehat{x}_i\widehat{x}_i'\right)\left(\widehat{X}'\widehat{X}\right)^{-1}$$

where again it is important that $\widehat{u}_i = y_i - x_i'\widehat{\beta}_{2SLS} \neq y_i - \widehat{x}_i\widehat{\beta}_{2SLS}$

Models with multiple explanatory variables

The requirements for identification (i.e. consistent estimation) of parameters using 2SLS and related instrumental variables estimators become more subtle in models with several explanatory variables

We first consider the case with one endogenous explanatory variable, several exogenous explanatory variables and several ('additional', 'external' or 'outside') instruments

$$y_i = \beta_1 + \beta_2 x_{2i} + ... + \beta_K x_{Ki} + u_i \qquad \text{for } i = 1, ..., N$$

where $E(x_{ki}u_i) = 0$ for $k = 2, ..., K-1$ but $E(x_{Ki}u_i) \neq 0$

and $E(z_{mi}u_i) = 0$ for $m = 1, ..., M$

We have 1 endogenous explanatory variable (x_{Ki}), $K-1$ exogenous explanatory variables (incl. the intercept), and M (additional) instruments

In this case there is one (relevant) first stage projection for x_{Ki}

$$x_{Ki} = \delta_1 + \delta_2 x_{2i} + ... + \delta_{K-1} x_{K-1,i} + \theta_1 z_{1i} + ... + \theta_M z_{Mi} + r_i$$

Identification of the parameter vector $\beta = (\beta_1, \beta_2, ..., \beta_K)'$ requires at least one of the θ_m coefficients to be non-zero in this first stage equation

$$x_{Ki} = \delta_1 + \delta_2 x_{2i} + ... + \delta_{K-1} x_{K-1,i} + \theta_1 z_{1i} + ... + \theta_M z_{Mi} + r_i$$

Notice that we condition here on all the exogenous explanatory variables $(1, x_2, ..., x_{K-1})$, as well as on the set of (additional) instruments

- i.e. all the variables that are specified to be uncorrelated with u_i are included in the first stage equation

Hence it is not sufficient to have one instrument whose simple correlation with the endogenous x_{Ki} is non-zero

- identification would fail if such an instrument was also linearly dependent on the exogenous explanatory variables $(1, x_{2i}, ..., x_{K-1,i})$

How does this identification condition extend to models with more than one endogenous explanatory variable?

We then have more than one relevant first stage equation

Each of these has a similar form to that illustrated above, with each of the endogenous explanatory variables projected on all of the variables that are specified to be uncorrelated with u_i

The complete system of first stage equations specifies a linear projection for each of the K explanatory variables (endogenous or exogenous) on all of the variables that are specified to be uncorrelated with u_i

In models with several endogenous explanatory variables, the sufficient condition for the parameter vector $\beta = (\beta_1, \beta_2, ..., \beta_K)'$ to be identified is expressed as a **rank condition** on the matrix of first stage regression coefficients in this complete system of first stage regression equations

We first state this for a general version of the model, then illustrate what this means for a specific example

Let $x_i' = (1, x_{2i}, ..., x_{K_i})$ be the $1 \times K$ row vector of explanatory variables, such that

$$y_i = x_i'\beta + u_i$$

with $E(x_i u_i) \neq 0$

And let z_i' be the $1 \times L$ row vector containing the intercept, any exogenous variables contained in x_i, and any additional instrumental variables that are uncorrelated with u_i (so that we have $E(z_i u_i) = 0$)

[In the previous example with one endogenous variable, we had $z_i' = (1, x_{2i}, ..., x_{K-1,i}, z_{1i}, ..., z_{Mi})$ s.t. $L = K - 1 + M$]

Order condition

$$L \geqslant K$$

This requires at least as many exogenous variables (including both exogenous explanatory variables and additional instruments) as the number of parameters we are trying to estimate

- the order condition is *necessary but not sufficient* for identification of β

Rank condition

The $L \times K$ matrix $E(z_i x_i')$ has full column rank

$$\text{rank } E(z_i x_i') = K$$

- the rank condition is *sufficient* for identification of β

The rank condition can be expressed equivalently as

$$\text{rank } \Pi = K$$

in the complete system of first stage equations

$$\begin{array}{ccccccc} X & = & Z & & \Pi & + & R \\ N \times K & , & N \times L, & & L \times K & , & N \times K \end{array}$$

obtained by regressing each element of x_i on the row vector z_i'

Note that in the special case with $K = L = 1$, the matrix Π becomes a scalar, and the rank condition simplifies to the previous requirement that $\pi \neq 0$

Example

Consider a model with $K = 4$

$$\begin{aligned} y_i &= \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + u_i \\ &= x_i'\beta + u_i \qquad\qquad \text{for } i = 1, ..., N \end{aligned}$$

with $E(u_i) = E(x_{2i}u_i) = 0$ but with $E(x_{3i}u_i) \neq 0$ and $E(x_{4i}u_i) \neq 0$, so that

$$E(x_i u_i) = E\begin{pmatrix} u_i \\ x_{2i}u_i \\ x_{3i}u_i \\ x_{4i}u_i \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ E(x_{3i}u_i) \\ E(x_{4i}u_i) \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Suppose we have 3 (additional) instruments that satisfy $E(z_{1i}u_i) = E(z_{2i}u_i) = E(z_{3i}u_i) = 0$

Then we have $L = 5$ and $E(z_i u_i) = 0$ where z_i' is the 1×5 vector $(1, x_{2i}, z_{1i}, z_{2i}, z_{3i})$, and

$$E(z_i u_i) = E \begin{pmatrix} u_i \\ x_{2i}u_i \\ z_{1i}u_i \\ z_{2i}u_i \\ z_{3i}u_i \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

The complete system of 4 first stage equations is then (for $i = 1, ..., N$)

$$
\begin{aligned}
1 &= 1 \\
x_{2i} &= x_{2i} \\
x_{3i} &= \delta_{31} + \delta_{32} x_{2i} + \theta_{31} z_{1i} + \theta_{32} z_{2i} + \theta_{33} z_{3i} + r_{3i} \\
x_{4i} &= \delta_{41} + \delta_{42} x_{2i} + \theta_{41} z_{1i} + \theta_{42} z_{2i} + \theta_{43} z_{3i} + r_{4i}
\end{aligned}
$$

which can be written as

$$
(1, x_{2i}, x_{3i}, x_{4i}) = (1, x_{2i}, z_{1i}, z_{2i}, z_{3i}) \begin{pmatrix} 1 & 0 & \delta_{31} & \delta_{41} \\ 0 & 1 & \delta_{32} & \delta_{42} \\ 0 & 0 & \theta_{31} & \theta_{41} \\ 0 & 0 & \theta_{32} & \theta_{42} \\ 0 & 0 & \theta_{33} & \theta_{43} \end{pmatrix} + (0, 0, r_{3i}, r_{4i})
$$

or

$$x_i' = z_i'\Pi + r_i' \qquad \text{for } i = 1, ..., N$$

where x_i' is 1×4, z_i' is 1×5, Π is 5×4 and r_i' is 1×4

Stacking across the N observations then gives

$$X = Z\Pi + R$$

where X is $N \times 4$, Z is $N \times 5$, Π is 5×4 and R is $N \times 4$

As before, the i^{th} row of X is x_i'

Similarly the i^{th} row of Z is z_i', and the i^{th} row of R is r_i'

Now consider

$$\Pi = \begin{pmatrix} 1 & 0 & \delta_{31} & \delta_{41} \\ 0 & 1 & \delta_{32} & \delta_{42} \\ 0 & 0 & \theta_{31} & \theta_{41} \\ 0 & 0 & \theta_{32} & \theta_{42} \\ 0 & 0 & \theta_{33} & \theta_{43} \end{pmatrix}$$

The rank condition here requires rank $\Pi = 4$

Note that this condition fails if we have either

$$\theta_{31} = \theta_{32} = \theta_{33} = 0$$

$$\text{or} \quad \theta_{41} = \theta_{42} = \theta_{43} = 0$$

The parameter vector β is not identified if we have one endogenous variable for which we have no informative (outside) instruments

The rank condition also fails if we have

$$\theta_{31} \neq 0 \quad \text{and} \quad \theta_{41} \neq 0$$

$$\text{with} \quad \theta_{32} = \theta_{33} = \theta_{42} = \theta_{43} = 0$$

The parameter vector β is not identified if we have only one informative (outside) instrument for two (or more) endogenous variables

This simplifies if there is only one endogenous variable in the model

Suppose we know that $\beta_4 = 0$ and we impose this restriction, so that now $K = 3$ and Π is the 5×3 matrix

$$\Pi = \begin{pmatrix} 1 & 0 & \delta_{31} \\ 0 & 1 & \delta_{32} \\ 0 & 0 & \theta_{31} \\ 0 & 0 & \theta_{32} \\ 0 & 0 & \theta_{33} \end{pmatrix}$$

Then the condition rank $\Pi = 3$ is satisfied if either $\theta_{31} \neq 0$ or $\theta_{32} \neq 0$ or $\theta_{33} \neq 0$, as we stated previously

The rank condition for identification can in principle be tested, although statistical tests on the rank of estimated matrices are not trivial, and the implementation of tests for identification in applications of instrumental variables is still not common

Inspection of the first stage equations for signs of an identification problem is good practice and is becoming more common

- given the increasing awareness of finite sample bias problems associated with 'weak instruments' or weak identification, that we discuss below

Testing over-identifying restrictions

In the just-identified case, when $L = K$, the GMM estimator sets $\frac{1}{N}\sum_{i=1}^{N} z_i u_i = 0$ whether or not the assumption that $E(z_i u_i) = 0$ is valid, and we cannot use this fact to test the assumption

In the over-identified case, when $L > K$, we can ask whether the minimized value of the GMM criterion function is 'small enough' to be consistent with the validity of our assumption that $E(z_i u_i) = 0$

This is known as a test of the over-identifying restrictions

For the 2SLS estimator in the linear model with conditional homoskedasticity, this test was proposed by Sargan (*Econometrica*, 1958), and is sometimes referred to as the Sargan test

For the linear model with conditional heteroskedasticity, and for non-linear models, this test was generalized for asymptotically efficient GMM estimators by Hansen (*Econometrica*, 1982), and is sometimes referred to as the Hansen test

The link is that 2SLS is the asymptotically efficient GMM estimator in the linear model with conditional homoskedasticity

[We have not discussed the asymptotically efficient GMM estimator in the linear model with conditional heteroskedasticity; you can find details in, for example, sections 3.4-3.5 of Hayashi (2000)]

For the linear model with conditional homoskedasticity, the Sargan test statistic has the form

$$J = \left(\frac{1}{\widehat{\sigma}^2}\right) \widehat{u}' Z (Z'Z)^{-1} Z' \widehat{u} \overset{a}{\sim} \chi^2(L-K)$$

under the null hypothesis that $E(z_i u_i) = 0$, where $\widehat{u}_i = y_i - x_i' \widehat{\beta}_{2SLS}$ and $\widehat{\sigma}^2$ is a consistent estimator of $E(u_i^2 | z_i) = \sigma^2$

Values in the upper tail of the $\chi^2(L-K)$ distribution are unlikely draws if the null hypothesis is true, and so would result in rejection of this null hypothesis

This version of the test of over-identifying restrictions is **not robust** to the presence of conditional heteroskedasticity, and should not be used if conditional heteroskedasticity is suspected

In that case, the heteroskedasticity-robust test of over-identifying restrictions proposed by Hansen (1982) can be used

Details can be found in section 3.6 of Hayashi (2000)

Testing for endogeneity/simultaneity bias

The motivation for using instrumental variables rather than OLS estimation is that we suspect the OLS estimates would be biased and inconsistent, as a result of correlation between the error term and one or more of the explanatory variable(s)

We hope we have used appropriate instruments, and that the 2SLS estimates are consistent, and not subject to important finite sample biases

If our instrumental variables are both valid (i.e. uncorrelated with the error term) and informative (i.e. satisfying the requirements for identification), then we should expect the 2SLS parameter estimates to be quite different from the OLS parameter estimates

Comparing these two sets of parameter estimates forms the basis for the Hausman (*Econometrica*, 1978) test for (the absence of) endogeneity or simultaneity bias

The null hypothesis is that there is no correlation between the error term and all the explanatory variables (i.e. no relevant omitted variables, no measurement error, ...)

$$H_0 : E(x_i u_i) = 0$$

The alternative hypothesis is that there is some correlation between the error term and at least one of the explanatory variables (e.g. due to omitted variables, measurement error, ...)

$$H_1 : E(x_i u_i) \neq 0$$

$$H_0 : E(x_i u_i) = 0$$

$$H_1 : E(x_i u_i) \neq 0$$

To motivate the test, we also assume that we have valid and informative instrumental variables available under the alternative

Under the null, $\widehat{\beta}_{OLS}$ is consistent

Under the alternative, $\widehat{\beta}_{OLS}$ is (biased and) inconsistent, but $\widehat{\beta}_{2SLS}$ is consistent for some choice of $z_i \neq x_i$ satisfying $E(z_i u_i) = 0$ and rank $\Pi = K$

Under the null, $\widehat{\beta}_{2SLS}$ is also a consistent estimator of β, although $\widehat{\beta}_{2SLS}$ is less efficient than $\widehat{\beta}_{OLS}$ if the null hypothesis is valid

- the set of instruments used to obtain $\widehat{\beta}_{2SLS}$ excludes some of the explanatory variables in x_i

- given that the null hypothesis is valid, these variables would be both valid and highly informative instruments, and we lose efficiency unnecessarily by not using them

The basic idea of the test is that we expect $\widehat{\beta}_{OLS}$ and $\widehat{\beta}_{2SLS}$ to be similar if the null hypothesis is correct (both are consistent estimators of the true β), but to be different if the null hypothesis is false

Equivalently, we expect the difference $(\widehat{\beta}_{2SLS} - \widehat{\beta}_{OLS})$ to be small under the null, but to be large under the alternative

The Hausman test formalizes this idea

$$h = (\widehat{\beta}_{2SLS} - \widehat{\beta}_{OLS})'[\widehat{avar}(\widehat{\beta}_{2SLS}) - \widehat{avar}(\widehat{\beta}_{OLS})]^{-1}(\widehat{\beta}_{2SLS} - \widehat{\beta}_{OLS})$$

$$\overset{a}{\sim} \chi^2(K) \qquad \text{under } H_0 : E(x_i u_i) = 0$$

where K is the number of parameters in β

$$h = (\widehat{\beta}_{2SLS} - \widehat{\beta}_{OLS})'[\widehat{avar}(\widehat{\beta}_{2SLS}) - \widehat{avar}(\widehat{\beta}_{OLS})]^{-1}(\widehat{\beta}_{2SLS} - \widehat{\beta}_{OLS})$$

Although $[\widehat{avar}(\widehat{\beta}_{2SLS}) - \widehat{avar}(\widehat{\beta}_{OLS})]$ is asymptotically positive definite, it is possible that we cannot invert our estimate of this covariance matrix in finite samples

- in this case we can either form a Hausman test statistic based on a subset of the parameter vector β (typically we focus on the coefficients on the explanatory variables that are treated as endogenous under the alternative), or use one of the alternative forms of the test discussed, for example, in Chap 6 of Wooldridge, *Econometric Analysis of Cross Section and Panel Data*

$$h = (\widehat{\beta}_{2SLS} - \widehat{\beta}_{OLS})'[\widehat{avar}(\widehat{\beta}_{2SLS}) - \widehat{avar}(\widehat{\beta}_{OLS})]^{-1}(\widehat{\beta}_{2SLS} - \widehat{\beta}_{OLS})$$

This version of the test is simple to compute, but requires $E(u_i^2|z_i) = \sigma^2$, and is **not robust** to conditional heteroskedasticity

The simplicity follows from the conditional homoskedasticity assumption, under which $\widehat{\beta}_{OLS}$ is asymptotically efficient under the null hypothesis, and $avar(\widehat{\beta}_{2SLS} - \widehat{\beta}_{OLS})$ simplifies to $[avar(\widehat{\beta}_{2SLS}) - avar(\widehat{\beta}_{OLS})]$

Alternative, heteroskedasticity-robust versions of the test are available, and are discussed, for example, in Chapter 6 of Wooldridge, *Econometric Analysis of Cross Section and Panel Data*

Finite sample problems

Instrumental variables estimators like 2SLS are consistent if the instruments used are both valid and informative, but they may be subject to important finite sample biases

We now look at two distinct sources of finite sample bias

- the use of 'too many' instruments relative to the available sample size

- the use of instruments that are only weakly related to the endogenous variable(s), resulting in 'weak identification' of the parameters of interest

Overfitting

Suppose we have a single endogenous explanatory variable x_i and we use L valid instruments $(z_{1i}, z_{2i}, ..., z_{Li})$ in a sample of size N

What would happen if $L = N$?

In the first stage regression

$$x_i = z_{1i}\pi_1 + z_{2i}\pi_2 + ... + z_{Li}\pi_L + r_i$$

we then have as many parameters π_l as we have data points

In this case, we obtain a perfect fit $(R^2 = 1)$...

...the predicted values are equal to the actual values $(\widehat{x}_i = x_i)$...

...and replacing the actual x_i by the predicted $\widehat{x}_i$ achieves nothing

In this case the second stage regression coincides with the standard OLS regression: $\widehat{\beta}_{2SLS} = \widehat{\beta}_{OLS}$

$\widehat{\beta}_{2SLS}$ will therefore have exactly the same bias as $\widehat{\beta}_{OLS}$

While no-one would use this many instruments in practice, a similar finite sample bias occurs in less extreme circumstances

As the number of instruments (L) approaches the sample size (N), the 2SLS estimator $\widehat{\beta}_{2SLS}$ tends towards the OLS estimator $\widehat{\beta}_{OLS}$

Since the OLS estimator is expected to be biased, the 2SLS estimator may also have a serious finite sample bias (in the direction of the OLS estimator) if 'too many' instruments are used relative to the sample size - even though all the instruments are valid

Thus if we find that $\widehat{\beta}_{2SLS}$ and $\widehat{\beta}_{OLS}$ are similar, we should check that this is not the result of using 'too many' instruments, before concluding that there is no inconsistency in the OLS estimates

One simple way to investigate this is to calculate a sequence of 2SLS estimates based on smaller and smaller subsets of the original instruments

- and check that there is no systematic tendency for the 2SLS estimates to move away from the OLS estimates, in a particular direction, as we reduce the dimension of the instrument set

Moral - avoid the temptation to use 'too many' instruments, particularly when this results in precise 2SLS estimates that are close to their OLS counterpart

Weak instruments

'Weak instruments' describes the situation where the instruments used only weakly identify the parameters of interest

When there is a single instrument for a single endogenous variable, this corresponds to the instrument being only weakly correlated with the endogenous variable

Weak identification is an important concern in many contexts

This introduces two distinct problems, depending on whether the weak instruments are **strictly** valid instruments, or not

i) finite sample bias

We first assume the instruments used are strictly valid ($E(z_i u_i) = 0$), in which case the 2SLS estimator is consistent if the parameters are formally identified

Still there can be an important finite sample bias if the instruments provide very little (additional) information about the endogenous variable(s)

To explore this, we focus on the model with one endogenous variable and one instrument, and consider the extreme case in which the instrument is completely uninformative (s.t. β is not identified)

In the first stage regression

$$x_i = z_i\pi + r_i$$

the *true* value of $\pi = 0$

However when we estimate π in any finite sample, we will generally obtain an *estimate* $\widehat{\pi} \neq 0$, as a result of sampling variation

In this case, the decomposition of the endogenous x_i into $\widehat{x}_i = z_i\widehat{\pi}$ and $(x_i - \widehat{x}_i)$ that we use in the second stage regression is essentially arbitrary

$$y_i = \widehat{x}_i\beta + (u_i + (x_i - \widehat{x}_i)\beta)$$

It is no surprise that the 2SLS estimator does not have desirable properties

In this case, the distribution of the 2SLS estimator has the same expected value as the OLS estimator - see Bound, Jaeger and Baker, *Journal of the American Statistical Association*, 1995

[The two estimators do not coincide in this case; the 2SLS estimator will have much higher variance than the corresponding OLS estimator]

More generally, when the instruments used are valid in the sense of being uncorrelated with u_i, but only weakly correlated with x_i, $\widehat{\beta}_{2SLS}$ can have an important finite sample bias in the direction of $\widehat{\beta}_{OLS}$

If the instruments used are weak enough, even sample sizes that appear enormous are not sufficient to eliminate the possibility of quantitatively important finite sample biases

ii) large inconsistency

The second problem associated with weak instruments is that any inconsistency of $\widehat{\beta}_{2SLS}$, which will be present if the instruments have *some* correlation with u_i, gets magnified if the instruments are *also* only weakly correlated with x_i

Intuition may suggest that the bias of 2SLS should be lower than the bias of OLS, if the correlation between the error term and the instruments is much lower than the correlation between the error term and the endogenous variable(s)

This intuition can be **seriously wrong** in situtations where the instruments are only weakly correlated with the explanatory variable(s)

To see this, recall that in the case of a single explanatory variable we have

$$p \lim_{N\to\infty} \widehat{\beta}_{OLS} = \beta + p \lim_{N\to\infty} \left(\frac{X'X}{N}\right)^{-1} p \lim_{N\to\infty} \left(\frac{X'u}{N}\right)$$

$$= \beta + \frac{p \lim_{N\to\infty} \left(\frac{X'u}{N}\right)}{p \lim_{N\to\infty} \left(\frac{X'X}{N}\right)}$$

Also recall that in the case of a single explanatory variable we have

$$p \lim_{N\to\infty} \widehat{\beta}_{2SLS} = \beta + p \lim_{N\to\infty} \left(\frac{\widehat{X}'X}{N} \right)^{-1} p \lim_{N\to\infty} \left(\frac{\widehat{X}'u}{N} \right)$$

$$= \beta + \frac{p \lim_{N\to\infty} \left(\frac{\widehat{X}'u}{N} \right)}{p \lim_{N\to\infty} \left(\frac{\widehat{X}'X}{N} \right)}$$

Now compare

$$p \lim_{N\to\infty} \widehat{\beta}_{OLS} = \beta + \frac{p\lim_{N\to\infty}\left(\frac{X'u}{N}\right)}{p\lim_{N\to\infty}\left(\frac{X'X}{N}\right)}$$

$$\text{and} \qquad p \lim_{N\to\infty} \widehat{\beta}_{2SLS} = \beta + \frac{p\lim_{N\to\infty}\left(\frac{\widehat{X}'u}{N}\right)}{p\lim_{N\to\infty}\left(\frac{\widehat{X}'X}{N}\right)}$$

A high covariance between x_i and u_i could translate into a modest inconsistency for $\widehat{\beta}_{OLS}$ if the variance of x_i is sufficiently large

While a small covariance between $\widehat{x}_i$ and u_i could translate into a large inconsistency for $\widehat{\beta}_{2SLS}$ if the covariance between $\widehat{x}_i$ and x_i is also very small

The latter is relevant when the instruments z_i are only **weakly correlated** with x_i (i.e. when $\widehat{x}_i = z_i'\widehat{\pi}$ with $\widehat{\pi} \approx 0$)

Morals

- avoid using weak instruments whenever possible

- avoid attaching too much significance to 2SLS estimates in situations where the best available instruments are nevertheless very weak

- look carefully at the first stage regression equations relating the endogenous x_i variables to the instruments (and any exogenous x_i variables) to check whether the (additional) instruments have reasonable explanatory power for each of the endogenous variables

- for discussion of some suggested diagnostic tests, see Bound, Jaeger and Baker (JASA, 1995) and the discussions in Chapter 5 of Wooldridge, *Econometric Analysis of Cross Section and Panel Data* or Chapter 4 of Cameron and Trivedi, *Microeconometrics: Methods and Applications*

The latest version of the ivreg2 command in Stata now implements a number of these suggestions