Instrumental Variables (IV)

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

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

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

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

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

First stage 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

$$\mathrm{E}(z_i u_i) = 0$$

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

We require

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

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

NB. From the linear projection

$$x_i = z_i\pi + r_i$$

we have

$$z_i x_i = z_i^2\pi + z_i r_i$$

with $\mathrm{E}(z_i r_i) = 0$ from the definition of a linear projection

Then

$$\mathrm{E}(z_i x_i) = \mathrm{E}(z_i^2)\pi$$

so that $\mathrm{E}(z_i x_i) \neq 0 \leftrightarrow \pi \neq 0$ provided $\mathrm{E}(z_i^2) = \mathrm{Var}(z_i) > 0$, which in turn is implied by $\mathrm{E}(z_i x_i) \neq 0$

We estimate the first stage coefficient π using OLS

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

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

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

Substituting $\widehat{x}_i$ for x_i, the second stage model is

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

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

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 coefficient β using OLS

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

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

The 2SLS estimator can also be written as

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

Notice that $\widehat{\beta}_{2\text{SLS}} = \widehat{\beta}_{\text{OLS}}$ 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 original model

In our current setting with a single valid and informative instrument (z_i) for a single endogenous explanatory variable (x_i), the scalar parameter β is said to be 'just-identified'

More generally, the $K \times 1$ parameter vector β is said to be **just-identified** when we have exactly K valid and informative instruments

β is **under-identified** if we have strictly less than K valid and informative instruments

β is **over-identified** if we have strictly more than K valid and informative instruments

[Requirements for instruments to be 'informative' when $K > 1$ will be discussed later]

If and only if β is just-identified, we can also express the 2SLS estimator as

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

In our current setting with $K = 1$, this follows from $\widehat{\pi}$ being a scalar in the estimated first stage linear projection used to obtain $\widehat{X} = Z\widehat{\pi}$, so that

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

Two Stage Least Squares (2SLS)

Before considering properties of this 2SLS estimator, we generalise to a model with a single endogenous explanatory variable (x_i) and a set of L instruments ($z_i = (z_{1i}, z_{2i}, ..., z_{Li})'$, an $L \times 1$ column vector) - all assumed to have mean zero for simplicity

As before, we have

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

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

Now the first stage linear projection is

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

$$X = Z\pi + r$$

in which z_i' is a $1 \times L$ row vector, Z is an $n \times L$ matrix, and π is an $L \times 1$ column vector

We now require

$$\mathrm{E}(z_i u_i) = 0$$

where $z_i u_i$ is an $L \times 1$ column vector, so that **all** of the instrumental variables in z_i are **valid**

We now require

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

where both π and $z_i x_i$ are $L \times 1$ column vectors, so that **at least one** of the instrumental variables in z_i is **informative**

This requirement can also be stated as

$$\text{rank } \pi = 1$$

or as

$$\text{rank } \mathrm{E}(z_i x_i) = 1$$

NB. From the linear projection

$$x_i = z_i'\pi + r_i$$

we now have

$$z_i x_i = (z_i z_i')\pi + z_i r_i$$

with $\mathrm{E}(z_i r_i) = 0$ from the definition of a linear projection

Then

$$\mathrm{E}(z_i x_i) = \mathrm{E}(z_i z_i')\pi$$

so that $\mathrm{E}(z_i x_i) \neq 0 \leftrightarrow \pi \neq 0$ provided $\mathrm{E}(z_i z_i') = \mathrm{Var}(z_i) > 0$, which in turn is implied by $\mathrm{E}(z_i x_i) \neq 0$

We estimate the first stage coefficient vector π using OLS

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

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

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

Substituting $\widehat{x}_i$ for x_i, the second stage model is again

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

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

We estimate the second stage coefficient β using OLS

As before, this 2SLS estimator of β may be expressed as

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

$$= (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y$$

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

Consistency

To establish conditions under which $\widehat{\beta}_{2\text{SLS}}$ is a consistent estimator of β, we express $\widehat{\beta}_{2\text{SLS}}$ in the form

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

Taking probability limits

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

We assume the data on (y_i, x_i, z_i') are independently and identically distributed, with $\mathrm{E}(z_i u_i) = 0$ and $\mathrm{E}(z_i x_i) \neq 0 \leftrightarrow \pi \neq 0$

From the Law of Large Numbers, the vector of sample means

$$\frac{1}{n}\sum_{i=1}^{n} z_i u_i \xrightarrow{\mathrm{P}} \mathrm{E}(z_i u_i) = 0$$

We can also write $\frac{1}{n}\sum_{i=1}^{n} z_i u_i = \frac{1}{n}(Z'u)$, so we have

$$\left(\frac{Z'u}{n}\right) \xrightarrow{\mathrm{P}} 0$$

$\widehat{\pi}$ is a consistent estimator of the coefficient vector π in the first stage linear projection, so we also have $\widehat{\pi} \xrightarrow{\mathrm{P}} \pi \neq 0$

Since $\widehat{X} = Z\widehat{\pi}$, we can express

$$\left(\frac{\widehat{X}'u}{n}\right) = \left(\frac{(Z\widehat{\pi})'u}{n}\right) = \left(\frac{(\widehat{\pi}'Z')u}{n}\right) = \widehat{\pi}'\left(\frac{Z'u}{n}\right)$$

Then using Slutsky's theorem, we have

$$\operatorname{p}\lim_{n\to\infty}\left(\frac{\widehat{X}'u}{n}\right) = \pi'0 = 0$$

Similarly, from the Law of Large Numbers, the vector of sample means

$$\frac{1}{n}\sum_{i=1}^{n} z_i x_i \xrightarrow{\text{P}} \text{E}(z_i x_i) = M_{ZX} \neq 0$$

where $M_{ZX} = \text{E}(z_i x_i)$ is an $L \times 1$ column vector

We can also write $\frac{1}{n}\sum_{i=1}^{n} z_i x_i = \frac{1}{n}(Z'X)$, so we have

$$\left(\frac{Z'X}{n}\right) \xrightarrow{\text{P}} M_{ZX} \neq 0$$

Since $\widehat{X} = Z\widehat{\pi}$, we can express

$$\left(\frac{\widehat{X}'X}{n}\right) = \left(\frac{(Z\widehat{\pi})'X}{n}\right) = \left(\frac{(\widehat{\pi}'Z')X}{n}\right) = \widehat{\pi}'\left(\frac{Z'X}{n}\right)$$

Then using Slutsky's theorem, we have

$$\text{p}\lim_{n\to\infty}\left(\frac{\widehat{X}'X}{n}\right) = \pi' M_{ZX} \neq 0$$

and

$$\text{p}\lim_{n\to\infty}\left(\frac{\widehat{X}'X}{n}\right)^{-1} = (\pi' M_{ZX})^{-1} \text{ is finite}$$

Now recalling that

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

we have shown:

i) $\operatorname{p}\lim_{n\to\infty}\left(\frac{\widehat{X}'u}{n}\right) = 0$, given the instrument validity condition $\mathrm{E}(z_i u_i) = 0$

ii) $\operatorname{p}\lim_{n\to\infty}\left(\frac{\widehat{X}'X}{n}\right)^{-1}$ exists and is finite, given the instrument informativeness condition $\mathrm{E}(z_i x_i) \neq 0 \leftrightarrow \pi \neq 0$

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

$$\operatorname{p}\lim_{n\to\infty}\widehat{\beta}_{2\text{SLS}} = \beta \quad \text{or} \quad \widehat{\beta}_{2\text{SLS}} \xrightarrow{\text{P}} \beta$$

Alternative interpretation

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

Write the model (NB. still with $K = 1$, for simplicity) as

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

The L orthogonality conditions $\mathrm{E}(z_i u_i) = 0$ are population moment conditions

GMM estimators work with the corresponding sample analogues, i.e. sample means

In the **just-identified case** (here with $L = 1$), the GMM estimator of β is the value of β which sets the sample mean

$$\frac{1}{n}\sum_{i=1}^{n} z_i u_i(\beta) = 0$$

or which makes the residuals $\widehat{u}_i = y_i - x_i\widehat{\beta}_{\text{GMM}}$ orthogonal to the instrument z_i in the sample

This coincides with the 2SLS estimator $\widehat{\beta}_{\text{2SLS}}$

In the **over-identified case** (here with $L > 1$), we cannot (in general) set L sample correlations equal to zero through the choice of a scalar parameter

In this case GMM estimators choose β to minimize a weighted quadratic distance

$$\left(\frac{1}{n}\sum_{i=1}^{n} u_i(\beta) z_i'\right) W_n \left(\frac{1}{n}\sum_{i=1}^{n} z_i' u_i(\beta)\right) = \left(\frac{u(\beta)'Z}{n}\right) W_n \left(\frac{Z'u(\beta)}{n}\right)$$

or, equivalently, to minimize

$$u(\beta)'ZW_nZ'u(\beta)$$

for some positive definite $L \times L$ weight matrix W_n

Intuitively, GMM estimators find the value of β that allows the residuals $\widehat{u}_i = y_i - x_i\widehat{\beta}_{\text{GMM}}$ to satisfy the population moment conditions $\text{E}(z_i u_i) = 0$ 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 $\text{E}(z_i u_i) = 0$

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

Efficiency depends on the choice of the weight matrix

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

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

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

In the over-identified case, $\widehat{\beta}_{\text{2SLS}}$ is thus *a* GMM estimator, for this particular choice of the weight matrix

For the linear model with independently and identically distributed observations and conditional homoskedasticity ($\text{E}(u_i^2|z_i) = \sigma^2$), this choice of the weight matrix $W_n = (Z'Z)^{-1}$ can be shown to give the asymptotically efficient estimator in this class

This asymptotically 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

Large sample distribution results

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

Derivations will be provided later in the course

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

With instruments that are both valid and informative, and observations that are independently and identically distributed, we have an asymptotic normality result

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

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

In large finite samples, hypothesis tests about the true value of one or more of the elements of the parameter vector β can thus be conducted using familiar standard normal or chi-squared Wald test statistics, provided a consistent estimator of V is available

Under the conditional homoskedasticity assumption that $\mathrm{E}(u_i^2|z_i) = \sigma^2$ for all $i = 1, 2, ..., n$, a consistent estimator of the limit variance matrix V is given by

$$\begin{aligned}\widehat{V} &= n\widehat{\sigma}^2(X'Z(Z'Z)^{-1}Z'X)^{-1} \\ &= n\widehat{\sigma}^2\left(\widehat{X}'\widehat{X}\right)^{-1}\end{aligned}$$

where σ^2 can be estimated consistently using

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

and $\widehat{u}_i = y_i - x_i'\widehat{\beta}_{2\mathrm{SLS}}$ [see next slide]

This gives an estimator of the variance of $\widehat{\beta}_{2\text{SLS}}$ as

$$\widehat{\text{Var}}(\widehat{\beta}_{2\text{SLS}}) = \frac{\widehat{V}}{n} = \widehat{\sigma}^2 \left(\widehat{X}'\widehat{X} \right)^{-1}$$

Notice that $\widehat{u}_i = y_i - x_i'\widehat{\beta}_{2\text{SLS}} \neq y_i - \widehat{x}_i'\widehat{\beta}_{2\text{SLS}}$, so that σ^2 is **not** estimated consistently using the residuals from the second stage specification (due to the additional $(x_i' - \widehat{x}_i')\beta$ component in the second stage 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 note on *IV Estimation Using Stata*

Importantly, we can also obtain **heteroskedasticity-consistent** (White) 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 common in applied research

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

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

Provided large finite samples are available - so that efficiency is not a key 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. $\mathrm{E}(u_i^2|z_i) \neq \sigma^2$ for all $i = 1, 2, ..., n$), the variance of $\widehat{\beta}_{2\mathrm{SLS}}$ can be estimated using

$$\widehat{\mathrm{Var}}(\widehat{\beta}_{2\mathrm{SLS}}) = \frac{\widehat{V}}{n} = \left(\widehat{X}'\widehat{X}\right)^{-1}\left(\sum_{i=1}^{n}\widehat{u}_i^2\widehat{x}_i\widehat{x}_i'\right)\left(\widehat{X}'\widehat{X}\right)^{-1}$$

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

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

and $\mathrm{E}(z_{mi} u_i) = 0$ 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 (non-trivial) 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 non-trivial 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 necessary and sufficient condition for the parameter vector $\beta = (\beta_1, \beta_2, ..., \beta_K)'$ to be identified can be expressed as a **rank condition** on the matrix of first stage coefficients in this complete system of first stage linear projections

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 $\mathrm{E}(x_i u_i) \neq 0$

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

[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 $\mathrm{E}(z_i x_i')$ has full column rank

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

- the rank condition is *necessary and 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 projecting 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 $\mathrm{E}(u_i) = \mathrm{E}(x_{2i}u_i) = 0$ but with $\mathrm{E}(x_{3i}u_i) \neq 0$ and $\mathrm{E}(x_{4i}u_i) \neq 0$, so that

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

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

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

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

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 estimated 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 will be discussed later in the course

NB. From the complete system of first stage projections

$$x_i' = z_i'\Pi + r_i'$$

we have

$$z_i x_i' = (z_i z_i')\Pi + z_i r_i'$$

with $\mathrm{E}(z_i r_i') = 0$ from the definition of linear projections

Then we have

$$\mathrm{E}(z_i x_i') = \mathrm{E}(z_i z_i')\Pi$$

from which follows the equivalence between the two statements of the rank condition: rank $\mathrm{E}(z_i x_i') = K$ and rank $\Pi = K$

With independently and identically distributed observations on (y_i, x_i', z_i'), the 2SLS estimator is consistent provided we have:

i) the instrument validity condition, $\mathrm{E}(z_i u_i) = 0$

ii) the instrument informativeness rank condition, rank $\mathrm{E}(z_i x_i') = K$

The proof of consistency in this more general setting follows essentially the same arguments used previously in the special case with $K = 1$

The limit distribution results and resulting estimators for the variance of $\widehat{\beta}_{\mathrm{2SLS}}$ were stated to cover the more general setting with $K > 1$