

The instrumental variable estimator, and the generalized instrumental variable estimator, and diagnostic tests

Econometrics for minor Finance, Lecture 7

Tunga Kantacı, Fall 2025

The IV estimator

\mathbf{z}_i

is $L \times 1$ vector of instruments.

\mathbf{x}_i

is $K \times 1$ vector of endogenous regressors.

Suppose that $L = K$. That is, there are **as many instruments as there are endogenous regressors**. In this case we say that the model is **exactly identified**. For ease of exposition we consider that L and K are 1. This leads to the

$\hat{\beta}_{IV}$

estimator.

The IV estimator

$$y_i = \beta x_i + u_i$$

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

$$\mathbb{E}[z_i y_i] = \mathbb{E}[\beta z_i x_i] + \mathbb{E}[z_i u_i]$$

$$\mathbb{E}[z_i y_i] = \beta \mathbb{E}[z_i x_i]$$

$$(\mathbb{E}[z_i x_i])^{-1} \mathbb{E}[z_i y_i] = \beta (\mathbb{E}[z_i x_i])^{-1} \mathbb{E}[z_i x_i]$$

$$(\mathbb{E}[z_i x_i])^{-1} \mathbb{E}[z_i y_i] = \beta$$

The IV estimator

This derivation uses two assumptions that we have already seen.

The relevance assumption:

$$E[z_i x_i] \neq 0$$

so that the inverse exists, and the exogeneity assumption:

$$E[z_i u_i] = 0$$

The IV estimator

$$\beta = \frac{E[z_i y_i]}{E[z_i x_i]} = \frac{\text{plim} \frac{1}{n} \sum_{i=1}^n z_i y_i}{\text{plim} \frac{1}{n} \sum_{i=1}^n z_i x_i}$$

Expected value terms are population terms, so they are unobserved. We can estimate them using sample data, which gives the IV estimator:

$$\hat{\beta}_{IV} = \frac{\frac{1}{n} \sum_{i=1}^n z_i y_i}{\frac{1}{n} \sum_{i=1}^n z_i x_i} = \frac{\sum_{i=1}^n z_i y_i}{\sum_{i=1}^n z_i x_i}$$

The IV estimator: Sampling distribution: Finite sample properties

In finite samples, the IV estimator

$$\hat{\beta}_{IV}$$

can be biased because the sample expectations in the numerator and denominator are noisy. This includes cases where the instruments are weakly correlated with the endogenous variable, or where the sample moments are imprecisely estimated due to limited sample size. The bias diminishes as the sample size increases, making the IV estimator consistent in large samples. Therefore, **we rely on the large sample properties of**

$$\hat{\beta}_{IV}$$

The IV estimator: Sampling distribution: Large sample properties: Consistency

$$\hat{\beta}_{IV}$$

is consistent if the standard IV model assumptions hold;
homoskedasticity is not required. We skip the proof.

The IV estimator: Sampling distribution: Large sample properties: Asymptotic efficiency

$$\hat{\beta}_{IV}$$

is asymptotically efficient. We skip the proof.

The IV estimator: Sampling distribution: Finite sample properties: Asymptotic normality

$$\hat{\beta}_{IV} \stackrel{a}{\sim} N \left[\beta, \frac{\sigma^2}{n} \frac{E[z_i z_i]}{E[z_i x_i] E[z_i x_i]} \right]$$

The IV estimator: Sampling distribution: Finite sample properties: Asymptotic normality

The asymptotic variance of the estimator

$$\text{Asy. Var} \left[\hat{\beta}_{IV} \right] = \frac{\sigma^2}{n} \frac{E[z_i z_i]}{E[z_i x_i] E[z_i x_i]}$$

is unobserved.

The IV estimator: Sampling distribution: Finite sample properties: Asymptotic normality

We can estimate

$$\sigma$$

with

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_{IV} x_i)^2$$

and

$$\frac{E[z_i z_i]}{E[z_i x_i] E[z_i x_i]}$$

with

$$\frac{\frac{1}{n} \sum_{i=1}^n z_i z_i}{\frac{1}{n} \sum_{i=1}^n z_i x_i \frac{1}{n} \sum_{i=1}^n z_i x_i}$$

The IV estimator: Sampling distribution: Finite sample properties: Asymptotic normality

$$\text{Est. Asy. Var} \left[\hat{\beta}_{IV} \right] = \hat{\sigma}^2 \frac{\sum_{i=1}^n z_i z_i}{\sum_{i=1}^n z_i x_i \sum_{i=1}^n z_i x_i}$$

The generalized IV estimator

\mathbf{z}_i

is $L \times 1$ vector of instruments, and

\mathbf{x}_i

is a $K \times 1$ vector of endogenous regressors.

Suppose that $L > K$. That is, there are **more instruments than there are endogenous variables**. That is, we have more information than we strictly need to proxy an endogenous variable. In this case we say that the model is **overidentified**.

The generalized IV estimator

We have

$$L > K$$

Should we then just use an arbitrary selection of K instruments, and throw away the remaining instruments? No. Throwing away useful information leads to an inefficient IV estimator. Linear combinations of the L instruments can also satisfy the relevance and exogeneity assumptions. This leads to an estimator at least as efficient as the $\hat{\beta}_{IV}$ estimator:

$$\hat{\beta}_{GIV}$$

The generalized IV estimator

Assume the simplest case where $L = 2$ and $K = 1$.

The generalized IV estimator

The linear regression model is

$$y_i = \beta x_i + u_i$$

where

$$x_i$$

is an endogenous variable, and

$$z_{1i}$$

and

$$z_{2i}$$

are two instruments available.

The generalized IV estimator

The generalized IV estimator

$$\hat{\beta}_{GIV}$$

is obtained in two stages.

The generalized IV estimator: Stage one

Using the OLS method, estimate

$$x_i = \pi_1 z_{1i} + \pi_2 z_{2i} + v_i$$

where

$$\pi_1$$

and

$$\pi_2$$

are the first-stage coefficients of the instruments.

The generalized IV estimator: Stage one

Using the OLS method, obtain the coefficient estimates

$$\hat{\pi}_1$$

and

$$\hat{\pi}_2$$

The generalized IV estimator: Stage one

Using the coefficient estimates, obtain the predicted values

$$\hat{x}_i = \hat{\pi}_1 z_{1i} + \hat{\pi}_2 z_{2i}$$

The generalized IV estimator: Stage two

Using the predicted values as a regressor, and using the OLS method, estimate the equation

$$y_i = \beta \hat{x}_i + u_i^*$$

where

$$u_i^* = \beta \hat{v}_i + u_i$$

The generalized IV estimator

How we end up with

$$u_i^* = \beta \hat{v}_i + u_i$$

Considering that there is only one endogenous variable,

$$x_i = \pi z_i + v_i$$

Then,

$$x_i = \hat{x}_i + \hat{v}_i$$

Replacing x_i in

$$y_i = \beta x_i + u_i$$

we have

$$y_i = \beta \hat{x}_i + \beta \hat{v}_i + u_i$$

and

$$u_i^* := \beta \hat{v}_i + u_i$$

We will use this slide to come up with a test of endogeneity.

The generalized IV estimator: Stage two

The OLS estimator in this model takes the familiar form

$$\hat{\beta} = \hat{\beta}_{GIV} = \frac{\sum_{i=1}^n \hat{x}_i y_i}{\sum_{i=1}^n \hat{x}_i^2}$$

The generalized IV estimator

Since the estimator is obtained in two stages, textbooks call it the [two-stage least squares estimator](#), and denote it as TSLS or 2SLS.

The generalized IV estimator: Sampling distribution: Finite sample properties

$$\hat{\beta}_{GIV}$$

is biased in a finite sample, just like the

$$\hat{\beta}_{IV}$$

Therefore, we rely on the asymptotic properties of the estimator.

The generalized IV estimator: Sampling distribution: Large sample properties: Consistency

$$\hat{\beta}_{GIV}$$

is consistent. The proof is very similar to that of

$$\hat{\beta}_{IV}$$

The generalized IV estimator: Sampling distribution: Large sample properties: Asymptotic efficiency

Asymptotic variance of

$$\hat{\beta}_{GIV}$$

is equal to or **smaller** than that of

$$\hat{\beta}_{IV}$$

That is,

$$\hat{\beta}_{GIV}$$

is at least as efficient as the

$$\hat{\beta}_{IV}$$

We do not prove this. Intuitively, by exploiting additional exogenous variation from multiple instruments, the generalized IV estimator achieves lower sampling variance.

The generalized IV estimator: Sampling distribution: Large sample properties: Asymptotic normality

Derivation of the asymptotic normality of

$$\hat{\beta}_{GIV}$$

is similar to that of

$$\hat{\beta}_{IV}$$

The generalized IV estimator: Asymptotic normality

For one endogenous regressor and multiple instruments:

$$\hat{\beta}_{GIV} \stackrel{a}{\sim} N \left[\beta, \frac{\sigma^2}{n} \left[\sum_{j=1}^m \frac{E[z_{ij}x_i] E[z_{ij}x_i]}{E[z_{ij}] E[z_{ij}]} \right]^{-1} \right]$$

The GIV estimator: Sampling distribution: Asymptotic normality

The asymptotic variance of the estimator

$$\text{Asy. Var} \left[\hat{\beta}_{GIV} \right] = \frac{\sigma^2}{n} \left[\sum_{j=1}^m \frac{\mathbb{E}[z_{ij}x_i]\mathbb{E}[z_{ij}x_i]}{\mathbb{E}[z_{ij}z_{ij}]} \right]^{-1}$$

is unobserved.

The GIV estimator: Sampling distribution: Asymptotic normality

We can estimate

$$\sigma^2$$

with

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_{GIV} x_i)^2$$

and

$$\frac{\mathbb{E}[z_{ij}x_i]\mathbb{E}[z_{ij}x_i]}{\mathbb{E}[z_{ij}z_{ij}]}$$

with

$$\frac{\frac{1}{n} \sum_{i=1}^n z_{ij}x_i \frac{1}{n} \sum_{i=1}^n z_{ij}x_i}{\frac{1}{n} \sum_{i=1}^n z_{ij}z_{ij}}$$

The GIV estimator: Sampling distribution: Asymptotic normality

$$\text{Est. Asy. Var} \left[\hat{\beta}_{GIV} \right] = \hat{\sigma}^2 \left[\sum_{j=1}^m \frac{\sum_{i=1}^n z_{ij} x_i \sum_{i=1}^n z_{ij} x_i}{\sum_{i=1}^n z_{ij} z_{ij}} \right]^{-1}$$

The generalized IV estimator: Note one

The estimated asymptotic variance can be written in an alternative form as

$$\text{Est. Asy. Var} \left[\hat{\beta}_{GIV} \right] = \frac{\hat{\sigma}^2}{\sum_{i=1}^n \hat{x}_i^2}$$

This form looks familiar from the standard OLS estimator except that instead of a regressor we have its predicted version as we are in the IV framework.

The generalized IV estimator: Note two

If $L = 1$ so that the number of instruments equals the number of endogenous regressors, the IV and GIV estimators coincide:

$$\hat{\beta}_{GIV} = \hat{\beta}_{IV} = \frac{\sum_{i=1}^n z_i y_i}{\sum_{i=1}^n z_i x_i}$$

The generalized IV estimator: Example

```
. reg lwage educ age age2 black
```

Source	SS	df	MS	Number of obs	=	2,220
Model	88.0908302	4	22.0227076	F(4, 2215)	=	143.09
Residual	340.908673	2,215	.153909108	Prob > F	=	0.0000
Total	428.999503	2,219	.193330105	R-squared	=	0.2053
				Adj R-squared	=	0.2039
				Root MSE	=	.39231

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.0385118	.0032895	11.71	0.000	.032061	.0449627
age	.1326507	.0555628	2.39	0.017	.0236901	.2416113
age2	-.0015523	.0009674	-1.60	0.109	-.0034494	.0003448
black	-.2127221	.0232691	-9.14	0.000	-.2583537	-.1670906
_cons	3.315457	.7883061	4.21	0.000	1.769561	4.861354

The generalized IV estimator: Example

```
. ivregress 2sls lwage (educ = motheduc fatheduc) age age2 black, first
```

First-stage regressions

Number of obs	=	2,220
F(5, 2214)	=	157.81
Prob > F	=	0.0000
R-squared	=	0.2628
Adj R-squared	=	0.2611
Root MSE	=	2.2244

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
educ					
age	.9804534	.314502	3.12	0.002	.3637036 1.597203
age2	-.0160649	.0054764	-2.93	0.003	-.0268043 -.0053256
black	-.1607076	.1376706	-1.17	0.243	-.4306846 .1092694
motheduc	.1975247	.0201066	9.82	0.000	.1580948 .2369545
fatheduc	.2230658	.0167964	13.28	0.000	.1901275 .2560042
_cons	-5.389924	4.472077	-1.21	0.228	-14.15983 3.379979

The generalized IV estimator: Example

Instrumental variables (2SLS) regression

Number of obs	=	2,220
Wald chi2(4)	=	503.26
Prob > chi2	=	0.0000
R-squared	=	0.1900
Root MSE	=	.39564

lwage	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
educ	.0600324	.0069201	8.68	0.000	.0464692 .0735955
age	.1094726	.0564143	1.94	0.052	-.0010974 .2200426
age2	-.0011585	.0009819	-1.18	0.238	-.003083 .0007659
black	-.1833938	.0248831	-7.37	0.000	-.2321638 -.1346237
_cons	3.354017	.7950635	4.22	0.000	1.795721 4.912313

Instrumented: educ

Instruments: age age2 black motheduc fatheduc

The generalized IV estimator: Example

Note that exogenous explanatory variables, age and black, are also used as instruments since they can offer exogenous variation to the endogenous variable x_i that is not related to the regression error u_i .

- They are already part of the model and observed directly.
- By definition, they are not correlated with u_i .
- Using them as instruments strengthens identification and improves efficiency.

The generalized IV estimator: Example

In the estimation output, why the standard normal, z , and not the t distribution, t , is used?

Test whether a regressor is in fact endogenous: The Hausman test

We had

$$x_i = \pi_1 z_{1i} + \pi_2 z_{2i} + v_i$$

where

$$\pi_1$$

and

$$\pi_2$$

are the first-stage coefficients associated with the two instruments.

Test whether a regressor is in fact endogenous: The Hausman test

From the estimation of

$$x_i = \pi_1 z_{1i} + \pi_2 z_{2i} + v_i$$

obtain the residuals

$$\hat{v}_i$$

These residuals represent the endogenous variation in x_i after the exogenous variation is netted out.

Test whether a regressor is in fact endogenous: The Hausman test

Include the first-stage residuals as a regressor in the main equation

$$y_i = \beta x_i + \delta \hat{v}_i + u_i$$

Then,

$$\hat{v}_i$$

should have no explanatory power if x_i is exogenous. If this is the case, it is just noise and it is fine if it ends up in u_i .

Test whether a regressor is in fact endogenous: The Hausman test

For the original model

$$y_i = \beta x_i + u_i$$

this gives us a test of whether x_i is endogenous:

H_0 : $\delta = 0$: The residuals do not matter: x_i is exogenous in the original model: OLS estimator is consistent.

H_1 : $\delta \neq 0$: The residuals matter: x_i contains an endogenous component correlated with u_i in the original model: IV is required.

Test whether the instruments are relevant: First-stage F test for instrument relevance

We test instrument relevance in the first-stage regression:

$$x_i = \pi_0 + \pi_1 z_{1i} + \pi_2 z_{2i} + \cdots + \pi_L z_{Li} + \nu_i$$

H_0 : $\pi_1 = \pi_2 = \cdots = \pi_L = 0$: Instruments do not explain x_i .

H_1 : At least one $\pi_j \neq 0$: Instruments are relevant.

This the F test of joint significance of instruments in the first stage. $F > 10$ indicates sufficiently strong instruments. Otherwise, that is, weak instruments inflate variance and bias.

Test whether the instruments are relevant: First-stage F test for instrument relevance: Example

```
. ivregress 2sls lwage (educ = motheduc fatheduc) age age2 black, first
```

First-stage regressions

Number of obs	=	2,220
F(5, 2214)	=	157.81
Prob > F	=	0.0000
R-squared	=	0.2628
Adj R-squared	=	0.2611
Root MSE	=	2.2244

educ	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	.9804534	.314502	3.12	0.002	.3637036 1.597203
age2	-.0160649	.0054764	-2.93	0.003	-.0268043 -.0053256
black	-.1607076	.1376706	-1.17	0.243	-.4306846 .1092694
motheduc	.1975247	.0201066	9.82	0.000	.1580948 .2369545
fatheduc	.2230658	.0167964	13.28	0.000	.1901275 .2560042
_cons	-5.389924	4.472077	-1.21	0.228	-14.15983 3.379979

Test whether the instruments are exogenous: Sargan Hansen J test

We test instrument exogeneity using the overidentification test:

H_0 : Instruments are exogenous: Uncorrelated with u_i .

H_1 : At least one instrument is endogenous: Correlated with u_i .

Procedure:

- Estimate the equation by IV or GIV.
- Obtain residuals \hat{u}_i .
- Regress \hat{u}_i on all instruments z_{ij} .
- Compute the test statistic $J = n \cdot R^2$.
- Decision rule: Under H_0 , $J \sim \chi^2_{L-K}$, where L = number of instruments and K = number of endogenous regressors.

Test whether the instruments are exogenous: Sargan
Hansen J test: Intuition

The intuition is that residuals are the leftover variation in y_i after using instruments. If instruments still explain this leftover, they are contaminated by the same forces causing endogeneity, so they are not valid.

Test whether the instruments are exogenous: Sargan Hansen J test: Intuition

An important note is that the J test can be used only if there are more instruments than there are endogenous variables. That is, if

$$L > K$$

If

$$L = K$$

we cannot test for exogeneity of the instruments. The reason is technical, but the rough intuition is that when $L = K$, the IV procedure fits the data exactly using all available instruments, so there is no remaining variation that could reveal whether the instruments are invalid.