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

Econometrics for minor Finance, Lecture 7

Tunga Kantarcı, Fall 2025

The IV estimator

$$\boldsymbol{z}_i$$

is $L \times 1$ vector of instruments.

$$\boldsymbol{x}_i$$

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. This leads to the

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

estimator.

The IV estimator

For ease of exposition, consider that

$$L = K = 1$$

The IV estimator

$$
\begin{aligned}
y_i &= \beta x_i + u_i \\
z_i y_i &= \beta z_i x_i + z_i u_i \\
\mathrm{E}[z_i y_i] &= \mathrm{E}\left[\beta z_i x_i\right] + \mathrm{E}\left[z_i u_i\right] \\
\mathrm{E}\left[z_i y_i\right] &= \beta \mathrm{E}\left[z_i x_i\right] \\
\left(\mathrm{E}\left[z_i x_i\right]\right)^{-1} \mathrm{E}\left[z_i y_i\right] &= \beta \left(\mathrm{E}\left[z_i x_i\right]\right)^{-1} \mathrm{E}\left[z_i x_i\right] \\
\left(\mathrm{E}\left[z_i x_i\right]\right)^{-1} \mathrm{E}\left[z_i y_i\right] &= \beta
\end{aligned}
$$

The IV estimator

This derivation uses two assumptions that we have already seen. The relevance assumption:

$$\mathrm{E}\left[z_i x_i\right] \neq 0$$

so that the inverse exists, and the exogeneity assumption:

$$\mathrm{E}\left[z_i u_i\right] = 0$$

The IV estimator

$$\beta = \frac{\mathrm{E}\left[z_i y_i\right]}{\mathrm{E}\left[z_i x_i\right]} = \frac{\operatorname{plim} \frac{1}{n} \sum_{i=1}^{n} z_i y_i}{\operatorname{plim} \frac{1}{n} \sum_{i=1}^{n} z_i x_i}$$

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

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

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

The asymptotic variance of the estimator

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

is unobserved.

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

We can estimate

$$\sigma$$

with

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

and

$$\frac{\mathrm{E}\left[z_i z_i\right]}{\mathrm{E}\left[z_i x_i\right]\mathrm{E}\left[z_i x_i\right]}$$

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

$$\boldsymbol{z}_i$$

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

$$\boldsymbol{x}_i$$

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

Suppose now that $L > K$. That is, there are more instruments than there are endogenous regressors. 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

We 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 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 form is familiar. It is just the OLS estimator on transformed

$$x_i$$

which is

$$\hat{x}_i$$

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

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

That is, the former is at least as efficient as the latter. 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:

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

The GIV estimator: Sampling distribution: Asymptotic normality

The asymptotic variance of the estimator

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

is unobserved.

The GIV estimator: Sampling distribution: Asymptotic normality

We can estimate

$$\sigma^2$$

with

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

and

$$\frac{\mathrm{E}\left[z_{ij}x_i\right]\mathrm{E}\left[z_{ij}x_i\right]}{\mathrm{E}\left[z_{ij}z_{ij}\right]}$$

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

```
. regress cigarcons cigarprice income
```

Source	SS	df	MS		
Model	27962.8385	2	13981.4193	Number of obs	= 96
				F(2, 93)	= 36.50
Residual	35622.9814	93	383.042811	Prob > F	= 0.0000
				R-squared	= 0.4398
				Adj R-squared	= 0.4277
Total	63585.8199	95	669.32442	Root MSE	= 19.571

cigarcons	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
cigarprice	-.3595778	.0486057	-7.40	0.000	-.4560992	-.2630565
income	-2.70e-08	1.77e-08	-1.53	0.130	-6.22e-08	8.12e-09
_cons	163.4619	6.885415	23.74	0.000	149.7888	177.135

The generalized IV estimator: Example

```
. ivregress 2sls cigarcons (cigarprice = cigartax cigartaxspecific) income, first

First-stage regressions
```

Number of obs	96
F(3, 92)	172.33
Prob > F	0.0000
R-squared	0.8489
Adj R-squared	0.8440
Root MSE	17.3340

cigarprice	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
income	1.61e-09	1.59e-08	0.10	0.919	-2.99e-08	3.32e-08
cigartax	-.6899014	.648958	-1.06	0.291	-1.978788	.5989853
cigartaxspecific	2.653167	.5462455	4.86	0.000	1.568276	3.738057
_cons	44.5173	5.146638	8.65	0.000	34.29564	54.73897

The generalized IV estimator: Example

Instrumental-variables 2SLS regression

Number of obs	=		**96**
Wald chi2(**2**)	=		**75.49**
Prob > chi2	=		**0.0000**
R-squared	=		**0.4362**
Root MSE	=		**19.324**

cigarcons	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
cigarprice	**-.3969339**	**.0526937**	**-7.53**	**0.000**	**-.5002116**	**-.2936562**
income	**-2.24e-08**	**1.77e-08**	**-1.27**	**0.204**	**-5.71e-08**	**1.22e-08**
_cons	**168.362**	**7.37324**	**22.83**	**0.000**	**153.9108**	**182.8133**

Endogenous: **cigarprice**
Exogenous: **income cigartax cigartaxspecific**

The generalized IV estimator: Example

Note that exogenous the explanatory variable, income, is also used as an instrument since it can offer exogenous variation to the endogenous variable x_i that is not related to the regression error u_i.

- It is already part of the model and observed directly.
- By definition, it is not correlated with u_i.
- Using it as an instrument 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 a regressor is in fact endogenous: The Hausman test: Example

```
. regress cigarprice cigartax income
```

Source	SS	df	MS	Number of obs	=	96
				F(2, 93)	=	198.48
Model	148248.931	2	74124.4657	Prob > F	=	0.0000
Residual	34731.305	93	373.454892	R-squared	=	0.8102
				Adj R-squared	=	0.8061
Total	182980.236	95	1926.10775	Root MSE	=	19.325

cigarprice	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
cigartax	2.410446	.1305053	18.47	0.000	2.151289	2.669604
income	1.40e-08	1.75e-08	0.80	0.423	-2.06e-08	4.87e-08
_cons	39.15717	5.604335	6.99	0.000	28.02807	50.28626

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

```
. predict vhat, resid
```

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

```
. regress cigarcons cigarprice income vhat
```

Source	SS	df	MS	Number of obs	=	96
				F(3, 92)	=	25.45
Model	28835.9644	3	9611.98814	Prob > F	=	0.0000
Residual	34749.8555	92	377.715821	R-squared	=	0.4535
				Adj R-squared	=	0.4357
Total	63585.8199	95	669.32442	Root MSE	=	19.435

cigarcons	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
cigarprice	-.3978933	.0544495	-7.31	0.000	-.5060348	-.2897518
income	-2.23e-08	1.78e-08	-1.25	0.214	-5.78e-08	1.31e-08
vhat	.1788652	.1176441	1.52	0.132	-.0547861	.4125164
_cons	168.4879	7.594561	22.19	0.000	153.4044	183.5713

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 cigarcons (cigarprice = cigartax cigartaxspecific) income, first

First-stage regressions
```

Number of obs	96
F(3, 92)	172.33
Prob > F	0.0000
R-squared	0.8489
Adj R-squared	0.8440
Root MSE	17.3340

cigarprice	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
income	1.61e-09	1.59e-08	0.10	0.919	-2.99e-08	3.32e-08
cigartax	-.6899014	.648958	-1.06	0.291	-1.978788	.5989853
cigartaxspecific	2.653167	.5462455	4.86	0.000	1.568276	3.738057
_cons	44.5173	5.146638	8.65	0.000	34.29564	54.73897

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.

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

Instrumental-variables 2SLS regression

Number of obs	=	**96**	
Wald chi2(**2**)	=	**75.49**	
Prob > chi2	=	**0.0000**	
R-squared	=	**0.4362**	
Root MSE	=	**19.324**	

cigarcons	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
cigarprice	**-.3969339**	**.0526937**	**-7.53**	**0.000**	**-.5002116**	**-.2936562**
income	**-2.24e-08**	**1.77e-08**	**-1.27**	**0.204**	**-5.71e-08**	**1.22e-08**
_cons	**168.362**	**7.37324**	**22.83**	**0.000**	**153.9108**	**182.8133**

Endogenous: **cigarprice**
Exogenous: **income cigartax cigartaxspecific**

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

```
. predict residuals, residuals
```

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

```
. regress residuals cigartax cigartaxspecific income
```

Source	SS	df	MS	Number of obs	=	96
				F(3, 92)	=	0.00
Model	2.22509246	3	.741697486	Prob > F	=	0.9999
Residual	35847.0092	92	389.641405	R-squared	=	0.0001
				Adj R-squared	=	-0.0325
Total	35849.2343	95	377.360361	Root MSE	=	19.739

residuals	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
cigartax	-.0557755	.7390117	-0.08	0.940	-1.523517	1.411966
cigartaxspecific	.0457516	.6220461	0.07	0.942	-1.189686	1.281189
income	-1.10e-10	1.81e-08	-0.01	0.995	-3.60e-08	3.58e-08
_cons	.1807147	5.86082	0.03	0.975	-11.45938	11.82081

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

```
. display "overidentification statistic:" e(N)*e(r2)
overidentification statistic:.00595853

. display "p-value:" chiprob(1,e(N)*e(r2))
p-value:.93847117
```