University of Ljubljana

SEB SCHOOL OF ECONOMICS AND BUSINESS

7. Instrumental Variables Estimation

Prof. Dr. Miroslav Verbič

miroslav.verbic@ef.uni-lj.si

www.miroslav-verbic.si

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION AMBA ACCREDITED

Ljubljana, October 2025

University of Ljubljana

SEB SCHOOL OF ECONOMICS AND BUSINESS

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION AMBA ACCREDITED

I am indebted to Shingo Takahashi from the International University of Japan and Todd A. Gormley from the Wharton School for permission to employ and modify their lecture materials.

Basic definitions

Endogeneity introduced by *violating* the following assumption:

$$E\left(u\middle|X\right)=0$$

CONDITIONAL MEAN INDEPENDENCE

(necessary for consistency and unbiasedness of the OLS estimator)

Basic definitions

Situations, in which endogeneity appears:

- Omitted explanatory variables, if correlated with the included explanatory variables;
- Simultaneity related to dependent and explanatory variables;
- Dynamics related to including lagged dependent variable(s) as explanatory variable(s);
- Measurement error in the dependent and/or explanatory variables.

One explanatory variable case (bivariate regression model)

- Consider the following regression:

$$\log(wage) = \beta_0 + \beta_1 educ + u; \quad u = \beta_2 abil + v$$

- Since ability is not observed, we can only run the following regression:

$$\log(wage) = \beta_0 + \beta_1 educ + u$$

- Since ability is correlated with *educ*, *educ* is endogenous (i.e., correlated with u). Thus, $\hat{\beta}_1$ will be biased.

One explanatory variable case (bivariate regression model)

- ✓Initially, we have two methods to eliminate the bias:

  (1) Plug in the proxy variable for ability, such as IQ.

  (2) Use panel data method (either the fixed effect or the first differenced model).

- ✓Instrumental variables estimator is another method to eliminate the bias.

Instrumental variables (IV) estimator

- Consider the following model:

$$y = \beta_0 + \beta_1 x + u$$

- Suppose that x is endogenous, that is $\text{Cov}(x, u) \neq 0$.
- Further, suppose that you have another variable, z, which satisfies the following conditions:

1. $\text{Cov}(z, u)=0$ (instrument exogeneity)
2. $\text{Cov}(z, x)\neq 0$ (instrument relevance)

- If the above conditions are satisfied, we call z an instrumental variable (an instrument).

Instrumental variables (IV) estimator

✓There are two ways to intuitively understand these conditions:

1. Instrumental variable is a variable that is *not* correlated with the omitted variable, but *is* correlated with the endogenous explanatory variable.

2. Instrumental variable z is a variable that affects y only through x.

Instrumental variables (IV) estimator

- The condition $\text{Cov}(z, u)=0$ involves unobserved u. Therefore, we cannot test this condition (when you have extra instrumental variables, you can test this; this will be discussed later).
- The condition $\text{Cov}(z, x)\neq 0$ is easy to test. Just run the following OLS regression:

  $x = \pi_0+\pi_1 z+v$ and then test: H_0: $\pi_1=0$.

Instrumental variables (IV) estimator

- The R^2 for IV regression is computed as:

$$R^2 = 1 - RSS/TSS,$$

where RSS is the sum of the squared IV residuals.

- Unlike in the case of OLS, RSS can be greater than TSS. Thus, R^2 can be negative.
- In IV regression, R^2 does not have a natural interpretation.

Finding an instrumental variable

- The most difficult part of the instrumental variable estimation is to find suitable instrumental variables.
- Consider the following regression:

$$\log(wage) = \beta_0 + \beta_1 educ + \underbrace{(\beta_2 abil + v)}_{u}$$

- Then, you have to find z that is correlated with *educ*, but not correlated with *abil*. What can be z?

Finding an instrumental variable

- ✓ Consider the father's education. Perhaps a person whose father is highly educated tends to take more education as well. So the father's education is likely correlated with *educ*.
- ✓ But, for father's education to be an instrument, this should not be correlated with the unobserved ability. A highly educated father may nurture his child better, so father's education may be correlated with the unobserved ability. If this is the case, father's education is not a good instrument.
- ✓ Nonetheless, many studies have used father's and mother's education as instruments.

An example, computer printout in Stata

We run the following regression using OLS:

$$\log(wage) = \beta_0 + \beta_1 educ + u$$

Using the father's education as an instrument for *educ*, we estimate the same model using IV regression. We also check if father's education is correlated with *educ*.

An example, computer printout in Stata

OLS regression:

```
. reg lwage educ
```

Source	SS	df	MS
Model	26.3264193	1	26.3264193
Residual	197.001022	426	.462443713
Total	223.327441	427	.523015084

Number of obs	428
F(1, 426)	56.93
Prob > F	0.0000
R-squared	0.1179
Adj R-squared	0.1158
Root MSE	.68003

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.1086487	.0143998	7.55	0.000	.0803451	.1369523
_cons	-.1851968	.1852259	-1.00	0.318	-.5492673	.1788736

An example, computer printout in Stata

IV regression:

```
. ivregress 2sls lwage (educ= fatheduc)

Instrumental variables (2SLS) regression          Number of obs   =        428
                                                  Wald chi2(1)    =       2.85
                                                  Prob > chi2     =     0.0914
                                                  R-squared       =     0.0934
                                                  Root MSE        =     .68778
```

lwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
educ	.0591735	.0350596	1.69	0.091	-.009542	.127889
_cons	.4411034	.4450583	0.99	0.322	-.4311947	1.313402

Instrumented: educ
Instruments: fatheduc

An example, computer printout in Stata

Check if father's education is correlated with *educ*:

```
. reg educ fatheduc, robust
```

Number of obs	=	428			
F(1, 426)	=	87.12			
Prob > F	=	0.0000			
R-squared	=	0.1726			
Adj R-squared	=	0.1706			
Root MSE	=	2.0813			

educ	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
fatheduc	.2694416	.0288675	9.33	0.000	.2127013	.326182
_cons	10.23705	.2718861	37.65	0.000	9.702646	10.77146

IV estimator versus the OLS estimator

What would happen if you use the IV method when the suspected endogenous variable is in fact exogenous?

- ✓ Controlling for endogeneity (i.e. using IV method) when it is actually exogenous is costly in terms of efficiency (precision).

IV estimator versus the OLS estimator

Poor instruments: What would happen if the instrumental variable does not satisfy the instrument conditions?

✓ Answer to this question is the following:

1. IV estimators are inconsistent.
2. The directions of the biases in IV estimators and OLS estimators can even be the opposite.
3. The bias in IV can be worse than OLS.

Extension to multiple regression model

- ✓ We will extend the discussion to the multiple regression model and explain the following three cases, step by step:

Case 1: One endogenous variable, one instrument.

Case 2: One endogenous variable, more than one instrument (two stage least squares).

Case 3: More than one endogenous variable, more than one instrument (two stage least squares).

One endogenous variable, one instrument

- Consider the following regression:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exp + u$$

- Suppose that *educ* is endogenous, but experience is exogenous.
- To explain IV estimation for multiple regression, it is often useful to use different notation for endogenous end exogenous variable.

One endogenous variable, one instrument

- Let us use y for endogenous variables (i.e. correlated with u) and z for exogenous variables (i.e. uncorreated with u).
- Then, we can write the model as:

$$y_1=\beta_0+\beta_1 y_2+\beta_2 z_1+u$$

where y_1 is log(*wage*), y_2 is *educ*, and z_1 is *exp*.

One endogenous variable, one instrument

- ✓This model is called the structural equation to emphasize that this equation shows the causal relationship. Of course, OLS cannot be used to consistently estimate the parameters, since y_2 is endogenous.
- ✓If you have an instrument for y_2, you can consistently estimate the model. Let us call this instrument z_2.

One endogenous variable, one instrument

- ✓ As before, z_2 should satisfy (i) instrument exogeneity, and (ii) instrument relevance.
- ✓ For a multiple regression model, these conditions are written as:
  1. **The instrument exogeneity:**

     $\text{Cov}(z_2, u)=0$

  2. **The instrument relevance:**

     $y_2=\pi_0+\pi_1 z_1+\pi_2 z_2+v$ and $\pi_2 \neq 0$

All the exogenous variables included. This equation is often called the reduced form equation.

- ✓ In addition, z_2 should **not** be a part of the structural equation. This is called **the exclusion restriction**.

Instrumental variables (IV) estimator

- Now, we have the following three (population) conditions that can be used to obtain the IV estimators:

$$E(u)=0$$
$$Cov(z_1, u)=0$$
$$Cov(z_2, u)=0$$ (this is from the instrument exogeneity)

The sample counterparts of these conditions are given on the next slide.

Instrumental variables (IV) estimator

$$\sum_{i=1}^{n} \hat{u}_i = 0 \qquad \sum_{i=1}^{n} (y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0$$

If you divide it by n, this is the sample average of $\hat{u}$.

$$\sum_{i=1}^{n} z_{i1}\hat{u}_i = 0 \qquad \sum_{i=1}^{n} z_{i1}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0$$

If you divide it by $n-1$, this is the sample covariance between z_1 and $\hat{u}$.

$$\sum_{i=1}^{n} z_{i2}\hat{u}_i = 0 \qquad \sum_{i=1}^{n} z_{i2}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0$$

If you divide it by $n-1$, this is the sample covariance between z_2 and $\hat{u}$.

- ✓ This is a set of three equations with three unknowns: $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$.
- ✓ The solutions to these equations are the **IV estimators**.

Instrumental variables (IV) estimator

- The above method can be easily extended to the case where there are more explanatory variables (but only one endogenous variable).
- Consider the following model:

$$y_1=\beta_0+\beta_1 y_2+\beta_2 z_1+\beta_3 z_2+\beta_4 z_3+\ldots+\beta_k z_{k-1}+u$$

- Suppose that z_k is the instrument for y_2. Then the IV estimators are the solution to the following equations:

Instrumental variables (IV) estimator

$$\sum_{i=1}^{n} (y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1} - \ldots - \hat{\beta}_k z_{ik-1}) = 0$$

$$\sum_{i=1}^{n} z_{i1} (y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1} - \ldots - \hat{\beta}_k z_{ik-1}) = 0$$

$$\vdots$$

$$\sum_{i=1}^{n} z_{ik} (y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1} - \ldots - \hat{\beta}_k z_{ik-1}) = 0$$

Solutions to the above equations are the IV estimators when there are many explanatory variables, but only one endogenous variable and one instrument.

An example, computer printout in Stata

✓Consider the following model:

$$\log(wage)=\beta_0+\beta_1 educ+\beta_2 exper+\beta_3 exper^2+$$
$$+\beta_3 smsa+\beta_4 south+u$$

Using college proximity *nearc4* (a dummy variable for someone who grew up near a four-year college) as an IV for education, estimate the model.

An example, computer printout in Stata

IV regression:

```
. ivregress 2sls lwage exper expersq  smsa south (educ=nearc4)

Instrumental variables (2SLS) regression          Number of obs =    3010
                                                  Wald chi2(5)  =  499.36
                                                  Prob > chi2   =  0.0000
                                                  R-squared     =  0.2051
                                                  Root MSE      =  .39562
```

lwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
educ	.13542	.0486085	2.79	0.005	.0401491	.230691
exper	.1067727	.0218136	4.89	0.000	.0640188	.1495266
expersq	-.0022553	.0003394	-6.64	0.000	-.0029205	-.00159
smsa	.1249987	.0284538	4.39	0.000	.0692302	.1807671
south	-.1409356	.0343705	-4.10	0.000	-.2083005	-.0735707
_cons	3.703427	.8201379	4.52	0.000	2.095986	5.310867

```
Instrumented:   educ
Instruments:    exper expersq smsa south nearc4
```

An example, computer printout in Stata

```
. reg  educ exper expersq  smsa south nearc4, robust

Linear regression                               Number of obs =    3010
                                                F(  5,  3004) =  675.83
                                                Prob > F      =  0.0000
                                                R-squared     =  0.4524
                                                Root MSE      =  1.9825
```

educ	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
exper	-.4258437	.0320651	-13.28	0.000	-.4887155	-.362972
expersq	.0009774	.0017044	0.57	0.566	-.0023646	.0043194
smsa	.3639914	.0863314	4.22	0.000	.1947167	.5332661
south	-.582683	.0743531	-7.84	0.000	-.7284712	-.4368948
nearc4	.3456458	.0824092	4.19	0.000	.1840616	.50723
_cons	16.68131	.1489113	112.02	0.000	16.38933	16.97329

Check if *nearc4* satisfies instrument relevance: using the *t*–test, we can reject the null hypothesis that *nearc4* is not correlated with educ, after controlling for all other exogenous variables.

One endogenous variable, more than one instrument

- Consider the following model with one endogenous variable:

$$y_1=\beta_0+\beta_1 y_2+\beta_2 z_1+u$$

- Now, suppose that you have two instruments for y_2 that satisfy the instrument conditions. Call them z_2 and z_3.

One endogenous variable, more than one instrument

- ✓You could apply the IV method using either z_2 or z_3. But this produces two different estimators. Moreover, they are not efficient.
- ✓There is a more efficient estimator in such a case.
- ✓First, it is important to lay out the instrument conditions.

One endogenous variable, more than one instrument

✓For z_2 and z_3 to be valid instruments, they have to satisfy the following two conditions.

1. **Instrument exogeneity:**

   $\text{Cov}(z_2, u)=0$ and $\text{Cov}(z_3, u)=0$

2. **Instrument relevance:**

   $y_2=\pi_0+\pi_1 z_1+ \pi_2 z_2+ \pi_3 z_3+v$ ← Include all the exogenous variables

   and $\pi_2 \neq 0$ or $\pi_3 \neq 0$

In addition, z_2 and z_3 should not be a part of the structural equation. These are called **the exclusion restrictions**.

One endogenous variable, more than one instrument

- ✓Now, let's explain the estimation method.
- ✓Instead of using only one instrument, we use a linear combination of z_2 and z_3 as the instrument.
- ✓Since a linear combination of z_2 and z_3 also satisfies the instrument conditions, this is a valid method.
- ✓The question is how to find the best linear combination of z_2 and z_3.

Two stage least squares (2SLS) estimator

- ✓ It turns out that the OLS regression of the following model provides the best linear combination:

$$y_2=\pi_0+\pi_1 z_1+\pi_2 z_2+\pi_3 z_3+v$$

- ✓ After you estimate this model, you get the predicted (fitted) value of y_2:

$$\hat{y}_2=\hat{\pi}_0+\hat{\pi}_1 z_1+\hat{\pi}_2 z_2+\hat{\pi}_3 z_3$$

- ✓ Since $\hat{y}_2$ is a combination of variables that are not correlated with u, $\hat{y}_2$ is not correlated with u either. At the same time, $\hat{y}_2$ is correlated with y_2. Thus it is a valid instrument.

Two stage least squares (2SLS) estimator

- We have the following three conditions that can be used to derive an IV estimator:

$$E(u)=0$$

$$Cov(z_1, u)=0$$

$$Cov(\hat{y}_2, u)=0$$

The sample counterparts of the above equations are given by:

Two stage least squares (2SLS) estimator

$$\sum_{i=1}^{n} \hat{u}_i = 0 \qquad \sum_{i=1}^{n} (y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0$$

$$\sum_{i=1}^{n} z_{i1} \hat{u}_i = 0 \qquad \sum_{i=1}^{n} z_{i1} (y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0$$

$$\sum_{i=1}^{n} \hat{y}_{i2} \hat{u}_i = 0 \qquad \sum_{i=1}^{n} \hat{y}_{i2} (y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0$$

- ✓ This is a set of three equations with three unknowns: $\hat{\beta}_0$, $\hat{\beta}_1$,$\hat{\beta}_2$.
- ✓ Solution to these equations represents a special type of the IV estimator, called **the two stage least squares estimator.**

Two stage least squares (2SLS) estimator

- ✓You can estimate these parameters by following the above procedure.
- ✓There is an alternative and equivalent procedure to estimate these parameters. This procedure will give you an idea why it is called the two stage least squares.

Two stage least squares (2SLS) estimator

Stage 1. Estimate the following model using OLS and get the predicted value for y_2: $\hat{y}_2$.

$$y_2=\pi_0+\pi_1 z_1+ \pi_2 z_2+ \pi_3 z_3+v$$

Make sure to insert all exogenous variables

Stage 2. Replace y_2 with $\hat{y}_2$, then estimate the following model using OLS:

$$y_1 = \beta_0 + \beta_1 \hat{y}_2 + \beta_2 z_1 + u$$

OLS estimators of the coefficients are the two stage least squares estimators (2SLS).

An example, computer printout in Stata

✓Consider the following model:

$$\log(wage)=\beta_0+\beta_1 educ+\beta_2 exper+\beta_3 exper^2+u$$

1. Suppose *educ* is endogenous, but *exper* and its square are exogenous. Using mother's and father's education as instruments, estimate the 2SLS model.
2. Manually estimate the model to check if you get the same coefficients (note that you will not get the correct standard errors).

An example, computer printout in Stata

```
. ivregress 2sls lwage exper expersq (educ = motheduc fatheduc), first

First-stage regressions
```

"First" option shows first stage and second stage

```
                                          Number of obs   =        428
                                          F(  4,    423)  =      28.36
                                          Prob > F        =     0.0000
                                          R-squared       =     0.2115
                                          Adj R-squared   =     0.2040
                                          Root MSE        =     2.0390
```

educ	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
exper	.0452254	.0402507	1.12	0.262	-.0338909	.1243417
expersq	-.0010091	.0012033	-0.84	0.402	-.0033744	.0013562
motheduc	.157597	.0358941	4.39	0.000	.087044	.2281501
fatheduc	.1895484	.0337565	5.62	0.000	.1231971	.2558997
_cons	9.10264	.4265614	21.34	0.000	8.264196	9.941084

First stage regression

```
Instrumental variables (2SLS) regression      Number of obs   =        428
                                              Wald chi2(3)    =      24.65
                                              Prob > chi2     =     0.0000
                                              R-squared       =     0.1357
                                              Root MSE        =     .67155
```

lwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
educ	.0613966	.0312895	1.96	0.050	.0000704	.1227228
exper	.0441704	.0133696	3.30	0.001	.0179665	.0703742
expersq	-.000899	.0003998	-2.25	0.025	-.0016826	-.0001154
_cons	.0481003	.398453	0.12	0.904	-.7328532	.8290538

2SLS results

```
Instrumented:  educ
Instruments:   exper expersq motheduc fatheduc
```

An example, computer printout in Stata

- ✓ Estimating 2SLS manually: When you regress the first stage manually on this data, more observations are used than the above 2SLS. To use exactly the same observations, first run the 2SLS and find the observations used in the regression.

```
. ivregress 2sls lwage exper expersq (educ = motheduc fatheduc)

Instrumental variables (2SLS) regression          Number of obs =     428
                                                  Wald chi2(3)  =   24.65
                                                  Prob > chi2   =  0.0000
                                                  R-squared     =  0.1357
                                                  Root MSE      =  .67155
```

lwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
educ	.0613966	.0312895	1.96	0.050	.0000704	.1227228
exper	.0441704	.0133696	3.30	0.001	.0179665	.0703742
expersq	-.000899	.0003998	-2.25	0.025	-.0016826	-.0001154
_cons	.0481003	.398453	0.12	0.904	-.7328532	.8290538

```
Instrumented:  educ
Instruments:   exper expersq motheduc fatheduc

. gen fullsample= e(sample)
```

e(sample) enable you to create a dummy if the observation is used

An example, computer printout in Stata

```
. reg educ exper expersq motheduc fatheduc if fullsample==1
```

Source	SS	df	MS
Model	471.620998	4	117.90525
Residual	1758.57526	423	4.15738833
Total	2230.19626	427	5.22294206

Number of obs	= 428
F(4, 423)	= 28.36
Prob > F	= 0.0000
R-squared	= 0.2115
Adj R-squared	= 0.2040
Root MSE	= 2.039

educ	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
exper	.0452254	.0402507	1.12	0.262	-.0338909	.1243417
expersq	-.0010091	.0012033	-0.84	0.402	-.0033744	.0013562
motheduc	.157597	.0358941	4.39	0.000	.087044	.2281501
fatheduc	.1895484	.0337565	5.62	0.000	.1231971	.2558997
_cons	9.10264	.4265614	21.34	0.000	8.264196	9.941084

Then, estimate the first stage regression. Note "if fullsample==1" tells STATA to use observations only if in the 2SLS.

```
. predict educ_hat, xb
```

After estimation, type this command. This will automatically create the predicted value of *educ*.

An example, computer printout in Stata

Finally, estimate the second stage regression. You can see that the coefficients are the same as before, but standard errors and *t*–statistics are different.

```
. reg lwage educ_hat exper expersq if fullsample==1
```

Source	SS	df	MS
Model	11.117828	3	3.70594266
Residual	212.209613	424	.50049437
Total	223.327441	427	.523015084

Number of obs	428
F(3, 424)	7.40
Prob > F	0.0001
R-squared	0.0498
Adj R-squared	0.0431
Root MSE	.70746

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ_hat	.0613966	.0329624	1.86	0.063	-.0033933	.1261866
exper	.0441704	.0140844	3.14	0.002	.0164865	.0718543
expersq	-.000899	.0004212	-2.13	0.033	-.0017268	-.0000711
_cons	.0481003	.4197565	0.11	0.909	-.7769624	.873163

More than one endogenous variable, more than one instrument

- Consider the following structural equation:

$$y_1=\beta_0+\beta_1 y_2+\beta_2 y_3+\beta_3 z_1+\beta_4 z_2+\beta_5 z_3+u$$

- There are two endogenous variables, y_2 and y_3. Thus, OLS will be biased. In order to estimate this model with the IV method, you need at least two instruments.
- When you have multiple endogenous variables, you need at least the same number of instruments as endogenous variables.

More than one endogenous variable, more than one instrument

- ✓Suppose you have 3 instruments: z_4, z_5, z_6. As usual, these instruments should satisfy the required conditions.
- ✓The first is that they should not be correlated with u (instrument exogeneity).
- ✓The second is that they should be correlated with endogenous variables (instrument relevance). When you have multiple endogenous variables, the second condition has a more complex expression, and it is called **the rank condition**.

Two stage least squares (2SLS) estimator

- *Stage 1*. Estimate the following two reduced form regressions:

$$y_2=\pi_{10}+\pi_{11}z_1+\pi_{12}z_2+\pi_{13}z_3+\pi_{14}z_4+\pi_{15}z_5+\pi_{16}z_6+v$$
$$y_3=\pi_{20}+\pi_{21}z_1+\pi_{22}z_2+\pi_{23}z_3+\pi_{24}z_4+\pi_{25}z_5+\pi_{26}z_6+w$$

Then obtain $\hat{y}_2$ and $\hat{y}_3$.

- *Stage 2*. Estimate the following "second stage regression":

$$y_1=\beta_0+\beta_1\hat{y}_2+\beta_2\hat{y}_3+\beta_3z_1+\beta_4z_2+\beta_5z_3+u$$

The estimated coefficients are the 2SLS estimators.

Two stage least squares (2SLS) estimator

- ✓Note that the second stage regression does not produce correct standard errors. Stata `ivregress` command automatically computes the correct standard errors.
- ✓In the 2SLS method, the *F*–statistic formula we used for OLS is no longer valid. STATA automatically computes a valid *F*–type statistic for 2SLS.

Specific model diagnostics

Statistic tests that need to be carried out *in addition* to the standard model diagnostics:

- The Sargan test, as the most standard **test for overidentifying restrictions**;
- The (Wu-)Hausman test, as the most standard **test for endogeneity**.

Testing overidentifying restrictions

- ✓ Usually, the instrument exogeneity cannot be tested.
- ✓ However, when we have extra instruments, we can effectively test this by using **the test of overidentifying restrictions**.

Testing overidentifying restrictions

- ✓Before presenting the procedure, we will give the basic idea of the test.
- ✓Consider the following model:

$$y_1=\beta_0+\beta_1 y_2+\beta_2 z_1+\beta_3 z_2+u$$

- ✓Suppose you have two instruments for y_2: z_3, z_4. If both instruments are valid instruments, using either z_3 or z_4 as an instrument will produce consistent estimates.

"brève"

- ✓Let $\breve{\beta}_1$ be the IV estimator when z_3 is used as an instrument. Let $\tilde{\beta}_1$ be the IV estimator when z_4 is used as an instrument.

"tilde"

Testing overidentifying restrictions

- The idea is to check if $\breve{\beta}_1$ and $\tilde{\beta}_1$ are similar. That is, you test $H_0: \breve{\beta}_1 - \tilde{\beta}_1 = 0$.
- If you reject this null, it means that either z_3 or z_4, or both of them are not exogenous. We do not know which one is not exogenous. So the rejection of the null typically means that your choice of instruments is invalid.

Testing overidentifying restrictions

- ✓On the other hand, if you fail to reject the null hypothesis, we can have some confidence in the overall set of instruments used.
- ✓However, caution is necessary. Even if you fail to reject the null, this does not always mean that the instruments are valid.
- ✓In general, for the validity of the test, at least one valid instrument is needed to start with.

The Sargan test

(i) Estimate the structural equation by 2SLS and obtain the 2SLS residuals, $\hat{u}$.

(ii) Regress $\hat{u}$ on all exogenous variables. Obtain the multiple determination coefficient.

(iii) Under the null that all IVs are uncorrelated with the structural disturbance term u, the test statistic $nR_1^2 \sim \chi_q^2$ can be used, where q is the number of extra instruments.

✓ This is called the **Sargan test statistic**.

✓ If we fail to reject the null, then we have some confidence about the instrument exogeneity.

✓ If we reject it, at least some of the instruments are not exogenous and thus not valid.

An example, computer printout in Stata

✓Consider the following model:

$$\log(wage)=\beta_0+\beta_1 educ+\beta_2 exper+\beta_3 exper^2+u$$

Estimate the above equation using *motheduc* and *fatheduc* as instruments for *educ*, and test the overidentifying restrictions.

An example, computer printout in Stata

✓First, conduct the test "manually":

```
. ivregress 2sls lwage exper expersq (educ=motheduc fatheduc)

Instrumental variables (2SLS) regression          Number of obs =     428
                                                  Wald chi2(3)  =   24.65
                                                  Prob > chi2   =  0.0000
                                                  R-squared     =  0.1357
                                                  Root MSE      =  .67155
```

1. First, estimate 2sls.

lwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
educ	.0613966	.0312895	1.96	0.050	.0000704	.1227228
exper	.0441704	.0133696	3.30	0.001	.0179665	.0703742
expersq	-.000899	.0003998	-2.25	0.025	-.0016826	-.0001154
_cons	.0481003	.398453	0.12	0.904	-.7328532	.8290538

```
Instrumented:  educ
Instruments:   exper expersq motheduc fatheduc
```

```
. predict uhat, resid
(325 missing values generated)
```

2. Second, generate the 2sls residuals. Call this *uhat*.

An example, computer printout in Stata

```
. reg uhat exper expersq motheduc fatheduc
```

Source	SS	df	MS
Model	.170503136	4	.042625784
Residual	192.84951	423	.455909007
Total	193.020013	427	.452037502

Number of obs	= 428
F(4, 423)	= 0.09
Prob > F	= 0.9845
R-squared	= 0.0009
Adj R-squared	= -0.0086
Root MSE	= .67521

uhat	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
exper	-.0000183	.0133291	-0.00	0.999	-.0262179	.0261813
expersq	7.34e-07	.0003985	0.00	0.999	-.0007825	.000784
motheduc	-.0066065	.0118864	-0.56	0.579	-.0299704	.0167573
fatheduc	.0057823	.0111786	0.52	0.605	-.0161902	.0277547
_cons	.0109641	.1412571	0.08	0.938	-.2666892	.2886173

3. Third, regress *uhat* on all the exogenous variables. Do not forget to include exogenous variables in the structural equation: *exper* and *expersq*.

```
. gen rsq=e(r2)

. su rsq
```

Variable	Obs	Mean	Std. Dev.	Min	Max
rsq	753	.0008833	0	.0008833	.0008833

4. Fourth, get the R-squared from this regression. You can use this, but it is rounded. To compute more precisely, type this.

5. Finally, compute the test statistic:

```
. gen n_rsq=428*rsq

. su n_rsq
```

Variable	Obs	Mean	Std. Dev.	Min	Max
n_rsq	753	.3780714	0	.3780714	.3780714

This is the so-called Sargan test statistic.

An example, computer printout in Stata

- ✓ The nR^2 statistic follows $\chi^2_{(1)}$. The degrees of freedom are equal to the number of extra instruments.

  In our case, there is only one endogenous variable. Thus, you need only one instrument. But we have two instruments. Therefore the number of extra instruments is 1.

- ✓ Since the 5% cutoff point for $\chi^2_{(1)}$ is 3.84, we do not reject the null hypothesis that the instrumental variables are not correlated with the structural disturbance term.

- ✓ Thus, we have some confidence in the choice of instruments. In other words, our instruments have 'passed' the test of overidentifying restrictions.

An example, computer printout in Stata

✓Now, let us conduct the test of overidentifying restrictions automatically:

```
. ivregress 2sls lwage exper expersq (educ=motheduc fatheduc)

Instrumental variables (2SLS) regression          Number of obs =     428
                                                  Wald chi2(3)  =   24.65
                                                  Prob > chi2   =  0.0000
                                                  R-squared     =  0.1357
                                                  Root MSE      =  .67155
```

lwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
educ	.0613966	.0312895	1.96	0.050	.0000704	.1227228
exper	.0441704	.0133696	3.30	0.001	.0179665	.0703742
expersq	-.000899	.0003998	-2.25	0.025	-.0016826	-.0001154
_cons	.0481003	.398453	0.12	0.904	-.7328532	.8290538

```
Instrumented:  educ
Instruments:   exper expersq motheduc fatheduc

. estat overid

  Tests of overidentifying restrictions:

  Sargan (score) chi2(1) =  .378071  (p = 0.5386)
  Basmann chi2(1)        =  .373985  (p = 0.5408)
```

This is the Sargan test statistic with the corresponding *p*–value.

Testing for endogeneity

- Consider again the following model:

$$y_1=\beta_0+\beta_1 y_2+\beta_2 z_1+\beta_3 z_2+u$$

where y_2 is the suspected endogenous variable and you have instruments z_3 and z_4.

- If y_2 is actually exogenous, OLS estimator is more efficient.
- You can test whether y_2 is exogenous, if you have valid instruments. Therefore, test the overidentifying restrictions first, if possible.

Testing for endogeneity

- ✓Before laying out the procedure, let us understand the basic idea behind the test:

  Structural equation: $y_1=\beta_0+\beta_1 y_2+\beta_2 z_1+\beta_3 z_2+u$

  Reduced equation: $y_2=\pi_0+\pi_1 z_1+\pi_2 z_2+\pi_3 z_3+\pi_4 z_4+v$

- ✓You can see that y_2 is correlated with u only if v is correlated with u (as $\hat{y}_2$ is not).
- ✓Further, let $u=\delta v+w$. Then u and v are correlated only if $\delta \neq 0$. Thus, consider:

$$y_1=\beta_0+\beta_1 y_2+\beta_2 z_1+\beta_3 z_2+\delta v+w$$

and then, test that $\delta = 0$.

The (Wu-)Hausman test

(i) Estimate the reduced form equation using OLS:

$$y_2=\pi_0+\pi_1 z_1+\pi_2 z_2+\pi_3 z_3+\pi_4 z_4+v$$

Then obtain the residual $\hat{v}$.

(ii) Add $\hat{v}$ to the structural equation and estimate it using OLS:

$$y_1=\beta_0+\beta_1 y_2+\beta_2 z_1+\beta_3 z_2+\alpha\hat{v}+w$$

(iii) Then, test H_0: $\alpha = 0$. If we reject H_0, then we conclude that y_2 is endogenous, as u and v are correlated.

An example, computer printout in Stata

✓Consider the following model:

$$\log(wage)=\beta_0+\beta_1 educ+\beta_2 exper+\beta_3 exper^2+u$$

Suppose that father's and mother's education satisfy the instrument exogeneity. Conduct the (Wu-)Hausman test of endogeneity to check if *educ* is exogenous.

An example, computer printout in Stata

✓First, conduct the test "manually":

```
. ivregress 2sls lwage exper expersq (educ=motheduc fatheduc)

Instrumental variables (2SLS) regression          Number of obs =     428
                                                  Wald chi2(3)  =   24.65
                                                  Prob > chi2   =  0.0000
                                                  R-squared     =  0.1357
                                                  Root MSE      =  .67155
```

lwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
educ	.0613966	.0312895	1.96	0.050	.0000704	.1227228
exper	.0441704	.0133696	3.30	0.001	.0179665	.0703742
expersq	-.000899	.0003998	-2.25	0.025	-.0016826	-.0001154
_cons	.0481003	.398453	0.12	0.904	-.7328532	.8290538

```
Instrumented:  educ
Instruments:   exper expersq motheduc fatheduc

. gen fullsample=e(sample)
```

To use the same observations as 2sls, run 2sls once and generate this variable

An example, computer printout in Stata

```
. reg educ exper expersq motheduc fatheduc  if fullsample==1
```

Source	SS	df	MS		
Model	471.620998	4	117.90525	Number of obs	= 428
Residual	1758.57526	423	4.15738833	F(4, 423)	= 28.36
				Prob > F	= 0.0000
				R-squared	= 0.2115
				Adj R-squared	= 0.2040
Total	2230.19626	427	5.22294206	Root MSE	= 2.039

educ	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
exper	.0452254	.0402507	1.12	0.262	-.0338909	.1243417
expersq	-.0010091	.0012033	-0.84	0.402	-.0033744	.0013562
motheduc	.157597	.0358941	4.39	0.000	.087044	.2281501
fatheduc	.1895484	.0337565	5.62	0.000	.1231971	.2558997
_cons	9.10264	.4265614	21.34	0.000	8.264196	9.941084

```
. predict uhat_reduced, resid

. reg lwage educ exper expersq uhat_reduced
```

Source	SS	df	MS		
Model	36.2573098	4	9.06432745	Number of obs	= 428
Residual	187.070131	423	.442246173	F(4, 423)	= 20.50
				Prob > F	= 0.0000
				R-squared	= 0.1624
				Adj R-squared	= 0.1544
Total	223.327441	427	.523015084	Root MSE	= .66502

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.0613966	.0309849	1.98	0.048	.000493	.1223003
exper	.0441704	.0132394	3.34	0.001	.0181471	.0701937
expersq	-.000899	.0003959	-2.27	0.024	-.0016772	-.0001208
uhat_reduced	.0581666	.0348073	1.67	0.095	-.0102501	.1265834
_cons	.0481003	.3945753	0.12	0.903	-.7274721	.8236727

Now run the reduced-form regression, then get the residual.

Then check if this coefficient is different from zero.

An example, computer printout in Stata

- ✓The coefficient on *uhat* is *not* significant at the 5% level. Thus, you *cannot reject* the null hypothesis that *educ* is **exogenous** (i.e. not correlated with the structural disturbance term) at the 5% level.
- ✓The coefficient is significant *only* at the 10% level. Thus, you can reject the null hypothesis that *educ* is exogenous *only* at the 10% level.
- ✓There is thus *only* moderate evidence that *educ* is **endogenous** and that, consequently, the 2SLS estimates should be reported.

An example, computer printout in Stata

✓Stata conducts the test of endogeneity automatically (somewhat differently):

```
. ivregress 2sls lwage exper expersq (educ=motheduc fatheduc)

Instrumental variables (2SLS) regression          Number of obs =     428
                                                  Wald chi2(3)  =   24.65
                                                  Prob > chi2   =  0.0000
                                                  R-squared     =  0.1357
                                                  Root MSE      =  .67155
```

lwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
educ	.0613966	.0312895	1.96	0.050	.0000704	.1227228
exper	.0441704	.0133696	3.30	0.001	.0179665	.0703742
expersq	-.000899	.0003998	-2.25	0.025	-.0016826	-.0001154
_cons	.0481003	.398453	0.12	0.904	-.7328532	.8290538

```
Instrumented:  educ
Instruments:   exper expersq motheduc fatheduc

. estat endog

  Tests of endogeneity
  Ho: variables are exogenous

  Durbin (score) chi2(1)          =  2.80707  (p = 0.0938)
  Wu-Hausman F(1,423)             =  2.79259  (p = 0.0954)
```

This is the Wu-Hausman test statistic with the corresponding *p*–value.

University of Ljubljana

SEB SCHOOL OF ECONOMICS AND BUSINESS

7. Instrumental Variables Estimation

Prof. Dr. Miroslav Verbič

miroslav.verbic@ef.uni-lj.si

www.miroslav-verbic.si

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION AMBA ACCREDITED

Ljubljana, October 2025