Elements of Econometrics. Lecture 12. Instrumental Variables.

FCS, 2022-2023

ASSUMPTION B7 VIOLATION:

1. Reasons: Measurement Errors (special case). More general – Endogeneity.

2. Consequences: inconsistent OLS estimators, standard statistics wrongly calculated, tests invalid.

4. Remedial measures: Instrumental Variables. Find another variable independently distributed with the disturbance term, but correlated with the endogenous (in special case – measured with error) explanatory variable

3. Detection: Durbin-Wu-Hausman test (standard, or Davidson-McKinnon version)

INSTRUMENTAL VARIABLES

$$Y = \beta_1 + \beta_2 X + u, \quad X \; related \; with \; u$$

$$Y_i - \bar{Y} = (\beta_1 + \beta_2 X_i + u_i) - (\beta_1 + \beta_2 \bar{X} + \bar{u}) = \beta_2 (X_i - \bar{X}) + u_i - \bar{u}$$

$$\hat{\beta}_2^Z = \frac{\sum (Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum (Z_i - \bar{Z})^2} = \frac{\sum (Z_i - \bar{Z})(\beta_2 [X_i - \bar{X}] + u_i - \bar{u})}{\sum (Z_i - \bar{Z})^2}$$

$$= \beta_2 \frac{\sum (Z_i - \bar{Z})(X_i - \bar{X})}{\sum (Z_i - \bar{Z})^2} + \frac{\sum (Z_i - \bar{Z})(u_i - \bar{u})}{\sum (Z_i - \bar{Z})^2}$$

$$\mathrm{E}\left(\hat{\beta}_2^Z\right) = \beta_2 \, \mathrm{E}\left(\frac{\sum (Z_i - \bar{Z})(X_i - \bar{X})}{\sum (Z_i - \bar{Z})^2}\right)$$

Suppose that X is related with u. Then the OLS estimators are inconsistent. Suppose that *Z* is related to *X* but unrelated to *u*.

Then Z can be used instead of X in the regression (as an instrument).

INSTRUMENTAL VARIABLES

$$\hat{\beta}_2^Z = \frac{\sum(Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum(Z_i - \bar{Z})^2} \qquad E\left(\hat{\beta}_2^Z\right) = \beta_2 E\left(\frac{\sum(Z_i - \bar{Z})(X_i - \bar{X})}{\sum(Z_i - \bar{Z})^2}\right)$$

$$\hat{\beta}_2^{\mathrm{IV}} = \frac{\sum(Z_i - \bar{Z})^2}{\sum(Z_i - \bar{Z})(X_i - \bar{X})}\hat{\beta}_2^Z = \frac{\sum(Z_i - \bar{Z})^2}{\sum(Z_i - \bar{Z})(X_i - \bar{X})}\frac{\sum(Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum(Z_i - \bar{Z})^2}$$

$$= \frac{\sum(Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum(Z_i - \bar{Z})(X_i - \bar{X})} \quad - \ IV\ estimator$$

INSTRUMENTAL VARIABLES: CONSISTENCY

$$Y = \beta_1 + \beta_2 X + u; \qquad Z \text{ is an instrument}$$

$$\hat{\beta}_2^{\mathrm{IV}} = \frac{\sum(Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum(Z_i - \bar{Z})(X_i - \bar{X})} = \beta_2 + \frac{\sum(Z_i - \bar{Z})(u_i - \bar{u})}{\sum(Z_i - \bar{Z})(X_i - \bar{X})}$$

$$\operatorname{plim}\left(\frac{\sum(Z_i - \bar{Z})(u_i - \bar{u})}{\sum(Z_i - \bar{Z})(X_i - \bar{X})}\right) = \operatorname{plim}\left(\frac{\frac{1}{n}\sum(Z_i - \bar{Z})(u_i - \bar{u})}{\frac{1}{n}\sum(Z_i - \bar{Z})(X_i - \bar{X})}\right)$$

$$= \frac{\operatorname{plim}\frac{1}{n}\sum(Z_i - \bar{Z})(u_i - \bar{u})}{\operatorname{plim}\frac{1}{n}\sum(Z_i - \bar{Z})(X_i - \bar{X})} = \frac{\operatorname{cov}(Z, u)}{\operatorname{cov}(Z, X)} = \frac{0}{\sigma_{ZX}} = 0$$

$$\operatorname{var}\left(\hat{\beta}_2^{\mathrm{IV}}\right) = \sigma^2_{\hat{\beta}_2^{\mathrm{IV}}} = \frac{\sigma_u^2}{\sum(X_i - \bar{X})^2} \times \frac{1}{r^2_{X,Z}}$$

INSTRUMENTAL VARIABLES: AN EXAMPLE

$$C_1^P = \beta_1 + \beta_2 Y_1^P \qquad C_2^P = \beta_1 + \beta_2 Y_2^P$$

$$C_1 = C_1^P + C_1^T \qquad C_2 = C_2^P + C_2^T$$

$$Y_1 = Y_1^P + Y_1^T \qquad Y_2 = Y_2^P + Y_2^T$$

$$C_2 - C_2^T = \beta_1 + \beta_2 (Y_2 - Y_2^T)$$

$$C_2 = \beta_1 + \beta_2 Y_2 + (C_2^T - \beta_2 Y_2^T)$$

$$\hat{\beta}_2^{\mathrm{IV}} = \frac{\sum (Y_{1i} - \bar{Y}_1)(C_{2i} - \bar{C}_2)}{\sum (Y_{1i} - \bar{Y}_1)(Y_{2i} - \bar{Y}_2)}$$

An example: consumption function with Friedman's Permanent Income Hypothesis (N.Liviatan).

Y_2 includes Y_2^T, so the explanatory variable is related with the disturbance term. Y_1 is an instrument, and consistent estimate is received.

ASYMPTOTIC DISTRIBUTIONS OF THE IV ESTIMATOR

$$Y = \beta_1 + \beta_2 X + u; \quad Z \text{ is an instrument}$$

$$\hat{\beta}_2^{\mathrm{IV}} = \frac{\sum(Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum(Z_i - \bar{Z})(X_i - \bar{X})} \qquad \operatorname{plim} \hat{\beta}_2^{\mathrm{IV}} = \beta_2$$

$$\sigma^2_{\hat{\beta}_2^{\mathrm{IV}}} = \frac{\sigma_u^2}{\sum(X_i - \bar{X})^2} \times \frac{1}{r_{X,Z}^2} = \frac{\sigma_u^2}{n\left(\frac{1}{n}\sum(X_i - \bar{X})^2\right)} \times \frac{1}{r_{X,Z}^2} = \frac{\sigma_u^2}{n\,\mathrm{MSD}(X)} \times \frac{1}{r_{X,Z}^2}$$

$$\hat{\beta}_2^{\mathrm{IV}} \sim N\left(\beta_2, \frac{\sigma_u^2}{n\mathrm{MSD}(X)} \times \frac{1}{r_{XZ}^2}\right); \qquad \left(\hat{\beta}_2^{\mathrm{IV}} - \beta_2\right) \sim N\left(0, \frac{\sigma_u^2}{n\mathrm{MSD}(X)} \times \frac{1}{r_{XZ}^2}\right)$$

$$\sqrt{n}\left(\hat{\beta}_2^{\mathrm{IV}} - \beta_2\right) \xrightarrow{d} N\left(0, \frac{\sigma_u^2}{\sigma_X^2} \times \frac{1}{r_{XZ}^2}\right)$$

By the central limit theorem $\sqrt{n}\left(\hat{\beta}_2^{\mathrm{IV}} - \beta_2\right)$ has the limiting normal distribution

FINITE-SAMPLE DISTRIBUTIONS OF THE IV ESTIMATOR:

MONTE CARLO EXPERIMENT

$$Y = \beta_1 + \beta_2 X + u \qquad Y = 10 + 5X + u$$

$$X = \lambda_1 Z + \lambda_2 V + u \qquad X = 0.5Z + 2.0V + u$$

Suppose that *Z*, *V*, and *u* are drawn independently from a normal distribution with mean zero and unit variance.

***Let Z* and *V* be variables, *u* - disturbance term in the model. λ_1 and λ_2 are constants.**

***X* is related with *u,* and Assumption B.7 is violated.**

***Z* is correlated with *X*, but independent of *u*, and can be an instrument.**

$$\hat{\beta}_2 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{\sum(X_i - \bar{X})(5[X_i - \bar{X}] + u_i - \bar{u})}{\sum(X_i - \bar{X})^2} =$$

$$= 5 + \frac{\sum(X_i - \bar{X})(u_i - \bar{u})}{\sum(X_i - \bar{X})^2} = 5 + \frac{\widehat{\mathrm{Cov}}\,(X, u)}{\widehat{\mathrm{Var}}\,(X)}.$$

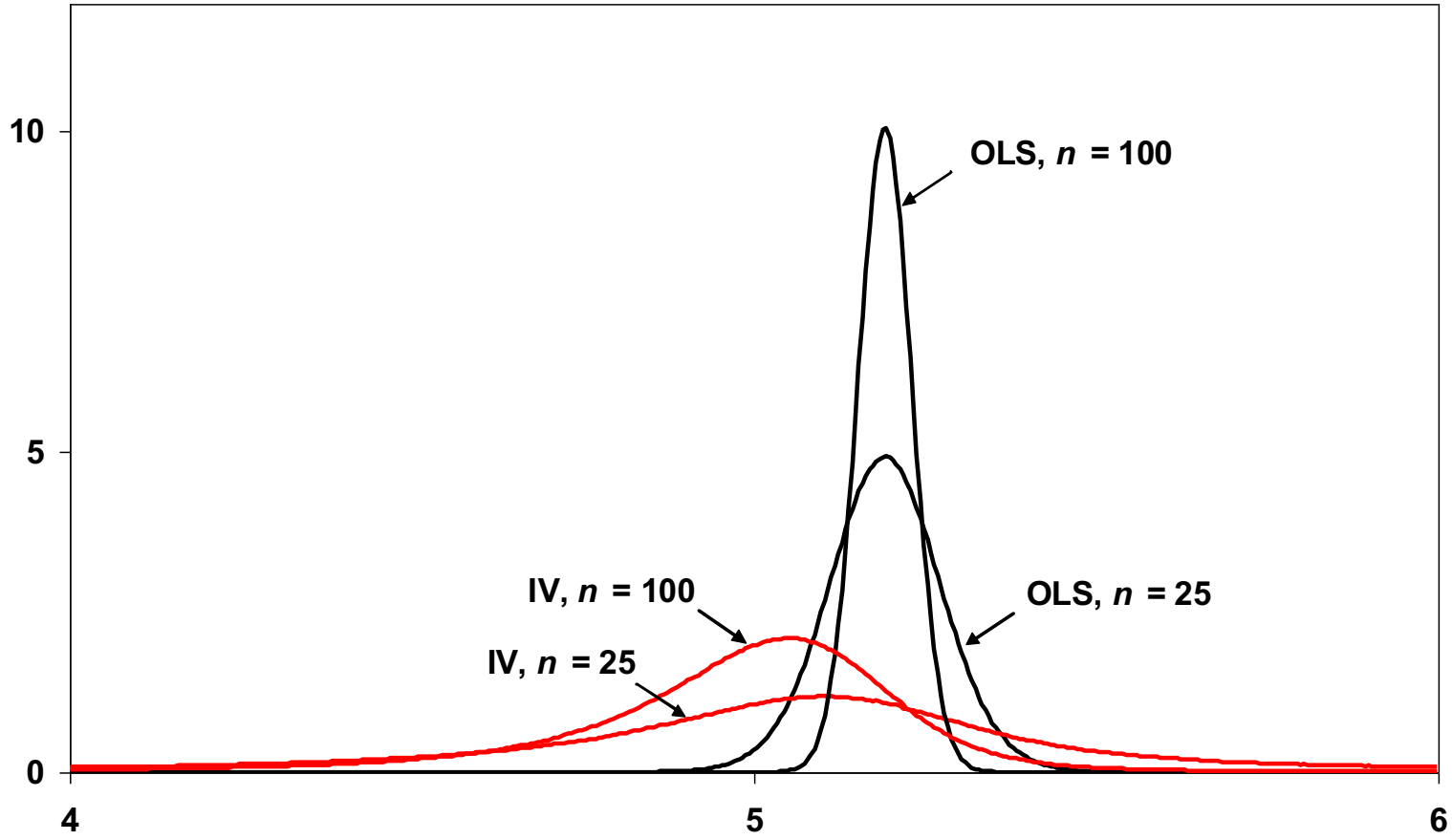
$$\mathrm{plim}(5 + \frac{\widehat{\mathrm{Cov}}\,(X, u)}{\widehat{\mathrm{Var}}\,(X)}) = 5 + \frac{\mathrm{plim}(\widehat{\mathrm{Cov}}\,(X, u))}{\mathrm{plim}(\widehat{\mathrm{Var}}\,(X))} = 5 + \frac{\mathrm{Cov}(X, u)}{\mathrm{Var}(X)} = 5 + \frac{1}{0.25 + 4 + 1} \approx 5.19$$

FINITE-SAMPLE DISTRIBUTIONS OF THE IV ESTIMATOR:

MONTE CARLO EXPERIMENT

$$Y = \beta_1 + \beta_2 X + u \qquad Y = 10 + 5X + u$$

$$X = \lambda_1 Z + \lambda_2 V + u \qquad X = 0.5Z + 2.0V + u$$



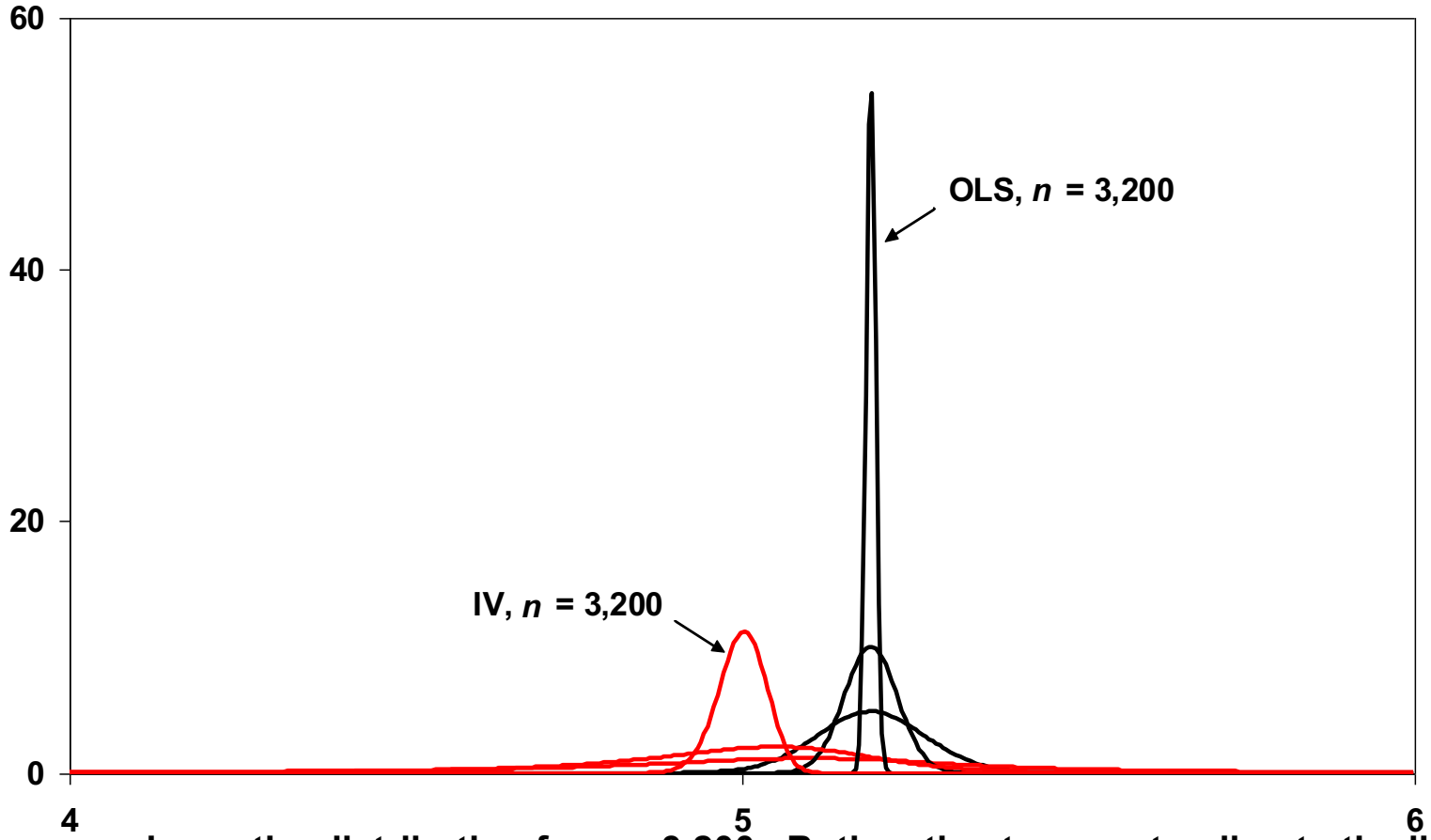
The diagram shows the distributions of the OLS and IV estimators of β_2 for $n = 25$ and $n = 100$, for 10 million samples in both cases. In this case $\text{plim}\hat{\beta}_2^{\text{OLS}} = 5.19$, and $\text{plim}\hat{\beta}_2^{\text{IV}} = 5.00$

FINITE-SAMPLE DISTRIBUTIONS OF THE IV ESTIMATOR:

MONTE CARLO EXPERIMENT

$$Y = \beta_1 + \beta_2 X + u \qquad Y = 10 + 5X + u$$

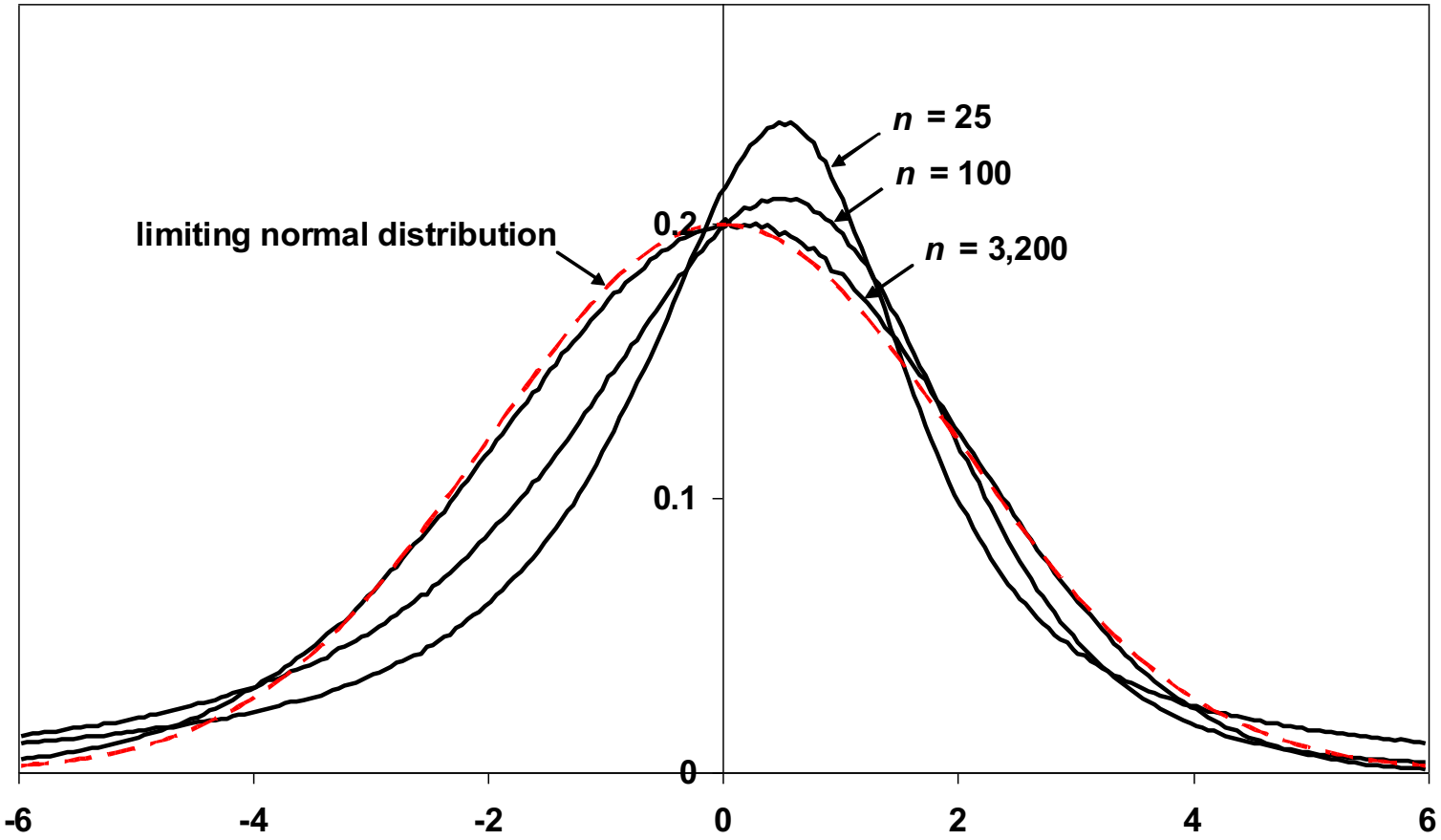
$$X = \lambda_1 Z + \lambda_2 V + u \qquad X = 0.5Z + 2.0V + u$$



This diagram shows the distribution for n = 3,200. Both estimators are tending to the limits 5 and 5.19 and turn into spikes (the IV estimator more slowly than the OLS, because it has a larger variance). The IV estimator is definitely better here.

ASYMPTOTIC AND FINITE-SAMPLE DISTRIBUTIONS OF THE IV ESTIMATOR

$$Y = \beta_1 + \beta_2 X + u \qquad Y = 10 + 5X + u$$
$$X = \lambda_1 Z + \lambda_2 V + u \qquad X = 0.5Z + 2.0V + u$$



This diagram shows the distribution of $\sqrt{n}\left(\hat{\beta}_2^{\mathrm{IV}} - \beta_2\right)$ for *n* = 25, 100, and 3,200. The dashed red line shows the limiting normal distribution predicted by the central limit theorem.

r^2_{XZ}=0.22, and for small samples the IV estimator does not seem good.

DURBIN–WU–HAUSMAN SPECIFICATION TEST.

TESTING FOR RELATIONSHIP OF EXPLANATORY VARIABLES AND DISTURBANCE TERM.

H_0: Assumption B.7 is valid (The disturbance term is distributed independently of the regressors).

Estimator b - consistent under H_0 and H_1

Estimator B - inconsistent under H_1, efficient under H_0

The IV estimator b is consistent under both the H_0 and H_1

The OLS estimator B is consistent (and unbiased), and more efficient than the IV estimator under the H_0, but it is inconsistent under H_1.

DURBIN–WU–HAUSMAN SPECIFICATION TEST:

IF THE COEFFICIENTS DIFFERENCE IS SYSTEMATIC?

Test: H_0: difference in coefficients is not systematic

$$\chi^2(k) = (b-B)'[(V\_b-V\_B)^{-1}](b-B)$$

(Here V_b, V_B – estimated covariance matrices).
Under the null hypothesis, the test statistic has a χ2 (chi-squared) distribution with degrees of freedom equal to the number of regressors tested for endogeneity.

However, if the test statistic is not significant, this does not necessarily mean that the null hypothesis is true. It could be that it is false, but the instruments used in IV are weak.

To perform the Regressor Endogeneity Test in EViews, click on View/IV Diagnostics and Tests/Regressor Endogeneity Test. Enter the list of regressors to test for endogeneity, hit OK and the test results are shown. The DWH statistics is indicated as "Difference in J-statistics".

DURBIN–WU–HAUSMAN SPECIFICATION TEST

Davidson-MacKinnon version of DWH (Hausman) test (EViews)

Estimate the initial model.

Estimate the regression of instrumented variable on the instrument(s), save the residuals.

Add the residuals as the additional regressor in the initial model.

If new variable is insignificant – then the difference in coefficients is not systematic (the OLS estimates are consistent); use the initial model. If it is significant – then the difference is systematic, use IV.

This is an asymptotic test, and the t-statistic should be compared with the critical values from the standard normal.

Both t-test and $\chi 2$ –test give usually close p-values and conclusions.

MEASUREMENT ERRORS AND INSTRUMENTAL VARIABLES: EXAMPLE

$$LGEARN = \beta_1 + \beta_2 S + \beta_3 ASVABC + u$$

Dependent Variable: LGEARN
Method: Least Squares
Included observations: 570

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.777056	0.132301	5.873384	0.0000
S	0.077404	0.010007	7.734779	0.0000
ASVABC	0.012379	0.002662	4.650030	0.0000

R-squared	0.227830	Mean dependent var	2.456463
S.D. dependent var	0.541347	S.E. of regression	0.476537
Sum squared resid	128.7586	F-statistic	83.64712
Durbin-Watson stat	1.728273		

We will illustrate the measurement errors using Earnings Function (EAEF 40). Assume first that in the true model *LGEARN* depends only on *S* and *ASVABC*. Let ASVABC be measured by ASVABM, and ASVABM=ASVABC+10*nrnd.

MEASUREMENT ERRORS AND INSTRUMENTAL VARIABLES: EXAMPLE

$$LGEARN = \beta_1 + \beta_2 S + \beta_3 ASVABM + u$$

Dependent Variable: LGEARN Method: Least Squares Included observations: 570

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.977490	0.122315	7.991580	0.0000
S	0.091519	0.009360	9.778091	0.0000
ASVABCM	0.004546	0.001604	2.834822	0.0047
R-squared	0.209586	Mean dependent var		2.456463

The coefficient of ASVABM seems to be downward biased as prescribed by the theory.

As a potential instrument we may use ASVAB2 which presents one component of ASVABC (arithmetic reasoning) measured without error.

MEASUREMENT ERRORS AND INSTRUMENTAL VARIABLES: EXAMPLE

$$ASVABM = \alpha_1 + \alpha_2 ASVAB2 + u$$

Dependent Variable: ASVABM Method: Least Squares

Included observations: 570

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.485441	2.468044	1.817407	0.0697
ASVAB2	0.910315	0.047805	19.04212	0.0000
R-squared	0.389643	Mean dependent var		50.69112

ASVAB2 is highly correlated with ASVABC, but is not correlated with the disturbance term.

We generate RES1=RESID (the residuals from this regression).

Davidson-MacKinnon version of DWH (Hausman) test (EViews): Example

$$LGEARN = \beta_1 + \beta_2 S + \beta_3 ASVABM + \beta_4 RES1 + u$$

Dependent Variable: LGEARN Method: Least Squares Included observations: 570

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.646442	0.133493	4.842518	0.0000
S	0.070694	0.009875	7.158975	0.0000
ASVABCM	0.016681	0.002698	6.182507	0.0000
RES1	-0.017528	0.003176	-5.518752	0.0000

R-squared	0.249946	Mean dependent var	2.456463

The coefficient of RES1 is significant. This means that instrumenting for ASVABM is needed. In practice, Two Stage Least Squares (TSLS) Method is used.

Two Stage Least Squares (TSLS): Stage 1

$$ASVABM = \beta_1 + \beta_2 S + \beta_3 ASVAB2 + u$$

Dependent Variable: ASVABM Method: Least Squares Included observations: 570

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.254778	2.909055	0.087581	0.9302
S	0.605411	0.223466	2.709184	0.0069
ASVAB2	0.830950	0.055842	14.88039	0.0000
R-squared	0.397443	Mean dependent var		50.69112

TSLS, Stage 1: Regressing instrumented variable on all potential instruments. We get the best instrument, with the highest potential correlation with ASVABM. ASVABMF presents the theoretical values of ASVABM from this regression: $ASVABMF = b_1 + b_2 S + b_3 ASVAB2$

Two Stage Least Squares (TSLS): Stage 2

$$LGEARN = \beta_1 + \beta_2 S + \beta_3 ASVABMF + u$$

Dependent Variable: LGEARN Method: Least Squares Included observations: 570

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.720170	0.127958	5.628170	0.0000
S	0.059069	0.010848	5.445056	0.0000
ASVABCMF	0.018356	0.002949	6.225240	0.0000

R-squared	0.249667	Mean dependent var	2.456463

TSLS, Eviews: tsls lgearn c s asvabcm @ c s asvab2

TSLS, Stage 2: Regressing dependent variable on the regressors, including ASVABMF.

Now we get consistent estimates of $\beta_1, \beta_2, \beta_3$

THE GENERAL INSTRUMENTAL VARIABLES REGRESSION MODEL AND TERMINOLOGY

The general IV regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i; \quad i = 1, \ldots, n \ (*)$$

where:

– Y_i is the dependent variable;
– u_i is the error term, which represents measurement error and/or omitted factors;
– $X_{1i}, \ldots, X_{ki}$ are k endogenous regressors, which are potentially correlated with u_i ;
– $W_{1i}, \ldots, W_{ri}$ are *r* included exogenous regressors, which are uncorrelated with u_i ;
– $\beta_0, \beta_1, \ldots, \beta_{k+r}$ are unknown regression coefficients;
– $Z_{1i}, \ldots, Z_{mi}$ are *m* instrumental variables.

The coefficients are overidentified if there are more instruments than endogenous regressors (*m* > *k*); they are underidentified if *m* < *k*; and they are exactly identified if *m* = *k*. Estimation of the IV regression model requires exact identification or overidentification.

TWO STAGE LEAST SQUARES

The TSLS estimator in the general IV regression model in Equation (*) with multiple instrumental variables is computed in two stages:

1. First-stage regression(s): Regress X_{1i} on the instrumental variables $Z_{1i}, \ldots, Z_{mi}$ and the included exogenous variables $W_{1i}, \ldots, W_{ri}$ using OLS. Compute the predicted values from this regression; call these $\hat{X}_{1i}$ Repeat this for all the endogenous regressors $X_{2i}, \ldots, X_{ki}$ thereby computing the predicted values $\hat{X}_{2i}, \ldots, \hat{X}_{ki}$.

2. Second-stage regression: Regress Y_i on the predicted values of the endogenous variables $\hat{X}_{2i}, \ldots, \hat{X}_{ki}$ and the included exogenous variables $W_{1i}, \ldots, W_{ri}$ using OLS. The TSLS estimators $\beta_0^{TSLS}, \beta_1^{TSLS}, \ldots, \beta_{k+r}^{TSLS}$ are the estimators from the second-stage regression.

In practice, the two stages are done automatically within TSLS estimation commands in modern econometric software.

EViews: TSLS Y C X1 … Xk W1 … Wr @ c Z1 … Zm W1 … Wr