Instrumental Variables Example

Ani Katchova

Instrumental Variables Example

- We want to study the factors influencing medical expenses (y_1) given the endogenous regressor of having health insurance (y_2) and exogenous regressors of illnesses, age, and income (x_1). Instruments are the SS income ratio and firm multiple locations (x_2).
- Data are from the Medical Expenditure Panel Survey (MEPS).

Single equation:

OLS regression: $y_1 = y_2^{'}\beta_1 + x_1^{'}\beta_2 + u$
2SLS, first-stage equation: $y_2 = x_1^{'}\gamma_1 + x_2^{'}\gamma_2 + e$
2SLS, second-stage equation: $y_1 = \widehat{y_2^{'}}\beta_1 + x_1^{'}\beta_2 + u$

Systems of equations:

$y_1 = y_2^{'}\beta_1 + z_1^{'}\gamma_1 + u_1$
$y_2 = y_1^{'}\beta_2 + z_2^{'}\gamma_2 + u_2$

2SLS estimation, just-identified case (1 endogenous variable, 1 instrument)

	OLS regression for y1 (log of med expenses)	2SLS: first stage for y2 (health insurance)	2SLS: second stage for y1 (log med expenses)
Have health insurance (endogenous variable y2)	0.075*	-	-0.852*
Illnesses (x1)	0.441*	0.011*	0.449*
Age (x1)	-0.003	-0.009*	-0.012*
Log income (x1)	0.017*	0.054*	0.098*
SS income ratio (instrument x2)	-	-0.200*	-
Constant	5.780*	0.959*	6.590*

- Interpretation of coefficient on the endogenous variable in OLS model: For individuals with health insurance, the medical expenses are 7.5% higher than those for individuals without health insurance.
- Interpretation of the coefficient of the endogenous variable in 2SLS. After instrumentation, for individuals with health insurance, their medical expenses are 85.2% lower than those for individuals without health insurance.
- Note that the 2SLS coefficient turned out quite different from the OLS coefficient.

2SLS estimation, over-identified case (1 endogenous variable, 2 instruments)

	OLS regression for y1 (log of med expenses)	2SLS: first stage for y2 (health insurance)	2SLS: second stage for y1 (log med expenses)
Have health insurance (endogenous variable y2)	0.075*	-	-0.970*
Illnesses (x1)	0.441*	0.012*	0.450*
Age (x1)	-0.003	-0.008*	-0.013*
Log income (x1)	0.017*	0.051*	0.108*
SS income ratio (instrument x2)	-	-0.191*	-
Firm location (instrument x2)	-	0.116*	-
Constant	5.780*	0.912*	6.692*

- With two instruments instead of one, the estimates changed only slightly from -0.852 to -0.970 for the coefficient on have health insurance.

Tests

- The Durbin-Wu-Hausman test compares OLS and the 2SLS model coefficients. The null hypothesis is that the regressors are exogenous is rejected. Therefore, the health insurance is an endogenous regressor and we need to use instrumental variables approach.
- The test for overidentifying restriction shows all instruments are valid.
- There is low correlation among instruments and endogenous variable, of about 0.1-0.25 in absolute value. The correlation is low, but not an indication of weak instruments.
- The test for weak instruments looks at the F statistic for joint significance of instruments. The number is 69 from the model with 1 instrument and 59 from the model with 2 instruments, which is larger than the rule of thumb of 10. Therefore, the instruments are not weak.

Systems of equations, 2SLS and 3SLS

	2SLS estimation for y1: (log of med expenses)	2SLS estimation for y2: (health insurance)	3SLS estimation for y1: (log of med expenses)	3SLS estimation for y2: (health insurance)
Log of med expenses (y1)	-	0.235*	-	0.235*
Have health insurance (y2)	-1.673*	-	-1.599*	-
Illnesses (x1, x12)	0.458*	-0.100*	0.456*	-0.100*
Age (x1)	-0.019*	-	-0.018*	-
Log income (x1)	0.142*	-	0.136*	-
SS income ratio (instrument x2)	-0.164	-	-0.134	-
Firm location (instrument x22)	-	-0.283*	-	-0.283*
Constant	7.377*	-0.972*	7.237*	-0.972*

- Interpretation of coefficients: for individuals with health insurance, medical expenses are 167.3% or 159.9% lower from the 2SLS or 3SLS models.
- The 3SLS results are different (even though only slightly) from the 2SLS because of different regressors and instruments.
- We will get identical results if the system is just identified, 1 instrumental variable for each endogenous variable, and the same exogenous variables in both equations.