Chapter 23

INSTRUMENTAL VARIABLE ESTIMATION

Anuja A. R., K. N. Singh, Shivaswamy G. P., Rajesh T. and Harish Kumar H. V.

INTRODUCTION

In general, research issues in the social sciences are causal. Impact assessment studies focus on the influence of treatment on outcome. For example, while assessing the impact of a welfare initiative on poverty reduction, the welfare program is the treatment and poverty reduction is the intended outcome. Here, allotting treatment randomly to the experimental units is not feasible. Estimation of a causal relationship under such circumstances is problematic as it is difficult to establish that the treatments are exogenous to the investigated system.

One of the basic assumptions of Ordinary Least Square (OLS) is that there is no correlation between independent variables and residuals. When the predictor variable X is correlated with the error term U, the estimation of the causal effect using observational data will be biased. The problem can be addressed by adding additional exogenous variables to the model. In social science, Instrumental Variable (IV) technique is helpful to estimate the causal effect when there exists endogeneity. The Wu-Hausman test can be used to check endogeneity of treatment variable. IV can be used to solve the problem of omitted variable bias and the classic errors-in-variables problem.

Endogeneity occurs when there exists a correlation between independent variables and the error term. Let us take an example to explain the situation. Suppose we want to assess the impact of years of schooling on the earning of individuals. We observe correlation between years of schooling and the outcome variable i.e. earnings of individuals. But this correlation not necessarily indicates a causal relationship. Suppose, there is some unobservable variable that influences the outcome here such as IQ of the individual. There is a possibility that a better IQ of the individual is positively influencing both the treatment (years of schooling) and outcome variables (earnings of the individual). Fig 1 depicts the situations where causal inference in observational studies will be valid. The instrumental variable technique is an important tool used in the impact assessment studies in agriculture.

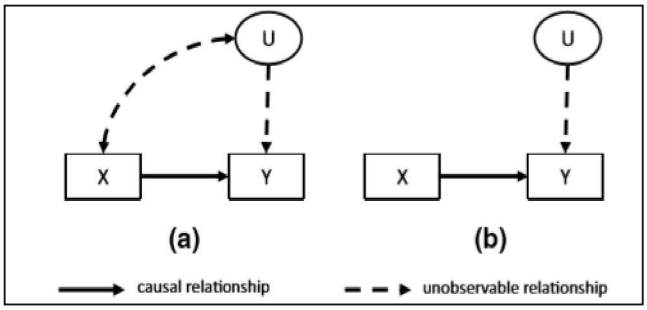

Fig 1: Examples of a situation where the modeling of causal relationships using observational data will be biased (a) and a situation where it will be valid (b)
(Adopted from Pokrope, 2016)

WHAT ARE THE INSTRUMENTAL VARIABLES?

Instrumental variable (IV) methods allow for endogeneity. An instrumental variable Z is an exogenous variable employed to assess the causal effect of variable X on Y (Fig 2).

A variable Z is an instrumental (relative to the pair (X, Y)) if

(i) Z is independent of all variables (including error terms) that have an influence on Y that is not mediated by X and

(ii) Z is not independent of X (Pearl 2000).

The first clause is referred to as the 'exclusion' and the second as the 'relevance'.

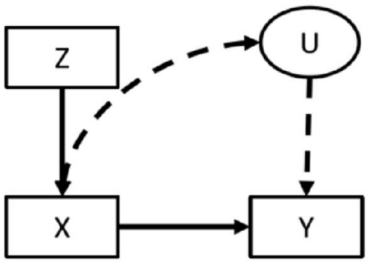

Fig 2: Situation where Z is a valid instrument (Pokrope, 2016)

Illustrating the application of instrumental variable technique in the agriculture

Birthal *et al.* (2015) employed IV technique to assess the impact of crop diversification on farm poverty in India. Unobserved features such a skill, motivation, etc. may lead to bias in the estimated coefficient. Using OLS regression to assess the impact may capture this unobserved heterogeneity and hence the estimates can suffer from bias. An instrumental variable was introduced into the model to mask unobserved heterogeneity at household level. As explained earlier, an ideal IV will not influence the outcome but will influence the treatment variable. In the study, the neighborhood effect based on geographical and social proximity was the IV. The logic of choosing the IV was that if the number of farmers growing high-value crops in the neighborhood is high it would positively influence the treatment variable i.e. area share of high-value crops. At the same time, the said IV would not affect the outcome variable of the model (farm poverty).

Selection of the instrumental variable

The selection of IV is of at most importance for the proper estimation of the causal effect. Finding a suitable instrumental variable for a large-scale database is a difficult task. Knowledge, experience and thorough understanding of the research issue can guide the researcher in finding proper IV for a situation. Weak instruments may worsen the bias in estimation (Khandker *et al.,* 2010). A value greater than 10 for the first stage F statistic indicates a strong instrument. This does not necessarily rule out a weak instrument issue.

Disadvantages of instrumental variable

There are many challenges associated with the application of IV variables in impact assessment. The very difficulty in finding a suitable IV following all the assumptions

is a major challenge. The poor performance of IV in small samples is another issue (Baum, 2008). The strength of the IV determines the precision. In comparison with the OLS estimates, IV estimates suffer from severe precision loss, if the instrument is weak. IV approaches are not immune from selection bias and the issue can be addressed by using the inverse probability of selection weights (Canan *et al.,* 2017)

IV technique using Two-Stage Least Squares (2SLS) regression

In the OLS regression, there is a basic assumption that all independent variables are uncorrelated with the error term. Two-Stage least squares (2SLS) regression analysis is employed when there exists problem of endogeneity (Gujarati *et al.,* 2012)

*Problematic causal variable***:** This is the independent variable that is correlated with the error term or it is the variable that is influenced by other variables in the model. This endogenous causal variable is replaced with an instrumental variable in the first stage of the analysis.

Instrument variable: An instrumental variable is a new variable used in 2SLS to account for unobserved behavior between variables.

Estimation stages

First stage: A new variable is created using the instrument variable

Second stage: Instead of actual values of the problematic predictors, estimated values from the earlier stage is used in an OLS model to estimate the impact of the treatment variable

First stage regression:-

$$x_i = I\alpha + Zv + \delta_i \quad \ldots \tag{1}$$

x_i – Vector of the endogenous variable i (where i = 1,…, N)

I- Matrix for Instrumental variables

Z- Matrix of the covariates

δ_i- Error term

The role of the instrumental variables finishes at the first stage of 2SLS. Covariates are included in the first stage of the estimation to ensure that there is no direct influence of IV on the outcome. More than one IV can be employed in the first stage considering the appropriateness of the variables.

Second stage regression: -

$$y = \hat{x}_i\beta_i + Z\beta + e \quad \ldots \tag{2}$$

y- Vector of the outcome variable

$\hat{x}_i$ - Vector of predicted values of x based on first stage regression

β_i - Parameter estimate of the causal effect of X on Y

Z - Matrix of the covariates

β - Vector of slope parameters for the covariates from Z

e - Error term.

Interpretation

The IV estimates indicate the local average treatment effect (LATE) instead of the average treatment effect (ATE). The ATE is the expected average effect of the treatment on outcome. The LATE provides information about the units that are likely to get the treatment if it is in the treatment group, but otherwise not take the treatment. The estimated LATE can be generalized for the population if there is no striking difference between the individuals influenced by the instrument and the population (Pokrope, 2016).

ILLUSTRATATION

Suppose we want to study the impact of having health insurance on medical expenses. In the given example, the dependent variable is 'medical expenses' (y_1), the endogenous regressor is 'having health insurance' (y_2) and exogenous regressors are illness, age, and income (x_1) of the individuals. In this example, social security income (ssi) ratio of the individual is used as an instrument (x_2). The IV represents variables assumed to affect 'the choice of having health insurance or not' but to have no direct effect on the outcome i.e. medical expenses. Table 1 indicates the sample data.

Table 1: Sample data

Number	Medical expenses	Health insurance	Age	Female	Income	Illnesses	ssi ratio
1	595	1	74	1	95	0	0.15
2	1783	1	73	0	36	3	0.40
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
n-1	720	0	69	1	29	1	0.15
n	809	1	90	1	21	1	0.36

Note: The data used in the illustrative example is a modified data from Katchova, A. (2013). Instrumental Variables in STATA. https://sites.google.com/site/econometricsacademy/econometrics-models/instrumental-variables.

OLS regression in STATA

First, define the dependent variable, independent variables, endogenous variable and instrumental variable. Command used for OLS regression in STATA – '***regress***'. Here the dependent variable is medical expenses (y_1). The endogenous regressor is 'having

health insurance' (y_2) and exogenous regressors are illness, age, and income (x_1) of the individuals. Table 2 illustrates the results of OLS regression. The results indicate that for individuals with health insurance, the medical expenses are 7.5% higher than those for individuals without health insurance.

STATA Command: regress y_1 y_2 x_{1list}

Table 2: Result of OLS regression

y_1: log of medical expenses	**Coef.**	**SE**	**t**	**P>t**	**[95% Conf. Interval]**	
Health insurance (y_2)	**0.075***	0.026	2.880	0.004	0.024	0.126
Illnesses (x_1)	0.441*	0.010	46.040	0.000	0.422	0.459
Age (x_1)	-0.003	0.002	-1.380	0.167	-0.006	0.001
Log of income (x_1)	0.017	0.014	1.250	0.211	-0.010	0.044
Constant	5.780*	0.151	38.310	0.000	5.484	6.076

* $p < 0.01$

2SLS estimation: - Command used for 2SLS regression using IV in STATA is as follows. '

Command: ivregress 2sls y_1 (y_2= x_2) x_{1list}

Table 3: Result of 2 SLS estimation

y_1: log of medical expenses	**Coef.**	**SE**	**t**	**P>t**	**[95% Conf. Interval]**	
Health insurance (y_2)	**-0.852***	0.198	-4.300	0.000	-1.241	-0.463
Illnesses (x_1)	0.449*	0.010	43.590	0.000	0.428	0.469
Age (x_1)	-0.012*	0.003	-4.230	0.000	-0.017	-0.006
Log of income (x_1)	0.098*	0.022	4.350	0.000	0.054	0.142
SS incomer ratio (instrument x_2)	-					
Constant	6.590*	0.235	28.090	0.000	6.130	7.050

* $p < 0.01$

X_{1list} – Indicates list of exogenous variables

Table 3 explains the results of 2SLS with IV model. After instrumentation, for individuals with health insurance, their medical expenses are 85.2% lower than those for individuals without health insurance. It is evident from the results that the 2SLS coefficient turned out quite different from the OLS coefficient.

The following tests can be employed to ascertain the strength and suitability of the instruments.

Durbin-Wu-Hausman test for endogeneity

The endogeneity in the model can be tested using the Durbin-Wu-Hausman test for endogeneity. The Null hypothesis of the Durbin-Wu-Hausman test is that the independent variables are exogenous in nature. Rejection of null-hypothesis indicates the presence of endogeneity. The presence of endogeneity necessitates the usage of IV approach.

In the given example *test for endogeneity* was performed using the following command in STATA.

quietly ivregress 2sls y_1 (y_2= x_2) x_{1list}, first
estat endogenous
quietly regress y_2 x_2 x_{1list}
quietly predict vhat, resid
quietly regress y_1 y_2 x_{1list} vhat
testvhat

```
F(  1, 10083) =   25.14
     Prob > F =    0.0000


Tests of endogeneity
Ho: variables are exogenous

Durbin (score) chi2(1)          =  25.0914   (p = 0.0000)
Wu-Hausman F(1,10083)           =   25.139   (p = 0.0000)
```

The rejection of null hypothesis confirmed the presence of endogeneity.

Correlation

The correlation between 'having health insurance' (endogenous variable) and ssi (IV) was tested and there was a negative correlation of -0.21. Here the correlation is weak and this may lead to biased estimates.

```
First-stage regression summary statistics
-------------------------------------------------------------------------
             |           Adjusted      Partial       Robust
    Variable |   R-sq.      R-sq.        R-sq.    F(1,10084)   Prob > F
-------------+-----------------------------------------------------------
  healthinsu |  0.0684     0.0680       0.0194        68.881     0.0000
-------------------------------------------------------------------------
```

Weak instrument test -F statistics

As a thumb rule, if the value of F statistics of the model is greater than 10, instruments are not weak. Following commands were used to estimate the F statistics.

quietly ivregress 2sls y_1 (y_2= x_2) x_{1list}, vce (robust)

estat first stage, forcenonrobust

As the value is 69 (which is greater than 10 as per thumb rule), the given instrument is not weak.

Validity of multiple instruments.

The test for over-identifying restriction can be used to check the validity of multiple instruments. In the given example we have employed a single instrument.

REFERENCES

Baum, C. F. (2008), Using Instrumental Variables Techniques in Economics and Finance. Boston College and DIW Berlin. German Stata Users Group Meeting, Berlin, June 2008. Online available at https://www.stata.com/meeting/germany08/Baum.DESUG8621.beamer.pdf

Birthal, P. S., D. Roy and D. S. Negi (2015), Assessing the Impact of Crop Diversification on Farm Poverty in India. World Development, Elsevier, 72(C), 70-92.

Canan, C., C. Lesko and B. Lau (2017), Instrumental Variable Analyses and Selection Bias. *Epidemiology (Cambridge, Mass.)*, *28*(3); 396–398. doi:10.1097/EDE.0000000000000639

Gujarati, D. N., D. C. Porter and S. Gunasekar (2012), Basic Econometrics. McGraw Hill Education (India) Private Limited.

Katchova, A. (2013), Instrumental Variables in STATA. Online available at https://sites.google.com/site/econometricsacademy/econometrics-models/instrumental-variables.

Khandker, S. R., G. B. Koolwal and H. A. Samad (2010), World Bank: Handbook on Impact Evaluation: Quantitative Methods and Practices. World Bank.

Pearl, J. (2000), Causality: Models, Reasoning and Inference. New York: Cambridge University Press.

Pokrope, A. (2016), Introduction to Instrumental Variables and their Application to Large-Scale Assessment Data. Large-scale Assessments in Education 4:4 DOI 10.1186/s40536-016-0018-2.