Instrumental Variables Regression

Instrumental Variables (IV) estimation is used when the model has endogenous X's.

IV can thus be used to address the following important threats to internal validity:

- Omitted variable bias from a variable that is correlated with X but is unobserved, so cannot be included in the regression;
- Simultaneous causality bias (endogenous explanatory variables; X causes Y, Y causes X);
- Errors-in-variables bias (X is measured with error)

Instrumental variables regression can eliminate bias from these three sources.

Terminology: Endogeneity and Exogeneity

An ***endogenous*** variable is one that is correlated with ε
An ***exogenous*** variable is one that is uncorrelated with ε
Historical note: "Endogenous" literally means "determined within the system," that is, a variable that is jointly determined with Y, that is, a variable subject to simultaneous causality. However, this definition is narrow and IV regression can be used to address omitted variable bias and errors-in-variable bias, not just simultaneous causality bias.

Instrumental Variables: Intuition

- An *instrumental variable*, Z is uncorrelated with the disturbance ε but is correlated with X (e.g., proximity to college might be correlated with schooling but not with wage residuals)
- With this new variable, the IV estimator should capture only the effects on Y of shifts in X induced by Z whereas the OLS estimator captures not only the direct effect of X on Y but also the effect of the included measurement error and/or endogeneity
- IV is not as efficient as OLS (especially if Z only weakly correlated with X, i.e. when we have so-called "weak instruments") and only has large sample properties (consistency)
    - IV results in biased coefficients. The bias can be large in the case of weak instruments (see below)

What Is an Instrumental Variable?

In order for a variable, z, to serve as a valid instrument for x, the following must be true

- The instrument must be exogenous
  - That is, $Cov(z, \varepsilon) = 0$
- The instrument must be correlated with the endogenous explanatory variable x
  - That is, $Cov(z, x) \neq 0$

The Validity of Instruments

- We have to use common sense and economic theory to decide if it makes sense to assume $Cov(z, \varepsilon) = 0$
  - Can't directly test this condition because don't have unbiased estimator for ε
  - OLS estimator of ε is presumed biased and the IV estimator of ε depends on the validity of $Cov(z, \varepsilon) = 0$ condition
- We can test if $Cov(z, x) \neq 0$
- By testing whether $p_1 = 0$ in the regression: $x = p_0 + p_1 z + v$
- This is sometimes referred to as the first-stage regression

The IV Estimator with a Single Regressor and a Single Instrument

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Loosely, IV regression breaks X into two parts: a part that might be correlated with ε, and a part that is not.

By isolating the part that is not correlated with ε, it is possible to estimate β_1.

- This is done using an ***instrumental variable***, Z_i, which is uncorrelated with ε_i.
- The instrumental variable detects movements in X_i that are uncorrelated with ε_i, and use these to estimate β_1.

Two conditions for a valid instrument

For an instrumental variable (an "***instrument***") Z to be valid, it must satisfy two conditions:

1. ***Instrument relevance***: $Cov(z, x) \neq 0$
2. ***Instrument exogeneity***: $Cov(z, \varepsilon) = 0$

Suppose for now that you have such a Z_i. How can you use Z_i to estimate β_1?

The IV Estimator, one *X* and one *Z*

Explanation #1: Two Stage Least Squares (TSLS)

As it sounds, TSLS has two stages – two regressions:

(1) First isolates the part of X that is uncorrelated with ε:

regress X on Z using OLS

$X_i = \pi_0 + \pi_1 Z_i + v_i$ (1)

- Because Z_i is uncorrelated with ε_i, $\pi_0 + \pi_1 Z_i$ is uncorrelated with ε_i. We don't know π_0 or π_1 but we have estimated them, so…
- Compute the predicted values of X_i, $\widehat{X_\iota}$, where $\widehat{X_\iota} = \widehat{\pi_0} + \widehat{\pi_1} Z_i$, $i = 1, \ldots, n$

(2) Replace X_i by $\widehat{X_\iota}$ in the regression of interest:

regress Y on $\widehat{X_\iota}$ using OLS:

$Y_i = \beta_0 + \beta_1 \widehat{X_\iota} + \varepsilon_i$ (2)

- Because $\widehat{X_\iota}$ is uncorrelated with ε_i in large samples, so the assumption A1 holds
- Thus β_1 can be estimated by OLS using regression (2)
- This argument relies on large samples (so that π_0 and π_1 are well estimated using regression (1))
- The resulting estimator is called the "Two Stage Least Squares" (TSLS or 2SLS) estimator, $\hat{\beta}_1^{TSLS}$.
- $\hat{\beta}_1^{TSLS}$ is a consistent estimator of β_1.

Inference using Two Stage Least Squares

Statistical inference proceeds in the usual way.

Note on standard errors:

- The OLS standard errors from the second stage regression aren't right – they don't take into account the estimation in the first stage ($\widehat{X_\iota}$ is estimated).
- Instead, use a single specialized command that computes the TSLS estimator and the correct *SE*s.
- As usual, use heteroskedasticity-robust *SE*s

Two Stage Least Squares – Finding the Best Instrument

- If we have more than one potential instrument, say $Z_2$and Z_3, then we could use either Z_2 or Z_3 as an instrument (the model is said to be over-identified)
- The best instrument however is a linear combination of all of the exogenous variables,

$$X_i = \pi_0 + \pi_1 Z_{2i} + \pi_2 Z_{3i} + v_i$$

- We can obtain the predicted values, $\widehat{X_\iota}$ by regressing X_i on Z_{2i} and Z_{3i}, which would be the first stage regression
- We could then once again replace X_i by $\widehat{X_\iota}$ in the original regression equation.
- Once again the standard errors from doing TSLS by hand are incorrect (statistical packages usually correct for this however).

Summary of IV Regression with a Single X and Z

A valid instrument *Z* must satisfy two conditions:
(1) *relevance*: $Cov(z, x) \neq 0$
(2) *exogeneity*: $Cov(z, \varepsilon) = 0$

- TSLS proceeds by first regressing X on Z to get $\hat{X}$, then regressing Y on $\hat{X}$.
- The key idea is that the first stage isolates the part of the variation in X that is uncorrelated with ε
- If the instrument is valid, then the large-sample sampling distribution of the TSLS estimator is normal, so inference proceeds as usual

The General IV Regression Model

The general IV regression model: notation and jargon

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \ldots + \beta_{k+r} W_{ri} + \varepsilon_i$$

- Y_i is the dependent variable
- $X_{1i}, \ldots, X_{ki}$ are the endogenous regressors (potentially correlated with ε_i)
- $W_{1i}, \ldots, W_{ri}$ are the ***included exogenous variables*** or ***included exogenous regressors*** (uncorrelated with ε_i)
- $\beta_0, \beta_1, \ldots, \beta_{k+r}$ are the unknown regression coefficients
- $Z_{1i}, \ldots, Z_{mi}$ are the m instrumental variables (the ***excluded exogenous variables***)

We need to introduce some new concepts and to extend some old concepts to the general IV regression model:

- Terminology: *identification* and *overidentification*
- TSLS with included exogenous variables
  - One endogenous regressor
  - Multiple endogenous regressors
- Assumptions that underlie the normal sampling distribution of TSLS
  - Instrument validity (relevance and exogeneity)
  - General IV regression assumptions

Identification

• In general, a parameter is said to be ***identified*** if different values of the parameter would produce different distributions of the data.

• In IV regression, whether the coefficients are identified depends on the relation between the number of instruments (*m*) and the number of endogenous regressors (*k*)

• Intuitively, if there are fewer instruments than endogenous regressors, we can't estimate $\beta_1, \ldots, \beta_{k+r}$

• For example, suppose *k* = 1 but *m* = 0 (we have no instruments)!

The coefficients $\beta_1, \ldots, \beta_k$ are said to be:

- ***Exactly Identified*** if $m = k$.

There are just enough instruments to estimate $\beta_1, \ldots, \beta_k$.

- ***Overidentified*** if $m > k$.

There are more than enough instruments to estimate $\beta_1, \ldots, \beta_k$.

If so, you can test whether the instruments are valid (*a test of the "overidentifying restrictions*")

- ***Underidentified*** if $m < k$.

There are too few enough instruments to estimate $\beta_1, \ldots, \beta_k$.

If so, you need to get more instruments!

General IV regression: TSLS with one Endogenous Regressor
The regression model takes the form,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 W_{1i} + \dots + \beta_{1+r} W_{ri} + \varepsilon_i$$

- Instruments: $Z_{1i}, \dots, Z_{mi}$
- First stage
  - Regress X_1 on *all* the exogenous regressors: regress X_1 on $W_1, \dots, W_r$ and $Z_{1i}, \dots, Z_{mi}$ using OLS
  - Compute predicted values $\widehat{X_\iota}$, $i = 1, \dots, n$
- Second stage
  - Regress *Y* on $\widehat{X_\iota}$, $W_1, \dots, W_r$ using OLS
  - The coefficients from this second stage regression are the TSLS estimators, but the standard errors are again wrong

General IV regression: TSLS with Multiple Endogenous Regressors

$$Y_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \dots + \beta_{k+r} W_{ri} + \varepsilon_i$$

- Instruments: $Z_{1i}, \dots, Z_{mi}$
- Now there are *k* first stage regressions:
    - Regress X_1 on *all* the exogenous regressors: regress X_1 on $W_1, \dots, W_r$ and $Z_{1i}, \dots, Z_{mi}$ using OLS
    - Compute predicted values $\widehat{X_{1\imath}}$, $i = 1, \dots, n$
    - Regress X_2 on *all* the exogenous regressors: regress X_2 on $W_1, \dots, W_r$ and $Z_{1i}, \dots, Z_{mi}$ using OLS
    - Compute predicted values $\widehat{X_{2\imath}}$, $i = 1, \dots, n$
    - Repeat for all X's, obtaining $\widehat{X_{1\imath}}$,..., $\widehat{X_{k\imath}}$.

- Second stage
    - Regress Y on $\widehat{X_{1\imath}}$,..., $\widehat{X_{k\imath}}$ and $W_1, \dots, W_r$ using OLS
    - The coefficients from this second stage regression are the TSLS estimators (but the standard errors are wrong)

Where Do Valid Instruments Come From?

- Valid instruments are (1) relevant and (2) exogenous
- One general way to find instruments is to look for exogenous variation – variation that is "as if" randomly assigned in a randomized experiment – that affects X.
    - Rainfall shifts the supply curve for butter but not the demand curve; rainfall is "as if" randomly assigned
    - Sales tax shifts the supply curve for cigarettes but not the demand curve; sales taxes are "as if" randomly assigned

“Weak Instruments”

- If $Cov(Z, X)$ is weak, IV no longer has such desirable asymptotic properties
- IV estimates are not unbiased, and the bias tends to be larger when instruments are weak (even with very large datasets)
- Weak instruments tend to bias the results towards the OLS estimates
- Adding more and more instruments to improve asymptotic efficiency does not solve the problem

Recommendation always test the ‘strength’ of your instrument(s) by reporting the *F*-test on the instruments in the first stage regression

Summary: IV Regression

- A valid instrument lets us isolate a part of X that is uncorrelated with ε, and that part can be used to estimate the effect of a change in X on Y
- IV regression hinges on having valid instruments:
  - A valid instrument isolates variation in X that is “as if” randomly assigned.
  - The critical requirement of at least m valid instruments cannot be tested – *you must use your head*.

Specification Tests

Testing for Endogeneity - Wu-Hausman Test

- Since OLS is preferred to IV (or TSLS) if we do not have an endogeneity problem, we'd like to be able to test for endogeneity
- If we do not have endogeneity, both OLS and IV are consistent, but IV is inefficient
- Idea of Hausman test is to see if the estimates from OLS and IV are different
- Auxilliary regression is easiest way to do this test

- Consider the following regression:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 W_{1i} + \beta_3 W_{2i} + \varepsilon_i$$

- With Z_{1i} and Z_{2i} as additional exogenous variables (i.e. additional instruments)
- If $X_1$is uncorrelated with *Y* we should estimate this equation by OLS
- Hausman (1978) suggested comparing the OLS and TSLS estimates and determining whether the differences are significant. If they differ significantly, we conclude that X_1 is an endogenous variable.
- This can be achieved by estimating the first stage regression:

$$X_{1i} = \alpha_0 + \alpha_1 Z_{1i} + \alpha_2 Z_{2i} + \alpha_3 W_{1i} + \alpha_4 W_{2i} + \upsilon_i$$

- Since each instrument is uncorrelated with ε_i, X_{1i} is uncorrelated with ε_i only if υ_i is uncorrelated with ε_i.
- To test this, we run the following regression using OLS:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 W_{1i} + \beta_3 W_{2i} + \delta_1 \hat{\upsilon}_i + error$$

- And test whether $\delta_1 = 0$ using a standard t-test (If we reject the null hypothesis we conclude that X_1 is endogenous, since υ_i and ε_i will be correlated).

Testing Overidentifying Restrictions

- IV must satisfy two conditions:

  (1) *relevance*: $Cov(z, x) \neq 0$

  (2) *exogeneity*: $Cov(z, \varepsilon) = 0$
- We cannot test (2) because it involves a correlation between the IV and an unobserved error.
- If we have more than one instrument however, we can effectively test whether some of them are uncorrelated with the structural error.
- Consider the above example:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 W_{1i} + \beta_3 W_{2i} + \varepsilon_i$$

- With Z_1 and Z_2 as additional exogenous variables (i.e. additional instruments)
- Estimate this equation by IV using only Z_1 as an instrument, and compute the residuals, $\hat{\varepsilon}_i$.
- We can now test whether Z_2 and $\hat{\varepsilon}_i$ are correlated. If they are, $Z_2$is not a valid instrument.
- This tells us nothing about whether Z_1 and $\hat{\varepsilon}_i$ are correlated (in fact, for this test to be relevant we have to assume that they are not)
- If however, the two instruments are chosen using the same logic (e.g. mother's and father's education levels) finding that Z_2 and $\hat{\varepsilon}_i$ are correlated casts doubt on the use of Z_1 as an instrument.
- Note: if we have a single instrument then there are no overidentifying restrictions and we cannot use this test; if we have two IVs for X_1 we have one overidentifying restriction; if we have three we have two overidentifying restrictions, and so on.

General Method for Testing Overidentifying Restrictions (The Sargan test)

- Estimate the structural equation by TSLS and obtain the residuals, $\hat{\varepsilon}_i$.
- Regress $\hat{\varepsilon}_i$ on all exogenous variables. Obtain the R-squared, R_1^2.
- Under the null hypothesis that all IVs are uncorrelated with ε, $nR_1^2 \sim \chi_q^2$, where q is the number of instrumental variables from outside the model minus the total number of endogenous explanatory variables.
- If the test statistic exceeds the critical value we reject the null hypothesis and conclude that at least some of the IVs are not exogenous.

Examples 1

Instrumental Variables and Omitted Variable Bias

- Example from Angrist and Krueger (1991)
- A wage regression of the form:
- $ln\,(Wage)\; = \;\beta_0 + \;\beta_1 School\; + \;\varepsilon$
- may suffer from a bias due to omitted ability
- Part of the covariation in schooling and wages is because both are affected by the ability of the person
- Angrist and Krueger suggest using quarter of birth as an instrument because part of the variation in school years is because of month of birth (with minimum leaving age laws).
  - e.g. if children are required to enter school in the September of the year in which they turn six and assume that December the 31st is the cut-off date, then children born in the first quarter will be 6 and ¾ when they enter school, while those born in the fourth quarter will be 5 and ¾. So children have different lengths of schooling due to their birth dates.
  - Thus, individuals born earlier in the year reach the minimum school leaving age (e.g. 16th or 17th birthday) at a lower grade than people born later in the year.
  - Therefore, those who want to drop out as soon as legally possible can leave school with less education.
- Wage effects of that part of the variation in school years is not due to the impact of omitted ability. Quarter of birth can thus be used as an instrument for schooling in the above equation.
- But, the correlation between quarter of birth and education is fairly weak (the F-statistic from the first stage regression is sometimes less than two), i.e. a weak instruments problem.

Example 2

- See Wooldridge, Example 15.4 and the dataset card.wf1
- Card (1995) used wage and education data for a sample of men in 1976 to estimate the return to education
- He used a dummy variable for whether someone grew up near a four-year college (nearc4) as an IV for education
- In a log(wage) equation he included other standard controls: experience, a black dummy, dummy variables for living in an SMSA (standard metropolitan statistical area) and living in the South, and a full set of regional dummies, and an SMSA dummy for where the man was living in 1966.
- In order for nearc4 to be a valid instrument it must be correlated with the error term (assume this)
- It must also be partially correlated with education (to check this regress educ on nearc4 and the other exogenous variables). The coefficient implies, holding other things constant, people who lived near a college had, on average, a third of a year more education than those who did not grow up near a college.
- Turning to the wage equation, if we compare the coefficients from the OLS and IV regression we find that the coefficient from the IV regression is nearly twice as large as the OLS estimate.
- The standard error of the IV estimate is much larger than that from the OLS estimate however – the larger confidence interval is the price we pay to get a consistent estimator of the return to education.