

Multiple Instrumental Variables

Our discussion to this point has emphasized some of the problems that IV is designed to address (measurement error, omitted variables) with the use of a single instrument for each X variable. We now consider the possibility of the availability of multiple instruments (with some or all being correlated with more than one variable in X) and set the stage for a discussion of the problems of *simultaneity* and *identification*.

Even though it is hard to find one truly valid instrument, it is often the case that there are many candidates for instruments whose claims are similar. If mother's education can be instrument for education, why not father's education? If distance to an education center, why not travel time and travel time by public transportation? If streams, rivers, valleys, why not mountains and canyons?

Generalized IV Estimation

When there is more than one instrument for an endogenous variable, and keeping in mind that an exogenous variable 'serves as an instrument for itself', the total number of instruments exceeds the number of X 's. That is, there are potentially more Z 's than X 's. This situation is often called 'GIV' or 'GIVE'.

(Now following Greene pp. 78-83.) The crucial things about instruments: not correlated with disturbance ($plim n^{-1}Z'\varepsilon = 0$); also correlated with X , structurally redundant (discussion: how do we interpret 'structurally redundant' for the exogenous X 's, which are 'instruments for themselves'?)

If elements of Z are ('asymptotically') uncorrelated with ε , then all linear combinations of Z are ('asymptotically') uncorrelated with ε . This suggests the 'projection of the columns of X in the column space of Z :'

$$\hat{X} = Z(Z'Z)^{-1}Z'X$$

'Composite IV' and 2SLS

What is $\hat{X} = Z(Z'Z)^{-1}Z'X$? It's the fitted values for X of the regression of X on Z . Recall the matrix P that turned y 's into \hat{y} 's, (Lecture 2, $P = X(X'X)^{-1}X'$; call that P_X). We can thus write $\hat{X} = P_Z X$ (this does not do us that much good right now, but is useful later.) By the foregoing arguments, if the Z 's are valid instruments, so are the columns of \hat{X} . And \hat{X} offers the following advantages: (1) it has the same number of columns (K) as X ; (2) it 'incorporates all the information' that was in Z (well there's at least an intuition there.)

Using this \hat{X} in place of Z in the formula $(Z'X)^{-1}Z'y$ we get:

$$\begin{aligned} b_{IV} &= (\hat{X}'X)^{-1}\hat{X}'y \\ &= [X'Z(Z'Z)^{-1}Z'X]^{-1}X'Z(Z'Z)^{-1}Z'y, \end{aligned}$$

which expresses the estimator in terms of X and Z .

Why It's Called 'Two Stage Least Squares'

You could calculate this estimator by regressing X on Z to get \hat{X} , and then regress y on \hat{X} . Thus 'two stage least squares.' You could do it this way, but you probably shouldn't: (1) the output from the second stage regression doesn't anything but the correct coefficients (e.g. the standard errors are 'wrong'); (2) one is prone to make mistakes in the first stage (not including all the X 's as regressors for each and every Z); (3) it doesn't generalize in any way (for example, to nonlinear models). To see that you can calculate this as 2SLS:

$$\begin{aligned}b_{IV} &= (\hat{X}'X)^{-1}\hat{X}'y \\b_{IV} &= (X'P_Z'X)^{-1}X'y \\b_{IV} &= (X'P_Z'P_ZX)^{-1}X'y \\b_{IV} &= (\hat{X}'\hat{X})^{-1}\hat{X}'y,\end{aligned}$$

since P_Z is symmetric and idempotent.

'Efficiency' of 2SLS

Suppose we thought about choosing not $\hat{X} = P_Z X = Z(Z'Z)^{-1}Z'X = ZG$ (say) as our instruments but ZF , a different linear combination of Z . We can calculate the variance of the resulting estimate of β and show that it (asymptotically) exceeds that of 2SLS. We will do this later, towards the end of the course, when we interpret GIVE as a GMM estimator that exploits the moment conditions $E(Z'\varepsilon) = 0$.

The Durbin–Wu–Hausman Test

(I think it made it a big splash when Hausman did it because he explained it better; Hausman (1978), Wu (1973), Durbin (1954).)

Suppose we are not sure whether a variable is endogenous or not. To treat it as endogenous is 'safer' but costly: it's hard to find good instruments and even if we do, we know from the Gauss Markov Theorem that the resulting estimator must be less efficient than OLS (i.e. OLS treating the variable as exogenous.) One notion behind the Hausman test is that if you compute b_{IV} and b_{OLS} and they are pretty much the same, you might as well take b_{OLS} as your estimate. More formally, under the null of exogeneity, both estimators are consistent; under the alternative, only b_{IV} is. The test is designed to exploit this difference.

The Durbin–Wu–Hausman Test (2)

Let $d = b_{IV} - b_{OLS}$. Then we can base a test on

$$H = d' \{Est.Asy.Var[d]\}^{-1} d;$$

this style of test, where we look at a quadratic form of a difference that is zero under the null with inverse of its estimated asymptotic variance is called a *Wald* test. We will study these further during our unit on ML estimation next week.

The next steps in the analysis are these: (1) what is the asymptotic variance of d ?; (2) what is the distribution of H ?. Hausman noted: **the covariance between an efficient estimator, b_E , of a parameter vector, β , and its difference from an inefficient estimator, b_I , of the same parameter vector, $b_E - b_I$, is zero.**

If this were not true, you could exploit the covariance to construct an estimator that was more efficient than either of them, thereby contradicting the hypothesis that b_E is efficient.

The Durbin–Wu–Hausman Test (3)

Exploiting the insight:

$$\begin{aligned}Cov(b_E, b_E - b_I) &= 0 \\Cov(b_E, b_E - b_I) &= E[(b_E - \bar{b}_E)(b_E - b_I)] \\&= V(b_E) - cov(b_E, b_I), \quad \text{so} \\V(b_E) &= cov(b_E, b_I), \quad \text{and}\end{aligned}$$

$$Asy.Var[b_{IV} - b_{LS}] = Asy.Var[b_{IV}] - Asy.Var[b_{OLS}],$$

all of this under the null, of course. So we now know what to substitute for $Est.Asy.Var[d]$ in $H = d' \{Est.Asy.Var[d]\}^{-1}d$. **HOWEVER**, this does not have K degrees of freedom (as one might think at first glance; K is number of X variables) but rather $K^* = K - K_0$ degrees of freedom, where K_0 is the number of X 's known (i.e. assumed) to have no correlation with the disturbance. This additionally means $Est.Asy.Var[d]$ does not have full rank and a generalized inverse must be used.

Wu's Regression

It is simpler in this particular context to test the hypothesis at hand by running Wu's regression:

$$y = X\beta + \hat{X}^*\gamma + \varepsilon^*,$$

where \hat{X}^* are the elements of \hat{X} corresponding to the endogenous X 's; this is a clever way of relaxing the orthogonality conditions $X'\varepsilon$ for the endogenous elements of X .

The problem with Wu's regression is that is not general, while the Hausman procedure is:

$$(\hat{\theta}_I - \hat{\theta}_E)' \{Est.Asy.Var[\hat{\theta}_I] - Est.Asy.Var[\hat{\theta}_E]\}^{-1} (\hat{\theta}_I - \hat{\theta}_E)$$

is 'generally' $\chi^2(J)$, where J is context dependent.

Systems of Regression Equations

We now want to think about systems of equations, which roughly come in two varieties. In the first type of system, there is no simultaneity: that is, each equation has a 'different' y and *no* y ever occurs on the right hand side of an equation. Moreover, the X 's in each of the individual regressions are orthogonal to their respective disturbances. In such cases, we *could* just estimate each equation separately by OLS. We can usually do better; let's look more carefully (following Greene pp. 339–347.)

$$\begin{aligned}y_1 &= X_1\beta_1 + \varepsilon_1 \\y_2 &= X_2\beta_2 + \varepsilon_2 \\&\vdots \\y_M &= X_M\beta_M + \varepsilon_M\end{aligned}$$

There are M equations and T observations.

The 'Stacked' Model

You can think of the problem posed by the system as a giant OLS problem to be solved all at once (the question will be whether solving it 'all at once' is advantageous; often it will be.) Write:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & X_M \end{bmatrix} + \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_M \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_M \end{bmatrix} = X\beta + \varepsilon$$

For the t^{th} observation let's write the $M \times M$ covariance matrix of the disturbances as:

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1M} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2M} \\ & & \vdots & \\ \sigma_{M1} & \sigma_{M2} & \cdots & \sigma_{MM} \end{bmatrix}$$

so the WHOLE covariance matrix can be written: $\Omega = \Sigma \otimes I$ and $\Omega^{-1} = \Sigma^{-1} \otimes I$.

Seemingly Unrelated Regressions

When Σ the contemporaneous covariance matrix of the disturbances is not diagonal, these regressions are 'seemingly' unrelated because there are no coefficient restrictions across equations. However, the covariance of the disturbances can be exploited to get a more efficient estimate than one obtains with equation by equation OLS. This estimator is called 'Generalized Least Squares' (GLS) and is given by:

$$\hat{\beta} = [X'\Omega^{-1}X]^{-1}X'\Omega^{-1}y$$

where Ω^{-1} restores 'homoscedasticity' (discussion; ie. it's like transforming X and y by $\Omega^{-1/2}$; single equation analogy.)

SUR: Two special cases.

Equation by equation OLS is fully efficient when

1. The equations are actually unrelated, in that Σ is diagonal.
2. The exact same explanatory variables appear in every equation.

The first case is fairly intuitive. The second is not (at least to me.)

Feasible GLS is the estimator that results when we do not know Σ and so must estimate it in a first stage. (Well, not must—might.) More precisely means any procedure which must estimate Σ .

Simultaneous Equations Models

Following Greene, pp. 378–395. We start from an example, more or less the simplest one possible, a system with a supply and a demand equation.

$$\begin{aligned}q_{d,t} &= \alpha_1 p_t + \alpha_2 x_t + \varepsilon_{d,t} \\q_{s,t} &= \beta_1 p_t + \quad \quad + \varepsilon_{s,t} \\q_{d,t} &= q_{s,t}\end{aligned}$$

This is called the structural form (discussion.) Assuming the disturbances are ‘well-behaved’, we can rewrite this so p and q are determined directly and explicitly by $x_t, \varepsilon_{d,t}, \varepsilon_{s,t}$, i.e.

$$\begin{aligned}p &= \frac{\alpha_2 x}{\beta_1 - \alpha_1} + \frac{\varepsilon_d - \varepsilon_s}{\beta_1 - \alpha_1} = \pi_1 x + v_1 \\q &= \frac{\beta_1 \alpha_2 x}{\beta_1 - \alpha_1} + \frac{\beta_1 \varepsilon_d - \alpha_1 \varepsilon_s}{\beta_1 - \alpha_1} = \pi_2 x + v_2\end{aligned}$$

Structural vs. Reduced Forms

$$q_{d,t} = \alpha_1 p_t + \alpha_2 x_t + \varepsilon_{d,t}$$

$$q_{s,t} = \beta_1 p_t + \quad + \varepsilon_{s,t}$$

$$q_{d,t} = q_{s,t}$$

$$p = \frac{\alpha_2 x}{\beta_1 - \alpha_1} + \frac{\varepsilon_d - \varepsilon_s}{\beta_1 - \alpha_1} = \pi_1 x + v_1$$

$$q = \frac{\beta_1 \alpha_2 x}{\beta_1 - \alpha_1} + \frac{\beta_1 \varepsilon_d - \alpha_1 \varepsilon_s}{\beta_1 - \alpha_1} = \pi_2 x + v_2$$

A reduced form shows the outcome of the endogenous variables for values of the exogenous variables and disturbances. In classical linear models (such as the one above) it writes the outcome as a linear function of the endogenous variables, and shows how to derive the coefficients of those linear functions from the structural model's coefficients. (Informal discussion anticipating identification.)