

Econometrics 1
Mette Ejrnæs and Hans Christian Kongsted
Institute of Economics
University of Copenhagen

April 20, 2004.

NOTES ON INSTRUMENTAL VARIABLE ESTIMATION.¹

1 Correlation is not causality.

Suppose we have n observations on two variables X and Y . We formulate a linear equation for Y in terms of X (with a constant):

$$Y_i = \alpha + \beta X_i + u_i \quad (1)$$

and estimate $\hat{\beta}_{OLS}$. Suppose we find that $\hat{\beta}_{OLS} \neq 0$ (which is exactly the same as saying that Y and X are correlated). Can we conclude that increasing X by one unit will cause Y to change by $\hat{\beta}_{OLS}$ units? Generally not, since there are many reasons why we might find a non-zero OLS slope coefficient (or, equivalently, a non-zero correlation between the sampled X and Y). Here is a list.

- **Chance.** Given *any* two sets of n numbers it is very unlikely that the OLS coefficient would be exactly zero. For example, suppose the sample is the group of people in this course. If we took Y to be your mark in this course and X to be a dummy variable that is one if your mother's first name starts with a letter between A and L in the alphabet (and zero otherwise) then we would almost certainly find that $\hat{\beta}_{OLS} \neq 0$. This is purely due to chance and if we took another sample we might as well find a coefficient with the opposite sign. It does not mean that changing your mother's name would help you to get a good mark! We deal with the possibility of chance correlations by using standard statistical inference procedures - basically looking at the 't-value' on the coefficient.
- X causes Y but Y does not cause X . Then (and only then) can we interpret the OLS coefficient to be the most likely estimate of the effect on the mean value of Y of changing X (and even then this depends on the functional form). Example: let Y be the height of an adult female and let X be the height of her mother (we would want to include other covariates, such as the height of the father etc.).
- Y causes X but X does not cause Y . In this case changing X would not cause Y to change. Example (the reverse of the previous example): let X

¹Most of the material in this note is borrowed from a note by Martin Browning for the 2002 course Econometrics.

be the height of an adult female and let Y be the height of her mother. If we now regress Y on X we will find a positive value for the OLS coefficient. We would not conclude that it is tall daughters who ‘cause’ tall mothers. Note that here theory (basic genetics and the fact that time flows one way) establishes the interpretation of the OLS coefficient. That is *always* true in econometrics: it is only by using theory that we can identify causal relationships.

- Y causes X and X causes Y . This is the simultaneous equation case. The classic example is if the two variables are the sales and price of a particular good.
- Y and X are caused by some other variables W which they do not cause. Example: let Y_i be the length of person i 's right leg and X_i be the length of person i 's left leg. For almost any sample of people we would have $\hat{\beta}_{OLS}$ close to one and ‘highly significant’ (a very small standard error). This does not mean that one leg is causing the growth of the other but that they are jointly determined by genes and early nutrition.

To repeat: correlation is not causality. In fact, it is almost never causality!

2 Stochastic right hand side variables.

Introductory courses in statistics consider the case in which the X variables are ‘given’ (or ‘fixed in repeated samples’) - that is: the probability that observation i for variable j is X_{ij} is one. Assuming that X is given allows us to assert, for example, the following:

$$E(X'u) = X'E(u) = 0 \text{ if } E(u) = 0 \quad (2)$$

which in turn gives that OLS is unbiased. The assumption that the X variables are given may be appropriate in an experimental setting where the investigator chooses the X values but it is never true for non-experimentally generated data.

This is the main reason that in the present course, we have treated both Y and X as a random sample from the population. This is Assumption MLR.2 (Wooldridge p. 85). Still, we have shown in Theorems 2.1 and 3.1 that OLS remains unbiased for a random sample from a population that is subject to the zero conditional mean assumption, $E(u|X) = 0$ (Assumption MLR.3, Wooldridge p. 85). For consistency, a weaker assumption of zero correlation between u and X is needed, $cov(u_i, X_{ij}) = 0, \quad j = 1, 2, \dots, k$. This is Assumption MLR.3', Wooldridge p. 168.

We have now seen a number of examples in which the explanatory variables do not satisfy MLR.3' (and therefore also not MLR.3). These are cases in which the explanatory variables are correlated with u and thus endogenous: Omitted variables (chapter 5), functional form misspecification (chapter 9) and measurement error (also chapter 9). This makes life considerably more difficult and we need to address large sample or asymptotic results. A brief review of the large sample theory we need for this is found in the Appendix to this note.

3 Exogeneity and endogeneity.

At this point it is useful to review the derivations to establish consistency of OLS. For this we apply consistency to random vectors and matrices. For example, suppose that we have an $n \times k$ data matrix X . Consider the $k \times k$ matrix $X'X$ and let n 'become large'. As we add more observations the elements of this matrix grow without bound. For example, for the second element in the diagonal we have:

$$(X'X)_{2,2} = \sum_{i=1}^n (X_{i2})^2 \quad (3)$$

The limit as n tends to infinity for this value is ∞ . If, however, we divide by n then the value will tend to 'settle down' to a particular value. Specifically we assume that

$$plim \left(\frac{1}{n} \sum_{i=1}^n (X_{i2})^2 \right) = \sigma_{22} < \infty \quad (4)$$

If we assume this for all the elements of $X'X$ then we have:

$$plim \left(\frac{1}{n} X'X \right) = \Sigma_{XX} \quad (5)$$

where Σ_{XX} is a $k \times k$ matrix. One really useful example of the so-called Slutsky property (see the Appendix) is that the probability limit of an inverse is equal to the inverse of the probability limit:

$$plim \left(\left(\frac{1}{n} X'X \right)^{-1} \right) = \left(plim \left(\frac{1}{n} X'X \right) \right)^{-1} = (\Sigma_{XX})^{-1} \quad (6)$$

We say that an $n \times m$ matrix Z is *asymptotically uncorrelated* with a random $n \times 1$ vector u if:

$$plim \left(\frac{1}{n} Z'u \right) = \mathbf{0}_m \quad (7)$$

where $\mathbf{0}_m$ is an $m \times 1$ vector of zeros.

With this apparatus we can consider OLS when we have stochastic right hand side variables. Suppose we have a theory model:

$$Y = X\beta + u \quad (8)$$

and that we assume:

$$\begin{aligned} plim \left(\frac{1}{n} X'X \right) &= \Sigma_{XX} \text{ a non-singular matrix} \\ plim \left(\frac{1}{n} X'u \right) &= \mathbf{0} \end{aligned} \quad (9)$$

The latter states that the X variables and the errors u are asymptotically uncorrelated. The OLS estimator is:

$$\hat{\beta}_{OLS} = (X'X)^{-1} X'Y = \beta + (X'X)^{-1} X'u \quad (10)$$

As noted above, assumption (9) is not sufficient to establish unbiasedness of OLS. We can, however, show that OLS is consistent since:

$$\begin{aligned}
plim(\hat{\beta}_{OLS}) &= plim\beta + plim\{(X'X)^{-1}X'u\} \\
&= \beta + plim\left(\left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{n}X'u\right)\right) \\
&= \beta + plim\left(\frac{1}{n}X'X\right)^{-1}plim\left(\frac{1}{n}X'u\right) \\
&= \beta + (\Sigma_{XX})^{-1}\mathbf{0}_k = \beta
\end{aligned} \tag{11}$$

Thus OLS is consistent if the right hand side variables are asymptotically uncorrelated with the error term.

Wooldridge gives a number of examples of when this is *not* a good assumption, e.g. the classical measurement error model of section 9.3 and the wage equation where education is correlated with unobserved ability, section 15.1. In each case the inconsistency of OLS arises from the fact that one of the right hand side variables is asymptotically correlated with the error term. We say that a variable X_j is endogenous if this is the case. That is, variable X_j is *endogenous* if:

$$plim\left(\frac{1}{n}\sum_{i=1}^n X_{ij}u_i\right) \neq 0 \tag{12}$$

If the asymptotic correlation is zero then we say that variable X_j is *exogenous*. There are, in fact, many varieties of exogeneity but this is the basic idea. Note that if at least one right hand side variable is endogenous then this causes problems for all the OLS coefficients. That is, even if variable X_j is exogenous, the OLS coefficient on this variable will not be consistent if some other variable on the right hand side is endogenous.

If we have an endogenous variable, such as education in the wage equation, what can we do about it? We cannot simply leave it out of the equation to be estimated since this leads to omitted variable bias. The next section presents a method for dealing with the endogeneity problem.

4 Instrumental variables.

The wage equation example looks pessimistic. Inconsistency arose from leaving out some relevant variables that are correlated with the right hand side variable of interest, education. One solution would be to include all relevant variables, but some of these, such as ability, intelligence or the propensity to work hard, are inherently very difficult to measure. Surprisingly, it turns out that an alternative solution is to find some variable that should *not* be included in the equation of interest but that is correlated with the endogenous right-hand side variable.

Consider again the classical measurement error model from Wooldridge section 9.3. As an example, take the theory model to be a model for consumption,

Y^* , and income, X^* :

$$Y^* = \alpha + \beta X^* + u \quad (13)$$

We only observe the variables with some measurement error, so that the measurement model is:

$$\begin{aligned} Y &= Y^* + \eta \\ X &= X^* + \varepsilon \end{aligned} \quad (14)$$

where the η and ε are random variables. Then we can only estimate the parameters of the linear model for the observed data. In deviations from the mean we have:

$$\begin{aligned} y_i &= \beta x_i + (u_i + \eta_i - \beta \varepsilon_i) \\ &= \beta x_i + v_i \end{aligned} \quad (15)$$

The problem arises because the right hand side variable, income, is measured with error. Suppose this income measure X_i is the response by person i to a question on income in a survey. Suppose that we also have tax records for each person and we can observe a variable Z_i which is the tax measure of income. One option would be to use this in our equation instead of the survey measure but it may also be measured with error. Remembering that the measurement error for X^* is ε , we assume the following (using deviations about the mean):

$$\begin{aligned} plim \left(\frac{1}{n} \sum_{i=1}^n z_i x_i \right) &= \sigma_{xz} \neq 0 \\ plim \left(\frac{1}{n} \sum_{i=1}^n z_i \varepsilon_i \right) &= 0 \end{aligned} \quad (16)$$

That is, the new variable Z (taxable income) is correlated with the observed value of X (self-reported income) but it is uncorrelated with the measurement error for X . We call a variable that is correlated with an endogenous variable but that is uncorrelated with the error term an *instrumental variable* (or, more simply, an instrument).

For the wage equation, no one has been able to find a convincing instrument for education. That is, a variable that partly determines the length of education of an individual but does not directly impact on wages. If you can think of one, then let us know and we can publish and become world famous (in economics).

To show how we use an instrument, consider equation (15) and an arbitrary coefficient value equal to $\tilde{\beta}$. This defines n errors $\tilde{v}_i = y_i - x_i \tilde{\beta}$. We would like to choose $\tilde{\beta}$ to set all these errors to zero, but that is impossible since we have n such errors and only one parameter β . What we can do, however, is to choose β so that the implied errors are uncorrelated with the z variables. That is, choose β so that:

$$\frac{1}{n} \sum_{i=1}^n z_i \tilde{v}_i = 0 \quad (17)$$

This gives one equation in one unknown. Solving gives the *instrumental variable estimator*:

$$\begin{aligned} 0 &= \frac{1}{n} \sum_{i=1}^n z_i \hat{v}_i = \frac{1}{n} \sum_{i=1}^n z_i (y_i - \hat{\beta}_{IVE} x_i) \Rightarrow \\ \hat{\beta}_{IVE} &= \frac{\sum_{i=1}^n z_i y_i}{\sum_{i=1}^n z_i x_i} \end{aligned} \quad (18)$$

This estimator is consistent. To see this, take probability limits:

$$plim \left(\hat{\beta}_{IVE} \right) = \beta + \frac{plim \frac{1}{n} \sum_{i=1}^n z_i v_i}{plim \frac{1}{n} \sum_{i=1}^n z_i x_i} = \beta \quad (19)$$

Thus the instrumental variable estimator $\hat{\beta}_{IVE}$ is consistent.

We now put all of this in a general framework with k right hand side variables:

$$Y = X\beta + u \quad (20)$$

Let the asymptotic correlation between X_j and the error term be given by:

$$plim \frac{1}{n} \sum_{i=1}^n X_{ij} u_i = \sigma_{ju} \quad (21)$$

If this is non-zero then variable X_j is endogenous and we need to find instruments. Suppose we have l endogenous variables and arrange the $n \times k$ matrix X so that these are the last l columns. Thus the first $k - l$ columns of X are exogenous variables. To proceed, we need at least l exogenous variables that are correlated with the l endogenous variables, that is, l instruments. If we have exactly as many instruments as endogenous variables then we say that the model is *just identified*. Arrange the instrumental variables in an $n \times k$ matrix Z where the first $k - l$ columns are the same as the first $k - l$ columns of X and the last l columns are the instruments. Thus:

$$\begin{aligned} X &= [X_1 \ X_2 \ \dots \ X_{k-l} \ X_{k-l+1} \ \dots \ X_k] \\ Z &= [X_1 \ X_2 \ \dots \ X_{k-l} \ Z_1 \ \dots \ Z_l] \end{aligned} \quad (22)$$

where X_j is the column vector for the j th variable (and we usually take X_1 to be a column of ones). There are sometimes problems with terminology here. We shall call the matrix X a matrix of ‘included variables’ and the matrix Z a matrix of instruments, where it is understood that an included variable X_j that is exogenous is an instrument for itself. By assumption we have:

$$plim \frac{1}{n} \sum_{i=1}^n Z_{ij} u_i = 0 \Rightarrow plim \frac{1}{n} Z' u = \mathbf{0} \quad (23)$$

and the X ’s and Z ’s are correlated so that the $k \times k$ matrices:

$$\begin{aligned} plim \frac{1}{n} X' Z &= \Sigma_{ZX} \\ plim \frac{1}{n} Z' Z &= \Sigma_{ZZ} \end{aligned} \quad (24)$$

are non-singular (have rank k). We now present four different ways to derive the IV estimator. These all give the same answer but they are useful in different places later.

Method 1. This follows the development above for the measurement error example. Set the residuals to be uncorrelated with the instruments which gives k equations in k unknowns:

$$\begin{aligned} 0 &= Z'\hat{u} = Z'(Y - X\hat{\beta}_{IVE}) \\ \Rightarrow \hat{\beta}_{IVE} &= (Z'X)^{-1} Z'Y \end{aligned} \quad (25)$$

Note that this requires that $Z'X$ be non-singular in the sample, which is not quite the same as the assumption made above that the probability limit of this matrix is non-singular. Taking probability limits we have:

$$\begin{aligned} plim\hat{\beta}_{IVE} &= \beta + plim\left(\frac{1}{n}Z'X\right)^{-1} plim\left(\frac{1}{n}Z'u\right) \\ &= \beta + \Sigma_Z^{-1}\mathbf{0} = \beta \end{aligned} \quad (26)$$

so that the instrumental variable estimator $\hat{\beta}_{IVE}$ is consistent for β . Note that if all the X variables are exogenous then $Z = X$ and we have the OLS formula. Thus OLS is a special case of IVE.

Method 2. Two stage least squares (2SLS). As the name suggests this is a two stage procedure. In the first stage we obtain predictions of the X variables from regressing (using OLS) on the Z variables. Thus take the system of equations:

$$X = Z\Pi + E \quad (27)$$

where Π is a $k \times k$ matrix of parameters and E is an $n \times k$ matrix of residuals. These regressions of the right hand side variables on the instruments are sometimes known as the *reduced forms* for X or as the *auxiliary equations* for the endogenous variables. If we use the OLS formula to estimate we have parameter estimates:

$$\hat{\Pi} = (Z'Z)^{-1} Z'X \quad (28)$$

and predicted values for the X variables:

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

where $P_Z = Z(Z'Z)^{-1} Z'$. The P_Z matrix is sometimes known as the projection matrix; it is symmetric ($P_Z = P_Z'$) and idempotent ($P_Z P_Z = P_Z$). Note that the predicted values for an exogenous variable given by this formula is just the exogenous variable itself. The important fact about these predictions is that

they are asymptotically uncorrelated with the error terms:

$$\begin{aligned}
plim \left(\frac{1}{n} \hat{X}'u \right) &= plim \left(\frac{1}{n} X'Z (Z'Z)^{-1} Z'u \right) \\
&= plim \left(\frac{1}{n} X'Z \left(\frac{1}{n} Z'Z \right)^{-1} \frac{1}{n} Z'u \right) \quad (29) \\
&= plim \left(\frac{1}{n} X'Z \right) plim \left(\frac{1}{n} (Z'Z)^{-1} \right) plim \left(\frac{1}{n} Z'u \right) \\
&= \Sigma_{ZX} (\Sigma_{ZZ})^{-1} \mathbf{0} = \mathbf{0} \quad (30)
\end{aligned}$$

Thus the predicted X 's are exogenous and can be used in a second OLS step:

$$\begin{aligned}
\hat{\beta} &= (\hat{X}'\hat{X})^{-1} \hat{X}'Y \\
&= (X'Z (Z'Z)^{-1} Z'Z (Z'Z)^{-1} Z'X)^{-1} X'Z (Z'Z)^{-1} Z'Y \\
&= (Z'X)^{-1} (Z'Z) (X'Z)^{-1} X'Z (Z'Z)^{-1} Z'Y \\
&= (Z'X)^{-1} Z'Y \quad (31)
\end{aligned}$$

which is exactly the IVE formula we had before. Note that in this derivation we could invert $X'Z$ since it is an $k \times k$ square matrix. We also used $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$ for square matrices.

Method 3. Perform step 1 as in 2SLS and then use the values as instruments. This is not strictly a new method, since it uses the IV formula. We have:

$$\hat{\beta} = (\hat{X}'X)^{-1} \hat{X}'Y = \hat{\beta}_{IVE} \quad (32)$$

You should complete this to show that this gives the same formula as before.

Method 4. Residual augmented regressions. Once again, do the first stage of 2SLS and find the predicted values \hat{X} and also the residuals from the first stage: $\hat{E} = X - Z\hat{\Pi}$. Note that the two sets of variables \hat{X} and \hat{E} are uncorrelated in the sense that $\hat{X}'\hat{E} = 0$ (you should show this as an exercise in matrix manipulations). In step two we regress Y on both \hat{X} and \hat{E} :

$$Y = \hat{X}\beta + \hat{E}\gamma + v \quad (33)$$

to give:

$$\begin{bmatrix} \hat{\beta}_{IVE} \\ \hat{\gamma}_{IVE} \end{bmatrix} = \begin{bmatrix} \hat{X}'\hat{X} & \hat{X}'\hat{E} \\ \hat{E}'\hat{X} & \hat{E}'\hat{E} \end{bmatrix}^{-1} \begin{bmatrix} \hat{X}'Y \\ \hat{E}'Y \end{bmatrix} \quad (34)$$

$$= \begin{bmatrix} \hat{X}'\hat{X} & 0 \\ 0 & \hat{E}'\hat{E} \end{bmatrix}^{-1} \begin{bmatrix} \hat{X}'Y \\ \hat{E}'Y \end{bmatrix} \quad (35)$$

so that $\hat{\beta}_{IVE}$ is the same as before (you should check this).

To finish off this section we note that if we want predicted values of the left hand side variable Y from the IV regression then we should take:

$$\hat{Y} = X\hat{\beta}_{IVE} \quad (36)$$

That is, we use the actual values of the X variables to predict and not the predicted values \hat{X} . As an exercise you should determine whether $Z'\hat{u} = 0$ and/or $X'\hat{u} = 0$ where \hat{u} is the vector of residuals $\hat{u} = Y - \hat{Y}$.

5 More instruments than endogenous variables.

In all of the last section we assumed that we had exactly as many instruments as endogenous variables (the ‘just identified case’). What if we have more instruments than endogenous variables? For example, suppose that in the measurement error example above we have two variables Z_1 and Z_2 both of which are valid instruments. To make things easy for ourselves we shall always assume that Z_1 and Z_2 have mean zero (or that they are deviations from the mean) and denote them z_1 and z_2 . If they are good instruments then:

$$\begin{aligned} plim \frac{1}{n} \sum_{i=1}^n z_{1i} X_i^* &\neq 0, \quad plim \frac{1}{n} \sum_{i=1}^n z_{2i} X_i^* \neq 0 \\ plim \frac{1}{n} \sum_{i=1}^n z_{1i} v_i &= 0, \quad plim \frac{1}{n} \sum_{i=1}^n z_{2i} v_i = 0 \end{aligned} \quad (37)$$

We now have two IV estimates of β :

$$\begin{aligned} \hat{\beta}_{IVE}^1 &= \frac{\sum_{i=1}^n z_{1i} y_i}{\sum_{i=1}^n z_{1i} x_i} \\ \hat{\beta}_{IVE}^2 &= \frac{\sum_{i=1}^n z_{2i} y_i}{\sum_{i=1}^n z_{2i} x_i} \end{aligned} \quad (38)$$

both of which are consistent. Thus we have two equations for one unknown. When we have more instruments than endogenous variables we say that the model is *over-identified*. Which of the two estimates should we use? We could just ignore one but that seems to be wasting information (it is ‘inefficient’). Alternatively we could take some weighted sum of the two, but how to choose the weights? It turns out that there is a ‘proper’ way to combine this information. We shall do this for the general model in which we have an $n \times k$ matrix X and an $n \times (k - l + g)$ matrix $Z = [X_1, X_2, \dots, X_{k-l}, Z_1, Z_2, \dots, Z_g]$, with $g > l$.

The preferred estimator can be derived as a variant on Method 3 above. We first regress the matrix X on Z to obtain the $n \times k$ matrix of predictions \hat{X} ($= Z(Z'Z)^{-1}Z'X = P_Z X$). Now use these k variables as instruments in a just

identified way:

$$\begin{aligned}
\hat{\beta}_{IVE} &= (\hat{X}'X)^{-1} \hat{X}'Y \\
&= (X'Z(Z'Z)^{-1}Z'X)^{-1} X'Z(Z'Z)^{-1}Z'Y \\
&= (X'P_ZX)^{-1} X'P_ZY
\end{aligned} \tag{39}$$

Since $X'Z$ is not square (if $g > l$) we cannot reduce this any further, as we did for the just identified case.

6 Inference with instrumental variables.

Up until now we have only considered parameter estimation using IVE. We also need to derive a covariance matrix so that we can test for linear restrictions and the significance of individual coefficients. The theory model is:

$$Y = X\beta + u \tag{40}$$

We denote:

$$\begin{aligned}
plim \left(\frac{1}{n} Z'Z \right) &= \Sigma_{ZZ} \\
plim \left(\frac{1}{n} X'Z \right) &= \Sigma_{XZ}
\end{aligned} \tag{41}$$

and assume that the Z 's are asymptotically uncorrelated with the error terms and that the latter are homoscedastic:

$$\begin{aligned}
plim \left(\frac{1}{n} Z'u \right) &= 0 \\
plim \left(\frac{1}{n^2} Z'u u' Z \right) &= \sigma^2 \Sigma_{ZZ}
\end{aligned} \tag{42}$$

We have:

$$\begin{aligned}
\hat{\beta}_{IVE} &= (X'P_ZX)^{-1} X'P_ZY \\
&= \beta + (X'P_ZX)^{-1} X'P_Zu
\end{aligned} \tag{43}$$

The *asymptotic covariance matrix* of the IV estimator is estimated by:

$$\begin{aligned}
Cov \left(\widehat{\beta}_{IVE} \right) &= \hat{s}^2 (X'P_ZX)^{-1} \\
\text{where } \hat{s}^2 &= \frac{1}{n} \sum_{i=1}^n (\hat{u}_i)^2
\end{aligned} \tag{44}$$

This estimated covariance matrix is easy to construct and can be used for testing purposes.

7 Testing for exogeneity.

In all of the above we have assumed that we know whether a variable is endogenous or not. In practice, we may suspect that a particular variable is endogenous, but we would like to test this. To illustrate how to do this suppose we have $k - l$ exogenous rhs variables $X^a = [X_1, X_2, \dots, X_{k-l}]$ and l variables $X^b = [X_{k-l+1}, \dots, X_k]$ that we think may be endogenous. We have:

$$Y = X^a \beta^a + X^b \beta^b + u \quad (45)$$

We wish to test whether X^b is endogenous. Suppose we have a valid set of instruments for X^b , the variables $[Z_1, Z_2, \dots, Z_g]$, where $g \geq l$.

If we take X^b to be exogenous then we would use OLS to estimate $(\hat{\beta}_{OLS}^a, \hat{\beta}_{OLS}^b)$ and use these estimates. Conversely, if X^b is taken to be endogenous then we would use the instruments and estimate $(\hat{\beta}_{IVE}^a, \hat{\beta}_{IVE}^b)$. If X^b is exogenous then OLS is best because it is minimum variance (see the Gauss-Markov theorem). If, on the other hand, X^b is endogenous then we should use IVE since OLS is inconsistent. Thus it is important to determine whether a particular variable is exogenous. There are two ways to test for exogeneity.

One widely used exogeneity test is known as the Wu-Hausman test. If X^b is exogenous then both OLS and IVE are consistent and we have $plim(\hat{\beta}_{OLS}^a, \hat{\beta}_{OLS}^b) = plim(\hat{\beta}_{IVE}^a, \hat{\beta}_{IVE}^b) = (\beta^a, \beta^b)$. Thus we could test whether the two sets of estimates are similar. If the OLS and IV estimates are close to each other then we conclude that X^b is exogenous. If, on the other hand, the two sets of estimates are quite different then it looks like instrumenting makes a difference and we conclude that X^b is endogenous. The formal version of this test is known as a Wu-Hausman test. Details are given e.g. in the textbook by Johnston and DiNardo (1997).

There is an ‘residual augmentation’ alternative to the Wu-Hausman test which is actually more useful and also allows us to write the full procedure for testing for and allowing for endogeneity. To develop this it is a very good idea to look again at the whole process, using the example we have taken here. We have two equations:

$$Y = X^a \beta^a + X^b \beta^b + u \quad (46)$$

$$X^b = Z \Pi^b + E^b \quad (47)$$

where the first equation is the equation of interest and the second is the auxiliary equation (reduced form) for the possibly endogenous variable, X^b . Clearly, X^b is perfectly correlated with E^b given Z . If E^b is also correlated with u then X^b is also correlated with u and so X^b is endogenous. Thus the endogeneity of X^b is equivalent to u and E^b being correlated. Let us write:

$$u_i = \rho E_i^b + \varepsilon_i \quad (48)$$

where E^b and ε are uncorrelated. Thus X^b is endogenous if and only if $\rho \neq 0$.

The theory model is now:

$$Y = X^a\beta^a + X^b\beta^b + E^b\rho + \varepsilon \quad (49)$$

If we could observe E^b then we could include it as a regressor and then X^b would be exogenous since it is uncorrelated with ε . But of course, we cannot include E^b in the equation of interest since we do not observe it. What we can do, however, is to include $\hat{E}^b = X^b - Z\hat{\Pi}^b$ based on consistent OLS estimates of Π^b from the auxiliary equation. This then allows for a test of exogeneity: the coefficient on \hat{E}^b ($= \hat{\rho}$) should be insignificant if X^b is exogenous.

We can generalise this to testing whether several right hand side variables are endogenous by including residuals from several auxiliary equations and performing joint tests for the significance of the included residuals in the equation of interest.

8 The IVE estimation procedure.

Based on this, we can now present a series of steps to follow in an empirical analysis. Some of these steps are similar to those given in method 4 to derive the IV estimator given above.

1. Write down the theory model and decide which right hand side variables are exogenous and which may be endogenous. We shall take the example above and suppose that some variables, X^a , are exogenous but another set of variables, X^b , may be endogenous.
2. Find potential instruments for the potentially endogenous variables. This is usually the most difficult part of any empirical analysis and the most important. It is also the step that we have least to say about! The usual place to look for justification of the exogeneity of a potential instrument is economic theory. Suppose that we can convince ourselves that a variable Z_j can be excluded from the equation of interest (that is, it is not a part of X^a nor X^b). Thus this variable is a potential instrument. This could, of course, also be a set of instruments.
3. Given one endogenous variable, we have one auxiliary (reduced form) equation (47) above. The variables in $Z = [X^a, Z_1, \dots, Z_g]$ are taken to be uncorrelated with the error term so we estimate the parameters of this equation by OLS. Now test whether the coefficients on $[Z_1, \dots, Z_g]$ are ‘significantly different from zero’. If that is not the case then you have a ‘weak instrument’ problem and you cannot test for exogeneity. In this case, you need either to find another potential instrument or to simply go ahead with OLS in the hope that X^b is exogenous. If the coefficients of $[Z_1, \dots, Z_g]$ in the reduced form equation are ‘significantly different from zero’ then we conclude that it is correlated with the endogenous variable X^b and hence are valid instruments. In this case, we proceed to the next step.

4. Given the OLS estimates of the auxiliary equations in (47), construct the residuals $\hat{E}^b = X^b - Z\hat{\Pi}^b$.
5. Estimate the parameters of the ‘residual augmented’ regression by OLS:

$$Y = X\beta + \hat{E}^b\rho + \varepsilon \quad (50)$$

Test whether $\hat{\rho}_{OLS}$ is significantly different from zero. If this is not the case, then we conclude that \hat{E}^b is not significant and that X^b is exogenous. If $\hat{\rho}_{OLS}$ is significantly different from zero then we conclude that X^b is endogenous.

6. If X^b is taken to be exogenous, run OLS for the equation of interest, equation (46). If X^b is taken to be endogenous, run IVE for this equation.

A Probability limits.

First a reminder of limits in the mathematical sense. We say that a sequence $\{a_n\}_{n=1,2,\dots,\infty}$ tends to a limit a if for ‘large’ n , all the values are very close to a . Formally if for any $\varepsilon > 0$ (no matter how small) we have that there is some number N such if $n > N$ then $|a_n - a| < \varepsilon$. We need a similar sort of definition when we have a sequence of estimators based on sample sizes of n so that $n > N$ corresponds to a large sample. To do this we have to take account of the fact that the a_n ’s are random variables. To see how we do this, consider the following example that should be familiar to you from your second year statistics. Suppose we sample independently from a distribution with unknown mean μ and variance σ^2 . Denote the sample mean by \bar{X}_n where the n subscript denotes explicitly that this statistic depends on the sample size. We have that the expected value of the sample mean $E(\bar{X}_n) = \mu$ and the variance is $V(\bar{X}_n) = \sigma^2/n$. Because of the latter, as n becomes large it becomes increasingly unlikely that \bar{X}_n is very different from μ . It is this idea of large sample convergence in a probabilistic sense that we wish to capture for any statistic L_n . Formally we say that a statistic L_n *converges in probability* to λ if for any $\varepsilon > 0$ we have:

$$\lim_{n \rightarrow \infty} \text{probability}(\lambda - \varepsilon \leq L_n \leq \lambda + \varepsilon) = 1 \quad (51)$$

This is sometimes written $\text{plim}L_n = \lambda$ or L_n is a *consistent estimator* of λ . One easy way to check consistency is to use the following sufficient condition:

$$\begin{aligned} \text{If } \lim_{n \rightarrow \infty} E(L_n) &= \lambda \text{ and } \lim_{n \rightarrow \infty} V(L_n) = 0 \\ \text{then } \text{plim}L_n &= \lambda \end{aligned} \quad (52)$$

Using this we see that \bar{X}_n is a consistent estimator of μ ($\text{plim}\bar{X}_n = \mu$).

Example 1 Consider the sample discussed above and define the sample variance by:

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

From your statistics course you should remember that this is a biased estimator of the population variance with:

$$E(S_n^2) = \frac{n-1}{n}\sigma^2 \quad (53)$$

However, as n becomes large the bias becomes small. In fact we can show that S_n^2 is consistent for σ^2 . That is, $\text{plim}S_n^2 = \sigma^2$. Thus the sample variance is a biased but consistent estimator of the population variance.

Suppose that we have two estimators a_n and b_n with $\text{plim}a_n = a$ and $\text{plim}b_n = b$. Then:

$$\begin{aligned} \text{plim}a_nb_n &= a \cdot b \\ \text{plim}\frac{a_n}{b_n} &= \frac{a}{b} \text{ if } b \neq 0 \end{aligned} \quad (54)$$

This is called the Slutsky property. It really is very useful; we shall use it over and over in the analysis.