# Topic 6: Endogeneity and Instrumental Variables

Dr. Ani Dasgupta
Department of International Business, Mass. Maritime Academy
and
Department of Economics, Boston University

*Draft lecture notes for a graduate econometrics course offered at Boston University, Spring 2018. Please do not circulate without permission.*

## 1 Motivation

From topic 3, we learnt that if $E(\varepsilon_i \,|\, \mathbf{x_i}) \neq 0$, we cannot expect unbiasedness of the OLS estimates. In the last topic we learnt that if instead of the exogeneity condition, if at least we have the orthogonality condition ($E(\mathbf{x_i}\,\varepsilon_i) = \mathbf{0}$), along with i.i.d.-ness of observations and a finite nonsingular second moment matrix of the regressors we can still get consistent estimates. Given zero mean errors, orthogonality simply means that the error term is uncorrelated with regressors. Unfortunately, in many scenarios even this is too much to ask for.

You have already encountered an example of this possibility. Recall that when we discussed the paper by Marc Nerlove on returns to scale in electricity production, I drew your attention to the curious fact that he chose to estimate the cost function and not the production function although, given his objective of examining returns to scale, the latter path would appear to be more natural. I pointed out one key reason for that: suppose the production function is $y = f(x, \beta) + \varepsilon$, and suppose that each firm has to produce a certain quantity of output (indirectly mandated by regulators) and they do so in a cost-minimizing fashion by choosing the inputs optimally. Now, think of the $\varepsilon$ term as an embodiment of technological efficiency (higher its realization is, the more efficient is our firm). So, clearly that term cannot be uncorrelated with the regressors; higher it is, ceteris paribus, lower are the regressors (assuming for example, a CRS technology and that the mandated output and input prices are not systematically correlated with firm efficiency).

We have just discussed an example of 'endogeneity' which is a serious problem in many applied econometric exercises (In fact, it is difficult to think of real-life problems where endogeneity will be completely absent). In the next section I provide you with a list of common situations where endogeneity occurs.

# 2 Causes of Endogeneity

## 2.1 Omitted Variables

This is perhaps the commonest source of the endogeneity problem. Suppose the true model is

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \gamma q + v \tag{1}$$

where $v$ is a zero-mean disturbance that is independent of all regressors. Suppose the variable $q$ is correlated with some of the $x$'s, and so $E(q|\mathbf{x})$ is some function of the $x$'s which is non-zero. Topic 4 has taught us that if we regress $y$ on $x_1, \ldots x_k$, we will not get unbiased estimates of $\beta_1, \ldots \beta_k$; we now show that we will get inconsistent estimates as well; moreover we will try to establish the extent by which the probability limits of the estimates will over or under estimate the betas.

To do this, let us write

$$q = \delta_0 + \delta_1 x_1 + \ldots + \delta_k x_k + r \tag{2}$$

where

$$\begin{pmatrix} \delta_1 \\ \vdots \\ \delta_k \end{pmatrix} = [Var(\mathbf{x})]^{-1} Cov(\mathbf{x}, q) \qquad \text{with} \qquad \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix} \tag{3}$$

$$\delta_0 = E(q) - \delta_1 E(x_1) - \ldots \delta_k E(x_k) \tag{4}$$

and

$$r = q - (\delta_0 + \delta_1 x_1 + \ldots + \delta_k x_k) \tag{5}$$

Equation 2 is known as the linear projection of $q$ on the $x$ variables, and it answers the following question: What linear function of the $x$'s is the best linear projector/forecaster of $q$, where 'best' is defined as producing the smallest expected mean squared error. This is an extremely useful formula and what it shows is that to do a (linear) forecast of $y$ via the $x$'s, you don't have to have a linear relation defined between those set of objects. In addition, as a byproduct of the way of defining the linear projection, one can also show that $E(r) = 0$ and $E(\mathbf{x}\, r) = \mathbf{0}$, so equation 2 looks just like a standard regression equation and $r$ looks just like a well-behaved error term in such an equation.

Now combining equations 1 and 2, we see that

$$y = (\beta_0 + \gamma\delta_0) + (\beta_1 + \gamma\delta_1)x_1 + \ldots + (\beta_k + \gamma\delta_k)x_k + (v + \gamma r) \tag{6}$$

It is now clear that if we regress $y$ on $x_1, \ldots x_k$, equation 6 is what we will be estimating. Since, $E(\mathbf{x}(v + \gamma r)) = \mathbf{0}$, we will obtain consistent estimates of the coefficients on the $x's$ in this equation. Hence the plim of the OLS coefficient on $x_j$ will be $\beta_j + \gamma\delta_j$ and not $\beta_j$.

To further examine the nature of this bias, suppose $q$ is correlated with only one of the $x$'s, say $x_k$. Then,

$$plim\, b_k = \beta_k + \gamma\, \theta Cov(x_k, q) \tag{7}$$

where $\theta$ is the last diagonal element of the $[Var(\mathbf{x})]^{-1}$ matrix (In the special case where $x_k$ is uncorrelated with the other $x_j$'s, $\theta = 1/Var(x_k)$). Now, since $\theta$ must be positive (why?) the sign of the bias depends on the signs of both $\gamma$ and the covariance between the omitted and the included variable. As an example, suppose we are regressing log wage on a bunch of variables such as age, experience, marital status, educational achievement etc. Suppose the omitted variable is 'innate ability'. If $x_k$ is education, and we assume that $\gamma$ is positive, then the bias will be positive if there is a positive correlation between educational level and innate ability. If that assumption is false (imagine a world full of Bill Gates), the bias will be negative.

## 2.2 Errors in Variable

When we measure variables for running a regression, we cannot always expect to measure them accurately. The question is does this affect the consistency of OLS estimates?

Suppose the true model is

$$y^* = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon \quad \text{with} \quad E(\varepsilon x_j) = 0 \quad \forall j \tag{8}$$

However, imagine that we do not have data on $y^*$; all we have is $y$, where $y = y^* + \varepsilon_y$, with $E(\varepsilon_y) = 0$. Then regressing $y$ against 1, $x_1, \ldots x_k$ is still okay, since we can rewrite the above equation as

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + (\varepsilon + \varepsilon_y) \tag{9}$$

As long as $E(\varepsilon_y x_j) = 0$ for all $j$, we can see that the new error term in the new equation is uncorrelated with the regressors and thus the estimated coefficients will come out consistently even when the lhs variable is measured with error.

Now suppose the true model is

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k^* + \varepsilon \quad \text{with} \quad E(\varepsilon x_j) = 0 \quad \forall j = 1, \ldots k-1 \quad \text{and } E(x_k^* \varepsilon) = 0 \tag{10}$$

and we only observe $x_k = x_k^* + \varepsilon_{x_k}$, with $E(\varepsilon_{x_k}) = 0$ and $E(x_k^*\, \varepsilon_{x_k}) = 0$ Now, rewriting $y$ in terms of $x_1, \ldots x_k$ gives

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + (\varepsilon - \beta_k \varepsilon_{x_k}) \tag{11}$$

Let $\theta = \varepsilon - \beta_k \varepsilon_{x_k}$. Then, we have

$$E(\theta x_k) = E(\varepsilon - \beta_k \varepsilon_{x_k})(x_k^* + \varepsilon_{x_k}) = -\beta_k E(\varepsilon_{x_k})^2 \tag{12}$$

which is nonzero (as long as $\beta_k$ is nonzero). Hence, we have endogeneity and inconsistency of OLS estimates. To see the direction of the (asymptotic) bias, think of this in terms of an omitted variable problem. While the true equation is equation 11, you are trying to regress $y$ on 1, $x_1, \ldots x_k$ only ($\varepsilon_{x_k}$ is the omitted variable). Hence, using equation 7, we see that

$$plim\, b_k = \beta_k - \beta_k \theta Cov(x_k, \varepsilon_{x_k}) = \beta_k - \beta_k \theta Var(\varepsilon_{x_k}) \tag{13}$$

which tells us that if the original coefficient is positive (negative), the estimated coefficient will tend to be less positive (negative). In the particular case where the $x$'s are uncorrelated, we will have

$$plim\, b_k = \beta_k - \beta_k \frac{Var(\varepsilon_{x_k})}{Var(x_k)} = \beta_k - \beta_k \frac{Var(\varepsilon_{x_k})}{Var(x_k^*) + Var(\varepsilon_{x_k})} \tag{14}$$

Thus if $\beta_k > 0$, the (probability limit of the) coefficient still remains positive but moves closer towards 0, and if $\beta_k < 0$, it still remains negative and the 'plim' moves closer towards 0. This phenomenon is known as 'attenuation bias' for the errors in variable model.

## 2.3 Simultaneous Equations

Suppose, we have a supply demand system as follows:

$$q_t = a + bp_t + cw_t + u_t \quad \text{(supply)} \tag{15}$$

$$q_t = e - fp_t + gi_t + v_t \quad \text{(demand)} \tag{16}$$

Here $q_t$ is quantity, $p_t$ is price, $w_t$ is a weather-related variable, $i_t$ is income, and $u_t, v_t$ are error terms. The data-generating process should be thought of as follows: the error terms, the weather variable and the income variable are all separately and independently generated (it is possible though to allow the supply error term and the demand error term to be correlated; nothing that follows depends on that correlation, if any). Once these are determined, price and quantity are simultaneously determined by solving equations 15 and 16. Geometrically speaking, $a$, the weather variable and $u_t$ determine a random intercept for the supply curve, while $b$, the income variable and $v_t$ determine a random intercept of the demand curve, and each period the intersections of the two random curves determine the actual realization of price and quantity. If the terms that affected the randomness of the supply equation were actually fairly stable, and demand-related factors were to fluctuate a lot, the actual price-quantity scatter will have a positive correlation.

Could we then consistently estimate the supply equation by regressing $q$ on $p$, $w$ and a constant? From the previous discussion, the answer should be easily seen to be negative. Technically, this can be seen by solving the above two equations for $p_t$ which gives

$$p_t = \frac{e-a}{b+f} - \frac{c}{b+f} w_t + \frac{g}{b+f} i_t + \frac{v_t - u_t}{b+f} \tag{17}$$

Clearly $p_t$ in the supply equation is correlated with the equation's error term and hence the supply equation cannot be consistently estimated via OLS.

Interestingly, however, the price equation above is a perfectly legitimate equation one can run a regression on. One can then obtain consisitent estimates of such quantities as: $\frac{e-a}{b+f}$, $\frac{c}{b+f}$, $\frac{g}{b+f}$. Similarly, one can solve for $q_t$ and obtain the equation

$$q_t = \frac{af+be}{b+f} + \frac{fc}{b+f}w_t + \frac{bg}{b+f}i_t + \frac{bv_t+fu_t}{b+f} \tag{18}$$

This equation can now be estimated via OLS which delivers consistent estimates for $\frac{af+be}{b+f}$, $\frac{fc}{b+f}$, $\frac{bg}{b+f}$. With the help of the six estimates obtained, from estimating the last two equations, parameters $a, b, c, e, f$ and $g$ can all be consistently estimated. This is however *not* the instrumental variable method that we are to describe in the next section, although the two methods will both provide consistent estimates.

Incidentally, equations 15 and 16 in these context are what are called 'structural equations' while equations 17 and 18 are what are called reduced form equations. Estimating parameters of the structural equations by first estimating the reduced form equations is what is known as the Indirect Least Squares method of estmating simultaneous equations.

## 2.4 Dynamic Models with Autocorrelated Errors

Suppose, we wish to estimate the expectation-augmented Phillips curve equation:

$$\dot{p}_t = \dot{p}_t^e + \mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_t \tag{19}$$

Here, $\dot{p}_t$ is inflation, $\dot{p}_t^e$ is expected inflation, $\mathbf{x}_t$ is a vector of other variables, $\boldsymbol{\beta}$ is a parameter vector and $\varepsilon_t$ is an error term. Now also assume that we have AR(1) errors, i.e.

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t \tag{20}$$

where $u_t$ is a sequence of i.i.d. zero mean random variables. In this model, $\dot{p}_t^e$ is often proxied by $\dot{p}_{t-1}$, and hence, one can estimate the model

$$(\dot{p}_t - \dot{p}_{t-1}) = \mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_t \tag{21}$$

and get consistent estimates for the betas.

However, if one assumes an adaptive expectations framework, where

$$\dot{p}_t^e = \dot{p}_{t-1} + \alpha(\dot{p}_{t-1} - \dot{p}_{t-2}) \tag{22}$$

then, one can rewrite the Phillips curve equation as

$$y_t = \alpha y_{t-1} + \mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_t \tag{23}$$

where $y_t = \dot{p}_t - \dot{p}_{t-1}$. But now we are in trouble because the error term is correlated with one of the regressors (do you see why?) - resulting in an endogeneity problem. Generally speaking, its ok to have as regressors lagged values of the left hand side variable, but not if the error term is autocorrelated.

### Section End Questions

1. In Section 2.1, I assert that the 'best' projection (or forecast) of $q$ we can have using a linear function of the $x$s is given by $\delta_0 + \mathbf{x}'\boldsymbol{\delta}$ where $\boldsymbol{\delta} = [Var(\mathbf{x})]^{-1}Cov(\mathbf{x}, q)$ and $\delta_0 = Eq - E(\mathbf{x})'\boldsymbol{\delta}$. Verify this by setting up a minimization problem and doing some matrix differentiation (be careful and creative about the order in which certain vectors and matrices are multiplied). Also, show that as claimed, that the projection/forecast error has zero mean and has zero covariance with the $x$'s.

2. In section 2.2, if we have measurement errors for both $y$ and $x_k$, does the expression for the omitted variable bias change?

3. In section 2.3, suppose weather does not affect supply, i.e. $c = 0$ (we could be talking about a manufacturing good). Will we be able to consistently estimate all other parameters via the Indirect Least Squares method?

# 3 IV Estimation

## 3.1 What is an instrumental variable?

So, suppose we do have endogeneity in our model; there are 'bad regressors' in our equation which are correlated with the error term along with some good regressors as well which are not.[1] What are we to do? What we do is we look for certain extra variables called 'instruments' (at least one for each bad regressor). An instrumental variable is one that is

a) uncorrelated with the error term and

b) correlated with the bad regressor in question (after netting out the effect of the good regressors).

[1] What we have been calling 'bad regressors' econometricians often refer to as endogenous variables. Similarly our euphemism for 'good regressors' is often described as exogenous covariates. Note that there is a slight problem with this terminology. The term endogenous is also commonly used to refer to variables determined inside the model and the word exogenous often refers to variables that are determined outside the model. However, a model may contain many equations and it is perfectly possible that a variable on the right hand side of an equation is exogenous in the first sense but endogenous in the second sense.

The qualifier in requirement b) is often mistakenly skipped in the description of instruments. What I mean by that qualifier is that it should not be the case that the instrument should have some explanatory power over the bad regressor solely because a good regressor affects both the instrument and the bad regressor in question - it should have independent explnatory power over the bad regressor. It is not too hard to check this requirement: if you suspect that you have a bad regressor in your model and if you also believe that you have a good instrument for that bad regressor, simply regress the latter on your instrument and the other good regressors and check whether you have a significant coefficient on the instrument.

We now explore how one might be go about looking for valid instruments by looking at some examples.

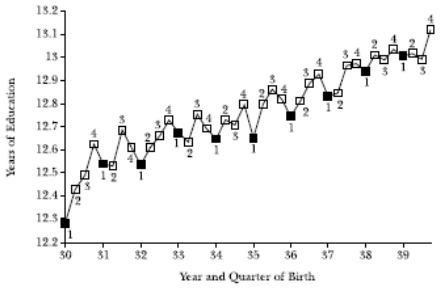
Consider, for example the endogeneity caused by simultaneity. Suppose, we are trying to estimate the demand equation as described by equation 16. We pointed out earlier that we cannot consistently estimate the parameters of that equation because the price variable is a bad regressor; it is correlated with the error term on the demand equation ($v_t$) since $p_t$ is endogenously obtained from the two-equation model in terms of $v_t$ (see equation 17). So, now we need to be looking for an extra variable not already featured in the demand equation that is uncorrelated with the error term $v_t$ but is correlated with $p_t$ (you can't use $i_t$ as it is already there in the equation; nor can you choose something that influences price, solely because it is affected by income. Other demand-related variables, such as proxies for preference patterns will not do either as they are part of the error term). One such candidate variable is the weather variable $w_t$. As equation 17 reveals, $p_t$ is influenced by $w_t$ (over and above being influenced by $i_t$), and also there should be no presumption that $w_t$ is correlated with $v_t$ (why should weather have anything to do with preference or demand shocks?). Thus the weather variable is a valid instrument in this context.

Now consider the problem created by omission of variables. Going back to the example of wage estimation where the right hand side variable educational achievement may be a bad regressor because it is correlated with the error term which may be capturing innate ability, we need to ask ourselves: can we come up with a new variable which is correlated with educational attainment but not with innate ability? One answer could be cost of higher education relative to family income at the person's place of origin and during the person's post high-school days. Such a variable will clearly be correlated with the amount of education the person obtains, but it is unlikely that it will be correlated with innate ability. A simpler candidate instrument is mother's educational achievement, but although well-educated parents tend to educate their offsprings more, thus giving credence to condition b), it is unclear that condition a) is satisfied. It all depends on whether you believe that well-educated mothers tend to have offsprings with higher innate ability.

Angrist and Krueger in their survey article on IV technique, quotes Maddala, a respectable econometrician and texbook author: "Where do you get such a variable?", and then goes on to comment, "Like most econometrics texts, he does not provide an answer. In our view, good

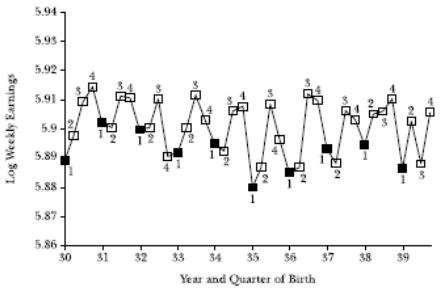
instruments often come from detailed knowledge of the economic mechanism and institutions determining the regressor of interest." To underline the importance of knowledge of economic institutions or natural experiments, consider the 1991 well-known study by the two authors where in the context of measuring returns to education, they use a novel instrument: the individual's quarter of birth. Why may this be a good instrument (at least in theory)? Because due to compulsory schooling laws, children enter school the year they turn six, and they are required to remain in school till their sixteenth birthday. This allows children born in the first quarter of the year to 'get away' with less education than those born in say the third and fourth quarter. Consequently, the birth-quarter dummies are correlated with years of schooling, but clearly they have nothing to do with the omitted ability variable. The following diagrams excerpted from an expository article in the Journal of Economic Perspectives show that there indeed is a strong relation between earnings, birth quarter and educational achievement vindicating their choice of instrument.

*Figure 1*
**Mean Years of Completed Education, by Quarter of Birth**




*Source:* Authors' calculations from the 1980 Census.

*Figure 2*
**Mean Log Weekly Earnings, by Quarter of Birth**




*Source:* Authors' calculations from the 1980 Census.

Relation between a) birth quarter and education and b) birth quarter and wages

## Section End Questions

1. Name two instrumental variables for estimating the supply equation.

2. It may be argued that wage depends on experience as well as innate ability and education. Could this create a problem for using the birth quarter dummies as instrumental variables?

### 3.2 Consistent Estimation via IVs

Let the model that we are interested in estimating be $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$ where $\mathbf{x}_i$ is a k-dimensional vector of regressors. Suppose, we have another k-dimensional random vector $\mathbf{z}_i$ of instruments (Note that it is possible to have some common variables among $x_i$ and $z_i$, namely, the good regressors). The following assumptions will be made in what follows:

A8: $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$ and $E(\mathbf{z}_i\varepsilon_i) = \mathbf{0}$

A9: $E(\mathbf{z}_i\mathbf{x}_i') = \boldsymbol{\Sigma}_{zx}$ is finite and invertible which, also implies $E(\mathbf{x}_i\mathbf{z}_i') = \boldsymbol{\Sigma}_{xz}$ is also finite and invertible.

A10: $E({\varepsilon_i}^2\mathbf{z}_i\mathbf{z}_i') = \boldsymbol{\Sigma}_{z\varepsilon}$ exists finitely.

A11: $E(\mathbf{z}_i\mathbf{z}_i') = \boldsymbol{\Sigma}_z$ is finite and invertible.

Now define the instrumental variable estimator $\mathbf{b}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{y})$. Then, we can claim:

**Theorem 1** *Under A3, A8, A9* $\mathbf{b}_{IV} \underset{p}{\longrightarrow} \boldsymbol{\beta}$.

**Proof:** $\mathbf{b}_{IV} = \boldsymbol{\beta} + (\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\varepsilon) = \boldsymbol{\beta} + (\frac{\mathbf{Z}'\mathbf{X}}{n})^{-1}(\frac{\mathbf{Z}'\varepsilon}{n})$

But A8 implies $plim\ \frac{\mathbf{Z}'\varepsilon}{n} = \mathbf{0}$, and by A9, $plim\ \frac{\mathbf{Z}'\mathbf{X}}{n} = \boldsymbol{\Sigma}_{zx}$ is finite and invertible so $plim\ (\frac{\mathbf{Z}'\mathbf{X}}{n})^{-1} = \boldsymbol{\Sigma}_{zx}^{-1}$ exists finitely, and so we have the result.

Let $Var(\varepsilon_i) = \sigma^2$. Then, a consistent estimate of this parameter can be found in the usual way...

**Theorem 2** *Under A3, A6, A8, A9 and the further assumption* $E(\mathbf{x_i}\,\varepsilon_i)$ *exists finitely,*

$$\hat{\sigma_n}^2 = \frac{(\mathbf{y} - \mathbf{X}\mathbf{b_{IV}})'(\mathbf{y} - \mathbf{X}\mathbf{b_{IV}})}{n} \underset{p}{\longrightarrow} \sigma^2$$

**Proof:** Omitted. Proceeds exactly along the line of argument presented in the proof of Theorem 2 in Topic 5 .

**Theorem 3** *Under A3, A8, A9, A10*

$$\sqrt{n}(\mathbf{b}_{IV} - \boldsymbol{\beta}) \underset{d}{\longrightarrow} N(0, (\boldsymbol{\Sigma}_{zx})^{-1}\boldsymbol{\Sigma}_{z\varepsilon}(\boldsymbol{\Sigma}_{xz})^{-1})$$

*Furthur, if we assume* $E({\varepsilon_i}^2|\mathbf{z}_i) = \sigma^2$*, we have:*

$$\sqrt{n}(\mathbf{b}_{IV} - \boldsymbol{\beta}) \underset{d}{\longrightarrow} N(0, \sigma^2(\boldsymbol{\Sigma}_{zx})^{-1}\boldsymbol{\Sigma}_{z}(\boldsymbol{\Sigma}_{xz})^{-1})$$

**Proof:** Omitted. See the proof of Theorem 3 in Topic 5 .

It then makes sense to estimate the variance of $\mathbf{b}_{IV}$ (not $\sqrt{n}\,\mathbf{b}_{IV}$ !) by $\frac{\hat{\sigma}^2}{n}(\frac{\mathbf{Z'X}}{n})^{-1}(\frac{\mathbf{Z'Z}}{n})(\frac{\mathbf{Z'X}}{n})^{-1}$ or, $\hat{\sigma}^2(\mathbf{Z'X})^{-\mathbf{1}}(\mathbf{Z'Z})(\mathbf{Z'X})^{-\mathbf{1}}$ and with the help of this, standard errors may be calculated.

You might be wondering at this stage, why the qualifier in requirement b) that was stated for instrumental variables features in all this. It is needed to ensure that A9 is valid (A8 is guaranteed by the requirement a) ). See if you can prove this assuming A11, i.e. there is no perfect multicollinearity among the variables in the $\mathbf{z}$ vector.

### 3.3 2SLS

Recall that we said that the vector of instruments $\mathbf{z}$ must be k-dimensional. What should we do if we have more instruments than there are regressors? The problem is that $\mathbf{b_{IV}} = (\mathbf{Z'X})^{-\mathbf{1}}(\mathbf{Z'Y})$ is not well-defined if $\mathbf{Z}$ is not $n \times k$ (so $\mathbf{Z'X}$ is not a square matrix).

Should we throw away the extra ones? If so, which ones? Generally, speaking, throwing away information is always a bad idea. What we will do is to 'combine' them in a particular way, so that we can create $k$ new variables. [2] So suppose, we have a matrix of observations on $l$ variables: $\mathbf{w}$. We assume the following:

A12: $E(\mathbf{w}_i\varepsilon_i) = \mathbf{0}$

A13: $E(\mathbf{w}_i\mathbf{x}_i') = \boldsymbol{\Sigma}_{wx}$ is finite with full column rank.

A14: $E(\mathbf{w}_i\mathbf{w}_i') = \boldsymbol{\Sigma}_{w}$ is finite and invertible.

Assuming that we have $n$ observations on the $w$ variables stacked in an $n \times l$ matrix $\mathbf{W}$, now let us define a $n \times k$ matrix $\mathbf{Z}$ as follows: $\mathbf{Z} = \mathbf{W}(\mathbf{W'W})^{-\mathbf{1}}\mathbf{W'X}$. Finally, let us define an

[2] P. G. Wright in 1928 used various instruments to estimate supply and demand elasticities of flaxseed, probably contributing to the first known application of IV technique. Although a pioneer, he did not know better than averaging the various estimates he obtained using the various instruments (one at a time). Isn't intellectual progress fascinating?

estimator as $\mathbf{b_{2SLS}} = (\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{Y})$. This is Henri Theil's 2SLS (sometimes written as TSLS) estimator.

Now, notice that $\mathbf{Z}$ could be written as $\mathbf{P_W X}$, where $\mathbf{P_W}$ is the projection matrix corresponding to the $\mathbf{W}$ space. In other words, $\mathbf{Z}$ simply contains the predicted values of regressing the variables in $\mathbf{X}$ on the variables that are in $\mathbf{W}$.

Notice one other thing. $\mathbf{Z}'\mathbf{X} = \mathbf{Z}'\mathbf{Z}$ (why?). Hence, we can write

$$\mathbf{b_{2SLS}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{Y} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y} \tag{24}$$

We now have a very pretty interpretation of this estimator which justifies its name:

1. In stage 1, regress the $\mathbf{x}$ variables on $\mathbf{w}$; obtain the predicted values, and stack them in $\mathbf{z}$.

2. In stage 2, regress $y$ on these new variables using OLS.

This also shows that an alternate expression for $\mathbf{b_{2SLS}}$ is $(\mathbf{X}'\mathbf{P_W}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{P_W}\mathbf{y})$.

In practice, when one estimates a single equation by 2SLS, the structure of the problem is typically as follows: $\mathbf{y} = \mathbf{X_1}\boldsymbol{\beta_1} + \mathbf{X_2}\boldsymbol{\beta_2} + \boldsymbol{\varepsilon}$ where $\mathbf{X_1}$ has the good regressors and $\mathbf{X_2}$ the bad ones. To take care of the potential inconsistency, we need another set of regressors which are in $\mathbf{X_3}$ (say). All the variables here satisfy the definition of instruments. We then regress the $[\mathbf{X_1}|\mathbf{X_2}]$ variables on $\mathbf{W} = [\mathbf{X_1}|\mathbf{X_3}]$ and store the fitted values in $\mathbf{Z}$ (obviously the good regressor observations are reproduced as is). And then in the second stage $\mathbf{y}$ is regressed on the variables in $\mathbf{Z}$.

We now show that $\mathbf{b_{2SLS}}$ is consistent for $\boldsymbol{\beta}$.

**Theorem 4** *For the model $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$, under A12 - A14, $\mathbf{b_{2SLS}} \xrightarrow[p]{} \boldsymbol{\beta}$.*

**Proof:** Following standard arguments, we need to show $plim\,(\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\boldsymbol{\varepsilon}) = \mathbf{0}$, which can be achieved by showing $plim\,\frac{(\mathbf{Z}'\mathbf{X})}{n}$ is finite and invertible, and $plim\,\frac{(\mathbf{Z}'\boldsymbol{\varepsilon})}{n}$ is $\mathbf{0}$. Since, $\mathbf{Z}'\boldsymbol{\varepsilon} = \mathbf{X}'\mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\boldsymbol{\varepsilon}$, by invoking assumptions A12-A14, we see that the latter is guaranteed. To see the former, note that $plim\,\frac{(\mathbf{Z}'\mathbf{X})}{n} = plim\,\frac{(\mathbf{X}'\mathbf{W})(\mathbf{W}'\mathbf{W})^{-1}(\mathbf{W}'\mathbf{X})}{n} = plim\,\frac{(\mathbf{X}'\mathbf{W})}{n}\,(plim\,\frac{(\mathbf{W}'\mathbf{W})}{n})^{-1}\,plim\,\frac{(\mathbf{W}'\mathbf{X})}{n}$ which by A13 and A14 is finite and invertible.

It is also pretty straightforward to demonstrate the asymptotic normality of the estimator under the additional assumption

A15: The random vector $\mathbf{w}_i\varepsilon_i$ has a finite variance matrix, say $\boldsymbol{\Sigma}_{w\varepsilon}$.

**Theorem 5** *For the model* $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$*, under A12 - A15,*

$$\sqrt{n}(\mathbf{b_{2SLS}} - \boldsymbol{\beta}) \underset{d}{\longrightarrow} \mathbf{N}(\mathbf{0}, (\mathbf{\Sigma_{xw}\Sigma_w}^{-1}\mathbf{\Sigma_{wx}})^{-1}(\mathbf{\Sigma_{xw}\Sigma_w}^{-1}\mathbf{\Sigma_{w\varepsilon}\Sigma_w}^{-1}\mathbf{\Sigma_{wx}})(\mathbf{\Sigma_{xw}\Sigma_w}^{-1}\mathbf{\Sigma_{wx}})^{-1})$$

**Proof:** Omitted.

To estimate the nasty-looking asymptotic variance consistently, we simply use the sample average counterparts of each of the $\Sigma$'s. In the special case of conditional homoskedasticity (which in this context means $E(\varepsilon_i|\mathbf{w}_i) = \sigma^2$, the expression for the asymptotic variance of the estimator reduces to $\sigma^2(\mathbf{\Sigma_{xw}\Sigma_w}^{-1}\mathbf{\Sigma_{wx}})^{-1}$ as can be easily verified.

## Section End Questions

1. Consider the simple bivariate model $y_i = \alpha + \beta x_i + \varepsilon_i$ where $x$ is a bad regressor and $\varepsilon_i$ is a zero mean error term. Suppose, $z_i$ is some random variable such that $cov(x, z)$ is non-zero and $cov(z, \varepsilon)$ is zero. Give a direct argument that the sample covariance between $y$ and $z$ divided by the sample covariance between $x$ and $z$ is a consistent estimate of $\beta$.
2. Now show that the estimator above is indeed the 2SLS estimator. Some simple matrix calculations are involved.
3. Work out the omitted proofs of Theorems 2, 3 and 5 in this section.

# 4 Practical Aspects of IV Estimation

The field of instrumental variable estimation throws up an interesting case study of comparison of estimators and the issue of tradeoff between bias, inconsistency and variance. IV estimators, even when they are consistent (under appropriate assumptions), have a huge problem: they are overdispersed. In fact, if $\mathbf{W}$ has $l$ variables and $\mathbf{X}$ has $k$ variables, then, in finite samples, the IV estimator (which from now on I will use interchangeably with the 2SLS estimator), has finite moments of order $m$ only if $m \leq l-k$. Incidentally, the quantity $l-k$ is called the degree of overidentification, and where this is zero, we say that the model is exactly identified.

From what has been said it follows that if you have exactly as many instruments as there are original variables in the model, the estimator has no expectation in finite samples! This may surprise you given that the estimator is consistent and has a well-defined asymptotic variance, but it is perfectly possible for to have a sequence of random variables, none of which have well-defined expectation, to converge (in probability) to a fixed object. Now, if a random variable's $m$th moment does not exist, neither does its higher moments, so you can imagine how 'overdispersed' your estimator will be.

This might send you scurrying towards finding lots of instruments. In fact, as has been alluded earlier, the asymptotic variance of an estimator with a larger set of instruments for the endogenous variables will always be smaller than that one with a smaller set. But the problem is in finite samples, as $l$ becomes large, the bias of the estimator typically increases and by the time $l$ reaches $n$, the IV estimator coincides with the OLS estimator.

Finally, if endogenous regressors correlate weakly with some of the instruments, we get again, poor finite sample properties for the estimators, a problem known in the literature, as the 'Weak Instrument' problem. Research by Staiger and Stock suggest that in the stage 1 regression if the F statistic for the joint test of null coefficients on the 'outside instruments' is less than 10, one might have a problem. An F-statistic of less than 5 is typically indicative of extreme finite sample bias.

### Section End Questions

1. Construct explicitly a sequence of random variables, none of which have finite mean and yet, they converge in probability to a fixed scalar.
2. Why does the 2SLS estimator become the OLS estimator when $l$ reaches $n$?

# 5 Hypothesis Testing

Finally, I will briefly mention here how you can test three types of hypotheses that appear in the IV context.

A. GENERAL HYPOTHESIS TESTING (of the form $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$).

These tests *can not* be conducted by imposing restrictions in the second stage of the two stage least squares procedure. The difficulty arises with the fact that the RSS obtained from the second stage (divided by $n$ or $n-k$) is not a consistent estimator of the variance of the disturbance term. For such a test, one needs to compute 3 RSS's. Let $\mathbf{b_{2sls}}$ be the 2SLS estimator.

$RSS_1$ is obtained as $\mathbf{e}'\mathbf{e}$ where $\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b_{2sls}}$.

$RSS_2$ is obtained by regressing $\mathbf{y}$ on $\mathbf{Z}$ variables with the restrictions.

$RSS_3$ is obtained by regressing $\mathbf{y}$ on $\mathbf{Z}$ variables without the restrictions.

The test statistic then is $\frac{(RSS_2 - RSS_3)/q}{(RSS_1)/(n-k)}$ and it is supposed to be distributed as $F_{q,n-k}$ under the null hypothesis.

B. TEST OF EXOGENEITY (WU-HAUSMAN TEST)

If you suspect that some of your regressors may be suffering from endogeneity and you think you have a good set of instruments for them, then you can actually test whether your suspicion is valid. For this purpose, a test statistic was proposed by Hausman in 1978 which tests whether the difference between $\boldsymbol{\theta} = (\mathbf{b_{IV}} - \mathbf{b_{OLS}})$ (which is sometime called the 'contrast' vector) is 'small' which is supposed to happen since, under the null hypothesis of no endogeneity, both estimators are consistent. In fact $\sqrt{n}\,\boldsymbol{\theta}$ has asymptotically normal distribution. This suggests that a test statistic of the form $n\boldsymbol{\theta}'(Avar(\boldsymbol{\theta}))^{-1}\boldsymbol{\theta}$ which will be distributed as a chi-squared variate. The implementability of the test is complicated by the need to calculate $Avar(\boldsymbol{\theta})$. Exploiting the efficiency of the OLS estimator, under the null (this is essentially the Gauss-Markov theorem), Hausmann shows that $Cov(\mathbf{b_{OLS}}, \boldsymbol{\theta}) = \mathbf{0}$ when one can write $Cov(\boldsymbol{\theta}) = Var(\mathbf{b}_{IV}) - Var(\mathbf{b_{OLS}})$. This reduces the Hausman statistic to

$$n(\mathbf{b_{IV}} - \mathbf{b_{OLS}})'[\mathbf{A\hat{var}}(\mathbf{b_{IV}}) - \mathbf{A\hat{var}}(\mathbf{b_{OLS}})]^{-1}((\mathbf{b_{IV}} - \mathbf{b_{OLS}})$$

A problem is $\boldsymbol{\theta}$ will typically not have a nonsingular variance matrix. As a consequence, one needs to use a generalized inverse rather than simple inverse. A second consequence of this is the test statistic is supposed to be a $\chi^2$ variate not with $k$ degrees of freedom, but with $k_2$ degrees of freedom under the null (where $k_2$ is the number of suspected bad regressors).

The Hausman test is a very general specification test. The setup needs two estimators. Under the null both are consistent but one is efficient. Under the alternative, their probability limits are different. This allows econometricians to specify both hypothesis in very general terms without specifying the exact reason for the posited behaviors of the estimators. Consequently the test is used in many different contexts, not just regarding the issue of endogeneity of regressors.

Despite its wide useage, the test is problemmatic. First, as we have mentioned the sandwiched matrix may not be invertible, and also, the test statistic often turns out to be negative which is strange for something that is supposed to be distributed as a chi-squared variate.

One solution is to carry out an asymptotically equivalent test suggested by Wu (1973). First we regress the suspected bad regressors on the instruments and create residuals. Next we regress $y$ on all the original $x$'s (both good and bad) plus these residuals. The test is simply an F-test for the coefficients on the residuals to be zero.

C. TEST OF OVERIDENTIFICATION

If we have strictly more instruments than regressors (which is the so called overidentification case), we can test the (joint) validity of the instruments via a simple F test or a chi-squared test. First we obtain the residuals $\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}_{\mathbf{2SLS}}$. We then regress these residuals on the instruments (the $W$'s). If the null hypothesis about the exogeneity of the instruments is valid then this regression will have little explanatory power. A simple way to create a test statistic was proposed by Sargan, which involves calculating $n$ times the uncentered $R^2$ from the second stage regression (in case the original model and the set of instruments included a constant, one can use the usual $R^2$ as well). Under the null, this statistic has an asymptotic chi-squared distribution with $l - k$ degrees of freedom, where $l$ is the number of instruments and $k$ the number of regressors.

## Section End Questions

1. Provide a justification to the Wu test of exogeneity using large sample theory.
2. Provide a justification for Sargan's test of overidentification using large sample theory.