MODULE TWO, PART ONE:
ISSUES OF ENDOGENEITY AND INSTRUMENTAL VARIABLES IN ECONOMIC EDUCATION RESEARCH

William E. Becker
Professor of Economics, Indiana University, Bloomington, Indiana, USA
Adjunct Professor of Commerce, University of South Australia, Adelaide, Australia
Research Fellow, Institute for the Study of Labor (IZA), Bonn, Germany
Editor, *Journal of Economic Education*
Editor, *Social Scinece Reseach Network: Economic Research Network Educator*

Federal Reserve Chair Ben Bernanke said that being an economist is like being a mechanic working on an engine while it is running. Economists typically do not have the convenience of random assignment as in laboratory experiments. However, in some situations they can take advantage of random events such as lotteries or nature. In other circumstances, they might be able to produce variables that have desired random components. When this is possible, they can use instrumental variable techniques and two-stage least squares estimation, which is the focus of the Module Two. Part One of Module Two is devoted to the general theoretical issues associated with endogeneity. Module Two, Parts Two, Three and Four provide the methods of instrumental variable estimation using LIMDEP (NLOGIT), STATA and SAS. To get started consider three types of problems for which instrumental variables are employed.[1]

The first problem of concern is omitted variables. When presenting regression results someone invariably proposes that an explanatory variable that is alleged to be relevant but was omitted is correlated with the included regressors. This renders the coefficient estimators of the included but correlated regressors biased and inconsistent. As stated in the introduction to these modules, examples of this can be traced back over one hundred years to a debate between statistician George Yule and economist Arthur Pigou, see Stephen Stigler (1986, pp. 356-357). Recall that Pigou criticized Yule's multiple regression (aimed at explaining the percentage of persons in poverty with the change in the percentage of disabled relief recipients, the percentage change in the proportion of old people, and the percentage change in the population) because it omitted the most important influences: superior program management and restrictive practices, which cannot be measured quantitatively.

Pigou identified the most enduring criticism of regression analysis; namely, the possibility that an unmeasured but relevant variable has been omitted from the regression and that it is this variable that is giving the appearance of a causal relationship between the dependent variable and the included regressors. As described by Michael Finkelstein and Bruce Levin (1990, pp. 363-364 and pp. 409-415), for example, defense attorneys continue to argue that the plaintiff's experts omitted relevant market and productivity variables when they use regression analysis to demonstrate that women are paid less than men. Modern academic journals are packed with articles that argue for one specification of a regression equation versus another for everything from the demand for places in higher education to the learning of economics in the introductory courses.

The second problem is errors in variables. The late Milton Friedman was awarded the Nobel prize in Economics in part because of his path-breaking work in estimating the relationship between consumption and permanent income, which is an unobservable quantity. His work was later applied in unrelated areas such as education research where a student's grade is hypothesized to be a function of his or her effort and ability, which are both unobservable. As we will see, unobserved explanatory variables for which index variables are created give rise to errors-in-variables problems. As seen in the early work of Becker and Salemi (1979), an outstanding example of this in economic education research occurs when the pretest is used as a proxy for existing knowledge, ability or prior understanding.

The third problem is simultaneity. At the aggregate level, estimating a Keynesian consumption function (in which consumption is a function of income) has problems caused by a second equation involving an accounting identity in which aggregate income must equal personal consumption plus other forms of aggregate expenditures. That is, for the nation as a whole there is a simultaneous relationship between income and consumption: consumption is a function of income and income is a function of consumption. Harvard/Stanford University researcher Caroline Hoxby (2000) identified a similar reverse causality problem in her study of the effect of competition among school districts on student performance, as reported in the *Wall Street Journal* (Oct 24, 2005). She hypothesized that more school districts in a community implied more competition and better schools. She also recognized, however, that there could be reverse causality in that a poor school district that could not be closed (because of state regulations, for example) would force politicians (through parental pressure) to start another school district. In economic education, Becker and Johnston (1999) identified a simultaneity problem in trying to explain scores on one type of test (say multiple choice) with scores on another (essay or free response), where causality is bidirectional. Students who score high on either are likely to score high on the other. As we will see, these are problems of simultaneity that involve endogenous regressors.

Omitted variables that are correlated with included explanatory variables, simultaneity and errors in variables are all examples of endogeneity problems for which single equation estimation is not sufficient.

PROBLEMS OF ENDOGENEITY

Put simply, the problem of endogeneity occurs when an explanatory variable is related to the error term in the population model of the data generating process, which causes the ordinary least squares estimators of the relevant model parameters to be biased and inconsistent. More precisely, for the least squared $\mathbf{b}$ vector to be a consistent estimator of the $\boldsymbol{\beta}$ vector in the population data generating model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the $\mathbf{X'X}$ matrix must be a positive definite matrix (defined by $\mathbf{Q}$, as the sample size n goes to infinity) and there can be no relationship between the vector of population error terms ($\boldsymbol{\varepsilon}$) and the regressors (explanatory variables) in **X**. Mathematically, if

$$\lim_{n\to\infty}\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)=\mathbf{Q} \text{ (a positive definite matrix) and } p\lim\left(\frac{1}{n}\mathbf{X}'\boldsymbol{\varepsilon}\right)=0,$$

then

$$p\lim\mathbf{b}=\boldsymbol{\beta}+\left(\mathbf{Q}\right)^{-1}p\lim\left(\frac{1}{n}\mathbf{X}'\boldsymbol{\varepsilon}\right)=\boldsymbol{\beta}$$

In words, if observations on the explanatory variables (the *X*s) are unrelated to draws from the error terms (in vector $\boldsymbol{\varepsilon}$), then the sampling distribution of each of the coefficients (the *b*s in **b**) will appear to degenerate to a spike on the relevant Beta, as the sample size increases. In probability limit, a *b* is equal to its β : $p\lim b=\beta$.

But if there is strong correlation between the *X*s and εs, and this correlation does not deteriorate as the sample size goes to infinity, then the least squares estimators are not consistent estimator of Betas and $p\lim\mathbf{b}\neq\boldsymbol{\beta}$. The **b** vector is an inconsistent estimator because of endogenous regressors. That is, the sampling distribution of at least one of the coefficient (one of the *b*s in **b**) will not degenerate to a spike on the relevant Beta, as the sample size continues to increase.

OMITTED VARIABLE

If someone asserts that a regression has omitted variable bias, he or she is saying that the population disturbance is related to an included regressor because a relevant explanatory variable is missing in the estimated regression and its effects must be in the disturbance. This is also known as **unobserved heterogeneity** because the effect of the omitted variable also leads to population error term heterogeneity. The straightforward solution is to include that omitted variable as a regressor, but often data on the missing variable are unavailable. For example, as described in Becker (2004), the U.S. Congressional Advisory Committee on Student Financial Assistance is interested in the functional relationship between the effects of financial variables (e.g., family income, loan availability, and/or grants) and the college-going decision, called persistence and measured by the probability of attending a post-secondary institution, number of post-secondary terms attempted and the like, in linear form:

$$Persistence = f\,(finances, random\ perturbation)$$

The U.S. Department of Education is concerned about getting students “college ready,” as measured by an index reflecting the completion of high school college prep courses, high school grades, SAT scores and the like:

$$Persistence = h(college\,ready, random\,error)$$

Putting the two interests together, where epsilon is the disturbance term, suggests that the appropriate linear model is

$$Persistence = \beta_1 + \beta_2(college\,ready) + \beta_3(finances) + \varepsilon$$

Information on college readiness is obtainable from Department of Education records but matching financial information is more difficult to obtain; thus, a researcher might consider estimating the parameters in

$$Persistence = \lambda_1 + \lambda_2(college\,ready) + u$$

Finances are now in the error term *u*. But students from wealthier families are known to be more college ready than those from less well-off families. Thus, the explanatory variable *college ready* is related to the error term *u*. If estimation is by OLS, bias and inconsistent estimation of λ_2 result:

$$\begin{aligned} \mathrm{E}[(college\,ready)u] &= \beta_3\mathrm{E}[(college\,ready)(finances)] + \mathrm{E}[(college\,ready)\varepsilon] \\ &= \beta_3\mathrm{E}[(college\,ready)(finances)] = \beta_3\mathrm{cov}[(college\,ready)(finances)] \neq 0 \end{aligned}$$

SIMULTANEITY

A classic case of simultaneity can be found in the most basic idea from microeconomics: that the competitive market of supply and demand determines the equilibrium quantity. The market data generating process is thus written as a three equation system:

Supply: $Qs = m + nP + U$

Demand: $Qd = a + bP + cZ + V$

Equilibrium $Q = Qd = Qs$

where *m, n, a, b* and *c* are parameters to be estimated. *P* is price. *Qd* and *Qs* are quantities demanded and supplied, which in equilibrium are equal to *Q*. *Z* is an exogenous variable and *U* and *V* are errors such that

$$\begin{aligned} &E(U) = E(V) = 0,\ E(UV) = 0, \\ &E(U^2) = \sigma_u^2,\ E(Y^2) = \sigma_v^2, \text{ and} \\ &E(VZ) = E(UZ) = 0. \end{aligned}$$

Suppose the supply curve now is to be estimated by OLS from observable market data for which it must be the case that quantity demand equals quantity supplied in equilibrium:

$$Q = m + nP + U.$$

The estimation slope coefficients in the supply equation would obtained as

$$\begin{aligned}\hat{n} &= (P'P)^{-1}P'Q \\ &= n + (P'P)^{-1}P'U.\end{aligned}$$

But from the market structure assumed to be generating the data we know

$$P = \frac{a-m}{n-b} + \frac{c}{n-b}Z + \frac{v-u}{n-b} = \beta_0 + \beta_1 Z + \varepsilon_2 .$$

Thus, $\mathrm{E}(P'U) \neq 0$

$$E(PU) = E[(\frac{a-m}{n-b})U + (\frac{c}{n-b}Z)U + (\frac{V-U}{n-b})U] = E(-\frac{U^2}{n-b}) = -\frac{\sigma_u^2}{n-b} \ .$$

The OLS estimator ($\hat{n}$) is downward biased; that is, the true population parameter is expected to be underestimated by the least squares estimator:

$$E(\hat{n}) = n - \frac{\sigma_u^2}{n-b} .$$

Next consider an example from macroeconomics in which an aggregate Keynesian consumption function is to be estimated.

$$C = A + BX + U$$

where C is consumption (realized and planned consumption are equal in equilibrium), X is current income and U is the disturbance term. A and B are parameters to be estimated. From the national income accounting rules, we know that

$X = C + V$, where V is other exogenous expenditure .

Thus, $X = (1-B)^{-1}(A + V + U)$. A shock in U causes a shock in X, and U and X are related by the algebra of the data generating process. The B cannot be estimated without bias using least squares.

Consider a third example of simultaneity that is more subtle. Carolyn Hoxby's problem in estimating the relationship between student performance and school competition was algebraically similar to the classic simultaneous equation problem of the Keynesian consumption function but yet quite a bit different in its theoretical origins.

She hypothesized that cities with many school districts provided more opportunity for parents to switch their children in the pursuit of better schools; thus, competition among school districts should lead to better schools as reflected in higher student test scores. Allowing for other explanatory variables, the implied regression is

$$Test\ scores = \beta_1 + \beta_2(number\ \ of\ \ school\ \ districts) + \ldots + \varepsilon \quad .$$

The causal effect of more school districts in a metropolitan area, however, may not be clearly discerned from this regression of mean metropolitan test scores on the number of school districts. Hoxby had anecdotal evidence that economy of scale arguments might lead to two good school districts being merged. At the other extreme, when districts were really bad they could not be merged with others and yet poor performance did not imply that the district would be shut down (it might be taken over by the state) even though a totally new district might be formed. That is, there is reverse causality: bad test performance leads to more districts and good performance leads to fewer.

As a final example of simultaneity, consider the Becker and Johnston (1999) study of the relationship between multiple-choice test and free-response test scores of economics understanding. Although these two form of tests are alleged to measure many different skills, matched scores are known to be highly correlated. Becker and Johnston assert that in part this is because both forms are a function of an unobservable ability that is cause in the error terms u and v in the following system of equations:

$$Multiple\text{-}choice\, score = \beta_1 + \beta_2(Free\text{-}response\, score) + \ldots + u\ .$$

$$Free\text{-}resonse\, score = \lambda_1 + \lambda_2(Multiple\text{-}choice\, score) + \ldots + v\ .$$

This system of equations should make the simultaneity apparent. As discussed in more detail later, the existence of the second equation (where both *u* and *v* include the effect of unobservable ability) makes the free-response test score an endogenous regressor in the first equation. Similarly, the existence of the first equation makes multiple-choice an endogenous regressor in the second.

ERRORS IN VARIABLES

Next consider an "errors in variables" problem that leads to regressor and error term correlation. In particular consider the example in which a student's *grade* on an exam in economics is hypothesized to be a function of *effort* and a random disturbance (*u*):

$$grade = A + B(effort) + u.$$

But effort is not observable (as was also the case for Milton Friedmen's permanent income). What is observable is the number of homework assignments completed, which may be either indicative of or the result of the amount of effort:

$$homework = C(effort) + v .$$

The equation to be estimated is then

$$grade = A + (B/C)homework + u^*, \text{ where } u^* = u - (1/C)v .$$

But a shock to *v* causes a shock to *homework*; thus, *homework* and *u** are correlated and the slope coefficient (*B/C*) cannot be estimated without bias via least squares.

A SINGLE VARIABLE INSTRUMENT

So what is the solution to these three problems of endogeneity? The instrumental variable (IV) solution is to find something that is highly correlated with the offending regressor but that is not correlated with the error term. In the case of Carolyn Hoxby's problem in estimating the relationship between student performance and school competition,

$$Test\ scores = \beta_1 + \beta_2(number\ of\ school\ districts) + \varepsilon ,$$

she observed that areas with a lot of school districts also had a lot of streams, possibly because the streams made natural boundaries for the school districts. She had what is become known as a **natural experiment**.[2] The number of streams was a random event in nature that had nothing to do with the population error term (ε) in the student performance equation but yet was highly related to number of school districts.[3]

For simplicity, ignoring any other variables in the student performance equation and measuring test scores, number of school districts and number of streams in deviation from their respective means, a consistent estimate of the effect of the number of school districts on test scores can be obtained with the **instrumental variable estimator**

$$b_2 = \frac{\sum(dev.\ in\ test\ scores)(dev.\ in\ number\ of\ streams)}{\sum\ (dev.\ in\ number\ of\ school\ districts)(dev.\ in\ number\ of\ steams)} .$$

To appreciate why the instrumental estimator works, consider the expected value of the terms in the numerator:

$$\begin{aligned}&E(\ deviations\ in\ test\ score)(deviations\ in\ number\ of\ streams)\\ &\quad = E\{[\beta_2(dev.\ in\ number\ of\ school\ districts) + \varepsilon\ ](dev.in\ number\ of\ streams)\}\\ &\qquad = \beta_2 \operatorname{cov}(number\ of\ school\ districts, number\ of\ streams)\ ,\end{aligned}$$

because the number of streams in an area is a purely random variable unrelated to epsilon.

In this example, deviations in one exogenous variable ($z - \bar{z}$: deviation in number of streams) could be used as an instrument for deviations in an endogenous explanatory variable ($x - \bar{x}$: deviations in number of school districts):

$$b_{IV} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})} \quad .$$

As with the OLS estimator, the IV estimator has an asymptotically normal distribution. The IV large sample variance is obtained by

$$s_{b_{IV}}^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 / n}{r_{x,z}^2 \sum_{i=1}^{n}(x_i - \bar{x})^2} \quad ,$$

where $r_{x,z}^2$ is the coefficient of determination (square of correlation coefficient) for *x* and *z*. Notice, if the correlation between *x* and *z* were perfect, the IV and OLS variance estimators would be the same. On the other hand, if the linear relationship between the *x* and *z* is weak, then the IV variance will greatly exceed that calculated by OLS.

Important to recognize is that a poor instrument is one that has a low $r_{x,z}^2$, causing the standard error of the estimated slope coefficient to be overly large, or has E($Z\varepsilon$) $\neq 0$, implying the Z was in fact endogenous. Unlike OLS estimators, the desired properties of IV estimators are all asymptotic; thus, to refer to small sample statistics like the *t* ratio is not appropriate. The appropriate statistic for testing with b_{IV} is the standard normal:

$$Z \cong \frac{B_{IV} - \beta}{S_{IV} / \sqrt{n}} \text{, for large } n.$$

It is important that this instrumental variable approach is not restricted to continuous endogenous variables. For example, Angrist (1990) was interested in the lifetime earnings effect of being a Vietnam War veteran. Measuring earnings in logarithmic form, Angrist's model was

$$Ln(earnings) = \beta_1 + \beta_2 veteran + ... + \varepsilon \, ,$$

where *veteran* is one if a veteran of the Vietnam War and zero otherwise. Angrist recognized that there was a sample selection problem (to be discussed in detail in a later module). It is likely that those who expected their earnings to be enhanced by the military experience are the ones who volunteer for service. That is, being a veteran is dependent on earning expectations at the time of joining. To the extent that all the factors that go into these earnings expectations and the decision to join are not captured in this single equation model they are in the epsilon error term. Thus, the error term must be correlated with being a veteran, $\mathrm{E}[(verteran)(\varepsilon)] \neq 0$.

For his instrument, Angrist observed that the lottery used to draft young men provided a natural experiment. Lottery numbers were assigned randomly; thus, the number received would not be correlated with ε. Men receiving lower numbers faced a higher probability of being drafted; thus, lottery numbers are correlated with being a Vietnam vet.

The use of these natural experiments has and likely will continue to be a source of instrumental variables for endogenous explanatory variables. Michael Murray (2006) provided a detailed but easily read review of natural experiments and the use of the IV estimator.

INSTRUMENTAL VARIABLE ESTIMATORS IN GENERAL

Often there are many exogenous variables that could be used as instruments for endogenous variables. Let matrix **Z** contain the set of all the endogenous variables that could serve as an instrument set of regressors. The instrumental variable estimator is now of the general form

$$\begin{aligned}&\mathbf{b}_{\mathbf{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}\\ &\mathbf{Var}(\mathbf{b}_{\mathbf{IV}}) = \boldsymbol{\sigma}^2(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{Z}(\mathbf{X}'\mathbf{Z})^{-1}\ .\end{aligned}$$

Unlike the selective replacement of a regressor with its instrument, for sets of regressors the typical estimation procedure involves the project of each of the columns of **X** in the column space of **Z**; at least conceptually we have

$$\hat{\mathbf{X}} = \mathbf{Z}[(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]\ .$$

This projected $\hat{\mathbf{X}}$ matrix is then substituted for **Z**.

$$\begin{aligned}&\mathbf{b}_{\mathbf{IV}} = (\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y}\\ &= [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}\\ &= [\mathbf{X}'(\mathbf{I}-\mathbf{M}_2)\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{I}-\mathbf{M}_2)\mathbf{y}\\ &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}\ ,\end{aligned}$$

which suggests a two step process: 1) regress the endogenous regressor(s) on all the exogenous variables; 2) use the predicted values from step 1 as replacement for the endogenous regressor in the original equation. This instrumental variable procedure is referred to as **Two-Stage Least Squares** (TSLS).

Unfortunately the standard errors associated with this TSLS estimation approach do not reflect the fact that the instrument is a combination of variables. That is, the standard errors obtained from the second step do not reflect the number of variables used in the first step predictions. In the case of a single instrument the difference between the variances of OLS and IV estimators was captured in the magnitude of $r^2_{x,z}$ and a similar adjustment must be made when multiple variables are used to form the instruments. Advanced econometrics programs like LIMDEP, SAS and SAS automatically do this in their TSLS programs.

The asymptotic variances correctly calculated can be extremely large if **Z** is not highly correlated with **X**; that is, $(\mathbf{Z'X})^{-1}$ is large if **X** and **Z** are not related. Also, for poor fitting instruments, it is possible to get negative R^2 when the typical computational formula [1 − (ResSS/TotalSS)] is used – recall that least squares minimized the ResSS so that it necessarily is less than or equal to TotalSS. But the IV estimator will have an ResSS greater than or equal to that of least squares. The fit of the IV can be so bad that its ResSS exceeds the Total SS. (For demonstration of this see Becker and Kennedy, 1992.)

DURBIN, HAUSMAN AND WU SPECIFICATION TEST APPLIED TO ENDOGENEITY

We wish to test $p\lim\left(\mathbf{X}'\boldsymbol{\varepsilon}/n\right)=0$, but cannot use the covariance between $n\times K$ matrix **X** and the *n* residuals ($e_i = y_i - \hat{y}_i$) in the $n\times 1$ vector **e** because $\mathbf{X'e}=0$ is a byproduct of least squares. Greene (2003, pp. 80-83) outlined the testing procedure originally proposed by Durbin (1954) and then extended by Wu (1973) and Hausman (1978). Davidson and MacKinnon (1993) are recognized for providing an algebraic demonstration of test statistic equivalence. Asymptotically, a **Wald (*W*) statistic** may be used in a Chi-square (χ^2) test with *K** degrees of freedom, or for smaller samples, an *F* statistic, with *K** and *n* − (*K* + *K**) degrees of freedom, can be used to test the joint significance of the contribution of the predicted values ($\hat{\mathbf{X}}^*$) of a regression of the *K** endogenous regressors, in matrix **X*,** on the exogenous variables (and a column of ones for the constant term) in matrix **Z:**

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

$$\text{where } \mathbf{X}^* = \mathbf{Z}\boldsymbol{\lambda} + \mathbf{u},\ \hat{\mathbf{X}}^* = \mathbf{Z}\hat{\boldsymbol{\lambda}}, \text{ and } \hat{\boldsymbol{\lambda}} \text{ is a least squares estimator of } \boldsymbol{\lambda}.$$

H_o : $\boldsymbol{\gamma}=0$, the variables in **Z** are exogenous.

H_A : $\boldsymbol{\gamma}\neq 0$, at least one of the variables in **Z** is endogenous.

As an example, consider the previously introduced economic exam grade equation that has the number of homework assignments as an explanatory variables:

$$grade = \beta_1 + \beta_2\, homework + \varepsilon\ .$$

The theoretical data generating process that gave rise to this model suggests that number of homeworks completed is an endogenous regressor. To test this we need truly exogenous variables – say x_2 and x_3 , which might represent student gender and race. The number of homeworks is then regressed on these two exogenous variables to get the least square equation

$$predicted\ homework = \hat{\boldsymbol{\lambda}}_1 + \hat{\boldsymbol{\lambda}}_2 x_2 + \hat{\boldsymbol{\lambda}}_3 x_3\ \ .$$

This predicted homework variable is then added to the exam grade equation to form the augmented regression

$$grade = \beta_1 + \beta_2\, homework + \boldsymbol{\gamma}\,(predicted\ homework) + \boldsymbol{\varepsilon}\,*$$

In this example, $K = 2$ (for β_1 *and* β_2) and $K^* = 1$ (for γ); thus, the degrees of freedom for the F statistic are 1 and $n - (K + K^*)$, which is also the square of a t statistic with $n - (K + K^*)$ degrees of freedom. That is, with only one endogenous variable and relatively small sample *n,* the t statistic printed by a computer program is sufficient to do the test. (Recall that asymptotically the t goes to the standard normal, with no adjustment for degrees of freedom required.) As with any other F, χ^2 , t or z test, calculated statistics greater than their critical values lead to the rejection of the null hypothesis. Important to keep in mind, however, is that failure to reject the null hypothesis at a specific probability of a Type I error does not prove exogeneity. The null hypothesis can always be rejected at some Type I error level.

Some introductory econometrics textbooks such as Wooldridge (2009, pp. 527-528) specify that the residuals from the auxiliary $\hat{\mathbf{X}}^* = \mathbf{Z}\hat{\boldsymbol{\lambda}}$ regression should be used in the augmented regression $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + (\mathbf{X}^* - \hat{\mathbf{X}}^*)\boldsymbol{\gamma} + \boldsymbol{\varepsilon}^*$. For example, in the case of the test scores model the augmented regression would be

$$grade = \beta_1 + \beta_2\, homework + \boldsymbol{\gamma}\,(homework - predicted\ homework) + \boldsymbol{\varepsilon}\,*$$

The additional calculation of this residual for inclusion in the augmented regression is not necessary because the absolute value of the estimate of γ and its standard error are identical regardless of whether predicted homework or the residual (= homework − predicted homework) is used.

Finally, keep in mind that you can use all the exogenous variables in the system to predict the endogenous variable. Some of these exogenous variables can even be in the original equation of interest – in the grade example, the grade equation might have been

$$grade = \beta_1 + \beta_2\, homework + \beta_3 x_3 + \varepsilon \ .$$

The auxiliary equation would still be

$$predicted\ homework = \hat{\boldsymbol{\lambda}}_1 + \hat{\boldsymbol{\lambda}}_2 x_2 + \hat{\boldsymbol{\lambda}}_3 x_3 \ .$$

As will become clear in the next section, the auxiliary equation should always have at least one more exogenous variable than the initial equation of interest.

IDENTIFICATION CONDITIONS

Whenever an instrumental variable estimator or two-stage least squares (2SLS) routine is employed consideration must be given to the identification conditions. To understand identification, consider a set of matched price and quantity observations (Figure 1, panel a) for which quantity values tend to rise as prices rise, as seen in the fitted OLS regression (Figure 1, panel b) The question to be asked: is this a supply relationship? As seen in Figure 1, panel c, the OLS line is not a supply curve. It is tracing equilibrium points.[4]

If a supply curve is to be estimated, more information than the observations that the quantity and price are positively related is needed. We need to identify a supply curve. This can be done if there is an exogenous variable that affects demand but does not affect supply. For example, household income likely affects demand but does not affect supply. In our previous simultaneous equation market model, for example,

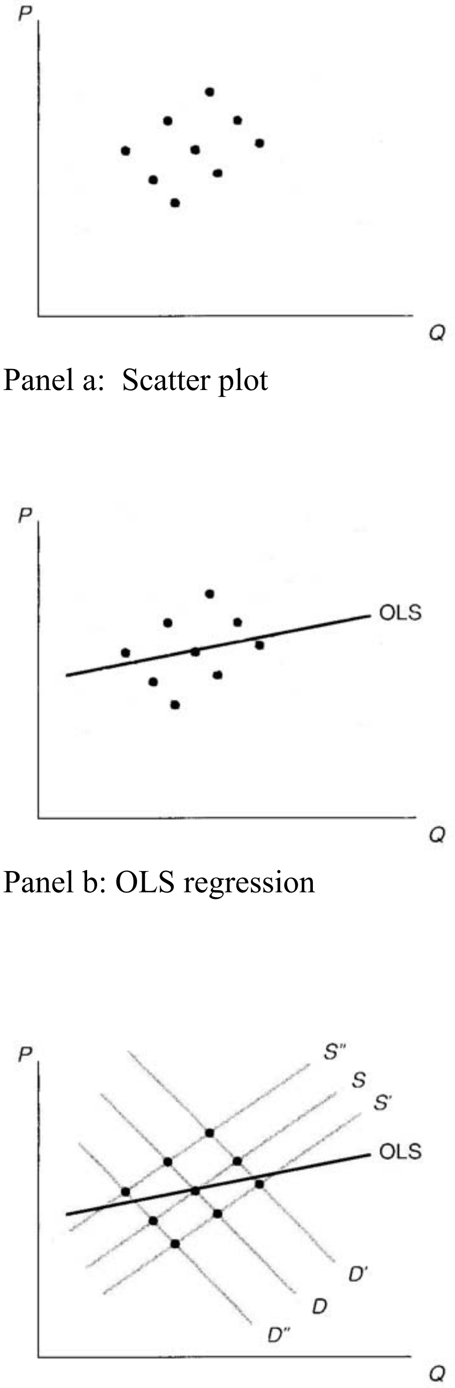
Supply in equilibrium: $Q = m + nP + U$

Demand in equilibrium: $Q=a + bP + cZ + V$

if Z is household income, then an increase in *Z* shifts the demand curve up, from *D* to *D*', but does not affect the supply curve (Figure 2); thus, the supply curve is identified by the change in equilibrium observations. Notice, however, that the demand curve is not identified because there is no unique exogenous variable in the supply equation.

Identification of this supply curve in this two endogenous variable system is achieved by an exclusionary or zero restriction -- the coefficient on income in the supply equation was restricted to zero**.** A necessary order condition for identification of any equation in a system is that the number of exogenous variables excluded from an equation must be at least as great as the number of endogenous variables less one. In this example, there were two endogenous variables (*Q* and *P*) and one exogenous variable (*Z*) excluded from the supply equation; thus, the necessary condition for identification was met: $2-1 \leq 1$. This necessary condition for identification is called the **order condition**.

Figure 1. Market data.

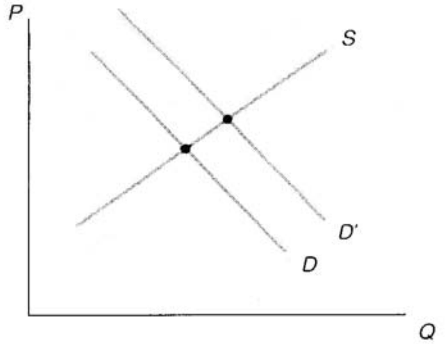


Panel a: Scatter plot

Panel b: OLS regression

Panel c: Demand and supply interaction

Figure 2. Supply curve is identified .



Any exogenous variable that is excluded from at least one equation in an equation system can be used as an instrumental variable. It can be used as an instrument in the equation from which it is excluded. For example, in the supply and demand equation system, the **reduced form** (no endogenous variables as explanatory variables) for *P* is

$$P = \frac{a-m}{n-b} + \frac{c}{n-b} Z + \frac{v-u}{n-b} = \beta_1 + \beta_2 Z + \varepsilon_2 \ .$$

And either the price predicted from this equation or *Z* itself can be used as the instrument for *P* in the supply equation. If there were more exogenous variables excluded from the supply equation then they could all be used to get predicted price from the reduced form equation.

Notice that the coefficient on *Z* in the reduced form equation for *P* must be nonzero for *Z* to be used as an instrument, which requires that $c \neq 0$ and $n - b \neq 0$. This requirement states that exogenous variable(s) excluded from the supply equation must have a nonzero population coefficient in the demand equation and that the effect of price cannot be the same in both demand and supply. This is known as the **rank condition**.

As an example of identification in economic education research consider the work of Becker and Johnston (1999). In addition to the multi-dimensional attributes of the Australian 12th grade test takers (captured in the explanatory *X* variables such as gender, age, English a second language, etc.), Becker and Johnston called attention to classroom and peer effects that might influence multiple-choice and essay type test taking skills in different ways. For example, if the student is in a classroom that emphasizes skills associated with multiple-choice testing (*e.g*., risk-taking behavior, question analyzing skills, memorization, and keen sense of judging between close alternatives), then the student can be expected to do better on multiple-choice questions. By the same token, if placed in a classroom that emphasizes the skills of essay test question answering (*e.g*., organization, good sentence and paragraph construction, obfuscation when uncertain, and logical argument), then the student can be expected to do better on the essay component. Thus, Becker and Johnston attempted to control for the type of class of which the student is a member. Their measure of “teaching to the multiple-choice questions” is the mean score on the multiple-choice questions for the school in which the i^{th} student took the 12th grade economics course. Similarly, the mean school score on the essay questions is their measure of the i^{th} student’s exposure to essay question writing skills.

In equation form, the two equations that summarize the influence of the various covariates on multiple-choice and essay test questions are written as the following **structural equations**:

$$M_i = \rho_{21} + \rho_{22} W_i + \rho_{23}\,\bar{M}_i + \sum_{j=4}^{J} \rho_{2j} X_{ij} + U_i^{*} \qquad .$$

$$W_i = \rho_{31} + \rho_{32} M_i + \rho_{33}\,\bar{W}_i + \sum_{j=4}^{J} \rho_{3j} X_{ij} + V_i^{*} \qquad .$$

M_i and W_i are the i^{th} student’s respective scores on the multiple-choice test and essay test. $\bar{M}_i$ and $\bar{W}_i$ are the mean multiple-choice and essay test scores at the school where the i^{th} student took the twelfth grade economics course. The X_{ij} variables are the other exogenous variables used to explain the i^{th} student’s multiple choice and essay marks, where the ρs are parameters to be estimated. U_I^{*} and V_i^{*} are assumed to be zero mean and constant variance error terms that may or may not each include an effect of unobservable ability.

Least squares estimation of the ρs will involve bias if the respective error terms U_i^{*} and V_i^{*} are related to regressors (W_i in the first equation, and M_i in second equation). Such relationships are seen in the **reduced form equations**, which are obtained by solving for M and W in terms of the exogenous variables and the error terms in these two equations:

$$M_i = \Gamma_{21} + \Gamma_{22}\,\bar{W}_i + \Gamma_{23}\,\bar{M}_i + \sum_{j=4}^{J} \Gamma_{2j} X_{ij} + U_i^{**} \qquad .$$

$$W_i = \Gamma_{31} + \Gamma_{32}\,\bar{M}_i + \Gamma_{33}\,\bar{W}_i + \sum_{j=4}^{J} \Gamma_{3j} X_{ij} + V_i^{**} \qquad .$$

The reduced form parameters (Γs) are functions of the ρs, and U^{**} and V^{**} are dependent on U^{*} and V^{*}:

$$U_i^{**} = \frac{U_i^* \; + \rho_{22} V_i^*}{1 - \rho_{22}\rho_{32}} \quad .$$

$$V_i^{**} = \frac{V_i^* \; + \rho_{32} U_i^*}{1 - \rho_{22}\rho_{32}} \quad .$$

In the reduced form error terms, it can be seen that a random shock in U^* causes a change in V^{**}, which causes a change in W in the reduced form. Thus, W and U^* are related in the essay structural equation, and consistent estimation of the parameters in this equation is not possible using least squares. Similarly, a shock in V^*, and a resulting change in U^{**}yields a change in M. Thus, M and V^* are dependent in the structural equation, and least squares estimators of the parameters in that equation are inconsistent.

The inclusion of $\bar{M}_i$ and $\bar{W}_i$ in their respective structural equations, and their exclusion from the other equation, enables both of the structural equations to be identified within the system. For example, if a student moves from a school with a low average multiple-choice test score to one with a higher average multiple-choice test score, then his or her multiple-choice score will rise via a shift in the *M*-*W* relationship in the first structural equation, but this shift is associated with a move along the *W*-*M* relationship in the second structural equation; thus, the second structural equation is identified. Similarly, if a student moves from a low average essay test score school to a higher one, then his or her essay test score will rise via a shift in the *W*-*M* relationship in second structural equation, but this shift implies a move along the *M*-*W* relationship in the first structural equation, and this first structural equation is thus identified. Most certainly, identification hinges critically on justifying the exclusionary rule employed.

To summarize, identification involved two conditions.

> **The order condition for identifying an equation in a model of *K* equations and *K* endogenous variables is that the equation exclude at least *K* – 1 variables that appear in the model. Alternatively, if the number of potential instruments (exogenous variables in the system but not in the equation) equals the number of endogenous regressors, the equation is exactly identified. If exactly *K* – 1 variables are excluded, then the equation is just identified. If more (less) than *K* – 1 variables are excluded, then the equation is over (under) identified.**
>
> **The order condition is a necessary condition, but not a sufficient condition for identification.**

> **The sufficient condition for identification is the rank condition. By the rank condition an equation is identified if and only if at least one nonzero determinant of order exists for the coefficients of the excluded variables that are included in the other equations of the model. This sufficient condition requires that variables excluded from the equation, but included in the other equations of the model, not be dependent. It ensures that the parameters can be estimated from the reduced form.**

CONCLUDING COMMENTS

Eagerness to employ natural experiments and instrumental variables to address problems of endogeneity have exploded within economics, but along with that growth has come questions of validity, as seen most recently in criticism of the work of Waldman, Nicholson and Adilov (2006) that suggests that TV watching causes autism. Economist Waldman recognized that he could not simply run a regression of incidence of autism on amount of TV watched because autism might in some way influence the TV watching. He observed, however, that TV watching and precipitation were highly correlated. Because rainfall is a natural occurrence unrelated to the error term in the autism regression, he had his instrument for TV watching. As reported in the *Wall Street Journal*, Whitehouse (2007), those who specialize in the study of autism were not impressed, labeling Waldman's work "irresponsible" (because it shifts responsibilty to parents when experts claim that it is genetic and beyond the control of parent) and "junk science."

When instrumental variables are used, that which is measured is unclear. Unanswered in the Waldman, Nicholson and Adilov study is how TV watching influences autism. Arm-chair speculation that children are distracted by television is not convincing to those who have devoted their lives to studying autism. Joseph Piven, Director of the Neurodevelopment Disorder Research Center at the University of North Carolina, is quoted in the *WSJ* article stating that "it is just too much of a stretch to tie (autism) to television-watching. Why not tie it to carrying umbrellas?" More damning still are the quotes from Nobel Laureate in Economics James Heckman, "There's a saying that ignorance is bliss," and IV econometrician guru Jerry Hausman, "I think that characterizes a lot of the enthusiasm for these instruments. If your instruments aren't perfect, you could go seriously wrong."

REFERENCES

Angrist, Joshua D. (1990). “Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records**, ”***American Economic Review*,” Vol. 80 (June): 313-336.

Angrist, Joshua D. and Alan B. Krueger (2001). "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments," *Journal of Economic Perspectives*, Vol. 15 (Fall),: 69-85.

Becker, William E. (2004). “Omitted Variables and Sample Selection Problems in Studies of College-Going Decisions,” *Public Policy and College Access: Investigating the Federal and State Role in Equalizing Postsecondary Opportunity*, Edward St. John (ed)*,* Vol. 19. NY: AMS Press: 65-86.

Becker, William E. and Carol Johnston (1999).“The Relationship Between Multiple Choice and Essay Response Questions in Assessing Economics Understanding,” *Economic Record* (Economic Society of Australia), Vol. 75 (December): 348-357.

Becker, William E. and Peter Kennedy (1995). “A Lesson in Least Squares and R Squared,” *American Statistician*, Vol. 55 (November): 282-283. Portions reprinted in Dale Poirier, *Intermediate Statistics and Econometrics* (Cambridge: MIT Press, 1995): 562-563.

Becker, William E. and Michael Salemi (1977). “The Learning and Cost Effectiveness of AVT Supplemented Instruction: Specification of Learning Models,” *Journal of Economic Education* Vol. 8 (Spring) : 77-92.

Davidson, Russell and James G. MacKinnon (1993). *Estimation and Inference in Econometrics*. New York, Oxford University Press.

Durbin, James. (1954). “Errors in Variables,” *Review of the International Statistical Institute*, Vol. 22(1): 23-32.

Finkelstein, Michael and Bruce Levin (1990). *Statistics for Lawyers*. New York, Springer-Verlag.

Greene, William (2003). *Econometric Analysis*. 5th Edition, New Jersey: Prentice Hall.

Hausman, Jerry (1978). “Specification Tests in Econometrics,” *Econometrica*, Vol. 46 (November): 1251-1271.

Hilsenrath, Jon (2005). “Novel Way to Assess School Competition Stirs Academic Row,” *Wall Street Journal* (October 24): A1 and A11

Hoxby, Caroline M. (2000). “Does Competition Among Public Schools Benefit Students and Taxpayers?’ *American Economic Review*. Vol. 90 (December): 1209-1238.

Kennedy, Peter (2003). *A Guide to Econometrics*, 5th Edition, United Kingdom: Blackwell.

Murray, Michael P. (2006). “Avoiding Invalid Instruments and Coping with Weak Instruments,” *Journal of Economic Perspectives*. Vol. 20 (Fall): 111-132.

Rosenzweig, Mark and Kenneth Wolpin (2000). "Natural 'Natural Experiments' in Economics," *Journal of Economic Literature*. Vol. 38 (December): 827-874.

Stigler, Stephen (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge: Harvard University Press.

Waldman, Michael, Sean Nicholson and Nodir Adilov (2006). “Does Television Cause Autism?” NBER Working Paper No. W12632 (October).

Whitehouse, Mark (2007). “Mind and Matter: Is an Economist Qualified to Solve Puzzle of Autism? Professor’s Hypothesis: Rainy Days and TV May Trigger Condition,” *Wall Street Journal*. February 27: A.1.

Wooldridge, Jeffrey M. (2009). *Introductory Econometrics: A Modern Approach.* 4th Edition Mason Oh; South-Western.

Working, E. J. (1927). “What Do Statistical ‘Demand Curves’ Show?” *Quarterly Journal of Economics*, 41(2) : 212-235.

Wu, De-Min. (1973). “Alternative Tests of Independence between Stochastic Regressors and Disturbances,” *Econometrica,* 41(July): 733-750.

ENDNOTES

[1] Conceptually there are more than three forms of endogeneity that could occur. For example, if there is a lagged dependent variable and the residuals are serially correlated, then the lagged dependent variable will be correlated with the error term. This is not a problem for the typical cross-section regressions considered by economic educators but does become a problem when time is introduced. To see this consider a data generating process in which knowledge of economics (Y_{it}) of the i^{th} student at time *t* is a linear function of the student's ability at time *t* (x_{it}) plus an error term (ε_{it}):

$$y_{it} = \beta_1 + \beta_2 x_{it} + \varepsilon_{it} .$$

At time t-1, knowledge is then given by

$$y_{it-1} = \beta_1 + \beta_2 x_{it-1} + \varepsilon_{it-1} .$$

If learning is assessed in the following equation, then the pretest $y_{it\text{-}1}$ regressor is endogenous by construction:

$$y_{it} = \beta 1(1-\rho) + \beta_2(x_{it} - \rho x_{it-1}) + \rho y_{it-1} + (\varepsilon_{it} - \rho\varepsilon_{it-1})$$

$$E[y_{it-1}(\varepsilon_{it} - \rho\varepsilon_{it-1})] = \rho E(y_{it-1}\varepsilon_{it-1}) \neq 0 .$$

As demonstrated in a later module, sample selection also leads to endogeneity problems. However, the sample selection form of endogeneity is typically associated with a truncation of the error term, which is a different problem than the three sources of endogeneity considered in the text of this module, where the error term is always assumed to be continuous.

[2] Natural experiments and instrumental variables are not synonymous but Rosenzweig and Wolpin (2000, pp.827-8) state "The most widely applied approach to identifying causal or treatment effects, which has a long history in economics, employs instrumental variable techniques . . .in standard instrumental variable studies, economists as well as researchers in other fields have sought out 'natural experiments,' random treatments that have arisen serendipitously . . ."

[3] Jon Hilsenrath reported in his *Wall Street Journal* (October 24, 2005, pp. A1 and A11) "Novel Way to Assess School Competition Stirs Academic Row," that Princeton University economist Jesse Rothstein questioned Hoxby's use of the instrumental variable technique because he could not replicate her count of streams, which aside from ethical questions posed by Hilsenrath introduces an added complication if her instrument has a measurement error problem.

[4] Working (1927) provided an early intuitive explanation of simultaneity and the identification problems that is still relevant today as seen in its modern rendition by Kennedy (2003).