A GUIDE TO ECONOMETRICS

SIXTH EDITION

PETER KENNEDY
Simon Fraser University

Blackwell
Publishing

Chapter 9

Violating Assumption Four: Instrumental Variable Estimation

9.1 Introduction

The fourth assumption of the classical linear regression (CLR) model specifies that the observations on the explanatory variables can be considered fixed in (conceptual) repeated samples. In many economic contexts the explanatory variables are themselves random/stochastic variables and thus could not possibly have the same values in repeated samples. A classic example is a simultaneous equation system with supply and demand curves. To estimate the demand curve we would regress quantity on price, among other variables. When we draw new error terms for the supply and demand equations to create a repeated sample, the intersection of the supply and demand curves changes and so the price changes: price is stochastic, it cannot remain fixed in repeated samples.

This assumption of fixed regressors is made mainly for mathematical convenience; if the regressors can be considered to be fixed in repeated samples, the desirable properties of the ordinary least squares (OLS) estimator can be derived quite straightforwardly. The role of this assumption in derivations of OLS estimator properties is to make the regressors and errors independent of one another. If this assumption is weakened to allow the explanatory variables to be stochastic but to be distributed independently of the error term, all the desirable properties of the OLS estimator are maintained; their algebraic derivation is more complicated, however, and their interpretation in some instances must be changed (for example, in this circumstance β^{OLS} is not, strictly speaking, a linear estimator). Even the maximum likelihood property of β^{OLS} is maintained if the disturbances are distributed normally and the distribution of the regressors does not involve the unknown parameters β and σ^2.

This fourth assumption can be further weakened at the expense of the small-sample properties of β^{OLS}. If the regressors are *contemporaneously uncorrelated* with the disturbance vector, the OLS estimator is biased but retains its desirable asymptotic properties. Contemporaneous uncorrelation in this context means that the nth observation on all regressors must be uncorrelated with the nth disturbance term, but it is allowed to be correlated with the disturbance terms associated with other observations. Suppose, for example, that a lagged value of the dependent variable, lagged y, appears as one of the explanatory variables. When we draw a new vector of error terms to create a repeated sample, all the dependent variable values, including lagged y, change because the error is a part of the equation determining the dependent variable. So the value of lagged y, one of the explanatory variables, is stochastic and cannot be considered as fixed in repeated samples. But in this example, although lagged y is correlated with the error in its own time period, period t–1, it is not correlated with the error in the following period, period t. The error in the equation being estimated is the error for period t, so no contemporaneous correlation exists between lagged y and the regression error. OLS will be biased, but consistent. In this case no alternative estimators are available with superior small-sample properties, so the OLS estimator is retained on the basis of its desirable asymptotic properties. Henceforth the "contemporaneous" qualification is dropped for expositional ease, so that the terminology "regressor correlated with the error" means contemporaneous correlation.

If the regressors are correlated with the error term, the OLS estimator is biased even asymptotically. (And this bias in general spills over to the estimates of all the slope coefficients, not just the slope of the regressor creating the problem!) The bias happens because the OLS procedure, in assigning "credit" to regressors for explaining variation in the dependent variable, assigns, in error, some of the disturbance-generated variation of the dependent variable to the regressor with which that disturbance is correlated. Consider as an example the case in which the correlation between the regressor and the disturbance is positive. When the disturbance is higher the dependent variable is higher, and owing to the correlation between the disturbance and the regressor, the regressor is likely to be higher, implying that too much credit for making the dependent variable higher is likely to be assigned to the regressor. This is illustrated in Figure 9.1. If the error term and the independent variable are positively correlated, negative values

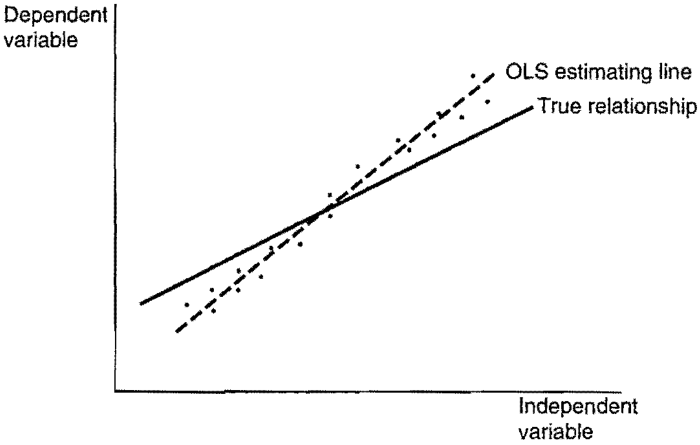

Figure 9.1 Positive contemporaneous correlation.

of the disturbance will tend to correspond to low values of the independent variable and positive values of the disturbance will tend to correspond to high values of the independent variable, creating data patterns similar to that shown in the diagram. The OLS estimating line clearly overestimates the slope of the true relationship. (This result of overestimation with positive correlation between the disturbance and regressor does not necessarily hold when there is more than one explanatory variable, however; the pattern of bias in the multivariate case is complicated.) Note that the estimating line provides a much better fit to the sample data than does the true relationship; this causes the variance of the error term to be underestimated.

When there is correlation between a regressor and the error term, that regressor is said to be *endogenous*; when no such correlation exists the regressor is said to be *exogenous*. Endogeneity gives rise to estimates biased even asymptotically, making economists very unhappy. Indeed, this is one of the features of economic data that distinguishes econometrics from other branches of statistics. The heart of the matter is that the data with which econometricians work seldom come from experiments designed to ensure that errors and explanatory variables are uncorrelated. Here are some examples of how this endogeneity problem can arise.

Measurement error in explanatory variables. Suppose $y = \alpha + \beta x + \varepsilon$ but that measured x, x_m, is $x + u$ where u is a random error. Add and subtract βu to find that the relationship between y and x_m, the explanatory variable used in the regression, is $y = \alpha + \beta x_m + (\varepsilon - \beta u)$. When a repeated sample is taken it must involve new values for measurement errors u in the data, as well as new values for the traditional error term ε. But clearly u affects both x_m and the composite error $(\varepsilon - \beta u)$; in this regression there is correlation between the explanatory variable x_m and the error term $(\varepsilon - \beta u)$. The general topic of measurement errors is discussed in chapter 10.

Autoregression with autocorrelated errors. Suppose the lagged value of the dependent variable, lagged y, is a regressor. When new errors are drawn for a repeated sample all values of the dependent variable change, including lagged y, so this regressor is stochastic. As noted earlier, lagged y is not contemporaneously correlated with the regression error. If the errors are autocorrelated, however, then this period's error is correlated with last period's error. But last period's error is a direct determinant of lagged y: this creates correlation between lagged y and this period's error. An obvious lesson here is that whenever lagged y appears as a regressor we should test for autocorrelated errors! Autoregression is discussed further in chapter 10.

Simultaneity. Suppose we are estimating a demand curve so that one of the explanatory variables is price. If the error term in this equation bumps up it shifts the demand curve and so through its simultaneity/intersection with the supply curve changes the price. This causes correlation between the demand curve errors and the explanatory variable price. In general, all endogenous variables in a system of simultaneous equations are correlated with all of the errors in that system. Simultaneity, sometimes referred to as *reverse causation*, is very common in econometric work.

Changes in policing, for example, could cause changes in crime rates, but changes in crime rates prompt changes in policing. Suppose we regress crime on policing. When the error term in this regression bumps up it directly increases crime. This increase in crime increases policing through the reverse causation (the simultaneity). This means that the error term is correlated with policing, so when we regress crime on policing we get biased estimates. In effect, when we regress crime on policing, some of the reverse influence of crime affecting policing gets into the coefficient estimate, creating simultaneity bias. Simultaneity is discussed at length in chapter 11.

Omitted explanatory variable. Whenever an explanatory variable has been omitted its influence is encompassed by the error term. But often the omitted explanatory variable is correlated with included explanatory variables. This makes these included explanatory variables correlated with the error term. Suppose, for example, that we are regressing wage on years of education but do not have an ability variable to include in the regression. People with higher ability will earn more than others with the same years of education, so they will tend to have high positive error terms; but because they have higher ability they will find it worthwhile to take more years of education. This creates correlation between the error term and the explanatory variable years of education. This is another way of viewing the omitted variable specification error discussed earlier in chapter 6.

Sample selection. Often people appear in a sample because they have chosen some option that causes them to be in the sample. Further, often this choice is determined by characteristics of these people that are unmeasured. Suppose you are investigating the influence of class size on student learning. Some parents may have gone to a lot of trouble to ensure that their child is in a small class; parents who take this trouble probably are such that they work hard with their child at home to enhance their child's learning, and thereby create for that child, other measured things being equal, a positive error term. A consequence of this sample selection phenomenon is that small classes are more likely to be populated by children with positive errors, creating (negative) correlation between class size and the error in the learning equation. This problem is sometimes referred to as *unobserved heterogeneity*; the observations in the sample are heterogeneous in unobserved ways that create bias. This is related to the omitted explanatory variable category above; if we could measure the causes of the heterogeneity we could include these measures as extra explanatory variables and so eliminate the bias. This sample selection problem is also addressed later in chapter 16 when discussing limited dependent variables.

The bottom line here is that correlation between explanatory variables and a regression's error term is not an unusual phenomenon in economics, and that this is a serious problem because it introduces bias into the OLS estimator that does not disappear in large samples. Unfortunately, there does not exist an alternative estimator which is unbiased; the best we can do is turn to estimation procedures that are unbiased asymptotically, or defend OLS using the mean square error (MSE) criterion. The purpose of this chapter is to exposit the instrumental variable (IV) estimator, the most common estimator employed as an alternative to OLS in this context. Why is a whole chapter devoted to this estimator? There are several reasons for this. First, this procedure is one which has a rich history in econometrics, and its frequent use is a major way in which econometrics is distinguished from other branches of statistics. Second, the procedure permeates a large part of econometrics in various ways, so a good understanding of it is of value. And third, there are lots of issues to be dealt with, some of which are quite problematic: How does the technique work? How do we find the instruments it requires? How do we test if these instruments are any good? How do we interpret the results? We begin by describing the IV procedure.

9.2 The IV Estimator

The IV procedure produces a consistent estimator in a situation in which a regressor is correlated with the error, but as noted later, not without cost. To facilitate exposition henceforth regressors that are correlated with the error are referred to as "troublesome" or "endogenous" explanatory variables. To use the IV estimator one must first find an "instrument" for each troublesome regressor. (If there is not at least one unique instrument for each troublesome variable the IV estimation procedure is not *identified*, meaning that it cannot produce meaningful estimates of the unknown parameters. Not being identified is like having more unknowns than equations – the equations can't be solved, in the sense that there is an infinite number of values of the unknowns that satisfy the equations!) This instrument is a new independent variable which must have two characteristics. First, it must be uncorrelated with the error; and second, it must be correlated (preferably highly so) with the regressor for which it is to serve as an instrument. The IV estimator is then found using a formula involving both the original variables and the IVs, as explained in the technical notes. (It is *not* found by replacing the troublesome variable with an instrument and running OLS, as too many students believe!) The general idea behind this estimation procedure is that it takes variation in the explanatory variable that matches up with variation in the instrument (and so is uncorrelated with the error), and uses only this variation to compute the slope estimate. This in effect circumvents the correlation between the error and the troublesome variable, and so avoids the asymptotic bias.

A major drawback to IV estimation is that the variance of the IV estimator is larger than that of the OLS estimator. It is easy to see why this happens. As explained earlier, only a portion of the variation in the troublesome variable (the portion matching up with the instrument) is used to produce the slope estimate; because less information is used, the variance is larger. This portion is larger (and so the variance smaller) the larger is the correlation between the troublesome variable and the instrument; this is why the "preferably highly so" was included earlier. This higher variance, which is sometimes dramatically higher, is the price paid for avoiding the asymptotic bias of OLS; the OLS estimator could well be preferred on the MSE criterion. A second major drawback is that if an instrument is "weak," meaning that its correlation with the troublesome variable is low, as explained later the IV estimates are unreliable, beyond just having a high variance.

How can we find instruments? At first glance it seems that finding suitable instruments is an impossible task, and sometimes it is. But in surprisingly many cases the context of the problem, economic theory, unexpected events (so-called "natural" experiments), and, especially, clever researchers, suggest instruments. Here are some examples.

1. If a troublesome variable measured with error has a small measurement error variance the measurement error may not affect the rank order of the magnitudes of the troublesome variable observations, in which case a variable created as the rank order (i.e., taking values $1, 2, 3, \ldots$) will be a good instrument: it is highly correlated with the troublesome variable and not correlated with the measurement error (and so not correlated with the regression error).
2. Suppose the troublesome variable is lagged y, the lagged value of the dependent variable, and there is another explanatory variable, say x, that is not troublesome (i.e., it is exogenous). One of the direct determinants of lagged y is lagged x, so lagged x is highly correlated with lagged y. Because x is exogenous, lagged x is uncorrelated with the error. So lagged x is a viable instrument for lagged y in this context.
3. Simultaneous equation systems contain endogenous variables, all of which are troublesome if they are serving as regressors. But these equation systems contain exogenous variables that are determined outside the system. Any change in an exogenous variable will shift one or more of the equations and so change the equilibrium values of all of the endogenous variables, so all exogenous variables are correlated with all endogenous variables. But because they are exogenous, they are all uncorrelated with the errors in the simultaneous equations. So any exogenous variable in a system of equations is a potential instrument for any endogenous/troublesome variable.
4. In a wage equation years of education is a troublesome variable if an ability variable is missing. Distance to the nearest college has been used as an instrument, on the grounds that other things equal those closer to college are more likely to attend but this distance should bear no relationship to ability (and so be uncorrelated with the error). Another, more controversial, instrument suggested here is quarter of year of birth. Depending on quarter of year of birth, some people are forced by legal regulations to spend more time in school, so quarter of year of birth should be correlated with years of education but have no relationship to ability.
5. Class size is a troublesome variable when estimating learning determinants because of selection problems. In some cases unexpected increases in enrolments have caused class sizes to be halved to meet legal restrictions on class sizes. A dummy variable equal to one for such classes and zero otherwise can capture such natural experiments. It is correlated with class size, but should not be correlated with the unmeasured characteristics of students, discussed earlier, that might otherwise be associated with smaller class sizes.
6. Consider a regression of incidence of violent crime on percentage of population owning guns, using data on US cities. Because gun ownership may be endogenous

(i.e., higher crime causes people to obtain guns), gun magazine subscriptions is suggested as an IV for gun ownership; this should be correlated with gun ownership, but not correlated with the error in the violent crime equation (i.e., an increase in the incidence of violent crime may cause more people to buy guns, but probably will not affect subscriptions to gun magazines). This actually turns out to be a bad instrument, for reasons explained later.

7. A high risk of expropriation could affect a country's per capita income by dampening entrepreneurial activity. But a higher per capita income could make a country feel it can afford to do away with such politically-determined constraints. Regressing per capita income on a risk of expropriation measure would not be convincing because of this reverse causation. The expected mortality of European settlers is suggested as an instrument for the risk of expropriation. Why? A high expected mortality of European settlers reduced the intensity of European colonization, which in turn increased the risk of expropriation. So the expected mortality of European settlers should be correlated with the risk of expropriation, but should not be correlated with the error in the income equation (i.e., if this error bumps up it may prompt changes in politically-determined constraints, but it will not affect the expected mortality rate of European settlers!)
8. Does a higher incarceration rate reduce crime? Regressing crime rate on incarceration rate will not be convincing because higher crime rates cause society to increase incarceration. This reverse causation problem requires IV estimation. In many states legal problems associated with overcrowding in state prisons forced the state to decrease its prison population. This is an exogenous change in the incarceration rate and so a variable capturing these events is used as an instrument. It is clearly correlated with the incarceration rate but because of the way it came about is not correlated with the error in the crime rate equation.

As these examples illustrate, sometimes instruments are readily available and sometimes they require considerable ingenuity on the part of a researcher. In all cases, researchers should investigate thoroughly the validity of an instrument. There are several means of doing so, beyond telling a good story based on the context of the problem, intuition, economic theory, or the serendipitous occurrence of a "natural experiment." First, tests for the validity of overidentifying instruments, explained later, can be undertaken. Second, we can check if the instrument has the anticipated sign, and is significant, when the troublesome variable is regressed on this instrument. Third, if alternative instruments are available we could check if similar estimates result from using IV estimation employing different instruments. Fourth, we should defend our implicit assumption that an instrument is not an explanatory variable in the equation being estimated, perhaps by referring to existing literature. Fifth, we should explain why our instrument is not correlated with an omitted explanatory variable (because if so it would be correlated with the error, which embodies this omitted variable!) Regardless of how cogently the validity of an instrument is defended, disputes can arise concerning the need for instruments, the validity of the instruments, and the interpretation of the IV coefficient estimates. The next section discusses these issues.

9.3 IV Issues

9.3.1 How can we test if errors are correlated with regressors?

A testing methodology popularized by Hausman (1978), and therefore called the *Hausman test,* is used for this purpose. To perform this test we need two estimators of the coefficients, both of which are consistent under the null of no correlation between the errors and the regressors, but only one of which is consistent when the null is false (i.e., when the error and the regressors are correlated). In particular, both OLS and IV estimators are consistent when the null is true, but only the IV estimator is consistent when the null is false. The idea behind this test is that if the null is true both estimates should be about the same (because they are both unbiased), whereas if the null is false there should be a substantive difference between the two estimates (because one is biased and the other is not). The Hausman test is based on seeing if there is a significant difference between the two estimates. In its original form this test is computationally awkward, but simple methods of conducting this test have been devised, as explained in the general notes.

9.3.2 How can we test if an instrument is uncorrelated with the error?

One of the requirements of an instrument is that it is uncorrelated with the error in the equation being estimated. This is not easy to check. Indeed, in the just identified case (only one instrument for each troublesome variable) it is impossible to test; in this circumstance we must rely on the logical reasons that lie behind the choice of instrument, based on economic theory, perhaps, or the context of the application.

In the overidentified case (more instruments than troublesome variables), however, a test of sorts is available. The "of sorts" qualification is added because the available tests do not actually test what we want to test. These tests assume that among the instruments is at least one valid instrument per troublesome variable so that IV estimation is identified and so legitimate. On the basis of this assumption they test only for the validity of the extra, overidentifying instruments, *without telling us which instruments these are*!

Here is the logic of this test. If we estimate the equation using IV we should get a "good" estimate of the parameters and so the resulting residuals should be "good" estimates of the original errors. These errors should not be correlated with the instruments, for two reasons – the instruments are not supposed to be explanatory variables in the original relationship, and to be valid instruments they are not supposed to be correlated with these errors. So if we regress these residuals on the instruments the coefficient estimates should test insignificantly different from zero. Note that this test is actually testing a dual null; it could reject the null either because the instruments are correlated with the errors or because there is a specification error and the instruments actually should have been included as explanatory variables in the equation being estimated. This test is often referred to as the *Sargan test*; see the technical notes for more detail.

It is worth repeating that the validity of this test depends on there being among the instruments enough legitimate instruments for identification. When accepting the null that the overidentifying restrictions are valid, it must be remembered that this does not necessarily endorse the validity of all the instruments.

9.3.3 How can we test if an instrument's correlation with the troublesome variable is strong enough?

Another requirement of an IV is that it is correlated, preferably highly so, with the troublesome variable. It is easy to check if an instrument is correlated with a troublesome variable simply by regressing the troublesome variable on the instrument and seeing if there is a substantive relationship, as indicated by the magnitude of R^2. Although there is some truth to this, this statement is vague because it does not tell us exactly what is meant by a "substantive" relationship. It is also misleading because it slides around the important issue of "weak" instruments.

To understand this issue we need to recall that although IV estimators are consistent, so that they are asymptotically unbiased, in small samples all IV estimators are biased (in the same direction as OLS). How big is this bias? It turns out that this bias can be quite large, even in very large samples, whenever an IV is "weak" in that it is not strongly correlated with the troublesome variable. Furthermore, using multiple weak instruments causes this bias to be worse. How strongly correlated with the troublesome variable do instruments need to be to avoid this problem? A popular rule of thumb that has been adopted here is to regress the troublesome variable on all the instruments and calculate the F statistic for testing the null that the slopes of all the instruments equal zero. If this F value exceeds 10 the IV bias should be less than 10% of the OLS bias. Like most rules of thumb this rule is too crude; see the general notes for more on this issue.

We know that the IV variance is greater than the OLS variance. An estimated IV variance dramatically larger than the OLS estimated variance is an indication that we are dealing with a weak instrument. (Only the variances of slope estimates of troublesome variables are substantively affected, however.) But there is another problem here. A weak instrument also causes the IV variance to be underestimated in small samples; this causes the true type I error rate to be higher than its chosen level. This is made worse by the fact that with weak instruments in small samples the distribution of the IV estimator is not approximated well by its asymptotic (normal) distribution. In short, weak instruments lead to unreliable inference.

Finally, when instruments are weak, even mild endogeneity of the instrument (i.e., the instrument being just slightly correlated with the error) can cause an IV estimate to exhibit more bias (even asymptotically) than OLS. With all these problems associated with weak instruments, if researchers cannot determine with some confidence that their instruments are strong, they should find another instrument, use another estimating procedure, or resign themselves to using OLS.

9.3.4 How should we interpret IV estimates?

An IV estimate is supposed to be an estimate of the exact same parameter that OLS is estimating, so it should have the same interpretation. But for this to be the case there is an implicit assumption operating that can be of substantive importance. The IV estimator works by picking out variations in the troublesome explanatory variable that match up with variations in the instrument and basing the slope estimate only on these variations. If the influence of the troublesome variable on the dependent variable is the same for all troublesome variable variations in the sample, the IV and OLS estimates are comparable. But if certain types of individuals in the sample react differently when the troublesome variable changes, the two estimates could be measuring different things. Suppose the sample contains observations on two types of individuals, A and B, whose slope coefficients on the troublesome variable are β_A and β_B. An OLS slope estimate will be an estimate of a weighted average of β_A and β_B, reflecting the relative variability of the troublesome variable for the two types of individuals in the sample. This may be exactly what we want to be measuring. But now suppose we have an IV that reflects the explanatory variable variations only of type B people. The IV estimate will be an estimate of β_B. This may not be what we want. The bottom line here is that if individuals respond in different ways to a change in a troublesome variable, IV estimation may produce a slope estimate that reflects an atypical group's behavior.

Here are some examples to illustrate this phenomenon. Suppose we are estimating a wage equation and the troublesome variable is years of education, troublesome because a measure of ability is missing. By using distance to the nearest college as the instrument we are implicitly estimating the influence on wage of an extra year of education caused by being close to a college. If this influence on wage is the same as the influence on wage of an extra year of education motivated by other exogenous reasons, the IV and OLS estimates have the same interpretation. As another example, suppose we are estimating the determinants of violent crime across cities and are using gun magazine subscriptions as an instrument for the troublesome variable gun ownership. The IV estimate of the slope on gun ownership measures the influence on crime of those who bought guns and also bought gun magazine subscriptions. Unfortunately, the IV gun subscriptions represent gun ownership that is culturally patterned, linked with a rural hunting subculture, and so does not represent gun ownership by individuals residing in urban areas, who own guns primarily for self-protection. Consequently, the resulting IV estimate is measuring something quite different from what a researcher may want it to measure.

Viewed in this way, choice of instruments should be constrained by what it is that a researcher wants to estimate, rather than, as is typically the case in econometrics textbooks, being focused exclusively on efficient avoidance of bias.

General Notes

9.1 Introduction

- The technical notes to section 3.2 discuss at some length the implications of weakening the assumption that the explanatory variables are fixed in repeated samples to read that they are independent of the error term. The bottom line is that OLS remains best linear unbiased estimator (BLUE), but that the formula for its variance–covariance matrix requires a different interpretation.
- Binkley and Abbott (1987) note that when the regressors are stochastic many of the standard results valid in the context of fixed regressors no longer hold. For example, when regressors are stochastic omission of a relevant regressor could increase, rather than decrease, the variance of

estimates of the coefficients of remaining variables. This happens because an omitted regressor has its influence bundled into the error term, making the variance of the new, composite error term larger. This problem did not arise when the regressors were fixed in repeated samples because then the contribution of the omitted regressor to the composite error was constant in repeated samples, influencing only the mean of the error, not its variance.

9.2 The IV Estimator

- Murray (2006b) has an excellent description of several applications of IV estimation and the imaginative IVs they employed, providing references for many of the examples cited earlier. Stock and Trebbi (2003) have a very interesting discussion of the historical development of IV estimation.
- Suppose Z is a set of instruments for the regressors X. The IV residuals are calculated as $y - X\beta^{IV}$, not as $y - Z\beta^{IV}$ or as $y - \hat{X}\beta^{IV}$ as too many students think.
- In the presence of troublesome explanatory variables in small samples both OLS and IV are biased, IV presumably less so because its bias disappears asymptotically. But IV has a larger variance than OLS, in both small and large samples. Because of this, on the mean square error criterion, in large or small samples, OLS could be superior to IV. It is also possible that IV using an instrument not uncorrelated with the error could be superior to OLS. Bartels (1991) offers some rules of thumb for selecting IV versus OLS. Lee (2001) offers some suggestions for how to alleviate the finite-sample bias of two- stage least squares (2SLS). OLS is not the only alternative to IV. When instruments are weak, Murray (2006b) suggests that an estimator due to Fuller (1977) may be a better alternative.
- The Ballentine of Figure 9.2 can be used to illustrate the rationale behind the IV estimator. Suppose that Y is determined by X and an error term ε (ignore the dashed circle Z for the moment), but that X and ε are not independent. The lack of independence between X and ε means that the yellow area (representing the influence of the error term) must now overlap with the X circle. This is represented by the red area; the action from the error is represented by the red plus yellow areas. Variation in Y in the red area is due to the influence of *both* the error term and the explanatory variable X. If Y were regressed on X, the information in the red-plus-blue-plus-purple area would be used to estimate β_x. This estimate is biased because the red area does not reflect variation in Y arising solely from variation in X. Some way must be found to get rid of the red area.

The circle Z represents an IV for X. It is drawn to reflect the two properties it must possess:

1. It must be independent of the error term, so it is drawn such that it does not intersect the yellow or red areas.
2. It must be as highly correlated as possible with X, so it is drawn with a large overlap with the X circle.

Suppose X is regressed on Z. The predicted X from this regression, $\hat{X}$, is represented by the purple-plus-orange area. Now regress Y on $\hat{X}$ to produce an estimate of β_x; this in fact defines the IV estimator. The overlap of the Y circle with the purple-plus-orange area is the purple area, so information in the purple area is used to form

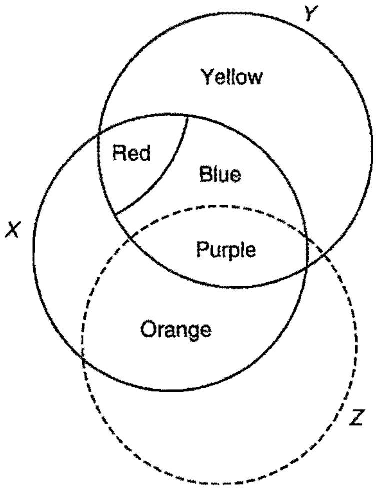

Figure 9.2 Using an instrumental variable Z.

this estimate; since the purple area corresponds to variation in Y arising entirely from variation in X, the resulting estimate of β_x is unbiased (strictly speaking, asymptotically unbiased).

Notice that, in constructing this estimate, although the bias arising from the red area is eliminated, the information set used to estimate β_x has shrunk from the red-plus-blue-plus-purple area to just the purple area. This implies that the variance of the IV estimator will be considerably higher than the variance of the OLS estimator, a reason why many researchers prefer to stick with OLS in spite of its asymptotic bias. It should now be apparent why the IV should be as highly correlated with X as possible: this makes the purple area as large as possible (at the expense of the blue area), reducing the variance of the IV estimator.

It is tempting to use the purple area by regressing Y on Z. This would produce an estimate of the "coefficient" of Z rather than the coefficient of X that is desired. Suppose, for example, that $y = \beta x + \varepsilon$ and $x = \theta z + u$. Substituting the second equation into the first gives $y = \beta\theta z + \beta u + \varepsilon$ so that regressing y on z will produce an estimate of $\beta\theta$ rather than an estimate of β. This is worth repeating: *Do not just replace a troublesome variable with an instrument and run OLS.* But you can replace a troublesome variable with its predicted value based on all the instruments and run OLS. The technical notes to this section spell this out.

- In some contexts there may be more than one IV available for a troublesome regressor. For example, as noted earlier every exogenous variable in a system of simultaneous equations is eligible to be used as an instrument for any endogenous variable serving as a regressor in any of the equations in that system. Because of this, often there will be more instruments than the minimum needed for identification. For identification purposes we must have one (different) instrument available for each of the troublesome variables in the equation we are estimating. When we have more than this bare minimum we could pick the best instruments and throw the others away, but this would be wasting information. A way must be found to exploit all the information at our disposal because by doing so we produce the most efficient estimates. Suppose both x and w are suitable instruments for p. This embarrassment of choice is resolved by using a linear combination of x and w. Since both x and w are uncorrelated with the error, any linear combination of them will be uncorrelated with the error. Since the variance of the IV estimator is smaller, the higher is the correlation of the instrument with p, we should choose the linear combination of x and w that is most highly correlated with p. This is $\hat{p}$, the predicted p obtained from regressing p on x and w. This procedure is called generalized instrumental variable estimation (GIVE): use all the available instruments to predict each of the troublesome variables in the equation being estimated, then use these predictions as instruments for the troublesome variables. Having more instruments than the bare minimum of one per troublesome variable is referred to as *overidentification.*
- Suppose we are regressing a dependent variable on two troublesome variables, $Y1$ and $Y2$, and three exogenous variables, and between them the troublesome variables have four instruments available. The IV procedure is undertaken by regressing $Y1$ and $Y2$ on the four instruments, producing predicted troublesome variables, $PY1$ and $PY2$, and then using $PY1$ and $PY2$ as instrumental variables in the IV estimating formula. Beware that, as warned above, we cannot use OLS with the troublesome variables replaced by their predictions. But, confusing to students, there is an alternative way of calculating the IV estimates that does involve replacing the troublesome variables by their predictions. This alternative method is called *two-stage least squares* (2SLS). In the first stage, $Y1$ and $Y2$ are regressed on the four instruments *and* the three exogenous variables to produce predicted troublesome variables $PY1^*$ and $PY2^*$. In the second stage the dependent variable is regressed on $PY1^*$, $PY2^*$, and the three exogenous variables, producing the same estimates that the IV formula would create. The logic here is that we are viewing all five of the explanatory variables as troublesome, and including the three exogenous variables among the instruments (for a total of seven instruments). What happens is that the three exogenous variables serving as

troublesome variables are instrumented/predicted perfectly by themselves and so are cancelled out of the troublesome category. Knowing this equivalence can avoid confusion when reading some textbook expositions, as well as the contents of the technical notes to this section. Unfortunately, as explained in the technical notes, this equivalence does not extend to estimation of variances; because of this it is wise to perform IV estimation using a software push-button.

- The discussion above suggests that it is a bit misleading to talk in terms of a specific instrument belonging to a specific troublesome variable, although that is the way in which instruments are thought about in the literature. All the instruments, including exogenous variables in the equation being estimated, contribute to the instrumentation of a troublesome variable, and the contribution of any specific instrument to this overall process reduces to any additional explanatory power it has beyond what is contributed by the other instruments. Suppose instrument Z has been included as an instrument for troublesome variable Y, and Z and Y are highly correlated. But what counts for the IV process is how much explanatory power Z has on Y after accounting for the influence of the other variables in the equation being estimated. Despite the high correlation of Z with Y, this might not be much if Z is highly correlated with these other explanatory variables.
- How many instruments should be found? This turns out to be an awkward question. On the one hand, if the number of instruments (including variables that can serve as their own instrument) is just equal to the number of troublesome variables (i.e., one instrument for each troublesome variable) β^{IV} has neither mean nor variance so we would expect it in some cases to have poor properties in finite samples. Nelson and Startz (1990a, 1990b) illustrated this dramatically. Adding an extra instrument allows it to have a mean, and one more allows it to have a variance, so it would seem desirable to have at least two more instruments than troublesome variables. On the other hand, as we add more and more instruments, in small samples the predicted troublesome variable becomes closer and closer to the actual troublesome variable and so runs the danger of introducing the bias that the IV procedure is trying to eliminate!
- What can one do when no IV is evident? Ebbes *et al.* (2005) suggest using a dummy variable separating the high and low values of the troublesome variable, or generalizations thereof. This and related procedures have been used in the context of measurement errors, as noted in chapter 10.
- Pesaran and Taylor (1999) examine a variety of diagnostics (testing for things like functional form or heteroskedasticity) in the context of IV estimation, stressing that traditional tests can be employed so long as forecasts of endogenous variables (calculated using only exogenous variables) are used in place of the endogenous variables themselves.
- Feldstein (1974) suggests forming a weighted average of OLS and IV estimators to help reduce (at the expense of some bias) the inevitably large variance associated with IV estimation. Feldstein shows his estimator to be desirable on the mean square error criterion.
- For inference (hypothesis testing) when using IV estimation, particularly in the context of weak instruments, Murray (2006b) recommends using the conditional likelihood ratio (CLR) test of Moreira (2003).

9.3 IV Issues

- Murray (2006b) and Baum, Schaffer, and Stillman (2003) have very good expositions of the IV issues discussed earlier, with references to recent literature; the latter is focused on how to use the Stata software to address these issues. Good discussions of the weak IV problem, and of the difference between asymptotic and small-sample properties of IV estimators, can be found in Bound, Jaeger, and Baker (1995), Zivot, Startz, and Nelson (1998), and Woglom (2001).
- Suppose a dependent variable y is being regressed on an exogenous variable x and troublesome variable w, for which we have an instrument z. When we regress y on x and w, the information in the collinearity between x and w (the red area in the

Ballentine in the general notes to section 3.3) is thrown away and only the remaining variation in w that it has in common with y is used to estimate the w slope parameter. Because of this, for IV estimation, the relevant correlation of z with w is its correlation with that part of w that is not collinear with x. To get at this we can regress w on z and x and look at the F test statistic for testing the slope of z against zero (i.e., after accounting for the influence of x, is z correlated with w?). In more general terms, we run the "reduced form" regression of the troublesome variable on all the instruments (including the exogenous variables that are instrumenting for themselves), and compute the F statistic for testing against zero the coefficients on the "outside" instruments. Stock and Yogo (2005) have provided special critical values that vary with the number of instruments and with the null that the bias is smaller than x% of the bias of OLS. (Tabled values of x are 10%, 15%, 20%, and 30%.) Another set of critical values is provided for the context of testing hypotheses, to address the problem of underestimated IV variances when instruments are weak. These critical values vary with the number of instruments and with the null that the 5% nominal size (type I error rate) of the test corresponds to an actual size less than x%. (Tabled values of x are 10%, 15%, 20%, and 25%.)

- Whenever there is more than one troublesome variable, the F test described above can mislead. Suppose, for example, there are two troublesome variables requiring instruments, and you have two instruments, just enough to allow IV estimation. But what if one of these instruments is highly correlated with both these explanatory variables, and the other has little or no correlation with these explanatory variables? The F test will be passed for each of these explanatory variables (thanks to the influence of the first of the instruments), but we really only have one legitimate instrument, not the two required; the second is weak but this has not been discovered. Stock and Yogo (2005) recognize this and for multiple troublesome variables provide critical values for a special F test designed to overcome this problem. Before the development of the Stock–Yogo test, partial R^2 statistics were used to address this problem; see Shea (1997) and Godfrey (1999).
- Problems interpreting IV coefficient estimates arise in the context of what are called heterogeneous response models, when not everyone reacts the same way to a change in the explanatory variable of interest. When this explanatory variable is a dummy representing some policy or "treatment," special terminology is employed. IV estimates capture the effect of treatment on the treated for those whose treatment status can be changed by the instrument at hand. In this sense, it is a sort of local effect, applicable only to these kinds of people, and so this measure is accordingly called *local average treatment effect* (LATE). It is to be distinguished from the *average treatment effect* (ATE), measuring the expected impact of the treatment on a randomly drawn person, and from the *average treatment effect on the treated* (ATET), measuring the expected impact of treatment on all those actually treated. The IV estimate is based on the behavior of those captured by the instrument and so may not reflect the behavior of others; furthermore, depending on the nature of the instrument, it may be impossible to identify any meaningful subpopulation whose behavior is being measured. Heckman, Tobias, and Vytlacil (2001) survey these measures; see also Angrist (2004) and Cobb-Clark and Crossley (2003).
- Suppose the context is our earlier example of the impact of incarceration on crime, so the treatment is releasing some prisoners from jail. The ATE is a measure of the expected impact on crime if a prisoner was chosen at random to be released. The ATET is the expected impact on crime resulting from releasing someone who was actually released. The two could differ if those actually released were not chosen randomly for release. The LATE is the expected impact on crime of releasing a prisoner because of the legal challenge that served as the instrument for incarceration. The LATE could be of interest here if we were considering a controlled release of prisoners using rules corresponding closely to those rules used for release during the legal challenges.

Technical Notes

9.1 Introduction

- Here is a crude way of seeing why only contemporaneous correlation between a regressor and the error creates asymptotic bias. Suppose $y_t = \beta y_{t-1} + \varepsilon_t$ where the intercept has been ignored for simplicity $\beta^{\text{OLS}} = \sum y_{t-1}y_t / \sum y_{t-1}^2$. Substituting for y_t we get $\beta^{\text{OLS}} = \beta + \sum y_{t-1}\varepsilon_t / \sum y_{t-1}^2$ so the bias is given by the expected value of the second term and the asymptotic bias is given by the plim of the second term. Using Slutsky's theorem (check appendix C) we can break the plim into two parts: plim $(\Sigma y_{t-1}\varepsilon_t/N)$/plim$(\Sigma y_{t-1}{}^2/N)$. The numerator is zero because y_{t-1} and ε_t are uncorrelated, so β^{OLS} is asymptotically unbiased. If y_{t-1} and ε_t were correlated (i.e., contemporaneous correlation between the regressor and the error), the numerator would not be zero and there would be asymptotic bias. When finding the expected value of $\sum y_{t-1}\varepsilon_t / \sum y_{t-1}^2$ we are dealing with small samples so that it cannot be broken into two parts. This means the εs in the numerator cannot be isolated from the εs in the denominator; we are stuck with having to find the expected value of a very awkward nonlinear function involving all the error terms. The expected value is not zero, so there is small sample bias, an expression for which is too difficult to calculate. For a more general formulation of this example replace y_{t-1} with x_t; only correlation between x_t and ε_t will cause asymptotic bias whereas correlation of x_t with any ε value will cause small sample bias.

9.2 IV Estimation

- Suppose $y = X\beta + \varepsilon$ where X contains K_1 columns of observations on exogenous variables that are uncorrelated with ε, including a column of ones for the intercept, and K_2 columns of observations on troublesome variables that are correlated with ε. The intercept and the exogenous variables can serve as their own (perfect) instruments, K_1 in number. New variables must be found to serve as instruments for the remaining explanatory variables. $K_3 \geq K_2$ such variables are required for the IV technique to work, that is, at least one instrument for each troublesome variable. This produces $K_1 + K_3$ instruments that are gathered together in a matrix Z. By regressing each column of X on Z we get $\hat{X}$ the predicted X matrix – the K_1 columns of X that are exogenous have themselves as their predictions (because variables in X that are also in Z will be predicted perfectly by the regression of X on Z!), and the K_2 columns of X that are troublesome have as predictions from this regression the best linear combination of all the possible instruments. Our earlier Ballentine discussion suggests that β^{IV} can be produced by regressing y on $\hat{X} = Z(Z'Z)^{-1}Z'X$ so that

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

- If Z is the same dimension as X, so that there is one instrument for each variable in X (exogenous variables in X would be instrumented by themselves), then algebraic manipulation of the formula above produces $\beta^{\text{IV}} = (Z'X)^{-1}Z'y$. Note that this is not (repeat, *not*) the same as $(Z'Z)^{-1}Z'y$, which is what many students want to use. (This same warning, with an example, was given at the end of the Ballentine discussion in the general notes.) This IV formula can also be derived as a method of moments estimator, using the moment conditions $EZ'\varepsilon = EZ'(y - X\beta) = 0$, just as the OLS estimator can be derived as a method of moments estimator using the moment conditions $EX'\varepsilon = 0$.
- When Z has more columns than X because there are more than exactly enough instruments, the moments $EZ'\varepsilon = 0$ are too numerous and so the GMM (generalized method of moments – see section 8.5) estimator must be used. This requires minimizing (with respect to β)

$$\begin{aligned}&(Z'\varepsilon)'[V(Z'\varepsilon)]^{-1}Z'\varepsilon \\ &\quad = (y - X\beta)'Z(Z'Z)^{-1}Z'(y - X\beta)/\sigma^2\end{aligned}$$

because $V(Z'\varepsilon) = Z'V(\varepsilon)Z = \sigma^2 Z'Z$.
This minimization produces exactly the IV formula given earlier.

- The preceding result suggests that when $y = f(X,\beta) + \varepsilon$ where f is a nonlinear function, the IV estimator can be found by minimizing

$$[y - f(X,\beta)]'Z(Z'Z)^{-1}Z'[y - f(X,\beta)]/\sigma^2.$$

Following Amemiya (1974), this is sometimes called nonlinear 2SLS, because if f is linear the estimator coincides with the 2SLS method of chapter 11. The choice of instruments is not clear here, as it is in the linear case, because the connection between instruments and explanatory variables may itself be nonlinear. Suppose a regressor takes the form $g(x)$ where g is a nonlinear function of the variable x (i.e., the regressor is ln$x$ or x^2, for example). We have an instrument z for x, but we need an instrument for $g(x)$. An obvious procedure would be to regress x on z to get $\hat{x}$ and then use $g(\hat{x})$ as the instrument for $g(x)$. Unfortunately, this produces an inconsistent estimator; we should instead regress $g(x)$ on z to get $\hat{g}$ to use as the instrument for $g(x)$.

- The variance–covariance matrix of β^{IV} is estimated by

$$\hat{\sigma}^2(\hat{X}'\hat{X})^{-1} = \hat{\sigma}^2[X'Z(Z'Z)^{-1}Z'X]^{-1}$$

which, when Z and X are of the same dimension, is written as

$$\hat{\sigma}^2(Z'X)^{-1}Z'Z(X'Z)^{-1}$$

It is tempting to estimate σ^2 by

$$s^2 = (y - \hat{X}\beta^{IV})'(y - \hat{X}\beta^{IV})/(N - K)$$

where K is the number of regressors. This is incorrect, however, because it is $y - X\beta^{IV}$ which estimates ε, not $y - \hat{X}\beta^{IV}$. Consequently, σ^2 is estimated using

$$\hat{\sigma}^2 = (y - X\beta^{IV})'(y - X\beta^{IV})/(N - K)$$

This has an important implication for F tests using the regression of y on $\hat{X}$. The numerator can continue to be the restricted minus unrestricted sums of squares divided by the number of restrictions, but now the denominator must be $\hat{\sigma}^2$ rather than s^2.

- An observant reader may have noticed that the discussion above was carefully worded so as not to give the formula for $V(\beta^{IV})$, only the formula used to estimate it. This is because with stochastic explanatory variables (the context in which IV is employed) the actual variance is too difficult to calculate. Instead the formula for the asymptotic variance is used. The usual way of calculating the asymptotic variance, as explained in appendix C, is as the inverse of the sample size N times the limit as N goes to infinity of N times the variance. For a spherical error this creates an asymptotic variance equal to the inverse of N times $\sigma^2[\Sigma_{X'Z}(\Sigma_{Z'Z})^{-1}\Sigma_{Z'X}]^{-1}$ where $\Sigma_{X'Z}$ is plim$(X'Z/N)$, $\Sigma_{Z'Z}$ is plim$(Z'Z/N)$, and $\Sigma_{Z'X}$ is plim$(Z'X/N)$. Remember plims from appendix C? Aren't you glad you asked? Throughout this book, for expositional reasons, results relating to stochastic explanatory variables, that technically should be qualified as holding only asymptotically, are presented as though they represent small-sample behavior.

- When estimating with IV, how should we create a heteroskedasticity-consistent estimate of its variance–covariance matrix? Following the logic of the presentation of the heteroskedasticity-consistent variance–covariance matrix in the technical notes to section 8.2, we can see that if $V(\varepsilon) = \sigma^2\Omega$ was known, we should estimate $V(\beta^{IV})$ by $(\hat{X}'\hat{X})^{-1}(\hat{X}'\sigma^2\Omega\hat{X})(\hat{X}'\hat{X})^{-1}$, with $\sigma^2\Omega$ replaced by a diagonal matrix with the squared IV residuals down the diagonal. The logic of this matches the logic used for OLS: We continue to use the traditional IV estimator but estimate its variance–covariance matrix with a different formula.

  An alternative way of proceeding is typically employed, however. By using the GMM approach to creating an IV estimate it turns out that we are able in the face of unknown heteroskedasticity both to improve upon traditional IV and to produce an appropriate estimator for the variance–covariance matrix of this improved estimator. From the technical notes to section 8.5 we saw that the GMM estimator minimizes $(y-X\beta)'Z(Z'\Omega Z)^{-1}Z'(y-X\beta)$. If we replace Ω with a diagonal matrix containing the squared residuals (from a preliminary consistent estimation procedure) down the diagonal, minimizing this expression creates an IV estimate that has been to some extent adjusted for the unknown heteroskedasticity, and has a variance–covariance estimate robust to that heteroskedasticity. It is a useful exercise to do the basic derivations here, to see the reason for the claims of GMM superiority.

- In the generalized least squares (GLS) model when $V(\varepsilon)=\sigma^2\Omega$, so that the appropriate moments are $Z'\Omega^{-1}\varepsilon$, if Ω is known the IV estimator can be found by minimizing

$$(y-X\beta)'\Omega^{-1}Z(Z'\Omega Z)^{-1}Z'\Omega^{-1}(y-X\beta)/\sigma^2$$

It is a useful exercise to do this minimization to discover that the IV formula in the presence of a nonspherical error of known form is not given by $(\hat{X}'\Omega^{-1}\hat{X})^{-1}\hat{X}'\Omega^{-1}y$ as one might have guessed.

- Dealing with autocorrelated errors in an IV context is a straightforward modification of procedures discussed in chapter 8. To test for a first-order autocorrelated error get the IV residuals, reestimate the equation (using IV) with the lagged residual as an extra explanatory variable, and test its coefficient against zero with a t test. To do EGLS, quasi-difference the data, and perform IV estimation using the quasi-differenced instrument as the IV.

9.3 IV Issues

How can we test if errors are correlated with regressors?

- The Hausman test appears in two forms. (Most of what follows rests on asymptotic arguments that are suppressed for expository purposes.) Suppose $Y=X\beta+\varepsilon$ and W is a set of instruments for X. Then $\beta^{IV}=(W'X)^{-1}W'Y=(W'X)^{-1}W'(X\beta^{OLS}+\varepsilon^{OLS})=\beta^{OLS}+(W'X)^{-1}W'\varepsilon^{OLS}$ so that $\beta^{IV}-\beta^{OLS}=(W'X)^{-1}W'\varepsilon^{OLS}$. Straightforward algebra on this yields $V(\beta^{IV}-\beta^{OLS})=V(\beta^{IV})-V(\beta^{OLS})$. (Intuitively, this comes about because the correlation between an efficient estimator and the difference between that efficient estimator and an inefficient estimator is zero – if not, the efficient estimator could be made more efficient by exploiting this correlation!) This suggests that a test of equality between β^{IV} and β^{OLS} could be formed by using the statistic

$$(\beta^{IV}-\beta^{OLS})'[V(\beta^{IV})-V(\beta^{OLS})]^{-1}(\beta^{IV}-\beta^{OLS})$$

which is distributed as a chi-square with degrees of freedom equal to the number of elements in β. This is the original form of the Hausman test.

Unfortunately, there are two problems with this form of the Hausman test, one theoretical, the other practical. First, whenever X and W overlap, as would normally be the case, $[V(\beta^{IV})-V(\beta^{OLS})]$ cannot be inverted in the normal way. In this case, we should really only be comparing the OLS and IV coefficient estimates of the troublesome regressor, rather than the full vector of coefficient estimates, and so should be using only the invertible part of $[V(\beta^{IV})-V(\beta^{OLS})]$. Second, the estimated $[V(\beta^{IV})-V(\beta^{OLS})]$ often turns out to have incorrect signs (although in theory $V(\beta^{OLS})$ is "smaller" than $V(\beta^{IV})$, their estimates may not preserve this result). Both these problems are avoided with the second variant of the Hausman test.

From above we have that $\beta^{IV}-\beta^{OLS}=(W'X)^{-1}W'\varepsilon^{OLS}$. This will be zero if W and ε^{OLS} are uncorrelated, which suggests testing if W and ε^{OLS} are uncorrelated. This can be done by running the regression: $Y=X\beta+W\theta+\varepsilon$ and testing $\theta=0$ with an F test. The intuition behind this is straightforward. Without W the regression would produce residuals ε^{OLS}. If W is to have a nonzero coefficient, it will have to "steal" some explanatory power from ε^{OLS}. (Try drawing a Ballentine to see this.) So if W has a nonzero coefficient, it must be the case that W and ε^{OLS} are correlated. Thus a test of $\theta=0$ is a test of W and $\varepsilon^{\mathrm{OLS}}$ being correlated which in turn is a test of $\beta^{\mathrm{IV}} - \beta^{\mathrm{OLS}} = 0$, which in turn is a test of contemporaneous correlation between the error and the regressors.

This is called the OV, or omitted variables, version of the Hausman test. It is computationally attractive, and there is no problem in figuring out the degrees of freedom because to run the OV regression W will have to be stripped of any variables that are serving as their own instruments (i.e., to avoid perfect multicollinearity if X and W have some elements in common). In the general case when W contains more variables than X, in this OV regression W is replaced by $\hat{X}$, the explained X from regressing X on W. To be precise here, regress each explanatory variable that cannot serve as its own instrument on all the instruments, calculate the predicted values of these explanatory variables from these regressions, add these predictions as extra explanatory variables in the regression of y on X, and do an F test of the null that the slopes of these predicted values are all zero.

- An algebraically equivalent form of the OV version of this test uses the residuals from a regression of X on W in place of W. This also has a good intuitive explanation. Think of X as having two parts, one part explained by W, and the other part the residuals from explaining X by W. If X and ε are correlated only the second part is correlated with ε (because we know that W is not correlated with ε). We can test if this second part is correlated with ε by seeing if it can steal some explanatory power from ε when it is added as an extra set of explanatory variables to the regression of y on X. To be precise here, regress each explanatory variable that cannot serve as its own instrument on all the instruments, calculate the residuals from these regressions, add these residuals as extra explanatory variables in the regression of y on X, and do an F test of the null that the slopes of these residuals are all zero.

  The algebraic equivalence of these two versions of the OV variant of the Hausman test is easy to see. In the first version, the regression equation is $y = X\beta + \hat{X}\theta + \varepsilon$. In the second version the regression equation is $y = X\beta + (X - \hat{X})\varphi + \varepsilon$ which can be rewritten as $y = X(\beta + \varphi) - \hat{X}\varphi + \varepsilon$.
- Yet another form of the OV variant of the Hausman test is also based on testing if W and $\varepsilon^{\mathrm{OLS}}$ are uncorrelated. Run the OLS regression and obtain the residuals. Then regress the residuals on X and W and test the slopes on W against zero. W will have to be stripped of its overlap with X to do this. The usual test statistic employed here is NR^2, distributed as a chi-square with degrees of freedom equal to the number of troublesome variables.
- The Hausman test becomes more complicated if some explanatory variables are known to be troublesome and we want to test if some additional explanatory variables are correlated with the error. Suppose $y = Y_1\delta_1 + Y_2\delta_2 + X\beta + \varepsilon$ and it is desired to test Y_2 for exogeneity, knowing that Y_1 is endogenous. This case is different from those examined earlier because rather than comparing OLS to IV, we are now comparing one IV to another IV. Spencer and Berk (1981) show that a regular Hausman test can be structured, to compare the 2SLS estimates with and without assuming Y_2 to be exogenous. An OV form of this test is also available, and defended on asymptotic grounds. Estimate the original equation by 2SLS assuming Y_2 to be exogenous, but add in the extra regressors Y_1^* and Y_2^*, the predicted (from the instruments) values of Y_1 and Y_2 that would be used if Y_2 were assumed to be endogenous. Test the coefficients of Y_1^* and Y_2^* jointly against zero. This test is tricky; see Davidson and MacKinnon (1993, chapter 7).
- Because the Hausman test is sensitive to several types of misspecification, Godfrey and Hutton (1994) recommend testing for general misspecification before applying the Hausman test, and recommend a test for doing so. Wong (1996) finds that bootstrapping the Hausman test improves its performance.

How can we test if an instrument is uncorrelated with the error?

- The Sargan test is used to test if overidentifying instruments are uncorrelated with the error.

The rationale behind this test was presented earlier. To calculate this test regress the IV residuals on all the instruments plus the exogenous variables (including the constant). The sample size N times the uncentered R^2 from this regression is distributed as a chi-square with degrees of freedom equal to the number of overidentifying instruments, namely the difference between the number of instruments and the number of troublesome variables. (The uncentered R^2 is $1 - \Sigma e^2/\Sigma y^2$ instead of $1 - \Sigma e^2/\Sigma(y - \bar{y})^2$.) This is in essence equivalent to an F test for zero slopes on the instruments in this regression. Wooldridge (2002, p.123) explains how to perform a heteroskedasticity-robust version of this test.

- A frustrating thing about the Sargan test is that if the test rejects the null of no correlation between the overidentifying instruments and the errors, we know that at least one instrument is correlated with the error but we do not know which one(s). Of some help in this regard is the difference-in-Sargan test, sometimes called the C test. This test tests the null that a subset of the overidentifying restrictions is uncorrelated with the error. It is calculated as the difference between the Sargan statistic for testing the validity of all the overidentifying instruments and the Sargan statistic for testing the validity of a smaller set of these instruments. If the dropped instruments are uncorrelated with the error, the Sargan statistic should not change much. This difference is distributed as a chi-square with degrees of freedom equal to the number of dropped instruments.
- As noted earlier, and in section 8.5, IV estimation can be undertaken via GMM. Suppose there are two troublesome variables and three exogenous variables in the equation being estimated. If we have two instruments, one for each of the troublesome variables, we have exact identification. In this case there are six moment conditions: the average of the errors times each instrument value equal to zero, the average of the errors times each exogenous variable equal to zero, and the average of the errors equal to zero (for the intercept). These six moment conditions can be solved to produce the IV estimates. If there are overidentifying instruments there are more moment conditions than parameters to be estimated and so not all the moment conditions can be set equal to zero, as discussed in section 8.5. This leads to GMM: choose the IV estimates to minimize the extent to which these moment conditions are collectively violated. With just enough instruments the moment conditions can all be set equal to zero, so the extent to which they are collectively violated is zero; the minimand is zero. Adding extra moments (i.e., extra instruments) causes these conditions to no longer be satisfied in the data, causing the minimand to be nonzero. If this violation is large, as measured by the magnitude of the minimand, then at least one of the moment conditions must be false. As noted in the technical notes to section 8.5, this minimand is in exactly the form of a chi-square statistic (called Hansen's J test) for testing the null that the overidentifying moment conditions are true; its degrees of freedom is the number of overidentifying instruments. This way of thinking about testing for the instruments being uncorrelated with the errors makes clearer the reason why we can only test for overidentifying restrictions and why we do not know which instrument is guilty if we reject the null.
- Stock and Watson (2007, pp. 443–5) suggest a different way of thinking about the Sargan test. If there is exact identification with exogenous instruments then a legitimate IV estimate is produced. Adding extra instruments produces a different IV estimate. If the two IV estimates are quite different from one another, we should be suspicious that one or more of these extra instruments is not exogenous. The Sargan test is implicitly making this comparison. When there is exact identification no comparison is possible because a different IV estimate cannot be calculated.
- A difference-in-J test, comparable to the difference-in-Sargan test, is available. Suppose you are confident that an identifying set of instruments is legitimate, and wish to test the legitimacy of an additional K questionable instruments, then the difference between the J statistics with and without this additional set of instruments

is distributed as a chi-square with K degrees of freedom.

- The Sargan test is identical to Hansen's J test whenever the errors are assumed to be spherical. When nonspherical errors are assumed, for example, when there is heteroskedasticity of unknown form, Hansen's J test is more popular. As noted earlier, GMM estimation in this context introduces more efficiency, as well as robustifying the variance–covariance estimate.