Wooldridge, Introductory Econometrics, 2d ed.

Chapter 15: Instrumental variables and two stage least squares

Many economic models involve endogeneity: that is, a theoretical relationship does not fit into the framework of $y-$on$-X$ regression, in which we can assume that the y variable is determined by (but does not jointly determine) X. Indeed, the simplest analytical concepts we teach in principles of economics–a demand curve in micro, and the Keynesian consumption function in macro–are relations of this sort, where at least one of the "explanatory" variables is endogenous, or jointly determined with the "dependent" variable. From a mathematical standpoint, the difficulties that this endogeneity cause for econometric analysis are identical to those which we have already considered, in two contexts: that of omitted variables, and that of errors-in-variables, or measurement error in the X variables. In each of these three cases, OLS is not capable of delivering consistent parameter estimates. We now turn to a general solution to the problem of endogenous regressors, which as we will see can also be profitably applied in other contexts, in which the omitted variable (or poorly measured variable) can be taken into account. The general concept is that of the **instrumental variables** estimator; a popular form of that estimator, often employed in the context of endogeneity, is known as **two-stage least squares (2SLS)**.

To motivate the problem, let us consider the omitted-variable

problem: for instance, a wage equation, which would be correctly specified as:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 abil + e \tag{1}$$

This equation cannot be estimated, because ability ($abil$) is not observed. If we had a proxy variable available, we could substitute it for $abil$; the quality of that equation would then depend on the degree to which it was a good proxy. If we merely ignore $abil$, it becomes part of the error term in the specification:

$$\log(wage) = \beta_0 + \beta_1 educ + u \tag{2}$$

If $abil$ and $educ$ are correlated, OLS will yield biased and inconsistent estimates. To consistently estimate this equation, we must find an **instrumental variable**: a new variable that satisfies certain properties. Imagine that variable z is uncorrelated with u, but is correlated with $educ$. A variable that meets those two conditions is an instrumental variable for $educ$. We cannot (ever) test the prior assumption, since we cannot observe u; but we can readily test the latter assumption, and should do so, by merely regressing the included explanatory variable on the instrument:

$$educ = \pi_0 + \pi_1 z + v \tag{3}$$

In this regression, we should easily reject $H_0 : \pi_1 = 0$. It should be clear that there is no unique choice of an instrument in this situation; many potential variables could meet these two conditions, of being uncorrelated with the unobservable factors influencing the wage (including $abil$) and correlated with $educ$. Note that in this context we are not searching for a proxy variable for $abil$; if we had a good proxy for $abil$, it would

not make a satisfactory instrumental variable, since correlation with $abil$ implies correlation with the error process u. What might serve in this context? Perhaps something like the mother's level of education, or the number of siblings, would make a sensible instrument. If we determine that we have a reasonable instrument, how may it be used?

Return to the misspecified equation (2), and write it in general terms of y and x :

$$y = \beta_0 + \beta_1 x + u \tag{4}$$

If we now take the covariance of each term in the equation with our instrument z, we find:

$$Cov(y, z) = \beta_1 Cov(x, z) + Cov(u, z) \tag{5}$$

We have made use of the fact that the covariance with a constant is zero. Since by assumption the instrument is uncorrelated with the error process u, the last term has expectation zero, and we may solve (5) for our estimate of β_1 :

$$b_1 = \frac{Cov(y, z)}{Cov(x, z)} = \frac{\sum (y_i - \bar{y})(z_i - \bar{z})}{\sum (x_i - \bar{x})(z_i - \bar{z})} \tag{6}$$

Note that this estimator has an interesting special case where $x = z$: that is, where an explanatory variable may serve as its own instrument, which would be appropriate if $Cov(x, u) = 0$. In that case, this estimator may be seen to be the OLS estimator of β_1. Thus, we may consider OLS as a special case of IV, usable when the assumption of exogeneity of

the x variable(s) may be made. We may also note that the IV estimator is consistent, as long as the two key assumptions about the instrument's properties are satisfied. The IV estimator is not an unbiased estimator, though, and in small samples its bias may be substantial.

Inference with the IV estimator

To carry out inference–compute interval estimates and hypothesis tests–we assume that the error process is homoskedastic: in this case, conditional on the instrumental variable z, not the included explanatory variable x. With this additional assumption, we may derive the asymptotic variance of the IV estimator as:

$$Var(b_1) = \frac{\sigma^2}{SST_x \rho_{xz}^2} \tag{7}$$

where n is the sample size, SST_x is the total sum of squares of the explanatory variable, and ρ_{xz}^2 is the R^2 (or squared correlation) in a regression of x on z : that is, equation (3). This quantity can be consistently estimated; σ^2 from the regression residuals, just as with OLS. Notice that as the correlation between the explanatory variable x and the instrument z increases, ceteris paribus, the sampling variance of b_1 decreases. Thus, an instrumental variables estimate generated from a "better" instrument will be more precise (conditional, of course, on the instrument having zero correlation with u). Note as well that this estimated variance must exceed that of the OLS estimator of b_1, since $0 \leq \rho_{xz}^2 \leq 1$. In the case where an explanatory variable may serve as its own instrument, the squared correlation

is unity. The IV estimator will always have a larger asymptotic variance than will the OLS estimator, but that merely reflects the introduction of an additional source of uncertainty (in the form of the instrument, imperfectly correlated with the explanatory variable).

What will happen if we use the instrumental variables with a "poor" instrument? A weak correlation between x and z will bring a sizable bias in the estimator. If there is any correlation between z and u, a weak correlation between x and z will render IV estimates inconsistent. Although we cannot observe the correlation between z and u, we can empirically evaluate the correlation between the explanatory variable and its instrument, and should always do so.

It should also be noted that an R^2 measure in the context of the IV estimator is not the "percentage of variation explained" measure that we are familiar with in OLS terms. In the presence of correlation between x and u, we can no longer decompose the variation in y into two independent components, SSE and SSR, and R^2 has no natural interpretation. In the OLS context, a joint hypothesis test can be written in terms of R^2 measures; that cannot be done in the IV context. Just as the asymptotic variance of an IV estimator exceeds that of OLS, the R^2 measure from IV will never beat that which may be calculated from OLS. If we wanted to maximize R^2, we would just use OLS; but when OLS is biased and inconsistent, we seek an estimation technique that will focus on providing consistent estimates of the regression parameters, and not mechanically find the "least

squares" solution in terms of inconsistent parameter estimates.

IV estimates in the multiple regression context

The instrumental variables technique illustrated above can readily be extended to the case of multiple regression. To introduce some notation, consider a **structural equation**:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1 \tag{8}$$

where we have suppressed the observation subscripts. The y variables are **endogenous**; the z variable is **exogenous**. The endogenous nature of y_2 implies that if this equation is estimated by OLS, the point estimates will be biased and inconsistent, since the error term will be correlated with y_2. We need an instrument for y_2 : a variable that is correlated with y_2, but not correlated with u. Let us write the endogenous explanatory variable in terms of the exogenous variables, including the instrument z_2 :

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v \tag{9}$$

The key identification condition is that $\pi_2 \neq 0$; that is, after partialling out z_1, y_2 and z_2 are still meaningfully correlated. This can readily be tested by estimating the auxiliary regression (9). We cannot test the other crucial assumption: that in this context, $cov(z_2, v) = 0$. Given the satisfaction of these assumptions, we may then derive the instrumental variables estimator of (8) by writing down the "normal equations" for the least squares problem, and solving them for the point estimates. In this context, z_1 serves as an instrument for itself.

We can extend this logic to include any number of additional exogenous variables in the equation; the condition that the analogue to (9) must have $\pi_2 \neq 0$ always applies. Likewise, we

could imagine an equation with additional endogenous variables; for each additional endogenous variable on the right hand side, we would have to find another appropriate instrument, which would have to meet the two conditions specified above.

Two stage least squares (2SLS)

What if we have a single endogenous explanatory variable, as in equation (8), but have more than one potential instrument? There might be several variables available, each of which would have a significant coefficient in an equation like (9), and could be considered uncorrelated with u. Depending on which of the potential instruments we employ, we will derive different IV estimates, with differing degrees of precision. This is not a very attractive possibility, since it suggests that depending on how we implement the IV estimator, we might reach different qualitatitive conclusions about the structural model. The technique of two-stage least squares (**2SLS**) has been developed to deal with this problem. How might we combine several instruments to produce the single instrument needed to implement IV for equation (8)? Naturally, by running a regression–in this case, an auxiliary regression of the form of equation (9), with all of the available instruments included as explanatory variables. The predicted values of that regression, $\hat{y}_2$, will serve as the instrument for y_2, and this auxiliary regression is the "first stage" of 2SLS. In the "second stage," we use the IV estimator, making use of the generated instrument $\hat{y}_2$. The IV estimator we developed above can be shown, algebraically, to be a 2SLS estimator; but although the IV estimator becomes non-unique in the presence of multiple

instruments, the 2SLS estimation technique will always yield a unique set of parameter values for a given instrument list.

Although from a pedagogical standpoint we speak of the two stages, we should not actually perform 2SLS "by hand." Why? Because the second stage will yield the "wrong" residuals (being computed from the instruments rather than the original variables), which implies that all statistics computed from those residuals will be incorrect (the estimate s^2, the estimated standard errors of the parameters, etc.) We should make use of a computer program that has a command to perform 2SLS (or, as some programs term it, instrumental variables). In Stata, you use the `ivreg` command to perform either IV or 2SLS estimation. The syntax of `ivreg` is:

```
ivreg depvar [varlist1]
(varlist2=varlist_iv)
```

where `depvar` is the dependent variable; `varlist1`, which may not be present, is the list of included exogenous variables (such as z_1 in equation (8); `varlist2` contains the included endogenous variables (such as y_2 in equation (8); and `varlist_iv` contains the list of instruments that are not included in the equation, but will be used to form the instrumental variables estimator. If we wanted to estimate equation (8) with Stata, we would give the command `ivreg y1 z1 (y2 = z2)`. If we had additional exogenous variables in the equation, they would follow `z1`. If we had additional instruments (and were thus performing 2SLS), we would list them after `z2`.

The 2SLS estimator may be applied to a much more complex

model, in which there are multiple endogenous explanatory variables (which would be listed after `y2` in the command), as well as any number of instruments and included exogenous variables. The constraint that must always be satisfied is related to the parenthesized lists: the **order condition for identification**. Intuitively, it states that for each included endogenous variable (e.g. `y2`), we must have at least one instrument—that is, one exogenous variable that does not itself appear in the equation, or satisfies an **exclusion restriction**. If there are three included endogenous variables, then we must have no fewer than three instruments after the equals sign, or the equation will not be **identified**. That is, it will not be possible to solve for a unique solution in terms of the instrumental variables estimator. In the case (such as the example above) where the number of included endogenous variables exactly equals the number of excluded exogenous variables, we satisfy the order condition with equality, and the standard IV estimator will yield a solution. Where we have more instruments than needed, we satisfy the order condition with inequality, and the 2SLS form of the estimator must be used to derive unique estimates, since we have more equations than unknowns: the equation is **overidentified**. If we have fewer instruments than needed, we fail the order condition, since there are more unknowns than equations. No econometric technique can solve this problem of **underidentification**. There are additional conditions for identification–the order condition is necessary, but not sufficient–since it must also be the case that each instrument has a nonzero partial correlation with the

dependent variable. This would fail, for instance, if one of our candidate instruments was actually a linear combination of the included exogenous variables.

IV and errors-in-variables

The instrumental variables estimator can also be used fruitfully to deal with the errors-in-variables model discussed earlier–not surprisingly, since the econometric difficulties caused by errors-in-variables are mathematically the same problem as that of an endogenous explanatory variable. To deal with errors-in-variables, we need an instrument for the mismeasured x variable that satisfies the usual assumptions: being well correlated with x, but not correlated with the error process. If we could find a second measurement of $x-$even one also subject to measurement error–we could use it as an instrument, since it would presumably be well correlated with x itself, but if generated by an independent measurement process, uncorrelated with the original $x's$ measurement error. Thus, we might conduct a household survey which inquires about disposable income, consumption, and saving. The respondents' answers about their saving last year might well be mismeasured, since it is much harder to track saving than, say, earned income. The same could be said for their estimates of how much they spent on various categories of consumption. But using income and consumption data, we could derive a second (mismeasured) estimate of saving, and use it as an instrument to mitigate the problems of measurement error in the direct estimate.

IV may also be used to solve proxy problems; imagine that

we are regressing $\log(wage)$ on education and experience, using a theoretical model that suggests that "ability" should appear as a regressor. Since we do not have a measure of ability, we use a test score as a proxy variable. That may introduce a problem, though, since the measurement error in the relation of test score to ability will cause the test score to be correlated with the error term. This might be dealt with if we had a second test score measure–on a different aptitude test–which could then be used as an instrument. The two test scores are likely to be correlated, and the measurement error in the first (the degree that it fails to measure ability) should not be correlated with the second score.

Tests for endogeneity and overidentifying restrictions

Since the use of IV will necessarily inflate the variances of the estimators, and weaken our ability to make inferences from our estimates, we might be concerned about the need to apply IV (or 2SLS) in a particular equation. One form of a test for endogeneity can be readily performed in this context. Imagine that we have the equation:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1 \tag{10}$$

where y_2 is the single endogenous explanatory variable, and the $z's$ are included exogenous variables. Imagine that the equation is overidentified for IV: that is, we have at least two instruments (in this case, z_3 and z_4) which could be used to estimate (10) via 2SLS. If we performed 2SLS, we would be estimating the following **reduced form** equation in the "first stage":

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + v \tag{11}$$

which would allow us to compute OLS residuals, $\hat{v}$. Those residuals will be that part of y_2 not correlated with the $z's$. If there is a problem of endogeneity of y_2 in equation (10), it will occur because $cov(v, u_1) \neq 0$. We cannot observe v, but we can calculate a consistent estimate of v as $\hat{v}$. Including $\hat{v}$ as an additional regressor in the OLS model

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta \hat{v} + \omega \tag{12}$$

and testing for the significance of δ will give us the answer. If $cov(v, u_1) = 0$, our estimate of δ should not be significantly different from zero. If that is the case, then there is no evidence that y_2 is endogenous in the original equation, and OLS may be applied. If we reject the hypothesis that $\delta = 0$, we should not rely on OLS, but should rather use IV (or 2SLS). This test may also be generalized for the presence of multiple included endogenous variables in (10); the relevant test is then an $F-$test, jointly testing that a set of δ coefficients are all zero. This test is available within Stata as the `ivendog` command (Baum et al., 2003), which may be installed via `ssc install ivendog`. It must be used after an IV regression (that is, following `ivreg`) since it must be able to construct the list of variables to be included (e.g. equation (11) for each included endogenous variable. See `help ivendog` for more details.

Although we can never directly test the maintained hypothesis that the instruments are uncorrelated with the error process

u, we can derive indirect evidence on the suitability of the instruments if we have an excess of instruments–that is, if the equation is overidentified, so that we are using 2SLS. The ivreg residuals may be regressed on all exogenous variables (included exogenous variables plus instruments). Under the null hypothesis that all IV's are uncorrelated with u, a Lagrange multiplier statistic of the nR^2 form will not exceed the critical point on a $\chi^2(r)$ distribution, where r is the number of **overidentifying restrictions** (i.e. the number of excess instruments). If we reject this hypothesis, then we cast doubt on the suitability of the instruments; at least some of them do not appear to be satisfying the condition of orthogonality with the error process. This test is available within Stata as the `overid` command (Baum, Wiggins and Stillman, 2001), which may be installed via `ssc install overid`. It must be used after an IV regression. See `help overid` for more details.

Applying 2SLS in a time series context

When there are concerns of included endogenous variables in a model fit to time series data, we have a natural source of instruments in terms of **predetermined** variables. For instance, if y_{2t} is an explanatory variable, its own lagged values, y_{2t-1}or y_{2t-2} might be used as instruments: they are likely to be correlated with y_{2t}, and they will not be correlated with the error term at time t, since they were generated at an earlier point in time. The one caveat that must be raised in this context relates to autocorrelated errors: if the errors are themselves autocorrelated, then the presumed exogeneity of predetermined variables will be

in doubt. Tests for autocorrelated errors should be conducted; in the presence of autocorrelation, more distant lags might be used to mitigate this concern.