CHAPTER 4. INSTRUMENTAL VARIABLES

1. INTRODUCTION

Consider the linear model $y = X\beta + \varepsilon$, where y is n×1, X is n×k, β is k×1, and ε is n×1. Suppose that *contamination* of X, where some of the X variables are correlated with ε, is suspected. This can occur, for example, if ε contains omitted variables that are correlated with the included variables, if X contains measurement errors, or if X contains endogenous variables that are determined jointly with y.

OLS Revisited: Premultiply the regression equation by X′ to get

$$X'y = X'X\beta + X'\varepsilon. \quad (1)$$

One can interpret the OLS estimate b_{OLS} as the solution obtained from (1) by first approximating $X'\varepsilon$ by zero, and then solving the resulting k equations in k unknowns,

$$X'y = X'Xb_{OLS}, \quad (2)$$

for the unknown coefficients. Subtracting (1) from (2), one obtains the condition

$$X'X(b_{OLS} - \beta) = X'\varepsilon, \quad (3)$$

and the error in estimating β is linear in the error caused by approximating $X'\varepsilon$ by zero. If $X'X/n \to_p A$ positive definite and $X'\varepsilon/n \to_p 0$, (3) implies the result that $b_{OLS} \to_p \beta$. What makes OLS consistent when $X'\varepsilon/n \to_p 0$ is that approximating $X'\varepsilon$ by zero is reasonably accurate in large samples. On the other hand, if one has instead $X'\varepsilon/n \to_p C \neq 0$, then b_{OLS} is not consistent for β, and instead $b_{OLS} \to_p \beta + A^{-1}C$.

Instrumental Variables: Suppose there is a n×j array of variables W, called *instruments*, that have two properties: (i) These variables are uncorrelated with ε; we say in this case that these instruments are *clean*. (ii) The matrix of correlations between the variables in X and the variables in W is of maximum possible rank (= k); we say in this case that these instruments are *fully correlated*. Call the instruments *proper* if they satisfy (i) and (ii). The W array should include any variables from X that are themselves clean. To be fully correlated, W must include at least as many variables as are in X, so that $j \geq k$. Another way of stating this necessary condition is that *the number of instruments in* W *that are excluded from* X *must be at least as large as the number of contaminated variables that are included in* X.

Instead of premultiplying the regression equation by X′ as we did for OLS, premultiply it by R′W′, where R is a j×k weighting matrix that we get to choose. (For example, R might select a subset of k from the j instrumental variables, or might form k linear combinations of these variables. The only restriction is that R must have rank k.) This gives

(4) $$R'W'y = R'W'X\beta + R'W'\varepsilon.$$

The idea of an instrumental variables (IV) estimator of β is to approximate $R'W'\varepsilon$ by zero, and solve

(5) $$R'W'y = R'W'X\,b_{IV}$$

for $b_{IV} = [R'W'X]^{-1}R'W'y$. Subtract (4) from (5) to get the IV analog of the OLS relationship (3),

(6) $$R'W'X(b_{IV} - \beta) = R'W'\varepsilon.$$

If $R'W'X/n$ converges in probability to a nonsingular matrix and $R'W'\varepsilon/n \to_p 0$, then $b_{IV} \to_p \beta$. Thus, in problems where OLS breaks down due to correlation of right-hand-side variables and the disturbances, you can use IV to get consistent estimates, provided you can find proper instruments.

The idea behind (5) is that W and ε are orthogonal in the population, a generalized moment condition. Then, (5) can be interpreted as the solution of a generalized method of moments problem, based on the sample moments $W'(y - X\beta)$. The properties of the IV estimator could be deduced as a special case of the general theory of GMM estimators. However, because the linear IV model is such an important application in economics, we will give IV estimators an elementary self-contained treatment, and only at the end make connections back to the general GMM theory.

2. OPTIMAL IV ESTIMATORS

If there are exactly as many instruments as there are explanatory variables, $j = k$, then the IV estimator is uniquely determined, $b_{IV} = (W'X)^{-1}W'y$, and R is irrelevant. However, if $j > k$, each R determines a different IV estimator. What is the best way to choose R? An analogy to the generalized least squares problem provides an answer: Premultiplying the regression equation by W' yields a system of $j > k$ equations in k unknown β's, $W'y = W'X\beta + W'\varepsilon$. Since there are more equations than unknowns, we cannot simply approximate all the $W'\varepsilon$ terms by zero simultaneously, but will have to accommodate at least j-k non-zero residuals. But this is just like a regression problem, with j observations, k explanatory variables, and disturbances $\nu = W'\varepsilon$. Suppose the disturbances ε have a covariance matrix $\sigma^2\Omega$, and hence the disturbances $\nu = W'\varepsilon$ have a non-scalar covariance matrix $\sigma^2W'\Omega W$. If this were a conventional regression satisfying $\mathbf{E}(\nu|W'X) = 0$, then we would know that the generalized least squares (GLS) estimator of β would be BLUE; this estimator is

(7) $$b_{GLSIV} = [X'W(W'\Omega W)^{-1}W'X]^{-1}X'W(W'\Omega W)^{-1}W'y.$$

This corresponds to using the weighting matrix $R = (W'\Omega W)^{-1}W'X$. In truth, the conditional expectation of ν given $W'X$ is not necessarily zero, but clean instruments will have the property that $(W'X)'\varepsilon/n \to_p 0$ because W and ε are uncorrelated in the population. This is enough to make the analogy work, so that (7) gives the IV estimator that has the smallest asymptotic variance among those that could be formed from the instruments W and a weighting matrix R.

If one makes the usual assumption that the disturbances ε have a scalar covariance matrix, $\Omega = I$, then the best IV estimator reduces to

(8) $$b_{2SLS} = [X'W(W'W)^{-1}W'X]^{-1}X'W(W'W)^{-1}W'y.$$

This corresponds to using the weighting matrix $R = (W'W)^{-1}W'X$. But this formula provides another interpretation of (8). If you regress each variable in X on the instruments, the resulting OLS coefficients are $(W'W)^{-1}W'X$, the same as R. Then, the best linear combination of instruments WR equals the fitted value $X^* = W(W'W)^{-1}W'X$ of the explanatory variables from a OLS regression of X on W. Further, you have $X'W(W'W)^{-1}W'X = X'X^* = X^{*\prime}X^*$ and $X'W(W'W)^{-1}W'y = X^{*\prime}y$, so that the IV estimator (8) can also be written

(9) $$b_{2SLS} = (X^{*\prime}X)^{-1}X^{*\prime}y = (X^{*\prime}X^*)^{-1}X^{*\prime}y.$$

This provides a two-stage least squares (2SLS) interpretation of the IV estimator: First, a OLS regression of the explanatory variables X on the instruments W is used to obtain fitted values X^*, and second a OLS regression of y on X^* is used to obtain the IV estimator b_{2SLS}. Note that in the first stage, any variable in X that is also in W will achieve a perfect fit, so that this variable is carried over without modification in the second stage.

The 2SLS estimator (8) or (9) will no longer be best when the scalar covariance matrix assumption $\mathbf{E}\varepsilon\varepsilon' = \sigma^2 I$ fails, but under fairly general conditions it will remain consistent. The best IV estimator (7) when $\mathbf{E}\varepsilon\varepsilon' = \sigma^2\Omega$ can be reinterpreted as a conventional 2SLS estimator applied to the transformed regression $Ly = LX\beta + \eta$ using the instruments $(L')^{-1}W$, where L is a Cholesky array that satisfies $L\Omega L' = I$. When Ω depends on unknown parameters, it is often possible to use a feasible generalized 2SLS procedure (FG2SLS): First estimate β using (8) and retrieve the residuals $u = y - Xb_{2SLS}$. Next use these residuals to obtain an estimate Ω^* of Ω. Then find a Cholesky transformation L satisfying $L\Omega^*L' = I$, make the transformations $\mathit{y} = Ly$, $\mathit{X} = LX$, and $\mathit{W} = (L')^{-1}W$, and do a 2SLS regression of y on X using W as instruments. This procedure gives a feasible form of (7), and is also called three-stage least squares (3SLS).

3. STATISTICAL PROPERTIES OF IV ESTIMATORS

IV estimators can behave badly in finite samples. In particular, they may fail to have moments. Their appeal relies on their behavior in large samples, although an important question is when samples are large enough so that the asymptotic approximation is reliable. We first discuss asymptotic properties, and then return to the issue of finite-sample properties.

We already made an argument that IV estimators are consistent, provided some limiting conditions are met. We did not show that IV estimators are unbiased, and in fact they usually are not. An exception where b_{IV} is unbiased is if the original regression equation actually satisfies Gauss-Markov assumptions. Then, no contamination is present, IV is not really needed, and if IV is used, its mean and variance can be calculated in the same way this was done for OLS, by first taking the conditional expectation with respect to ε, given X and W. In this case, OLS is BLUE, and since IV is another linear (in y) estimator, its variance will be at least as large as the OLS variance.

We show next that IV estimators are asymptotically normal under some regularity conditions, and establish their asymptotic covariance matrix. This gives a relatively complete large-sample theory for IV estimators. Let $\sigma^2\Omega$ be the covariance matrix of ε, given W, and assume that it is finite and of full rank. Make the assumptions:

[1] rank(W) = j ≥ k
[2a] $W'W/n \to_p H$, a positive definite matrix
[2b] $W'\Omega W/n \to_p F$, a positive definite matrix
[3] $X'W/n \to_p G$, a matrix of rank k
[4] $W'\varepsilon/n \to_p 0$
[5] $n^{-1/2}W'\varepsilon \to_d \mathcal{N}(0,\sigma^2F)$

Assumption [1] can always be met by dropping linearly dependent instruments, and should be thought of as true by construction. Assumption [1] implies that $W'W/n$ and $W'\Omega W/n$ are positive definite; Assumption [2] strengthens these to hold in the limit. Proper instruments have $X'W/n$ of rank k from the fully correlated condition and $\mathbf{E}(W'\varepsilon/n) = 0$ by the clean condition. Assumption [3] strengthens the fully correlated condition to hold in the limit. Assumption [4] will usually follow from the condition that the instruments are clean by applying a weak law of large numbers. For example, if the ε are independent and identically distributed with mean zero and finite variance, given W, then Assumption [2a] plus the Kolmogorov WLLN imply Assumption [4]. Assumption [5] will usually follow from Assumption [2b] by applying a central limit theorem. Continuing the i.i.d. example, the Lindeberg-Levy CLT implies Assumption [5]. There are WLLN and CLT that hold under much weaker conditions on the ε's, requiring only that their variances and correlations satisfy some bounds, and these can also be applied to derive Assumptions [4] and [5]. Thus, the statistical properties of IV can be established in the presence of many forms of heteroskedasticity and serial correlation.

Theorem: Assume that [1], [2b], [3] hold, and that an IV estimator is defined with a weighting matrix R_n that may depend on the sample n, but which converges to a matrix R of rank k. If [4] holds, then $b_{IV} \to_p \beta$. If both [4] and [5] hold, then

$$(10) \qquad n^{1/2}(b_{IV} - \beta) \to_d \mathcal{N}(0, \sigma^2(R'G')^{-1}R'FR(GR)^{-1}).$$

Suppose $R_n = (W'W)^{-1}W'X$ and [1]-[5] hold. Then the IV estimator specializes to the 2SLS estimator b_{2SLS} given by (8) which satisfies $b_{2SLS} \to_p \beta$ and

$$(11) \qquad n^{1/2}(b_{2SLS} - \beta) \to_d \mathcal{N}(0, \sigma^2(GH^{-1}G')^{-1}(GH^{-1}FH^{-1}G')(GH^{-1}G')^{-1}).$$

Suppose $R_n = (W'\Omega W)^{-1}W'X$ and [1]-[5] hold. Then the IV estimator specializes to the GLSIV estimator b_{GLSIV} given by (7) which satisfies $b_{GLSIV} \to_p \beta$ and

$$(12) \qquad n^{1/2}(b_{GLSIV} - \beta) \to_d \mathcal{N}(0, \sigma^2(GF^{-1}G')^{-1}).$$

Further, the GLSIV estimator is the minimum asymptotic variance estimator; i.e., $\sigma^2(R'G')^{-1}R'FR(GR)^{-1} - \sigma^2(GF^{-1}G')^{-1}$ is positive semidefinite. If $\Omega = I$, then the 2SLS and GLSIV estimators are the same, and the 2SLS estimator has limiting distribution (12) and is asymptotically best among all IV estimators that use instruments W.

The first part of this theorem is proved by dividing (6) by n and using assumptions [2], [3], and [4], and then dividing (6) by $n^{1/2}$ and applying assumptions [2], [3], and [5]. Substituting the definitions of R for the 2SLS and GLSIV versions then gives the asymptotic properties of these estimators. Finally, a little matrix algebra shows that the GLSIV estimator has minimum asymptotic variance among all IV estimators: Start with the matrix $I - F^{-1/2}G'(GF^{-1}G')^{-1}GF^{-1/2}$ which equals its own square, so that it is idempotent, and therefore positive semidefinite. Premultiply this idempotent matrix by $(R'G')^{-1}R'F^{1/2}$, and postmultiply it by the transpose of this matrix; the result remains positive semidefinite, and equals $(R'G')^{-1}R'FR(GR)^{-1} - (GF^{-1}G')^{-1}$. This establishes the result.

In order to use the large-sample properties of b_{IV} for hypothesis testing, it is necessary to find a consistent estimator for σ^2. The following estimator works: Define IV residuals

$$u = y - Xb_{IV} = [I - X(R'W'X)^{-1}R'W']y = [I - X(R'W'X)^{-1}R'W']\varepsilon,$$

the *Sum of Squared Residuals* SSR = $u'u$, and $s^2 = u'u/(n-k)$. If $\varepsilon'\varepsilon/n \to_p \sigma^2$, then s^2 is consistent for σ^2. To show this, simply write out the expression for $u'u/n$, and take the probability limit:

$$\begin{aligned}(13)\qquad \text{plim } u'u/n &= \text{plim } \varepsilon'\varepsilon/n - 2 \text{ plim } [\varepsilon'W/n]R([X'W/n]R)^{-1}[X'\varepsilon/n] \\ &\quad + [\varepsilon'W/n]R([X'W/n]R)^{-1}[X'X/n](R'[W'X/n])^{-1}R'[W'\varepsilon/n] \\ &= \sigma^2 - 2\cdot 0\cdot R\cdot(GR)^{-1}C + 0\cdot R\cdot(GR)^{-1}A(R'G')^{-1}R'\cdot 0 = \sigma^2.\end{aligned}$$

We could have used n-k instead of n in the denominator of this limit, as it makes no difference in large enough samples. The consistency of the estimator s^2 defined above holds for any IV estimator, and so holds in particular for the 2SLS or GLSIV estimators. Note that this consistent estimator of σ^2 substitutes the IV estimates of the coefficients into the original equation, and uses the original values of the X variables to form the residuals. When working with the 2SLS estimator, and calculating it by running the two OLS regression stages, you might be tempted to estimate σ^2 using a regression program printed values of SSR or the variance of the second stage regression, which is based on the residuals $\hat{u} = y - X^*b_{2SLS}$. It tuns out that this estimator is not consistent for σ^2: A few lines of matrix manipulation shows that $\hat{u}'\hat{u}/n \to_p \sigma^2 + \beta'[A - GF^{-1}G']\beta$. The second term is positive semidefinite, so this estimator is asymptotically biased upward.

Suppose $\mathbf{E}\varepsilon\varepsilon' = \sigma^2 I$, so that 2SLS is best among IV estimators using instruments W. The sum of squared residuals SSR = $u'u$, where $u = y - Xb_{2SLS}$, can be used in hypothesis testing in the same way as in OLS estimation. For example, consider the hypothesis that $\beta_2 = 0$, where β_2 is a $r\times 1$ subvector of β. Let SSR_0 be the sum of squared residuals from the 2SLS regression of y on X with $\beta_2 = 0$ imposed, and SSR_1 be the sum of squared residuals from the unrestricted 2SLS regression of y on X. Then, $[(SSR_0 - SSR_1)/m]/[SSR_1/(n-k)]$ has an approximate F-distribution under the null with m and n-k degrees of freedom. There are several cautions to keep in mind when considering use of this test statistic. This is a large sample approximation, rather than an exact distribution, because it is derived from the asymptotic normality of the 2SLS estimator. Its actual size in small samples could differ substantially from its nominal (asymptotic) size. Also, the large sample distribution of the statistic assumed that the disturbances ε have a scalar covariance matrix. Otherwise, it is mandatory to do a FGLS transformation before computing the test statistic above. For example, if $y = X\beta + \varepsilon$ represents a stacked system of equations such as structural equations in a simultaneous equations system, or if ε exhibits serial correlation, as may be the case in time-series or panel data, then one should estimate β consistently using 2SLS, retrieve the residuals $u = y - Xb_{2SLS}$ and use them to make an estimate Ω^* of $\Omega = \mathbf{E}\varepsilon\varepsilon'$, make the transformations $\mathcal{y}$ = Ly, $\mathcal{X}$ = LX, ν = Lε, and $\mathcal{W}$ = $(L')^{-1}W$ where L is a Cholesky matrix such that $L\Omega^*L'$ is proportional to an identity matrix, and finally apply 2SLS to the regression $\mathcal{y} = \mathcal{X}\beta + \nu$ with $\mathcal{W}$ as instruments and carry out the hypothesis testing using this model. The reason for the particular transformation of W is that one has $\mathcal{W}'\nu = W'\varepsilon$, so that the original property that the instruments were uncorrelated with the disturbances is preserved. The 3SLS procedure just described corresponds to estimating β using a feasible version of the GLSIV estimator.

What are the finite sample properties of IV estimators? Because you do not have the condition $\mathbf{E}(\varepsilon|X) = 0$ holding in applications where IV is needed, you cannot get simple expressions for the moments of $b_{IV} = [R'W'X]^{-1}R'W'y = \beta + [R'W'X]^{-1}R'W'\varepsilon$ by first taking expectations of ε conditioned on X and W. In particular, you cannot conclude that b_{IV} is unbiased, or that it has a covariance matrix corresponding to its asymptotic covariance matrix. In fact, b_{IV} can have very bad small-sample properties. To illustrate, consider the case where the number of instruments equals the number of observations, j = n. (This can actually arise in dynamic models, where often all lagged values of the exogenous variables are legitimate instruments. It can also arise when the candidate instruments are not only uncorrelated with ε, but satisfy the stronger property that $\mathbf{E}(\varepsilon|w) = 0$. In this case, all functions of w are also legitimate instruments.) In this case, W is a square matrix, and

$$\begin{aligned} b_{2SLS} &= [X'W(W'W)^{-1}W'X]^{-1}X'W(W'W)^{-1}W'y \\ &= [X'WW^{-1}W'^{-1}W'X]^{-1}X'WW^{-1}W'^{-1}W'y = [X'X]^{-1}X'y = b_{OLS}. \end{aligned}$$

We know OLS is inconsistent when $\mathbf{E}(\varepsilon|X) = 0$ fails, so clearly the 2SLS estimator is also biased if we let the number of instruments grow linearly with sample size. This shows that for the IV asymptotic theory to be a good approximation, n must be much larger than j. One rule-of-thumb for IV is that n - j should exceed 40, and should grow linearly with n in order to have the large-sample approximations to the IV distribution work well.

Considerable technical analysis is required to characterize the finite-sample distributions of IV estimators analytically; the names associated with this problem are Nagar, Phillips, and Mariano. However, simple numerical examples provide a picture of the situation. Consider first a regression $y = x\beta + \varepsilon$ where there is a single right-hand-side variable, and a single instrument w, and assume x, w, and ε have the simple joint distribution given in the table below, where λ is the correlation of x and w, ρ is the correlation of x and ε, and $0 \le \lambda,\rho$ and $\lambda + 2\rho < 1$:

x	w	ε	Prob
1	1	1	(1+λ)/8
-1	1	1	(1-λ)/8
1	-1	1	(1-λ+2ρ)/8
-1	-1	1	(1+λ-2ρ)/8
1	1	-1	(1+λ)/8
-1	1	-1	(1-λ)/8
1	-1	-1	(1-λ-2ρ)/8
-1	-1	-1	(1+λ+2ρ)/8

These random variables then satisfy $\mathbf{E}x = \mathbf{E}w = \mathbf{E}\varepsilon = 0$, $\mathbf{E}x\varepsilon = \rho$, $\mathbf{E}xw = \lambda$, and $\mathbf{E}w\varepsilon = 0$, and their products have the joint distribution

xw	$w\varepsilon$	$x\varepsilon$	Prob
1	1	1	$(1+\lambda+\rho)/4$
-1	-1	1	$(1-\lambda+\rho)/4$
-1	1	-1	$(1-\lambda-\rho)/4$
1	-1	-1	$(1+\lambda-\rho)/4$

Least squares is biased if $\rho \neq 0$, and IV is consistent if $\lambda \neq 0$. Suppose n = 2. Then the exact distribution of the relevant random variables is

$\sum xw$	$\sum w\varepsilon$	$\sum x\varepsilon$	$b_{OLS}-\beta$	$b_{IV}-\beta$	Prob
2	2	2	1	1	$(1+\lambda+\rho)^2/16$
0	0	2	1	0	$((1+\rho)^2-\lambda^2)/8$
0	2	0	0	$+\infty$	$(1-(\lambda+\rho)^2)/8$
2	0	0	0	0	$((1+\lambda)^2-\rho^2)/8$
-2	-2	2	1	1	$(1-\lambda+\rho)^2/16$
-2	0	0	0	0	$((1-\lambda)^2-\rho^2)/8$
0	-2	0	0	$-\infty$	$(1-(\lambda-\rho)^2)/8$
-2	2	-2	-1	-1	$(1-\lambda-\rho)^2/16$
0	0	-2	-1	0	$((1-\rho)^2-\lambda^2)/8$
2	-2	-2	-1	-1	$(1+\lambda-\rho)^2/16$

Note first that there is a positive probability that b_{IV} is not defined; hence, technically it has no finite moments. Collecting terms from this table, the exact CDF of $b_{OLS} - \beta$ and $b_{IV} - \beta$ satisfy

c	$\mathrm{Prob}(b_{OLS}-\beta\leq c)$	$\mathrm{Prob}(b_{IV}-\beta\leq c)$
$-\infty$	0	$(1-(\lambda-\rho)^2)/8$
-1	$(1-\rho)^2/4$	$(1-\lambda(1-\rho))/4$
0	$(1-\rho)(3+\rho)/4$	$(3-\lambda(1-\rho))/4$
1	1	$(\lambda+\rho)^2/2$
$+\infty$	1	1

Also, $\mathrm{Prob}(|b_{IV}-\beta| > |b_{OLS}-\beta|) = (1-\lambda^2-\rho^2)/4$. Then for this small sample there is a substantial probability that the IV estimator will be further away from the true value than the OLS estimator. As an exercise, carry through this example for n = 3, and show that in this case b_{IV} will always exist, but there continues to be a large probability that b_{OLS} is closer to β than b_{IV}. As n increases, the probability that b_{OLS} is closer than b_{IV} shrinks toward zero, but there is always a positive probability that the IV estimator is worse than the OLS estimator, and for n odd a positive probability that the IV estimator is infinite, so it never has any finite moments.

The second example is the one-variable model $y = x\beta + \varepsilon$ with one instrument w where (x,w,ε) are jointly normal with zero means, unit variances, $\mathbf{E}wx = \lambda$, $\mathbf{E}x\varepsilon = \rho$, and $\mathbf{E}w\varepsilon = 0$. A difficult technical analysis can be used to derive the exact distribution of the IV estimator in terms of a non-central Wishart distribution. However, for purposes of getting an idea of how IV performs, it is much simpler to do a small computer simulation. For the values $\rho = .2$ and $\lambda = .8$, the table below gives the results of estimating a true value $\beta = 1$ in 1000 samples of sizes n = 5, 10, 20, or 40. Because the denominator in the IV estimator is small with some probability, the IV estimator tends to produce large deviations that lead to a large mean square error (MSE). In this example, the probability that the IV estimator is closer to β than the OLS estimator exceeds 0.5 only for samples of size 20 or greater, and the IV estimator has a smaller MSE only for samples of size 40 or larger. The smaller ρ or λ, the larger the sample size needed to make IV better than OLS in terms of MSE.

Sample Size	Mean Bias in b_{OLS} (1000 samples)	Mean Bias in b_{IV} (1000 samples)	MSE of b_{OLS} (1000 samples)	MSE of b_{IV} (1000 samples)	Frequency of b_{IV} as good as b_{OLS} (1000 samples)
5	0.18	-0.15	0.25	63.5	39.6%
10	0.19	-0.04	0.15	0.70	45.7%
20	0.20	-0.02	0.09	0.10	54.6%
40	0.20	-0.00	0.07	0.04	69.2%

In practice, in problems where sample size minus the number of instruments exceeds 40, the asymptotic approximation to the distribution of the IV estimator is reasonably good, and one can use it to compare the OLS and IV estimates. To illustrate, continue the example of a regression in one variable, $y = x\beta + \varepsilon$. Suppose as before that x and ε have a correlation coefficient $\rho \neq 0$, so that OLS is biased, and suppose that there is a single proper instrument w that is uncorrelated with ε and has a correlation $\lambda \neq 0$ with x. Then, the OLS estimator is asymptotically normal with mean $\beta + \rho\sigma_\varepsilon/\sigma_x$ and variance $\sigma_\varepsilon^2/n\sigma_x^2$. The 2SLS estimator is asymptotically normal with mean β and variance $\sigma_\varepsilon^2/n\sigma_x^2\lambda^2$. The mean squares of the two estimators are then, approximately,

$$MSE_{OLS} = (\rho^2 + 1/n)\sigma_\varepsilon^2/\sigma_x^2$$
$$MSE_{2SLS} = \sigma_\varepsilon^2/n\sigma_x^2\lambda^2.$$

Then, 2SLS has a lower MSE than OLS when

$$1 < \rho^2\lambda^2 n/(1-\lambda^2) \approx (b_{2SLS}-b_{OLS})^2/(V(b_{2SLS})-V(b_{OLS})),$$

or approximately $n > (1 - \lambda^2)/\rho^2\lambda^2$. When $\lambda = 0.8$ and $\rho = 0.2$, this asymptotic approximation suggests that a sample size of about 14 is the tip point where b_{IV} should be better than b in terms of MSE. However, the asymptotic formula underestimates the probability of very large deviations arising from a denominator in b_{IV} that is near zero, and as a consequence is too quick to reject b_{OLS}. The right-hand-side of this approximation to the ratio of the MSE is the Hausman test statistic for exogeneity, discussed below; for this one-variable case, one should reject the null hypothesis of exogeneity when the statistic exceeds one. Under the null, the statistic is approximately chi-square with one degree of freedom, so that this criterion corresponds to a type I error probability of 0.317.

4. RELATION OF IV TO OTHER ESTIMATORS

The 2SLS estimator can be interpreted as a member of the family of *Generalized Method of Moments* (GMM) estimators. You can verify by differentiating to get the first-order condition that the 2SLS estimator of the equation $Ly = LX\beta + L\varepsilon$ using the instruments $(L')^{-1}W$, where $\mathbf{E}\varepsilon\varepsilon' = \sigma^2\Omega$ and L is a Cholesky matrix satisfying $L\Omega L' = I$, solves

$$\text{Min}_\beta\ (y\text{-}X\beta)'W(W'\Omega W)^{-1}W'(y\text{-}X\beta). \tag{14}$$

In this quadratic form objective function, $W'(y\text{-}X\beta)$ is the moment that has expectation zero in the population when β is the true parameter vector, and $(W'\Omega W)^{-1}$ is a "distance metric" in the center of the quadratic form. Define $P = (L')^{-1}W(W'\Omega W)^{-1}W'(L)^{-1}$, and note that P is idempotent, and thus is a projection matrix. Then, the GMM criterion chooses β to minimize the length of the vector $L(y\text{-}X\beta)$ projected onto the subspace spanned by P. The properties of GMM hypothesis testing procedures follow readily from the observation that $L(y\text{-}X\beta)$ has mean zero and a scalar covariance matrix. In particular, $\text{Min}_\beta\ (y\text{-}X\beta)'W(W'\Omega W)^{-1}W'(y\text{-}X\beta)/\sigma^2$ is asymptotically chi-squared distributed with degrees of freedom equal to the rank of P.

It is possible to give the 2SLS estimator a *pseudo-MLE* interpretation. Premultiply the regression equation by $W'L^{-1}$ to obtain $W'y = W'X\beta + W'\varepsilon$. Now treat $W'\varepsilon$ *as if* it were normally distributed with mean zero and j×j covariance matrix $\lambda^2 W'\Omega W$, conditioned on $W'X$. Then, the log likelihood of the sample would be

$$\begin{aligned} L = &- (j/2)\log 2\pi - (j/2)(\tfrac{1}{2})\log\lambda^2 - (\tfrac{1}{2})\log\det(W'\Omega W) \\ &- (1/2\lambda^2)(W'y\text{-}W'X\beta)'(W'\Omega W)^{-1}(W'y\text{-}W'X\beta). \end{aligned}$$

The first-order condition for maximization of this pseudo-likelihood is the same as the condition defining the 2SLS estimator.

5. TESTING EXOGENEITY

Sometimes one is unsure whether some potential instruments are clean. If they are, then there is an asymptotic efficiency gain from including them as instruments. However, if they are not, estimates will be inconsistent. Because of this tradeoff, it is useful to have a specification test that permits one to judge whether suspect instruments are clean or not. To set the problem, consider a regression $y = X\beta + \varepsilon$, an array of proper instruments Z, and an array of instruments W that includes Z plus other variables that may be either clean or contaminated.

Several superficially different problems can be recast in this framework:

(1) The regression may be one in which some right-hand-side variables are known to be exogenous and others are suspect, Z is an array that contains the known exogenous variables and other clean instruments, and W contains Z and the variables in X that were excluded from Z because of the possibility that they might be dirty. In this case, 2SLS using W reduces to OLS, and the problem is to test whether the regression can be estimated consistently by OLS.

(2) The regression may contain known endogenous and known exogenous variables, Z is an array that contains the known exogenous variables and other clean instruments, and W is an array that contains Z and additional suspect instruments from outside the equation. In this case, one has a consistent 2SLS estimator using instruments Z, and a 2SLS estimator using instruments W that is more efficient under the hypothesis that W is exogenous, but inconsistent otherwise. The question is whether to use the more inclusive array of instruments.

(3) The regression may contain known endogenous, known exogenous, and suspect right-hand-side variables, Z is an array that contains the known exogenous variables plus other instruments from outside the equation, and W is an array that contains Z plus the suspect variables from the equation. The question is whether it is necessary to instrument for the suspect variables, or whether they are clean and can themselves be used as instruments.

In the regression $y = X\beta + \varepsilon$, you can play it safe and use only the Z instruments. This gives $b_Q = (X'QX)^{-1}X'Qy$, where $Q = (L')^{-1}Z(Z'\Omega Z)^{-1}Z'(L)^{-1}$. Alternately, you use W, including the suspect instruments, taking a chance with inconsistency to gain efficiency. This gives

$$b_P = (X'PX)^{-1}X'Py, \text{ where } P = (L')^{-1}W(W'\Omega W)^{-1}W'(L)^{-1}.$$

If the suspect instruments are clean and both estimators are consistent, then b_Q and b_P should be close together, as they are estimates of the same β; further, b_P is efficient relative to b_Q, implying that the covariance matrix of $(b_Q - b_P)$ equals the covariance matrix of b_Q minus the covariance matrix of b_P. However, if the suspect instruments are contaminated, b_P is inconsistent, and $(b_Q - b_P)$ has a nonzero probability limit. This suggests a test statistic of the form

$$(15) \qquad (b_Q - b_P)'[V(b_Q) - V(b_P)]^{-}(b_Q - b_P),$$

where $[\cdot]^{-}$ denotes a generalized inverse, could be used to test if W is clean. This form is the exogeneity test originally proposed by Hausman. Under the null hypothesis that W is clean, this statistic will be asymptotically chi-square with degrees of freedom equal to the rank of the covariance matrix in the center of the quadratic form.

Another formulation of an exogeneity test is more convenient to compute, and can be shown (in one manifestation) to be equivalent to the Hausman test statistic. This alternative formulation has the form of an omitted variable test, with appropriately constructed auxiliary variables. We describe the test in the case $\mathbf{E}\varepsilon\varepsilon' = \sigma^2 I$ and leave as an exercise the extension to the case .$\mathbf{E}\varepsilon\varepsilon' = \sigma^2\Omega$.

First do an OLS regression of X on Z and retrieve fitted values $X^* = QX$, where $Q = Z(Z'Z)^{-1}Z'$. (This is necessary only for variables in X that are not in Z, since otherwise this step just returns the original variable.) Second, using W as instruments, do a 2SLS regression of y on X, and retrieve the sum of squared residuals SSR_1. Third, do a 2SLS regression of y on X <u>and</u> a subset of m columns of X^* that are linearly independent of X, and retrieve the sum of squared residuals SSR_2.

Finally, form the statistic $[(SSR_1 - SSR_2)/m]/[SSR_2/(n-k)]$. Under the null hypothesis that W is clean, this statistic has an approximate F-distribution with m and n-k degrees of freedom, and can be interpreted as a test for whether the m auxiliary variables from X^* should be omitted from the regression. When a subset of X^* of maximum possible rank is chosen, this statistic turns out to be asymptotically equivalent to the Hausman test statistic. Note that if W contains X, then the 2SLS in the second and third steps reduces to OLS.

We next show that this test is indeed an exogeneity test. Consider the 2SLS regression

$$y = X\beta + X_1^*\gamma + \eta,$$

where X_1^* is a subset of $X^* = QX$ such that $[X,X_1^*]$ is of full rank. The 2SLS estimates of the parameters in this model, using W as instruments, satisfy

$$\begin{bmatrix} b_P \\ c_P \end{bmatrix} = \begin{bmatrix} X'PX & X'QX_1 \\ X_1'QX & X_1'QX_1 \end{bmatrix}^{-1} \begin{bmatrix} X'Py \\ X_1'Qy \end{bmatrix} = \begin{bmatrix} \beta \\ 0 \end{bmatrix} + \begin{bmatrix} X'PX & X'QX_1 \\ X_1'QX & X_1'QX_1 \end{bmatrix}^{-1} \begin{bmatrix} X'P\varepsilon \\ X_1'Q\varepsilon \end{bmatrix} .$$

But $X'Q\varepsilon/n \to_p \text{plim}(X'Z/n)\cdot(\text{plim}(Z'Z/n))^{-1}\cdot\text{plim}(Z'\varepsilon/n) = 0$ by assumptions [1]-[4] when Z is clean. Similarly, $X'P\varepsilon/n \to_p GH^{-1}\cdot\text{plim}(W'\varepsilon/n) = 0$ when W is clean, but $X'P\varepsilon/n \to_p GH^{-1}\cdot\text{plim}(W'\varepsilon/n) \neq 0$ when W is contaminated. Define

$$\begin{bmatrix} X'PX/n & X'QX_1/n \\ X_1'QX/n & X_1'QX_1/n \end{bmatrix}^{-1} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} .$$

From the formula for a partitioned inverse,

$$A_{11} = (X'[P - QX_1(X_1'QX_1)^{-1}X_1'Q]X/n)^{-1}$$

$$A_{22} = (X_1'Q[I - X(X'PX)^{-1}X']QX_1/n)^{-1}$$

$$A_{21} = -(X_1'QX_1)^{-1}X_1'QX\cdot A_{11} = -A_{22}(X_1'QX)(X'PX)^{-1} = A_{12}'$$

Hence,

$$(16) \qquad c_P = A_{22}\cdot\{X_1'Q\varepsilon/n - (X_1'QX)(X'PX)^{-1}\cdot X'P\varepsilon/n\}.$$

If W is clean and satisfies assumptions [4] and [5], then $c_P \to_p 0$ and $n^{1/2}c_P$ is asymptotically normal. On the other hand, if W is contaminated, then c_P has a non-zero probability limit. Then, a test for $\gamma = 0$ using c_P is a test of exogeneity.

The test above can be reinterpreted as a Hausman test involving differences of b_P and b_Q. Recall that $b_Q = \beta + (X'QX)^{-1}X'Q\varepsilon$ and $b_P = \beta + (X'PX)^{-1}X'P\varepsilon$. Then

$$(17) \qquad (X'QX)(b_Q - b_P) = \{X'Q\varepsilon/n - (X'QX)(X'PX)^{-1}\cdot X'P\varepsilon/n\}.$$

Then in particular for a linearly independent subvector X_1 of X,

$$A_{22}(X_1'QX)(b_Q - b_P) = A_{22}\{X_1'Q\varepsilon/n - (X_1'QX)(X'PX)^{-1}\cdot X'P\varepsilon/n\} = c_P.$$

Thus, c_P is a linear transformation of $(b_Q - b_P)$. Then, testing whether c_P is near zero is equivalent to testing whether a linear transformation of $(b_Q - b_P)$ is near zero. When X_1 is of maximum rank, this equivalence establishes that the Hausman test in its original form is the same as the test for c_P.

6. EXOGENICITY TESTS ARE GMM TESTS FOR OVER-IDENTIFICATION

The Hausman Exogeneity Test. Consider the regression model $y = X\beta + \varepsilon$, and suppose one wants to test the exogeneity of p variables X_1 in X. Suppose R is an array of instruments, including X_2; then $Z = P_RX_1$ are instruments for X_1. Let W = [Z X] be all the variables that are orthogonal to ε in the population under the null hypothesis that X and ε are uncorrelated. As in the omitted variables problem, consider the test statistic for over-identifying restrictions, $2nQ_n = \min_b u'P_Wu/\sigma^2$, where u = y - Xb. Decompose $P_W = P_X + (P_W - P_X)$. Then $u'(P_W - P_X)u = y'(P_W - P_X)y$ and the minimizing b sets $u'P_Xu = 0$, so that $2nQ_n = y'(P_W - P_X)y/\sigma^2$. Since $P_W - P_X = P_{Q_XW}$, one also has $2nQ_n = y' P_{Q_XW} y$. This statistic is the same as the test statistic for the hypothesis that the coefficients of Z are zero in a regression of y on X and Z; thus the test for over-identifying restrictions is an omitted variables test. One can also write $2nQ_n = \|\hat{y}_W - \hat{y}_X\|^2/\sigma^2$, so that a computationally convenient equivalent test is based on the difference between the fitted values of y from a regression on X and Z and a regression on X alone. Finally, we will show that the statistic can be written

$$2nQ_n = (b_{1,2SLS} - b_{1,OLS})[V(b_{1,2SLS}) - V(b_{1,OLS})]^{-1}(b_{1,2SLS} - b_{1,OLS}).$$

In this form, the statistic is the Hausman test for exogenicity in the form developed by Hausman and Taylor, and the result establishes that the Hausman test for exogeneity is equivalent to a GMM test for over-identifying restrictions.

Several steps are needed to demonstrate this equivalence. Note that $b_{2SLS} = (X'P_MX)^{-1}X'P_My$, where $M = [Z\ X_2]$. Write

$$\begin{aligned} b_{2SLS} - b_{OLS} &= (X'P_MX)^{-1}X'P_My - (X'X)^{-1}X'y \\ &= (X'P_MX)^{-1}[X'P_M - X'P_MX(X'X)^{-1}X']y \\ &= (X'P_MX)^{-1}X'P_MQ_Xy. \end{aligned}$$

Since X_2 is in M, $P_MX_2 = X_2$, implying $X'P_MQ_X = \begin{bmatrix} X_1'P_MQ_X \\ X_2'P_MQ_X \end{bmatrix} = \begin{bmatrix} X_1'P_MQ_X \\ X_2'Q_X \end{bmatrix} = \begin{bmatrix} X_1'P_MQ_X \\ 0 \end{bmatrix}$.

Also, $X'P_MX = \begin{bmatrix} X_1'P_MX_1 & X_1'P_MX_2 \\ X_2'P_MX_1 & X_2'P_MX_2 \end{bmatrix} = \begin{bmatrix} X_1'P_MX_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}$. Then $\begin{bmatrix} X_1'P_MQ_Xy \\ 0 \end{bmatrix} = (X'P_MX)(b_{2SLS}$

- $b_{OLS}) \equiv \begin{bmatrix} X_1'P_M X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{bmatrix} b_{1,2SLS} - b_{1,OLS} \\ b_{2,2SLS} - b_{2,OLS} \end{bmatrix}$. From the second block of equations, one obtains the result that the second subvector is a linear combination of the first subvector. This implies that a test statistic that is a function of the full vector of differences of 2SLS and OLS estimates can be written equivalently as a function of the first subvector of differences. From the first block of equations, substituting in the solution for the second subvector of differences expressed in terms of the first, one obtains

$$[X_1'P_M X_1 - X_1'X_2(X_2'X_2)^{-1}X_2'X_1](b_{1,2SLS} - b_{1,OLS}) = X_1'P_M Q_X y$$

The matrix on the left-hand-side can be rewritten as $X_1'P_M \ Q_{X_2} \ P_M X_1$, so that

$$b_{1,2SLS} - b_{1,OLS} = (X_1'P_M \ Q_{X_2} \ P_M X_1)^{-1} X_1'P_M Q_X y.$$

Next, we calculate the covariance matrix of b_{2SLS} - b_{OLS}, and show that it is equal to the difference of $V(b_{2SLS}) = \sigma^2(X'P_M X)^{-1}$ and $V(b_{OLS}) = \sigma^2(X'X)^{-1}$. From the formula $b_{2SLS} - b_{OLS} = (X'P_M X)^{-1}X'P_M Q_X y$, one has $V(b_{2SLS} - b_{OLS}) = \sigma^2(X'P_M X)^{-1}X'P_M Q_X P_M X(X'P_M X)^{-1}$.
On the other hand,

$$\begin{aligned} V(b_{2SLS}) - V(b_{OLS}) &= \sigma^2(X'P_M X)^{-1}\{X'P_M X - X'P_M X(X'X)^{-1}X'P_M X\}(X'P_M X)^{-1} \\ &= \sigma^2(X'P_M X)^{-1}\{X'P_M[I - X(X'X)^{-1}X']P_M X\}(X'P_M X)^{-1} \\ &= \sigma^2(X'P_M X)^{-1}X'P_M Q_X P_M X(X'P_M X)^{-1}. \end{aligned}$$

Thus, $V(b_{2SLS} - b_{OLS}) = V(b_{2SLS}) - V(b_{OLS})$. This is a consequence of the fact that under the null hypothesis OLS is efficient among the class of linear estimators including 2SLS. Expanding the center of this expression, and using the results $P_M X_2 = X_2$ and hence $Q_X P_M X_2 = 0$, one has

$$X'P_M Q_X P_M X = \begin{bmatrix} X_1'P_M Q_X P_M X_1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Hence, $V(b_{2SLS}) - V(b_{OLS})$ is of rank p; this also follows by noting that $b_{2,2SLS} - b_{2,OLS}$ could be written as a linear transformation of $b_{1,2SLS} - b_{1,OLS}$.

Next, use the formula for partitioned inverses to show for N = M or N = I that the northwest corner of $\begin{bmatrix} X_1'P_N X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1}$ is $(X_1'P_N Q_{X_2} P_N X_1)^{-1}$. Then,

$$V(b_{1,2SLS} - b_{1,OLS}) = \sigma^2(X_1'P_M \ Q_{X_2} \ P_M X_1)^{-1}X_1'P_M Q_X P_M X_1(X_1'P_M \ Q_{X_2} \ P_M X_1)^{-1}.$$

Using the expressions above, the quadratic form can be written

$$(b_{1,2SLS} - b_{1,OLS})V(b_{1,2SLS} - b_{1,OLS})^{-1}(b_{1,2SLS} - b_{1,OLS})$$

$$= y'Q_XP_MX_1(X_1'P_MQ_XP_MX_1)^{-1}X_1'P_MQ_Xy/\sigma^2.$$

Finally, one has, from the test for over-identifying restrictions,

$$2nQ_n = y'(P_W - P_X)y/\sigma^2 = y'P_{Q_XW}y\ /\sigma^2$$

$$\equiv y'Q_XP_MX_1(X_1'P_MQ_XP_MX_1)^{-1}X_1'P_MQ_Xy/\sigma^2,$$

so that the two statistics coincide.

A Generalized Exogenicity Test: Consider the regression $y = X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + \varepsilon$, and the null hypothesis that X_1 is exogenous, where X_2 is known to be exogenous, and X_3 is known to be endogenous. Suppose N is an array of instruments, including X_2, that are sufficient to identify the coefficients when the hypothesis is false. Let $W = [N\ X_1]$ be the full set of instruments available when the null hypothesis is true. Then the best instruments under the null hypothesis are $X_o = P_WX \equiv [X_1\ \ X_2\ \ X_3^*]$, and the best instruments under the alternative are $X_u = P_NX \equiv [X_1^*\ \ X_2\ \ X_3^*]$. The test statistic for over-identifying restrictions is $2nQ_n = y'(P_{X_o} - P_{X_u})y/\sigma^2$, as in the previous cases. This can be written $2nQ_n = (SSR_{X_o} - SSR_{X_u})/\sigma^2$, with the numerator the difference in sum of squared residuals from a OLS regression of y on X_u and a OLS regression of y on X_o. Also, $2nQ_n = \|\hat{y}_{X_o} - \hat{y}_{X_u}\|^2/\sigma^2$, the difference between the fitted values of y from a regression on X_u and a regression on X_o. Finally,

$$2nQ_n = (b_{2SLS_o} - b_{2SLS_u})'[V(b_{2SLS_u}) - V(b_{2SLS_o})]^- (b_{2SLS_o} - b_{2SLS_u}),$$

an extension of the Hausman-Taylor exogeneity test to the problem where some variables are suspect and others are known to be exogenous. One can show that the quadratic form in the center of this quadratic form has rank equal to the rank of X_1, and that the test statistic can be written equivalently as a quadratic form in the subvector of differences of the 2SLS estimates for the X_1 coefficients, with the ordinary inverse of the corresponding submatrix of differences of variances in the center of the quadratic form.

7. INSTRUMENTAL VARIABLES IN TIME-SERIES MODELS

The treatment of IV estimation up to this point applies in principle to observations made either in cross section or over time. For example, if the observations correspond to time periods and $\mathbf{E}(\varepsilon\varepsilon'|W) = \sigma^2\Omega$ with Ω either known or estimated, the 2SLS estimator (2) or the two-stage feasible generalized least squares estimator (10) with Ω estimated using residuals obtained by application of (2), can be applied to problems where the structure of Ω comes from serial correlation. However, for time series applications it is useful to examine in more detail the structure of W and the orthogonality conditions used in forming IV estimators. In particular, one should ask how conventional sources of contamination in explanatory variables such as omitted variables or measurement error and conventional sources of serial correlation such as behavioral lags in adjustment are likely to affect the serial correlation structure of disturbances and the correlation of contemporaneous disturbances with explanatory variables for various transformations of the model.

Start with the example of a linear model with measurement error in explanatory variables, and suppose that in the absence of this measurement error problem the disturbance in the equation would follow an AR1 process. Let z_t denote the ideal variables without measurement error, and $x_t = z_t + \eta_t$ denote the observed explanatory variables. Then, the model can be written

$$y_t = z_t\beta + \varepsilon_t \text{ with } \varepsilon_t = \rho\varepsilon_{t-1} + \nu_t,$$

or

$$y_t = x_t\beta + \nu_t - \eta_t\beta + \rho\nu_{t-1} + \rho^2\nu_{t-2} + ..., \tag{18}$$

where the ν_t are i.i.d. innovations and $\rho^2 < 1$. This model can also be written

$$y_t = y_{t-1}\rho + x_t\beta - x_{t-1}\beta\rho + (\nu_t - \eta_t\beta + \eta_{t-1}\beta\rho). \tag{19}$$

The form (19) removes the serial correlation in the ideal equation disturbance, but in doing so introduces a moving average of the measurement errors. Only in the unlikely case that all components of η_t follow an AR1 process with the same ρ as the ε_t process will serial correlation be fully removed.[1] Application of OLS to either (18) or (19) will then in general result in inconsistent estimates. The issue for application of IV methods is whether proper instruments can be found. In (18), the variables in x_t that are measured with error would require instrumenting. If the z_t are serially correlated, and the η_t are not, then x_{t-1}, x_{t-2},... are potential clean instruments for x_t. However, if there is serial correlation in the measurement errors, one would need to find proper instruments from outside the model. In (19), all of the explanatory variables y_{t-1}, x_t, and x_{t-1} are contaminated, but if the z_t are correlated with a sufficiently long lag and the η_t are uncorrelated, then x_{t-2}, x_{t-3}, x_{t-4},... are potential clean instruments. It is important to not introduce x's with too high lags as instruments, because this requires truncating the sample in order to observe the instruments for each date used in the estimation, and the good statistical properties of the IV method begins to break down as the number of instruments ceases to be small relative to the remaining sample size.

Omitted variables leads to models similar to (18) and (19). In this case, interpret the disturbance in the model $y_t = x_t\beta + \varepsilon_t$ as including the omitted variables. If these omitted variables are themselves serially correlated, then they will induce serial correlation in ε_t, perhaps adding to serial correlation in a disturbance component that arises for reasons other than omitted variables. A transformation of the model in this case may be able to remove serial correlation in the disturbance, but does not remove the contamination. The issue will be to find proper instruments. If the included x's are themselves serially correlated and the final disturbance is AR1, then the

[1] The situation in which all the variables in a model follow the same AR process does has some chance of arising in stationary state equilibria, because equilibrium pressures may force all variables to move nearly in lock-step along a dynamic path determined by the largest root of the system.

equation $y_t = y_{t-1}\rho + x_t\beta - x_{t-1}\beta\rho + \varepsilon_t - \rho\varepsilon_{t-1}$ obtained by partial differencing will have y_{t-1}, x_{t-1}, x_{t-2},... as potential clean instruments. For this to work, the AR1 specification for ε_t must be correct, and x_t must not have the same AR1 process.

The preceding examples illustrate several important points about the use of IV methods in time-series models. First, there is likely to be an interaction between the source of the contamination and the nature of the serial correlation in the model. Second, the process followed by the explanatory variables will determine what variables are clean (i.e., uncorrelated with the contemporaneous disturbance) and what variables might be available as instruments. Third, choice of instruments is not clear-cut, and may involve the question of what variables are potential clean instruments and how many potential instruments to introduce given the fairly poor small sample properties of IV. The use of lags of y_t or x_t as instruments exacerbates the sample size problem, since it decreases the operating sample size as the number of instruments rises. Further, lagged variables may fail to be proper instruments, either because assumptions of zero correlation are not robust and fail due to a more complex pattern of serial correlation than the econometrician assumes, or because these lagged variables are not correlated with the variables they are instrumenting. Together, these observations suggest that careful consideration of the nature of contamination and serial correlation is needed in time-series applications of IV, and that this method be used with caution.

8. INSTRUMENTAL VARIABLES IN NONLINEAR MODELS

The method of instrumental variables in its most commonly used 2SLS form is applied to models linear in variables and in parameters, $y = X\beta + \varepsilon$. If there are proper instruments W for X and if $\mathbf{E}(\varepsilon|W) = \sigma^2\mathbf{I}$, then the 2SLS estimator (2) is consistent for β and efficient among all IV estimators using these instruments; see the theorem in Section 3. However, the orthogonality conditions invoked to justify the IV method do not necessarily extend to nonlinear transformations, because expectations are not preserved. For example, economic applications may postulate a zero correlation between variables for behavioral reasons, such as the rational expectations hypothesis that intertemporally optimized consumption is a random walk whose innovations are uncorrelated with history. This is not sufficient to guarantee that innovations in a nonlinear transformation of consumption are uncorrelated with history. To investigate what happens without linearity, consider three cases of nonlinearity:

(a) Models nonlinear in parameters only: $y = x\beta(\theta) + \varepsilon$
(b) Models nonlinear in variables only: $y = f(x)\beta + \varepsilon$
(c) Models nonlinear in both variables and parameters: $y = h(x,\theta) + \varepsilon$

A case such as (a) might arise for example when partial differencing is done to handle AR1 serial correlation. In this case, $y = x\alpha + \eta$ and $\eta = \rho\eta_{-1} + \nu$ with ν i.i.d., and transformation yields $y = \rho y_{-1} + x\alpha - x_{-1}\alpha\rho + \nu$, a model that has i.i.d. disturbances, but the parameters α and ρ appearing in nonlinear combination. Suppose in the model (a) that one first does an OLS regression of x on proper instruments w, and retrieves fitted values x^*, and second does a nonlinear least squares regression for the model $y = x^*\beta(\theta) + \varepsilon^*$. Examine the first-order conditions for the last regression, and show as an exercise that orthogonality of the instruments and the disturbances in the original regression implies consistency, just as in the fully linear case.[2] , It is the linearity of the first-order condition in the instruments and in ε that guarantees that the initial condition that the instruments be uncorrelated with ε continues to suffice.

Next consider the case y = f(x)β + ε with nonlinear transformation of the explanatory variables but linearity in parameters. If instruments w are available that are uncorrelated with ε and fully correlated with f(x), then GMM estimation using the criterion function

$$\left[\sum_{i=1}^{N} w_i(y_i - f(x_i)\beta)\right]' \cdot \left[\sum_{i=1}^{N} w_i w_i'\right]^{-1} \cdot \left[\sum_{i=1}^{N} w_i(y_i - f(x_i)\beta)\right] , \tag{20}$$

will be consistent; see Chapter 3. Solution of this GMM problem can be given a 2SLS interpretation: First do an OLS regression of $f(x_i)$ on w_i, and retrieve fitted values f^*, then do an OLS regression of y_i on f^*. Then, the form and computation of the IV estimator are not affected by nonlinearity in variables. However, there are substantial issues regarding specification of the instruments. In particular, given an initial set of "raw" instruments z, should they be given nonlinear transformations to improve the efficiency of the IV estimator? An initial issue is whether postulated orthogonality of z and ε will be preserved for nonlinear transformations of z. This will depend on the economic application and the nature of z. If the application can guarantee only that z is uncorrelated with ε, this property will not in general be preserved under nonlinear transformation, and the only clean instruments w will be the untransformed z. However, if the application can guarantee that z is statistically independent of ε, then any nonlinear transformation of z will be uncorrelated with ε, and is a potential clean instrument. For the remainder of this section, assume that z and ε are statistically independent.

What transformations of z make good instruments? In some cases it is feasible to apply the nonlinear transformation f to z_i, and tempting to use $f(z_i)$ to instrument $f(x_i)$. For example, if x_i is a variable measured with error, and z_i is an independent measurement of the same variable, then provided one is persuaded that the error in z_i is statistically independent of ε_i, $f(z_i)$ seems to be a reasonable instrument for $f(x_i)$; e.g., $\log(z_i)$ seems to be a natural instrument for $\log(x_i)$. This is a practical thing to do, and will often give a more precise IV estimator than one that just uses the raw instruments. However, it will not in general yield the most efficient possible IV estimator. The reason for this is the proposition that expectations are not preserved under nonlinear transformations.

The best instruments are given by the conditional expectation of $f(x_i)$ given z_i: $w^* \equiv \omega(z_i) = \mathbf{E}(f(x_i)|z_i)$. To see this, first observe that the asymptotic covariance matrix for the IV estimator using instruments w_i that are any specified transformations of z_i is

$$\sigma^2[(\mathbf{E}w'f(x))'(\mathbf{E}w'w)^{-1}(\mathbf{E}w'f(x))]^{-1}. \text{ But } \mathbf{E}w'f(x) = \mathbf{E}_z w'\mathbf{E}_{x|z}f(x) = \mathbf{E}_z w'w^*.$$

The asymptotic covariance matrix of this IV estimator can be written

[2] The usual limiting regularity conditions are assumed to hold, as in Section 3, and the parameter θ is assumed to be identified in the sense the mapping from θ to β is one-to-one for β in its range.

$$\sigma^2[(\mathbf{E}w'w^*)'(\mathbf{E}w'w)^{-1}(\mathbf{E}w'w^*)]^{-1}.$$

If w = w^*, this covariance matrix reduces to $\sigma^2(\mathbf{E}w^{*\prime}w^*)^{-1}$. It is a standard exercise to show that w = w^* minimizes the asymptotic variance. Let F = $\mathbf{E}w^{*\prime}w^*$, G = $\mathbf{E}w'w^*$, and H = $\mathbf{E}w'w$. Then the quadratic form

$$[I\ \ -G'H^{-1}]\cdot \begin{bmatrix} F & G' \\ G & H \end{bmatrix} \cdot[I\ \ -G'H^{-1}]' = F - G'H^{-1}G$$

is positive semidefinite, which implies that $[G'H^{-1}G]^{-1} - F^{-1}$ is positive semidefinite. From this result, the IV estimator using the instruments w^* is called the best nonlinear 2SLS estimator (BN2SLS).

In general, the BN2SLS estimator is not practical in applications because computation of the conditional expectation $\mathbf{E}_{x|z}f(x)$ is intractable. Obviously, in any application where direct computation of $\mathbf{E}_{x|z}f(x)$ is tractable, it should be used. In the remaining cases, it is possible to approximate $\mathbf{E}_{x|w}f(x)$. A method proposed by Kelejian (1971) and Amemiya (1974) is to make an approximation in terms of low-order polynomials in the raw instruments z; i.e., regress $f(x_i)$ on z_i, squares and cross-products of components of z_i, third-order interactions, and so forth. One interpretation of this procedure is that one is making a series approximation using the leading terms in a Taylor's expansion of $\mathbf{E}_{x|w}f(x)$, or in other words the low order conditional moments of x given w. This method can be implemented in the LSQ procedure in TSP by expanding the list of specified instruments in the command to include the desired low-order polynomials in the raw instruments. Viewed more generally, the expression $\mathbf{E}_{x|z}f(x)$ can be written as

$$(21) \qquad \mathbf{E}_{x|z}f(x) = \int_x f(x)\cdot g(x|z)\cdot dx \equiv \psi(z),$$

where g(x,z) is the joint density of x and z, and g(x|z) is the conditional density of x given z. If g(x|z) is known (or can be estimated consistently as a parametric function), but analytic computation of the integral is intractable, it may be possible to use simulation methods, drawing a "pseudo-sample" x_{ij} from $g(x|z_i)$ for j = 1,...,J and estimating $\mathbf{E}_{x|z}f(x)$ as the mean of $f(x_{ij})$ in this pseudo-sample. If the pseudo-sample size J grows at a sufficient rate with sample size (typically, faster than $N^{1/2}$), then IV using this approximation will have the same asymptotic covariance matrix as BN2SLS. If the conditional density is itself not known or tractable, it may be possible to estimate it nonparametrically, say using a kernel estimator; see Chapter 7. Alternately, viewing ψ(z) as a nonparametric function of z, the problem can be approached as a nonparametric regression $f(x_i) = \psi(z_i) + \zeta_i$, and ψ estimated by a variety of nonparametric procedures; again see Chapter 7. In particular, one approach to nonparametric regression is series approximation, where $\psi(z_i)$ is approximated by a linear combination of initial terms in a series approximation. In particular, the Kelejian-Amemiya method falls within this class, and nonparametric estimation theory provides a guide to choice of the truncation level as a function of sample size. The bottom line is that by simulation or nonparametric procedures, one may be able to "adaptively" achieve the asymptotic covariance matrix of the BN2SLS estimator without having to solve an intractable problem of determining $\mathbf{E}_{x|z}f(x)$ analytically. Existing software may not be sufficiently "adaptive" to automatically achieve the BN2SLS asymptotic efficiency level, so that it is up to the user to specify instruments in a form that achieves this adaptation. In practice, the issue of adaptiveness has no real bite in determining a good set of instruments in a given finite data set, and the properties of the asymptotic approximation may not tell you much about the actual finite-sample distribution of your estimators. Bootstrap methods, discussed in Chapter 7, may be one useful way to give a better approximation to finite-sample distributions and guide choice among estimators using different sets of instruments.

Finally, consider models that are nonlinear in both variables and parameters, $y = h(x,\theta) + \varepsilon$. First observe that if there are proper raw instruments z, then minimizing the GMM criterion

$$(22) \qquad \left[\sum_{i=1}^{N} z_i(y_i - h(x_i,\theta))\right]'\cdot \left[\sum_{i=1}^{N} z_i z_i'\right]^{-1}\cdot \left[\sum_{i=1}^{N} z_i(y_i - h(x_i,\theta))\right]$$

in θ will produce a consistent initial estimator θ_N for θ. There is an iterative procedure that can be used to calculate θ_N. From starting values $\theta^{(0)}$, suppose one has reached $\theta^{(r)}$. Linearize the model about $\theta^{(r)}$, obtaining

$$(23) \qquad y_i - h(x_i,\theta^{(r)}) = f^{(r)}(x_i)\cdot(\theta - \theta^{(r)}) + \upsilon_i,$$

where $f^{(r)}(x_i) = \nabla_\theta h(x_i,\theta^{(r)})$ and υ_i is a disturbance that includes the remainder from the linear approximation. Apply conventional 2SLS to this model, with the instruments z_i. The estimated coefficients provide the adjustments that produce the next iterate $\theta^{(r+1)}$. For a suitably chosen starting point, the iterates $\theta^{(r)}$ will converge to a limit at θ_N. It may be necessary to consider alternative starting values to obtain convergence to the minimand of the GMM criterion.

Start from the consistent initial estimator θ_N, and the linearized model (23) evaluated at θ_N, with $f_N(x) = \nabla_\theta h(x_i,\theta_N)$. Treating θ_N as a vector of constants, (23) now has the same form as the model that is nonlinear in variables but linear in parameters that was discussed above. As in the previous case, estimate this model using 2SLS and an approximation to the best instruments $\mathbf{E}_{x|z}f_N(x)$; this will approximate the BN2SLS estimator. This procedure, with the best instruments approximated by user-specified combinations of the raw instrumental variables, is used by the LSQ command in TSP. It is possible to iterate the procedure described in this paragraph, but the first application of the procedure is already asymptotically equivalent to the BN2SLS estimator (provided the approximation to the best instruments is adaptive), and there is no further gain in (first-order) asymptotic efficiency from iteration.

Exercise 1. The usual asymptotic analysis of IV estimators assumes that the full correlation condition holds in the limit (see Section 3, assumption [3]). In some applications, the degree of correlation of instruments and explanatory variables is so weak that this is a poor asymptotic approximation, and a better one is $X'W/n - G_1 - G_2/n^{1/2} \to_p 0$, where G_1 is a matrix of rank less than k, but for each finite n, $G_1 + G_2/n^{1/2}$ is of full rank k. What is the limiting distribution of the IV estimator under this "asymptotically weak instrument" assumption? Do the analysis for the simple case of a single explanatory variable that is contaminated and a single instrument.

Exercise 2. An econometrician implements an IV procedure by first running a OLS regression of the contaminated explanatory variables on a <u>subset</u> of the available instruments that <u>excludes</u> some uncontaminated explanatory variables, then a second stage OLS. Is this consistent?