Exogenicity Test – A Simple Case

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. Partition $X = [X_1\ X_2]$, where X_1 is n×p. Suppose that X_2 is believed to be orthogonal to the disturbance ϵ in the population, but that X_1 is suspected of *contamination*, making it non-orthogonal to ϵ in the population. This can occur, for example, if ε contains omitted variables that are correlated with the included variables in X_1, if X_1 contains measurement errors, or if X_1 contains endogenous variables that are determined jointly with y.

Suppose that there is a n×m array of proper instruments $Z = [V\ X_2]$, where V is a n×r array of instruments that are excluded from X, and one has $m \geq n$, or equivalently $r \geq p$. If X_1 is clean, then the broader array $W = [Z\ X_1] = [V\ X]$ also constitute proper instruments.

Suppose one calculates a 2SLS estimator of β using either the narrow array of instruments Z, or the broad array of instruments W. The first estimator will always be consistent, while the second estimator will be consistent only if all the instruments in W are proper; i.e., if X_1 is clean. On the other hand, if X_1 is clean, the second estimator will be more efficient.

The first estimator is

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

The second estimator using W as instruments simplifies to b_{OLS}. To see this, think of doing 2SLS with W as instruments by first regressing X on W, and then regressing y on the fitted values of X from the first stage. But W includes X, so that the first stage returns fitted values for X equal to the observed values, and the second stage is just the OLS regression of y on X. The test statistic for the Hausman exogenicity test then specializes to

$$(b_{2SLS} - b_{OLS})'[V(b_{2SLS}) - V(b_{OLS})]^{-}(b_{2SLS} - b_{OLS}),$$

where $[\cdot]^{-}$ denotes a generalized inverse. When X is not contaminated, so that W is clean, this statistic is 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, as an omitted variable test with appropriately constructed auxiliary variables, is more convenient to compute, and is asymptotically equivalent to the Hausman test statistic.

First do an OLS regression of X_1 on Z and retrieve fitted values $X_1^* = Z(Z'Z)^{-1}Z'X_1$. Second, do an OLS regression on the auxiliary regression model $y = X\beta + X_1^*\gamma + \eta$. Then test the null hypothesis that the coefficients γ are zero; i.e., an omitted variable test for the variables X_1^*. This test can be done as a conventional F-test for omitted variables. The numerator degrees of freedom will be p.

We next show that this test is indeed an exogeneity test. First, the OLS estimates of the parameters in the model $y = X\beta + X_1^*\gamma + \eta$ satisfy

$$\begin{bmatrix} b \\ c \end{bmatrix} = \begin{bmatrix} X'X & X'QX_1 \\ X_1'QX & X_1'QX_1 \end{bmatrix}^{-1} \begin{bmatrix} X'y \\ X_1'Qy \end{bmatrix} = \begin{bmatrix} \beta \\ 0 \end{bmatrix} + \begin{bmatrix} X'X & X'QX_1 \\ X_1'QX & X_1'QX_1 \end{bmatrix}^{-1} \begin{bmatrix} X'\varepsilon \\ X_1'Q\varepsilon \end{bmatrix}.$$

Where $Q = Z(Z'Z)^{-1}Z'$, implying $QX_1 = X_1^*$, and $QX_2 = X_2$. But $X'Q\varepsilon/n \rightarrow_p \text{plim}(X'Z/n)\cdot(\text{plim}(Z'Z/n))^{-1}\cdot\text{plim}(Z'\varepsilon/n) = 0$ when Z is clean. Similarly, $X'\varepsilon/n \rightarrow_p 0$ when X_1 is clean and $X_2'\varepsilon/n \rightarrow_p 0$ since X_2 is clean by assumption, but $X_1'\varepsilon/n \rightarrow_p C \neq 0$ when X_1 is contaminated. Define

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

From the formula for a partitioned inverse,

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

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

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

Hence,

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

If X_1 is clean, then $c_P \rightarrow_p 0$ and $n^{1/2}c_P$ is asymptotically normal. On the other hand, if X_1 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_{OLS} and b_{2SLS}. Recall that $b_{2SLS} = \beta + (X'QX)^{-1}X'Q\varepsilon$ and $b_{OLS} = \beta + (X'X)^{-1}X'\varepsilon$. Then

$$(X'QX)(b_{2SLS} - b_{OLS}) = \{X'Q\varepsilon/n - (X'QX)(X'X)^{-1}\cdot X'\varepsilon/n\}. \quad (17)$$

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

$$A_{22}(X_1'QX)(b_{2SLS} - b_{OLS}) = A_{22}\{X_1'Q\varepsilon/n - (X_1'QX)(X'X)^{-1}\cdot X'\varepsilon/n\} = c_P.$$

Thus, c_P is a linear transformation of $(b_{2SLS} - b_{OLS})$. Then, testing whether c_P is near zero is equivalent to testing whether a linear transformation of $(b_{2SLS} - b_{OLS})$ is near zero. When X_1 is of rank p, this equivalence establishes that the Hausman test in its original form is the same as the test that c_P is zero.

RELATION TO GMM TEST FOR OVER-IDENTIFICATION

Let $W = [Z\ X_1] = [V\ X]$ be all the variables that are orthogonal to ε in the population under the null hypothesis that X and ε are uncorrelated. Let P_W denote the projection operator onto the subspace spanned by W; i.e., $P_W = W(W'W)^{-1}W'$. 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 $2nQ_n = y'(P_W - P_X)y/\sigma^2$. This statistic is the same as the test statistic for the hypothesis that the coefficients of X_1^* are zero in a regression of y on X and X_1^*; 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 X_1^* 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 $M = [V\ X_2]$. Write

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

where $Q_M = I - P_M$. 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

$\begin{bmatrix} X_1'P_MQ_Xy \\ 0 \end{bmatrix} = (X'P_MX)(b_{2SLS} - b_{OLS}) \equiv \begin{bmatrix} X_1'P_MX_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{bmatrix} b_{1,2SLS} - b_{1,OLS} \\ b_{2,2SLS} - b_{2,OLS} \end{bmatrix}$. 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

$$[X_1'P_MX_1 - X_1'X_2(X_2'X_2)^{-1}X_2'X_1](b_{1,2SLS} - b_{1,OLS}) = X_1'P_MQ_Xy$$

The matrix on the left-hand-side can be rewritten as $X_1'P_M\ Q_{X_2}\ P_MX_1$, so that

$$b_{1,2SLS} - b_{1,OLS} = (X_1'P_M \ \ Q_{X_2} \ \ P_MX_1)^{-1}X_1'P_MQ_Xy.$$

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

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

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 $P_MX_2 = X_2$ and hence $Q_XP_MX_2 = 0$, one has

$$X'P_MQ_XP_MX = \begin{bmatrix} X_1'P_MQ_XP_MX_1 & 0 \\ 0 & 0 \end{bmatrix}.$$

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 $\begin{bmatrix} X_1'P_NX_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1}$ is $(X_1'P_NQ_{X_2}P_NX_1)^{-1}$. Then,

$$V(b_{1,2SLS} - b_{1,OLS}) = \sigma^2(X_1'P_M \ \ Q_{X_2} \ \ P_MX_1)^{-1}X_1'P_MQ_XP_MX_1(X_1'P_M \ \ Q_{X_2} \ \ P_MX_1)^{-1}.$$

Using the expressions above, the quadratic form can be written

$$\begin{aligned} &(b_{1,2SLS} - b_{1,OLS})V(b_{1,2SLS} - b_{1,OLS})^{-1}(b_{1,2SLS} - b_{1,OLS}) \\ &= y'Q_XP_MX_1(X_1'P_MQ_XP_MX_1)^{-1}X_1'P_MQ_Xy/\sigma^2. \end{aligned}$$

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

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

so that the two statistics coincide.