A comment on under-identification 15 August 2007

D A Freedman Statistics 215

Let (P_i, X_i, ϵ_i) be IID, jointly normal, with positive variances, and $E(P_i) = E(X_i) = E(\epsilon_i) = 0$. Suppose P_i and X_i are correlated, as are P_i and ϵ_i; however, X_i and ϵ_i are uncorrelated, i.e., $X_i \perp \epsilon_i$, viz., $E(X_i \epsilon_i) = 0$. Thus, P_i is "endogenous" and X_i is "exogenous." (For jointly normal variables, uncorrelated and independent are synonymous.) Let a, b be real parameters, and $Q_i = aP_i + bX_i + \epsilon_i$. We think of Q_i, P_i, X_i as observable, ϵ_i as unobservable.

Claim. The parameters a, b cannot be identified from the joint distribution of Q_i, P_i, X_i.

Let $\alpha = \text{cov}(X_i, P_i)/\text{var}(X_i)$, so that $\delta_i = P_i - \alpha X_i \perp X_i$. Check that $\delta_i \neq 0$—otherwise, P_i would be exogenous. Let c be a real number. Check that

$$Q_i = (a - c)P_i + (b + \alpha c)X_i + (c\delta_i + \epsilon_i)$$

and $X_i \perp c\delta_i + \epsilon_i$. Thus, (a, b) and $(a - c, b + \alpha c)$ lead to the same joint distribution for the observables, Q_i, P_i, X_i. Matters would be otherwise, of course, if ϵ_i were observable—but it isn't, so it is legitimate to change the disturbance term along with the parameters.

The extension to p-dimensional X_i is easy. Suppose X_i is $p \times 1$, and $C = \text{cov}(X_i)$ is full rank; C is a $p \times p$ matrix. Let $D = \text{cov}(X_i, Y_i)$, viewed as a $p \times 1$-vector. We continue to assume that (P_i, X_i, ϵ_i) are IID and jointly normal, with expectation 0; that P_i and ϵ_i have positive variance, that P_i and X_i are correlated ($D \neq 0$), as are P_i and ϵ_i; that $X_i \perp \epsilon_i$. Let a be scalar whilst b is $p \times 1$. Let $\alpha = C^{-1}D$. The rest of the construction is the same: $Q_i = aP_i + X_i b + \epsilon_i$.

Take II

Let's redo this from a slightly different perspective. Again, units are IID. For a typical unit, the response variable is Y, a scalar. The $1 \times p$ vector of explanatory variables is X, which may be endogenous. There is $1 \times q$ vector of variables Z, which are proposed for use as instruments, with $q \geq p \geq 1$. The (unobservable) disturbance term is ϵ. The variables Z, X, Y are assumed to be jointly normal, with expectation 0. Let Γ be the variance-covariance matrix of Z, X, Y; this is assumed to have rank $q + p + 1$, and the $q \times p$ matrix $M = E(Z'X)$ is assumed to have rank p. Notice that Γ determines—and is determined by—the joint distribution of the observables Z, X, Y. The matrix M is a sub-matrix of Γ.

Let $\alpha = E(Z'\epsilon)$; this is a $q \times 1$ vector of nuisance parameters. Let β be $p \times 1$ with

$$Y = X\beta + \epsilon \tag{1}$$

This β is a parameter vector.

Claim. Γ does not determine α or β.

Choose any β whatsoever; then simply define $\epsilon = Y - X\beta$. Thus, Γ does not determine β. Let $N = E(Z'Y)$, a $q \times 1$ sub-matrix of Γ. Let H be the column space of M translated by N; this is a p-dimensional hyperplane in R^q. Plainly, $\alpha = E(Z'\epsilon) = E(Z'Y) - M\beta = N - M\beta$ is in H. Because M has rank p, as β runs through all p vectors, α runs through all of H; thus, α cannot be determined from Γ, which completes the proof.

Interestingly, if $0_{q\times 1} \notin H$—i.e., α cannot be $0_{q\times 1}$—then Z cannot be exogenous. If $0_{q\times 1} \in H$, then Z can be exogenous, but need not be so. After all, H is p-dimensional, and $0_{q\times 1}$ is but a single point. In short, additional information is needed to determine exogeneity, beyond the joint distribution of the observables.

Corollary. Γ can determine that $\alpha \neq 0$; however, Γ cannot determine that $\alpha = 0$.

To get a specific example where Γ determines that $\alpha \neq 0$, take $q = 2$ and $p = 1$. Let $X = \theta_1 Z_1 + \theta_2 Z_2 + U$ and $Y = \psi_1 Z_1 + \psi_2 Z_2 + X + U + V$. Here, Z_1, Z_2, U, V are independent standard normal variables, $\theta_1, \theta_2, \psi_1, \psi_2$ are free parameters. Since

$$Y = (\theta_1 + \psi_1)Z_1 + (\theta_2 + \psi_2)Z_2 + 2U + V$$

we have

$$M = E(Z'X) = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \qquad N = E(Z'Y) = \begin{pmatrix} \theta_1 + \psi_1 \\ \theta_2 + \psi_2 \end{pmatrix}$$

Thus, N is in the column space of M—i.e., N is proportional to M—only if (ψ_1, ψ_2) is proportional to (θ_1, θ_2). On the other hand, suppose in equation (1) that the "structural parameter" is $\beta = 1$, and $\epsilon = U + V$. Then X is indeed endogenous, being correlated with ϵ. But Z_1 and Z_2 can be used as instruments only when $\psi_1 = \psi_2 = 0$; otherwise, the "exclusion restrictions" are violated, i.e., Z_1 and Z_2 should appear in the equation.