The biases in the OLS estimates of a model like (8.10) arise because the error terms are correlated with some of the regressors. The IV estimator solves this problem asymptotically, because the projections of the regressors on to $\mathcal{S}(\boldsymbol{W})$ are asymptotically uncorrelated with the error terms. However, there must always still be some correlation in finite samples, and this causes the IV estimator to be biased.

Systems of Equations

In order to understand the finite-sample properties of the IV estimator, we need to consider the model (8.10) as part of a system of equations. We therefore change notation somewhat and rewrite (8.10) as

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta}_1 + \boldsymbol{Y}\boldsymbol{\beta}_2 + \boldsymbol{u}, \quad \mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top) = \sigma^2\mathbf{I}, \tag{8.38}$$

where the matrix of regressors $\boldsymbol{X}$ has been partitioned into two parts, namely, an $n \times k_1$ matrix of exogenous and predetermined variables, $\boldsymbol{Z}$, and an $n \times k_2$ matrix of endogenous variables, $\boldsymbol{Y}$, and the vector $\boldsymbol{\beta}$ has been partitioned conformably into two subvectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. There are assumed to be $l \geq k$ instruments, of which k_1 are the columns of the matrix $\boldsymbol{Z}$.

The model (8.38) is not fully specified, because it says nothing about how the matrix $\boldsymbol{Y}$ is generated. For each observation t, $t = 1, \ldots, n$, the value y_t of the dependent variable and the values $\boldsymbol{Y}_t$ of the other endogenous variables are assumed to be determined by a set of linear simultaneous equations. The variables in the matrix $\boldsymbol{Y}$ are called **current endogenous variables**, because they are determined simultaneously, row by row, along with $\boldsymbol{y}$. Suppose that all the exogenous and predetermined explanatory variables in the full set of simultaneous equations are included in the $n \times l$ instrument matrix $\boldsymbol{W}$, of which the first k_1 columns are those of $\boldsymbol{Z}$. Then, as can easily be seen by analogy with the explicit result (8.09) for the demand-supply model, we have for each endogenous variable $\boldsymbol{y}_i$, $i = 0, 1, \ldots, k_2$, that

$$\boldsymbol{y}_i = \boldsymbol{W}\boldsymbol{\pi}_i + \boldsymbol{v}_i, \quad \mathrm{E}(v_{ti} \,|\, \boldsymbol{W}_t) = 0. \tag{8.39}$$

Here $\boldsymbol{y}_0 \equiv \boldsymbol{y}$, and the $\boldsymbol{y}_i$, for $i = 1, \ldots, k_2$, are the columns of $\boldsymbol{Y}$. The $\boldsymbol{\pi}_i$ are l–vectors of unknown coefficients, the $\boldsymbol{v}_i$ are n–vectors of error terms that are innovations with respect to the instruments, v_{ti} is the t^{th} element of $\boldsymbol{v}_i$, and $\boldsymbol{W}_t$ is the t^{th} row of $\boldsymbol{W}$.

Equations like (8.39), which have only exogenous and predetermined variables on the right-hand side, are called **reduced form equations**, in contrast with equations like (8.38), which are called **structural equations**. Writing a model as a set of reduced form equations emphasizes the fact that all the endogenous variables are generated by similar mechanisms. In general, the error terms for the various reduced form equations display **contemporaneous correlation**: If v_{ti} denotes a typical element of the vector $\boldsymbol{v}_i$, then, for observation t, the reduced form error terms v_{ti} are generally correlated among themselves and correlated with the error term u_t of the structural equation.

A Simple Example

In order to gain additional intuition about the properties of the IV estimator in finite samples, we consider the very simplest nontrivial example, in which the dependent variable $\boldsymbol{y}$ is explained by only one variable, which we denote by $\boldsymbol{x}$. The regressor $\boldsymbol{x}$ is endogenous, and there is available exactly one exogenous instrument, $\boldsymbol{w}$. In order to keep the example reasonably simple, we suppose that all the error terms, for both $\boldsymbol{y}$ and $\boldsymbol{x}$, are normally distributed. Thus the DGP that simultaneously determines $\boldsymbol{x}$ and $\boldsymbol{y}$ can be written as

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{x}\beta_0 + \sigma_u \boldsymbol{u}, \\ \boldsymbol{x} &= \boldsymbol{w}\pi_0 + \sigma_v \boldsymbol{v}, \end{aligned} \tag{8.40}$$

where the second equation is analogous to (8.39). By explicitly writing σ_u and σ_v as the standard deviations of the error terms, we can define the vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ to be multivariate standard normal, that is, distributed as $\mathrm{N}(\mathbf{0}, \mathbf{I})$. There is contemporaneous correlation of $\boldsymbol{u}$ and $\boldsymbol{v}$. Therefore, $\mathrm{E}(u_t v_t) = \rho$ for some correlation coefficient ρ such that $-1 < \rho < 1$. The result of Exercise 4.4 shows that the expectation of u_t conditional on v_t is ρv_t, and so we can write $\boldsymbol{u} = \rho\boldsymbol{v} + \boldsymbol{u}_1$, where $\boldsymbol{u}_1$ has mean zero conditional on $\boldsymbol{v}$.

In this simple, just identified, setup, the IV estimator of the parameter β is

$$\hat{\beta}_{\mathrm{IV}} = (\boldsymbol{w}^\top\boldsymbol{x})^{-1}\boldsymbol{w}^\top\boldsymbol{y} = \beta_0 + \sigma_u(\boldsymbol{w}^\top\boldsymbol{x})^{-1}\boldsymbol{w}^\top\boldsymbol{u}. \tag{8.41}$$

This expression is clearly unchanged if the instrument $\boldsymbol{w}$ is multiplied by an arbitrary scalar, and so we can, without loss of generality, rescale $\boldsymbol{w}$ so that $\boldsymbol{w}^\top\boldsymbol{w} = 1$. Then, using the second equation in (8.40), we find that

$$\hat{\beta}_{\mathrm{IV}} - \beta_0 = \frac{\sigma_u \boldsymbol{w}^\top\boldsymbol{u}}{\pi_0 + \sigma_v \boldsymbol{w}^\top\boldsymbol{v}} = \frac{\sigma_u \boldsymbol{w}^\top(\rho\boldsymbol{v} + \boldsymbol{u}_1)}{\pi_0 + \sigma_v \boldsymbol{w}^\top\boldsymbol{v}}.$$

Let us now compute the expectation of this expression conditional on $\boldsymbol{v}$. Since, by construction, $\mathrm{E}(\boldsymbol{u}_1 \,|\, \boldsymbol{v}) = \mathbf{0}$, we obtain

$$\mathrm{E}(\hat{\beta}_{\mathrm{IV}} - \beta_0 \,|\, \boldsymbol{v}) = \frac{\rho\sigma_u}{\sigma_v}\frac{z}{a + z}, \tag{8.42}$$

where we have made the definitions $a \equiv \pi_0/\sigma_v$, and $z \equiv \boldsymbol{w}^\top\boldsymbol{v}$. Given our rescaling of $\boldsymbol{w}$, it is easy to see that $z \sim \mathrm{N}(0, 1)$.

When $\rho = 0$, the right-hand side of equation (8.42) vanishes, and so, conditional on $\boldsymbol{v}$, $\hat{\beta}_{\mathrm{IV}}$ is unbiased. In fact, since $\boldsymbol{v}$ is independent of $\boldsymbol{u}$ in this case, and $\boldsymbol{w}$ is exogenous, it follows that $\boldsymbol{x}$ is itself exogenous. With both $\boldsymbol{x}$ and $\boldsymbol{w}$ exogenous, the IV estimator is like the estimators dealt with in Exercise 3.17, which are unbiased conditional on these exogenous variables. If $\rho \neq 0$, however, $\boldsymbol{x}$ is not exogenous, and the estimator is biased conditional on $\boldsymbol{v}$. The *unconditional* expectation of the estimator does not even exist. To see this, let us try to calculate the expectation of the random variable $z/(a+z)$. If the expectation existed, it would be

$$\mathrm{E}\Big(\frac{z}{a+z}\Big) = \int_{-\infty}^{\infty} \frac{x}{a+x}\,\phi(x)\,dx, \tag{8.43}$$

where, as usual, $\phi(\cdot)$ is the density of the standard normal distribution. It is a fairly simple calculus exercise to show that the integral in (8.43) diverges in the neighborhood of $x = -a$.

If $\pi_0 = 0$, then $a = 0$. In this extreme case, the model is not asymptotically identified, and $\boldsymbol{x} = \sigma_v \boldsymbol{v}$ is just noise, as though it were an error term. As a consequence, $\boldsymbol{w}$ is not a valid instrument, and the IV estimator is inconsistent.

When $a \neq 0$, which is the usual case, the IV estimator (8.41) is neither biased nor unbiased, because it has no expectation for any finite sample size n. This may seem to contradict the result according to which $\hat\beta_{\rm IV}$ is asymptotically normal, since all the moments of the normal distribution exist. However, the fact that a sequence of random variables converges to a limiting random variable does not necessarily imply that the *moments* of the variables in the sequence converge to those of the limiting variable; see Davidson and MacKinnon (1993, Section 4.5). The estimator (8.41) is a case in point. Fortunately, this possible failure to converge of the moments does not extend to the CDFs of the random variables, which do indeed converge to that of the limit. Consequently, P values and the upper and lower limits of confidence intervals computed with the asymptotic distribution are legitimate approximations, in the sense that they become more and more accurate as the sample size increases.

A less simple calculation can be used to show that, in the overidentified case, the first $l-k$ moments of $\hat{\boldsymbol{\beta}}_{\rm IV}$ exist; see Kinal (1980). This is consistent with the result we have just obtained for an exactly identified model, where $l-k=0$, and the IV estimator has no moments at all. When the mean of $\hat{\boldsymbol{\beta}}_{\rm IV}$ exists, it is almost never equal to $\boldsymbol{\beta}_0$. Readers will have a much clearer idea of the impact of the existence or nonexistence of moments, and of the bias of the IV estimator, if they work carefully through Exercises 8.10 to 8.13, in which they are asked to generate by simulation the EDFs of the estimator in different situations.

The General Case

We now return to the general case, in which the structural equation (8.38) is being estimated, and the other endogenous variables are generated by the reduced form equations (8.39) for $i = 1, \ldots, k_2$, which correspond to the first-stage regressions for 2SLS. We can group the vectors of fitted values from these regressions into an $n \times k_2$ matrix $\boldsymbol{P_W Y}$. The generalized IV estimator is then equivalent to a simple IV estimator that uses the instruments $\boldsymbol{P_W X} = [\boldsymbol{Z}\;\; \boldsymbol{P_W Y}]$. By grouping the l–vectors $\boldsymbol{\pi}_i$, $i = 1, \ldots, k_2$ into an