CHAPTER 1

Motivation

1.1 Introduction

The method of instrumental variables (IV) has traditionally been viewed as a response to a common problem in regression contexts, namely where one or more of the regressors on the right-hand side of the proposed equation are correlated with the equation disturbance. If this happens, the method of ordinary least squares suffers from consistency problems. The instrumental variables methods were developed to overcome these problems. It could legitimately be objected that the focus on consistency alone as a criterion of statistical effectiveness is misplaced. Thus it is often the case that estimators that are consistent possess inferior mean-square error properties to those that are not. Remarkably enough, however, the IV methodology can in many circumstances provide estimators that have superior efficiency properties all round. Indeed, it will be one of the themes of the later chapters of this book that the method of maximum likelihood may, in certain contexts of importance, itself be regarded as an instrumental variables estimator, so that IV estimators are asymptotically fully efficient. In the present chapter, however, we shall lower our sights a little and consider the motivation for instrumental variables as arising from requirements of statistical consistency.

On regressors and disturbances

Although the method of instrumental variables was not originally developed in the specific context of regression theory, it has since been viewed essentially as an adjunct to regression analysis. This is certainly the

easiest way to motivate the method and is, by and large, the path toward its exposition followed throughout this book. Putting historical matters temporarily aside, let us begin by considering circumstances under which standard regression methods do not work.

Suppose that we are to fit the very simple regression model

$$y_i = \beta x_i + u_i, \qquad i = 1, \ldots, n, \tag{1.1}$$

where the u_i are disturbances that we shall suppose have zero expectation and common variance σ^2. We do not at this stage impose the assumption that successive disturbances are serially uncorrelated. Likewise, we shall not yet impose any particular specification or structure on the right-hand variables x_i. Now the least-squares estimator of the parameter β in equation (1.1) is

$$b = \sum y_i x_i / \sum x_i^2 = \beta + \sum x_i u_i / \sum x_i^2, \tag{1.2}$$

where we have utilized $y_i = \beta x_i + u_i$. If the x_i can be regarded as drawings from a distribution function of a random variable that is independent of the u_i, then the sum $(1/n) \sum x_i u_i$ tends in probability to zero and we assume that the sum $(1/n) \sum x_i^2$ tends in probability to a constant, say, σ_x^2. The same will be true, with reinterpretations as appropriate, if the x_i are regarded as fixed, that is, nonrandom variates that obey suitable regularity conditions. (At this point we shall not be explicit about what constitutes such conditions.) Under either of these circumstances, it will follow that the least-squares estimate b tends in probability to β; indeed, the convergence will be almost sure.[1]

Suppose, however, that the observations x_i are random variables that are correlated with the disturbances u_i, at least in the weak sense that $\operatorname{plim}(1/n) \sum x_i u_i \neq 0$. We shall tacitly assume that the limit in probability exists and is nonzero. It follows that the least squares estimate b does not tend in probability to the true value β; or more precisely, we cannot assert that it will. The term $\operatorname{plim}(1/n) \sum x_i u_i / \sigma_x^2$ represents the inconsistency of the estimator, namely $\operatorname{plim}(b - \beta)$.

The method of instrumental variables is basically a way of proceeding with remedial action in such situations, which possesses the merit of consistency, and which does so in a manner that preserves the regression framework, in a sense that will become clear as development proceeds. However, before turning to these methods, it is useful to review certain situations that do exhibit the problem of regressor–disturbance correlation. We do this for two reasons: first, because to do so now will save us some rather elementary exposition that would unduly interrupt the analysis at later stages, and second, to demonstrate that this particular problem occurs in a wide variety of models. Since these models are all

amenable in varying degree to instrumental variables estimation, this will help to motivate the development of a general IV methodology.

All four applications are familiar to students of econometrics. We do suggest, however, that such readers take time out to look at the errors in variables material, which as a matter of interpretation improves on existing textbook discussions, and at the self-selection problem, which is discussed at length in Section 2.6.

Errors in variables and latent variables

Suppose that we have a bivariate data series x_i, y_i; $i = 1, \ldots, n$, of mutually independent random variables with means χ_i, $\alpha + \beta\chi_i$ and variances ϕ^2, σ^2, respectively. An equivalent structural formulation is

$$x_i = \alpha + \beta\chi_i + u_i; \qquad \mathcal{E}u_i = 0, \quad \mathcal{E}u_i^2 = \sigma^2; \tag{1.3a}$$

$$x_i = \chi_i + \epsilon_i; \qquad \mathcal{E}\epsilon_i = 0, \quad \mathcal{E}\epsilon_i^2 = \phi^2; \tag{1.3b}$$

with in addition $\mathcal{E}u_i\epsilon_i = 0$. This is the classic errors-in-variables model, the idea being that the true values of the variables are observed only with errors (u_i, ϵ_i). These "errors" may represent not only errors of observation but also a substantive hypothesis about the behavior of the agents concerned. The most celebrated of this latter class of theories is Friedman's (1957) permanent-income hypothesis. Here y_i would denote observed consumption and χ_i would be "permanent income," representing a smoothing by the household of temporary fluctuations in accordance with some desired pattern of lifetime consumption. Observed income differs from permanent income by a supposedly random disturbance ϵ_i. Although not a particularly convincing piece of consumption theory, models of this kind have been empirically fitted by Friedman and others.

Now the variates χ_i that appear in the errors in variables model (1.3) have the status of unknown parameters, even if of the nuisance or incidental variety, and it is readily apparent that the parameter space is potentially of infinite dimension. It is this fact that has led, over the years, to a voluminous statistical literature on an apparently very simple model. Only relatively recently has a fuller understanding of its properties been arrived at.

To start with, it has long been known that the method of maximum likelihood fails for this model. Some writers have labeled this failure an identifiability problem, but this is not really the case. To see this, denote by θ the vector $(\beta, \sigma^2, \phi^2; \chi_i, i = 1, \ldots, n)'$ of parameters and write the normal log-likelihood function (apart from a constant) in the form

$$\frac{1}{n}l(\theta) = -\left[\log(\sigma^2\phi^2) + \frac{1}{n\phi^2}\sum_i (x_i - \chi_i)^2 + \frac{1}{n\sigma^2}\sum_i (y_i - \beta\chi_i)^2\right]. \tag{1.4}$$

For simplicity we have assumed $\alpha = 0$. The sample size is written as n, a convention that will be followed, except where otherwise indicated, throughout the book. Let us fix the sample size n and consider the identifiability of the parameters. Following the Kullback–Bowden information criterion for identification (Bowden 1973), let θ_0 denote the true parameter values and denote by $\mathcal{E}^0$ the operation of taking expectations, on the premise that the true parameter values are θ_0. By using decompositions such as $x_i - \chi_i = (x_i - \chi_i^0) + (\chi_i^0 - \chi_i)$, we derive

$$h(\theta;\theta_0) = \mathcal{E}^0\left[\frac{1}{n}l(\theta)\right] = -\left[\log(\sigma^2\phi^2) + \frac{\phi_0^2}{\phi^2} + \frac{\sigma_0^2}{\sigma^2} + \frac{1}{n\phi^2}\sum_i(\chi_i - \chi_i^0)^2 + \frac{1}{n\sigma^2}\sum_i(\beta\chi_i - \beta_0\chi_i^0)^2\right]. \tag{1.5}$$

The parameter values θ_0 will be globally identified if and only if the function $h(\theta;\theta_0)$ has an isolated global maximum at $\theta = \theta_0$. It is apparent by an inspection of expression (1.5) that such a global maximum is unique and is attained at $\theta = \theta_0$, a conclusion that can be verified by the usual methods of differential calculus.

There is therefore no lack of identification in the model, a fact that has been pointed out by previous commentators. Suppose, however, that one proceeds to maximize the likelihood function (1.4) as a problem in estimation. It turns out that there are two critical (stationary) points, according to the sign of β, and the value of the likelihood function is the same at each. Conventionally, one chooses the sign of β so that $\beta(\sum x_i y_i) > 0$. Apart from this problem of arbitrariness, however, further problems arise. By setting $\partial l/\partial\chi_i = 0$ one obtains (see Johnston 1963)

$$y_i - \hat{\beta}\chi_i = -\frac{\hat{\sigma}^2}{\hat{\beta}\hat{\phi}^2}(x_i - \chi_i).$$

Squaring and summing over i, one obtains

$$\hat{\beta}^2 = \hat{\sigma}^2/\hat{\phi}^2.$$

Clearly at least one of the maximum-likelihood estimates $\hat{\beta}^2, \hat{\sigma}^2, \hat{\phi}^2$ cannot be consistent.

Yet as we have seen, the model is identified. The resolution of this apparent paradox is that the likelihood function (1.4) is not a classical log likelihood based upon independent, identically distributed (i.i.d.) elements. It is instead a data density, constructed from probability elements that, because the means $(\chi_i, \beta\chi_i)$ differ, are independent but not

identically distributed. Now if our observations took the form of independent replications of the entire vector $(\mathbf{x}', \mathbf{y}') = (x_1 \ldots x_n, y_1 \ldots y_n)$ for fixed n, no difficulty would arise. The problem is that we have only one replication. In other contexts, the stationarity assumptions often imply that one can derive ensemble properties from a single infinite realization. The presence of an infinite number of parameters χ_i precludes this convenient property in the errors-in-variables model.

As shown by Solari (1969) (see also Sprent 1970), the upshot is that the likelihood function – or better, the data density (1.5) – does not possess a proper maximum. The critical points noted above are in fact saddle points, with the main axis of the saddle oriented along a line in χ space joining the points $\mathbf{x}$ and $(1/\beta)\mathbf{y}$. Moreover, the likelihood function can be made to assume any value between $\pm\infty$ by appropriate parameter choices. The likelihood approach can be resurrected but needs some prior information, either of the "hard-and-fast" variety (say, assuming that the ratio $\lambda = \sigma^2/\phi^2$ is known) or else by specifying an informative prior for the Bayesian analysis of the problem (e.g., Lindley and El-Sayyad 1968).

The evident failure of the method of maximum likelihood does not by itself rule out an alternative methodology. Here again, however, a certain amount of caution is called for. In particular, the pitfalls of applying ordinary least squares (OLS) to the observable variables are well known. One may write the relationship between the observables as

$$y_i = \beta x_i + v_i,$$

where the new disturbance $v_i = u_i - \beta\epsilon_i$. Let us assume that the χ_i are bounded and that $\lim(1/n)\sum \chi^2 = s_\chi^2$, say. It is then easy to show that $\operatorname{plim}(1/n)\sum x_i u_i = -\beta\sigma^2$ and that the asymptotic inconsistency of the estimator is $-\beta\phi^2/(s_\chi^2 + \phi^2)$. This is directly related to the variance of the observation error on the dependent variable.

For the errors-in-variables model, therefore, the methods of maximum likelihood and ordinary least squares both fail, the one in absolute terms and the other in its inconsistency. No estimation method is known that satisfactorily handles this problem without introducing additional information in some way. It turns out that instrumental variables methods are available that are rather weak in their informational requirements. The application of such methods to this model is discussed in detail in Section 2.8.

More general models of the type (1.3) are now used extensively in the communications and control literature. In this context, the emphasis would be on time series ($i = t$) and the χ_i would represent an unobservable signal; the object is to filter out the "noise" – in this case the

disturbances u_i, ϵ_i – to detect the signal sequence, which is usually specified as itself obeying some designated recursive relationship in time. The latter specification effectively helps to rid the model of the infinite-parameters drawback.

A somewhat different resolution to the problem of unobservable variates is followed in the literature on latent variables, the term indicating variables that are themselves unobservable or even simply imputations of some kind, but that determine, perhaps stochastically, variables that are measurable. An example is the "partial-expectations" model, studied by authors such as Duck et al. (1976), McDonald (1977), and Wickens (1982). This has the canonical form

$$\mathbf{y} = \beta \mathbf{z}^* + X\gamma + \boldsymbol{\epsilon},$$

$$\mathbf{z} = \mathbf{z}^* + \mathbf{v} = W\alpha + \mathbf{v},$$

where X and W are data matrices of observations on nonstochastic exogenous variables; $\boldsymbol{\epsilon}$ and $\mathbf{v}$ are independent, normally distributed random vectors with zero means and covariance matrices $\sigma_\epsilon^2 I$ and $\sigma_v^2 I$, respectively. The vector $\mathbf{z}^*$ represents expectations held by economic agents as to the values of the economic variable z. The data matrix W contains observations on variables, considered exogenous, that are important in the formation of expectations; it is assumed that expectations are formed in unbiased fashion, giving rise to the above specification on the error $\mathbf{v}$. Related models have been considered in other contexts, where the latent variables may represent "desired" magnitudes, "true" values, or imputed equilibrium values depending upon the application.[2] All these models share the property that, unlike the pure errors-in-variables model, one is prepared to specify the generation of the latent variables in terms of observable magnitudes, even if some contamination with noise or error occurs. More or less conventional identification problems aside, the resulting parameter space is of finite dimension.

Most latent variable models can be handled by the use of instrumental variables techniques. Observe, for instance, that the above partial-expectation model may be rewritten in the following form:

$$\mathbf{y} = \beta \mathbf{z} + X\gamma + \boldsymbol{\epsilon} - \beta \mathbf{v} = H\delta + \mathbf{u}, \tag{1.6a}$$

$$\mathbf{z} = W\alpha + \mathbf{v}, \tag{1.6b}$$

where $H = (\mathbf{z}\ X)$, $\delta = (\beta\ \gamma')'$, and $\mathbf{u} = \boldsymbol{\epsilon} - \beta \mathbf{v}$. Since the elements of H and $\mathbf{u}$ are evidently correlated (both depend upon $\mathbf{v}$), we cannot simply apply OLS directly to equation (1.6a). However, it is always open to us to use the set (X, W) as instruments for H. In particular, if the elements of W include those of X, the model (1.6) constitutes a limited-information

simultaneous model. The instrumental variables approach to such models is reviewed in detail in Section 4.3.

The self-selection problem

Suppose that we are trying to fit earnings functions that relate wages or earnings (y_i) of individual i to a set of individual characteristics represented by the vector of variables $\mathbf{x}_i$. We have observations on individuals in two sectors, say the factory sector (I) and the casual sector of employment (II), and we are interested in testing the hypothesis that the response of earnings to the variables $\mathbf{x}_i$ is the same in both sectors. We might set up the model

$$\begin{aligned} y_i &= \mathbf{x}_i'(\boldsymbol{\beta}+\boldsymbol{\delta})+u_i \qquad \text{if } i \in \text{I}, \\ &= \mathbf{x}_i'\boldsymbol{\beta}+u_i \qquad \text{if } i \in \text{II}. \end{aligned}$$

This can be expressed more compactly as

$$\begin{aligned} y_i = \mathbf{x}_i\boldsymbol{\beta}+d_i\mathbf{x}_i'\boldsymbol{\delta}+u_i, \qquad d_i &= 1 \qquad \text{if } i \in \text{I}, \\ &= 0 \qquad \text{otherwise}. \end{aligned} \tag{1.7}$$

We are interested in testing hypotheses on the elements of $\boldsymbol{\delta}$.

Suppose, however, that one of the sectors – the factory sector (I) – is viewed as being a more desirable work environment. Now the disturbance u_i represents unobservable variations in the individual's earning power that cannot be accounted for in terms of the observable variables $\mathbf{x}_i$. Given the more attractive workplace of sector I, it is rather more likely that an individual with $u_i \gg 0$ will choose sector I as his place of employment.

In terms of equation (1.7) we can interpret this preference as a positive correlation between the disturbance term u_i and the binary sector variable d_i. This means that an attempt to test hypotheses on $\boldsymbol{\delta}$ with the use of ordinary least squares will fail, since these parameters are the coefficients of those variables most at risk from regressor–disturbance correlation. The use of IV methods for this problem is discussed at length in Section 2.6.

Self-selection problems are in fact quite pervasive in empirical work. To mention just one further example, studies that purport to measure the effect of unionization on wage rates typically contain (or in effect contain) a right-hand dummy variable indicating whether or not the individual belongs to a union. The application of OLS to such an equation would be valid (in the sense of consistency) only if unionized industries exhibited no tendency to pay more. Moreover, self-selection situations

are by no means confined to representations like the model (1.7). The allocation of observations to excess demand or excess supply in simple disequilibrium models may be regarded[3] in the light of a self-selection procedure. Instrumental variables methods may also be applied to such problems, which are briefly considered in Section 5.2.

The simultaneous-equations model

Consider the following simple macroeconomic model of a closed economy, where the subscript t denotes time:

$$C_t = \beta_0 + \beta_1 Y_t + u_t, \qquad u_t \text{ i.i.d. with mean zero,} \tag{1.8a}$$

$$Y_t = C_t + I_t. \tag{1.8b}$$

The first equation is behavioral, to the effect that income Y_t determines consumption C_t. The second is definitional and says that consumption determines income, along with investment I_t, which is treated as exogenously determined outside the model. The evident circularity clearly renders void any assumption that Y_t is independent of the disturbance term u_t. Indeed, on solving for Y_t in terms of I_t and u_t we have

$$Y_t = \frac{\beta_0}{1-\beta_1} + \frac{I_t}{1-\beta_1} + \frac{u_t}{1-\beta_1},$$

from which it follows that

$$\operatorname{plim} \frac{1}{n} \sum Y_t u_t = \frac{1}{1-\beta_1} \operatorname{plim} \frac{1}{n} \sum u_t^2 = \frac{\sigma_u^2}{1-\beta_1} \neq 0.$$

In general, we can expect any model in which the right-hand variables are simultaneously determined along with the dependent or left-hand variables to create problems of regressor–disturbance correlation. Such circularity often arises from equilibrium considerations, in which a set of structural equations simultaneously determines equilibrium values. Thus a system comprising a demand equation and a supply equation for a certain good will have the equilibrium price and quantity transacted determined jointly by these equations. Or the population of lynxes may be determined by that of snowshoe hares, and the population of hares by that of lynxes, in a Volterra-type predator–prey nexus. Simultaneity bias may continue to apply even if the equilibrium values are not directly observable but are treated as moving targets, where the adjustment to equilibrium is not complete in every period. We note in passing, however, that there is a school of thought that maintains that structural relationships should always be specified in continuous rather than discrete

time and that simultaneity may be an artifice that results from the necessity of measuring flows over some finite period of time (Bergstrom 1966). Whatever one's views on this, it is unavoidable as a matter of practice that in any system designed to determine an equilibrium or an adjustment to equilibrium, problems of regressor–disturbance correlation will arise.

The simultaneity effect may owe its genesis to individual decision problems as well as consideration of dynamic or market equilibrium of a more macro kind. For example, earlier work on the supply of labor assumed that the individual's supply of hours depended upon, in addition to nonlabor income and specific worker characteristics, the wage rate (where the latter could be treated as statistically exogenous). More recently, however, several authors (e.g., Rosen 1976, Hausman and Wise 1976, Burtless and Hausman 1978) have considered the wage rate to be endogenous. Because of the effects of progressive income tax rates, income-tested social security, and other welfare payments, the wage rate may in fact depend upon the supply of hours. Thus to simply regress hours worked upon the wage rates, together with other relevant individual characteristics, will result in problems of consistency, if not of specification and identification.

The existence of simultaneity is in fact pervasive in empirical work in economics. In some instances, one may be able to specify a complete model that accounts for all aspects of the simultaneity. Thus the multiequation macroeconometric models typically aim to describe the joint probability distribution of every variable considered as endogenous. More often, however, one will be able to recognize that a given variable included among the right-hand regressors is in turn influenced by the left-hand or dependent variable, yet have no very precise model for the joint generation of the two variables concerned. In the parlance of stochastic simultaneous-equation theory, the latter is a "limited-information" situation. Since they are relatively robust with respect to specification error, IV methods are particularly suited to such incomplete-information contexts. Their application to both limited- and full-information models is considered in detail in Chapter 4, which deals with linear models, and Chapter 5, which concerns nonlinear models.

Time series problems

Consider the following distributed-lag model, which distributes or smoothes over time the effect of one variable (w_t) on a dependent variable:

$$y_t = \beta_0 + \beta_1(1-\lambda)(w_t + \lambda w_{t-1} + \lambda^2 w_{t-2} + \cdots) + u_t, \tag{1.9}$$

where $0 \leq \lambda < 1$ and the disturbances u_t have mean zero and are i.i.d. ("white noise," in the terminology), with variance σ^2. Thus investment plans may be determined in terms of an exponentially smoothed series of sales figures w_t. Equation (1.9) may be transformed by appropriate lag operations into

$$y_t = \lambda y_{t-1} + (1-\lambda)\beta_0 + (1-\lambda)\beta_1 w_t + v_t, \tag{1.10}$$

where in this autoregressive version the new disturbance is the moving average process $v_t = u_t - \lambda u_{t-1}$. As a regression formulation, equation (1.10) is plainly more convenient than the original. However, the effect of the transformation is that the regressor y_{t-1} and the new error v_t are correlated. Indeed,

$$\operatorname{plim} \frac{1}{n} \sum y_{t-1} v_t = \operatorname{plim} \frac{1}{n} \sum y_{t-1}(u_t - \lambda u_{t-1}) = -\lambda\sigma^2 \neq 0,$$

so that regressor and error in equation (1.10) are negatively correlated.

The above model provides an example of the transformation of a structural equation with a white-noise disturbance term into an estimating form in which the disturbance is no longer white noise. There are many examples of transformations of this type in the time series literature. As a general observation, it is satisfying to think that one is fitting an underlying structural model in which the disturbance term is purely random white noise, if only because this indicates that all sources of systematic variation, or information, are incorporated among the regressors. In many instances, however, it may be unduly restrictive to specify white-noise residuals, even if the researcher has no very precise theory as to why they should be serially correlated. Thus one should fit the structural model allowing for the presence of serially correlated residuals as a procedure in which the case of zero serial correlation can be appropriately nested. Given the widespread presence of lagged dependent variables arising from expectational or partial adjustment effects, this means that the estimation procedure should explicitly allow for the possibility of regressor–error correlation. The application of instrumental variable techniques to such models is considered in detail in Sections 3.3 and 3.4.

1.2 The instrumental variables estimator: a first approach

The models outlined in the previous sections may all be subsumed under the following forms:

(a) In the case of linear structures, a linear regression framework:

$$\mathbf{y} = X\boldsymbol{\beta} + \mathbf{u}. \tag{1.11}$$