

Comments on: Panel data analysis—advantages and challenges

Marc Nerlove

Published online: 13 March 2007

© Sociedad de Estadística e Investigación Operativa 2007

In most applications of statistical analysis in the sciences, the process by which the observed data are generated is transparent having usually been determined by the investigator by design. In contrast, in many applications in the social sciences, especially in economics, the mechanism by which the data are generated is opaque. In such circumstances, estimation of the parameters of the statistical model of the process and testing specific hypotheses about it are only half the problem of inference. My own view is that understanding the process by which the observations at hand are generated is of equal importance. Were the data, for example, obtained from a sample of firms selected by stratified random sampling from a census of all firms in the United States in 2000? Were they obtained from regulatory activity? In the case of time series, the data are almost always “fabricated,” in one way or another, by aggregation, interpolation, or extrapolation, or by all three. The nature of the sampling frame or the way in which the data are fabricated must be part of the model specification on which parametric inference or hypothesis testing is based.

In his exemplary survey of panel data analysis, Cheng Hsiao focuses primarily on problems of estimation and inference from a parametrically well-specified model of how the observed data were generated. In my commentary, I would like briefly to address some of the issues associated with the other half of the problem. Since such a discussion is data specific, it is possible only to deal with the issues in the context of a specific, although possibly abstract, example. Suppose a longitudinal household survey in which the same households are questioned over time about their actions in, say, a number of consecutive months or years and, initially, about various

This comment refers to the invited paper available at:
<http://dx.doi.org/10.1007/s11749-007-0046-x>.

M. Nerlove (✉)

Department of Agricultural and Resource Economics, University of Maryland, Maryland, USA
e-mail: mnerlove@arec.umd.edu

demographic and economic characteristics. These households differ in various ways some of which we observe and many which we do not. Some of these differences are the result of their past behavior or past circumstances (path dependence), some are differences in tastes or other unobserved characteristics which may be assumed to be permanent (individual heterogeneity), and some are due to peculiarities not permanently associated with time or individual.¹

Let me turn to the questions of what, in the context of these data, can be considered as random, what is the population from which we may consider the data a sample, and what is a parameter, and what a random variable.

Statistical and, *a fortiori*, econometric analysis, are usually based on the idea of sampling from a *population* in order to draw inferences from the underlying population. But what is the population from which economic data may be supposed to be a sample? In his famous 1944 monograph, *The Probability Approach in Econometrics*, in which Haavelmo laid the foundation for modern econometrics, Haavelmo (1944, p. 56) wrote “. . .the class of populations we are dealing with does not consist of an infinity of different individuals, it consists of an infinity of possible decisions which might be taken. . . .” In their recent text, *Econometric Theory and Methods*, Davidson and Mackinnon (2004, pp. 30–31) make the same point: “In econometrics, the use of the term population is simply a metaphor. A better concept is that of a *data-generating process*, or DGP. By this term, we mean whatever mechanism is at work in the real world of economic activity giving rise to the numbers in our samples, that is, precisely the mechanism that our econometric model is supposed to describe. A data-generating process is thus an analog of a population in biostatistics. Samples may be drawn from a DGP just as they may be drawn from a population. In both cases, the samples are assumed to be representative of DGP or population from which they are drawn.”

What is a random variable in this context and what is not? Whether or not a particular variable can be considered a random draw from some population or not, in principle, can be decided by applying the principle of “exchangeability” introduced by de Finetti (1930).² In a nutshell, the idea, very Bayesian in flavor, is to ask whether we can exchange two elements in a sample and still maintain the same subjective distribution. Thus, in a panel study of households, are any two households in the sample exchangeable without affecting the distribution, from which we imagine household observables and unobservables to be drawn. In a panel of state data, are California and Maryland exchangeable without affecting the subjective distribution of the state effects? It is a dicey question (sometimes).

From the standpoint of a Bayesian I suppose there is no real distinction between a parameter and a random variable, but in this context I would say that a parameter is an unobserved variable which affects the distribution of the random variables of the model and is unaffected by the particular values such variables take on. It is what we wish to estimate and about which we wish to make inferences. A related concept

¹In his wonderfully titled paper, “Identifying the Hand of the Past: Distinguishing State Dependence from Heterogeneity,” Heckman (1991) argues that in general it is not possible to distinguish. The ability to do so rests critically “on maintaining explicit assumptions about the way in which observables and unobservables interact.” I return to this particular issue at the end of this commentary.

²See also de Finetti (1970, trans. 1990, vol. 2, pp. 211–224) and Lindley and Novick (1981).

is that of an exogenous variable, to which I return below. But note here that such an exogenous variable is still a random variable and not a parameter.

In general, in the formulation of econometric models, i.e., the DGP for the process yielding the particular set of data we want to “explain,” the distinction between what can be observed and what is not is fundamental. Linear functions are often used to describe such a DGP. To get more precisely to the issues posed by the formulation of the DGP for a sample of economic data, we need to include several observable variables. Suppose that we draw a random sample of N individuals over T time periods; for example, a household survey in which we collect observations on the income, x_{it} , and consumption of household i , y_{it} , for many households N , in year t over a brief period T . From the survey we have observations on the pairs $\{x_{it}, y_{it}\}$. Since the households are chosen at random for the survey, but the years over which they are observed are not, the lists $\{x_{i1}, y_{i1}, \dots, x_{iT}, y_{iT}\}$ are exchangeable, but the order within each list is not.

Imagine we are estimating a consumption function and assume a linear relationship subject to error:

$$y_{it} = a + bx_{it} + \varepsilon_{it}. \quad (1)$$

This would be the case if, for example, the joint distribution of variables could be assumed normal and we were trying to estimate the mean of y_{it} for a particular year t conditional on x_{it} . We might then write ε_{it} as

$$\varepsilon_{it} = \mu_i + \lambda_t + u_{it}, \quad (2)$$

where ε_{it} is an unobserved random variable which is the sum of three effects, all of which are also unobserved: λ_t is a year effect, arguably nonrandom and therefore a parameter to be estimated for each year, t ;³ μ_i is a household effect, which, in view of the way the observations are drawn, should surely be treated as random, and, finally, u_{it} is a random variable to represent all the rest.

We are far from done yet, however. The question remains as to what we should assume about the observable variables, x_{it} . They are clearly random variables jointly distributed with the variable y_{it} . If not subject to errors of measurement, an assumption difficult to justify in the context of an economic survey, are they also independent of, or at least uncorrelated with, the disturbances ε_{it} in (1)? This question clearly affects not only what we can say about the DGP which generates our observations, but also how many and what parameters must be considered. Let us examine the regression with some care. Since λ_t is not a random variable but a parameter, consider it to be a constant for each t and add it to the constant a in the regression equation (1):

$$y_{it} = a_t^* + bx_{it} + v_{it}, \quad (3)$$

where $a_t^* = a + \lambda_t$ and $v_{it} = \mu_i + u_{it}$.

³Of course, in a sequence of years, one might expect the λ 's for adjacent years to be more alike than those for distant years, hence to follow some form of functional dependence, which would enforce this expectation. It would be natural to try to approximate the resulting behavior by an autoregressive relation among years or a spline.

Suppose that, given t , v_{it} is distributed with mean zero and variance-covariance matrix Σ_t . Suppose further that Σ_t does not depend on t . If x_{it} is *strictly exogenous* in the regression (3), which means

$$E[v_{it} | x_{it}] = 0, \quad \text{all } i \text{ and } t, \tag{4}$$

then (3) is the usual panel model. This means that b can be estimated by GLS or ML with a dummy variable for each t . *Weak exogeneity* is a related concept introduced by Engle et al. (1983). In the context of regression (3), we say that x_{it} is weakly exogenous if v_{it} is distributed independently of $\{x_{is}, y_{is}, \text{ all } i \text{ and } s \leq t - 1\}$, if the marginal distribution of $\{x_{is}, y_{is}, \text{ all } i \text{ and } s \leq t - 1\}$ does not depend on any unknown parameters in Σ or on b or the λ s, and the pdf of $x_{it} | \{x_{is}, y_{is}, \text{ all } i \text{ and } s \leq t\}$ and $x_{it} | \{x_{is}, y_{is}, \text{ all } i \text{ and } s \leq t - 1\}$ does not depend on any unknown parameters in Σ or on b or the λ s. If regression (3) satisfies the conditions of weak exogeneity, the likelihood function for the whole sample of observations on x and y factors into two pieces, one of which is the usual regression likelihood and the other is a function of x but not of the parameters in Σ or on b or the λ s. In that sense we can treat the observations on x as fixed.

But is the exogeneity, weak or strict, a reasonable assumption? Here is what Wooldridge (2002, p. 252) says:

Traditional unobserved components panel models take the x_{it} as fixed. We will never assume the x_{it} are nonrandom because potential feedback from y_{it} to x_{is} for $s > t$ needs to be addressed explicitly.

The assumption that the explanatory variables in the regression are exogenous is generally impossible. If the vector of explanatory variables includes any lagged values of y_{it} , either explicitly or implicitly, the strict or weak exogeneity is generally impossible. Any meaningful DPG describing individual economic behavior is intrinsically dynamic in the sense that the “hand of the past,” whether as a result of path dependence or of individual heterogeneity, is ever present. To put the point more explicitly if, among the observed variables, there are any initial conditions related to past values of the observed y_{it} ’s or to unobservables affecting present and past behavior, at least one of the components of x_{it} must be correlated with μ_{it} . A Hausman test will reject the exogeneity of the x ’s almost certainly. A rejection of exogeneity does not, of course, imply that the unobserved components $\{\mu_{it}\}$ of the errors in (3) are not random (RE) but fixed (FE). Unfortunately, as Hsiao points out, this leaves the econometrician between Scylla and Charybdis: We are damned if we do and damned if we do not. I quote directly from Hsiao’s paper, making some minor changes in his notation to conform to mine:

The advantages of random effects (RE) specification are: (a) The number of parameters stay constant when sample size increases. (b) It allows the derivation of efficient estimators that make use of both within and between (group) variation. (c) It allows the estimation of the impact of time-invariant variables. The disadvantage is that one has to specify the conditional density of μ_i given $x_i = (x_{i1}, \dots, x_{iT})$, $f(\mu_i | x_i)$, while μ_i are unobservable. A common assumption is that $f(\mu_i | x_i)$ is identical to the marginal density $f(\mu_i)$. However, if the effects are correlated with x_{it} or if there is a fundamental difference among

individual units, i.e., conditional on x_{it} , y_{it} cannot be viewed as a random draw from a common distribution, common RE model is misspecified and the resulting estimator is biased.

The advantages of fixed effects (FE) specification are that it can allow the individual-and/or time specific effects to be correlated with explanatory variables x_{it} . Neither does it require an investigator to model their correlation patterns. The disadvantages of the FE specification are: (a') The number of unknown parameters increases with the number of sample observations. In the case where T (or N for λ_t) is finite, it introduces the classical incidental parameter problem (e.g. Neyman and Scott 1948). (b') The FE estimator does not allow the estimation of the coefficients that variables are time-invariant.

In other words, the advantages of RE specification are the disadvantages of FE specification, and the disadvantages of RE specification are the advantages of FE specification.

So what is one to do? As Heckman, quoted above, says, one must be willing to make “explicit assumptions about the way in which observables and unobservables interact.” But most econometricians are not willing to specify such interactions as part of the DGP. Hence, the random effects are treated as parameters rather than random variables. They are viewed as incidental parameters, and the object is to get rid of them without distorting the estimates of the structural parameters β . There is no universally accepted way of doing so in all contexts, especially not in explicitly dynamic or nonlinear contexts, and, in my view, no right way of doing so. Hsiao gives the best survey of the many approaches which have been tried econometrically I have seen until now.

References

- Davidson R, Mackinnon JG (2004) *Econometric theory and methods*. Oxford University Press, New York
- de Finetti B (1930) Problemi determinati e indeterminati nel calcolo delle probabilità. *Rend R Accad Naz Lincei Ser 6* 12(9)
- de Finetti B (1970) *Teoria delle Probabilità: sintesi introduttiva con appendice critica*. Giulio Einaudi Editorial, Torino. Translated as *Theory of Probability*, Chichester, 1990
- Engle RF, Hendry DF, Richard J-F (1983) Exogeneity. *Econometrica* 51:277–304
- Haavelmo T (1944) The probability approach in econometrics. *Econometrica* 12 (supplement)
- Heckman JJ (1991) Identifying the hand of the past: distinguishing state dependence from heterogeneity. *Am Econ Rev* 81(2):75–79
- Lindley DV, Novick MR (1981) The role exchangeability in inference. *Ann Stat* 9:45–58
- Neyman J, Scott EL (1948) Consistent estimates based on partially consistent observations. *Econometrica* 16:1–32
- Wooldridge JM (2002) *Econometric analysis of cross section and panel data*. MIT Press, Cambridge