Sociedad de Estadística e Investigación Operativa
Test (0000) *Vol.* **00**, *No.* 0, pp. 1–63

Panel Data Analysis — Advantages and Challenges

Cheng Hsiao*
Department of Economics, University of Southern California, USA
Wang Yanan Institute for Studies in Economics, Xiamen University, China

Abstract

We explain the proliferation of panel data studies in terms of (i) data availability, (ii) the more heightened capacity for modeling the complexity of human behavior than a single cross-section or time series data can possibly allow, and (iii) challenging methodology. Advantages and issues of panel data modeling are also discussed.

Key Words: Panel data, longitudinal data, unobserved heterogeneity, random effects, fixed effects.

AMS subject classification:62-02

1 Introduction

Panel data or longitudinal data typically refer to data containing time series observations of a number of individuals. Therefore, observations in panel data involve at least two dimensions; a cross-sectional dimension, indicated by subscript i, and a time series dimension, indicated by subscript t. However, panel data could have a more complicated clustering or hierarchical structure. For instance, variable y may be the measurement of the level of air pollution at station ℓ in city j of country i at time t (e.g. Antweiler, 2001; Davis, 2002). For ease of exposition, I shall confine my presentation to a balanced panel involving N cross-sectional units, $i = 1, \ldots, N$, over T time periods, $t = 1, \ldots, T$.

There is a proliferation of panel data studies, be it methodological or empirical. In 1986, when Hsiao's (1986) first edition of *Panel Data Analysis* was published, there were 29 studies listing the key words: "panel data or

*Correspondence to: Cheng Hsiao. Department of Economics, University of Southern California, Los Angeles, CA 90089-0253, USA. E-mail: chsiao@usc.edu

longitudinal data", according to Social Sciences Citation index. By 2004, there were 687 and by 2005, there were 773. The growth of applied studies and the methodological development of new econometric tools of panel data have been simply phenomenal since the seminal paper of Balestra and Nerlove (1966).

There are at least three factors contributing to the geometric growth of panel data studies. (i) data availability, (ii) greater capacity for modeling the complexity of human behavior than a single cross-section or time series data, and (iii) challenging methodology. In what follows, we shall briefly elaborate each of these one by one. However, it is impossible to do justice to the vast literature on panel data. For further reference, see Arellano (2003), Baltagi (2001), Hsiao (2003), Mátyás and Sevestre (1996), and Nerlove (2002), etc.

2 Data availability

The collection of panel data is obviously much more costly than the collection of cross-sectional or time series data. However, panel data have become widely available in both developed and developing countries.

The two most prominent panel data sets in the US are the National Longitudinal Surveys of Labor Market Experience (NLS) and the University of Michigan's Panel Study of Income Dynamics (PSID). The NLS began in the mid 1960's. It contains five separate annual surveys covering distinct segments of the labor force with different spans: men whose ages were 45 to 59 in 1966, young men 14 to 24 in 1966, women 30 to 44 in 1967, young women 14 to 24 in 1968, and youth of both sexes 14 to 21 in 1979. In 1986, the NLS expanded to include annual surveys of the children born to women who participated in the National Longitudinal Survey of Youth 1979. The list of variables surveyed is running into the thousands, with emphasis on the supply side of market.

The PSID began with collection of annual economic information from a representative national sample of about 6,000 families and 15,000 individuals in 1968 and has continued to the present. The data set contains over 5,000 variables (Becketti et al., 1988). In addition to the NLS and PSID data sets, there are many other panel data sets that could be of interest to economists, see Juster (2000).

In Europe, many countries have their annual national or more frequent surveys such as the Netherlands Socio-Economic Panel (SEP), the German Social Economics Panel (GSOEP), the Luxembourg Social Panel (PSELL), the British Household Panel Survey (BHS), etc. Starting in 1994, the National Data Collection Units (NDUS) of the Statistical Office of the European Committees have been coordinating and linking existing national panels with centrally designed multi-purpose annual longitudinal surveys. The European Community Household Panel (ECHP) are published in Eurostat's reference data base New Cronos in three domains: health, housing, and income and living conditions.

Panel data have also become increasingly available in developing countries. In these countries, there may not have been a long tradition of statistical collection. It is of special importance to obtain original survey data to answer many significant and important questions. Many international agencies have sponsored and helped to design panel surveys. For instance, the Dutch non-government organization (NGO), ICS, Africa, collaborated with the Kenya Ministry of Health to carry out a Primary School Deworming Project (PDSP). The project took place in Busia district, a poor and densely-settled farming region in western Kenya. The 75 project schools include nearly all rural primary schools in this area, with over 30,000 enrolled pupils between the ages of six to eighteen from 1998-2001. Another example is the Development Research Institute of the Research Center for Rural Development of the State Council of China, in collaboration with the World Bank, which undertook an annual survey of 200 large Chinese township and village enterprises from 1984 to 1990.

3 Advantages of panel data

Panel data, by blending the inter-individual differences and intra-individual dynamics have several advantages over cross-sectional or time-series data:

(i) More accurate inference of model parameters. Panel data usually contain more degrees of freedom and more sample variability than cross-sectional data which may be viewed as a panel with $T = 1$, or time series data which is a panel with $N = 1$, hence improving the efficiency of econometric estimates (e.g. Hsiao et al., 1995).

(ii) Greater capacity for capturing the complexity of human behavior than a single cross-section or time series data. These include:

(ii.a) Constructing and testing more complicated behavioral hypotheses. For instance, consider the example of Ben-Porath (1973) that a cross-sectional sample of married women was found to have an average yearly labor-force participation rate of 50 percent. These could be the outcome of random draws from a homogeneous population or could be draws from heterogeneous populations in which 50% were from the population who always work and 50% never work. If the sample was from the former, each woman would be expected to spend half of her married life in the labor force and half out of the labor force. The job turnover rate would be expected to be frequent and the average job duration would be about two years. If the sample was from the latter, there is no turnover. The current information about a woman's work status is a perfect predictor of her future work status. A cross-sectional data is not able to distinguish between these two possibilities, but panel data can because the sequential observations for a number of women contain information about their labor participation in different subintervals of their life cycle.

Another example is the evaluation of the effectiveness of social programs (e.g. Heckman et al., 1998; Hsiao et al., 2006; Rosenbaum and Rubin, 1985). Evaluating the effectiveness of certain programs using cross-sectional sample typically suffers from the fact that those receiving treatment are different from those without. In other words, one does not simultaneously observe what happens to an individual when she receives the treatment or when she does not. An individual is observed as either receiving treatment or not receiving treatment. Using the difference between the treatment group and control group could suffer from two sources of biases, selection bias due to differences in observable factors between the treatment and control groups and selection bias due to endogeneity of participation in treatment. For instance, Northern Territory (NT) in Australia decriminalized possession of small amount of marijuana in 1996. Evaluating the effects of decriminalization on marijuana smoking behavior by comparing the differences between NT and other states that were still non-decriminalized could suffer from either or both sorts of bias. If panel data over this time period are available, it would allow the possibility of observing the before- and affect-effects on individuals of decriminalization as well as providing the possibility of isolating the effects of treatment from other factors affecting the outcome.

(ii.b) Controlling the impact of omitted variables. It is frequently argued that the real reason one finds (or does not find) certain effects is due to ignoring the effects of certain variables in one's model specification which are correlated with the included explanatory variables. Panel data contain information on both the intertemporal dynamics and the individuality of the entities may allow one to control the effects of missing or unobserved variables. For instance, MaCurdy's (1981) life-cycle labor supply model under certainty implies that because the logarithm of a worker's hours worked is a linear function of the logarithm of her wage rate and the logarithm of worker's marginal utility of initial wealth, leaving out the logarithm of the worker's marginal utility of initial wealth from the regression of hours worked on wage rate because it is unobserved can lead to seriously biased inference on the wage elasticity on hours worked since initial wealth is likely to be correlated with wage rate. However, since a worker's marginal utility of initial wealth stays constant over time, if time series observations of an individual are available, one can take the difference of a worker's labor supply equation over time to eliminate the effect of marginal utility of initial wealth on hours worked. The rate of change of an individual's hours worked now depends only on the rate of change of her wage rate. It no longer depends on her marginal utility of initial wealth.

(ii.c) Uncovering dynamic relationships.

"Economic behavior is inherently dynamic so that most econometrically interesting relationship are explicitly or implicitly dynamic". (Nerlove, 2002). However, the estimation of time-adjustment pattern using time series data often has to rely on arbitrary prior restrictions such as Koyck or Almon distributed lag models because time series observations of current and lagged variables are likely to be highly collinear (e.g. Griliches, 1967).

With panel data, we can rely on the inter-individual differences to reduce the collinearity between current and lag variables to estimate unrestricted time-adjustment patterns (e.g. Pakes and Griliches, 1984).

(ii.d) Generating more accurate predictions for individual outcomes by pooling the data rather than generating predictions of individual outcomes using the data on the individual in question. If individual behaviors are similar conditional on certain variables, panel data provide the possibility of learning an individual's behavior by observing the behavior of others. Thus, it is possible to obtain a more accurate description of an individual's behavior by supplementing observations of the individual in question with data on other individuals (e.g. Hsiao et al., 1993, 1989).

(ii.e) Providing micro foundations for aggregate data analysis.

Aggregate data analysis often invokes the "representative agent" assumption. However, if micro units are heterogeneous, not only can the time series properties of aggregate data be very different from those of disaggregate data (e.g. Granger, 1990; Lewbel, 1994; Pesaran, 2003), but policy evaluation based on aggregate data may be grossly misleading. Furthermore, the prediction of aggregate outcomes using aggregate data can be less accurate than the prediction based on micro-equations (e.g. Hsiao et al., 2005). Panel data containing time series observations for a number of individuals is ideal for investigating the "homogeneity" versus "heterogeneity" issue.

(iii) Simplifying computation and statistical inference.

Panel data involve at least two dimensions, a cross-sectional dimension and a time series dimension. Under normal circumstances one would expect that the computation of panel data estimator or inference would be more complicated than cross-sectional or time series data. However, in certain cases, the availability of panel data actually simplifies computation and inference. For instance:

(iii.a) Analysis of nonstationary time series.

When time series data are not stationary, the large sample approximation of the distributions of the least-squares or maximum likelihood estimators are no longer normally distributed, (e.g. Anderson, 1959; Dickey and Fuller, 1979, 1981; Phillips and Durlauf, 1986). But if panel data are available, and observations among cross-sectional units are independent, then one can invoke the central limit theorem across cross-sectional units to show that the limiting distributions of many estimators remain asymptotically normal (e.g. Binder et al., 2005; Im et al., 2003; Levin et al., 2002; Phillips and Moon, 1999).

(iii.b) Measurement errors.

Measurement errors can lead to under-identification of an econometric model (e.g. Aigner et al., 1984). The availability of multiple observations for a given individual or at a given time may allow a researcher to make different transformations to induce different and deducible changes in the estimators, hence to identify an otherwise unidentified model (e.g. Biørn, 1992; Griliches and Hausman, 1986; Wansbeek and Koning, 1989).

(iii.c) Dynamic Tobit models. When a variable is truncated or censored, the actual realized value is unobserved. If an outcome variable depends on previous realized value and the previous realized value are unobserved, one has to take integration over the truncated range to obtain the likelihood of observables. In a dynamic framework with multiple missing values, the multiple integration is computationally unfeasible. With panel data, the problem can be simplified by only focusing on the subsample in which previous realized values are observed (e.g. Arellano et al., 1999).

4 Methodology

Standard statistical methodology is based on the assumption that the outcomes, say $\underset{\sim}{y}$, conditional on certain variables, say $\underset{\sim}{x}$, are random outcomes from a probability distribution that is characterized by a fixed dimensional parameter vector, $\underset{\sim}{\theta}$, $f(\underset{\sim}{y} \mid \underset{\sim}{x}; \underset{\sim}{\theta})$. For instance, the standard linear regression model assumes that $f(\underset{\sim}{y} \mid \underset{\sim}{x}; \underset{\sim}{\theta})$ takes the form that

$$E(y \mid \underset{\sim}{x}) = \alpha + \underset{\sim}{\beta}' \underset{\sim}{x}, \tag{4.1}$$

and

$$\mathrm{Var}(y \mid \underset{\sim}{x}) = \sigma^2, \tag{4.2}$$

where $\underset{\sim}{\theta}' = (\alpha, \underset{\sim}{\beta}', \sigma^2)$. Typical panel data focuses on individual outcomes. Factors affecting individual outcomes are numerous. It is rare to be able to assume a common conditional probability density function of y conditional on $\underset{\sim}{x}$ for all cross-sectional units, i, at all time, t. For instance, suppose that in addition to $\underset{\sim}{x}$, individual outcomes are also affected by unobserved individual abilities (or marginal utility of initial wealth as in MaCurdy (1981) labor supply model discussed in (iib) on Section 3), represented by α_i, so that the observed $(y_{it}, \underset{\sim}{x}_{it}), i = 1, \ldots, N, t = 1, \ldots, T$, are actually generated by

$$y_{it} = \alpha_i + \underset{\sim}{\beta}'\underset{\sim}{x}_{it} + u_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{matrix} \tag{4.3}$$

as depicted by Figure 1 in which the broken-line ellipses represent the point scatter of individual observations around their respective mean, represented by the broken straight lines. If an investigator mistakenly imposes the homogeneity assumption (4.1) - (4.2), the solid lines in those figures would represent the estimated relationships between y and $\underset{\sim}{x}$, which can be grossly misleading.

If the conditional density of y given $\underset{\sim}{x}$ varies across i and over t, the fundamental theorems for statistical inference, the laws of large numbers and central limit theorems, will be difficult to implement. One way to restore homogeneity across i and/or over t is to add more conditional variables, say $\underset{\sim}{z}$,

$$f(y_{it} \mid \underset{\sim}{x}_{it}, \underset{\sim}{z}_{it}; \underset{\sim}{\theta}). \tag{4.4}$$

However, the dimension of $\underset{\sim}{z}$ can be large. A model is a simplification of reality, not a mimic of reality. The inclusion of $\underset{\sim}{z}$ may confuse the fundamental relationship between y and $\underset{\sim}{x}$, in particular, when there is a shortage of degrees of freedom or multicollinearity, etc. Moreover, $\underset{\sim}{z}$ may not be observable. If an investigator is only interested in the relationship between y and $\underset{\sim}{x}$, one approach to characterize the heterogeneity not captured by $\underset{\sim}{x}$ is to assume that the parameter vector varies across i and over t, $\underset{\sim}{\theta}_{it}$, so that the conditional density of y given $\underset{\sim}{x}$ takes the form $f(y_{it} \mid \underset{\sim}{x}_{it}; \underset{\sim}{\theta}_{it})$. However, without a structure being imposed on $\underset{\sim}{\theta}_{it}$, such a model only has descriptive value. It is not possible to draw any inference about $\underset{\sim}{\theta}_{it}$.

The methodological literature on panel data is to suggest possible structures on $\underset{\sim}{\theta}_{it}$ (e.g. Hsiao, 2003). One way to impose some structure on $\underset{\sim}{\theta}_{it}$ is to decompose $\underset{\sim}{\theta}_{it}$ into $(\underset{\sim}{\beta}, \underset{\sim}{\gamma}_{it})$, where $\underset{\sim}{\beta}$ is the same across i and over t, referred to as *structural parameters*, and $\underset{\sim}{\gamma}_{it}$ as *incidental parameters* because

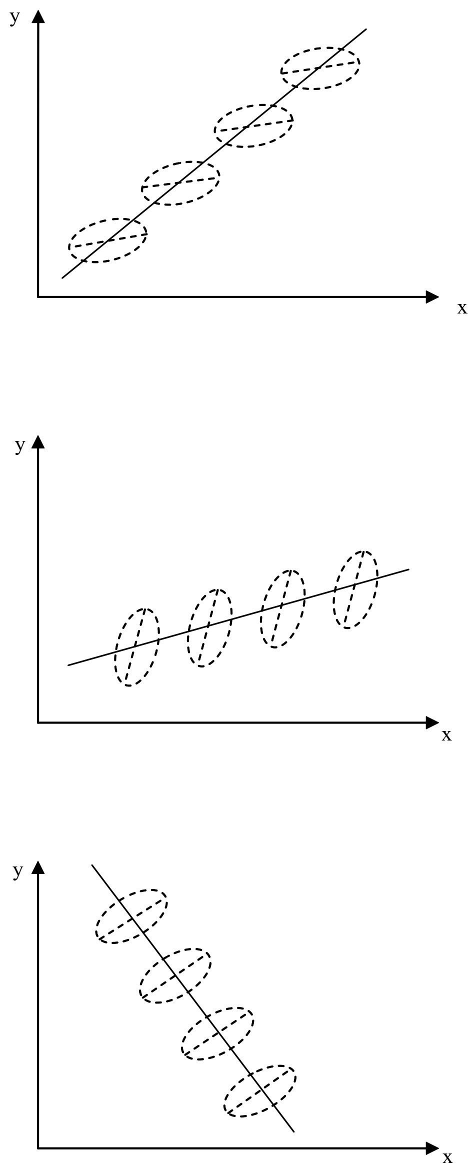

Figure 1: *Scatter Diagrams of* $(y(i,t),\, x(i,t))$

when cross-section units, N and/or time series observations, T increases, so does the dimension of $\underset{\sim}{\gamma}_{it}$. The focus of panel data literature is to make inference on $\underset{\sim}{\beta}$ after controlling the impact of $\underset{\sim}{\gamma}_{it}$.

Without imposing a structure for $\underset{\sim}{\gamma}_{it}$, again it is difficult to make any inference on $\underset{\sim}{\beta}$ because estimation of $\underset{\sim}{\beta}$ could depend on γ_{it} and the estimation of the unknown $\underset{\sim}{\gamma}_{it}$ probably will exhaust all available sample information. Assuming that the impacts of observable variables, $\underset{\sim}{x}$, are the same across i and over t, represented by the structure parameters, $\underset{\sim}{\beta}$, the incidental parameters $\underset{\sim}{\gamma}_{it}$ represent the heterogeneity across i and over t that are not captured by $\underset{\sim}{x}_{it}$. They can be considered composed of the effects of omitted individual time-invariant, α_i, period individual-invariant, λ_t, and individual time-varying variables, δ_{it}. The individual time-invariant variables are variables that are the same for a given cross-sectional unit through time but vary across cross-sectional units such as individual-firm management, ability, gender, and socio-economic background variables. The period individual-invariant variables are variables that are the same for all cross-sectional units at a given time but vary through time such as prices, interest rates, and wide spread optimism or pessimism. The individual time-varying variables are variables that vary across cross-sectional units at a given point in time and also exhibit variations through time such as firm profits, sales and capital stock. The effects of unobserved heterogeneity can either be assumed as random variables, referred to as the *random effects* model, or fixed parameters, referred to as the *fixed effects* model, or a mixture of both, refereed to as the *mixed effects* model.

The challenge of panel methodology is to control the impact of unobserved heterogeneity, represented by the incidental parameters, γ_{it}, to obtain valid inference on the structural parameters $\underset{\sim}{\beta}$. A general principle of obtaining valid inference of $\underset{\sim}{\beta}$ in the presence of incidental parameters $\underset{\sim}{\gamma}_{it}$ is to find proper transformation to eliminate $\underset{\sim}{\gamma}_{it}$ from the specification or to integrate out the effects of $\underset{\sim}{\gamma}_{it}$. Since proper transformations depend on the model one is interested, as illustrations, I shall try to demonstrate the fundamental issues from the perspective of linear static models, dynamic models, nonlinear models, models with cross-sectional dependencies and models with large N and large T.

For ease of exposition, I shall assume for the most time that there are no time-specific effects, λ_t and the individual time-varying effects, δ_{it}, can be represented by a random variable u_{it}, that is treated as the error of an equation. In other words, only individual-specific effects, α_i, are present. The individual-specific effects, α_i, can either be assume as random or fixed. The standard assumption for random effects specification is that they are randomly distributed with a common mean and are independent of fixed $\underset{\sim}{x}_{it}$.

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 a conditional density of α_i given $\underset{\sim}{x}_i' = (\underset{\sim}{x}_{it}, \ldots, \underset{\sim}{x}_{iT}), f(\alpha_i \mid \underset{\sim}{x}_i)$, while α_i are unobservable. A common assumption is that $f(\alpha_i \mid \underset{\sim}{x}_i)$ is identical to the marginal density $f(\alpha_i)$. However, if the effects are correlated with $\underset{\sim}{x}_{it}$ or if there is a fundamental difference among individual units, i.e., conditional on $\underset{\sim}{x}_{it}, y_{it}$ 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 $\underset{\sim}{x}_{it}$. 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 when 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 are time-invariant.

In order words, the advantages of RE specification are the disadvantages of FE specification and the disadvantages of RE specification are the advantages of FE specification. To choose between the two specifications, Hausman (1978) notes that if the FE estimator (or GMM), $\underset{\sim}{\theta}_{FE}$, is consistent whether α_i is fixed or random and the commonly used RE estimator (or GLS), $\underset{\sim}{\theta}_{RE}$, is consistent and efficient only when α_i is indeed uncorrelated with $\underset{\sim}{x}_{it}$ and is inconsistent if α_i is correlated with $\underset{\sim}{x}_{it}$. Therefore, he suggests using the statistic

$$(\underset{\sim}{\theta}_{FE} - \underset{\sim}{\theta}_{RE})' \left[\text{Cov}\,(\underset{\sim}{\theta}_{FE}) - \text{Cov}\,(\underset{\sim}{\theta}_{RE})\right]^{-} (\underset{\sim}{\theta}_{FE} - \underset{\sim}{\theta}_{RE}) \tag{4.5}$$

to test RE vs FE specification. The statistic (4.5) is asymptotically chi-square distributed with degrees of freedom equal to the rank of $[\text{Cov}(\underset{\sim}{\theta}_{FE}) - \text{Cov}(\underset{\sim}{\theta}_{RE})]$.

4.1 Linear Static Models

A widely used panel data model is to assume that the effects of observed explanatory variables, $\underset{\sim}{x}$, are identical across cross-sectional units, i, and over time, t, while the effects of omitted variables can be decomposed into the individual-specific effects, α_i, time-specific effects, λ_t, and individual time-varying effects, $\delta_{it} = u_{it}$, as follows:

$$y_{it} = \underset{\sim}{\beta}'\underset{\sim}{x}_{it} + \alpha_i + \lambda_t + u_{it}, \begin{array}{l} i = 1, \ldots, N, \\ t = 1, \ldots, T. \end{array} \tag{4.6}$$

In a single equation framework, individual-time effects, u, are assumed random and uncorrelated with $\underset{\sim}{x}$, while α_i and λ_t may or may not correlated with $\underset{\sim}{x}$. When α_i and λ_t are treated as fixed constants as coefficients of dummy explanatory variables, $d_{it} = 1$ if the observation corresponds to ith individual at time t, and 0 otherwise, whether they are correlated with $\underset{\sim}{x}$ is not an issue. On the other hand, when α_i and λ_t are treated as random, they become part of the error term and are typically assumed to be uncorrelated with $\underset{\sim}{x}_{it}$.

For ease of exposition, we shall assume that there are no time-specific effects, i.e., $\lambda_t = 0$ for all t and u_{it} are independently, identically distributed (i.i.d) across i and over t. Stack an individuals T time series observations of $(y_{it}, \underset{\sim}{x}'_{it})$ into a vector and a matrix, (4.6) may alternatively be written as

$$\underset{\sim}{y}_i = X_i\underset{\sim}{\beta} + \underset{\sim}{e}\alpha_i + \underset{\sim}{u}_i, i = 1, \ldots, N, \tag{4.7}$$

where $\underset{\sim}{y}_i = (y_{i1}, \ldots, y_{iT})', X_i = (\underset{\sim}{x}_{i1}, \ldots, \underset{\sim}{x}_{iT})', \underset{\sim}{u}_i = (u_{i1}, \ldots, u_{iT})'$, and $\underset{\sim}{e}$ is a $T\times 1$ vector of 1's.

Let Q be a $T\times T$ matrix satisfying the condition that $Q\underset{\sim}{e} = \underset{\sim}{0}$. Premultiplying (4.7) by Q yields

$$Q\underset{\sim}{y}_i = QX_i\underset{\sim}{\beta} + Q\underset{\sim}{u}_i, \quad i = 1.\ldots, N. \tag{4.8}$$

Equation (4.8) no longer involves α_i. The issue of whether α_i is correlated with $\underset{\sim}{x}_{it}$ or whether α_i should be treated as fixed or random is no longer relevant for (4.8). Moreover, since X_i is exogenous, $E(QX_i\underset{\sim}{u}'_iQ') = QE(X_i\underset{\sim}{u}'_i)Q' = \underset{\sim}{0}$ and $EQ\underset{\sim}{u}_i\underset{\sim}{u}'_iQ' = \sigma_u^2 QQ'$. An efficient estimator of $\underset{\sim}{\beta}$ is the generalized least squares estimator (GLS),

$$\underset{\sim}{\beta} = \left[\sum_{i=1}^N X'_i(Q'Q)^- X_i\right]^{-1} \left[\sum_{i=1}^N X'_i(Q'Q)^- \underset{\sim}{y}_i\right], \tag{4.9}$$

where $(Q'Q)^{-}$ denotes the Moore-Penrose generalized inverse (e.g. Rao, 1973).

When $Q = I_T - \frac{1}{T}\underset{\sim}{e}\underset{\sim}{e}'$, Q is idempotent. The Moore-Penrose generalized inverse of $(Q'Q)^{-}$ is just $Q = I_T - \frac{1}{T}\underset{\sim}{e}\underset{\sim}{e}'$ itself. Premultiplying (4.8) by Q is equivalent to transforming (4.6) into a model

$$(y_{it} - y_i) = \underset{\sim}{\beta}'(\underset{\sim}{x}_{it} - \underset{\sim}{x}_i) + (u_{it} - u_i), \begin{matrix} i = 1, \dots, N, \\ t = 1, \dots, T, \end{matrix} \tag{4.10}$$

where $y_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$, $\underset{\sim}{x}_i = \frac{1}{T}\sum_{t=1}^{T} \underset{\sim}{x}_{it}$ and $u_i = \frac{1}{T}\sum_{t=1}^{T} u_{it}$. The transformation is called *covariance transformation*. The least squares estimator (LS) (or a generalized least squares estimator (GLS)) of (4.10),

$$\underset{\sim}{\beta}_{cv} = \left[\sum_{i=1}^{N}\sum_{t=1}^{T}(\underset{\sim}{x}_{it} - \underset{\sim}{x}_i)(\underset{\sim}{x}_{it} - \underset{\sim}{x}_i)'\right]^{-1}\left[\sum_{t=1}^{N}\sum_{t=1}^{T}(\underset{\sim}{x}_{it} - \underset{\sim}{x}_i)(y_{it} - y_i)\right], \tag{4.11}$$

is called *covariance* estimator or *within* estimator because the estimation of $\underset{\sim}{\beta}$ only makes use of within (group) variation of y_{it} and $\underset{\sim}{x}_{it}$ only. The covariance estimator of $\underset{\sim}{\beta}$ turns out to be also the least squares estimator of (4.10). It is the best linear unbiased estimator of $\underset{\sim}{\beta}$ if α_i is treated as fixed and u_{it} is i.i.d.

If α_i is random, transforming (4.7) into (4.8) transforms T independent equations (or observations) into $(T-1)$ independent equations, hence the covariance estimator is not as efficient as the efficient generalized least squares estimator if $E\alpha_i\underset{\sim}{x}'_{it} = \underset{\sim}{0}'$. When α_i is independent of $\underset{\sim}{x}_{it}$ and is independently, identically distributed across i with mean $\underset{\sim}{0}$ and variance σ_α^2, the best linear unbiased estimator (BLUE) of $\underset{\sim}{\beta}$ is GLS,

$$\underset{\sim}{\beta} = \left[\sum_{i=1}^{N} X_i'V^{-1}X_i\right]^{-1}\left[\sum_{i=1}^{N} X_i'V^{-1}\underset{\sim}{y}_i\right]. \tag{4.12}$$

where $V = \sigma_u^2 I_T + \sigma_\alpha^2 \underset{\sim}{e}\underset{\sim}{e}'$, $V^{-1} = \frac{1}{\sigma_u^2}\left[I_T - \frac{\sigma_\alpha^2}{\sigma_u^2 + T\sigma_\alpha^2}\underset{\sim}{e}\underset{\sim}{e}'\right]$. Let $\psi = \frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2}$, the GLS is equivalent to first transforming the data by subtracting a fraction $(1 - \psi^{1/2})$ of individual means y_i and $\underset{\sim}{x}_i$ from their corresponding y_{it} and $\underset{\sim}{x}_{it}$, then regressing $[y_{it} - (1 - \psi^{1/2})y_i]$ on $[\underset{\sim}{x}_{it} - (1 - \psi^{1/2})\underset{\sim}{x}_i]$. (for detail, see Baltagi, 2001; Hsiao, 2003).

If a variable is time-invariant, like gender dummy, $x_{kit} = x_{kis} = x_{ki}$, the covariance transformation eliminates the corresponding variable from the specification. Hence, the coefficients of time-invariant variables cannot be estimated. On the other hand, if α_i is random and uncorrelated with $\underset{\sim}{x}_i, \psi \neq 1$, the GLS can still estimate the coefficients of those time-invariant variables.

4.2 Dynamic models

When the regressors of a linear model contains lagged dependent variables, say, of the form (e.g. Balestra and Nerlove, 1966)

$$\underset{\sim}{y}_i = \underset{\sim}{y}_{i,-1}\gamma + X_i\underset{\sim}{\beta} + \underset{\sim}{e}\alpha_i + \underset{\sim}{u}_i = Z_i\underset{\sim}{\theta} + \underset{\sim}{e}\alpha_i + \underset{\sim}{u}_i, \quad i = 1, \ldots, N. \tag{4.13}$$

where $\underset{\sim}{y}_{i,-1} = (y_{i0}, \ldots, y_{i,T-1})'$, $Z_i = (\underset{\sim}{y}_{i,-1}, X_i)$ and $\underset{\sim}{\theta} = (\gamma, \underset{\sim}{\beta}')'$. For ease of notation, we assume that y_{i0} are observable. Technically, we can still eliminate the individual-specific effects by premultiplying (4.13) by the transformation matrix Q ($Q\underset{\sim}{e} = \underset{\sim}{0}$),

$$Q\underset{\sim}{y}_i = QZ_i\underset{\sim}{\theta} + Q\underset{\sim}{u}_i. \tag{4.14}$$

However, because of the presence of lagged dependent variables, $EQZ_i\underset{\sim}{u}_i'Q'$ $\neq 0$ even with the assumption that u_{it} is independently, identically distributed across i and over t. For instance, the covariance transformation matrix $Q = I_T - \frac{1}{T}\underset{\sim}{e}\underset{\sim}{e}'$ transforms (4.13) into the form

$$(y_{it} - \bar{y}_i) = (y_{i,t-1} - \bar{y}_{i,-1})\gamma + (\underset{\sim}{x}_{it} - \bar{\underset{\sim}{x}}_i)'\underset{\sim}{\beta} + (u_{it} - \bar{u}_i), \quad \begin{array}{l} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{array} \tag{4.15}$$

where $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}, \bar{y}_{i,-1} = \frac{1}{T}\sum_{t=1}^{T} y_{i,t-1}$ and $\bar{u}_i = \frac{1}{T}\sum_{t=1}^{T} u_{it}$. Although, $y_{i,t-1}$ and u_{it} are uncorrelated under the assumption of serial independence of u_{it}, the covariance between $\bar{y}_{i,-1}$ and u_{it} or $y_{i,t-1}$ and $\bar{u}_i$ is of order (1/T) if $|\gamma| < 1$. Therefore, the covariance estimator of $\underset{\sim}{\theta}$ creates a bias of order (1/T) when $N \to \infty$ (Anderson and Hsiao, 1981, 1982; Nickell, 1981). Since most panel data contain large N but small T, the magnitude of the bias can not be ignored (e.g. with T=10 and γ=0.5, the asymptotic bias is -0.167).

When $EQZ_i\underset{\sim}{u}_i'Q' \neq \underset{\sim}{0}$, one way to obtain a consistent estimator for $\underset{\sim}{\theta}$ is to find instruments W_i that satisfy

$$EW_i\underset{\sim}{u}_i'Q' = \underset{\sim}{0}, \tag{4.16}$$

and

$$\text{rank } (W_i Q Z_i) = k, \tag{4.17}$$

where k denotes the dimension of $(\gamma, \underset{\sim}{\beta}')'$, then apply the generalized instrumental variable or generalized method of moments estimator (GMM) by minimizing the objective function

$$\left[\sum_{i=1}^{N} W_i(Q\underset{\sim}{y}_i - QZ_i\underset{\sim}{\theta})\right]' \left[\sum_{i=1}^{N} W_i Q \underset{\sim}{u}_i \underset{\sim}{u}_i' Q' W_i'\right]^{-1} \left[\sum_{i=1}^{N} W_i(Q\underset{\sim}{y}_i - Q\underset{\sim}{Z}_i\underset{\sim}{\theta})\right], \tag{4.18}$$

with respect to $\underset{\sim}{\theta}$. (e.g. Ahn and Schmidt, 1995; Arellano, 2003; Arellano and Bond, 1991; Arellano and Bover, 1995). For instance, one may let Q be a $(T-1)\times T$ matrix of the form

$$D = \begin{bmatrix} -1 & 1 & 0 & \cdot & \cdot \\ 0 & -1 & 1 & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & -1 & 1 \end{bmatrix}, \tag{4.19}$$

then the transformation (4.14) is equivalent to taking the first difference of (4.13) over time to eliminate α_i for $t = 2, \ldots, T$,

$$\Delta y_{it} = \Delta y_{i,t-1}\gamma + \Delta \underset{\sim}{x}_{it}'\underset{\sim}{\beta} + \Delta u_{it}, \begin{matrix} i = 1, \ldots, N, \\ t = 2, \ldots, T, \end{matrix} \tag{4.20}$$

where $\Delta = (1 - L)$ and L denotes the lag operator, $Ly_t = y_{t-1}$. Since $\Delta u_{it} = (u_{it} - u_{i,t-1})$ is uncorrelated with $y_{i,t-j}$ for $j \geq 2$ and $\underset{\sim}{x}_{is}$, for all s, when u_{it} is independently distributed over time and $\underset{\sim}{x}_{it}$ is exogenous, one can let W_i be a $T(T-1)[K + \frac{1}{2}]\times(T-1)$ matrix of the form

$$W_i = \begin{bmatrix} \underset{\sim}{q}_{i2} & \underset{\sim}{0} & \cdot & \cdot \\ \underset{\sim}{0} & \underset{\sim}{q}_{is} & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \underset{\sim}{q}_{iT} \end{bmatrix}, \tag{4.21}$$

where $\underset{\sim}{q}_{it} = (y_{i0}, y_{i1}, \ldots, y_{i,t-2}, \underset{\sim}{x}_i')'$, $\underset{\sim}{x}_i = (\underset{\sim}{x}_{i1}', \ldots, \underset{\sim}{x}_{iT}')'$, and $K = k-1$. Under the assumption that $(\underset{\sim}{y}_i', \underset{\sim}{x}_i')$ are independently, identically distributed across i, the Arellano and Bover (1995) GMM estimator takes the form

$$\underset{\sim}{\theta}_{AB,GMM} = \left\{ \left[\sum_{i=1}^{N} Z_i' D' W_i' \right] \left[\sum_{i=1}^{N} W_i A W_i' \right]^{-1} \left[\sum_{i=1}^{N} W_i D Z_i \right] \right\}^{-1} \left\{ \left[\sum_{i=1}^{N} Z_i' D' W_i' \right] \left[\sum_{i=1}^{N} W_i A W_i' \right]^{-1} \left[\sum_{i=1}^{N} W_i D \underset{\sim}{y}_i \right] \right\}, \tag{4.22}$$

where A is a $(T-1)\times(T-1)$ matrix with 2 on the diagonal elements, -1 on the elements above and below the diagonal elements and 0 elsewhere.

The GMM estimator has the advantage that it is consistent and asymptotically normally distributed whether α_i is treated as fixed or random because it eliminates α_i from the specification. However, the number of moment conditions increases at the order of T^2 which can create severe downward bias in finite sample (Ziliak, 1997). An alternative is to use a (quasi-) likelihood approach which has the advantage of having a fixed number of orthogonality conditions independent of the sample size. It also has the advantage of making use of all the available sample, hence may yield more efficient estimator than (4.22) (e.g. Binder et al., 2005; Hsiao et al., 2002). However, the likelihood approach has to formulate the joint likelihood function of $(y_{i0}, y_{i1}, \ldots, y_{iT})$ (or the conditional likelihood function $(y_{i1}, \ldots, y_{iT} \mid y_{i0})$). Since there is no reason to assume that the data generating process of initial observations, y_{i0}, to be different from the rest of y_{it}, the initial y_{i0} depends on previous values of $\underset{\sim}{x}_{i,-j}$ and α_i which are unavailable. Bhargava and Sargan (1983) suggest to circumscribe this missing data problem by conditioning y_{i0} on $\underset{\sim}{x}_i$ and α_i if α_i is treated as random. If α_i is treated as a fixed constant, Hsiao et al. (2002) propose conditioning $(y_{i1} - y_{i0})$ on the first difference of $\underset{\sim}{x}_i$.

4.3 Nonlinear models

When the unobserved individual specific effects, α_i, (and or time-specific effects, λ_t) affect the outcome, y_{it}, linearly, one can avoid the consideration of random versus fixed effects specification by eliminating them from the specification through some linear transformation such as the covariance transformation (4.8) or first difference transformation (4.20). However, if α_i affects y_{it} nonlinearly, it is not easy to find transformation that can eliminate α_i. For instance, consider the following binary choice model where the observed y_{it} takes the value of either 1 or 0 depending on the latent response function

$$y_{it}^* = \underset{\sim}{\beta}' \underset{\sim}{x}_{it} + \alpha_i + u_{it}, \tag{4.23}$$

and

$$y_{it} = \begin{cases} 1, & \text{if } y_{it}^* > 0, \\ 0, & \text{if } y_{it}^* \leq 0, \end{cases} \tag{4.24}$$

where u_{it} is independently, identically distributed with density function $f(u_{it})$. Let

$$y_{it} = E(y_{it} \mid \underset{\sim}{x}_{it}, \alpha_i) + \epsilon_{it}, \tag{4.25}$$

then

$$\begin{aligned} E(y_{it} \mid \underset{\sim}{x}_{it}, \alpha_i) &= \int_{-(\underset{\sim}{\beta}' \underset{\sim}{x}_{it} + \alpha_i)}^{\infty} f(u) du \\ &= [1 - F(-\underset{\sim}{\beta}' \underset{\sim}{x}_{it} - \alpha_i)]. \end{aligned} \tag{4.26}$$

Since α_i affects $E(y_{it} \mid \underset{\sim}{x}_{it}, \alpha_i)$ nonlinearly, α_i remains after taking successive difference of y_{it},

$$\begin{aligned} y_{it} - y_{i,t-1} = {} & [1 - F(-\underset{\sim}{\beta}' \underset{\sim}{x}_{it} - \alpha_i)] \\ & - [1 - F(-\underset{\sim}{\beta}' \underset{\sim}{x}_{i,t-1} - \alpha_i)] + (\epsilon_{it} - \epsilon_{i,t-1}). \end{aligned} \tag{4.27}$$

The likelihood function conditional on $\underset{\sim}{x}_i$ and α_i takes the form,

$$\Pi_{i=1}^{N} \Pi_{t=1}^{T} [F(-\underset{\sim}{\beta}' \underset{\sim}{x}_{it} - \alpha_i)]^{1-y_{it}} [1 - F(-\underset{\sim}{\beta}' \underset{\sim}{x}_{it} - \alpha_i)]^{y_{it}}. \tag{4.28}$$

If T is large, consistent estimator of $\underset{\sim}{\beta}$ and α_i can abe obtained by maximizing (4.28). If T is finite, there is only limited information about α_i no matter how large N is. The presence of incidental parameters, α_i, violates the regularity conditions for the consistency of the maximum likelihood estimator of $\underset{\sim}{\beta}$.

If $f(\alpha_i \mid \underset{\sim}{x}_i)$ is known, and is characterized by a fixed dimensional parameter vector, consistent estimator of $\underset{\sim}{\beta}$ can be obtained by maximizing the marginal likelihood function,

$$\Pi_{i=1}^{N} \int \Pi_{t=1}^{T} [F(-\underset{\sim}{\beta}' \underset{\sim}{x}_{it} - \alpha_i)]^{1-y_{it}} [1 - F(-\underset{\sim}{\beta}' \underset{\sim}{x}_{it} - \alpha_i)]^{y_{it}} f(\alpha_i \mid \underset{\sim}{x}_i) d\alpha_i. \tag{4.29}$$

However, maximizing (4.29) involves T-dimensional integration. Butler and Moffitt (1982); Chamberlain (1984); Heckman (1981), etc., have suggested methods to simplify the computation.

The advantage of RE specification is that there is no incidental parameter problem. The problem is that $f(\alpha_i \mid \underset{\sim}{x}_i)$ is in general unknown. If a wrong $f(\alpha_i \mid \underset{\sim}{x}_i)$ is postulated, maximizing the wrong likelihood function will not yield consistent estimator of $\underset{\sim}{\beta}$. Moreover, the derivation of the marginal likelihood through multiple integration may be computationally infeasible. The advantage of FE specification is that there is no need to specify $f(\alpha_i \mid \underset{\sim}{x}_i)$. The likelihood function will be the product of individual likelihood (e.g. (4.28)) if the errors are i.i.d. The disadvantage is that it introduces incidental parameters.

A general approach of estimating a model involving incidental parameters is to find transformations to transform the original model into a model that does not involve incidental parameters. Unfortunately, there is no general rule available for nonlinear models. One has to explore the specific structure of a nonlinear model to find such a transformation. For instance, if $f(u)$ in (4.23) is logistic, then

$$\text{Prob }(y_{it} = 1 \mid \underset{\sim}{x}_{it}, \alpha_i) = \frac{e^{\underset{\sim}{\beta}' \underset{\sim}{x}_{it} + \alpha_i}}{1 + e^{\underset{\sim}{\beta}' \underset{\sim}{x}_{it} + \alpha_i}}. \tag{4.30}$$

Since, in a logit model, the denominators of $\text{Prob}(y_{it} = 1 \mid \underset{\sim}{x}_{it}, \alpha_i)$ and $\text{Prob}(y_{it} = 0 \mid \underset{\sim}{x}_{it}, \alpha_i)$ are identical and the numerator of any sequence $\{y_{i1}, \ldots, y_{iT}\}$ with $\sum_{t=1}^{T} y_{it} = s$ is always equal to $\exp(\alpha_i s) \cdot \exp\{\sum_{t=1}^{T} (\underset{\sim}{\beta}' \underset{\sim}{x}_{it}) y_{it}\}$, the conditional likelihood function conditional on $\sum_{t=1}^{T} y_{it} = s$ will not involve the incidental parameters α_i. For instance, consider the simple case that $T = 2$, then

$$\begin{aligned} \text{Prob}(y_{i1} = 1, y_{i2} = 0 \mid y_{i1} + y_{i2} = 1) &= \frac{e^{\underset{\sim}{\beta}' \underset{\sim}{x}_{i1}}}{e^{\underset{\sim}{\beta}' \underset{\sim}{x}_{i1}} + e^{\underset{\sim}{\beta}' \underset{\sim}{x}_{i2}}} \\ &= \frac{1}{1 + e^{\underset{\sim}{\beta}' \Delta \underset{\sim}{x}_{i2}}}, \end{aligned} \tag{4.31}$$

and

$$\text{Prob}(y_{i1} = 0, y_{i2} = 1 \mid y_{i1} + y_{i2} = 1) = \frac{e^{\underset{\sim}{\beta}' \Delta \underset{\sim}{x}_{i2}}}{1 + e^{\underset{\sim}{\beta}' \Delta \underset{\sim}{x}_{i2}}}, \tag{4.32}$$

(Chamberlain, 1980; Hsiao, 2003).

This approach works because of the logit structure. In the case when $f(u)$ is unknown, Manski (1987) exploits the latent linear structure of (4.23) by noting that for given i,

$$\underset{\sim}{\beta}'\underset{\sim}{x}_{it} \gtreqless \underset{\sim}{\beta}'\underset{\sim}{x}_{i,t-1} \Longleftrightarrow E(y_{it} \mid \underset{\sim}{x}_{it}, \alpha_i) \gtreqless E(y_{i,t-1} \mid \underset{\sim}{x}_{i,t-1}, \alpha_i), \tag{4.33}$$

and suggests maximizing the objective function

$$H_N(b) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=2}^{T} sgn(\underset{\sim}{b}'\Delta\underset{\sim}{x}_{it})\Delta y_{it}, \tag{4.34}$$

where $sgn(w) = 1$ if $w > 0, = 0$ if $w = 0$, and -1 if $w < 0$. The advantage of the Manski (1987) maximum score estimator is that it is consistent without the knowledge of $f(u)$. The disadvantage is that (4.33) holds for any $c\underset{\sim}{\beta}$ where $c > 0$. Only the relative magnitude of the coefficients can be estimated with some normalization rule, say $\parallel \underset{\sim}{\beta} \parallel = 1$. Moreover, the speed of convergence is considerably slower ($N^{1/3}$) and the limiting distribution is quite complicated. Horowitz (1992) and Lee (1999) have proposed modified estimators that improve the speed of convergence and are asymptotically normally distributed.

Other examples of exploiting specific structure of nonlinear models to eliminate the effects of incidental parameters α_i include dynamic discrete choice models (Chamberlain, 1993; Honoré and Kyriazidou, 2000; Hsiao et al., 2006), symmetrically trimmed least squares estimator for truncated and censored data (Tobit models) (Honoré, 1992), sample selection models (or type II Tobit models) (Kyriazidou, 1997), etc. However, often they impose very severe restrictions on the data such that not much information of the data can be utilized to obtain parameter estimates. Moreover, there are models such that there does not appear to possess consistent estimator when T is finite.

An alternative to consider consistent estimators is to consider bias reduced estimator. The advantage of such an approach is that the bias reduced estimators may still allow the use of all the sample information so that from a mean square error point of view, the bias reduced estimator may still dominate a consistent estimators because the latter often have to throw away a lot of sample, thus tend to have large variances.

Following the idea of Cox and Reid (1987), Arellano (2001)) and Carro (2005) propose to derive the modified MLE by maximizing the modified log-likelihood function

$$L^*(\underset{\sim}{\beta}) = \sum_{i=1}^{N} \left[\ell_i^*(\underset{\sim}{\beta},\ \alpha_i(\underset{\sim}{\beta})) - \frac{1}{2} \log \ell_{i,d_i d_i}^*(\beta_1\ \alpha_i(\underset{\sim}{\beta}) \right], \tag{4.35}$$

where $\ell_i^*(\underset{\sim}{\beta},\ \alpha_i(\underset{\sim}{\beta}))$ denotes the concentrated log-likelihood function of $\underset{\sim}{y}_i$ after substituting the MLE of α_i in terms of $\underset{\sim}{\beta}$, $\alpha_i(\underset{\sim}{\beta})$, (i.e., the solution of $\frac{\partial \log L}{\partial \alpha_i} = 0$ in terms of $\underset{\sim}{\beta}, i = 1, \ldots, N$), into the log-likelihood function and $\ell_{i,\alpha_i\alpha_i}^*(\underset{\sim}{\beta},\ \alpha_i(\underset{\sim}{\beta}))$ denotes the second derivative of ℓ_i^* with respect to α_i. The bias correction term is derived by noting that to the order of $(1/T)$ the first derivative of ℓ_i^* with respect to $\underset{\sim}{\beta}$ converges to $\frac{1}{2}\frac{E[\ell_{i,\beta\alpha_i\alpha_i}^*(\underset{\sim}{\beta},\alpha_i)]}{E[\ell_{i,\alpha_i\alpha_i}^*(\underset{\sim}{\beta},\alpha_i)]}$. By subtracting the order (1/T) bias from the likelihood function, the modified MLE is biased only to the order of $(1/T^2)$, without increasing the asymptotic variance.

Monte Carlo experiments conducted by Carro (2005) have shown that when $T = 8$, the bias of modified MLE for dynamic probit and logit models are negligible. Another advantage of the Arellano-Carro approach is its generality. For instance, a dynamic logit model with time dummy explanatory variable can not meet the Honoré and Kyriazidou (2000) conditions for generating consistent estimator, but can still be estimated by the modified MLE with good finite sample properties.

4.4 Modeling cross-sectional dependence

Most panel studies assume that apart from the possible presence of individual invariant but period varying time specific effects, λ_t, the effects of omitted variables are independently distributed across cross-sectional units. However, often economic theory predicts that agents take actions that lead to interdependence among themselves. For example, the prediction that risk averse agents will make insurance contracts allowing them to smooth idiosyncratic shocks implies dependence in consumption across individuals. Ignoring cross-sectional dependence can lead to inconsistent estimators, in particular when T is finite (e.g. Hsiao and Tahmiscioglu, 2005). Unfortunately, contrary to the time series data in which the time label gives a natural ordering and structure, general forms of dependence for cross-sectional dimension are difficult to formulate. Therefore, econometricians have relied on strong parametric assumptions to model cross-sectional dependence. Two approaches have been proposed to model cross-sectional dependence: economic distance or spatial approach and factor approach.

In regional science, correlation across cross-section units is assumed to follow a certain spatial ordering, i.e. dependence among cross-sectional units is related to location and distance, in a geographic or more general economic or social network space (e.g. Anselin, 1988; Anselin and Griffith, 1988; Anselin et al., 2006). A known spatial weights matrix, $W = (w_{ij})$ an $N \times N$ positive matrix in which the rows and columns correspond to the cross-sectional units, is specified to express the prior strength of the interaction between individual (location) i (in the row of the matrix) and individual (location) j (column), w_{ij}. By convention, the diagonal elements, $w_{ii} = 0$. The weights are often standardized so that the sum of each row, $\sum_{j=1}^{N} w_{ij} = 1$.

The spatial weight matrix, W, is often included into a model specification to the dependent variable, to the explanatory variables, or to the error term. For instance, a *spatial lag* model for the $NT \times 1$ variable $\underset{\sim}{y} = (\underset{\sim}{y}_1', \dots, \underset{\sim}{y}_N')', \underset{\sim}{y}_i = (y_{i1}, \dots, y_{iT})'$, may take the form

$$y = \rho(W \otimes I_T)\underset{\sim}{y} + X\underset{\sim}{\beta} + \underset{\sim}{u} \tag{4.36}$$

where X and $\underset{\sim}{u}$ denote the $NT \times K$ explanatory variables and $NT \times 1$ vector of error terms, respectively, and $\otimes$ denotes the Kronecker product. A *spatial error* model may take the form,

$$\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{v}, \tag{4.37}$$

where $\underset{\sim}{v}$ may be specified as in a *spatial autoregressive* form,

$$\underset{\sim}{v} = \theta(W \otimes I_T)\underset{\sim}{v} + \underset{\sim}{u}, \tag{4.38}$$

or a *spatial moving average* form,

$$\underset{\sim}{v} = \gamma(W \otimes I_T)\underset{\sim}{u} + \underset{\sim}{u}. \tag{4.39}$$

The spatial model can be estimated by the instrumental variables (generalized method of moments estimator) or the maximum likelihood method. However, the approach of defining cross-sectional dependence in terms of "economic distance" measure requires that the econometricians have information regarding this "economic distance" (e.g. Conley, 1999). Another approach to model cross-sectional dependence is to assume that the error of a model, say model (4.37) follows a linear factor model,

$$v_{it} = \sum_{j=1}^{r} b_{ij} f_{jt} + u_{it}, \tag{4.40}$$

where $\underset{\sim}{f}_t = (f_{1t}, \ldots, f_{rt})'$ is a $r \times 1$ vector of random factors, $\underset{\sim}{b}_i' = (b_{i1}, \ldots, b_{ir})$, is a $r \times 1$ nonrandom factor loading coefficients, u_{it}, represents the effects of idiosyncratic shocks which is independent of $\underset{\sim}{f}_t$ and is independently distributed across i. (e.g. Bai and Ng, 2002; Moon and Perron, 2004; Pesaran, 2004). The conventional time-specific effects model is a special case of (4.40) when $r = 1$ and $b_i = b_\ell$ for all i and ℓ.

The factor approach requires considerably less prior information than the economic distance approach. Moreover, the number of time-varying factors, r, and factor load matrix $B = (b_{ij})$ can be empirically identified if both N and T are large. The estimation of a factor loading matrix when N is large may not be computationally feasible. Pesaran (2004) has therefore suggested to add cross-sectional means $\bar{y}_t = \frac{1}{N}\sum_{i=1}^{N} y_{it}, \bar{\underset{\sim}{x}}_t = \frac{1}{N}\sum_{i=1}^{N} \underset{\sim}{x}_{it}$ as additional regressors with individual-specific coefficients to (4.37) to filter out cross-sectional dependence. This approach is very appealing because of its simplicity. However, it is not clear how it will perform if N is neither small nor large. Neither is it clear how it can be generalized to nonlinear models.

4.5 Large-N and large-T panels

Our discussion has been mostly focusing on panels with large N and finite T. There are panel data sets, like the Penn-World tables, covering different individuals, industries, and countries over long periods. In general, if an estimator is consistent in the fixed-T, large-N case, it will remain consistent if both N and T tend to infinity. Moreover, even in the case that an estimator is inconsistent for fixed T and large N, (say, the MLE of dynamic model (4.13) or fixed effects probit or logit models (4.26)), it can become consistent if T also tends to infinity. The probability limit of an estimator, in general, is identical irrespective of how N and T tend to infinity. However, the properly scaled limiting distribution may depend on how the two indexes, N and T, tend to infinity.

There are several approaches for deriving the limits of large-N, large-T panels:

a. Sequential limits — First, fix one index, say N, and allow the other, say T, to go to infinity, giving an intermediate limit, then, let N go to infinity.

b. Diagonal-path limits — Let the two indexes, N and T, pass to infinity along a specific diagonal path, say $T = T(N)$ as $N \longrightarrow \infty$.

c. Joint limits — Let N and T pass to infinity simultaneously without placing specific diagonal path restrictions on the divergence.

In many applications, sequential limits are easy to derive. However, sometimes sequential limits can give misleading asymptotic results. A joint limit will give a more robust result than either a sequential limit or a diagonal-path limit, but will also be substantially more difficult to derive and will apply only under stronger conditions, such as the existence of higher moments. Phillips and Moon (1999) have given a set of sufficient conditions that ensures that sequential limits are equivalent to joint limits.

When T is large, there is a need to consider serial correlations more generally, including both short-memory and persistent components. For instance, if unit roots are present in y and x (i.e. both are integrated of order 1),, but are not cointegrated, Phillips and Moon (1999) show that if N is fixed but $T \longrightarrow \infty$, the least squares regression of y on x is a nondegenerate random variables that is a functional of Brownian motion that does not converge to the long-run average relation between y and x, but it does if N also tends to infinity. In other words, the issue of spurious regression will not arise in panel with large N (e.g. Kao, 1999).

Both theoretical and applied researchers have paid a great deal attention to unit root and cointegration properties of variables. When N is finite and T is large, standard time series techniques can be used to derive the statistical properties of panel data estimators. When N is large and cross-sectional units are independently distributed across i, central limit theorems can be invoked along the cross-sectional dimension. Asymptotically normal estimators and test statistics (with suitably adjustment for finite T bias) for unit roots and cointegration have been proposed (e.g. Baltagi and Kao, 2000; Im et al., 2003; Levin et al., 2002). They, in general, gain statistical power over their standard time series counterpart (e.g. Choi, 2001).

When both N and T are large and cross-sectional units are not independent, a factor analytic framework of the form (4.40) has been proposed to model cross-sectional dependency and variants of unit root tests are proposed (e.g. Moon and Perron, 2004). However, the implementation of those panel unit root tests is quite complicated. When $N \longrightarrow \infty$, $\frac{1}{N}\sum_{i=1}^{N} u_{it} \longrightarrow 0$, (4.40) implies that $\bar{v}_t = \bar{\underset{\sim}{b}}' \underset{\sim}{f}_t$, where $\bar{\underset{\sim}{b}}'$ is the cross-sectional average of $\underset{\sim}{b}_i' = (b_{i1}, \ldots, b_{ir})$ and $\underset{\sim}{f}_t = (f_{1t}, \ldots, f_{rt})$. Pesaran (2004, 2005) suggests a simple approach to filter out the cross-sectional dependency by augmenting the cross-sectional means, $\bar{y}_t$ and $\bar{\underset{\sim}{x}}_t$ to the regression model (4.37),

$$y_{it} = \underset{\sim}{x}_{it}'\underset{\sim}{\beta} + \alpha_i + \bar{y}_t c_i + \bar{\underset{\sim}{x}}_t' d_i + e_{it}, \tag{4.41}$$

or $\bar{y}_t, \Delta\bar{y}_{t-j}$ to the Dickey and Fuller (1979) type regression model,

$$\begin{aligned}\Delta y_{it} = \alpha_i + \delta_i t + \gamma_i y_{i,t-1} + \sum_{\ell=1}^{p_i} \phi_{i\ell}\Delta y_{i,t-\ell} + c_i\bar{y}_{t-1} \\ + \sum_{\ell=1}^{p_i} d_{i\ell}\Delta\bar{y}_{t-\ell} + e_{it},\end{aligned} \tag{4.42}$$

for testing of unit root, where $\bar{y}_t = \frac{1}{N}\sum_{i=1}^{N} y_{it}$, $\bar{\underset{\sim}{x}}_t = \dfrac{1}{N}\sum_{i=1}^{N}\underset{\sim}{x}_{it}$, $\Delta\bar{y}_{t-j} = \dfrac{1}{N}\sum_{i=1}^{N}\Delta y_{i,t-j}$ and $\Delta = (1-L)$, L denotes the lag operator. The resulting pooled estimator will again be asymptotically normally distributed.

When cross-sectional dependency is of unknown form, Chang (2002) suggests to use nonlinear transformations of the lagged level variable, $y_{i,t-1}$, $F(y_{i,t-1})$, as instruments (IV) for the usual augmented Dickey and Fuller (1979) type regression. The test static for the unit root hypothesis is simply defined as a standardized sum of individual IV t-ratios. As long as $F(\cdot)$ is regularly integrable, say $F(y_{i,t-1}) = y_{i,t-1}e^{-c_i|y_{i,t-1}|}$, where c_i is a positive constant, the product of the nonlinear instruments $F(y_{i,t-1})$ and

$F(y_{j,t-1})$ from different cross-sectional units i and j are asymptotically uncorrelated, even the variables $y_{i,t-1}$ and $y_{j,t-1}$ generating the instruments are correlated. Hence, the usual central limit theorems can be invoked and the standardized sum of individual IV t-ratios is asymptotically normally distributed.

For further review of the literature on unit roots and cointegration in panels, see Breitung and Pesaran (2006) and Choi (2006). However, a more fundamental issue of panel modeling with large N and large T is whether the standard approach of formulating unobserved heterogeneity for the data with finite T remains a good approximation to the true data generating process with large T?.

5 Concluding remarks

In this paper we have tried to provide a summary of advantages of using panel data and the fundamental issues of panel data analysis. Assuming that the heterogeneity across cross-sectional units and over time that are not captured by the observed variables can be captured by period-invariant individual specific and/or individual-invariant time specific effects, we surveyed the fundamental methods for the analysis of linear static and dynamic models. We have also discussed difficulties of analyzing nonlinear models and modeling cross-sectional dependence. There are many important issues such as the modeling of joint dependence or simultaneous equations models, varying parameter models (e.g. Hsiao, 1992, 2003; Hsiao and Pesaran, 2006), unbalanced panel, measurement errors (e.g. Griliches and Hausman, 1986; Wansbeek and Koning, 1989), nonparametric or semiparametric approach, repeated cross-section data, etc. that are not discussed, but are of no less importance.

Although panel data offer many advantages, they are not panacea. The power of panel data to isolate the effects of specific actions, treatments or more general policies depends critically on the compatibility of the assumptions of statistical tools with the data generating process. In choosing a proper method for exploiting the richness and unique properties of the panel, it might be helpful to keep the following factors in mind: First, what advantages do panel data offer us in investigating economic issues over data sets consisting of a single cross section or time series? Second, what are the limitations of panel data and the econometric methods that have been proposed for analyzing such data? Third, are the assumptions underlying the statistical inference procedures and the data-generating process compatible. Fourth, when using panel data, how can we increase the efficiency of parameter estimates and reliability of statistical inference?

Acknowledgements

I would like to thank Irene C. Hsiao for helpful discussion and editorial assistance and Kannika Damrongplasit for drawing the figures. Some of the arguments presented here also appear in Hsiao (2005, 2006).

References

Ahn, S. C. and Schmidt, P. (1995). Efficient estimation of models for dynamic panel data. *Journal of Econometrics*, 68:5–27.

Aigner, D. J., Hsiao, C., Kapteyn, A., and Wansbeek, T. (1984). Latent variable models in econometrics. In Z. Griliches and M. D. Intriligator, eds., *Handbook of Econometrics*, Vol. 2, pp. 1322–1393. North–Holland, Amsterdam.

Anderson, T. W. (1959). On asymptotic distributions of estimates of parameters of stochastic difference equations. *Annals of Mathematical Statistics*, 30:676–687.

Anderson, T. W. and Hsiao, C. (1981). Estimation of dynamic models with error components. *Journal of the American Statistical Association*, 767:598–606.

Anderson, T. W. and Hsiao, C. (1982). Formulation and estimation of dynamic models using panel data. *Journal of Econometrics*, 18:47–82.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers, Dordrecht.

Anselin, L. and Griffith, D. A. (1988). Do spatial effects really matter in regression analysis? *Papers of the Regional Science Association*, 65:11–34.

Anselin, L., Le Gallo, J., and Jayet, H. (2006). Spatial panel econometrics. In L. Mátyás and P. Sevestre, eds., *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory*, Chap. 18. Kluwer Academic Publishers, Dordrecht, 3rd ed.

Antweiler, W. (2001). Nested random effects estimation in unbalanced panel data. *Journal of Econometrics*, 101:295–313.

Arellano, M. (2001). Discrete choice with panel data. Working Paper 0101, CEMFI, Madrid.

Arellano, M. (2003). *Panel Data Econometrics*. Oxford University Press, Oxford.

Arellano, M. and Bond, S. R. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies*, 58:277–297.

Arellano, M. and Bover, O. (1995). Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics*, 68:29–51.

Arellano, M., Bover, O., and Labeaga, J. (1999). Autoregressive models with sample selectivity for panel data. In C. Hsiao, K. Lahiri, L. F. Lee, and M. H. Pesaran, eds., *Analysis of Panels and Limited Dependent Variable Models*, pp. 23–48. Cambridge University Press, Cambridge.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70:91–121.

Balestra, P. and Nerlove, M. (1966). Pooling cross-section and time series data in the estimation of a dynamic model: The demand for natural gas. *Econometrica*, 34:585–612.

Baltagi, B. H. (2001). *Econometric Analysis of Panel Data*. John Wiley and Sons, New York, 2nd ed.

Baltagi, B. H. and Kao, C. (2000). Nonstationary panels, cointegration in panels and dynamic panel, a survey. In B. H. Baltagi, ed., *Nonstationary Panels, Panel Cointegration, and Dynamic Panels*, Vol. 15 of *Advances in Econometrics*, pp. 7–51. JAI Press, Amsterdam.

Beckettii, S., Gould, W., Lillard, L., and Welch, F. (1988). The panel study of income dynamics after fourteen years: An evaluation. *Journal of Labor Economics*, 6:472–492.

Ben-Porath, Y. (1973). Labor force participation rates and the supply of labor. *Journal of Political Economy*, 81:697–704.

Bhargava, A. and Sargan, J. D. (1983). Estimating dynamic random effects models from panel data covering short time periods. *Econometrica*, 51:1635–1659.

Binder, M., Hsiao, C., and Pesaran, M. H. (2005). Estimation and inference in short panel vector autoregressions with unit roots and cointegration. *Econometric Theory*, 21:795–837.

Biørn, E. (1992). Econometrics of panel data with measurement errors. In L. Mátyás and P. Sevestre, eds., *The Econometrics of Panel Data: Theory and Applications*, pp. 152–195. Kluwer Academic Publishers, Dordrecht, 1st ed.

Breitung, J. and Pesaran, M. H. (2006). Unit roots and cointegration in panels. In L. Mátyás and P. Sevestre, eds., *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory*, Chap. 8. Kluwer Academic Publishers, Dordrecht, 3rd ed.

Butler, J. S. and Moffitt, R. (1982). A computationally efficient quadrature procedure for the one factor multinominal probit model. *Econometrica*, 50:761–764.

Carro, J. M. (2005). Estimating dynamic panel data discrete choice models with fixed effects. *Journal of Econometrics*. In press.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies*, 47:225–238.

Chamberlain, G. (1984). Panel data. In Z. Griliches and M. Intriligator, eds., *Handbook of Econometrics*, Vol. 2, pp. 1247–1318. North Holland, Amsterdam.

Chamberlain, G. (1993). Feedback in panel data models. Discussion paper, Department of Economics, Harvard University.

Chang, Y. (2002). Nonlinear IV unit root tests in panels with cross-sectional dependency. *Journal of Econometrics*, 110:261–292.

Choi, I. (2001). Unit root tests for panel data. *Journal of International Money and Finance*, 20:249–272.

Choi, I. (2006). Nonstationary panels. In K. Patterson and T. C. Mills, eds., *Palgrave Handbooks of Econometrics*, Vol. 1, pp. 511–539. Palgrave MacMillan, New York.

Conley, T. G. (1999). GMM estimation with cross-sectional dependence. *Journal of Econometrics*, 92:1–45.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B*, 49:1–39.

Davis, P. (2002). Estimating multi-way error components models with unbalanced panel data structure. *Journal of Econometrics*, 106(1):67–95. Reprinted in "Recent Developments in the Econometrics of Panel Data" B. H. Baltagi and G. Sumney (eds.), Edward Elgar Publishing, UK.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74:427–431.

Dickey, D. A. and Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica*, 49:1057–1072.

Granger, C. W. J. (1990). Aggregation of time-series variables: A survey. In T. Barker and M. H. Pesaran, eds., *Disaggregation in Econometric Modeling*. Routledge, London.

Griliches, Z. (1967). Distributed lags: A survey. *Econometrica*, 35:16–49.

Griliches, Z. and Hausman, J. A. (1986). Errors-in-variables in panel data. *Journal of Econometrics*, 31:93–118.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46:1251–1271.

Heckman, J. J. (1981). Statistical models for discrete panel data. In C. F. Manski and D. McFadden, eds., *Structural Analysis of Discrete Data with Econometric Applications*, pp. 114–178. MIT Press, Cambridge.

Heckman, J. J., Ichimura, H., Smith, J., and Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica*, 66:1017–1098.

Honoré, B. (1992). Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects. *Econometrica*, 60:533–567.

Honoré, B. and Kyriazidou, E. (2000). Panel data discrete choice models with lagged dependent variables. *Econometrica*, 68:839–874.

Horowitz, J. L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica*, 60:505–531.

Hsiao, C. (1986). *Analysis of Panel Data*, Vol. 11 of *Econometric Society Monographs*. Cambridge University Press, New York.

Hsiao, C. (1992). Random coefficient models. In L. Mátyás and P. Sevestre, eds., *The Econometrics of Panel Data*, pp. 223–241. Kluwer Academic Publishers, Dordrecht, 1st ed. Reprinted in 2nd ed. (1996) pp. 410–428.

Hsiao, C. (2003). *Analysis of Panel Data*, Vol. 34 of *Econometric Society monographs*. Cambridge University Press, Cambridge, 2nd ed.

Hsiao, C. (2005). Why panel data? *Singapore Economic Review*, 50(2):1–12.

Hsiao, C. (2006). Longitudinal data analysis. In *The New Palgrave Dictionary of Economics*. Palgrave MacMillan. In press.

Hsiao, C., Appelbe, T. W., and Dineen, C. R. (1993). A general framework for panel data analysis — with an application to canadian customer dialed long distance service. *Journal of Econometrics*, 59:63–86.

Hsiao, C., Luke Chan, M. W., Mountain, D. C., and Tsui, K. Y. (1989). Modeling ontario regional electricity system demand using a mixed fixed and random coefficients approach. *Regional Science and Urban Economics*, 19:567–587.

Hsiao, C., Mountain, D. C., and Ho-Illman, K. (1995). Bayesian integration of end-use metering and conditional demand analysis. *Journal of Business and Economic Statistics*, 13:315–326.

Hsiao, C. and Pesaran, M. H. (2006). Random coefficients models. In L. Mátyás and P. Sevestre, eds., *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory*, Chap. 5. Kluwer Academic Publishers, Dordrecht, 3rd ed.

Hsiao, C., Pesaran, M. H., and Tahmiscioglu, A. K. (2002). Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods. *Journal of Econometrics*, 109:107–150.

Hsiao, C., Shen, Y., and Fujiki, H. (2005). Aggregate vs disaggregate data analysis — a paradox in the estimation of money demand function of japan under the low interest rate policy. *Journal of Applied Econometrics*, 20:579–601.

Hsiao, C., Shen, Y., Wang, B., and Weeks, G. (2006). Evaluating the effectiveness of washington state repeated job search services on the employment rate of prime-age female welfare recipients. Working paper.

Hsiao, C. and Tahmiscioglu, T. (2005). Estimation of dynamic panel data models with both individual and time specific effects.

Im, K., Pesaran, M. H., and Shin, Y. (2003). Testing for unit roots in heterogeneous panels. *Journal of Econometrics*, 115:53–74.

Juster, T. (2000). Economics/micro data. In *International Encyclopedia of Social Sciences*.

Kao, C. (1999). Spurious regression and residual-based tests for cointegration in panel data. *Journal of Econometrics*, 90:1–44.

Kyriazidou, E. (1997). Estimation of a panel data sample selection model. *Econometrica*, 65:1335–1364.

Lee, M. J. (1999). A root-n-consistent semiparametric estimator for related effects binary response panel data. *Econometrica*, 67:427–433.

Levin, A., Lin, C., and Chu, J. (2002). Unit root tests in panel data: Asymptotic and finite-sample properties. *Journal of Econometrics*, 108:1–24.

Lewbel, A. (1994). Aggregation and simple dynamics. *American Economic Review*, 84:905–918.

MaCurdy, T. E. (1981). An empirical model of labor supply in a life cycle setting. *Journal of Political Economy*, 89:1059–1085.

Manski, C. F. (1987). Semiparametric analysis of random effects linear models from binary panel data. *Econometrica*, 55:357–362.

Mátyás, L. and Sevestre, P. (1996). *The Econometrics of Panel Data — Handbook of Theory and Applications*. Kluwer Academic Publishers, Dordrecht, 2nd ed.

Moon, H. R. and Perron, B. (2004). Testing for a unit roots in panels with dynamic factors. *Journal of Econometrics*, 122:81–126.

Nerlove, M. (2002). *Essays in Panel Data Econometrics*. Cambridge University Press, Cambridge.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16:1–32.

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica*, 49:1399–1416.

Pakes, A. and Griliches, Z. (1984). Estimating distributed lags in short panels with an application to the specification of depreciation patterns and capital stock constructs. *Review of Economic Studies*, 51:243–262.

Pesaran, M. H. (2003). On aggregation of linear dynamic models: An application to life-cycle consumption models under habit formation. *Economic Modeling*, 20:227–435.

Pesaran, M. H. (2004). Estimation and inference in large heterogeneous panels with cross-sectional dependence.

Pesaran, M. H. (2005). A simple panel unit root test in the presence of cross-section dependence. Working Paper 0346, Cambridge University DAE.

PHILLIPS, P. C. B. and DURLAUF, S. N. (1986). Multiple time series regression with integrated processes. *Review of Economic Studies*, 53:473–495.

PHILLIPS, P. C. B. and MOON, H. R. (1999). Linear regression limit theory for nonstationary panel data. *Econometrica*, 67:1057–1111.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*. John Wiley and Sons, New York, 2nd ed.

ROSENBAUM, P. and RUBIN, D. (1985). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79:516–524.

WANSBEEK, T. J. and KONING, R. H. (1989). Measurement error and panel data. *Statistica Neerlandica*, 45:85–92.

ZILIAK, J. P. (1997). Efficient estimation with panel data when instruments are predetermined: An empirical comparison of moment-condition estimators. *Journal of Business and Economic Statistics*, 15:419–431.

DISCUSSION

Marc Nerlove
Department of Agricultural and Resource Economics
University of Maryland, USA

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 the testing of 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 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

[1] In 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.

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 the 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's 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 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

[2] See also de Finetti (1970, trans. 1990, Vol. 2, pp. 211-224) and Lindley and Novick (1981).

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 $\{x_{i1}, y_{i1}, \ldots, x_{iT}, y_{iT}\}$ 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:

$$y_{it} = a + bx_{it} + \varepsilon_{it}. \tag{1}$$

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

$$\varepsilon_{it} = \mu_i + \lambda_t + u_{it} \tag{2}$$

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're 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

[3] Of 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.

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):

$$y_{it} = a_t^* + bx_{it} + \nu_{it} \tag{3}$$

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

Suppose that, given t, ν_{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

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

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 the regression (3), we say x_{it} is weakly exogenous if ν_{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 λs, and the pdf of $x_{it} \mid \{x_{is}, y_{is}, \text{all } i \text{ and } s \leq t\}$ and $x_{it} \mid \{x_{is}, y_{is}, \text{all } i \text{ and } s \leq t-1\}$ does not depend on any unknown parameters in Σ or on b or the λs. 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 λs. In that sense we can treat the observations on x as fixed.

But is exogeneity, weak or strict, a reasonable assumption? Here's 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 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 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're damned if we do, and damned if we don't. 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 a conditional density of μ_i given $x_i = (x_{i1}, \ldots, x_{iT})$, $f(\mu_i \mid x_i)$, while μ_i are unobservable. A common assumption is that $f(\mu_i \mid x_i)$ 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 when 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 order 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.

Additional references

DAVIDSON, R. and MACKINNON, J. G. (2004). *Econometric Theory and Methods*. Oxford University Press, New York.

DE FINETTI, B. (1930). Problemi determinati e indeterminati nel calculo delle probabilità. *Rendiconti della R. Accademia Nazionale dei Lincei. Series 6*, 12(9).

DE FINETTI, B. (1970). *Teoria delle Probabilità: sintesi introduttiva con appendice critica*. Giulio Einaudi Editorial, Torino. Translated as *Theory of Probability*, John Wiley & Sons, Chichester, 1990.

ENGLE, R. F., HENDRY, D. F., and RICHARD, J.-F. (1983). Exogeneity. *Econometrica*, 51:277–304.

HAAVELMO, T. (1944). *The Probability Approach in Econometrics*, Vol. 12, Supplement.

HECKMAN, J. J. (1991). Identifying the hand of the past: Distinguishng state dependence from heterogeneity. *American Economic Review*, 81(2):75–79.

LINDLEY, D. V. and NOVICK, M. R. (1981). The role of exchangeability in inference. *The Annals of Statistics*, 9:45–58.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge.

Tom Wansbeek[1]
Department of Econometrics
University of Groningen, The Netherlands
Erik Meijer
RAND Corporation and
Department of Econometrics, University of Groningen, The Netherlands

Professor Hsiao is to be congratulated with an excellent article surveying many, sometimes quite recent issues regarding panel data. He does so with great virtuosity, accuracy, and detail. Evidently, as he already mentions himself, no survey can do justice to the huge literature in the field. We would like to supplement his article with three points that we think are useful for applied researchers.

Structural equation models and software

In his Section 4, Professor Hsiao discusses several, mostly linear, models and their estimation. It is suggested that a considerable amount of technical analysis of specific cases is necessary to find a satisfactory estimator. However, many models, here and elsewhere, can be viewed as *structural equation models* (SEMs), for which widely available software can provide efficient estimators. SEM is a general framework for models with latent variables. There are several equivalent general model structures, of which the so-called LISREL model (after the LISREL program) is the most well-known:

$$\begin{aligned}
\eta_i &= \alpha + B\eta_i + \Gamma\xi_i + \zeta_i \\
y_i &= \tau_y + \Lambda_y\eta_i + \varepsilon_i \\
x_i &= \tau_x + \Lambda_x\xi_i + \delta_i,
\end{aligned}$$

where η_i is a vector of endogenous latent (unobserved) variables for observation i, ξ_i is a vector of exogenous latent variables, y_i and x_i are vectors

[1]The authors thank Arie Kapteyn for his helpful comments

of observed variables, ζ_i, ε_i, and δ_i are disturbances or errors, α, τ_y, and τ_x are vectors of intercept parameters, B and Γ are matrices of regression coefficients among the latent variables, and Λ_y and Λ_x are matrices of coefficients, called *factor loadings*, linking observed and latent variables. The first equation is a simultaneous equations regression model for the latent variables, whereas the second and third equations are factor analysis submodels, also jointly called the *measurement model*. Through the latter, errors-in-variables models fall into this class (e.g., Wansbeek, 2001), but less tangible latent constructs like technical efficiency (e.g., Ahn et al., 2001) can also be tackled in this way. By imposing restrictions on the general SEM structure a wide variety of specific models can be generated, including dynamic ones. Already Jöreskog (1978) showed how panel data models can be written as SEMs. Nevertheless, the potential of SEMs in econometrics in general, and in panel data analysis in particular, remains underexploited to a surprising degree.

Originally, the SEM framework included only linear models, but state of the art software also includes facilities for ordinal dependent variables, in which y_i and x_i are replaced by the latent variables y_i^* and x_i^* and the relations between the starred and unstarred variables are of the familiar threshold type, as in ordinal probit models. Mixture models, stratified and clustered samples, and full information estimation with missing data have also been studied in the literature and are features of some of the more advanced programs. See, e.g., Wansbeek and Meijer (2000) and the references therein for an extensive discussion of this type of model. The most widely used SEM software packages are LISREL (`http://www.ssicentral.com`), EQS (`http://www.mvsoft.com`), Mplus (`http://www.statmodel.com`), Mx (`http://www.vcu.edu/mx/`), and Amos (`http://www.spss.com/amos/`).

The actual way in which a particular model can be written as a SEM is sometimes quite complicated algebraically. But fortunately, applied researchers typically do not have to make these translations explicitly, because the software allows for a more intuitive model specification, either through almost literally writing the equations or through graphical user interfaces. This approach works best with large N and small T and can be applied with random individual effects and fixed time effects. Some programs also allow random coefficients (across individuals) in addition to random effects. The Stata package GLLAMM (`http://www.gllamm.org`) and the related book by Skrondal and Rabe-Hesketh (2004) are based on this idea of viewing random effects and random coefficients as latent variables.

Attrition

An important problem in panel data analysis, which is not explicitly mentioned in Professor Hsiao's article, is *attrition* or dropout of the study, so that for some respondents only measures on the first few time points are available. This is most problematic if the probability of dropout is related to the variables of interest; as elsewhere in econometrics, selection on an endogenous variable induces inconsistency of estimators if no precaution is taken. So endogenous attrition must be explicitly modeled to obtain consistent estimators of the parameters of interest. The topic was pioneered by Hausman and Wise (1979), who investigated the potential effect, on the estimation of earnings functions, of attrition in the Gary income maintenance experiment. The attrition process they considered was generalized by Ridder (1990), who considered the possibility of attrition depending on lagged variables. Recent contributions include Hirano et al. (2001), who show the potential of using refreshment samples in order to distinguish between various forms of attrition, and Das (2004), who provides a nonparametric approach. A recent overview of the topic of incomplete panels, of which the attrition literature forms an important subset, is given by Baltagi and Song (2006).

Panel data on aging, retirement, and health

Professor Hsiao mentions some well-known panel data sets that have proven useful for economic analysis, most notably the NLS and the PSID. A fairly recent exciting development for economists, epidemiologists, sociologists, and researchers in many other fields, is the emergence of a worldwide concerted effort of collecting panel data about aging, retirement, and health in many countries. This started with the Health and Retirement Study in the USA (HRS; http://www.rand.org/labor/aging/dataprod/, http://hrsonline.isr.umich.edu/), which is a bi-annual panel data set, with currently seven waves available (1992–2004). It was followed by the English Longitudinal Study of Ageing (ELSA; currently 2002 and 2004 available; http://www.ifs.org.uk/elsa/) and the Survey of Health, Ageing, and Retirement in Europe (SHARE; currently the first wave, 2004, available; http://www.share-project.org/), which covers 11 continental European countries, but more European countries, as well as Israel, will be added. Other countries are developing similar projects, in particular several Asian countries.

These data sets are collected with a multidisciplinary view, and thus contain lots of information about people of (approximately) 50 years and over and their households. Among others, this involves labor history and present labor force participation, income from various sources (labor, self-employment, pensions, social security, assets), wealth in various categories (stocks, bonds, pension plans, housing), various aspects of health (general health, diseases, problems with activities of daily living and mobility), subjective predictions of retirement, and actual retirement. These studies are set up such that the data are highly comparable across countries, so that in addition to cross-sectional comparisons and comparisons over time, comparisons across countries can be made as well.

Using these data, researchers can study various substantive questions that cannot be studied from other (panel) studies, such as the development of health at older age, and the relation between health and retirement. Furthermore, due to the highly synchronized questionnaires across a large number of countries, it becomes possible to study the role of institutional factors, like pension systems, retirement laws, and social security plans, on labor force participation and retirement.

Additional references

Ahn, S. C., Lee, Y. H., and Schmidt, P. (2001). GMM estimation of linear panel data models with time-varying individual effects. *Journal of Econometrics*, 101:219–255.

Baltagi, B. H. and Song, S. H. (2006). Unbalanced panel data: a survey. *Statistical Papers*, 47:493–523.

Das, M. (2004). Simple estimators for nonparametric panel models with sample attrition. *Journal of Econometrics*, 120:159–180.

Hausman, J. A. and Wise, D. A. (1979). Attrition bias in experimental and panel data: the Gary income maintenance experiment. *Econometrica*, 47:455–473.

Hirano, K., Imbens, G. W., Ridder, G., and Rubin, D. B. (2001). Combining panel data sets with attrition and refreshment samples. *Econometrica*, 69:1645–1659.

Jöreskog, K. G. (1978). An econometric model for multivariate panel data. *Annales de l'INSEE*, 30–31:355–366.

Ridder, G. (1990). Attrition in multi-wave panel data. In J. Hartog, G. Ridder, and J. Theeuwes, eds., *Panel Data and Labor Market Studies*, pp. 45–68. North–Holland, Amsterdam.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall/CRC, Boca Raton, FL.

Wansbeek, T. J. (2001). GMM estimation in panel data models with measurement error. *Journal of Econometrics*, 104:259–268.

Wansbeek, T. J. and Meijer, E. (2000). *Measurement Error and Latent Variables in Econometrics*. North–Holland, Amsterdam.

Badi H. Baltagi
Department of Economics and Center for Policy Research,
Syracuse University, USA, and
Department of Economics, University of Leicester, UK

This paper provides an excellent review of panel data methods and their application in economics. Professor Hsiao is the leading authority on panel data econometrics and he wrote the first textbook on the subject that appeared as an econometric society monograph in 1986 with the second edition appearing in 2003. As the author points out in the introduction, it is impossible to do justice to the vast and growing literature on panel data. In my discussion, I will try to complement his review with additional references that the reader might want to read. These papers are discussed in the panel data textbooks cited in the paper including Hsiao (2003), Arellano (2003) and Baltagi (2005).

The paper starts by reviewing the advantages of panel data arguing that the proliferation of panel applications in economics is due to the wider availability of panel data in both developed and developing countries. Because of space limitations, the paper does not go into pseudo-panels. These are panels constructed from consumer surveys which may not involve the same individuals or households. It does so, by focusing on cohorts, see Deaton (1985). Also, the paper does not have the space to discuss problems of attrition in panels which can be somewhat alleviated with refreshment samples, or rotating panels, see Biorn (1981).

Advantages of panel data over time series data or cross-section data is more degrees of freedom, less multicollinearity, and more variation in the data that results in more efficiency of the estimators. In addition, panel data allows us to control for heterogeneity, study dynamics, and test more complicated behavioral hypotheses than is possible with a single time series or cross-section. Panel data generate better predictions and provide micro-foundations for aggregate data analysis. The paper gives examples where panel data simplifies computation and statistical inference including the analysis of non-stationary time series, measurement error and dynamic Tobit models.

In the methodology section, the paper reviews the fixed and random effects specifications and the Hausman (1978) test. It also discusses the dynamic panel data model and the generalized method of moments (GMM) method used to estimate it. Also, non-linear panel data models including the fixed effects conditional logit model and the random effects probit model, and extensions of these models, see the Handbook chapter by Arellano and Honoré (2001) for an extensive review. The paper also discusses recent attempts at modeling cross-section dependence in panels, using spatial econometrics; see Baltagi et al. (2006) for a special issue of the *Journal of Econometrics* on spatial dependence. Also, using factor models, see the recent surveys by Choi (2006) and Breitung and Pesaran (2006). The paper also discusses large N and large T panels and how one should carry the asymptotics as discussed in Phillips and Moon (1999). It also highlights the growing literature in macro-panels on panel unit root tests and panel cointegration.

As the paper points out, panel data is not a panacea and will not solve all the problems that a time series or a cross-section study could not handle. In fact, while the paper emphasizes the advantages of panel data, one should be reminded of its limitations. These include problems in the design, data collection, and data management of panel surveys. Problems of coverage (incomplete account of the population of interest), nonresponse (due to lack of cooperation of the respondent or because of interviewer error), recall (respondent not remembering correctly), frequency of interviewing, interview spacing, reference period, the use of bounding to prevent the shifting of events from outside the recall period into the recall period, and time-in-sample bias. Measurement errors may arise because of faulty response due to unclear questions, memory errors, deliberate distortion of responses (e.g., prestige bias), inappropriate informants, misrecording of responses, and interviewer effects. Although these problems can occur in cross-section studies, they are aggravated in panel data studies. Panel data sets may also exhibit bias due to sample selection problems and attrition, see Wooldridge (1995). For the initial wave of the panel, respondents may refuse to participate, or the interviewer may not find anybody at home. This may cause some bias in the inference drawn from this sample. Although this nonresponse can also occur in cross-section data sets, it is more serious with panels because subsequent waves of the panel are still subject to nonresponse. Respondents may die, move, or find that the cost of responding is high. The rate of attrition differs across panels and usually increases from one wave to the next, but the rate of increase declines over time.

Collecting panel data is quite costly, and there is always the question of how often should one interview respondents. For example, some economists argue that economic development is far from instantaneous, so that changes from one year to the next are probably too noisy and too short-term to be really useful. They conclude that the payoff for panel data is over long time periods, five years, ten years, or even longer. In contrast, for health and nutrition issues, especially those of children, one could argue the opposite case, i.e., those panels with a shorter time span are needed in order to monitor the health and development of these children.

Users of panel data argue that these data provide several advantages worth their cost. However, as with economic data in general, the more we have of it, the more we demand of it, see Griliches (1986). The economist using panel data or any data for that matter has to know its limitations.

Additional references

Arellano, M. and Honoré, B. (2001). Panel data models: Some recent developments. Vol. 5 of *Handbook of Econometrics*, Chap. 53, pp. 3229–3296. North–Holland, Amsterdam.

Baltagi, B. H. (2005). *Econometric Analysis of Panel Data*. John Wiley & Sons, Chichester, 3rd ed.

Baltagi, B. H., Kelejian, H., and Prucha, I. (2006). Analysis of spatially dependent data. *Journal of Econometrics*, Special Issue. Forthcoming.

Biorn, E. (1981). Estimating economic relations from incomplete cross-section/time-series data. *Journal of Econometrics*, 16:221–236.

Deaton, A. (1985). Panel data from time series of cross-sections. *Journal of Econometrics*, 30:109–126.

Griliches, Z. (1986). Economic data issues. Vol. 3 of *Handbook of Econometrics*, Chap. 25, pp. 1466–1514. North–Holland, Amsterdam.

Wooldridge, J. M. (1995). Selection corrections for panel data models under conditional mean independence assumptions. *Journal of Econometrics*, 68:115–132.

Manuel Arellano
CEMFI – Madrid, Spain

This paper provides an impressive, yet compact and easily accessible review of the econometric literature on panel data analysis. Professor Cheng Hsiao has succeeded in surveying, in a coherent manner, classic results as well as more challenging recent developments on nonlinear models, cross-sectional dependence, and long time-series panels. The coverage of topics in the article reflects the breadth of Professor Hsiao's important contributions to panel data econometrics over many years.

In my comments I will focus on two aspects of recent work on nonlinear models. Firstly, I will discuss state dependence and dynamics from a treatment effect perspective. Secondly, I will consider the issue of choice of population framework and its implications for identifiability. Both comments are closely interconnected.

State dependence and treatment effects

Take a random sample of binary sequences $(y_{i1}, ..., y_{iT})$. Unit i chooses 1 or 0 in period t. This choice may depend on the choice in $t-1$. The purpose is to measure this dependence.

The problem can be cast into the framework of potential outcomes:

$$y_{it} = \begin{cases} y_{it}(1) & \text{if } y_{it-1} = 1 \\ y_{it}(0) & \text{if } y_{it-1} = 0. \end{cases}$$

The "treatment" is y_{it-1} and the potential outcomes are $y_{it}(1), y_{it}(0)$. The causal effect for person i is $y_{it}(1) - y_{it}(0)$. A measure of population state dependence is provided by the average treatment effect $E[y_{it}(1) - y_{it}(0)]$. We may also consider a conditional average given some exogenous variables or covariates. A discussion on identified bounds in this setting is in Manski (2006).

Because $(y_{i1}, ..., y_{iT})$ is a sequence of outcomes it is difficult to imagine a conceptual experiment that would justify a non-structural treatment-effects formulation. One could assign initial conditions randomly and regard the rest of the time series as a vector of outcomes, but this is not typically the intention when seeking to measure the extent of state dependence. Thus, it is natural to regard the potential outcome representation as describing a structural decision rule.

I wish to discuss an aspect of the identifying content of time-varying covariates, which is standard in the context of linear models but has not received attention from a potential outcome perspective. I consider a non-parametric partial adjustment structural model that exploits exclusion restrictions in a time-varying strictly exogenous covariate $x_i^T = (x_{i1}, ..., x_{iT})$.

The idea is that $\Pr[y_{it}(s)]$ $(s = 0, 1)$ is conditional on x_i^T, but we would expect $\Pr[y_{it}(s)]$ to be more sensitive to x_{it} than to $x's$ from other periods. A drastic but convenient implementation of this notion is:

$$\Pr[y_{it}(s) \mid x_{i1}, ..., x_{iT}] = \Pr[y_{it}(s) \mid x_{it}].$$

So, using x_{it-1}, we have the instrumental-variable (IV) assumption

$$\{y_{it}(0), y_{it}(1)\} \perp x_{it-1} \mid x_{it}.$$

As an example, think of y_{it} as smoking status, and suppose that cigarette prices x_{it-1} and x_{it} are set exogenously. The IV assumption says that, given current prices, (past smoking-induced) potential smoking outcomes are independent of past prices. This is the type of situation discussed in the local average treatment effect (LATE) literature (c.f. Imbens and Angrist, 1994).

Using a potential outcome formulation for y_{it-1} and a binary x_{it-1}:

$$y_{it-1} = \begin{cases} y_{it-1}^{[1]} & \text{if } x_{it-1} = 1 \\ y_{it-1}^{[0]} & \text{if } x_{it-1} = 0, \end{cases}$$

we can distinguish between compliers (those induced to quit smoking by changing x_{it-1} from 0 to 1: $y_{it-1}^{[0]} - y_{it-1}^{[1]} = 1$), stayers, and defiers (those with $y_{it-1}^{[0]} - y_{it-1}^{[1]} = -1$). If we rule out defiers, the distributions of $y_{it}(0)$ and $y_{it}(1)$ for compliers are point identified:

$$\Pr\left(y_{it}(s) \mid y_{it-1}^{[0]} - y_{it-1}^{[1]} = 1, x_{it}\right) \qquad (s = 0, 1).$$

Given this, we can get measures of state dependence (addiction) and price effects on smoking. Note that we have defined two different sequences of potential outcomes, $y_{it}^{[s]}$ and $y_{it}(s)$.

Exogeneity

An alternative conditional exogeneity assumption is

$$\{y_{it}(0), y_{it}(1)\} \perp y_{it-1} \mid x_{it}.$$

This is a strong assumption because y_{it-1} is not randomly assigned. A linear version of this is the standard partial adjustment model without serial correlation.

Fixed effects

The previous discussion can be thought of as being conditional on time-invariant observable covariates. The panel literature has emphasized parametric situations where the results hold conditional on a time-invariant unobserved effect α_i:

$$\{y_{it}(0), y_{it}(1)\} \perp y_{it-1} \mid x_{it}, \alpha_i \tag{1}$$

or

$$\{y_{it}(0), y_{it}(1)\} \perp x_{it-1} \mid x_{it}, \alpha_i,$$

thus allowing for "fixed-effects endogeneity" of y_{it-1} or x_{it-1}.

In situations of this kind, we only have fixed-T point identification for particular objects in certain models. An example of (1) is the binary autoregressive formulation

$$y_{it}(s) = 1(\gamma s + \alpha_i + v_{it} \geq 0) \qquad (s = 0, 1) \tag{2}$$

where v_{it} are iid across i and t, independent of α_i, with logit or probit *cdf* F.

The average treatment effect in this case is

$$\phi \equiv E[y_{it}(1) - y_{it}(0)] = E_{\alpha_i}[F(\gamma + \alpha_i) - F(\alpha_i)].$$

There is point identification of γ for logit if $T \geq 4$, but not for probit, although the identified set for γ seems to be small (Honoré and Tamer, 2006). There is set identification for ϕ for both logit and probit.

Unobserved heterogeneity and identification

Take just one individual time series and think of it as the realization of a well defined, suitably stable, but individual-specific, stochastic process. A descriptive measure of unit's i persistence is the first-order autocorrelation:

$$\begin{aligned}
\rho_i &= \mathcal{P}_i(y_{it} = 1 \mid y_{it-1} = 1) - \mathcal{P}_i(y_{it} = 1 \mid y_{it-1} = 0) \\
&= \operatorname*{plim}_{T\to\infty} \left(\frac{1}{T_1} \sum_{y_{it-1}=1} y_{it} - \frac{1}{T_0} \sum_{y_{it-1}=0} y_{it} \right)
\end{aligned}$$

where $T_1 = \sum_{t=2}^{T} y_{it-1}$ and $T_0 = \sum_{t=2}^{T}(1 - y_{it-1})$. On the other hand, a time-series average of causal effects is:

$$r_i = \operatorname*{plim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} [y_{it}(1) - y_{it}(0)].$$

In general ρ_i and r_i are different concepts. Note that

$$y_{it} = [y_{it}(1) - y_{it}(0)]\, y_{it-1} + y_{it}(0),$$

so that we have $r_i = \rho_i$ if $[y_{it}(0), y_{it}(1)]$ are independent of y_{it-1} over time. For example, this is true for the binary fixed-effect autoregressive model (2).

A cross-sectional measure of persistence in a two-period panel is

$$\pi_t = \Pr(y_{it} = 1 \mid y_{it-1} = 1) - \Pr(y_{it} = 1 \mid y_{it-1} = 0)$$
$$= \operatorname*{plim}_{N\to\infty} \left(\frac{1}{N_1} \sum_{y_{it-1}=1} y_{it} - \frac{1}{N_0} \sum_{y_{it-1}=0} y_{it} \right)$$

where $N_1 = \sum_{i=1}^{N} y_{it-1}$ and $N_0 = \sum_{i=1}^{N} (1 - y_{it-1})$. If ρ_i and π_t are constant for all i and t, they will coincide, but not otherwise.

The microeconometric literature on "genuine versus spurious" state dependence has been concerned with approximating summary measures of ρ_i from short panels. This may still be a descriptive pursuit, although $E(\rho_i)$ is arguably more informative than π_t because it distinguishes between cross-sectional unobserved heterogeneity and unit-specific time-series persistence.

Even for some of these descriptive objects we lack point identification under fixed T. However, the fact that ρ_i or cross-sectional functionals of it are not point identified from a fixed-T perspective, reflects a limitation of this perspective when T is statistically informative. I now turn to discuss this problem.

Population framework and identification

Fixed T identification may be problematic because it rules out statistical learning from individual time series data. For micro panels of moderate time dimension, approximate solutions to the incidental parameter problem from a time-series perspective (reviewed in Arellano and Hahn, 2006) are a promising avenue for progress.

In this literature three different approaches can be distinguished. One approach is to construct and analytical or numerical bias correction of a fixed effects estimator. A second approach is to consider estimators from bias corrected moment equations. The third one is to consider estimation from a bias corrected objective function relative to some target criterion. The latter is particularly attractive for its simplicity, specially in models with multiple fixed effects.

By way of illustration, suppose a likelihood model for independent data with common parameter θ and a potentially vector-valued individual effect α_i, where the log likelihood for individual i is $\sum_{t=1}^{T} \ell_{it}(\theta, \alpha_i)$. A modified concentrated likelihood that produces estimates of the common parameters with bias of order $1/T^2$ or less is given by

$$L_M(\theta) = \sum_{i=1}^{N}\left[\sum_{t=1}^{T}\ell_{it}(\theta, \widehat{\alpha}_i(\theta)) + \frac{1}{2}\ln\det H_i(\theta) - \frac{1}{2}\ln\det \Upsilon_i(\theta)\right]$$

where $\widehat{\alpha}_i(\theta)$ is the maximum likelihood estimate of α_i for given θ, $H_i(\theta) = -\sum_{t=1}^{T}\partial^2\ell_{it}(\theta, \widehat{\alpha}_i(\theta))/\partial\alpha_i\partial\alpha_i'$, and $\Upsilon_i(\theta) = \sum_{t=1}^{T} q_{it}(\theta) q_{it}(\theta)'$ where $q_{it}(\theta) = \partial\ell_{it}(\theta, \widehat{\alpha}_i(\theta))/\partial\alpha_i$. Thus, as discussed in Arellano and Hahn (2006), the adjustment depends exclusively on the sample Hessian $H_i(\theta)$ and the sample outer product of score term $\Upsilon_i(\theta)$.

Final remarks

In panel data analysis there is a choice of population framework, which may lead to conflicting identification arrangements. In situations of this kind there is much to be learned from research on both partial identification and estimability issues.

Additional references

Arellano, M. and Hahn, J. (2006). Understanding bias in nonlinear panel models: Some recent developments. In R. Blundell, W. Newey, and T. Persson, eds., *Advances in Economics and Econometrics, Ninth World Congress*. Cambridge University Press, Cambridge. Forthcoming.

Honoré, B. E. and Tamer, E. (2006). Bounds on parameters in panel dynamic discrete choice models. *Econometrica*, 74:611–629.

Imbens, G. W. and Angrist, J. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62:467–475.

Manski, C. F. (2006). Two problems of partial identification with panel data. Paper presented at 13th International Conference on Panel Data. Faculty of Economics, University of Cambridge.

Byeong U. Park
Department of Statistics
Seoul National University, Korea
Seuck Heun Song
Department of Statistics
Korea University, Korea

We congratulate Professor Hsiao on an excellent survey of recent issues in panel data. In this review paper the author discusses several important methodological contributions in the filed. We supplement Hsiao's paper by adding some additional literature and discussions on related issues that give further insights into the panel data models.

Semiparametric efficient estimation in panel models

The model (4.7) is the semiparametric model studied by Park and Simar (1994) and Park et al. (1998) if one allows both X_i and α_i to be random with unknown densities while restricts the distribution of the error vectors $\underset{\sim}{u}_i$ to be $N(0, \sigma_u^2 I_T)$. In particular, when no particular structure of dependence is assumed between X_i and α_i, it was shown by Park et al. (1998) that the within estimator given at (4.11) is semiparametric efficient. Semiparametric efficiency is an optimality property that no 'regular' estimator can have a smaller asymptotic mean squared error than an efficient estimator, see Bickel et al. (1993) for a precise definition.

When one assumes that X_i and α_i are independent, the within estimator is no longer semiparametric efficient. For this model, Park and Simar (1994) derived the asymptotic efficient bound that regular estimators can achieve and constructed an efficient estimator of $\underset{\sim}{\beta}$. Let h denote the density of α_i and write $\bar{\sigma}_u^2 = \sigma_u^2/T$. Define

$$w(z) \;=\; \int \phi_{\bar{\sigma}_u}(z-u)h(u)\,du$$

which is the density of $\alpha_i + u_i$, where $\phi_\sigma(\cdot) = \sigma^{-1}\phi(\cdot/\sigma)$. Then, the semiparametric efficiency bound is given by $\mathcal{I}^{-1} \equiv \left(\bar{\sigma}_u^{-2}\Sigma_{\text{within}} + \mathcal{I}_0\Sigma_{\text{between}}\right)^{-1}$, where

$$\mathcal{I}_0 = \int \frac{\left(w^{(1)}\right)^2}{w}(z)\,dz,$$

$$\Sigma_{\text{within}} = E\Big[T^{-1}\sum_{t=1}^{T}\big(\underset{\sim}{x}_{1t} - \underset{\sim}{x}_1\big)\big(\underset{\sim}{x}_{1t} - \underset{\sim}{x}_1\big)'\Big],$$

$$\Sigma_{\text{between}} = E\Big[\big(\underset{\sim}{x}_1 - E\underset{\sim}{x}_1\big)\big(\underset{\sim}{x}_1 - E\underset{\sim}{x}_1\big)'\Big] = \text{var}\big(\underset{\sim}{x}_1\big).$$

One may compare the asymptotic variance of the GLS estimator with the above efficiency bound. Note that the estimator given at (4.12) is GLS in the case where the mean of α_i equals 0. If one allows the mean, say μ, to take an arbitrary real value, then one can rewrite the model (4.7) as

$$\underset{\sim}{y}_i = X_i^*\underset{\sim}{\gamma} + \underset{\sim}{e}\alpha_i + \underset{\sim}{u}_i,$$

where $X_i^* = (\underset{\sim}{e}, X_i)$, $\underset{\sim}{\gamma}' = (\mu, \underset{\sim}{\beta}')$ and α_i continue to have mean 0. Thus, the GLS estimator of $\underset{\sim}{\gamma}$ is given by

$$\widehat{\underset{\sim}{\gamma}} = \left[\sum_{i=1}^{N} X_i^{*\prime} V^{-1} X_i^*\right]^{-1} \sum_{i=1}^{N} X_i^{*\prime} V^{-1} \underset{\sim}{y}_i.$$

By direct algebraic manipulations from this or by standard projection arguments, one can derive

$$\begin{aligned}\widehat{\underset{\sim}{\beta}} &= \left[\sum_{i=1}^{N}\left(X_i - \underset{\sim}{e}\,\underset{\sim}{x}'\right)' V^{-1}\left(X_i - \underset{\sim}{e}\,\underset{\sim}{x}'\right)\right]^{-1} \sum_{i=1}^{N}\left(X_i - \underset{\sim}{e}\,\underset{\sim}{x}'\right)' V^{-1}\underset{\sim}{y}_i \\ &= \underset{\sim}{\beta} + \left[\sum_{i=1}^{N}\left(X_i - \underset{\sim}{e}\,\underset{\sim}{x}'\right)' V^{-1}\left(X_i - \underset{\sim}{e}\,\underset{\sim}{x}'\right)\right]^{-1} \sum_{i=1}^{N}\left(X_i - \underset{\sim}{e}\,\underset{\sim}{x}'\right)' V^{-1} \\ &\qquad \times \left(\underset{\sim}{e}\alpha_i + \underset{\sim}{u}_i\right),\end{aligned}$$

where $\underset{\sim}{x} = N^{-1}\sum \underset{\sim}{x}_i$. It is then clear to see that, as $N \to \infty$ with T held fixed,

$$\sqrt{N}\left[N^{-1}\sum_{i=1}^{N}\left(X_i - \underset{\sim}{e}\,\underset{\sim}{x}'\right)' V^{-1}\left(X_i - \underset{\sim}{e}\,\underset{\sim}{x}'\right)\right]^{1/2}\left(\widehat{\underset{\sim}{\beta}} - \underset{\sim}{\beta}\right) \Longrightarrow N(\underset{\sim}{0}, I_d).$$

Furthermore, it can be verified that

$$\begin{aligned} N^{-1} \sum_{i=1}^{N} \left(X_i - \underset{\sim}{e}\,\underset{\sim}{x}' \right)' V^{-1} \left(X_i - \underset{\sim}{e}\,\underset{\sim}{x}' \right) &= \bar{\sigma}_u^{-2} N^{-1} T^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\underset{\sim}{x}_{it} - \underset{\sim}{x}_i \right) \left(\underset{\sim}{x}_{it} - \underset{\sim}{x}_i \right)' \\ &\quad + \left(\bar{\sigma}_u^2 + \sigma_\alpha^2 \right)^{-1} N^{-1} \sum_{i=1}^{N} \left(\underset{\sim}{x}_i - \underset{\sim}{x} \right) \left(\underset{\sim}{x}_i - \underset{\sim}{x} \right)' \\ &\xrightarrow{p} \bar{\sigma}_u^{-2} \Sigma_{\text{within}} + \left(\bar{\sigma}_u^2 + \sigma_\alpha^2 \right)^{-1} \Sigma_{\text{between}}. \end{aligned}$$

We claim that $\mathcal{I}_0 \geq \left(\bar{\sigma}_u^2 + \sigma_\alpha^2 \right)^{-1}$. This means that the asymptotic variance of the GLS estimator is greater than or equal to the asymptotic efficiency bound, as it should be. To see the claim, note that $\bar{\sigma}_u^2 + \sigma_\alpha^2$ is nothing else than the second moment of the density w. The claim follows since

$$\left(\int x^2 f(x)\, dx \right) \left(\int \frac{\left(f^{(1)}(x) \right)^2}{f(x)}\, dx \right) \geq \left(\int x \frac{f^{(1)}(x)}{f(x)} f(x)\, dx \right)^2 = 1$$

by Hölder inequality, for all absolutely continuous densities f with finite Fisher information and second moment. In the special case where $\alpha_i \sim N(0, \sigma_\alpha^2)$, one can find that $\mathcal{I}_0 = \left(\bar{\sigma}_u^2 + \sigma_\alpha^2 \right)^{-1}$.

Park et al. (2003) extended the work of Park and Simar (1994) and Park et al. (1998) by considering a more general model than (4.7) that allows the error terms to have an AR(1) dependence over time. They continued to work on the dynamic panel model (4.13). It would be interesting to compare the GMM estimators discussed in Hsiao's paper with the semiparametric efficient estimator of Park et al. (2007). The question of efficient estimation with the GMM technique in dynamic panel models was addressed by Arellano and Bond (1991), Arellano and Bover (1995) and Ahn and Schmidt (1995). It is worthwhile to note that Park et al. (2007) constructed an estimator which attains the semiparametric efficiency bound under minimal assumptions. A possible direction for future research with panel models is to work on semiparametric efficient estimation for nonlinear models discussed in Section 4.3. A major difficulty in derivation of efficiency bounds and construction of efficient estimators is the fact that the likelihood involves integration with respect to the density of the unobservable random effects α_i.

Spatial dependence in panel models

In Section 4.4, Professor Hsiao discusses the cross sectional dependence with a spatial weight matrix. Spatial dependence models deal with spatial autocorrelation and spatial heterogeneity primarily in cross-section data, see Anselin (1988). Spatial dependence models use a metric of economic distance. There is an extensive literature on estimating these spatial models using maximum likelihood methods (e.g. Anselin, 1988) and generalized method of moments (e.g. Kelejian and Prucha, 1999; Conley, 1999). Testing for spatial dependence is also extensively studied, see Anselin (2001, 1988) and Anselin and Bera (1998).

An important problem in spatial panel data analysis, which is not explicitly mentioned in Hsiao's paper, is estimation and test of model specification. In the studies of regional science, urban economic and environmental economic research, spatial panel data models are becoming increasingly attractive. Spatial panels can include issues related to interaction between observational units collected at different spatial and time scales. In recent years, there has been a growing interest in the specification and estimation of econometric relationships based on spatial panel data models with spatial error autocorrelation or a spatially lagged dependent variable. Elhost (2001, 2003) provides a survey of issues arising in maximum likelihood estimation in spatial extensions to the four panel data models used in applied research: the fixed effects model, the random effects model, the fixed coefficient model, the random coefficients model. Since the likelihood functions involve the determinant of a matrix whose dimension grows as the sample size (the number of spatial units) increases and depends on unknown parameters, maximum likelihood estimation may be computationally difficult, particularly when the sample size is large. Kapoor et al. (2007) derived a GMM estimator for spatial panel data models that is simple to compute and remains computationally feasible even for large sample sizes. The use of semiparametric methods has seen a recent increase and is an area of very active research in spatial panels (Chen and Conley, 2001; Kelejian and Prucha, 2007).

While most of the related works in spatial panel data are focused on the method of spatial estimation, applied researchers are confronted with a host of model specification problems which are treated rather inattentively in spatial panel regression models. These might involve comparison of models that are based on different spatial weight matrices and spatial model specifications.

The standard error component panel data model assumes that the disturbances have a homoskedastic variance and no spatial correlation. This may be a restrictive assumption in many panel data applications. A more general approach is to take into account spatial correlation in the model, as was studied by Anselin (1988), Baltagi et al. (2003), and Kapoor et al. (2007). In particular, Baltagi et al. (2003) considered the problem of testing jointly for random regional effects and spatial correlation across the regions. However, they did not consider the problem of serial correlation in the error terms. Baltagi et al. (2006) considered a spatial panel regression model that takes into account serial correlation over time for each spatial unit as well as spatial dependence across the units at each time point. Furthermore, the model allows heterogeneity across the spatial units through random effects. They derived joint, conditional and marginal LM and LR tests for serial, spatial correlation and heterogeneity, and studied their performances using Monte Carlo experiments. On the other hand, Holly and Gardiol (2000) derived an LM statistic which tests for homoskedasticity of the disturbances in a one-way random effect panel data model. Baltagi et al. (2007) extended the Holly and Gardiol (2000) model by allowing spatial correlation in the disturbances. They derived a joint LM test for homoskedasticity and no spatial correlation in spatial panels. However, they did not consider alternative forms of spatial lag dependence and the asymptotic properties of the test statistics, which is an interesting subject for future research.

Additional references

Anselin, L. (2001). Rao's score tests in spatial econometrics. *Journal of Statistical Planning and Inference*, 97:113–139.

Anselin, L. and Bera, A. K. (1998). Spatial dependence in linear regression models with an introduction to spatial econometrics. In A. Ullah and D. E. A. Giles, eds., *Handbook of Applied Economic Statistics*, pp. 237–289. Marcel Dekker, New York.

Baltagi, B. H., Song, S. H., Jung, B. C., and Koh, W. (2007). Testing for serial correlation, spatial autocorrelation and random effects using panel data. *Journal of Econometrics*. In press.

Baltagi, B. H., Song, S. H., and Koh, W. (2003). Testing panel data regression models with spatial error correlation. *Journal of Econometrics*, 117:123–150.

Baltagi, B. H., Song, S. H., and Kwon, J. H. (2006). Testing for heteroskedasticity and spatial correlation in a random effects panel data model. Mimeo.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation in Non- and Semi-parametric Models*. Johns Hopkins University Press, Baltimore.

Chen, X. and Conley, T. G. (2001). A new semiparametric spatial model for panel time series. *Journal of Econometrics*, 105:59–83.

Elhost, J. P. (2001). Dynamic models in space and time. *Geographical Analysis*, 33:119–140.

Elhost, J. P. (2003). Specification and estimation of spatial panel data models. *International Regional Science Review*, 26:244–268.

Holly, A. and Gardiol, L. (2000). A score test for individual heteroscedasticity in a one-way error components model. In J. Krishnakumar and E. Ronchetti, eds., *Panel Data Econometrics: Future Directions*, pp. 199–211. North-Holland, Amsterdam.

Kapoor, M., Kelejian, H. H., and Prucha, I. R. (2007). Panel data models with spatial correlated error components. *Journal of Econometrics*. In press.

Kelejian, H. H. and Prucha, I. R. (1999). A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review*, 40:509–533.

Kelejian, H. H. and Prucha, I. R. (2007). HAC estimation in a spatial framework. *Journal of Econometrics*. In press.

Park, B. U., Sickles, R. C., and Simar, L. (1998). Stochastic frontiers: a semiparametric approach. *Journal of Econometrics*, 84:273–301.

Park, B. U., Sickles, R. C., and Simar, L. (2003). Semiparametric efficient estimation of AR(1) panel data models. *Journal of Econometrics*, 117:279–309.

Park, B. U., Sickles, R. C., and Simar, L. (2007). Semiparametric efficient estimation of dynamic panel data models. *Journal of Econometrics*, 136:281–301.

Park, B. U. and Simar, L. (1994). Efficient semiparametric estimation in a stochastic frontier models. *Journal of the American Statistical Association*, 89:929–936.

Yongcheol Shin
Leeds University Business School
University of Leeds, U.K.

I am glad to have an opportunity to discuss the paper by Hsiao on Panel data analysis - Advantages and Challenges. This paper not only provides a comprehensive review of the panel data studies over the last four decades but also addresses the few challenging methodological issues currently being encountered. The scope of the paper is too broad, so I decide to focus only on a couple of challenging modelling issues.

1 Estimation of the impacts of individual specific regressors

Consider the panel-data model,

$$y_{it} = \beta'\mathbf{x}_{it} + \gamma'\mathbf{z}_i + \varepsilon_{it},\ i = 1, ..., N,\ t = 1, ..., T, \tag{1}$$

where $\mathbf{x}_{it}$ is a $k \times 1$ vector of variables that vary over individuals and time periods, $\mathbf{z}_i$ is a $g \times 1$ vector of time-invariant individual-specific variables, and β and γ are conformably defined column vectors of parameters. The main advantage of the panel data analysis is to control for unobserved heterogeneous individual and time effects. Following recent studies highlighting the importance of explicitly modelling the cross-sectional dependency (e.g. Pesaran, 2006), we assume that unobservable errors, ε_{it}, follow a two-way heterogeneous error components structure:

$$\varepsilon_{it} = \alpha_i + \sum_{j=1}^{r} \lambda_{ij}' f_{jt} + u_{it}, \tag{2}$$

where α_i is an individual effect that might be correlated with explanatory variables, $\mathbf{f}_t = (f_{1t}, ..., f_{rt})'$ is an $r \times 1$ vector of unobserved time-specific common factors and u_{it} is a zero mean idiosyncratic disturbance uncorrelated across cross section units and over time periods. The distinguishing feature is that this model allows a certain degree of cross-sectional dependency of ε_{it} via heterogeneous individual factor loading coefficients, $\lambda_i = (\lambda_{i1}, ..., \lambda_{ir})'$. Most recent studies have focussed on consistent estimation of β only, implicitly overlooking the issue of consistent estimation of γ which might be of interest to policy makers. This can be easily handled in a two-step approach (HT) proposed by Hausman and Taylor (1981). In this regard, Serlenga and Shin (2007) develop a generalized HT estimator for a heterogeneous panel given by (1) and (2). Their approach specifically extends the correlated common effect pooled estimator of Pesaran (2006) and proposes consistent estimation of the impact of both time-varying and time-invariant regressors. Their empirical results for a gravity equation of bilateral trade amongst 15 EU member countries over the period 1960-2001 yield much more sensible results than assuming homogeneous fixed time effects. However, their approach requires that both N and T are sufficiently large. Therefore, it is worth investigating how to extend this heterogeneous approach into (possibly) dynamic panels with large N and fixed T. Ahn et al. (2007) analyse almost the same model (but without directly including time-invariant variables) and provide an alternative GMM estimator which is valid for large N and small T. Alternatively, we may develop a different approach which does not require to consistently estimate nuisance parameters capturing cross-sectional correlation structure. In this regard the subsampling approach applied to the panel data-based inference by Choi and Chue (2007) and the bias corrected estimation procedure based on the iterative bootstrap applied to homogeneous dynamic panels with fixed time dimension by Eberaert and Pozzi (2006) would be a promising candidate. Hence it would be a fruitful challenge to develop a generalized HT estimation procedure in these contexts.

2 Nonlinear regime-switching panel data modelling

In recent time series literature there have been many studies that examine the implications of the existence of a particular kind of nonlinear asymmetric dynamics. Examples of popular nonlinear models are Markov-Switching, Smooth Transition and Threshold Autoregression Models. The popularity of these models is that they allow to draw inferences about the underlying data generating process or to yield reliable forecasts in a manner that is not possible using only linear models. However, until recently, most econometric analysis has stopped short of studying these issues explicitly within a panel data context. Hansen (1999) develops the panel threshold regression model where regression coefficients can take on a small number of different values, depending on the value of other exogenous stationary variable. González et al. (2005) generalise and develop a panel smooth transition regression model which allows the coefficients to change gradually from one regime to another. In a broad context these models may be a specific example of various panel data approaches that allow coefficients to vary over time and across cross-sectional units as considered by Hsiao (2003, Chapter 6). Both approaches are static in nature, thus need to be extended into dynamic panels. Eventually, we should address the most challenging issue as how best to model nonlinear asymmetric dynamics mechanism with possibly nonstationary variables, cross-sectional heterogeneity and interdependency, simultaneously. As an example, consider the dynamic panel nonlinear error correction model where the long-run relationship is given by the linear cointegration while the associated error correction adjustment is governed by the two-regime threshold switch mechanism:

$$\begin{aligned}\Delta y_{it} =& \rho_{1i}\xi_{i,t-1}1\left(\xi_{i,t-1} < \gamma_i\right) + \rho_{2i}\xi_{i,t-1}1\left(\xi_{i,t-1} \geq \gamma_i\right) \\ &+ \omega_i'\Delta \mathbf{x}_{it} + \sum_{j=1}^{p-1} \psi_{ij}'\Delta \mathbf{z}_{i,t-j} + \varepsilon_{it}, \end{aligned} \tag{3}$$

where $\xi_{it} = y_{it} - \beta_i'\mathbf{x}_{it}$, $\mathbf{z}_{it} = (y_{it}, \mathbf{x}_{it}')'$ and $1(\cdot)$ is an indicator function and γ_i's are unknown threshold parameters. For simplicity we assume $\beta_i = \beta$ and $\gamma_i = \gamma$, which implies that the long-run cointegrating parameter is homogeneous and the associated threshold parameters are common across cross-section units. This is an extension of the panel error correction model advanced by Pesaran et al. (1999). One might imagine an example where a common threshold might be in effect in, for example, firm growth rates where the costs of listing on the stock market might create a threshold size above which firms can access relatively cheap capital and grow quickly. It would be an important challenge to allow ε_{it} in (3) to follow a two-way heterogeneous error components structure, as in (2), and develop a rigorous estimation and inference theory for the cases both with large N and large T and with large N and small T. In addition the panel-based non-linear testing procedures for unit root and cointegration would make an important contribution too. Finally, there has been some attempts to model heterogeneity and dynamics for both conditional means and conditional variances in panels. For example, Meghir and Pistaferri (2004) model the conditional variance of the income shocks as parsimonious ARCH with observable and unobservable heterogeneity and find strong evidence of ARCH effects and of unobserved heterogeneity using the data set drawn from the 1967-1992 PSID. Following various asymmetric modelling approaches to the time-varying conditional heteroskedasticity it would be interesting and challenging to embed the possibility of asymmetry of response to time-varying shocks in (3).

Additional references

Ahn, S. C., Lee, Y. H., and Schmidt, P. (2007). Panel data models with multiple time varying individual effects. *Journal of Productivity Analysis*. Forthcoming.

Choi, I. and Chue, T. (2007). Subsampling hypothesis tests for nonstationary panels with applications to exchange rates and stock prices. *Journal of Applied Econometrics*. Forthcoming.

Eberaert, G. and Pozzi, L. (2006). Bootstrap based bias correction for dynamic panels. Ghent University. mimeo.

González, A., Teräsvirta, T., and van Dijk, D. (2005). Panel smooth transition model and an application to investment under credit constraints. Working paper, Stockholm School of Economics.

Hansen, B. E. (1999). Threshold effects in non-dynamic panels: Estimation, testing and inference. *Journal of Econometrics*, 93:345–368.

Hausman, J. and Taylor, W. (1981). Panel data and unobservable individual effect. *Econometrica*, 49:1377–1398.

Meghir, C. and Pistaferri, L. (2004). Income variance dynamics and heterogeneity. *Econometrica*, 72:1–32.

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74:967–1012.

Pesaran, M. H., Shin, Y., and Smith, R. P. (1999). Pooled mean group estimation of dynamic panels. *Journal of the American Statistical Association*, 94:621–634.

Serlenga, L. and Shin, Y. (2007). Gravity models of intra-EU trade: Application of the CCEP-HT estimation in heterogeneous panels with unobserved common time-specific factors. *Journal of Applied Econometrics*. Forthcoming.