# Why Panel Data?

By

**Cheng Hsiao**

**September 2005**

**IEPR WORKING PAPER 05.33**

**INSTITUTE OF ECONOMIC POLICY RESEARCH**

**UNIVERSITY OF SOUTHERN CALIFORNIA**

**http://www.usc.edu/iepr**

# Why Panel Data?

**Cheng Hsiao**[*]

**Department of Economics**

**Nanyang Technological University, Singapore**

**and**

**University of Southern California**

**Los Angeles, CA 90089-0253**

**September 6, 2005**

## 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.

Keywords: Panel data; Longitudinal data; Unobserved heterogeneity; Random effects; Fixed effects

[*] I would like to thank Irene C. Hsiao for helpful discussion and editorial assistance and Kannika Damrongplasit for drawing the figures

## 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 $\ell$ in city $j$ of country $i$ at time $t$ (e.g. Antweiler (2001), Davis (1999)). 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 longitudinal data", according to Social Sciences Citation index. By 2003, there were 580, and in 2004, there were 687. The growth of applied studies and the methodological development of new econometric tools of panel data have been simply phenomenal since the seminar 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), Matyas and Sevester (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, Gould, Lillard and Welch (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 less multicollinearity 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.

(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. 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. Panel data has the advantage that it is possible to observe the before- and after-effects of receiving the treatment of the same individual as well as providing the possibility of isolating the effects of treatment from other factors affecting the outcome. For instance, consider the "three-strike law" for deterring crimes in California. The "three-strike law" stipulates that if a suspect is convicted three times, then she will be jailed for life. Evaluating the effects of such a law on deterring crimes by comparing the crime rates of California after the introduction of such law, denoted by $y_{it}$, and the crime rates of another state without the "three-strike law" at the same time, say the state of Oregon, denoted by $y_{jt}$, can be misleading because although the many other factors that also affect the crime rates may be very different between the two States. On the other hand, simply comparing the crime rates of California before the introduction of the "three strike law", denoted by $y_{is}$, and after the introduction using California time series data can also be misleading because although factors that stay more or less constant over this time period, say climate and demographics, are eliminated by differencing $y_{it}$ and $y_{is}$, but the effects of those factors that vary over time, say, unemployment rate, remain, so $(y_{it} - y_{is})$ represents the combined effects of "three-strike law" and the changes of unemployment rate. However, if those time-varying factors move in a similar fashion in California and Oregon, then the further differencing of the differences between California and Oregon $[(y_{it} - y_{is}) - (y_{jt} - y_{js})]$ will be able to isolate the effects of "three-strike" law from other factors that also affect crimes. This *difference-in-difference* method can work only if panel data are available (Lee (2005)).

(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 the logarithm of the worker's marginal utility of initial wealth from the model specification because it is unobserved can lead to seriously biased inference on the wage elasticity on hours worked because 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, Appelbe and Dineen (1993), Hsiao, Chan Mountain and Tsui (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 (1992); 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, Shen and Fujiki (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 (1979,81), 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, Hsiao and Pesaran (2005), Levin, Lin and Chu (2002), Im, Pesaran and Shin (2004), Phillips and Moon (1999)).

(iii.b) Measurement errors.

Measurement errors can lead to underidentification of an econometric model (e.g. Aigner, Hsiao, Kapteyn and Wansbeek (1985)). 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. Biorn (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, Bover, and Labeager (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 welath as in MaCurdy (1981) labor supply model discussed in (iib) on section 3), represented by $\alpha_i$, so that the observed $(y_{it}, \underset{\sim}{x}_{it}), i = 1, \ldots, N, t = 1, \ldots, T$, are actually generated by

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

as depicted by Figure 1, 2 and 3 in which the broken-line ellipses represent the point scatter of individual observations around the 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 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 not possible to make any inference on $\underset{\sim}{\beta}$ because the unknown $\underset{\sim}{\gamma}_{it}$ 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, $\alpha_i$, period individual-invariant, $\lambda_t$, and individual time-varying variables, $\delta_{it}$. 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.

The challenge of panel methodology is to control the impact of unobserved heterogeneity, represented by the incidental parameters, $\gamma_{it}$, to obtain valid inference on the structural parameters $\underset{\sim}{\beta}$. For ease of exposition, I shall assume $\gamma_{it} = \alpha_i$, that is, there are only unobserved individual-specific effects present. The unobserved heterogeneity can affect the outcomes linearly or nonlinearly. Model (4.3) is an example of unobserved heterogeneity, $\gamma_{it} = \alpha_i$, that affects the outcome linearly. If $y_{it}$ is unobservable, the observed data instead take the form of $(d_{it}, \underset{\sim}{x}_{it})$, $i = 1, \ldots, N$ and $t = 1, \ldots, T$, where

$$d_{it} = \begin{cases} 1, \text{ if } y_{it} > 0, \\ 0, \text{if } y_{it} \leq 0. \end{cases} \tag{4.5}$$

Then, conditional on $\underset{\sim}{x}_{it}$ and $\alpha_i$,

$$\begin{aligned} E(d_{it} \mid \underset{\sim}{x}_{it}, \alpha_i) &= \text{Prob}(d_{it} = 1 \mid \underset{\sim}{x}_{it}, d_i) \\ &= \int_{-(\underset{\sim}{\beta}' \underset{\sim}{x}_{it} + \alpha_i)}^{\infty} f(u) du, \end{aligned} \tag{4.6}$$

where $f(u)$ denotes the probability density of $u$, is an example of unobserved individual-specific effects affecting the outcome nonlinearly.

Since $\alpha_i$ is unobserved, there are two approaches. One is to assume $\alpha_i$ as a random variable (RE model). If the conditional density of $\alpha_i$ given $\underset{\sim}{x}_i = (\underset{\sim}{x}_{i1}, \ldots, \underset{\sim}{x}_{iT})$ is known, one can integrate out $\alpha_i$ to obtain the (marginal) conditional density of $\underset{\sim}{y}_i' = (y_{it}, \ldots, y_{iT})$ or $\underset{\sim}{d}_i' = (d_{i1}, \ldots, d_{iT})$ given $\underset{\sim}{x}_i$. The other is to treat $\alpha_i$ as unknown parameters (FE model).

The advantage of RE specification is that the number of unknown parameters stay constant as sample size increases. The disadvantages are that the (marginal) conditional density of $\underset{\sim}{y}_i$ or $\underset{\sim}{d}_i$ given $\underset{\sim}{x}_i$ involves $T$-dimensional integrations. It may be computationally unwieldy. In addition, the individual-specific effects, $\alpha_i$, are unobserved. If the conditional density of $\alpha_i$ given $\underset{\sim}{x}_i$ is misspecified, the marginal likelihood of $y_i$ given $\underset{\sim}{x}$, will be misspecified. Statistical inferences based on a wrong likelihood function may be misleading.

The FE specification eliminates the needs to specify the conditional density of $\alpha_i$ given $\underset{\sim}{x}_i$, hence there is no need to evaluate the $T$-dimensional integration. However, typically there are not enough degrees of freedom to obtain precise information on the incidental parameters, $\alpha_i$, as many panel data sets are of large $N$ and small $T$ type. When the inference on the structural parameters, $\underset{\sim}{\beta}$, depends on the incidental parameters, $\alpha_i$, the imprecisely estimated $\alpha_i$, affects the inference of $\underset{\sim}{\beta}$.

A general rule for obtaining consistent estimators of structural parameters, $\underset{\sim}{\beta}$, in the presence of incidental parameters $(\alpha_1, \ldots, \alpha_N)$ is to find transformations so that the likelihood function of the transformed model does not depend on the incidental parameters. If the individual-specific effects affect the outcome linearly as in (4.3), because $\alpha_i$ is time-invariant, it is possible to eliminate $\alpha_i$ from the specification by taking some linear transformation, say, taking the time difference of an individual equation. If $\alpha_i$ affects the outcomes nonlinearly, unfortunately, there is no general rule to transform a nonlinear model to eliminate the incidental parameters. The conditional maximum likelihood estimator of logit model (Chamberlain (1980)), the maximum score estimator of binary choice model (Manski (1987)), the symmetrically trimmed least squares estimator for truncated or censored data (Honoré (1992)) etc. are examples of exploiting the specific structures of nonlinear models to find transformations that do not depend on the incidental parameters $\alpha_i$. While all these methods are ingenious, they also often impose very severe restrictions on the data that it may be difficult to extract useful information in many empirical studies. For instance, Hsiao, Shen, Wang and Weeks (2005) have proposed a transitional probability model to evaluate the effectiveness of Washington State repeated job search services on the employment rate of prime-age female welfare recipients. However, in order to control the impact of unobserved individual-specific effects, they have to impose the conditions that out of the observed sample only (i) those individuals with $T \geq 4$, (ii) the first period and fourth period outcomes of an individual's employment status are identical, and (iii) the conditional variables must be identical for the third and fourth period also, can be used. As a result, less than 10% of the observed sample roughly satisfy these conditions. It appears that to devise simple, yet efficient estimators of nonlinear panel data models that do not put such stringent conditions on sample remains a challenge for econometricians.

## 5. Concluding Remarks

Although panel data offer many advantages, they are not a 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 a panel, it might be helpful to keep the following factors in mind: First, what advantages do the panel data offer us in investigating economic issues of interest ? Second, what are the limitations of the 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, how can we increase the efficiency of parameter estimators?

# REFERENCES

Aigner, D.J., C. Hsiao, A. Kapteyn and T. Wansbeek (1985), "Latent Variable Models in Econometrics", in *Handbook of Econometrics*, vol. II., ed. By Z. Griliches and M.D. Intriligator, Amersterdam: North-Holland, 1322-1393.

Anderson, T.W. (1959) "On Asymptotic Distributions of Estimates of Parameters of Stochastic Difference Equations", *Annals of Mathematical Statistics* 30, 676-687.

Antweiler, W. (2001). "Nested Random Effects Estimation in Unbalanced Panel Data", *Journal of Econometrics*, 101, 295-313.

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

Arellano, M., O. Bover and J. Labeaga (1999), "Autoregressive Models with Sample Selectivity for Panel Data", in *Analysis of Panels and Limited Dependent Variable Models*, ed., by C. Hsiao, K. Lahiri, L.F. Lee and M. H. Pesaran, Cambridge: Cambridge University Press, 23-48.

Balestra, P. and M. Nerlove (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. (2001), *Econometric Analysis of Panel Data*, 2nd ed., New York: Wiley.

Becketti, S., W. Gould, L. Lillard and F. Welch (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 Labor", *Journal of Political Economy*, 81, 697-704.

Binder, M., C. Hsiao and M. H. Pesaran (2005). "Estimation and Inference in Short Panel Vector Autoregressions with Unit Roots and Cointegration", *Econometric Theory*, 21, 795-837.

Biorn, E. (1992), "Econometrics of Panel Data with Measurement Errors" in *Econometrics of Panel Data: Theory and Applications*, ed. By. L. Mátyás and P. Sevestre, Klumer, 152-195.

Chamberlain, G. (1980), "Analysis of Covariance with Qualitative Data", *Review of Economic Studies*, 47, 225-238.

Davis, P. (1999). “Estimating Multi-way Error Components Models with Unbalanced Panel Data Structure”, mimeo, MIT Sloan School.

Dickey, D.A. and W.A. Fuller (1979), “Distribution of the Estimators for Autoregressive Time Series with a Unit Root”, *Journal of the American Statistical Association*, 74, 427-431.

____________ (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 *Disaggregation in Econometric Modelling*, ed. By. T. Barker and M.H. Pesaran, London: Routledge.

Griliches, Z. (1967), “Distributed Lags: A Survey”, Econometrica, 35, 16-49.

____________ and J.A. Hausman (1986), “Errors-in-Variables in Panel Data”, *Journal of Econometrics*, 31, 93-118.

Honoré, B.E. (1992) “Trimmed LAD and Lest Squares Estimation of Truncated and Censored Regression Models with Fixed Effects”, *Econometrica*, 60, 553-567.

Hsiao, C., (1986) “*Analysis of Panel Data*, Econometric Society monographs No. 11, New York: Cambridge University Press.

____________ (2003), *Analysis of Panel Data*, 2nd edition, Econometric Society Monograph 36, New York: Cambridge University Press.

Hsiao, C., T.W. Appelbe, and C.R. Dineen (1993), “A General Framework for Panel Data Analysis—With an Application to Canadian Customer Dialed Long Distance Service”, *Journal of Econometrics*, 59, 63-86.

____________, Y. Shen and H. Fujiki (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.

____________ M.W. Luke Chan, D.C. Mountain and K.Y. Tsui (1989), “Modeling Ontario Regional Electricity System Demand Using Mixed Fixed and Random Coefficients Approach”, *Regional Science and Urban Economics*, 19, 567-587.

____________Y. Shen, B. Wang and G. Weeks (2005), “Evaluating the Effectiveness of Washington State Repeated Job Search Services on the Employment Rate of Prime-age Female Welfare Recipients”, *mimeo*.

Im, K., M.H. Pesaran and Y. Shin (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*, (forthcoming).

Lee, M.J. (2005), *Micro-Econometrics for Policy, Program and Treatment Analysis*, Oxford: Oxford University Press.

Lewbel, A. (1994), “Aggregation and Simple Dynamics”, *American Economic Review*, 84, 905-918.

Levin, A., C. Lin, and J. Chu (2002), “Unit Root Tests in Panel Data: Asymptotic and Finite-Sample Properties”, *Journal of Econometrics*. 108, 1-24.

MaCurdy, T.E. (1981), “An Empirical Model of Labor Supply in Life Cycle Setting”, *Journal of Political Economy*, 89, 1059-85.

Manski, C.F. (1987), “Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data”, *Econometrica,* 55, 357-362.

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

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

Pakes, A. and Z. Griliches (1984), “Estimating Distributed Lags in Short Panels with 15an 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.

Phillips, P.C.B. and S.N. Durlauf (1986), “Multiple Time Series Regression with Integrated Processes”, *Review of Economic Studies*, 53, 473-495.

____________ and H.R. Moon (1999), “Linear Regression Limit Theory for Nonstationary Panel Data”, *Econometrica*, 67, 1057, 1111.

Wansbeek, T. J. and R.H. Koning (1989), “Measurement Error and Panel Data”, *Statistica Neerlandica*, 45, 85-92.

Scatter Diagrams of (y(i,t),x(i,t))

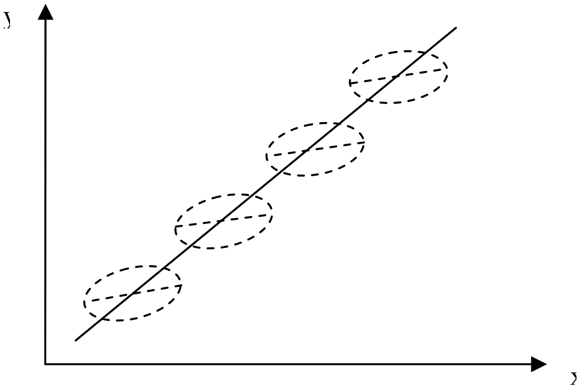

Figure 1

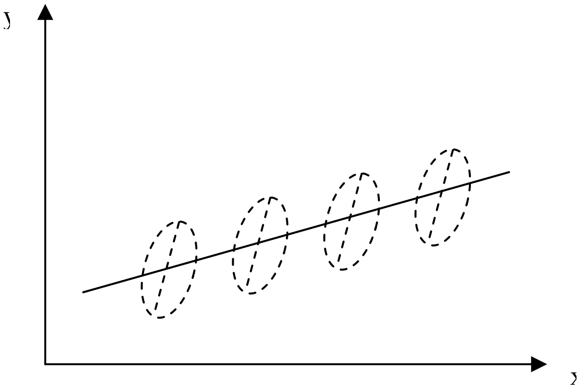

Figure 2

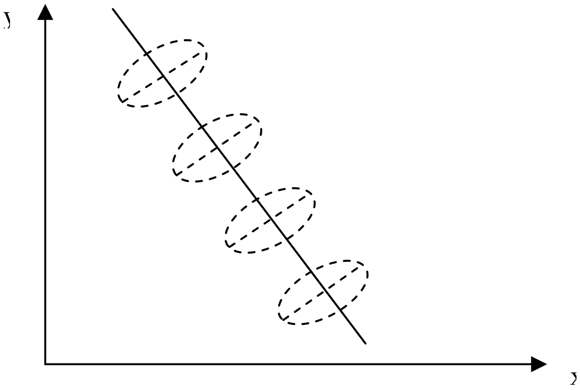

Figure 3