Applied Microeconometrics (L5): Panel Data-Basics

Nicholas Giannakopoulos

University of Patras
Department of Economics

ngias@upatras.gr

November 10, 2015

Overview

What are Panel Data?

Panel data: a type of longitudinal data (i.e., data collected at different points in time)

- Three main types of longitudinal data
  - **Time series data**. Examples: stock price trends, aggregate national statistics.
  - **Pooled cross sections**. Examples: General Social Surveys, IPUMS Census extracts, US Current Population Surveys
  - **Panel data**. Example: Panel surveys of households and individuals (PSID, NLSY, EU-ECHP, EU-SILC), Data on organizations and firms at different time points (AMADEUS-European Firms, iMentor-Greek Firms), Aggregated regional data over time (Eurostat), Country aggregates over time (OECD Statistics, Barro-Lee 1960-2010)

Why to analyse Panel Data?

- Describe changes over time
  - social change, e.g. changing attitudes, behaviors, social relationships
  - individual growth or development, e.g. life-course studies, child development, career trajectories, school achievement
  - occurrence (or non-occurrence) of events
- Estimate trends in social phenomena
  - Panel models can be used to inform policy e.g. health, obesity
  - Multiple observations on each unit can provide superior estimates as compared to cross-sectional models of association
- Estimate causal models
  - Policy evaluation
  - Estimation of treatment effects

Time series data

- Many observations (large t) on as few as one unit (small N)
  - $(x_t, t = 1, ..., T)$ univariate series, e.g. a price series: Its path over time is modeled. The path may also depend on third variables.
  - Multivariate, e.g. several price series: Their individual as well as their common dynamics is modeled. Third variables may be included.

Cross sectional data

- Two or more independent samples of many units (large N) drawn from the same population at different time periods. Are observed at a single point of time for several individuals, countries, assets, etc.,
  - $(x_i, i = 1, ..., N)$
  - Researcher's aim: model the distinction of single individuals (i.e., the heterogeneity across individuals)

Pooling Data

Two or more independent data sets of the same type.

- Pooled time series
  - we observe e.g. return series of several sectors, which are assumed to be independent of each other, together with explanatory variables. The number of sectors, N, is usually small.
  - Observations are viewed as repeated measures at each point of time. So parameters can be estimated with higher precision due to an increased sample size.
- Pooled cross sections
  - Mostly these type of data arise in surveys, where people are asked about e.g. their attitudes to political parties. This survey is repeated, T times, before elections every week. T is usually small.
  - So we have several cross sections, but the persons asked are chosen randomly. Hardly any person of one cross section is member of another one. The cross sections are independent.
  - Only overall questions can be answered, like the attitudes within males or women, but no individual (even anonymous) paths can be identified.

Panel Data

Two or more observations (small t) on many units (large N). A panel data set (also longitudinal data) has both a cross-sectional and a time series dimension, where all cross section units are observed during the whole time period.

- Structure: $(x_{it},\ i = 1, ..., N\ ;\ t = 1, ..., T)$
- Types of panel data
  - balanced
  - unbalanced
- Example: The Greek Household Budget Survey (HBS) is a household survey, with the same size of 4000 hh each year. It collects information on households' composition, members' employment status, living conditions and members' expenditure on goods and services as well as on their income. The main purpose of the HBS is to determine in detail the household expenditure pattern in order to revise the Consumer Price Index.

Panel Data

- Balanced panel: all individuals are present in all periods
- Unbalanced panel: individuals are observed a different number of times, e.g. because of missing values
- Efficiency gains using panel data vs pooling cross-sections: in panel data (especially in balanced panels) the observation of one individual for several periods reduces the variance compared to repeated random selections of individuals
- While the mechanics of the unbalanced case are similar to the balanced case, a careful treatment of the unbalanced case requires a formal description of why the panel may be unbalanced, and the sample selection issues can be somewhat subtle (issues of sample selection and attrition)

Micro-Panel vs Macro-Panel

- A micro-panel data set is a panel for which the time dimension T is largely less important than the individual dimension N (e.g., Panel Study of Income Dynamics, PSID with 15,000 individuals observed since 1968): $T << N$
- A macro-panel data set is a panel for which the time dimension T is similar to the individual dimension N (example: a panel of 100 countries with quarterly data since 1945): $T \simeq N$

Advantages of Panel Data

Example: Suppose that a cross-sectional sample of married women is found to have an average yearly labor-force participation rate of 50% [Ben-Porath, Y. (1973), "Labor Force Participation Rates and the Supply of Labor", Journal of Political Economy, 81, 697-704.)].

- It might be interpreted as implying that each woman in a homogeneous population has a 50 percent chance of being in the labor force in any given year.
- It might imply that 50 percent of the women in a heterogeneous population always work and 50 percent never work.
- To discriminate between these two models, we need to utilize individual labor-force histories (the time dimension) to estimate the probability of participation in different sub-intervals of the life cycle.
- Panel data allows to control for omitted (unobserved or mis-measured) variables.

Advantages of Panel Data

[Hsiao, C. (2014). *Analysis of Panel Data*, Third edition, Econometric Society Monographs 54. Cambridge University Press.]

1. More accurate inference of model parameters/efficiency of econometric estimates
   - Large number of data points, increased degrees of freedom, reduced collinearity
2. Greater capacity for constructing more realistic behavioral hypotheses
   - For instance, panel data overcome the problem imposed by the typical assumption in cross-sectional data i.e., $E(y_i|x_i = a) = E(y_j|x_j = a)$ (individuals with same x have the same expected value)
3. Uncovering dynamic relationships
   - "economic behavior is inherently dynamic" (Nerlove, 2000): due to institutional or technological rigidities or inertia in human behavior
   - Microdynamic and macrodynamic effects cannot be estimated using cross-sectional data.
   - Distributed-lag time-series models cannot provide good estimates: $y_t = \sum_{\tau=0}^{h} \beta_\tau x_{t-\tau}$, $t = 1, ..., T$ where x_t is near x_{t-1} and $x_{t-1} + (x_{t-1} - x_{t-2})$

Advantages of Panel Data

4. Controlling the impact of omitted variables (or individual or time heterogeneity)
5. Generating more accurate predictions for individual outcomes by pooling data of individual outcomes rather that predicting the outcomes using certain variables
6. Providing micro-foundations for aggregate data analysis. Panel data containing time series observations for a number of individuals are ideal for investigating the "homogeneity" versus "heterogeneity" issue. Panel data overcome the problem of aggregation bias of the macro assumption of "representative agent"
7. Simplify computation and statistical inference resulting from:
   - non-stationary time series
   - measurement errors
   - truncated or censored data

Main issues in utilizing panel data

- Heterogeneity bias: Ignoring heterogeneity (in slope and/or constant) could lead to inconsistent or meaningless estimates of interesting parameters.
  - When important factors peculiar to a given individual are left out, the typical assumption that an economic variable is generated by a parametric probability distribution function may not be a realistic one
  - Ignoring the individual or time-specific effects that exist among cross-sectional or time-series units but are not captured by the included explanatory variables can lead to parameter heterogeneity in the model specification
- Dynamic panel data models

Dealing with unobserved heterogeneity

- The focus of panel data analysis is how to control the impact of unobserved heterogeneity to obtain valid inference on the common parameters (β).
- Linear regression framework: unobserved heterogeneity is individual specific and time invariant
- The individual-specific effect on the outcome variable y_{it} could either be invariant with the explanatory variables x_{it} or interact with x_{it}.
- Realistic Model: $y_{it} = \alpha_i^* + \beta_i x_{it} + u_{it}$, $i = 1, ..., N$, $t = 1, ..., T$ where u_{it} is the error term, uncorrelated with x and $u_{it}, i.i.d.(0, \sigma_u^2)$
  - α_i^* and β_i are different for different cross-sectional units (although time-invariant)
- OLS provide misleading estimates: Model: $y_{it} = \alpha^* + \beta x_{it} + u_{it}$, $i = 1, ..., N$, $t = 1, ..., T$
  - Heterogeneous intercepts, homogeneous slope
  - Heterogeneous intercepts and slopes

Econometric Model: Static Linear

$$y_{it} = \alpha + \beta x_{it} + \gamma z_i + (c_i + u_{it})$$

- y_{it}: dependent variable
- x_{it}: K-dimensional row vector of time-varying explanatory variables
- z_i: M-dimensional row vector of time-invariant explanatory variables excluding the constant
- α intercept
- β: K-dimensional column vector of parameters
- γ: M-dimensional column vector of parameters
- c_i: individual-specific effect
- u_{it}: idiosyncratic error term

Econometric Model: Static Linear

T observations for individual i

$$y_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{it} \\ \vdots \\ y_{iT} \end{bmatrix}_{T\times 1} \quad X_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{it} \\ \vdots \\ x_{iT} \end{bmatrix}_{T\times K} \quad Z_i = \begin{bmatrix} z_{i1} \\ \vdots \\ z_{it} \\ \vdots \\ z_{iT} \end{bmatrix}_{T\times M} \quad u_i = \begin{bmatrix} u_{i1} \\ \vdots \\ u_{it} \\ \vdots \\ u_{iT} \end{bmatrix}_{T\times 1}$$

NT observations for all individuals

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_N \end{bmatrix}_{NT\times 1} \quad X = \begin{bmatrix} X_1 \\ \vdots \\ X_i \\ \vdots \\ X_N \end{bmatrix}_{NT\times K} \quad Z = \begin{bmatrix} Z_1 \\ \vdots \\ Z_i \\ \vdots \\ Z_N \end{bmatrix}_{NT\times M} \quad u = \begin{bmatrix} u_1 \\ \vdots \\ u_i \\ \vdots \\ u_N \end{bmatrix}_{NT\times 1}$$

Econometric Model: Static Linear

Simple case: no time-invariant explanatory variables, i.e., $z_i = 0$

$$y_{it} = \alpha + \beta x_{it} + (c_i + u_{it})$$

- Assumptions
  - Cross-sectional independence: Observations on (y_i, x_i) are independent over $i =, ..., N$
  - Slope parameter homogeneity: The parameters in β are common to all $i =, ..., N$
- The form of unobserved heterogeneity relates to the individual-specific intercept term c_i which links y_{it} to x_{it}. This treatment is known as "fixed effects" of "random effects", depending on whether they are assumed to be correlated or uncorrelated with the explanatory variables in x_{it}.

Ordinary Least Squares

$$y_{it} = \alpha + \beta x_{it} + (c_i + u_{it})$$
$$y_{it} = \alpha + \beta x_{it} + \epsilon_{it}$$
$$\epsilon_{it} = c_i + u_{it}$$

- Using OLS we get $\hat{\beta}_{OLS} = (X'X)^{-1}X'y$
- Assumption: $E[x_{it}u_{it}] = 0$
- Properties of $\hat{\beta}_{OLS}$ depend on $E[x_{it}c_i]$

Omitted Variables Problem in Cross-Sectional Data

- What to do?
  - find a proxy
  - find a valid IV correlated with elements of x, elements correlated with c
- omitted variables problem-Panel Data
  - under relatively strong assumptions we can use Random Effects (RE) models
  - eliminating c using Fixed Effects (FE) methods

Assumption: Strict Exogeneity

- To estimate RE or FE we assume strict exogeneity

$$E(y_{it}|x_{it}, c_i) = \beta x_{it} + c_i$$

  - once x_{it} and c_i are controlled for, x_{is} has no partial effect on y_{it} for $s \neq t$.

- Thus, in our model $y_{it} = \alpha + \beta x_{it} + (c_i + u_{it})$ the strict exogeneity assumption in terms of idiosyncratic errors is: $E[u_{it}|x_{it}, c_i] = 0$, $t = 1, ..., T$.
- This implies that explanatory variables in each time period are uncorrelated with the idiosyncratic error in each time period ($E[x_{is}u_{it} = 0]$, $s, t = 1, ..., T$ vs. $E[x_{it}u_{it} = 0]$, $t = 1, ..., T$)

Random Effects-RE

- Random effects models effectively put c_i in the error term under the assumption that c_i is orthogonal to x_{it} and then accounts for the serial correlation in the composite error
- Thus in RE we impose two assumptions (a) strict exogeneity and (b) orthogonality between c_i and x_{it}:
  1. $E[u_{it}|x_i, c_i = 0],\ t = 1, ..., T$
  2. $E[c_i|x_i] = E[c_i] = 0]$
- the assumption $E[c_i] = 0]$ is included as an intercept in x_{it}
- OLS consistent but not efficient (use Generalized Least Squares-GLS)

Random Effects (using GLS)

- RE accounts for the serial correlation in $(c_i + u_{it})$ and thus $\epsilon_{it} = c_i + u_{it}$
- Now the model reads as $y_{it} = \alpha + \beta x_{it} + \epsilon_{it}$
- The RE assumptions imply $E[\epsilon_{it}|x_i] = 0]$, $t = 1, ..., T$
- Applying GLS methods now account for the error structure in ϵ_{it}
- If the RE assumptions are satisfied it is consistent and efficient
- More flexible Feasible GLS models can be estimated that allow for heteroscedasticity and autocorrelation (e.g., $\epsilon_{it} = \rho\epsilon_{it-1} + \omega_{it}$)

Fixed Effects

- Random effects assumes that c_i orthogonal to x_{it}: Very strong assumption
- But in reality there are many arbitrary correlations between c_i and x_{it}:usefulness of Panel Data
- FE explicitly deals with the fact that c_i and x_{it} may be correlated
- FE assumes strict exogeneity $E[u_{it}|x_i, c_i = 0]$ BUT it does not assume orthogonality $E[c_i|x_i] = E[c_i]$
- Thus in FE we impose weaker assumptions than in RE
- Cost of using FE (and not RE): we cannot include time-invariant variables in x_{it}

How to deal with unobserved heterogeneity in FE (ways to eliminate c_i)

- 3 Ways
- Aim: Eliminate c_i because it causes the error term to be correlated with the regressors
  1. Within estimates (FE transformation)
  2. Estimating c_i with dummies
  3. First differencing

Within Estimator-FE

- Basic equation

$$y_{it} = \alpha + \beta x_{it} + c_i + u_{it} \tag{1}$$

- Averaging the basic equation (over T)

$$\bar{y}_i = \alpha + \beta \bar{x}_i + c_i + \bar{u}_i \tag{2}$$

- Substract equation (2) from equation (1):

$$y_{it} - \bar{y}_i = \beta(x_{it} - \bar{x}_i) + (u_{it} - \bar{u}_i) \tag{3}$$

$$\tilde{y}_i = \beta \tilde{x}_i + \tilde{u}_{it} \tag{4}$$

- Estimate 4 using pooled OLS in order to get estimates for β. Standard errors should be adjusted - Stata this do automatically)

Dummy Variables Estimator-LSDV

- Estimate c_i using a set of dummies for every i in the sample (especially in cases with small N, $i = 1, ..., N$)
- Include N dummies (one fo reach i) in the regression

$$y_{it} = \alpha + \beta x_{it} + c_i + u_{it} \tag{5}$$

- Estimate using OLS (usual standard errors)
- Computationally intensive method when N is large

First Difference Estimator-FD

- Assume strict exogeneity. Basic equation

$$y_{it} = \alpha + \beta x_{it} + c_i + u_{it} \tag{6}$$

- Use 1-period lagged values in the basic model

$$y_{it-1} = \alpha + \beta x_{it-1} + c_i + u_{it-1} \tag{7}$$

- Subtract lagged model from basic model

$$y_{it} - y_{it-1} = (\beta x_{it} - \beta x_{it-1}) + (u_{it} - u_{it-1}) \tag{8}$$

$$\Delta y_{it} = \beta \Delta x_{it} + \Delta u_{it} \tag{9}$$

- First differencing eliminates c_i
- Smaller sample: $T - 1$ periods -instead of T- for each i
- Estimate using pooled OLS (typical standard errors)

Specification test (Hausman, 1978)

$$H_0 : Cov(u_{it}, x_{it} = 0) \text{ (RE)}$$
$$H_1 : Cov(u_{it}, x_{it} \neq 0) \text{ (FE)}$$

- Fixed effects estimator is consistent under H0 and H1
- Random effects estimator is efficient under H0, but inconsistent under H1
- Hausman Test Statistic

$$H = (\hat{\beta}_{RE} - \hat{\beta}_{FE})'[Var(\hat{\beta}_{RE}) - Var(\hat{\beta}_{FE})]^{-1}(\hat{\beta}_{RE} - \hat{\beta}_{FE}) \sim \chi^2$$

- Stata do this automatically

Stata implementation of Static Panel Data Estimators

```
clear all
webuse nlswork.dta, clear
use nlswork.dta
*Define individuals (variable idcode) and time periods (variable year)
xtset idcode year
*Fixed effects estimator (xtereg with the option fe)
generate ttl_exp2 = ttl_exp^2
xtreg ln_wage ttl_exp ttl_exp2, fe
*Cluster-robust Huber/White standard errors are reported with the vce option
xtreg ln_wage ttl_exp ttl_exp2, fe vce(cluster idcode)
*Automatically estimate robust standard errors with clustering
xtreg ln_wage ttl_exp ttl_exp2, fe vce(robust)
*Random effcts estimator (xtereg with the option re)
xtreg ln_wage grade ttl_exp ttl_exp2, re vce(robust)
*The Hausman test is calculated by
xtreg ln_wage grade ttl_exp ttl_exp2, re
estimates store b_re
xtreg ln_wage ttl_exp ttl_exp2, fe
estimates store b_fe
hausman b_fe b_re, sigmamore
*Pooled OLS estimator with corrected standard errors
egen ttl_exp_mean = mean(ttl_exp), by(idcode)
egen ttl_exp2_mean = mean(ttl_exp2), by(idcode)
reg ln_wage grade ttl_exp ttl_exp2, vce(cluster idcode)
regress ln_wage grade ttl_exp ttl_exp2 ///
ttl_exp_mean ttl_exp2_mean, vce(cluster idcode)
*Least squares dummy variables estimator
drop if idcode > 50
xi: regress ln_wage ttl_exp ttl_exp2 i.idcode
*get suppressed results of LSDV
areg ln_wage ttl_exp ttl_exp2, absorb(idcode)
```