Panel data: models and estimators

B.D. in Business Administration and Economics
Course in Quantitative Methods III

Rosario Barone
University of Rome "Tor Vergata"

rosario.barone@uniroma2.it

Panel data

Panel data (also longitudinal data or repeated measures) are repeated observations on the same cross section observed for several time periods.

Notation: let suppose to observe N individuals T times. We indicate the observation of the individual i at time t as $y_{i,t}$ with $i = 1, \ldots, N$ and $t = 1, \ldots, T$.

Short panel: large cross section of individuals observed for a few time periods, rather than a long panel such as a small cross section of countries observed for many time periods.

More precision in estimates

First advantage: increased precision in estimation.

- this is the result of an increase in the number of observations ;
- for a valid inference there is the needing to control for likely correlation of regression model errors over time for a given individual;
- the usual OLS estimators in a pooled OLS regression typically underestimated standard errors and t-statistics that can be greatly inflated.

Control for unobserved effects

Università di Roma Tor Vergata

Second advantage: possibility of control for the consistent unobserved individual-specific effects:

- Fixed effects model: allows for unobserved individual heterogeneity data from a short panel, if heterogeneity is assumed to be additive and time-invariant.
- Random effects model: treat any unobserved individual heterogeneity as a realization of a random variable distributed independently of the regressors.

Informative about individual dynamics

Università di Roma Tor Vergata

Third advantage: possibility of learning more about the dynamics of individual behavior than is possible from a single cross section.

Example: across section may yield a poverty rate of 20% but we need panel data to determine whether the same 20% are in poverty each year.

Overview

Università di Roma Tor Vergata

- Panel data provide information on individual behavior in two-dimensions: across time and across individuals.
- Standard panel data analysis uses a much wider range of models and estimators than is the case with cross-section data.
- Obtaining correct standard errors of estimators is also more complicated than in the cross-section case.
- One needs to control for correlation over time in errors for a given individual, in addition to possible heteroskedasticity.

- **Panel Data Models**

- Panel Data Estimators

Panel Data Models

- Pooled Model
- Individual and Time Dummies
- Fixed Effects
- Random Effects Models

Pooled Model

The most restrictive model is a pooled model that specifies constant coefficients, the usual assumption for cross-section analysis, so that

$$y_{it} = \alpha + x_{it}'\beta + u_{it}$$

If this model is correctly specified and regressors are uncorrelated with the error then it can be consistently estimated using pooled OLS.

Pooled Model

- The error term is likely to be correlated over time for a given individual: usual reported standard errors should not be used as they can be greatly downward biased.

- The pooled OLS estimator is inconsistent if the fixed effects model, defined in the following, is appropriate.

Individual and Time Dummies

A simple variant of the model permits intercepts to vary across individuals and over time while slope parameters do not.

$$y_{it} = \alpha_i + \gamma_t + x'_{it}\beta + u_{it},$$

which may be rewritten as

$$y_{it} = \sum_{j=1}^{N} \alpha_j d_{j,it} + \sum_{s=2}^{T} \gamma_s d_{s,it} + x'_{it}\beta + u_{it},$$

where:

- $d_{j,it} = 1$ if $i = j$ and 0 otherwise;
- $d_{s,it} = 1$ if $s = t$ and 0 otherwise.

Individual and Time Dummies

This model has $N + (T - 1) + \dim[x]$ parameters that can be consistently estimated if both $N \to \infty$ and $T \to \infty$.

In short panel $N \to \infty$ but T does not:

- no problem in estimating consistently γ_s
- the estimation of the individual intercepts α_j is challenging: one possibility is to instead have dummies for groups of observations.

Individual-specific effects model

The individual-specific effects model allows each cross-sectional unit to have a different intercept term though all slopes are the same, so that

$$y_{it} = \alpha_i + x'_{it}\beta + \epsilon_{it}$$

where ϵ_{it} is $i.i.d$ over i and t and α_i are random variables that capture unobserved heterogeneity.

Strong exogeneity or strict exogeneity assumption:

$$E[\epsilon_{it}|\alpha_i, x_{i1}, \ldots, x_{iT}] = 0, \quad t = 1, \ldots, T,$$

so that the error term is assumed to have mean zero conditional on past, current, and future values of the regressors.

Fixed Effects

Fixed Effects models treats α_i as an unobserved random variable that is **potentially correlated** with the observed regressors. The model speification is:

$$y_{it} = \alpha_i + x'_{it}\beta + \epsilon_{it}$$

and it is called fixed effects (FE) model as early treatments modeled these effects as parameters $\alpha_1, \ldots, \alpha_N$ to be estimated.

Fixed Effects

If fixed effects are present and correlated with x_{it} then many estimators such as pooled OLS are inconsistent.

Instead, alternative estimation methods that eliminate the α_i are needed to ensure consistent estimation of β in a short panel.

Random Effects Models

Random Effects models treats the unobservable individual effectsα_i as an unobserved random variable that are distributed **independently** of the regressors. The model speification is:

$$y_{it} = \alpha_i + x'_{it}\beta + \epsilon_{it}$$

and it usually makes the additional assumptions that:

$$\alpha_i \sim [\alpha, \sigma^2_\alpha] \text{ and } \epsilon_{it} \sim [0, \sigma^2_\epsilon]$$

- both the random effects and the error term are assumed to be $i.i.d.$;
- no specific distributions is specified.

Fixed vs Random Effect

The terms fixed and random effect are potentially misleading .

To avoid such confusion, M-J Lee (2002) calls a fixed effect a **related effect** and a random effect an **unrelated effect**.

It must be clear that α_i is a **random variable in both fixed and random effects models**.

Equicorrelated Model

The Random Effects model can be viewed as a special case of the pooled model: Let consider the model

$$y_{it} = \alpha_i + x'_{it}\beta + \epsilon_{it};$$

If α_i be subsumed into the error term, such that $u_i t = \alpha_i + \epsilon_{it}$, we get

$$y_{it} = x'_{it}\beta + u_{it};$$

This model id called *Equicorrelated model* because

$$cov[(\alpha_i + \epsilon_{it})(\alpha_i + \epsilon_{it})] = \begin{cases} \sigma^2_\alpha & t \neq s \\ \sigma^2_\alpha + \sigma^2_\epsilon & t = s \end{cases},$$

i.e. $cor(u_i, u_s) = \sigma^2_\alpha/(\sigma^2_\alpha + \sigma^2_\epsilon)$ for $i \neq s$ does not vary in time.

- Panel Data Models

- **Panel Data Estimators**

Panel Data Estimators

- Pooled OLS estimator
- Between Estimator
- Within Estimator or Fixed Effects Estimator
- First-Differences Estimator
- Random Effects Estimator

Panel Data Estimators

We now introduce several commonly used panel data estimators of β which differ in the extent to which cross-section and time-series variation in the data are used.

Their properties vary according to whether or not the fixed effects model is the appropriate model.

A regressor x_i t may be either time-invariant. For some estimators (within and first differences estimators) only the coefficients of time-varying regressors are identified.

Pooled OLS estimator

The pooled OLS estimator is obtained by stacking the data over i and t into one long regression with $N \times T$ observations, and estimating by OLS.

$$y_{it} = \alpha + x'_{it}\beta + u_{it}$$

If $cov(x_{it}, u_{it}) = 0$, then either $N \to \infty$ or $T \to \infty$ is sufficient for consistency. The pooled OLS estimator uses variation over both time and cross-sectional to estimate β.

Understatement of standard errors

For a given individual we expect considerable correlation in y over time, so that $cor(y_{it}, y_{is})$ is high.

Even after inclusion of regressors $cor(u_{it}, u_{is})$ may remain nonzero, and it often can still be quite high: the usual OLS variance matrix based on $i.i.d$ errors is not appropriate.

The usual OLS output treats each of the T years as independent pieces of information, but the information content is less than this given the positive error correlation. This leads to **overstatement of estimator precision**.

Inconsistency of pooled OLS

The pooled OLS estimator is inconsistent if the true model is the fixed effects model. Let consider:

$$y_{it} = \alpha + x'_{it}\beta + (\alpha_i - \alpha + \epsilon_{it})$$

by Fixed Effects model definition, α_i is correlated with x_{it}: therefore, the error term is correlated with the regressors.

Between Estimator

The between estimator in short panels instead uses just the cross-sectional variation. Begin with the individual-specific effects model. Let consider

$$y_{it} = \alpha_i + x'_{it}\beta + \epsilon_{it};$$

by avereging over all years we get

$$\bar{y}_i = \alpha_i + \bar{x}'_i\beta + \bar{\epsilon}_i,$$

which can be rewritten as the **between model**

$$\bar{y}_i = \alpha + \bar{x}'_i\beta + (\alpha_i - \alpha + \bar{\epsilon}_i).$$

It exloits variation between different individuals for estimating β (it is a cross section regression).

Consistency

The Between Estimator is consistent in case of uncorrelation of the error term $(\alpha_i - \alpha + \bar{\epsilon}_i)$ and the regressor $\bar{x}_i$. Therefore:

- Under Fixed Effect model condition $\rightarrow$ inconsistent;
- Under Random Effect model $\rightarrow$ consistent.

Within Estimator

The within estimator is an estimator that exploits the special features of panel data.

In a short panel it measures:

- the association between individual-specific deviations of regressors from their time averaged values
- individual-specific deviations of the dependent variable from its time-averaged value.

This is done by using the variation in the data over time.

Within Estimator

Università di Roma Tor Vergata

By considering the individual specific effect model,

$$y_{it} = \alpha_i + x'_{it}\beta + \epsilon_{it};$$

we take the the average over time:

$$\bar{y}_i = \alpha_i + \bar{x}'_i\beta + \bar{\epsilon}_i.$$

Then, we define the **within model** as $y_{it} - \bar{y}_i$, that is:

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)'\beta + (\epsilon_{it} - \bar{\epsilon}_i).$$

The within estimator is then the OLS estimator in the within model, and it yields consistent estimates of β in the fixed effects model, whereas the pooled OLS and between estimators do not.

Notes on the Within Estimator

- Also called Fixed Effect estimator (...efficient estimator for β in the fixed effect models);
- By this representation, individual effect α_i ignored and treated as nuisance parameters (we focus on the consistent estimation of β);
- If instead the fixed effects α_i are of interest they can also be estimated (in short panels estimates are inconsistent).

No time-invariant regressors allowed

Major limitation: the coefficients of time-invariant regressors are not identified in the within model (if $x_{it} = x_i$, then $(x_i - \bar{x}_i) = 0$).

Often we would need to estimate the effect of time-invariant regressors (e.g. gender, ...).

For this reason, sometimes Pooled OLS or random effects estimators are preferred, since they permit estimation of coefficients of time-invariant regressors, but...these estimators are inconsistent if the fixed effects model is the correct model.

First-Differences Estimator

The first-differences estimator also exploits the special features of panel data.

In a short panel it measures:

- the association between individual-specific one-period changes in regressors;
- individual-specific one-period changes in the dependent variable;

First-Differences Estimator

Let consider the individual specific effect model,

$$y_{it} = \alpha_i + x'_{it}\beta + \epsilon_{it};$$

and *lagging* of one period, we get

$$y_{i,t-1} = \alpha_i + x'_{i,t-1}\beta + \epsilon_{i,t-1}.$$

Therefore, we define the first-differences model as:

$$y_{it} - y_{i,t-1} = (x_{it} - x_{i,t-1})'\beta + (\epsilon_{it} - \epsilon_{i,t-1}).$$

The first-differences estimator is the OLS estimator in the first-differences model.

Notes on the First-Differences Estimator

- this estimator yields consistent estimates of β in the fixed effects model;
- the coefficients of time-invariant regressors are not identified;
- individual effect are ignored;
- less efficient than within estimator for $T > 2$ if ϵ_{it} is $i.i.d.$

Random Effects Estimator

Università di Roma Tor Vergata

Also the random effects estimator exploits the special features of panel data. Again, let consider the individual-specific effects model rewritten as

$$y_{it} = \mu + \alpha_i + x'_{it}\beta + \epsilon_{it};$$

and assume a random effects model where:

$$\alpha_i \sim [\alpha, \sigma^2_\alpha] \text{ and } \epsilon_{it} \sim [0, \sigma^2_\epsilon].$$

Then:

- pooled OLS is consistent;
- pooled GLS consistent and more efficient.

GLS estimator and Feasible GLS*

Generalized least squares (GLS) is a technique for estimating the unknown parameters in a linear regression model

$$y = X'\beta + u$$

when there residuals $u \sim [0, \Omega]$ are not independent and heteroscedastic. Let the variance error matrix as

$$\Omega \neq \sigma^2 \times \mathbb{I};$$

GLS estimator and Feasible GLS*

If Ω is *known* and *nonsingular*, we can premultiply the model by $\Omega^{-1/2}$

$$\Omega^{-1/2}y = \Omega^{-1/2}X'\beta + \Omega^{-1/2}u.$$

The errors in this transformed model are zero mean, uncorrelated and homoscedastic.

β is estimated via OLS regression of $\Omega^{-1/2}y$ on $\Omega^{-1/2}X$, obtaining the **generalized least-squares estimator**

$$\beta_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y.$$

GLS estimator and Feasible GLS*

In practice, Ω is unknown.

By defining $\Omega = \Omega(\gamma)$, where γ is a finite dimensional parameter vector, we obtain a consinstent estimate $\hat{\gamma}$ of γ, and then

$$\hat{\Omega} = \Omega(\hat{\gamma}).$$

The **feasible generalized least-squares** (FGLS) estimator is

$$\beta_{GLS} = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}y.$$

Random Effect Estimator

The feasible GLS estimator of the RE model (random effects estimator), can be calculated from OLS estimation of the transformed model

$$y_{it} - \hat{\lambda}\bar{y}_i = (1 - \hat{\lambda})\mu + (x_{it} - \hat{\lambda}\bar{x}_i)'\beta + \nu_{it}$$

where $\nu_{it} = (1 - \hat{\lambda})\alpha_i + (\epsilon_{it} - \hat{\lambda}\bar{\epsilon}_i)$ is asymptotically $i.i.d.$ and $\hat{\lambda}$ is consistent for

$$\lambda = 1 - \frac{\sigma_\epsilon}{\sqrt{\sigma_\epsilon^2 + T\sigma_\alpha^2}}.$$

Random Effect Estimator

Note that:

- $\hat{\lambda} = 0$ corresponds to pooled OLS;
- $\hat{\lambda} = 1$ corresponds to within estimator;
- $\hat{\lambda} \to 1$ as $T \to \infty$;
- the RE estimator is fully efficient under the RE model;
- It is inconsistent if the fixed effects model is the correct model.

Hausman Test

If individual effects are fixed:

- the within estimator $\hat{\beta}_W$ is consistent;
- the random effect estimator $\tilde{\beta}_{RE}$ is inconsistent.

Here β refers to the vector of coefficients of just the time-varying regressors.

One can therefore test whether fixed effects are present by using a Hausman test:

$$\begin{cases} H_0: & cor(\alpha_i, x_i t) = 0 \\ H_1: & cor(\alpha_i, x_i t) \neq 0 \end{cases}$$

Hausman Test

- A large value of the Hausman test statistic leads to rejection of the null hypothesis that the individual-specific effects are uncorrelated with regressors and to the conclusion that fixed effects are present;
- if regressors are correlated with individual-specific effects caused by omitted variables, then one can add further regressor and again perform a Hausman test in this larger model to see whether fixed effects are still necessary;
- even if such correlation persists it may be possible to estimate a random effects model using instrumental variables methods.

Panel-Robust Statistical Inference

The various panel models include error terms denoted with u_{it}, ϵ_i and α_i. The errors are potentially:

1. serially correlated (i.e., correlated over t for given i);
2. heteroskedastic.

Valid statistical inference requires controlling for both of these factors.

Panel Robust Standard Errors

- control only for heteroskedasticity:
  - White heteroskedastic consistent estimator extension for Panel
- controls for both serial correlation and heteroskedasticity:
  - Panel-Robust Sandwich Standard Errors
  - Panel Bootstrap Standard Errors