Introduction to Panel Data Econometrics

With applications in R

Giovanni Millo[1]

[1]DEAMS, University of Trieste

October 9, 2025

Outline of the talk

1. Introduction: why panel data
2. Motivating examples
3. The panel model as a specification problem
4. Estimation methods
5. Diagnostic testing

Structure of the lessons

Start from motivating examples, then address how to identify the Data Generating Process (DGP), then how to estimate parameters optimally

- why panel data
- how to model the double dimension
- how to (optimally) estimate each specification
- post-estimation diagnostics
- robust testing with clustered data
- introduction to the **R** software
- how to do all this in **R**

Outline of the talk

1. Introduction: why panel data
2. Motivating examples
3. The panel model as a specification problem
4. Estimation methods
5. Diagnostic testing

Outline: why panel data

Panel data have two dimensions: one is usually time, the other can be individuals, firms, countries etc.
In this section we will introduce the broad subject of panel data econometrics through its features and advantages over pure cross-sectional or time-series methods. According to Baltagi (2005), panel data allow to:

- control for individual heterogeneity
- exploit greater variability for more efficient estimation
- study adjustment dynamics
- identify effects one could not detect from cross-section data
- improve measurement accuracy (micro data instead of aggregated)
- use one dimension to infer about the other (as in panel time series)

Unobserved heterogeneity vs. error components

According to Arellano (2003), panel data techniques address two broad families of issues:

- unobserved heterogeneity
- error components

The former is typically dealt with through *fixed effects* techniques, aiming at controlling for unobserved variables possibly biasing estimation.

The latter is the domain of *random effects* methods, and of generalized least squares (GLS) in general; its main goal is to discriminate between various components of the error term, varying along the different directions of cross section and/or time, increasing the estimator's efficiency.

Outline of the talk

1. Introduction: why panel data
2. Motivating examples
3. The panel model as a specification problem
4. Estimation methods
5. Diagnostic testing

Outline: motivating examples

In this section we will provide some examples of interesting research questions from various disciplines that can take advantage from panel data

- is public capital productive?
- how to stop people from drinking and driving
- how to optimize rice crops in Indonesia
- does the purchasing power parity theory hold in the real world?
- are strong unions a curse or a blessing for economic growth?

Public capital productivity

Munnell (1990) specifies a Cobb-Douglas production function that relates the gross social product (gsp) of a given US state to the input of public capital (pcap), private capital (pc), labor (emp) and state unemployment rate (unemp) added to capture business cycle effects:

$$log(gsp) = \beta_0 + \beta_1 log(pcap) + \beta_2 log(pc) + \beta_3 log(emp) + \beta_4 unemp$$

- 48 US States allow for a relatively large cross-sectional sample
- 17 years give a fairly long timespan for variation in public capital to take effect
- neither dimension would have been sufficient for reliable inference if taken in isolation

Beer taxes and road fatalities

The Fatality dataset from Stock and Watson, *Introduction to Econometrics*, is a good example of the importance of individual heterogeneity and time effects in a spatially referenced setting.
The research question is whether taxing alcoholics can reduce the road death toll. The basic specification relates the road fatality rate to the (real) beer tax in a classical regression setting:

$$mrall_i = \alpha + \beta beertax_i + \epsilon_i$$

Data are 1982 to 1988 for each of the continental US States.

- cross-sectional models yield nonsense coefficients
- panel data allow to control for state heterogeneity reconciling economic theory and evidence

Rice farming

In the rice farming study by Druska and Horrace (2004), 171 rice farms in Indonesia are observed over six growing seasons, three wet and three dry, between 1975 and 1983. The farms are located in six different villages of the Chimanuk River basin in West Java. The production frontier equation relates rice output to the following inputs: seed, urea, phosphate (tsp), labour hours (lab) and land (size).
Dummy variables account for the use of high yield varieties of seed (high), or for a mix of seed varieties (mixed) and for the use of pesticides.
Dummy variables are also added for the six villages and for the season being a wet one.

- the problem is to control for idiosyncratic characteristics of soil and position, and to consider different seasons
- focusing on many small farms which can reasonably be seen as randomly drawn from a bigger population, provides a good example of the usefulness of random effects methods.

Purchasing Power Parity (PPP)

Coakley, Fuertes and Smith (2006) present a purchasing power parity (PPP) regression on quarterly data 1973Q1 to 1998Q4 for 17 developed countries, so that $N = 17$ and $T = 104$ which is fairly typical of a "long" panel.
The estimated model is

$$\Delta s_{it} = \alpha + \beta(\Delta p - \Delta p^*)_{it} + \nu_{it}$$

where s_{it} is the relative exchange rate against USD and $(\Delta p - \Delta p^*)_{it}$ is the inflation differential between the country and the US.

- the hypothesis of interest is $\beta = 1$
- the "main" dimension is the time series; pooling time series for different countries allows to characterize the overall behaviour of exchange rates, averaging out individual idiosyncracies

Economic performance and labour organization

Alvarez, Garrett and Lange (1991) estimate a model where economic performance in a panel of 16 countries over 15 years is related to political and labour organization variables. They use the FGLS estimator, finding out that economic performance is enhanced where strong unions coexist with an important presence of leftist movements in government or in the opposite situation (rightist governments with weak unions), being less satisfactory for in-between cases.
Beck et al. (1993) re-examine the data using OLS estimation of a dynamic model with time fixed effects and time-clustered errors, upholding previous conclusions as regards the effects on growth but rendering mixed evidence for inflation and unemployment.

- specification issues are often crucial
- importance of model specification diagnostics

...and more

Examples from many disciplines abound

- land use and forestry
- coal strata and the industrial revolution
- income and democracy
- the resource curse
- vaccine effectiveness
- social networks and the Arab Spring
- ...

Recap: long, short, wide, ... panels

From the above axamples we gather that panel data can take much different shapes according to the behaviour of n vs. T.

- *large* (*wide*) panels have many individuals observed over few time periods (years, bienniel "waves" of surveys); asymptotics resemble that of cross-sectional data
- *long* panels have a limited number of individual units (countries, firms) observed over many time periods (years, quarters; for financial data, even weeks or days); methods tend towards those of time series
- *macro* panels, a.k.a. *pooled time series*, have emerged recently in macroeconomics and political science where a moderate number of countries (e.g., OECD) are observed over a moderate number of years

As a general rule of thumb, one must use methods whose asymptotic properties exploit the *large* dimension

Practicals 1: panel data in **R**

- Panel data manipulation and model estimation in **R**
  - data conventions
  - formulae
  - estimators
  - object orientation

Outline of the talk

1. Introduction: why panel data
2. Motivating examples
3. The panel model as a specification problem
4. Estimation methods
5. Diagnostic testing

Outline: panel model specification

The specification problem:

- find a DGP that *might have generated* the data
- according to Occam's Razor, simpler is better

In this section we will address the issue of how to model panel data in a general-to-specific framework

- start from a general (maybe unfeasible, or at least inefficient) formulation that be statistically admissible
- drill down to the most parsimonious and efficient admissible specification by means of diagnostic testing (Hendry's famous advice: *test, test, test!*)

The panel model, general to specific

The basic linear panel models used in econometrics can be described through suitable restrictions of the following general model:

$$y_{it} = \alpha_{it} + \beta_{it}^{\top} x_{it} + u_{it} \tag{1}$$

where $i = 1, \ldots n$ is the individual (group, country, ...) index, $t = 1, \ldots, T$ is the time index and u_{it} a random disturbance term of mean 0.

(Full) Parameter homogeneity

Of course the latter is not estimable with $N = n \times T$ data points. A number of assumptions are usually made about the parameters, the errors and the exogeneity of the regressors, giving rise to a taxonomy of feasible models for panel data.
The most common one is parameter homogeneity, which means that $\alpha_{it} = \alpha$ for all i, t and $\beta_{it} = \beta$ for all i, t. The resulting model

$$y_{it} = \alpha + \beta^{\top} x_{it} + u_{it} \tag{2}$$

is a standard linear model pooling all the data across i and t.

Homogeneous β, heterogeneous intercepts

To model individual heterogeneity, one often assumes that the error term has two separate components, one of which is specific to the individual and does not change over time. This is called the unobserved effects model:

$$y_{it} = \alpha + \beta^\top x_{it} + \mu_i + \epsilon_{it} \tag{3}$$

For the sake of exposition we are considering only the individual effects case here. There may also be time effects, which is a symmetric case, or both of them, so that the error has three components: $u_{it} = \mu_i + \lambda_t + \epsilon_{it}$.

Correlated or uncorrelated heterogeneity

The appropriate estimation method for this model depends on the properties of the two error components.

- The idiosyncratic error ϵ_{it} is usually assumed well-behaved and independent from both the regressors x_{it} and the individual error component μ_i.
- The individual component may be in turn either independent from the regressors or correlated.

Independence of x_{it} and μ_i is a crucial assumption, because estimation consistency requires $E(u|X) = 0$, and $u = \mu + \epsilon$

Correlated heterogeneity: FE

If it were correlated, the ordinary least squares (OLS) estimator for β would be inconsistent, so it is customary to treat the μ_i as a further set of n parameters to be estimated, as if in the general model $\alpha_{it} = \alpha_i$ for all t.

$$y_{it} = \alpha_i + \beta_{it}^{\top} x_{it} + u_{it} \tag{4}$$

This is called the fixed effects (also known as *within* or *least squares dummy variables*) ~~model~~ "situation", and is usually estimated by OLS on transformed data, which gives consistent estimates for β.

Estimates of α are *T-consistent*!

Uncorrelated heterogeneity: RE

If the individual-specific component μ_i is uncorrelated with the regressors, a situation which is usually termed random effects, the overall error $u_{it} = \mu_i + \epsilon_{it}$ also is, so the OLS estimator is consistent.

Nevertheless, the common error component over individuals induces correlation across the composite error terms, making OLS estimation inefficient, so that for efficiency one has to resort to some form of feasible generalized least squares (GLS) estimators. This is based on the estimation of the variance of the two error components, for which there are a number of different procedures available.

No heterogeneity: OLS

If the individual component is missing altogether, $u_{it} = \epsilon_{it}$ is i.i.d. so that pooled OLS is the most efficient estimator for β.

This set of assumptions is usually labelled *pooling* model, although this actually refers to the errors' properties and the appropriate estimation method rather than the model itself.

First differences

Another way of estimating unobserved effects models through removing time-invariant individual components is by first-differencing the data: lagging the model and subtracting, the time-invariant components (the intercept and the individual error component) are eliminated, and the model

$$\Delta y_{it} = \beta^{\top} \Delta x_{it} + \Delta u_{it} \tag{5}$$

(where $\Delta y_{it} = y_{it} - y_{i,t-1}$, $\Delta x_{it} = x_{it} - x_{i,t-1}$ and, from (3), $\Delta u_{it} = u_{it} - u_{i,t-1} = \Delta \epsilon_{it}$ for $t = 2, ..., T$) can be consistently estimated by pooled OLS. This is called the *first-difference*, or FD estimator. Its relative efficiency, and so reasons for choosing it against other consistent alternatives, depends on the properties of the error term. The FD estimator is usually preferred if the errors u_{it} are strongly persistent in time, because then the Δu_{it} will tend to be serially uncorrelated.

Averaged data: BE

Lastly, the *between* model, which is computed on time (group) averages of the data,

$$y_{i.} = \alpha + \beta^{\top} x_{i.} + u_{i.}$$

discards all the information due to intragroup variability but is consistent in some settings (e.g., non-stationarity) where the others are not, and is often preferred to estimate long-run relationships.

The between estimator does *not* allow to control for unobserved heterogeneity!

Practicals 2: individual heterogeneity (beer tax data)

- Cross-sectional estimation of the beer tax data
- Between estimation
- Unbiased Fixed Effects estimation

Outline of the talk

1. Introduction: why panel data
2. Motivating examples
3. The panel model as a specification problem
4. Estimation methods
5. Diagnostic testing

Estimation methods

Each specification can be consistently estimated by an appropriate method

- Ordinary Least Squares (pooled, FE, FD, BE)
- Generalized Least Squares (RE, GGLS)
- Maximum Likelihood (RE, GLS, RE+AR(1), spatial...)

RE turns out to be estimable by OLS on transformed data as well.

OLS on transformed data: FE

The estimation methods for the basic models in panel data econometrics, the pooled OLS, random effects and fixed effects (or within), and first difference models, can all be described inside the OLS estimation framework.
In fact, while pooled OLS simply pools data, the standard way of estimating fixed effects models with, say, group (time) effects entails transforming the data by subtracting the average over time (group) to every variable, which is usually termed *time-demeaning* and is defined as:

$$y_{it} - \bar{y}_i = (X_{it} - \bar{X}_i)\beta + (u_{it} - \bar{u}_i) \tag{6}$$

where $\bar{y}$ and $\bar{X}$ denote time means of y and X

OLS on transformed data: FD

In turn, FD methods rely on transforming the data by subtracting the previous observation in time to every variable, which is termed *first-differencing* and is defined as:

$$y_{it} - y_{i,t-1} = (X_{it} - X_{i,t-1})\beta + (u_{it} - u_{i,t-1}) \tag{7}$$

so that, denoting $\Delta y_{it} = y_{it} - y_{i,t-1}$, the model becomes

$$\Delta y_{it} = \Delta X_{it}\beta + \Delta u_{it} \tag{8}$$

Notica that in both cases (FE and FD) any time-invariant heterogeneity is eiminated altogether, as $\bar{\alpha}_i = \Delta\alpha_i = \alpha_i - \alpha_i = 0$

OLS on transformed data: RE

In the random effects case, GLS estimation is equivalent to OLS on *partially demeaned* data, where partial demeaning is defined as:

$$y_{it} - \theta \bar{y}_i = (X_{it} - \theta \bar{X}_i)\beta + (u_{it} - \theta \bar{u}_i) \tag{9}$$

where $\theta = 1 - [\sigma_u^2/(\sigma_u^2 + T\sigma_e^2)]^{1/2}$, $\bar{y}$ and $\bar{X}$ denote time means of y and X, and the disturbance $v_{it} - \theta \bar{v}_i$ is homoskedastic and serially uncorrelated.

Thus the feasible RE estimate for β may be obtained estimating $\hat{\theta}$ and running an OLS regression on the transformed data.

The other estimators can be computed as special cases:

- for $\theta = 1$ one gets the fixed effects estimator
- for $\theta = 0$ the pooled OLS one.

The RE estimator as Generalized Least Squares

Despite the similarities in computing β_{RE} and β_{FE}, the RE estimator is conceptually much different.

- OLS (and to some extent FE, FD, ...) are justified by the Gauss-Markov theorem: least squares are the BLUE estimator for i.i.d. errors
- GLS are based on the Aitken theorem: they are optimal for a given error covariance matrix

The Gauss-Markov th. is a particular case of the Aitken th. with $V = \sigma^2 I$
The GLS estimator (for known $cov(u) = V$) is

$$\hat{\beta}_{GLS} = (X'V^{-1}X)^{-1}X'V^{-1}y; \qquad cov(\hat{\beta_{GLS}}) = (X'V^{-1}X)^{-1}$$

(remember: $\hat{\beta}_{OLS} = (X'X)^{-1}X'y$ and $cov(\hat{\beta_{OLS}}) = \sigma^2_\epsilon(X'X)^{-1}$)

Meet the errors' covariance matrix - 1

It is necessary, at this point, to consider the structure of the covariance matrix of (composite) errors $cov(u) = E(uu^\top) = V$

The random effects model is characterized by the assumption that the individual specific effect c_i is random and uncorrelated both with the explanatory variables and with the idiosyncratic error. Under this specification and under homoskedasticity in both c_i and u_{it} and no serial correlation in u_{it}, the variance-covariance matrix of the errors becomes

$$V = {\sigma_c}^2(I_N \otimes J_T) + {\sigma_u}^2(I_N \otimes I_T) \tag{10}$$

i.e., the errors' covariance is block-diagonal with

$$V = I_N \otimes \Omega \tag{11}$$

Meet the errors' covariance matrix - 2

(continued) where

$$\Omega = \begin{bmatrix} \sigma_u^2 + \sigma_c^2 & \sigma_c^2 & \dots & \sigma_c^2 \\ \sigma_c^2 & \sigma_u^2 + \sigma_c^2 & \dots & \vdots \\ \dots & & \ddots & \sigma_c^2 \\ \sigma_c^2 & & & \sigma_u^2 + \sigma_c^2 \end{bmatrix} \tag{12}$$

By Aitken's theorem (see, e.g., Greene 2003, 10.5), generalized least squares (GLS) are the efficient estimator for the model parameters if Ω is known. The estimator is then

$$\hat{\beta}_{GLS} = (X'V^{-1}X)^{-1}(X'V^{-1}y) \tag{13}$$

Feasible GLS for the RE model

The *feasible* GLS estimator is based on estimating $\hat{V}$. If the estimator for V is consistent, then the FGLS estimator for β_{FGLS} is consistent and efficient.
Substitute a consistent estimate $\hat{V}$ for V:

$$\hat{\beta}_{FGLS} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}y; \qquad cov(\hat{\beta}_{FGLS}) = (X'\hat{V}^{-1}X)^{-1}$$

Various methods have been devised for estimating the two variance components σ_μ and σ_e. The best quadratic unbiased estimators for the total error variance σ_u and the idiosyncratic error variance σ_e are (Baltagi 2005): $\sigma_u = T\sum_i \bar{u}_{i.}^2/n$ and $\sigma_e = \sum_i \sum_t (u_{it} - \bar{u}_{i.})^2/(n(T-1))$
but u_{it} is unobserved.

Estimating the variance components of RE

Residuals of various types can be used:

- Wallace and Hussein use the residuals from the pooled model $\hat{u}_{OLS}$ to estimate both σ_e^2 and $\sigma_u^2 = T \sum_t \bar{u}_{i.}^2 / n$; hence $\sigma_\mu^2 = (\sigma_u^2 - \sigma_e^2)/T$
- Amemiya uses the LSDV residuals as above
- Swamy and Arora estimate σ_u^2 from the between residuals $\hat{u}_{BE}$ and σ_e^2 from the within residuals $\hat{u}_{FE}$
- Nerlove estimates σ_μ^2 on the basis of the dummy coefficients in the LSDV regression, and σ_e^2 from the within residuals; then $\sigma_u^2 = T\sigma_\mu^2 + \sigma_e^2$
- Wooldridge uses pooled residuals $\hat{u}_{OLS}$ and estimates total error variance as $\sigma_u^2 = \sum_i \sum_t u^2/(nT - K)$ and σ_μ^2 as average of nondiagonal error products.

(Two-step) Maximum Likelihood

Maximum Likelihood (ML) is not needed for the simple OLS and, to some extent, for the RE case

- If the errors are i.i.d, the OLS estimator is also the ML estimator.
- If the errors are RE, then the GLS estimator using the true V matrix is also the ML estimator
- (Iterative) ML can be more efficient
- ML can allow for more complicated error structures
  - serial correlation
  - spatial correlation
  - nested effects
  - ...

Outline of the talk

1. Introduction: why panel data
2. Motivating examples
3. The panel model as a specification problem
4. Estimation methods
5. Diagnostic testing

Diagnostic testing

The hypotheses on parameters and error terms (and hence the choice of the most appropriate estimator) are usually tested by means of:

- *pooling* tests to check poolability, i.e., the hypothesis that the same intercept applies across all individuals (OLS vs. FE)
- if the homogeneity assumption over the coefficients is established, the next step is to establish the presence of unobserved effects (OLS vs. RE)
- the choice between fixed and random effects specifications is based on Hausman-type tests, comparing the two estimators under the null of no significant difference: if this is not rejected, the more efficient random effects estimator is chosen (RE vs. FE)
- even after this step, departures of the error structure from sphericity can further affect inference, so that either screening tests or robust diagnostics are needed.

Testing for correlated individual heterogeneity

The Hausman test is a test for $H_0 : E(\mu_i|X) = 0$; it is based on the comparison between a consistent estimator (here, FE) which is

- consistent under H_0
- consistent under H_A

and another one (here: RE) which is:

- consistent *and efficient* under H_0
- inconsistent under H_A

The Hausman test is based on the fact that
$Var(\beta_{FE} - \beta_{RE}) = Var(\beta_{FE}) - Var(\beta_{RE})$ so that

$$(\beta_{FE} - \beta_{RE})'[Var(\beta_{FE}) - Var(\beta_{RE})]^{-1}(\beta_{FE} - \beta_{RE}) \sim \chi^2_k$$