Panel data

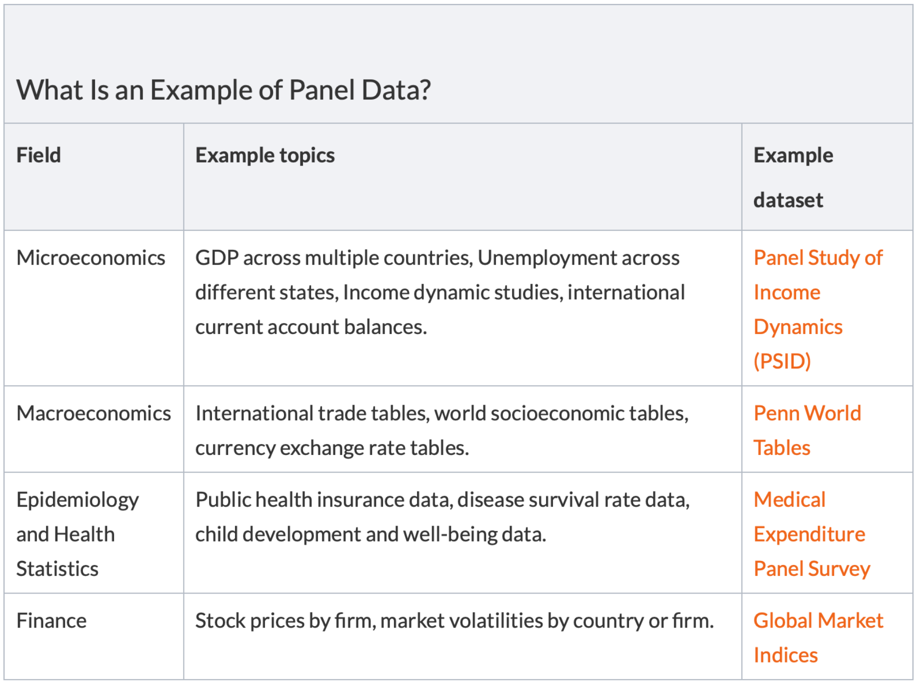
Federico Nutarelli

Panel data, sometimes referred to as longitudinal data, is data that contains observations about different cross sections across time. Those observations can be either regularly observed over time (balanced panel) or irregularly observed (unbalanced) either randomly or for specific reasons.

There are a number of advantages of panel data:

- Panel data can model both the common and individual behaviors of groups.
- Panel data contains more information, more variability, and more efficiency than pure time series data or cross-sectional data.
- Panel data can detect and measure statistical effects that pure time series or cross-sectional data can't (avoid Omitted Variable bias).
- Panel data can minimize estimation biases that may arise from aggregating groups into a single time series.

Here you can find several examples of panel data:



What Is an Example of Panel Data?

Field	Example topics	Example dataset
Microeconomics	GDP across multiple countries, Unemployment across different states, Income dynamic studies, international current account balances.	Panel Study of Income Dynamics (PSID)
Macroeconomics	International trade tables, world socioeconomic tables, currency exchange rate tables.	Penn World Tables
Epidemiology and Health Statistics	Public health insurance data, disease survival rate data, child development and well-being data.	Medical Expenditure Panel Survey
Finance	Stock prices by firm, market volatilities by country or firm.	Global Market Indices

1 Panel data heterogeneity

Panel data series modeling centers around addressing the likely **dependence across data observations within the same group** (sounds familiar? Think about the clustering example of individuals divided into classes. Here is the same. The only difference is that ”the groups” are not classes but the same individual repeated in time). In fact, the primary difference between panel data models and time series models, is that panel data models allow for heterogeneity across groups and introduce individual-specific effects.

To provide an example, let’s say that units are 3 countries (USA, UK and Italy) and Y_{it} is the GDP. If a recession happens worldwide, then all 3 countries’ GDPs are likely to be affected. If, instead, elections happen in UK, then only the UK GDP is affected over time. This is an individual specific effect which can be taken into account using panel techniques!

To summarize: panel data are useful in that they take into account the within group effects (individual specific effects) and allow for across groups’ heterogeneity (e.g. elections in UK and not in USA and Italy).

2 Models for panel data

Panel data methods can be split into two broad categories:

- Homogeneous (or pooled) panel data models assume that the model parameters are common across individuals.
- Heterogeneous models allow for any or all of the model parameters to vary across individuals. Fixed effects and random effects models are both examples of heterogeneous panel data models.

Within these groups, the assumptions made about the variation of the model across individuals are the primary drivers for which model to use.
To make an example, let's take the general model

$$y_{it} = \alpha + \beta x_{it} + \varepsilon_{it}$$

The latter representation is an homogeneous model since α, β are the same across groups and time. Moreover **Any differences across groups enter the model only through the error term ε_{it}.**
Alternatively, we could believe that groups share common coefficients on regressors but there are group-specific intercepts, as is captured in the fixed effects or least squares dummy variable (LSDV) model:

$$y_{it} = \alpha_i + \beta x_{it} + \varepsilon_{it}$$

The latter model is heterogeneous because the constants, α_i, are group-specific.

2.1 Popular panel data models

The 4 most popular panel data models are:

- Pooled OLS (covered here)
- One-way Fixed effects (covered by Millo I guess, in case let me know.)
- One way random effects (covered by Millo)
- Random effects (covered by Millo)

Throughout we will assume that the dgp looks as follows:

$$y_{it} = \beta x_{it} + \delta z_i + \varepsilon_{it}$$

In this model, X represents the observed characteristics such as age, firm size, expenditures, and Z represents **unobserved** characteristics, such as management quality, growth opportunities, or skill.

2.2 Pooled OLS

In some cases, there are no unobservable individual-specific effects, and is constant across individuals. This is a **strong assumption** (notice in fact that in the dgp there are!) and implies that all the observations within groups are

independent of one another.
In these cases, the (assumed) model becomes

$$y_{it} = \beta x_{it} + \alpha + \varepsilon_{it}$$

This implies that when there is no dependence within individual groups, the panel data can be treated as one large, pooled dataset (so eaach observation is "as if" it is a different individual. The same individuaal observed in two different time periods, t_1 and t_2 are considered as 2 different individuals in this setting!). The model parameters, β, and, α , can be directly estimated using pooled ordinary least squares.

Linear independence within the groups of a panel is unlikely and pooled OLS is rarely acceptable for panel data models.