Econometrics **April 16, 2024**

Topic 6: DID and TWFE

by Sai Zhang

Key points: This note is on the causal panel data, building upon Arkhangelsky and Imbens (2023).

Disclaimer: *This note is compiled by Sai Zhang.*

6.1 Panel Data Configurations

6.1.1 Data Types

6.1.1.1 Panel Data

For observations on N units, indexed by $i = 1, \cdots, N$, over T periods, indexed by $t = 1, \cdots, T$, the outcome of interest is denoted by Y_{it}, the treatment W_{it}. These observations may themselves consist of averages over more basic units:

$$\mathbf{Y} = \begin{pmatrix} Y_{11} & \cdots & Y_{1T} \\ \vdots & \ddots & \vdots \\ Y_{N1} & \cdots & Y_{NT} \end{pmatrix} \qquad \mathbf{W} = \begin{pmatrix} W_{11} & \cdots & W_{1T} \\ \vdots & \ddots & \vdots \\ W_{N1} & \cdots & W_{NT} \end{pmatrix}$$

we may also observe exogenous variables X_{it} or X_i. Typically, we focus on a balanced panel where for all units $i = 1, \cdots, N$ we observe outcomes for all $t = 1, \cdots, T$.

6.1.1.2 Grouped Repeated Cross-Section Data

In a GRCS data, we have observations on N units, each observed only once in period T_i for unit i. Different units may be observed at diffrent points in time, T_i typically takes on only a few values, with many units sharing the same value for T_i. The outcome Y_i and treatment W_i are indexed by the unit index i. The set of units is **partitioned** into 2 or more groups, with the group that unit i belongs to denoted by $G_i \in \mathcal{G} = \{1, 2, \cdots, G\}$.

Define the average outcomes for each group-time-period pair:

$$\overline{Y}_{gt} \equiv \frac{\sum_{i=1}^{N} \mathbf{1}_{G_i=g, T_i=t} Y_i}{\sum_{i=1}^{N} \mathbf{1}_{G_i=g, T_i=t}}$$

for treatment

$$\overline{W}_{gt} \equiv \frac{\sum_{i=1}^{N} \mathbf{1}_{G_i=g, T_i=t} W_i}{\sum_{i=1}^{N} \mathbf{1}_{G_i=g, T_i=t}}$$

then treat the $G \times T$ group averages $\overline{Y}_{gt}$ and $\overline{W}_{gt}$ as the unit of observation, then the grouped data is just a panel. The major issue in practice is that the number of groups is very small comparing to proper panel data.

6.1.1.3 Row and Column Exchangeable Data

The data are doubly indexed by $i = 1, \cdots, N$ and $j = 1, \cdots, J$, with outcomes Y_{ij}. They are different from panel data in that there is **no time ordering** for the second index. Many methods developed for panel data are also applicable here.

6.1.2 Shapes of Data Frames

Panel data can also be loosely classified by the shape:

- **Thin Frames** ($N \gg T$), where the number of cross-section units is large relative to the number of time periods:
  - unit-specific parameters (individual FEs) can not be estimated consistently due to the short time series
  - REs might be more suitable since they place a stocahstic structure on the individual components
- **Fat Frames** ($N \ll T$), where the number of cross-section units is large relative to the number of time periods.
- **Square** $N \simeq T$, where the number of units and time periods is comparable.

6.1.3 Assignment Mechanisms

6.1.3.1 The General Case

In the most general case, the treatment may vary both across units and over time, with units **switching** in and out of the treatment group:

References

Dmitry Arkhangelsky and Guido Imbens. Causal models for longitudinal and panel data: A survey. Technical report, National Bureau of Economic Research, 2023.