Panel Data Analysis: A Simplified Summary

Stephen Zamore

Hauge School of Management

NLA University College, Kristiansand

stephen.zamore@nla.no

December 2022

Abstract

The purpose of this article is to provide a summary of linear panel data analysis in simple language without mathematical expression. The summary serves as a quick reference for many researchers who want to utilize panel data techniques in their research. The article starts by explaining what a panel dataset is and how it differs from time series and cross-section datasets. Next, it explains the main linear panel data models which are then summarized graphically. Panel data techniques solve two main problems: (1) omitted variable bias (individual-fixed effects or unobserved heterogeneity) and (2) endogeneity bias. As shown in Figure 1, dynamic (GMM) models and some instrumental variable (IV) models can solve these two problems.

Keywords: Panel Data; Fixed Effects, Random Effects, Endogeneity, GMM

JEL codes: C00, C10, C23, C26

1. Panel data

Panel data (or longitudinal data) concern a collection of individuals, households, firms, countries, among others observed over time where the time dimension is usually shorter than the number of units (i.e., individuals, firms, etc.). For instance, if we collect income and other information from 200 people in a country for the period 2019–2022, we get a total of 800 (200 x 4 years) individual-year observations assuming no missing observations in the data set. Whenever the number of years is greater than the number of units, the data set is called time series or long panel. Assuming we collect the data from the 200 people only for 2019, the data set is called cross-section and the total observations will be 200.

Hsiao (2014) lists advantages of panel data, over cross-section or time series, including control for individual heterogeneity, analysis of dynamic adjustments, more accurate estimates due to availability of more informative data, more degree of freedom and less multicollinearity, among others. For details, read Hsiao (2014) and Baltagi (2021).

2. Panel data models

Figure 1 below summarizes the main linear panel data models[1]. Generally, panel data techniques are utilized to address two main problems: (1) individual-fixed effects (unobserved individual heterogeneity, hereafter 'heterogeneity') and (2) endogeneity bias. Figure 1 shows which problem (s) each panel data model solves. Panel models are grouped into two: static and dynamic models. Next, static models are discussed.

2.1 Static panel data models

Static panel data models assume that previous values of the dependent variable or the independent variables have no influence on current values of these variables. Thus, issues such as simultaneity or reversed causality are not taken into consideration. Some static models (IV-models in Figure 1) can solve endogeneity problem using standard instrumental variable (IV) regression technique and others (non-IV models in Figure 1) just solve only heterogeneity problem. The latter models are presented next.

2.1.1 Non-IV models

These models are employed when there is no endogeneity bias to address in the research.

Fixed effects model: The fixed effects (FE) model assumes that heterogeneity is correlated with the independent variables and must be removed. Thus, the FE model is designed to solve this heterogeneity problem by mean-differencing transformation (Cameron & Trivedi, 2022).

Random effects model: Random effects (RE) model assumes that heterogeneity is uncorrelated with the independent variables, hence, the original RE model does not solve heterogeneity problem. However, other versions of RE model do solve heterogeneity bias.

Mundlak correction model: Mundlak (1978) utilizes RE technique to solve heterogeneity problem by adding binary indicators (individual dummies) in the model.

[1] For details, read Baltagi (2021) and Cameron & Trived, (2022).

Extended Mundlak correction model: Wooldridge (2021) extends the Mundlak model by adding time-fixed effects (time dummies) in the model, making it a two-way error component model which yields estimates similar to two-way FE model. Next, IV models are discussed.

2.1.2 Instrumental variable (IV) models

IV models address endogeneity problem by assuming that external instruments exist and should be used to solve the problem. The instruments are considered external because they do not form part of the list of independent variables in the model specification.

Fixed effects model: solves heterogeneity and endogeneity problems.

Random effects model: regular RE model solves only endogeneity problem.

Hausman-Taylor model: Hausman and Taylor (1981) model assumes that: (1) heterogeneity bias is uncorrelated with some independent variables (i.e., exogenous variables), and (2) values of exogenous independent variables in other periods (e.g., t+1) than the current period can be used as instruments. Thus, Hausman-Taylor (HT) model utilizes RE and IV techniques to solve heterogeneity and endogeneity problems. Figure 2 summarizes HT model. Next, dynamic panel models are discussed.

2.2 Dynamic panel data models

Dynamic models assume that previous values of the dependent variable or the independent variables have influence on current values. Thus, issues such as reversed causality are considered using generalized method of moments (GMM). Dynamic models solve both heterogeneity and endogeneity problems.

2.2.1 Difference GMM

Arellano and Bond (1991) model accounts for endogeneity problem by subtracting last year's value of a variable from the current year's value. The new (differenced) equation is then used in the regression analysis where instruments are employed. However, the first-differencing transformation creates more gaps in the dataset, making the method not suitable for unbalanced panel because it leads to more data losses (Roodman, 2009).

2.2.2 System GMM

System GMM, based on Arellano and Bover (1995) and Blundell and Bond (1998), utilizes two sets of equations: (1) the original equation and (2) a transformed equation either via first-differencing or forward orthogonal deviations transformations (Roodman, 2009). By using two equations, more instruments are employed in System GMM, hence, it is more robust than Difference GMM.

Differencing transformation: First differencing is explained above; the differenced equation and the original equation are used in the System GMM regression.

Orthogonal forward deviation transformation. This approach subtracts the average of all future available observations of a variable from current year's value (Roodman, 2009). The new equation and the original equation are utilized in the System GMM regression. This approach minimizes data loss in unbalanced panel.

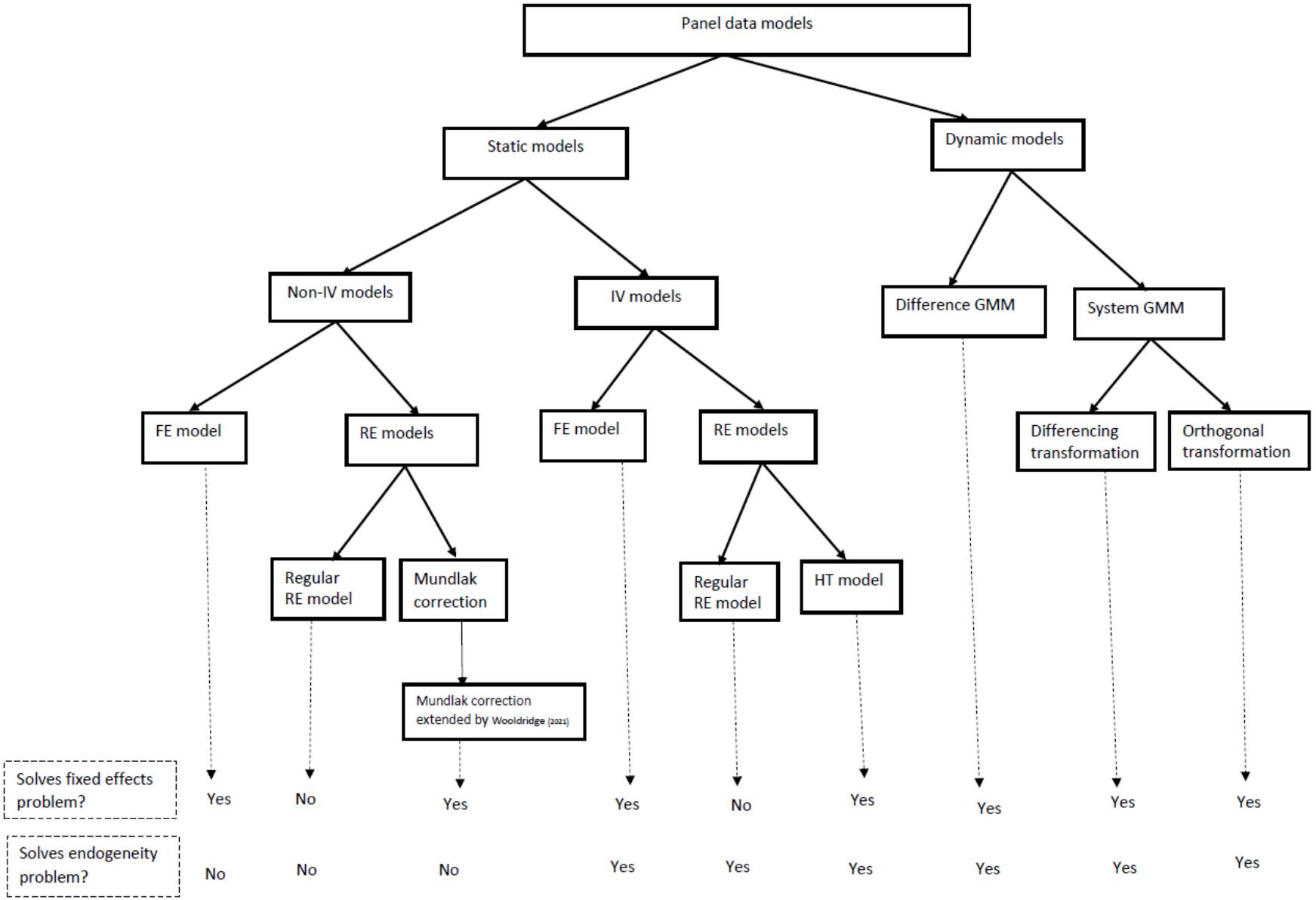

Figure 1: Panel data models

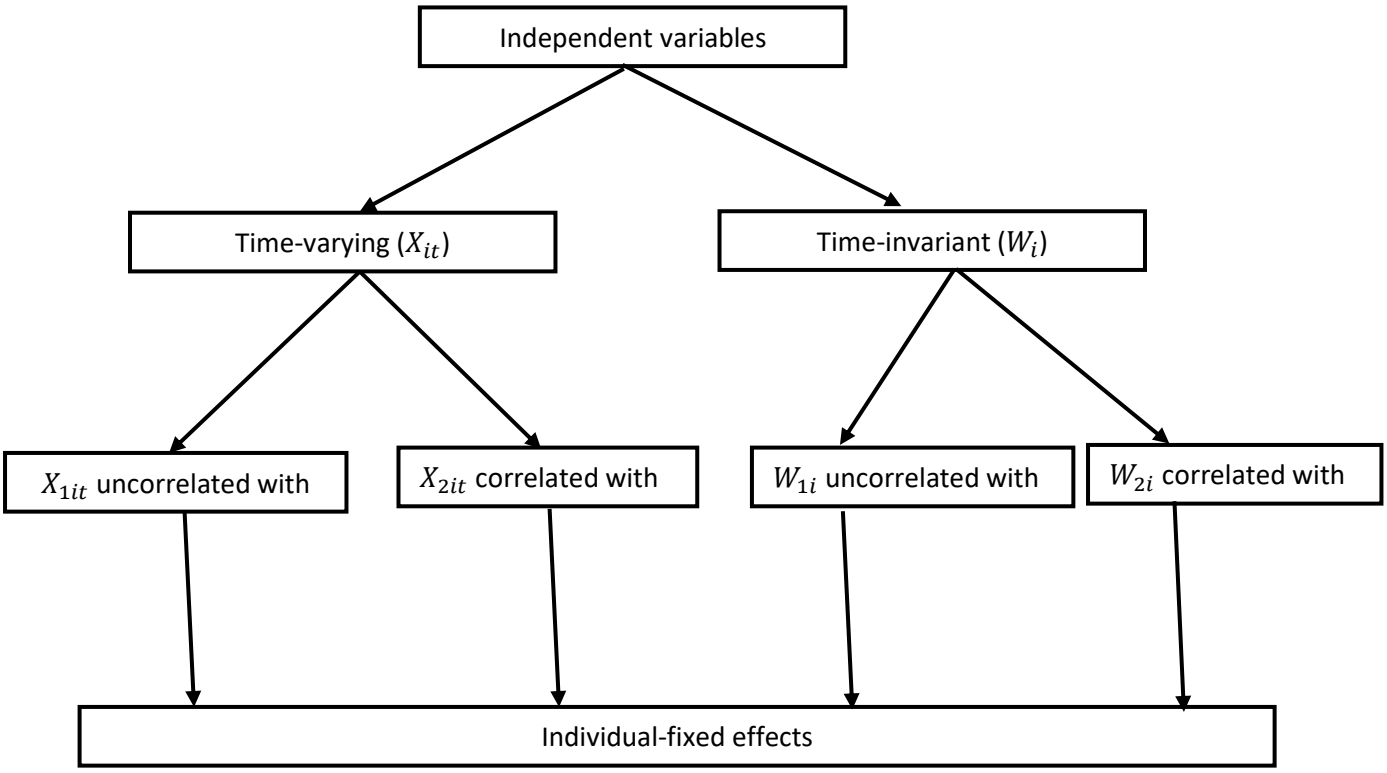

Figure 2: Hausman and Taylor model

References

Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies, 58*(2), 277-297.

Arellano, M., & Bover, O. (1995). Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics, 68*(1), 29-51.

Baltagi, B. H. (2021). *Econometric Analysis of Panel Data* (6th ed.). Cham: Springer.

Blundell, R., & Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics, 87*(1), 115-143.

Cameron, A. C., & Trivedi, P. K. (2022). *Microeconometrics using Stata* (2nd ed. Vol. I). College Station: Stata Press.

Hausman, J. A., & Taylor, W. E. (1981). Panel data and unobservable individual effects. *Econometrica, 49*(6), 1377-1398.

Hsiao, C. (2014). *Analysis of Panel Data* (3rd ed.). New York, NY: Cambridge University Press.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica, 46*(1), 69-85.

Roodman, D. (2009). How to do xtabond2: An introduction to difference and system GMM in Stata. *Stata Journal, 9*(1), 86-136.

Wooldridge, J. M. (2021). Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators. Working paper. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3906345