PANEL DATA ANALYSIS

A brief Introduction

Dr Love Odion Idahosa

OUTLINE

- Quick refresh – Estimators and CLRM Assumptions
- Panel data - Introduction
  - Types of panel data
  - Some Advantages of panel data
  - Some Limitations of panel data
- Basic panel data models for individual effects
  - One-way vs two way individual effects consideration
  - Pooled OLS model
  - Fixed Effects model & LSDV
  - Random Effects Model
- Other considerations
  - Some Factors to consider before choosing a panel data model
  - Some relevant tests to run (in consideration of the CLRM assumptions).

QUICK REFRESHER

FROM THE FAMILIAR – REFRESH OUR MEMORY

- Objective of statistics – typically to describe or make inference of a population
- Objective of econometrics – to make inferences of relationships in an unobserved population from the sample. **Causality is difficult to infer as typical datasets are not random.**
- In exploring the relationship between X and Y, the simple linear regression model is represented with the following equation:

$$y = \alpha + X'\beta + u$$ equation 1

  - Where y = the dependent/outcome variable
  - X' is a matrix of independent/predictor variables
  - β is the slope coefficient for each predictor variable.
  - u is the residual/error term; and α is the intercept/constant.

FROM THE FAMILIAR – REFRESH OUR MEMORY

- **What estimator is the best?** The one that best approximates the population parameters – i.e., is most representative.
- **ENTERS – the Classical Linear Regression Model (CLRM)** – Classical refers to the following assumptions:
    - 1. The regression model is linear, correctly specified, and has an additive error term.
    - 2. The error term has a zero population mean, meaning that the average error is zero at any specific value of the independent variable(s).
    - 3. All explanatory variables are uncorrelated with the error term – error is random
    - 4. Observations of the error term are uncorrelated with each other (**no serial correlation** - the error term doesn't exhibit a systematic relationship over time).
    - 5. The error term has a constant variance (**no heteroskedasticity**).
    - 6. No **perfect multicollinearity –** no explanatory variable is a perfect linear function of any other explanatory variables. That is, no independent variable can be expressed as a linear function of any other independent variables).

FROM THE FAMILIAR – REFRESH OUR MEMORY

- In summary the assumptions suggest that the linear model should produce residuals (error terms) that have a mean of zero [error on average should be zero], have a constant variance [errors are uniform/consistent across samples], and are not correlated with themselves or other variables [they do not influence each other or the other variables in the model so that the effect of each variable estimated is the pure effect].
- If these assumptions hold true, then the OLS is the Best Linear Unbiased Estimator [BLUE] (i.e., the OLS procedure creates the best possible estimates that are unbiased and efficient (have the smallest variance) – Gauss Markov Theorem.
- If these assumptions are violated, the OLS is no longer BLUE and a different model needs to be adopted.

PANEL DATA - INTRODUCTION

PANEL DATA - INTRODUCTION

- **What does panel data have to do with the CLRM Assumptions?**

- ***Cross Sectional data:*** *Observations for different entities* for the same period, stacked*

Entity	Period	$Variable_{observations}$			
		X1	X2	…	Xk
Entity A	1	$X1_a$	$X2_a$	…	XK_a
Entity B	1	$X1_b$	$X2_b$	…	XK_b
Entity C	1	$X1_c$	$X2_c$	…	$X3_c$
…	…	…	…	…	…
Entity N	1	$X1_n$	$X2_n$		XK_n

- ***Time series:*** *Observations for one entity* over different periods, stacked together.*

Entity	Period	$Variable_{observations}$			
		X1	X2	…	X_k
Entity A	1	$X1_1$	$X2_1$	…	XK_1
Entity A	2	$X1_2$	$X2_2$	…	XK_2
Entity A	3	$X1_3$	$X2_3$	…	XK_3
Entity A	…	…	…	…	…
Entity A	t	$X1_t$	$X2_t$		XK_t

**'Entities' will be used to refer to individuals, firms, countries, etc. And may be used interchangeably*

PANEL DATA - INTRODUCTION

- ***Panel:*** *Observations for different entities for multiple time periods, stacked together.*

Entity	Period	$Variable_{observations}$			
		X1	**X2**	**…**	**XK**
Entity A	1	$X1_{a,1}$	$X2_{a,1}$		$XK_{a,1}$
Entity A	2	$X1_{a,2}$	$X2_{a,2}$		$XK_{a,2}$
Entity A	3	$X1_{a,3}$	$X2_{a,3}$		$XK_{a,3}$
…	…	…	…		…
Entity A	t	$X1_{a,t}$	$X2_{a,t}$		$XK_{a,t}$
Entity B	1	$X1_{b,1}$	$X2_{b,1}$		$XK_{b,1}$
Entity B	2	$X1_{b,2}$	$X2_{b,2}$		$XK_{b,2}$
Entity B	3	$X1_{b,3}$	$X2_{b,3}$		$XK_{b,3}$
…	…	…	…		…
Entity B	t	$X1_{b,t}$	$X2_{b,t}$		$XK_{b,t}$
…					
Entity N	1	$X1_{n,1}$	$X2_{n,1}$		$XK_{n,1}$
Entity N	2	$X1_{n,2}$	$X2_{n,2}$		$XK_{n,2}$
Entity N	3	$X1_{n,3}$	$X2_{n,3}$		$XK_{n,3}$
…	…	…	…		…
Entity N	t	$X1_{n,t}$	$X2_{n,t}$		$XK_{n,t}$

PANEL DATA - INTRODUCTION

- **The panel regression equation:**

$$y_{it} = \alpha + X'_{it}\beta + u_{it} \qquad i = 1, \ldots., N; t = 1, \ldots T \qquad \text{equation 2}$$

$$u_{it} = \mu_i + v_{it}$$ - One-way error component model (mostly used)

$$u_{it} = \mu_i + \vartheta_t + v_{it}$$ - Two way error component model

Where μ_i is the unobserved individual specific effect; ϑ_t is the unobserved time effect; i denotes the entities (cross-sectional dimension); t denotes time (time-series dimension); and v_{it} is the residual error in the regression model.

Types of Panel Data

- Cross-section/micro & Time-series/macro panel
- Balanced and Un-balanced panel data

- Certain panel data models are only valid for balanced datasets.
- IMPORTANT to consider: are the missing data points random or non-random? – most standard panel models will yield correct coefficients if the missing values are random.
- To avoid dropping too many data points due to missing values, some unbalanced data models are available.

PANEL DATA - INTRODUCTION

Some advantages of panel data

Hsiao (2003) and Klevmarken (1989) list several benefits from using panel data. These include the following (see Baltagi, 2005):

1. Allows the flexibility of entities being heterogeneous and controls for this, as opposed to time-series and cross-section studies.
2. *'Panel data give more information, more variability', less collinearity* (as opposed to time-series) *among the variables, more degrees of freedom, and more efficiency'*, hence more reliable parameter estimates.
3. They are better able to evaluate dynamics of adjustment compared to cross-sectional data which seems relatively stable, but hide multitude of changes (e.g., unemployment spells, income mobility, etc.).
4. *'Panel data models allow us to construct and test more complicated behavioral models than purely cross-section or time-series data'*.

PANEL DATA - INTRODUCTION

Some limitations of panel data (see Baltagi, 2005)

1. '*Design and data collection problems* – collecting representative panel data is tedious and requires complex design.
2. '*Distortions of measurement errors* – the responses captured may be biased and not capture what is intended to be measured, resulting in inconsistencies.
3. *'Selection problems – self-selection, nonresponse, attrition'*
4. *'Possibility of cross-section dependence (e.g., cross-country dependence) in macro panels which must be accounted for in panel models.'*

BASIC PANEL DATA MODELS FOR INDIVIDUAL EFFECTS

BASIC PANEL DATA MODELS FOR INDIVIDUAL EFFECTS

POLS – POOLED ORDINARY LEAST SQUARE

- Pooled OLS assumes no panel effect in the data.
- Treats each observation as unique, bringing new information to the analysis.
- Problematic because observations within individuals are often correlated, and between individuals are correlated over time.

FIXED EFFECTS (FE) MODEL

- The individual effects are assumed to be fixed parameters, unique to each individual, to be estimated, an the remaining error term v_{it} is independent of/uncorrelated with each other, and with the regressors.

$$\text{i.e., the } \mu_i \text{ in } [u_{it} = \mu_i + v_{it}] \text{ is fixed}$$

- *Designed to study the causes of change within an entity.*
- Explores the relationship between the dependent and independent variable and ***assumes that something (which does not vary over time) within the individual*** may bias/impact the independent variable.

BASIC PANEL DATA MODELS FOR INDIVIDUAL EFFECTS

- **FIXED EFFECTS MODEL (contd.)**

- FE is appropriate when we assume that something within the individual entity may impact or bias the predictor or outcome variables and we need to control for this.
- FE removes these time-invariant characteristics so that the net/true effect of the independent variable can be evaluated (E.g., minimalist/maximalist behaviour in sustainability studies).
- Where fixed individual effects assumptions are violated, FE may lead to incorrect estimates – alternative model should be sought.
- The **Least Square Dummy Variable model (LSDV)** provides a good way to understand fixed effects.
- It includes a dummy for each entity in the simple OLS regression, allowing the pure effect of *X1* to be estimated (by controlling for the unobserved heterogeneity).
- Each dummy absorbs the effects particular to each entity. This becomes inefficient when N is large.
- It is also useful when carrying out a two-way error component FE model. '*Control for time effects whenever unexpected variation or special events my affect the outcome variable' (Torres-Reyna, 2007).*

BASIC PANEL DATA MODELS FOR INDIVIDUAL EFFECTS

RANDOM EFFECTS MODEL

- It assumes, unlike the fixed effects model, that the variation across entities (entity/individual effect) is random and uncorrelated with the independent variables included in the model; and is independent of the stochastic error term.

$$\text{i.e., the } \mu_i \text{ in } [u_{it} = \mu_i + v_{it}] \text{ is random}$$

- That is, individual effects are not related to/correlated with the independent variables being observed, or the error term, and can, hence, be independently investigated/evaluated as its own explanatory/independent variable; whereas they are absorbed in the intercept in the FE model.
- If one assumes that the individual effects might affect uniquely the dependent variable being evaluated, then random effects is best.
- It is, however, imperative to specify all those individual characteristics which might influence the explanatory/independent variables.
- As this is not always the case, one always runs the risk of omitted variable bias.

BASIC PANEL DATA MODELS FOR INDIVIDUAL EFFECTS

OTHER PANEL MODEL

- Given the machine learning focus of this class, I think it important to mention there are two other panel data models that might be applicable to the research you do:

1. The Random Coefficients Model (aka. Multilevel/Hierarchical Model)

- As opposed to the aforementioned panel models, the random coefficients model does not assume that the coefficients on regressors are the same across all individuals.
- It allows for different intercepts and slopes for different entity and/or time groups.
- Also allows for the relationship between entity and time to be included in the regression specification.

2. Dynamic Panel Data Models

- Panel datasets include a time series component.
- *In pure time series models we are able to model dynamics using lagged dependent variables which correct for autocorrelation between observations in the dataset at different points in time*.
- Because panel datasets include a time series component, it is also important to address the possibility of autocorrelation in panel data.
- The dynamic panel model adds dynamics (lagged components) to the entity effects framework.

OTHER CONSIDERATIONS

OTHER CONSIDERATIONS

SOME FACTORS TO CONSIDER BEFORE CHOOSING A PANEL DATA MODEL

1. Is it a cross-sectional panel or a time-series panel?
   - Is there serial correlation in the data – do effects in previous periods affect the current period?
2. Is the dataset balanced or unbalanced? If unbalanced, are the observation missing at random?
3. What is the nature of the dependent variable (continuous or limited – (e.g., dummy variable, categorical variable, truncated variable, etc.))?

OTHER CONSIDERATIONS

RELEVANT TESTS TO RUN

1. Hausman test – Fixed vs random effects
2. Test for time fixed effects – are the dummies for all years equal to 0?
3. Test for random effects – should a RE model be used or is OLS sufficient?
4. Test for heteroskedasticity – occurs often in micro panels
5. Test for serial/contemporaneous correlation/autocorrelation – peculiar to macro panels
6. Test for unit root/random walk with drift/non-stationarity – peculiar to a panels

OTHER CONSIDERATIONS

Concluding remarks:

- *'Panel data is not a panacea and will not solve all the problems that a time series or a cross-section study could not handle*.' – Baltagi (2005, page 8)
- As in most non-random data analysis, panel data will not always address issues of causality – in fact, it seldom does. As such, in interpreting the coefficients, care should be taken to not interpret slope coefficients as causal relationships.

References:

- Baltagi, B., 2005. *Econometric analysis of panel data*. John Wiley & Sons.
- Torres-Reyna, O., 2007. **Panel Data Analysis Fixed and Random Effects using Stata (v. 4.2).**
- Wooldridge, J.M., 2016. *Introductory econometrics: A modern approach*. Nelson Education.