IRAQI STATISTICIANS JOURNAL

https://isj.edu.iq/index.php/isj

ISSN: 3007-1658 (Online)

An Overview of Panel Data Analysis in Statistical and Econometric Research : A Review

Ban Ghanim Al-Ani [1,*], and Khalida Ahmed Mohammed [2]

[1,2] Department of Statistics and Informatics, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq

ARTICLE INFO

Article history:

Received	1 November 2025
Revised	1 November 2025,
Accepted	5 January 2026,
Available online	5 January 2026

Keywords:
PR Model
FE Model
RE Model
GMM Estimator
MG Estimator
PMG Estimator

ABSTRACT

Panel data combine features of both time series and cross-sectional data. Data on many economic or social units are observed in more than one time period. This framework enables the analysis of both within unit processes and between-unit heterogeneity, increases the efficiency of the estimations, and improves the fit of economic and statistical models. There are two essential types of panel data models: static panel models and dynamic panel models. Static models consider the dependent variable to be independent of its own past values, meaning that it does not contain any extrinsic dynamics. They are most commonly applied when there is no explicit time dependence to the phenomena of interest. These include the Pooled Regression Model (PRM), the Fixed Effect Model (FEM), and the Random Effect Model (REM). Dynamic models, on the other hand, consider that lagged values of the dependent variable exert some influence on the present values of the dependent variable, incorporating the lagged dependent variable as an explanatory variable. Since this variable is correlated with the endogenous error term, the estimation of such models must employ sophisticated techniques in order to prevent biased results, such as the (GMM): Arellano-Bond Difference (GMM), Arellano-Bover System (GMM) estimators or the Mean Group (MG) and Pooled Mean Group (PMG) estimators. Using static and dynamic panel models, the influences of several of the most commonly used indicators related to issues of poverty in Indonesia were examined between the years of (2012-2019), by looking at the impact of Human Development Index (HDI), Gini coefficient of income inequality, and the unemployment rate on poverty. As the later section of the paper shows, the dynamic model was able to better account for variations in poverty levels than the model that incorporated only contemporaneous data, supporting the hypothesis of a strong temporal form of endogeneity in the dynamics of poverty.

1. Introduction

Panel data are an important statistical tool in studies and research across various disciplines. They combine time series and cross-sectional data that allow for accurate analysis of changes in units over time for a more precise understanding of changes in units over time and enhancing accuracy in making estimates. Panel data often include simultaneous observations of multiple variables to broaden the scope of possible statistical and economic analysis. Common examples include countries (in international studies), organizations or firms (in management or financial studies), demographic groups (such as households or age groups), and individuals (in behavioral or social studies).

Panel data (also known as longitudinal data or pooled time-series data) are repeated measurements on entities (persons, families, companies, countries,…) over time. In other words, panel data have both a time series and a

* Corresponding author. drbanalani@uomosul.edu.iq
https://doi.org/10.62933/skqedv73

cross-sectional dimension, where (a relatively large number of) cross-section entities are observed during a (relatively short) period of time.

A cross-sectional data consists of observations recorded on a specific number of variables at a single point in time, while time series data consists of observations on one or more variables collected over several consecutive time periods. Panel data may have a very limited number of repeated measurements on the same variables and on the same sample or population. This is known in the design literature as a "one-shot" design [1]. Panel data can be represented geometrically as a three-dimensional structure for each variable, representing: the primary vertical dimension: time; the secondary vertical dimension: the cross-sectional units; and the horizontal dimension: the recorded values of the variables for each unit in each time period. This type of data has recently gained significant attention, especially in economic studies, because it takes into account both the effect of time change and the effect of the change in the difference between cross-sectional units [2].

To illustrate the concept and structure of panel data, we can use an applied example from the agricultural sector. We look at data covering multiple countries over multiple years. We begin by analyzing cross-sectional data representing a single point in time (such as a specific year), then move on to panel data by adding additional cross-sections representing subsequent years for the same countries - that is, we combine the time dimension with the cross-section dimension. Suppose our analytical goal is to explore the impact of farm labor, degree of farm mechanization, and fertilizer consumption (as the independent variables) on the grain yield (as the dependent variable). Table (1) presents a cross-sectional data set comprising three observations recorded on four variables. This data represents a cross-sectional snapshot of the status of the units under study in 2014 only, without any time dimension, reflecting the nature of cross-sectional data in capturing phenomena at a specific point in time.

Table 1. Cross-Sectional Data: Agricultural data for some Arab countries for the year 2014

Observation	Grain yield (tons)	Agricultural labor (1000 workers)	Agricultural mechanization (1000 hp)	Consumed fertilizer (tons)
Iraq	7093.85	1583.00	4533	194.939
Egypt	23713.40	6694.00	7662	2248.000
Syria	2694.21	1352.00	6801	39.577
Sudan	2871.00	3371.49	1913	195.590

Therefore, we cannot determine dynamic variations in sector agricultural performance over time. By contrast, the time-series data, shown in Table (2), provides variations in the observations across years. It describes the same four variables in Iraq for the three years between 2014 and 2016, so we can now track how the variables under study have changed over the past three years, providing a dynamic view that cross-sectional data alone do not provide. However, excluding other countries from the analysis and limiting the analysis to Iraq is a methodological limitation with significant analytical costs, as it limits our ability to generalize the results or isolate country specific effects. The idea of relying on just one country or even a very limited number of countries is statistically or methodologically unconvincing. Given this limitation, it becomes difficult to draw strong or statistically significant conclusions about the relationship between the three independent variables and grain productivity the central goal of our analysis.

To create a panel dataset, we combined the cross-sectional data recorded annually for each country, or in other words, applied the time series data for each country into a unified framework encompassing all cross-sectional units (countries). Table (3) shows the panel dataset, which was created by combining the information provided in Tables (1) and (2).

It is worth noting that each observation for four countries is repeated over three consecutive years (2014–2016), with values for four variables recorded in each year. This is more useful than time series or cross-sectional data alone.

For example, if a country's agricultural labor force changed during 2015, the panel data structure allows us to study how this change affected grain productivity in the following year or over the same period.

Table 2. Time Series Data: Agricultural data for Iraq during (2014-2016)

year	Grain yield (tons)	Agricultural labor (1000 workers)	Agricultural mechanization (1000 hp)	Consumed fertilizer (tons)
2014	7093.85	1583.00	4533	194.939
2015	3309.28	1623.200	4434	180.375
2016	3960.27	1664.420	4000	281.396

Table 3. Panel Data: Agricultural data for some Arab countries during (2014-2016) [3]

Country	year	Grain yield (1000 tons)	Agricultural labor (1000 workers)	Agricultural mechanization (1000 hp)	Consumed fertilizer (tons)
Iraq	2014	7093.85	1583.00	4533	194.939
	2015	3309.28	1623.200	4434	180.375
	2016	3960.27	1664.420	4000	281.396
Egypt	2014	23713.40	6694.00	7662	2248.000
	2015	23248.43	6403.00	7998	2248.000
	2016	23298.13	6486.26	8105	2248.000
Syria	2014	2694.21	1352.00	6801	39.577
	2015	4568.28	1297.63	6853	42.105
	2016	2761.27	1287.62	6875	24.931
Sudan	2014	2871.00	3371.49	1913	195.590
	2015	7841.00	3713.10	1935	110.030
	2016	4085.00	3825.49	1941	390.072

2. Material and Methods

2.1 Scope of using Panel Data

A. Economics fields:

The following are the most prominent scope of use of panel data, with real-life examples in each field, illustrating how this type of data is employed to analyze economic fields with greater accuracy.

1. Macroeconomics
- Main use: Analyzing the relationship between fiscal policy and economic growth.
- Case Study: For example, an analytical study could be conducted to examine the impact of inflation rate on real GDP growth in a sample of 10 developing countries, using panel data covering the period from 1980 to 2020.

2. Public Finance
- Primary use: Understanding oil revenue and development.
- OPEC countries as a case study to measure the effects of oil revenues on development, over a 15-year period.

3. Industrial Economics
- Primary use: Trends in sales by firm.
- A case study on the influence of raw-materials quality and quantity on the profits along a 25-year time span.

4. Financial Markets
- Primary application: Analysis of the influence of economic variables on stock prices or returns.
- A case study on the effect of interest rates and oil prices on stock returns of energy firms in the Gulf countries from 2000 to 2020.

5. Health Economics
- Key application: The influence of health policies on health outcomes.
- Example: Analysis of the effects of public health expenditure on infant mortality across a number of African countries.

6. Education Economics
- Primarily used for: analysis of education levels' effect on labor market.

- Example: The impact of education levels on employment within OCED countries over the past twenty years.

7. Development Studies
- Main use: Studying the relationship between foreign direct investment and economic growth.
- Case study: Impact of inflowing foreign direct investment on per capita GDP by sub-Saharan Africa from 1990 to 2020.

8. Political Economy & Governance
- The primary use is for the evaluation of the relationship between institutional indicators of development. evaluation of the relationship between institutional indicators of development.
- Case study: The impact of five institutional indicators on economic malaise in developing countries: government effectiveness, quality of regulations, rule of law, corruption control and voice & accountability.

9. Labor Economics
- Main use: Measuring the impact of minimum wages on unemployment.
- Case study Studying the relationship between the increases in minimum wage and unemployment rates in various US states over 25 years in various US states.

10. Urban & Regional Economics
- Main use: Analyzing the relationship between infrastructure and economic opening.
- Case study: Analyzing the impact of investment in transportation on imports and exports between some of Asian countries.

B. Non-Economics fields:

Some of the most prominent of these fields are:

1. Psychology
- Main use: A study of students' academic performance over time.
- Case study: Analyzing the impact of time spent browsing social media applications on high school students' academic performance over a five-year period.

2. Sociology
- Main use: Studying the impact of income inequality on education.
- Case study: Analyzing the impact of the Gini coefficient of income inequality on educational outcomes in cities of a developing country over a 10-year period.

3. Public Health
- Main use: Evaluating the impact of spending on health sector on population health.
- Case study: Tracking the impact of government spending on health sector on number of patients in US states.

4. Education
- Main use: Analyzing the impact of educational programs on student performance.
- Case study: Evaluating the impact of educational staff sizes on student performance in multiple schools over a 10-year period.

5. Environmental Science
- Main use: Studying the relationship between industrial units and pollution.
- Example: Analyzing the impact of increasing factory numbers on drinking water pollution in industrialized countries over a period of two decades.

6. Epidemiology
- Main use: Studying the relationship between climate change and the spread of some diseases and epidemics over time.
- Case study: Studying the impact of rising temperatures on Malaria infections in Sudanese states within 20 years.

7. Political Science
- Main use: Analyzing the relationship between the trust in government, democracy, or political participation across countries.
- Example: Evaluation impact of media freedom on voting rates in developing countries over 10 years.

8. Law & Criminology
- Main use: Evaluating the impact of the control of corruption as well as rule of law in relation to crime rates.
- Case study: Analyzing the impact of gun ownership and circulation laws on homicide rates in different cities over a 20-year period.

9. Agriculture & Natural Resources
- Main use: Tracking farm productivity or the impact of agricultural practices on sustainability.
- Case study: Studying the effect of organic fertilizer use on crop yields across multiple farms over five years.

10. Demographic Studies
- Main use: Studying the relationship between population growth on crude death rate.
- Case study: Analyzing the relationship between population growth and normal death cases in poor countries.

2.2 Advantages of Using Panel Data

1. Panel data contains more analytical information than cross-sectional or time series data alone.
2. Increasing sample size: leading to more degrees of freedom and then more accurate analysis.
3. Monitoring time dynamics: such as tracking the impact of policies over time.
4. It can measure and measure statistical impacts that cannot be easily detected in cross-sectional or time series data.
5. The collective behaviors of groups can be modeled through panel data.
6. Panel data can reduce biases in estimation that might arise from aggregating groups into a time series.
7. It enables a more in-depth dynamic analysis of economic relationships.
8. Controlling for heterogeneity: Unobserved characteristics of individual units can be controlled.
9. Improving statistical efficiency: by combining the time and cross-sectional dimensions.
10. Testing causality: The possibility of applying causality tests such as Granger causality.

2.3 Potential Challenges and Problems

- Dealing with missing data in unbalanced panels.
- The possibility of autocorrelation and heteroskedasticity.
- Determining the most appropriate model (fixed or random).
- The difficulty of analysis and programming in comparison to time-series or cross-sectional data.

2.4 Characteristics of Panel Data

- Multidimensionality: Allows tracking the same units (such as individuals or firms) across multiple time periods.
- Controlling for unobserved heterogeneity: Enables controlling for unobserved fixed factors that may influence the results.
- Analyzing dynamic relationships: Enables studying the lagged and cumulative effects of variables.

2.5 Types of Panel Data

Panel data come in many different types, each with its own analytical characteristics, advantages, and limitations. The most common classification in the statistical and economic literature is the distinction between balanced panel data and unbalanced panel data. Within each of these two types, there is long panel data (Macro Panel), so (N) which is the number of cross-sectional units, is relatively small, but the time period (T) is short. Short panel data (Micro Panel), if (N) is high, and (T) is short.

i- Balanced panel data: A Balanced panel data is defined as a set of observations collected for a fixed sample of sample units (such as companies or individuals) over an equal number of time periods. In other words, in a balanced panel data, it is required that each entity in the sample be observed in all the periods covered by the study. For example, a study tracking the annual sales of a sample of 100 companies over a period of 30 years is a balanced board provided that annual data are available for each company for the entire period. Balanced panels are distinguished by their analytical importance, as they enable researchers to apply fixed effects models that control the unobserved variation associated with the sample units. Technically, a data structure is considered a balanced panel when the number of observations is constant over all time periods. By contrast, in unbalanced plates which is common in random surveys the number of observations varies from one time interval to another, which may affect the efficiency of statistical estimates [4].

ii- Unbalanced panel data: Unbalanced panel is a set of data with a time periods for individuals or entities. For example, a study of how inflation affected economic growth for 15 Arab countries during the specified period. (1980-2020) is considered an unbalanced panel if data are not available for some countries for one or more of the study years. Another example of unbalanced panel data is to study the monthly rainfall for some regions in Nineveh Government in Iraq for the period (1990-2022), since rainfall in the summer months in some areas is almost non-existent, the readings for those areas in the summer months of each year are excluded, resulting in unbalanced panel data [5]. Unbalanced panel Data is of analytical interest, because it allows researchers to estimate random effects models that include observed and unobserved variance factors with sample units. An unbalanced data structure usually arises as a result of the absence of certain observations of certain units at specific time intervals during the data collection process. These gaps in the data stem from multiple reasons, the most prominent of which is the phenomenon of “attrition” and “selection bias”.

iii- Pooled cross-sectional data**:** Panel Data is a statistical collection that integrates cross-sectional observations from multiple sample units over a series of time periods, enabling analysis of changes within and between units over time. For example, a study comparing expenditure level of households’ groups in a country using data from the 2010, 2015, and 2020 data creates a pooled cross-sectional dataset. This type of data is very useful because it enables researchers to estimate the impacts of time-invariant variables on the outcome of interest.

iv- Between-group variation panel data: A data set that includes individuals or entities over multiple time periods, and is used in contexts where the analysis focuses more on the variation between sample units (Between-Group Variation) than the variation within each unit (Within-Group Variation). For example, a study comparing social characteristics (such as income or educational level) between individuals of different races over several years may generate a panel dataset highlighting Intergroup differences. The analytical significance of this type of data lies in enabling researchers to isolate and estimate the influence of variables that change over time on the dependent variable of interest, taking into account the structure of temporal and unitary dependence in the data.

2.6 Historical Background of Panel Data

With the growing interest in data science and its various types over the past few decades, longitudinal data has become the important type of data in researches due to its numerous advantages in terms of analysis, prediction, accuracy, and efficiency of results. This is particularly evident in many econometric studies. The following is a historical sequence of the development of panel data:

Phase One: Origins (1930s - 1950s)

The concept of panel data first appeared in social and economic research, particularly in studies of household and consumer behavior. Some surveys, such as the Harvard Longitudinal Studies, used this type of data, but without systematic statistical analysis. The problem with limited data in this era was the lack of computers, which made processing this type of data difficult and limited its use. The lack of statistical tools and computing power also limited the widespread use of this data.

Phase Two: Theoretical Crystallization (1960s - 1970s)

This period witnessed the first theoretical attempts to address panel data, as researchers began developing econometric models that combined cross-sectional and time-series data. Among the first authors to address this topic were [6] and [7], who discussed fixed-effects models. The first formal models for panel data were the fixed and random models, which were developed to address unobserved variation among units.

Phase Three: Theoretical and Technical Maturity (1980s - 1990s)

The characteristics of this phase include, first, the proliferation of economic studies and research based on panel data, as researchers realized the importance of this type of data in reducing bias resulting from unobserved variables. Second, the Hausman test became a standard tool for selection between a random model and a fixed model, leading to the use of more accurate models [8]. Third, the emergence of specialized computer software, such as “TSP” and “SHAZAM”, and later “STATA” and “EVIEWS”, facilitated the estimation of panel data models.

Phase Four: Expansion and Complexity (2000s - present)

This phase witnessed the development of estimation methods, with the introduction of dynamic models such as dynamic panel models (GMM) using the Arellano-Bond estimator to handle dynamic data and address endogeneity issues. The use of nonlinear models and long-run panel data also increased. As data collection techniques evolved, panels became larger and more detailed, particularly in fields such as macroeconomics, financial markets, energy, and education. Panel data applications also diversified, and they became widely used in forecasting, public policy analysis, economic growth, migration, poverty, finance, health, and more [9].

Current Phase: Artificial Intelligence and Big Data (2020s)

- Integrating Panel Data with Machine Learning techniques: Developing flexible and efficient estimation algorithms for complex models.
- Using neural networks and advanced analytics to estimate nonlinear models.
- New Challenges: Dealing with large and complex datasets requires the development of advanced analytical tools and models.
- Transition to Big Dynamic Panel Data:
- The emergence of large databases such as those used in the global economy, corporate data, and stock markets.

2.7 Methodology

The standard panel data model is [10]:

$$\boldsymbol{y} = \boldsymbol{\beta_0} + \boldsymbol{x\beta} + \boldsymbol{\varepsilon}$$

where:

N : is the number of cross-sections

T_i : is the time period for the section i, $i=1,2,\ldots,N$

k : is the number of independent variables (regressors)

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \\ \vdots \\ \boldsymbol{y_N} \end{bmatrix}, \ \boldsymbol{y_i} = \begin{bmatrix} y_{i,1} \\ y_{i,2} \\ \vdots \\ y_{i,T_i} \end{bmatrix}, \boldsymbol{\beta_0} = \begin{bmatrix} \beta_{\boldsymbol{0}_1} \\ \beta_{\boldsymbol{0}_2} \\ \vdots \\ \beta_{\boldsymbol{0}_{T_i}} \end{bmatrix},$$

$$\boldsymbol{x} = [\boldsymbol{x_1} \quad \boldsymbol{x_2} \quad \cdots \quad \boldsymbol{x_k}] \ , \ \boldsymbol{x_j} = \begin{bmatrix} x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,T_i} \end{bmatrix},$$

$j = 1,2,\ldots,k$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \boldsymbol{\varepsilon_1} \\ \boldsymbol{\varepsilon_2} \\ \vdots \\ \boldsymbol{\varepsilon_N} \end{bmatrix} \ , \ \boldsymbol{\varepsilon_i} = \begin{bmatrix} \varepsilon_{i,1} \\ \varepsilon_{i,2} \\ \vdots \\ \varepsilon_{i,T_i} \end{bmatrix},$$

$i = 1,2,\ldots,N$

$\boldsymbol{y}$ is a $\sum_{i=1}^{N} T_i \times 1$ matrix, $\boldsymbol{\beta_0}$ is a $\sum_{i=1}^{N} T_i \times 1$ matrix, $\boldsymbol{x}$ is a $\sum_{i=1}^{N} T_i \times k$ matrix, $\boldsymbol{\beta}\ is\ a\ k \times 1\ matrix$, and $\boldsymbol{\varepsilon}$ is a $\sum_{i=1}^{N} T_i \times 1$ matrix

Generally, panel data models are utilized to address two main problems: individual-fixed effects (unobserved individual heterogeneity) and endogeneity bias.

2.7.1 Static Panel Data Models

This type of models assumes that previous values of the dependent variable or the independent variables have no influence on current values of these variables. Some static models can solve endogeneity problem using standard instrumental variable and others just solve only heterogeneity problem. Assuming there are (N) cross-sections and (T) time units, the general form of balanced panel data models is:

$$y_{i,t} = \beta_{0,i} + \sum_{j=1}^{k} \beta_j x_{j(i,t)} + \varepsilon_{i,t}\ ; i = 1,2,\ldots,N\ ; \quad t = 1,2,\ldots,T \tag{1}$$

$x_{i,t}$ and $y_{i,t}$ are observations for each cross-sectional unit during time period t.

Panel data analysis models can be classified into two essential groups in terms of homogeneity assumptions in the parameters of the model. The first group includes pooled models or what are sometimes known as homogeneous panel models, which assume that the parameters of the relationship (such as slope and constant) are the same across all units (such as individuals or companies). The second group includes heterogeneous models, which allow the presence of a difference in one or all parameters of the model between the units. The most striking examples of this category are Fixed Effect Model (FEM) and Random Effect Model (REM). In this context, William Greene presented an analytical framework that includes three basic models of panel data, all of which assume that the individual effect of each cross-sectional unit is constant over time, although the way it is treated statistically differs whether represented as a specific parameter (in static effects) or as a random variable (in random effects) [11].

1. Pooled Regression Model (PRM)

Also known as the Common Effects Model (CEM), in this model the individual effect of any cross-sectional unit is constant for all cross-sections, meaning there is no role for time periods (any effect of time is neglected) [12]. Therefore, in this model the constant or intercept $\beta_{0,i}$ and the regression coefficients β_j will be the same for all cross-sections. The formula for this model is:

$$y_{i,t} = \beta_0 + \sum_{j=1}^{k} \beta_j x_{j(i,t)} + \varepsilon_{i,t} \ ;$$

$$i = 1,2,3,\ldots,N \ ; \ t = 1,2,3,\ldots,T \qquad (2)$$

for this model, the mean and variance of error are $E(\varepsilon_{i,t}) = 0$ and $V(\varepsilon_{i,t}) = \sigma_\varepsilon^2$. Ordinary Least Squares (OLS) method is used to estimate this model.

2. Fixed Effect Model (FEM)

This model assumes that each cross-section has its own fixed intercept, with all regression coefficients being identical for all cross-sections. In other words, this model assumes that there are as many parallel regression lines as there are cross-sections . The form of this model is:

$$y_{i,t} = \beta_{0,i} + \sum_{j=1}^{k} \beta_j x_{j(i,t)} + \varepsilon_{i,t} \ ;$$

$$i = 1,2,3,\ldots,N \ ; \ t = 1,2,3,\ldots,T \qquad (3)$$

for this model is $E(\varepsilon_{i,t}) = 0$ and $V(\varepsilon_{i,t}) = \sigma_\varepsilon^2$. Least Squares Dummy Variables (LSDV) method is used to estimate this model, where (N-1) dummy variables are added to the model $D_j \ ; \ j = 2,3,\ldots,N$, such that:

$$D_j = \begin{cases} 1 & \text{for section j} \\ 0 & \text{for other sections} \end{cases} \qquad \text{j} = 2,3,\ldots,\text{n} \qquad (4)$$

therefore, model (3) becomes

$$y_{it} = \alpha_1 + \sum_{j=1}^{N} \alpha_j D_j + \sum_{j=1}^{k} \beta_j x_{j(it)} + \varepsilon_{it} \ ;$$

$$i = 1,2,3,\ldots,N \ ; \ t = 1,2,3,\ldots,T$$

α_1 represents the first cross-section intercept. The amount $\alpha_1 + \sum_{j=1}^{N} \alpha_j D_j$ represents the change in the intercept parameter β_0, and since α_1 is constant quantity, it can be dropped from the equation, so the model becomes as follows:

$$y_{i,t} = \sum_{j=1}^{N} \alpha_j D_j + \sum_{j=1}^{k} \beta_j x_{j(i,t)} + \varepsilon_{i,t}$$

$$i = 1,2,3,\ldots,N \ ; \ t = 1,2,3,\ldots,T \qquad (5)$$

3. Random Effect Model (REM)

In panel data models based on the finite element theory (FEM) framework, model errors are assumed to follow a normal distribution with zero mean and variance σ_ε^2. To ensure efficient and unbiased parameter estimators, the assumptions of homoscedasticity and no autocorrelation are required: that is, the variance of error terms must remain homogeneous across all cross-sectional units, and that there is no correlation between the errors of observations over time for the same unit or between different units in any given time period.If one of the above assumptions is not met, then a (REM) is considered appropriate for the data, since this model has a intercept parameter as a random variable with mean μ, i.e.:

$$\beta_{0,i} = \mu + u_i \ ; \ i = 1,2,\ldots,N \qquad (6)$$

substituting (6) in (1), to get:

$$y_{it} = \mu + \sum_{j=1}^{k} \beta_j x_{j(it)} + u_i + \varepsilon_{it} \ ;$$

$$i = 1,2,\ldots,N \ ; \ t = 1,2,\ldots,T \qquad (7)$$

u_i is the error variable in the i-th cross-section data. The (REM) is sometimes called the error components model since model (7) contains two error components, v_i and $\varepsilon_{i,t}$. To estimate equation (7), the Generalized Least Squares

(GLS) method is used, where the basic condition for estimating this type of model is that (N > k) where (k) is the number of regressors (or slopes).

Selecting the best Model

To select the most suitable static model, different tests can be used, such as:

1. Chow Test

The Chow test is used to select between (PRM) and (FEM), which is also called the restricted F-Test. The null and alternative hypotheses are

H_0: Choose (PRM) if ($p \geq 0.05$)

H_1: Choose (FEM) if ($p < 0.05$)

The test statistics used is [13] and [14]:

$$F = \frac{(R^2_{FEM} - R^2_{PRM})/(N-1)}{(1 - R^2_{FEM})/(NT-k)} \tag{8}$$

where R^2_{FEM} and R^2_{PRM} are the determinant coefficients of (FEM) and (PRM) respectively.

Another formula for this test is [15]:

$$F = \frac{(RRSS - URSS)/(N-1)}{URSS/(NT-N-k)} \tag{9}$$

(*RRSS)* is the Restricted Residual Sum of Squares for (PRM), and (*URSS)* is the Unrestricted Residual Sum of Squares for (FEM).

2. Lagrange Multiplier Test

The Lagrange Multiplier (LM) test, also known as the Breusch-Pagan LM test, was developed by Breusch and Pagan to compare the (REM) and (PRM) to determine which is more appropriate in the context of panel data analysis [16]. It should be noted that this test is not applied in cases where the Chow and Hausman tests indicate the superiority of the (FEM). Rather, the LM test is only used in the scenario where the Chow test favors the (FEM) and the Hausman test favors the (REM). The LM test is then used as an aid to making the final decision between the remaining two models: the (PRM) and the (REM). These are the hypothesis in the LM test:

H_0: Choose (PRM) if ($p \geq 0.05$)

H_1: Choose (REM) if ($p < 0.05$)

The (LM) Test statistics is:

$$LM = \frac{NT}{2(T-1)}\left[\frac{\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\hat{\varepsilon}_{i,t}\right)^2}{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\varepsilon}^2_{i,t}} - 1\right] \tag{10}$$

It is found that under H_0, the statistic (LM) follows $\chi^2_{(1)}$.

3. Hausman Test

Hausman test is used to select between (FEM) and (REM). The hypotheses of this test are

H_0: Choose (REM) if ($p \geq 0.05$)

H_1: Choose (FEM) if ($p < 0.05$)

The test statistics used is [8] and [17]:

$$H = \left(\hat{\boldsymbol{\beta}}_{FEM} - \hat{\boldsymbol{\beta}}_{REM}\right)'\left[\boldsymbol{V}\left(\hat{\boldsymbol{\beta}}_{FEM}\right) - \boldsymbol{V}(\hat{\boldsymbol{\beta}}_{REM})\right]^{-1}\left(\hat{\boldsymbol{\beta}}_{FEM} - \hat{\boldsymbol{\beta}}_{REM}\right) \tag{11}$$

where $\hat{\boldsymbol{\beta}}_{FEM}$ and $\hat{\boldsymbol{\beta}}_{REM}$ represent the vectors of estimated parameters for the (FEM) and (REM) models, respectively, and $\boldsymbol{V}\left(\hat{\boldsymbol{\beta}}_{FEM}\right)$ and $\boldsymbol{V}(\hat{\boldsymbol{\beta}}_{REM})$ represent the variance-covariance matrices of the estimated parameters for the (FEM) and (REM), respectively. It is found that under H_0, the statistic (H) follows $\chi^2_{(k)}$ [18].

The above three tests to choose the best model, can be summarized in the graph below:

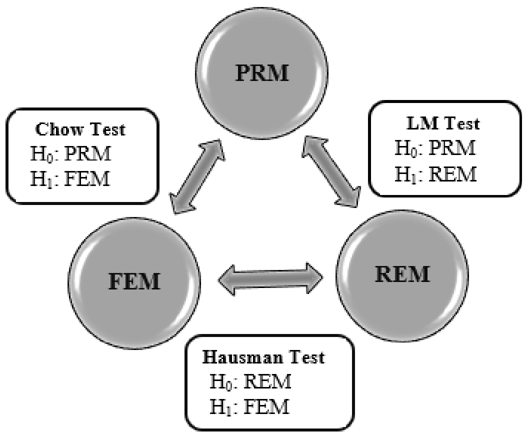

Figure 1. Choosing the best static model

2.7.2 Dynamic Panel Data Models

Recently, the use of dynamic panel data models has spread in various fields of research, particularly in economics, where their application has generated new and profound analytical insights. Recent examples of the use of these models include Flannery and Hankins study on corporate finance [19], Lee *et al*. analysis of the factors influencing economic growth [20], Dutta et al. research on the impact of foreign aid [21], and Pugh *et al*. study exploring the relationship between school spending and educational performance [22].

Dynamic panel data models extend the scope of static models by including the lagged dependent variable as an independent variable, demonstrating the persistence and adjustment of the dependent variable over time. These models are essential for analyzing temporal dependencies, but their complexity requires specialized estimation techniques, such as the (GMM: Generalized Method of Moments), due to the inconsistency of standard methods, such as (OLS) method, because the endogeneity problems resulting from the interaction between the lagged dependent variable and the error term.

- Why are Dynamic Models better than Static Models?

Dynamic models are considered better and more important than static models in economics for several scientific and methodological reasons, the most important of which are:

1. They reflect economic reality more accurately.
   - Economic phenomena do not occur all at once, but rather over time.
   - Dynamic models take into account the lag effect; that is, the impact of variables is not instantaneous but is distributed over time.
2. The ability to analyze causal and temporal relationships.
   - Dynamic models allow us to understand the true direction of causality (Granger causality).
   - They can distinguish between short-term and long-term effects.
3. Better forecasting.
   - Because they take into account past values of variables (lags), dynamic models provide higher forecasting accuracy than static models.
4. Handling time series data.
   - Most macroeconomic and microeconomic data are temporal, making dynamic models more suitable for their analysis.
   - They are also used to estimate models such as Autoregressive Distributed Lag (ARDL) model, Vector Autoregressive (VAR) model and Error Correction Model (ECM) model.
5. Understanding the Dynamics of Adjustment to Equilibrium
   - Kinetic models explain how an economy adjusts to a shock and how quickly it returns to equilibrium (for example, using the correction factor in the (ECM)).
6. Addressing the Autoregressive Problem
   - Many economic variables exhibit autocorrelation behavior, and kinetic models naturally address this by introducing lagged variables.

Dynamic panel data models allowing for the study of intertemporal relationships and individual heterogeneity. They are particularly useful when the dependent variable depends on its own past values or when there are unobserved individual-specific effects. These models can be used to control for the effects of unobserved or missing variables, disentangle components of variance, and estimate transition probabilities. Dynamic models solve both heterogeneity and endogeneity problems. The general form of the linear dynamic panel data model is:

$$y_{i,t} = \delta y_{i,t-1} + \boldsymbol{x_{i,t}\beta} + v_{i,t} \;\; ; \;\; i=1,2,3,\ldots,N \;, \;\; t=1,2,3,\ldots,T \qquad (12)$$

δ is a scalar and unknown parameter, $\boldsymbol{x_{i,t}}$ is $k \times 1$ matrix of regressors and $\boldsymbol{\beta}$ is $1 \times k$ vector of unknown parameters. It assumes that:

$$v_{i,t} = u_i + \varepsilon_{i,t}$$

u_i are iid such that $E(u_i) = 0$ and $V(u_i) = \sigma_u^2$, $\varepsilon_{i,t}$ are iid with $E(\varepsilon_{i,t}) = 0$ and $V(\varepsilon_{i,t}) = \sigma_\varepsilon^2$ and u_i and $\varepsilon_{i,t}$ are independent of each other. Model (12) has been extensively studied by Anderson and Hsiao [23] and Sevestre and Trognon [24]. Since $y_{i,t}$ is a function of $y_{i,t-1}$, $v_{i,t}$ and then for u_i, therefore $y_{i,t-1}$, a right hand regressor in (12), is correlated with the error term $\varepsilon_{i,t}$. This render the (OLS) estimator biased and inconsistent even if the $\varepsilon_{i,t}$ are not correlated.

Dynamic panel data models are characterized by two main sources of dynamism (or time dependence): the first is autocorrelation resulting from the inclusion of the lagged dependent variable as an explanatory variable, and the second is individual effects, which capture unobserved differences between cross-sectional units across time. To estimate this type of model, which suffers from consistency problems resulting from the correlation of the lagged dependent variable with the individual effects, the generalized method of moments (GMM) methodology is typically used, pioneered by Hansen [25] and later developed in the context of panel data by Arellano and Bond [9].

Dynamic Panel with Fixed T

Consider the following simple autoregre-ssive model AR(1):

$$y_{i,t} = \delta y_{i,t-1} + \varepsilon_{i,t} \ ; \quad i=1,2,\dots,N \ , t=2,3,\dots,T \tag{13}$$

such that $|\delta| < 1$, (OLS) estimator of δ is:

$$\hat{\delta} = \frac{\sum_{i=1}^{N}\sum_{t=2}^{T}\left(y_{i,t-1} - \bar{y}_{i,-1}\right)\left(y_{i,t} - \bar{y}_i\right)}{\sum_{i=1}^{N}\sum_{t=2}^{T}\left(y_{i,t-1} - \bar{y}_{i,-1}\right)^2} \tag{14}$$

where

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{i,t} \quad \text{and} \quad \bar{y}_{i,-1} = \frac{1}{T}\sum_{t=2}^{T} y_{i,t-1}$$

and transformation $(y_{i,t-1} - \bar{y}_{i,-1})$ is independent of $u_{i,t}$ and correlated with $\varepsilon_{i,t}$; that is:

$$Cov(y_{i,t-1} - \bar{y}_{i,-1}, \varepsilon_{i,t}) = -\frac{\sigma_\varepsilon^2}{T}$$

so, as $N \to \infty$ for fixed T the estimator will be biased and inconsistent. To analyze this, we can substitute (13) into (14) and get:

$$\hat{\delta} = \delta + \frac{\sum_{i=1}^{N}\sum_{t=2}^{T}\left(\varepsilon_{i,t-1} - \bar{\varepsilon}_{i,-1}\right)\left(y_{i,t} - \bar{y}_i\right)}{\sum_{i=1}^{N}\sum_{t=2}^{T}\left(y_{i,t-1} - \bar{y}_{i,-1}\right)^2} \tag{15}$$

As we mentioned above, $y_{i,t}$ is correlated with $\varepsilon_{i,t}$, therefore $\hat{\delta}$ is biased estimator and inconsistent, but (Nickell, 1981) has shown that, the fixed effects estimator will be consistent when the dynamic error components model is used, at $T \to \infty$ [26].

A) First-difference IV Estimation

Anderson & Hsiao proposed this model, the method suggests that to take the first difference of (13) to wipes out the individual effects u_i: [23]

$$\Delta y_{i,t} = \delta \Delta y_{i,t-1} + \Delta\varepsilon_{i,t} \ ; \quad t=2,3,\dots,T \tag{16}$$

where $\Delta\varepsilon_{i,t}$ constitutes a first-order moving average MA(1) process, the ordinary least squares (OLS) estimator derived from equation (16) is inconsistent, since there is a correlation between $\Delta y_{i,t-1}$ and the error term $\Delta\varepsilon_{i,t}$. This correlation arises primarily from the correlation between $y_{i,t-1}$ and $\varepsilon_{i,t-1}$, resulting from the presence of idiosyncratic or endogenous dynamic effects in the model. To overcome this problem, Anderson & Hsiao proposed using either $\Delta y_{i,t-2}$ or $y_{i,t-2}$ as instrumental variables (IVs) for $y_{i,t-1}$ [27]. Since the error disturbances $\varepsilon_{i,t}$ are assumed to be independent across time (i.e., $\varepsilon_{i,t}$ is uncorrelated with $\varepsilon_{i,s}$ for all $t \neq s$), these instruments will not be correlated with $\Delta\varepsilon_{i,t}$, thus satisfying the validity condition for the instrumental variables. For instance, if we use $y_{i,t-2}$ as an instrument for $\Delta y_{i,t-1}$], the resulting consistent IV estimator can be formulated as:

$$\hat{\delta}_{IV} = \frac{\sum_{i=1}^{N}\sum_{t=2}^{T} y_{i,t-2}\Delta y_{i,t}}{\sum_{i=1}^{N}\sum_{t=2}^{T} y_{i,t-2}\Delta y_{i,t-1}} \tag{17}$$

any $y_{i,s}$, $s \leq t$-2 can be used as an instrument. Take in accounts that the necessary condition for consistency is that:

$$\underset{N\to\infty}{plim} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T} y_{i,t-2}\Delta\varepsilon_{i,t} = 0$$

B) IV-GMM Estimator

This method is known also as Arellano and Bond estimator due to their work in (1991). They point out that instruments can be drawn from the orthogonality conditions that arise among lagged values. of $y_{i,t}$ and $\varepsilon_{i,t}$. Recall model (12): [8]

$$y_{i,t} = \delta y_{i,t-1} + \boldsymbol{x_{i,t}\beta} + u_i + \varepsilon_{i,t} \ ; i=1,2,\dots,N \ , \ t=2,3,\dots,T$$

the first difference is:

$$\Delta y_{i,t} = \delta \Delta y_{i,t-1} + \Delta\boldsymbol{x_{i,t}\beta} + \Delta\varepsilon_{i,t} \tag{18}$$

Assume that:

$$E(\boldsymbol{x'_{i,t}}\varepsilon_{i,s}) = 0 \ \ for\ all\ t, s = 1,2,\cdots,T$$

and assume that all $\boldsymbol{x_{i,t}}$ are independent of u_i. Then, all $\boldsymbol{x_{i,t}}$ are valid instruments for (18) in which case the matrix of $(1 \times kT)$ is defined by:

$$\boldsymbol{x}_i^* = (\boldsymbol{x_{i,1}} \quad \boldsymbol{x_{i,2}} \quad \dots \quad \boldsymbol{x_{i,T}})$$

Since both $y_{i,1}$ and $y_{i,2}$ are not correlated with $\Delta\varepsilon_{i,4}$, then they are valid instrument for $\Delta y_{i,3}$. By continuing this procedure for the period T, the set $(y_{i,1}, y_{i,2}, \dots, y_{i,T-2})$ will becomes valid instruments. Now define the matrix of $(T\text{ - }2)\times[(1+2+\ldots+T\text{-}2)+kT(T\text{ - }2)]$:

$$\boldsymbol{Z_i}=\begin{bmatrix}(y_{i,1},\boldsymbol{x}_i^*) & 0 & \dots\dots & 0\\ 0 & (y_{i,1},y_{i,2,},\boldsymbol{x}_i^*) & \dots\dots & 0\\ \vdots & \vdots & \vdots & \vdots\\ \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & \dots\dots & (y_{i,1},y_{i,2},\dots,y_{i,T-2},\boldsymbol{x}_i^*)\end{bmatrix}$$

The set of instruments is represented as a matrix, since each row contains the set of instruments valid for a given time period. Therefore, the entire set of moment conditions can be expressed in a concise form as follows:
$E(\boldsymbol{Z}_i'\Delta\boldsymbol{\varepsilon_i})=\boldsymbol{0}$
where $\Delta\boldsymbol{\varepsilon_i}=(\Delta\varepsilon_3 \quad \Delta\varepsilon_4 \quad \cdots \quad \Delta\varepsilon_T)'$, or:
$E\left(\boldsymbol{Z}_i'(\Delta\boldsymbol{y}_i-\delta\Delta\boldsymbol{y}_{i,-1})\right)=\boldsymbol{0}$
where $\Delta\boldsymbol{y_i}=(\Delta y_3 \quad \Delta y_4 \quad \cdots \quad \Delta y_T)'$ and $\Delta\boldsymbol{y_{i,-1}}=(\Delta y_2 \quad \Delta y_3 \quad \cdots \quad \Delta y_{T-1})'$ are $(T\times 2)$ vectors, respectively. A total number of moment conditions adds to $(1+2+3+\ldots+T\text{-}2)$. Next, define the $N(T\text{-}2)\times(1+2+3+\ldots+T\text{-}2)$ matrix of instruments by:

$$\boldsymbol{Z}=\begin{bmatrix}\boldsymbol{Z}_1\\ \boldsymbol{Z}_2\\ \vdots\\ \boldsymbol{Z}_N\end{bmatrix}$$

model (18) can be rewritten in the matrix form as follows:

$$\Delta\boldsymbol{y}=\boldsymbol{\delta}\Delta\boldsymbol{y_{-1}}+\Delta\boldsymbol{X\beta}+\Delta\boldsymbol{\varepsilon} \qquad (19)$$

where:

$$\Delta\boldsymbol{y}=\begin{bmatrix}\Delta\boldsymbol{y}_1\\ \Delta\boldsymbol{y}_2\\ \vdots\\ \Delta\boldsymbol{y}_N\end{bmatrix}_{N(T\text{-}2)\times 1},\ \Delta\boldsymbol{y_{-1}}=\begin{bmatrix}\Delta\boldsymbol{y_{1,-1}}\\ \Delta\boldsymbol{y_{2,-1}}\\ \vdots\\ \Delta\boldsymbol{y}_{N,-1}\end{bmatrix}_{N(T\text{-}2)\times 1},$$

$$\Delta\boldsymbol{X}=\begin{bmatrix}\Delta\boldsymbol{x}_{1,t}\\ \Delta\boldsymbol{x}_{2,t}\\ \vdots\\ \Delta\boldsymbol{x}_{N,t}\end{bmatrix}_{N(T\text{-}2)\times k},\ \boldsymbol{\beta}=\begin{bmatrix}\beta_1\\ \beta_2\\ \vdots\\ \beta_k\end{bmatrix}_{k\times 1},$$

$$\Delta\boldsymbol{\varepsilon}=\begin{bmatrix}\Delta\boldsymbol{\varepsilon_1}\\ \Delta\boldsymbol{\varepsilon_2}\\ \vdots\\ \Delta\boldsymbol{\varepsilon_N}\end{bmatrix}_{N(T\text{-}2)\times 1}$$

multiply (19) by $\boldsymbol{Z}'$

$$\boldsymbol{Z}'\Delta\boldsymbol{y}=\boldsymbol{\delta}\boldsymbol{Z}'\Delta\boldsymbol{y_{-1}}+\boldsymbol{Z}'\Delta\boldsymbol{X\beta}+\boldsymbol{Z}'\Delta\boldsymbol{\varepsilon} \qquad (20)$$

From (20), the (Arellano & Bond, 1991) suggested that the two-step Generalized Least Squares (GLS) estimator can then be obtained by: [8]

$$\begin{pmatrix}\hat{\delta}_{GLS,2}\\ \widehat{\boldsymbol{\beta}}_{GLS,2}\end{pmatrix}=\left(\Delta\boldsymbol{Q}'\boldsymbol{Z}\widehat{\boldsymbol{W}}_N^{-1}\boldsymbol{Z}'\Delta\boldsymbol{Q}\right)^{-1}\left(\Delta\boldsymbol{Q}'\boldsymbol{Z}\widehat{\boldsymbol{W}}_N^{-1}\boldsymbol{Z}'\Delta\boldsymbol{Q}\right) \qquad (21)$$

where $\Delta Q=(\Delta\boldsymbol{y}_{-1} \quad \Delta\boldsymbol{X})$ and

$$\widehat{\boldsymbol{W}}_N^{-1}=\sum_{i=1}^{N}\boldsymbol{Z}_i'\Delta\hat{\boldsymbol{\varepsilon}}_i\Delta\hat{\boldsymbol{\varepsilon}}_i'\boldsymbol{Z}_i$$

C) System GMM Estimator

This method is also known as the Arellano and Bover estimator, due to their work in (1995), where they developed a unifying IV-GMM framework for dynamic panel data models. Consider the following static panel model [28]:

$$y_{i,t}=\boldsymbol{x_{i,t}\beta}+q_i\gamma+v_{i,t} \quad ; \quad i=1,2,3,\ldots,N\ , \quad t=1,2,3,\ldots,T \qquad (22)$$

in vector form

$$\boldsymbol{y}_i=\boldsymbol{X_i}\ \boldsymbol{\beta}+\boldsymbol{Q}_i\boldsymbol{\gamma}+\boldsymbol{v}_i=\boldsymbol{S}_i\boldsymbol{\theta}+\boldsymbol{v}_i \qquad (23)$$

where

$\boldsymbol{v}_i=u_i\boldsymbol{e}_T+\boldsymbol{\varepsilon}_i$

and

$$\boldsymbol{y}_i=\begin{bmatrix}y_{i,1}\\ y_{i,2}\\ \vdots\\ y_{i,T}\end{bmatrix}_{T\times 1},\ \boldsymbol{X}_i=\begin{bmatrix}\boldsymbol{x}_{i,1}\\ \boldsymbol{x}_{i,2}\\ \vdots\\ \boldsymbol{x}_{i,T}\end{bmatrix}_{T\times k},\ \boldsymbol{Q}_i=\begin{bmatrix}\boldsymbol{q}_1\\ \boldsymbol{q}_2\\ \vdots\\ \boldsymbol{q}_i\end{bmatrix}_{T\times g},$$

$$\boldsymbol{v}_i=\begin{bmatrix}v_{i,1}\\ v_{i,2}\\ \vdots\\ v_{i,T}\end{bmatrix}_{T\times 1},\ \boldsymbol{S}_i=[\boldsymbol{X}_i,\boldsymbol{Q}_i]_{T\times(k+g)}$$

$$\boldsymbol{\theta}=\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\end{bmatrix}_{(k+g)\times 1},\ \boldsymbol{e}_T=\begin{bmatrix}1\\ 1\\ \vdots\\ 1\end{bmatrix}_{T\times 1},\ \boldsymbol{\varepsilon}_i=\begin{bmatrix}\varepsilon_{i,1}\\ \varepsilon_{i,2}\\ \vdots\\ \varepsilon_{i,T}\end{bmatrix}_{T\times 1}$$

using the following non-singular matrix transformation, Arellano and Bover transform model (23) as follows:

$$\boldsymbol{H}=\begin{bmatrix}\boldsymbol{C}\\ \boldsymbol{e}_T'/T\end{bmatrix}_{T\times T}$$

matrix $\boldsymbol{C}$ in any $(T\text{-}1)\times T$ of rank (T-1) where $\boldsymbol{Ce}_T = \boldsymbol{0}$.

Multiply $\boldsymbol{H}$ from the right by $\boldsymbol{v}_i$, then we obtain the transformed disturbances:

$$\boldsymbol{v}_i^* = \boldsymbol{H}\boldsymbol{v}_i = \begin{bmatrix} \boldsymbol{C}\boldsymbol{v}_i \\ \bar{v}_i \end{bmatrix} \qquad (24)$$

It is worth noting that the first (T-1) transformed disturbances are free of the individual effects u_i by construction Therefore, all exogenous variables are considered valid instruments for the first (T−1) equations in equation (23).

Define the $1\times(kT+g)$ vector $\boldsymbol{z}_i = (\boldsymbol{x}_{i,1} \quad \boldsymbol{x}_{i,2} \quad \cdots \quad \boldsymbol{x}_{i,T} \quad \boldsymbol{q}_i)$, and let m_i denote the row vector consisting of a subset of variables in z_i, assumed to be uncorrelated with the single effect u_i , such that $\dim(m_i) = \mathrm{m} \geq \dim(\gamma) = \mathrm{g}$. Then, the valid instrument matrix is given by:

$$\boldsymbol{Z_i} = \begin{bmatrix} \boldsymbol{z}_i & 0 & \dots\dots & 0 & 0 \\ 0 & \boldsymbol{z}_i & \dots\dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & 0 \\ 0 & 0 & \dots\dots & \boldsymbol{z}_i & \boldsymbol{m}_i \end{bmatrix}_{T\times(kT+g+m)} \qquad (25)$$

and the given moment conditions are:

$$E(\boldsymbol{Z}_i'\boldsymbol{H}\boldsymbol{v}_i) = 0 \qquad (26)$$

write (23) in matrix form

$$\boldsymbol{y} = \boldsymbol{S\theta} + \boldsymbol{v} \qquad (27)$$

where

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}_1 \\ \boldsymbol{S}_2 \\ \vdots \\ \boldsymbol{S}_N \end{bmatrix}_{NT\times(k+g)}, \quad \boldsymbol{v} = \begin{bmatrix} \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \\ \vdots \\ \boldsymbol{v}_N \end{bmatrix}_{NT\times 1}$$

Defining the matrix of instruments:

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_1 \\ \boldsymbol{Z}_2 \\ \vdots \\ \boldsymbol{Z}_N \end{bmatrix}_{NT\times(kT+g+m)}$$

Pre-multiplying (27) by $\boldsymbol{Z}'\overline{\boldsymbol{H}}$, where $\overline{\boldsymbol{H}} = \boldsymbol{I_N} \otimes \boldsymbol{H}$ is a $NT \times NT$ matrix, we obtain:

$$\boldsymbol{Z}'\overline{\boldsymbol{H}}\boldsymbol{y} = \boldsymbol{Z}'\overline{\boldsymbol{H}}\boldsymbol{S\theta} + \boldsymbol{Z}'\overline{\boldsymbol{H}}\boldsymbol{v} \qquad (28)$$

The optimal Arellano–Bover estimator (GMM) for the parameter θ, based on the moments condition in equation (26), is defined as the optimal (GLS) estimator applied to the system described in equation (28), as follows:

$$\widehat{\boldsymbol{\theta}}_{GLS} = (\boldsymbol{S}'\overline{\boldsymbol{H}}\boldsymbol{Z}\boldsymbol{W}^{-1}\boldsymbol{Z}'\overline{\boldsymbol{H}}\boldsymbol{S})^{-1} (\boldsymbol{S}'\overline{\boldsymbol{H}}\boldsymbol{Z}\boldsymbol{W}^{-1}\boldsymbol{Z}'\overline{\boldsymbol{H}}\boldsymbol{y}) \qquad (29)$$

where

$$Var(\boldsymbol{v}) = \boldsymbol{I}_N \otimes E(\boldsymbol{v}_i\boldsymbol{v}_i') = \boldsymbol{I}_N \otimes \boldsymbol{\Omega}$$

$$\boldsymbol{W} = Var(\boldsymbol{Z}'\overline{\boldsymbol{H}}\boldsymbol{v}) = \boldsymbol{Z}'\overline{\boldsymbol{H}}(\boldsymbol{I}_N \otimes \boldsymbol{\Omega})\overline{\boldsymbol{H}}'\boldsymbol{Z} = \boldsymbol{Z}'(\boldsymbol{I}_N \otimes \boldsymbol{H\Omega H}')\boldsymbol{Z}$$

Dynamic Panel with both N and T are Large

A) The Mean Group (MG) Estimator

The traditional dynamic panel model, which is applied in situations where there are a small number of temporal periods (small T), addresses the need to have the constants vary over cross-sectional units by means of adding unit-specific effects. However, this approach fails to consider one of the most important features of a panel dataset, which is the fact that slope coefficients can be differentiated by units or over time. There are three broad reasons for estimating heterogeneous dynamic panel models in which the slope coefficients are allowed to vary.

First, “in the context of obtaining an unbiased estimate of the average effect of exogenous variables, the presence of heterogeneity in slopes is not a prohibitive limitation”. If the slope variables are exogenous and their coefficients vary consistently around a random distribution across the cross-sectional units, analysis done by Zellner in 1969 demonstrates that pooled estimators, such as fixed and random effects estimators, yield an unbiased estimate of the average effect of those variables [29].

Second, in the case of dynamic panel model with small T, the scope for slope variance analysis appears to be significantly limited [30].

Third, Long-run responses are more common as the dependent variable in econometric analyses, and are less affected by the LVAR’s characteristic of allowing slope coefficients to vary across units. This necessitates an understanding of the relationship between coefficient estimates from time series models in the long-run, cross-sectional estimates, and panel data estimates.

In reality the heterogeneity of slope coefficients in the cross-sectional dimension can be to such an extent that pooling is incorrect or even misleading. In this regard, a more recent emerging practice is that of seeking estimates and making inferences about the so-called "group average" or the average over panel units, which is offered as a different target of inference that allows for a non-restrictive assumption of homogeneity.

Consider the following dynamic heterogeny-eous panels:

$$y_{i,t} = \delta_i y_{i,t-1} + \boldsymbol{x}_{i,t}\boldsymbol{\beta}_i + v_{i,t} \;\; ; \;\; i=1,2,...,N \;\; , \;\; t=2,3,...,T \qquad (30)$$

where we assume:

i. $\boldsymbol{x_{i,t}}$ and $v_{i,s}$ are independent for all *t* and *s*.
ii. $\delta_i \sim iidN(\delta, \sigma_\delta^2)$.
iii. $\boldsymbol{\beta}_i \sim iidN(\boldsymbol{\beta}, \boldsymbol{\Omega})$.
iv. δ_i and $\boldsymbol{\beta}_i$ are independently distributed with $y_{i,t}, \boldsymbol{x}_{i,t}$ and $v_{i,t}$ for all *t*.
v. $\boldsymbol{x}_{i,t}$ are covariance stationary processes.

The (MG) estimator proposed by Pesaran and Smith refers to the method that relies on estimating regression equations separately for each cross-sectional unit, that is [31]:

$$\widehat{\boldsymbol{\beta}}_{MG} = N^{-1}\sum_{i=1}^{\boldsymbol{N}} \widehat{\boldsymbol{\beta}}_i \qquad (31)$$

$\widehat{\boldsymbol{\beta}}_i$ is the OLS estimator of $\boldsymbol{\beta}_i$.

B) Pooled Mean Group (PMG) Estimator

According to N and T, there are three alternative estimation methods for heterogeneous dynamic panels:

1. **Small N and Large T:** When $\mathrm{N} > 1$, the SURE (Seemingly Uncorrelated Regression Equation) technique is often used. The basic advantage of this technique is that it allows simultaneous error variances to be freely estimated. However, this is only possible when N is small relative to T.
2. **Large N and Small T:** (Pesaran & Smith, 1995) show that traditional procedures for estimating models, such as fixed effects, random effects, instrumental variables (IVs), or (GMM) estimators, can produce inconsistent, and potentially highly misleading, estimates of mean coefficient values unless the slope coefficients are homogeneous. However, in most panels of this type, tests indicate that these coefficients vary significantly across groups [31].
3. **Large N and T:** Hsiao, Pesaran and Tahmiscioglu, considered Bayesian estimation of short-run coefficients in dynamic heterogeneous panels [32], and proved the asymptotic equivalence of the Bayesian estimator and the ensemble mean estimator [33], showing that the ensemble mean estimator is asymptotically normal for large N and large T, as long as $\sqrt{N}/T \to 0$ where both N and $T \to \infty$.

3. Results

3.1 The Data and Variables

Macroeconomic and socioeconomic variables were used for 34 provinces in Indonesia (2012–2019), as published by the (BPS: Badan Pusat Statistik) Central Statistics Agency in Indonesia [34]. The variables used in this study are:

POV: Poverty, the percentage of poor people living below the poverty line.

HDI: Human Development Index, which shows how well a population achieves development outcomes in terms of income, health, and education.

GII: Gini Index, a tool to measure the degree of income inequality among a population. Its value ranges from zero to one, where zero indicates equal income and one indicates high inequality.

UNR: Unemployment Rate (total unemployed divided by total labor force).

The general study model is:

$$POV_{i,t} = F(HDI_{i,t}, GII_{i,t}, UNR_{i,t}) + \varepsilon_{i,t} \;\; ; \;\; i = 1,2,\dots,N \;\; , \;\; t = 1,2,\dots,T$$

Here $\mathrm{N} = 34$ and $T = 8$, the total observations is $\mathrm{NT} = 272$. So, the structure of panel data is of the balanced and micro type ($\mathrm{N} > \mathrm{T}$).

3.2 Estimation of the Panel Data Models

A. Static Models

Table (4) shows the results of the three static models' estimation (PRM, FEM and REM).

The choice of the correct model was made by applying the Chow test, Lagrange Multiplier (LM) test and Hausman test as follows. (8), (10) and (11) respectively.

Table 4. Panel static models' estimation

Variables	PRM Model		FEM Model		REM Model	
	Estimator	Prob.(t)	Estimator	Prob.(t)	Estimator	Prob.(t)
HDI_t	-0.9404	0.000	-0.4482	0.000	-0.4620	0.000
GII_t	45.9917	0.000	-1.7624	0.496	-1.5078	0.567
UNR_t	0.0188	0.891	0.0328	0.530	0.0206	0.697
$\bar{R}^2$	0.51		0.44		0.48	
F	94.65		82.88		254.86	
Prob.(F)	0.000		0.000		0.000	

Table 5. Panel models selection tests

Panel Test	Statistics	Value	Prob.	Decision	Conclusion
Chow Test	*F_Stat.*	357.210	0.000	Reject H_0	FEM is better than PRM
LM Test	χ_1^2	723.680	0.000	Reject H_0	REM is better than PRM
Hausman Test	χ_3^2	14.208	0.003	Reject H_0	FEM is better than REM

Based on the results of the three tests performed, it can be concluded that the (FEM) is the most appropriate model for representing the relationship under study. However, before proceeding to interpret the parameters of this model, diagnostic tests must be performed to verify that it meets the basic assumptions of linear regression, these tests include: Jarque-Bera test to verify the assumption of normal distribution of the model's residuals; Breusch-Pagan test to examine the presence of homoskedasticity versus heteroskedasticity; evaluating multicollinearity by examining the variance inflation factor (VIF) values; and testing for autocorrelation using the Durbin-Watson statistic. The results of these four diagnostic tests were as follows:

Table 6. Diagnostic Checking tests

Assumptions	Statistics (Tests)	Prob.	Decision	Conclusion
Normality	Jarque-Bera	0.0869	Accept H_0	Normal distributed errors
Heteroscedasticity	Bruesch-Pagan	0.2011	Accept H_0	Homoscedasticity errors
Multicollinearity	Variance Inflation Factor	$VIF < 5$	Accept H_0	No Multicollinearity
Autocorrelation	Durbin-Watson	0.000	Reject H_0	Autocorrelation errors

It is clear from Table (6) that the assumptions of the static models are met, except the autocorrelation assumption, which usually appears in models that include time series data. To overcome this problem, dynamic models are used, which adding the lagged dependent variable as an independent variable with a lags, such as the Arellano-Bond model.

B. Dynamic Model

Table (7) shows the results of the Arellano-Bond model estimation.

Variables	Estimators	Prob.(t)
POV_{t-1}	0.5292	0.000
POV_{t-2}	0.1807	0.000
HDI_t	-0.5682	0.000
GII_t	3.6880	0.007
UNR_t	-0.0094	0.739
$\bar{R}^2$	0.99	
F	1249.90	
Prob.(F)	0.000	

Table 7. Arellano-Bond model estimators

The F-test is significant, so the model is fit, meaning that the regressors have impact on poverty. According to the largest adjusted coefficient of determination ($\bar{R}^2$), the dynamic model is the best fit. Furthermore, the autocorrelation hypothesis on the lag data is met.

Table 8. Dynamic Model autocorrelation

Lags	t-test	Prob.(t)	Decision	Conclusion
1	-2.52905	0.011	Reject H_0	Autocorrelation
2	-1.27538	0.202	Accept H_0	No Autocorrelation
3	0.96767	0.333	Accept H_0	No Autocorrelation

From the above results, it can be seen that the probability values of Arrellano-Bond test on lags 2 and 3 are greater than 5% means that the model is free of autocorrelation problem.

4. Discussion

Based on the t-test and its probability values for the dynamic model estimators and comparing them at the (5%) significance level, the human development index (*HDI*) has a significant negative impact on poverty. When (*HDI*) increases by one point, poverty declines by (0.53%). This result suggests that any increasing in (*HDI*), the poverty will decrease, meaning that the improve of life quality through investments in health and education will increase income, and then lead to a decline in poverty. There is also a significant positive impact of the Gini coefficient of income inequality on poverty. When the Gini coefficient increases by one point, poverty increases by (3.69%). The greater the income inequality (gap) among the population, the greater the poverty rate. Regarding the impact of the unemployment rate on poverty, this study does not provide sufficient evidence that the unemployment rate in Indonesia affects poverty, suggesting that the ratio of unemployed people does not have a significant impact on poverty. This result can be justified by the fact that in some Indonesian regions, the unemployment that occurs due to the voluntary unemployment where a person chooses to be unemployed rather than having to work for an unsatisfactory wage or an undesirable job.

5. Conclusion

This paper reviewed the concept of panel data, its historical review, its advantages and disadvantages, its various uses, and the econometric models used for it. It became clear from the narrative that dynamic models may be better than static models in many fields, as they are more realistic and reflect the role of time in the phenomenon under study, as many variables may not have an immediate effect, but their actual impact becomes apparent over a period of time. They also enable prediction and causal analysis, and can address the problems of heteroscedasticity and autocorrelation in residuals.

Acknowledgments

The authors would like to express his appreciation to the academic staff of the Department of Statistics and Informatics, College of Computer Science and Mathematics, University of Mosul.

References

[1] Campbell, D.T. & Stanley, J. (1963). Experimental and Quasi-Experimental Designs for Research, Chicago, IL: Rand McNally.

[2] Frees, E.W. (2004). Longitudinal and Panel Data: Analysis and Applications for the Social Sciences, Cambridge University Press, London.

[3] League of Arab States (2018). Arab Organization for Agricultural Development, Arab Agricultural Statistics Yearbook, Vol. (37).

[4] Algamal, Z.Y. (2012). Selecting Model in Fixed and Random Panel Data Models, *Iraqi Journal for Computer Science and Mathematics*, 21: 266-285.

[5] Al-Ani, B.G. (2025), Modeling the Count Panel Data for Vehicle Accidents in Iraq, IV International Rimar Congress of Pure Applied Sciences, 22-44.

[6] Mundlak, Y. (1961). Empirical Production Function Free of Management Bias, Journal of Farm Economics, 43(1): 44–56.

[7] Hildreth, C. & Houck, J.P. (1968). Some Estimators for a Linear Model with Random Coefficients, Journal of American Statistical Association, 63: 584-595.

[8] Hausman, J. (1978). Specification Test in Econometrics, Econometrica, 46(6): 1251-1271.

[9] 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.

[10] Hsiao, C. (2003). Analysis of Panel Data, Cambridge University Press, New York.

[11] Greene, W.H. (2012). Econometric Analysis, 7th edition, Pearson, USA.
[12] Al-Taae, M.A. & Al-Ani, B.G. (2022), Application of Poisson's Hierarchical Regression Model to the Deaths of Covid-19 in Mosul City Hospitals, Iraqi Journal of Statistical Sciences, 19(2): 1-14.
[13] Chairunnisa, A.D. & Fauzan, A. (2023). Implementation of Panel Data Regression in the Analysis of Factors Affecting Poverty Levels in Bengkulu Province in 2017-2020, Journal of Sciences and Data Analysis, 4(1): 40-45.
[14] AL-Dabbagh, N.K. & AL-Iragi, B.A. (2021). The Impact of the Public Budget Deficit in Promoting the Development of the Financial Sector Gulf Cooperation Council Countries as a Model, Tannmyat Al-Rafidain Journal, 40(130): 50-76.
[15] Greene, W.H. (2008), Econometric Analysis, 6th edition, Upper Saddle River, NJ, Prentice-Hall.
[16] Breusch, T. & Pagan, A. (1980). The Lagrange Multiplier Test and its Application to Model Specification in Econometrics, Review of Economic Studies, 47: 239-254.
[17] Al-Ani, B.G. (2023), Probability Modeling of Rainfall for Some Regions in Nineveh, Journal of AL-Rafidain University College for Sciences, 54: 134–148.
[18] Bai, J. (2009). Panel Data Models with Interactive Fixed Effects, Econometrica, 77: 1229–1279.
[19] Flannery, M.J. & Hankins, K.W. (2013). Estimating Dynamic Panel Models in Corporate Finance, Journal of Corporate Finance, 19: 1-19.
[20] Lee, S.H., Levendis, J. & Gutierrez, L. (2012). Telecommunications and Economic Growth: An Empirical Analysis of Sub-Saharan Africa, Applied Economics, 44(4): 461-469.
[21] Dutta, N., Leeson, P.T. & Williamson, C.R. (2013). The Amplification Effect: Foreign Aid's Impact on Political Institutions, Kyklos 66(2): 208-228.
[22] Pugh, G., Mangan, J., Blackburn, V. & Radicic, D. (2014). School Expenditure and School Performance: Evidence from New South Wales Schools using a Dynamic Panel Analysis, British Educational Research Journal, 41(2): 244-264.

[23] Anderson, T.W. & Hsiao, C. (1981). Estimation of Dynamic Models with Error Components, American Statistics Association, 76(375): 598-606.
[24] Sevestre, P. & Trognon, A. (1985). A Note on Autoregressive Error Component Models, Journal of Econometrics, 28: 231-245.
[25] Hansen, L.P. (1982). Large Sample Properties of Generalized Method of Moments Estimators, Econometrica, 50(4): 1029-1054.
[26] Nickell, S. (1981). Biases in Dynamic Models with Fixed Effects, Econometrica, 49: 1417-1426.
[27] Anderson, T.W. & Hsiao, C. (1982). Formulation and Estimation of Dynamic Models using Panel Data, Journal of Econometrics, 18: 47-82.
[28] Arellano, M. & Bover, O. (1995). Another Look at the Instrumental Variable Estimation of Error Components Models, Journal of Econometrics 68: 29-51.
[29] Zellner, A. (1969). On the Aggregation Problem: A new Approach to a Troublesome Problem. In: Fox, K.A., Sengupta, J.K., Narasimham, G.V.L., (Eds.), Economic Models, Estimation and Risk Programming: Essays in Honor of Gerhard Tintner, Springer-Verlag, New York.
[30] Balestra, P. & Nerlove, M. (1966). Pooling Cross-Section and Time-Series Data in the Estimation of a Dynamic Model: The Demand for Natural Gas, Econometrica, 34: 585–612.
[31] Pesaran, M.H. & Smith, R.P. (1995). Estimating Long-run Relationships from Dynamic Heterogenous Panels, Journal of Econometrics, 68: 79-113.
[32] Hsiao, C., Pesaran, M.H. & Tahmiscioglu A.K. (1999). Bayes Estimation of short-run Coefficients in Dynamic Panel Data Models, in Analysis of Panels and Limited Dependent Variable Models: A Volume in Honour of G.S. Maddala edited by C. Hsiao, K. Lahiri, L-F Lee and Pesaran, M.H., Cambridge: Cambridge University Press.
[33] Swamy, P.A.V.B. (1970). Efficient Inference in a Random Coefficient Regression Model, Econometrica 38: 311-323.
[34] Sihombing, P.R. & Arsani, A.M. (2021). Static and Dynamic Panel Models: Which is better? (Case study: Poverty data in Indonesia 2012-2019). IOP Conf. Ser.: Earth Environ. Sci., 739: 1-10.