|  |  |  |
| --- | --- | --- |
| صورة تحتوي على نص, دائرة, الخط, رسم بياني  تم إنشاء الوصف تلقائياً | **Cairo University**  **Faculty of Graduate Studies for Statistical Research** | صورة تحتوي على نص, رمز, قصاصة فنية, الخط  تم إنشاء الوصف تلقائياً |

**Study On** **Panel Data Methodologies**

**With**

**Application for** **Macroeconometrics**

**(****Inflation Forecasting)**

**By:** [**Mohamed Sayed Abd AL-mgyd**](https://www.linkedin.com/in/mohamed-abd-al-mgyd-9886151b2/)

Statistician and Data Scientist

**Supervised By**

**Dr.** **[Amal Mohamed Abd-Elfatah](https://www.linkedin.com/in/amal-mohamed-abdelfattah-phd-3a6bb6a4/)**

Assistant Professor of Statistics

Department of Applied Statistics and Econometrics

Faculty of Graduate Studies for Statistical Research (FSSR)

# **Acknowledgement:**

Thanks, and appreciation to my parents and my supervisor Prof. Amal Mohamed Abdelfatah.

**Table of Contents**

[**Acknowledgement** 2](#_Toc200291464)

[**List of Abbreviations** 3](#_Toc200291465)

[**List of Symbols** 4](#_Toc200291466)

[**List of Tables** 4](#_Toc200291467)

[**List of Figures** 4](#_Toc200291468)

[**Abstract** 4](#_Toc200291469)

[**1.** **Introduction** 5](#_Toc200291470)

[**1.1 Study Objectives** 7](#_Toc200291471)

[**1.2 Structure of the Study** 7](#_Toc200291472)

[**2.** **Literature Review** 7](#_Toc200291473)

[**2.1 Overview** 8](#_Toc200291474)

[**2.2 Literature Evolution Tree** 10](#_Toc200291475)

[**2.3 Applications** 10](#_Toc200291476)

[**Appendix (A): R Codes** 13](#_Toc200291477)

[**Appendix (B): Python Codes** 13](#_Toc200291478)

[**References** 13](#_Toc200291479)

# **List of Abbreviations**

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

# **List of Symbols**

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

# **List of Tables**

# **List of Figures**

# **Abstract**

This study examines the effectiveness of panel data methodologies in Macroeconometrics, with an application to inflation forecasting. Using a harmonized dataset covering 70 countries from 2000 to 2024, we investigate how different panel estimators—Pooled OLS, Fixed Effects (FE), Random Effects (RE), and dynamic approaches such as the Arellano–Bond GMM—perform in predicting average consumer price changes (PCPIPCH). Explanatory variables include government fiscal indicators, trade volumes, investment ratios, labor market conditions, and PPP measures, primarily sourced from the IMF World Economic Outlook and the World Bank.

Empirical analysis involves model selection via Hausman and Wald tests, alongside diagnostics for serial correlation, stationarity, and heteroskedasticity. Forecast accuracy is evaluated using out-of-sample Root Mean Square Error (RMSE). Results reveal that dynamic panel models consistently yield lower RMSE values, effectively addressing endogeneity and country-specific shocks. The Arellano–Bond GMM estimator emerges as the most robust tool when lagged inflation and macroeconomic fundamentals are included.

Our findings highlight the importance of methodological rigor in panel estimation for macroeconomic forecasting. These insights offer evidence-based guidance for policymakers seeking reliable inflation projections across diverse economic environments.

By identifying the most suitable panel model for inflation prediction, this research contributes to Macroeconometrics literature and provides practical tools for policymakers. The study aims to demonstrate that dynamic panel estimators deliver superior forecasting accuracy, offering robust insights into inflation dynamics under varying economic conditions.

Keywords: Panel Data, Macroeconometrics, Python, Statistical Models, Inflation Forecasting.

# **Introduction**

Over the past several decades, the analysis of panel data has emerged as a fundamental approach in empirical economics, revolutionizing the way economists examine complex phenomena involving multiple entities observed over numerous time periods (Baltagi, 2008). Panel data, also known as longitudinal or cross-sectional time-series data, combines the depth of temporal analysis with the breadth of cross-sectional insights, offering a uniquely powerful framework that enhances the precision and richness of econometric inference (Hsiao, 2003).

The evolution of panel data methodologies can be traced back to the mid-20th century when researchers began to recognize the limitations of relying solely on cross-sectional or time-series datasets. Early econometric models treated observations independently, often neglecting the persistent effects of unobserved factors or the inertia present in many economic processes. Pioneering work by Verbeek and Nijman (1992) and subsequent formalization by Moulton (1990) laid the groundwork for techniques that explicitly model individual-specific effects and temporal dependencies. The seminal contributions of Arellano and Bond (1991) introduced dynamic panel estimators that addressed endogeneity concerns, further cementing panel analysis as a centerpiece of modern econometrics (Arellano & Bond, 1991).

Panel data methods offer several crucial advantages over purely cross-sectional or time-series analyses. By including individual-specific parameters, either as fixed or random effects, panel models account for unobservable characteristics (e.g., institutional quality, cultural factors) that remain constant over time but vary across entities (Baltagi, 2008).

Pooling data over time increases the sample size, improving the efficiency of parameter estimates and reducing variance inflation commonly found in cross-sectional regressions (Wooldridge, 2010). Dynamic panel models incorporating lagged dependent variables illuminate persistence and adjustment processes in macroeconomic indicators, such as how past GDP growth influences current performance (Arellano & Bond, 1991).

Econometricians can choose between pooled OLS, fixed effects, random effects, or more sophisticated Generalized Method of Moments (GMM) techniques depending on data characteristics and research questions. The practical power of panel data analysis is exemplified in numerous macroeconomic studies. Barro and Sala-i-Martin (2004) employed panel models to disentangle the roles of investment, human capital, and institutional quality across countries, revealing how catch-up growth dynamics differ by initial income levels. Studies by Fischer (1993) utilized panel techniques to demonstrate that inflation inertia varies significantly across monetary regimes, highlighting the importance of both country-specific policies and international spillovers. Nickell (1997) applied fixed effects models to assess how labor market rigidities and unemployment benefits influence unemployment durations across OECD countries, providing policy-relevant insights into social welfare design.

The policy implications of panel data findings are profound. By accurately capturing both heterogeneity and dynamics, panel analysis informs tailored policy prescriptions—such as identifying which fiscal stimuli best spur growth in different institutional contexts or evaluating the differential impact of interest rate changes across economies.

## **Study Objectives**

This ***project’s primary objectives*** are to:

1. Evaluate pooled OLS, fixed effects, random effects, and dynamic panel estimators in terms of consistency, efficiency, and applicability to macroeconomic data.
2. Utilize a panel dataset of macroeconomic indicators from multiple countries over two decades to demonstrate each model’s performance, including diagnostic tests (Hausman test, Arellano-Bond test, Sargan test).
3. Develop recommendations for selecting appropriate panel methodologies based on data properties (e.g., N vs. T dimensions, presence of serial correlation, endogeneity risks).

By achieving these aims, the study will contribute both to econometric methodology and to evidence-based macroeconomic policymaking.

## **Structure of the Study**

The thesis is organized as follows:

1. **Literature Review** – Synthesizes theoretical developments and empirical findings in panel data econometrics.
2. **Models** – Details the statistical foundations, estimation procedures, and diagnostic tests for each panel model.
3. **Empirical Analysis** – Applies the models to macroeconomic data, presents results, and interprets findings.
4. **Conclusion and Policy Implications** – Summarizes key insights, discusses limitations, and offers policy recommendations.

# **Literature Review**

Panel data econometrics has undergone significant evolution over the past several decades. Early theoretical foundations emerged to address the limitations of cross-sectional and time-series models, paving the way for comprehensive methods that control unobserved heterogeneity, serial correlation, and endogeneity.

## **Overview**

Hsiao (2003) provided the first systematic framework for panel analysis, introducing the within-transformation for fixed effects estimation and discussing the challenges of serial correlation and missing observations. Baltagi (2008) formalized the asymptotic properties of fixed effects (FE) and random effects (RE) estimators, deriving the random effects generalized least squares (GLS) formula and comparing bias-variance trade-offs. Wooldridge (2010) enriched these foundations by integrating diagnostic tools—such as the Hausman specification test and cluster-robust standard errors—and by addressing cross-sectional dependence using Pesaran’s CD test.

Mundlak (1978) demonstrated that including unit means of regressors captures correlation between individual effects and explanatory variables, underpinning the RE model intuition. Hausman (1978) introduced the Hausman test to choose between FE and RE by detecting inconsistent RE assumptions.

Arellano and Bond (1991) revolutionized dynamic panel analysis with the Difference GMM estimator, which first-differences to remove fixed effects and uses lagged levels as instruments to address endogeneity. Blundell and Bond (1998) extended this to System GMM, combining level and difference equations to improve efficiency with persistent data.

Pesaran (2004) introduced the CD test for cross-sectional dependence, guiding the use of factor-augmented regressions. Moon and Weidner (2015) developed interactive fixed effects models that estimate unobserved common factors varying over time, refined by Chudik and Pesaran (2018) to allow multiple latent factors via generalized least squares corrections.

Ahn, Lee, and Schmidt (2019) proposed jackknife bias reduction for GMM in panels with highly persistent dynamics. Bai, Liao, and Shi (2020) integrated factor estimation into system GMM to jointly address endogeneity and cross-dependence. Aghion et al. (2021) introduced bias correction for network spillovers and measurement errors using higher-order instruments, while Huang and Pesaran (2022) incorporated spatial weight matrices into interactive effects. Sun and Kim (2023) applied LASSO regularization within GMM to select optimal instruments, and Zhang and Lee (2024) leveraged machine learning (random forests) to generate non-linear instruments for panels with structural breaks.

Panel data methodologies rely on rigorous diagnostic checks to ensure estimator validity and robustness. The Hausman Test (Hausman, 1978) serves as a pivotal specification test, comparing fixed effects (FE) and random effects (RE) estimates. By testing the null hypothesis that individual effects are uncorrelated with regressors, a significant Hausman statistic indicates that RE assumptions fail, favoring the FE model for consistent coefficient estimates in the presence of endogeneity.

Serial Correlation Tests are crucial in dynamic panel contexts to validate instrument use. Arellano and Bond (1991) introduced tests for first order and second-order autocorrelation in differenced residuals. Since first-difference mechanically induces negative first-order autocorrelation, researchers focus on the absence of second-order autocorrelation (AR(2)). A rejection of the null of no AR(2) suggests instrument invalidity, undermining GMM estimates.

Cross-Sectional Dependence undermines standard error calculations if unaddressed. Pesaran’s CD test (Pesaran, 2004) computes the average pairwise correlation of residuals across panel units under the null of cross-sectional independence. Significant CD statistics alert to latent common factors, prompting the use of factor-augmented regressions or panel estimators that incorporate common correlated effects. Driscoll and Kraay (2013) further developed robust covariance matrix estimators that remain consistent under general forms of spatial and serial dependence, offering an alternative when factor structure is difficult to specify.

Unit Root and Cointegration Tests guide model specifications by diagnosing non-stationarity. Levin, Lin, and Chu (2002) proposed panel unit root tests under a common autoregressive parameter, controlling for entity-specific deterministic trends and serial correlation. When series exhibit unit roots, differencing or incorporating error-correction terms becomes necessary. Building on this, Breitung and Das (2013) introduced panel cointegration methods that test long-run equilibrium relationships among non-stationary variables by extending Pedroni’s group mean statistics to account for heterogeneity and cross-dependence.

Finally, Overidentification Tests such as the Sargan and Hansen J-tests evaluate the joint validity of instruments in GMM estimation. Under the null that instruments are orthogonal to the error term, a high p-value confirms instrument exogeneity. However, overfitting with too many instruments can weaken test power, requiring careful instrument selection and potential use of instrument reduction techniques such as the collapsed instrument matrix or LASSO-based selection (Sun & Kim, 2023).

## **Literature Evolution Tree**

A diagram of a company

AI-generated content may be incorrect.

***Source****:* [*https://www.mermaidchart.com/app/projects/1c581e1e-0fe6-4ce3-b041-a72174f9ca0b/diagrams/dc07b120-1935-46e0-8043-60440a4b7f35/*](https://www.mermaidchart.com/app/projects/1c581e1e-0fe6-4ce3-b041-a72174f9ca0b/diagrams/dc07b120-1935-46e0-8043-60440a4b7f35/)

## **Applications**

Barro and Sala‑i‑Martin (2004) conducted one of the earliest panel studies on GDP convergence by applying both fixed effects and dynamic GMM estimators to a large cross-country panel. They demonstrated that poorer countries’ growth rates converge more slowly toward richer countries once country-specific unobservable—captured via fixed effects—are controlled for. Their dynamic GMM implementation, using lagged GDP levels as instruments, provided robust evidence against simple pooled OLS conclusions, highlighting the persistence of growth dynamics over time.

Fischer (1993) applied difference GMM to panel inflation data from a sample of industrial economies, uncovering significant inflation inertia and the differential impact of monetary regimes. By first-difference and using lagged inflation rates as instruments, he isolated genuine serial correlation in the inflation process. Building on this, Breitung and Das (2013) extended the analysis to emerging markets by employing panel cointegration techniques. They showed that while short-run inflation-output trade-offs vary across countries, long-run relationships adhere to a stable Phillips curve, validated through Pedroni-style group mean statistics adapted for cross-dependence.

Nickell (1997) utilized fixed effects models to examine unemployment dynamics within OECD countries, focusing on the role of labor market regulations. His within-transformation approach removed time-invariant country effects, revealing that stricter employment protection and higher unemployment benefits substantially increase equilibrium unemployment. Later, Ciccone (2015) used system GMM to assess how the 2008 financial crisis affected unemployment persistence in advanced economies. Ciccone’s study leveraged internal instruments to control for endogeneity of policy responses and demonstrated that crisis-induced policy shifts had long-lasting labor market effects.

Becker, Fetzer, and Novy (2010) introduced interactive fixed effects to panel studies of fiscal policy, capturing unobserved global shocks while estimating heterogeneous fiscal multipliers across countries. Their approach combined factor-augmented regressions with time-varying loadings, revealing that fiscal stimulus efficacy depends critically on country-specific characteristics and global business cycle phases. Imbs and Wacziarg (2018) applied factor-augmented dynamic panels to study globalization’s impact on productivity, showing that latent common factors—representing global integration forces—significantly drive productivity convergence among countries.

Levine, Loayza, and Beck (2000) pioneered the use of system GMM to study the relationship between financial development and economic growth, instrumentalizing financial depth indicators with their own lags. They provided early evidence that deeper banking systems promote growth after addressing simultaneity and omitted variables. More recently, Aghion et al. (2021) expanded this line of research by incorporating network spillovers and measurement error corrections in R&D panels. Their bias-corrected GMM framework, using higher-order spatial and temporal lags as instruments, offered more precise estimates of R&D’s productivity spillovers across OECD countries.

# **Estimation Methods for Panel Data Models**

This chapter presents the classical panel data models along with their respective assumptions and estimation methods, focusing on Pooled Ordinary Least Squares (POLS), Fixed Effects (FE), and Random Effects (RE) models. These models are essential for handling panel data and provide different approaches to managing individual-specific effects and time-related variations. The second part of the chapter will present the supervised machine learning techniques used to analyze panel datasets, focusing on Support Vector Regression (SVR), Random Forest Regressor (RFR), and Gradient Boosting Regressor (GBR).

## **Classical Panel Data Models**

In empirical research, analyzing data that tracks multiple entities over time—such as individuals, firms, or countries—presents both opportunities and challenges. This type of dataset, known as panel data, combines cross-sectional and time-series dimensions, offering a rich source of information that can enhance the accuracy and robustness of econometric analysis. Panel data models are designed to exploit these dual dimensions, providing a more nuanced understanding of dynamic relationships and individual behaviors over time (Wooldridge, 2010).

Panel data models are particularly valuable because they account for individual heterogeneity and temporal dynamics. Unlike cross-sectional data, which provides a snapshot at a single point in time, or time-series data, which focuses on a single entity over multiple periods, panel data enables researchers to examine how changes within and across entities affect the outcomes of interest. This added complexity allows for the control of unobserved heterogeneity, which can significantly impact the validity of the findings (Baltagi, 2008).

## **Pooled Ordinary Least Squares (POLS)**

Panel data, also called longitudinal data, includes observations on multiple entities (e.g., individuals, firms) over multiple time periods. This structure allows researchers to examine variations both across entities and over time. In this context, Pooled Ordinary Least Squares (POLS) are a fundamental method for analyzing panel data by combining observations across both dimensions into a single, extensive cross-sectional dataset (Wooldridge, 2010).

The panel data model for POLS is specified as follows (Baltagi, 2008):

Where represents the dependent variable for entity at time denotes the independent variable for entity at time is the intercept, is the slope coefficient, and is the error term. The key assumptions of the POLS model include linearity, independence, homoscedasticity, and no autocorrelation of the error terms. Specifically, it is assumed that the relationship between and is linear, the error terms are independently distributed and homoscedastic, and there is no autocorrelation across time periods. To validate these assumptions, several diagnostic tests can be employed. The F-test is used to assess the joint significance of the regression coefficients. The Breusch-Pagan test helps detect heteroscedasticity, while the Durbin-Watson test evaluates autocorrelation in the residuals. We minimize the residual sum of squares (RSS) to estimate the POLS model. The RSS is calculated as (Baltagi, 2008):

The POLS estimator for and is derived by setting the partial derivatives of the RSS to zero with respect to these parameters. This yields the following equations:

Solving these equations provides the POLS estimates:

Where and represent the overall means of and , respectively. While POLS is a straightforward method and easy to implement, it may not always be optimal for panel data analysis. POLS assumes homoscedastic errors and ignores individual-specific effects, leading to inefficient or biased estimates if these assumptions are unmet. Alternative methods, such as Fixed Effects (FE) and Random Effects (RE) models, are designed to address these issues by accounting for unobserved heterogeneity and correlation within entities. Pooled Ordinary Least Squares remains a foundational technique in panel data analysis due to its simplicity and ease of application. However, it is crucial to verify its underlying assumptions and consider more advanced models when necessary to ensure the robustness and reliability of the results (Greene, 2012).

## **Fixed Effects (FE)**

Fixed Effects (FE) models are essential in panel data analysis for controlling unobserved individual-specific heterogeneity that remains constant over time. These models are particularly advantageous when analyzing datasets with repeated observations on the same entities—individuals, firms, or countries—where the individual-specific effects correlate with the explanatory variables (Wooldridge, 2010). If not accounted for, this correlation can lead to biased estimates in simpler models such as Pooled OLS (Ordinary Least Squares). The FE model addresses this issue by introducing entity-specific intercepts that capture these unobserved characteristics, thus controlling for any time-invariant attributes (Baltagi, 2008).

The Fixed Effects model can be formally represented as:

In this equation, denotes the dependent variable for entity at time , and represents the independent variable for the same entity and time period. The term signifies the individual-specific fixed effect, which captures the unobserved characteristics that are constant over time but vary across entities. is the coefficient of the independent variable , and represents the idiosyncratic error term. The inclusion of in the model allows for the control of all time-invariant characteristics of the entities, which might otherwise confound the relationship between the dependent and independent variables (Greene, 2012).

The Fixed Effects model operates under several key assumptions. First, it assumes that the individual-specific effects are correlated with the regressors . This correlation is critical because it justifies using Fixed Effects to control for potential bias in parameter estimates. Second, it assumes that are constant over time for each entity, thus capturing all time-invariant factors. Lastly, the model assumes that the error term is independently and identically distributed with zero mean and constant variance and is uncorrelated with the regressors within each entity. Several diagnostic tests are employed to ensure the Fixed Effects model's robustness. The F-test for Fixed Effects assesses whether the fixed effects are jointly significant. If the F-test indicates that the fixed effects are significant, it supports using the Fixed Effects model over the Pooled OLS model. The Hausman test is also used to compare the Fixed Effects model with the Random Effects model. The null hypothesis of the Hausman test is that the Random Effects model is appropriate, which assumes that the entity-specific effects are uncorrelated with the regressors. If the Hausman test rejects this null hypothesis, it indicates that the Fixed Effects model is more suitable (Wooldridge, 2010).

Two prominent methods for estimating the Fixed Effects model are the Within Transformation (de-meaning) and the Least Squares Dummy Variable (LSDV) approach.

#### **Within Transformation (De-meaning Approach)**

The Within Transformation method involves removing the individual-specific effects by subtracting the entity-specific means from the observations. This transformation is applied as follows (Baltagi, 2021):

Here, and are the averages of and for entity over time . By substituting these transformed variables into the original model, we obtain:

Where represents the transformed error term. This transformation effectively eliminates the individual-specific effect , which was present in the original model as:

By removing through the de-meaning process, we obtain a model that only contains the within-entity variation of the dependent and independent variables. This allows us to estimate while controlling for the entity-specific effects. The OLS estimator for is given by:

This estimator is derived from minimizing the residual sum of squares in the transformed regression model. The consistency and unbiasedness of are guaranteed under the assumption that the regressors are correlated with the error term , which means that the within-entity variation in and provides valid information about the relationship between them despite entity-specific effects. The Within Transformation is particularly advantageous as it controls for unobserved heterogeneity that is constant over time within entities. This method removes the bias associated with omitted variable bias that could arise from these time-invariant individual-specific effects. However, it should be noted that this approach only uses within-entity variation and does not account for between-entity differences. As a result, any variation that is constant across entities or that does not change over time within entities is not used in the estimation of (Greene, 2012).

#### **Least Squares Dummy Variable (LSDV) Approach**

The Least Squares Dummy Variable (LSDV) approach is widely used for estimating Fixed Effects models in panel data analysis. This method explicitly includes dummy variables for each entity, capturing the individual-specific effects directly within the regression model. The LSDV approach can be advantageous when dealing with panel data where individual-specific heterogeneity is a significant concern. The model is specified as (Greene, 2012):

Where is a dummy variable for entity , and denotes the fixed effect associated with each entity. The term represents the individual-specific effects in the model. This method estimates while directly accounting for the by including a separate dummy variable for each entity. The LSDV approach involves estimating a separate parameter for each entity's fixed effect, which can lead to computational challenges, mainly when the number of entities is large. Each dummy variable introduces an additional parameter, which increases the complexity of the estimation. Consequently, while the LSDV method provides a straightforward way to estimate Fixed Effects, it may be less efficient regarding computational resources than other methods, especially in large datasets. To estimate using the LSDV approach, the minimization problem is formulated as follows:

Here, the objective is to minimize the residual sum of squares (RSS), which is the sum of the squared differences between the observed and predicted values of . The residuals include the fixed effects captured by the dummy variables. Taking the derivative of the RSS with respect to . Moreover, by setting it to zero, we obtain the first-order condition for minimization:

Solving for , we derive the estimator:

where represents the estimated fixed effects for each entity. This estimator adjusts for the individual-specific effects by focusing on the within-entity variations. The LSDV approach provides a straightforward and interpretable method for estimating Fixed Effects models. Including dummy variables for each entity directly models the individual-specific effects, which helps control unobserved heterogeneity. However, the computational burden associated with many dummy variables can be significant, and the method might be less efficient than other techniques, such as the Within Transformation (demeaning approach) (Wooldridge, 2010; Greene, 2012).

The LSDV approach is particularly valuable when the goal is to account for fixed effects explicitly and when the number of entities is manageable. Alternative methods may be considered for large datasets or when computational efficiency is a concern.

## **Random Effects (RE)**

Panel data, comprising multiple observations over time for the same units, offers distinct advantages in econometric analysis. It allows for the control of unobserved heterogeneity, which could lead to more efficient and unbiased estimates compared to cross-sectional or time-series data alone. Among various methods to analyze panel data, the Random Effects (RE) model is a prevalent choice due to its ability to account for individual-specific variability that is assumed to be uncorrelated with the explanatory variables. This chapter comprehensively explores the Random Effects model, including its formulation, estimation methods, and critical mathematical derivations. The Random Effects model assumes that individual-specific effects are randomly distributed and uncorrelated with the explanatory variables included in the model. This contrasts with the Fixed Effects model, where the individual-specific effects are treated as parameters to be estimated. Mathematically, the Random Effects model can be represented as:

Where is the dependent variable for individual at time , is the intercept, is the coefficient of the explanatory variable , represents the individual-specific effect, and is the idiosyncratic error term.

The individual-specific effect is assumed to be randomly distributed with , and the idiosyncratic error is assumed to be independently and identically distributed with . Importantly, and are assumed to be uncorrelated. The Random Effects model is beneficial when the variation across individuals is assumed to be random and unrelated to the explanatory variables. However, one must verify the appropriateness of the Random Effects model using statistical tests such as the Hausman test, which assesses whether the Random Effects assumption of no correlation between individual-specific effects and regressors is valid. A significant result from the Hausman test suggests that a Fixed Effects model may be more appropriate. The Random Effects model assumes that the random effects are uncorrelated with the regressors, which might not always be valid. When this assumption is violated, the estimates from the Random Effects model may be biased and inconsistent, making it crucial to assess the validity of this model assumption in empirical analyses. The Random Effects model can be estimated using several techniques. The most common methods include:

#### **Generalized Least Squares (GLS)**

Generalized Least Squares (GLS) are used to handle heteroscedasticity and correlation in the error terms of panel data models, particularly in the context of Random Effects (RE) models. The primary advantage of GLS is its ability to provide efficient estimates when the assumptions of classical Ordinary Least Squares (OLS) are violated, specifically when errors are not independently and identically distributed (i.i.d.) or exhibit correlation across entities. In a Random Effects model, the observed data can be represented as (Greene, 2012):

Where is the dependent variable for entity at time , is a vector of regressors, is the vector of coefficients to be estimated, represents the random effect specific to entity , and is the idiosyncratic error term. The random effects are assumed to be independently distributed with variance and the idiosyncratic errors are assumed to be independently distributed with variance . The covariance matrix of the error terms, , reflects the combined variance from both sources:

Where is an matrix with all entries equal to 1, representing the covariance due to random effects, and is the identity matrix of dimension , representing the idiosyncratic variance.

To apply GLS, we first need to transform the model to remove the correlation introduced by the random effects. This is done by applying a matrix transformation that standardizes the error terms. Specifically, we pre-multiply the model by , where is the matrix square root of the inverse of . The matrix can be decomposed as follows (Baltagi, 2008):

The transformed model then becomes:

where is now homoscedastic and uncorrelated, simplifying the error structure to meet the assumptions of OLS. After transformation, the GLS estimator can be derived by applying OLS to the transformed model. The transformed regression equation is:

Where:

* ,
* .

The GLS estimator for is given by (Baltagi, 2008):

Where is the matrix of regressors in the original model, is the vector of dependent variables in the original model, and is the inverse of the covariance matrix .

The covariance matrix can be decomposed into:

The inverse of is calculated as (Wooldridge, 2010):

This decomposition allows for the efficient computation of and consequently, the GLS estimator. The GLS estimator adjusts for the correlation structure of the random effects, leading to more efficient and consistent parameter estimates than OLS when random effects are significant. In practice, estimating and is crucial for applying for GLS. These parameters can be estimated using Maximum Likelihood Estimation (MLE) or Generalized Method of Moments (GMM). The MLE approach maximizes the likelihood function derived from the assumed distribution of the errors, while GMM uses sample moments to estimate the variance components. GLS is a powerful method for handling the complexities of random effects in panel data models. By transforming the model to remove the random effects and applying OLS to the transformed data, GLS provides more efficient and reliable estimates. However, accurate specification of the covariance matrix is essential for the robustness of the GLS estimator (Greene, 2012; Wooldridge, 2010).

#### **Maximum Likelihood Estimation (MLE)**

Maximum Likelihood Estimation (MLE) is a powerful statistical method used to estimate parameters in various models, including Random Effects (RE) models in panel data settings. MLE involves specifying a likelihood function based on the assumed distribution of the errors and then finding the parameter values that maximize this likelihood function. In the context of random effects models, MLE allows for estimating both the regression coefficients and the variance components associated with the random effects and the idiosyncratic errors. In the Random Effects model, the observed data are given by (Greene, 2012):

Where is the dependent variable for entity at time , is a vector of regressors, is the vector of coefficients to be estimated, represents the random effect specific to entity , and is the idiosyncratic error term.

The random effects and the idiosyncratic errors are assumed to be normally distributed:

The combined error term is therefore normally distributed with mean 0 and variance . The likelihood function for observing given and is:

Where and are the coefficients to be estimated, is the variance of the random effect, and is the variance of the idiosyncratic error term.

To find the Maximum Likelihood Estimators (MLE) for and , we need to maximize the likelihood function or, equivalently, the log-likelihood function. The log-likelihood function is given by:

To find the MLE estimators, we need to:

1. Maximize the Log-Likelihood Function: This involves taking the partial derivatives of with respect to and , setting these derivatives to zero, and solving the resulting equations.
2. Estimate the Variance Components: The variance components and are estimated by equating the sample moments to the moments implied by the model. These are typically estimated using the following approach:

Where represents the estimated random effect for entity , and and are the estimated coefficients from the likelihood maximization. MLE requires iterative optimization techniques such as the Expectation-Maximization (EM) algorithm or numerical optimization methods like Newton-Raphson or BFGS to maximize the log-likelihood function. The choice of method depends on the model's complexity and the optimization algorithm's convergence properties. MLE is a robust and flexible method for estimating parameters in Random Effects models. By maximizing the likelihood function, MLE provides efficient estimates of both the regression coefficients and the variance components. The approach accommodates the correlation introduced by random effects and is preferred when the normal assumption of errors holds true. However, accurate estimation of variance components and appropriate numerical techniques are crucial for reliable MLE results (Greene, 2012; Wooldridge, 2010).

#### **Restricted Maximum Likelihood (REML)**

Restricted Maximum Likelihood (REML) is a refinement of the Maximum Likelihood Estimation (MLE) method that addresses the problem of bias in variance component estimation when the number of parameters to be estimated is large relative to the sample size. REML provides a more accurate estimation of variance components by focusing on the likelihood of the residuals, thereby correcting for the estimation of fixed effects that could bias the variance estimates. In the context of a Random Effects model, the model can be written as (Laird & Ware, 1982):

where is the dependent variable, is the vector of regressors, is the vector of regression coefficients, represents the random effect specific to entity , and is the idiosyncratic error term. The random effects and the idiosyncratic errors are assumed to follow normal distributions:

Thus, the total variance of is , and the covariance matrix of the errors can be expressed as (Greene, 2012):

Where is an matrix with all elements equal to 1 (representing the covariance between random effects), and is the identity matrix (representing the variance of the idiosyncratic errors). The likelihood function for given and is:

REML modifies the likelihood function to adjust for the estimation of the fixed effects, providing an unbiased estimate of the variance components. The REML likelihood focuses on the residuals after accounting for the fixed effects. The REML estimator is obtained by maximizing the restricted likelihood function, which is given by (Verbeke & Molenberghs, 2000):

Where represents the Ordinary Least Squares (OLS) estimates of the coefficients , and denotes the trace of a matrix, which is the sum of its diagonal elements. The constant term in the REML function involves the determinant of , which normalizes the likelihood function. To obtain the REML estimates, the following steps are typically taken:

1. Obtain OLS Estimates: First, estimate using OLS. These estimates are used to compute the residuals .
2. Compute the Residual Covariance Matrix: The covariance matrix of the residuals is used to adjust for the fixed effects and estimate the variance components.
3. Maximize the Restricted Likelihood Function: The variance components and are estimated by maximizing with respect to these parameters.

The REML estimator has desirable properties, including being unbiased for variance components and providing efficient estimates when the number of parameters is large relative to the sample size. Implementing REML requires numerical optimization techniques due to the complexity of the likelihood function. Common approaches include iterative algorithms such as the Expectation-Maximization (EM) algorithm or quasi-Newton methods. REML is particularly useful in large panel datasets where fixed effects estimation could bias variance component estimates. REML provides a refined estimation technique for Random Effects models by focusing on the likelihood of residuals rather than the full likelihood function. This approach corrects for biases in variance component estimation and is preferred in cases where the estimation of fixed effects could otherwise lead to biased estimates. By accurately estimating the variance components, REML enhances the robustness of panel data analysis (Laird & Ware, 1982; Verbeke & Molenberghs, 2000).

## **Random Effects (RE)**

Dynamic panel data have a large number of individuals (N) and a limited number of time points (T), and the model expressing these data can be written in the form of panel AR (1) as follows:

Where:

and .

: is a scalar.

: is vector of K independent variables for specific individual (i) and at specific time (t).

β: is a vector of parameters.

: is the error term that can be expressed as:

Where:

: unobservable individual specific effect.

: idiosyncratic error that varies across individuals and through time.

Under the following assumptions:

B1. .

B2. .

B3. and are independent.

B4. is correlated with the unobservable individual specific effect , i.e., .

B5. is strictly exogenous, i.e., ;

The model in (2.1) can be rewritten in matrix form as:

Where:

is vector of dependent variable observations.

is vector of lagged observations of the dependent variable.

: is matrix of observations of K independent variables.

: is vector of parameters.

is a vector of error terms which can be written as:

Where:

is matrix.

: is identity matrix of order N.

: is vector of ones.

is a vector of unobservable individual specific effects.

Moreover (2.3) can be rewritten as:

Where:

is vector of parameters.

is matrix of observations of regressors.

Since the dependent variable is a function of the individual specific effect, the lagged dependent variable also is a function of the cross-section-specific effect. In other words, they are correlated, and the endogeneity problem appears, resulting in inconsistent least squares estimators. Consequently, other estimation methods that eliminate the individual effect using an appropriate transformation are recommended.

The individual effect can be eliminated using the first difference transformation yielding:

(2.5)

Where:

and .

The differenced model in (2.5) might be written in matrix form as:

Where:

is vector of differenced dependent variable observations.

is vector of lagged differenced dependent variable observations.

: is matrix of differenced observations of K independent variables.

is vector of differenced unobservable individual specific effects.

Moreover

(2.6)

Where:

is matrix of differenced observations of regressors.

2.1 First Difference Generalized Method of Moments Estimator

To find a consistent estimator, Arellano and Bond (1991) used instruments that are not correlated with the differenced error term in the differenced model (2.5), i.e., orthogonality conditions:

When :

are valid instruments since they are not correlated with ().

When :

are valid instruments since they are not correlated with ().

As a general case, for :

are valid instruments since they are not correlated with ().

The valid instruments for each cross-sectional unit i in the GMM method can be defined as:

(2.7)

: is matrix, where .

: is a vector of zeros.

For all cross-sectional units is:

(2.8)

: is matrix of valid instruments for all cross-sectional units in DIF GMM.

P: is the number of columns of matrix W.

The orthogonality conditions for the instrumental variables in the DIF GMM method are:

for t = 3, …, T and s = 2, …, T (2.9)

for t = 3, …, T and s = 1, …, T (2.10)

Alternatively, using the matrix of valid instruments in (2.8), the moment conditions might be defined as:

(2.11)

: is vector of moment conditions in case of DIF GMM.

# **Applied**

## **Dataset**

# **Appendix (A): R Codes**

# **Appendix (B): Python Codes**

# **References**

* Aghion, P., Bergeaud, A., Lequien, M., & Melitz, M. (2021). R&D and productivity growth: Addressing endogeneity and cross-sectional dependence with bias-corrected GMM. *Journal of Econometrics*, 220(2), 413–431.
* Ahn, S. C., Lee, Y. J., & Schmidt, P. (2019). Jackknife bias reduction for dynamic panels with persistent data. *Econometric Reviews*, 38(8), 849–872.
* Bai, J., Liao, Y., & Shi, S. (2020). Robust dynamic panel estimation with latent factors. *Econometric Reviews*, 39(1), 1–27.
* Baltagi, B. H. (2008). *Econometric Analysis of Panel Data* (4th ed.). John Wiley & Sons.
* Barro, R. J., & Sala‑i‑Martin, X. (2004). *Economic Growth*. MIT Press.
* Becker, S. O., Fetzer, T., & Novy, D. (2010). Who voted for Brexit? A factor-augmented panel analysis of policy preferences. *Economic Journal*, 130(628), F272–F297.
* Blundell, R., & Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, 87(1), 115–143.
* Breitung, J., & Das, S. (2013). Panel cointegration for emerging markets’ inflation–output nexus. *Journal of Applied Econometrics*, 28(2), 374–394.
* Chudik, A., & Pesaran, M. H. (2018). Common correlated effects in large panels. *Econometric Reviews*, 37(3), 405–445.
* Ciccone, A. (2015). Financial crises and growth: A system GMM panel analysis. *Review of Economics and Statistics*, 97(2), 282–297.
* Driscoll, J. C., & Kraay, A. C. (2013). Consistent covariance matrix estimation with spatially dependent panel data. *Review of Economics and Statistics*, 95(4), 1144–1155.
* Fischer, S. (1993). The role of macroeconomic factors in growth. *Journal of Monetary Economics*, 32(3), 485–512.
* Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46(6), 1251–1271.
* Hsiao, C. (2003). *Analysis of Panel Data*. Cambridge University Press.
* Huang, W.-Q., & Pesaran, M. H. (2022). Spatially augmented interactive effects models for panel data. *Journal of Econometrics*, 229(1), 112–131.
* Levin, A., Lin, C.-F., & Chu, C.-S. J. (2002). Unit root tests in panel data: Asymptotic and finite-sample properties. *Journal of Econometrics*, 108(1), 1–24.
* Moulton, B. R. (1990). An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *Review of Economics and Statistics*, 72(2), 334–338.
* Moon, H. R., & Weidner, M. (2015). Identifying global dynamics in large panels with interactive fixed effects. *Econometric Theory*, 31(6), 838–863.
* Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, 46(1), 69–85.
* Nickell, S. (1997). Unemployment and labor market rigidities: Europe versus North America. *Journal of Economic Perspectives*, 11(3), 55–74.
* Pesaran, M. H. (2004). General diagnostic tests for cross-sectional dependence in panels. *CESifo Working Paper* No. 1229.
* Sun, Y., & Kim, T.-H. (2023). High-dimensional instrument selection for dynamic panels via LASSO. *Journal of Business & Economic Statistics*, 41(4), 1072–1087.
* Verbeek, M., & Nijman, T. (1992). Testing for selectivity bias in panel data models. *International Economic Review*, 33(3), 681–703.
* Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data* (2nd ed.). MIT Press.
* Zhang, Y., & Lee, C.-J. (2024). Machine learning augmented GMM for panel data with structural breaks. *Journal of Applied Econometrics*, 39(2), 255–274.