Demonstrating A Python-based Cross-platform Novel Tool for Panel Data Analysis In Econometrics

A Thesis

submitted to

Indian Institute of Science Education and Research Pune
in partial fulfillment of
the requirements for the BS-MS Dual Degree Programme

by

Harsh Jinger

Indian Institute of Science Education and Research Pune
Dr. Homi Bhabha Road,
Pashan, Pune 411008, INDIA.

April, 2025

Supervisor: Dr. Shailesh Rastogi

Harsh Jinger

Certificate

This is to certify that this dissertation entitled **Demonstrating A Python-Based Cross-Platform Novel Tool For Panel Data Analysis In Econometrics** towards the partial fulfilment of the BS-MS dual degree programme at the *Indian Institute of Science Education and Research, Pune* represents study/work carried out by *Harsh Jinger* at *Symbiosis Institute of Business Management, Nagpur (Symbiosis International University, Pune)* under the supervision of Dr. Shailesh Rastogi, Director & Professor, Symbiosis Institute of Business Management, Nagpur, during the academic year 2024-2025.

Dr. Shailesh Rastogi

Dr. Bejoy Thomas

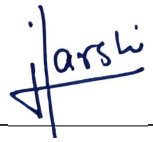
This thesis is dedicated to my mom and my dad for their unshakable belief in me and my endeavours—always acting as my shield; to my brother, whose unwavering belief in me has always been the silver lining after all the dark storms; and most importantly, to my dearest friend, Simmi, who never left my side since the day we met and has been my northern star in the darkest of nights ever since.

Declaration

I hereby declare that the academic matter embodied in the report, duly entitled

"Demonstrating A Python-Based Cross-Platform Novel Tool For Panel Data Analysis In Econometrics,"

are the results of the work carried out by me at the Symbiosis Institute of Business Management (SIBM) Nagpur, Symbiosis International University (SIU) Pune, under the supervision of Dr. Shailesh Rastogi, as a part of partial fulfillment of the BS-MS Programme at Department of Humanities and Social Science, IISER Pune and the same has not been submitted elsewhere for any other degree. Wherever others have contributed, every effort is made to indicate this clearly, with due reference to the literature and acknowledgment of collaborative research and discussions.



Harsh Jinger

20191022

Contents

List of Tables

List of Figures

Abstract

Background

Panel data is prevalent in the contemporary economic landscape because of the digitalization of academic and financial resources. However, advanced econometric models in panel data analysis are out of the reach of early-stage and under-funded researchers due to technical and institutional barriers. Panel Data Analysis requires special software, such as STATA and SPSS, which are considered expensive resources that are helpful for a competitive advantage in well-funded institutions and for senior researchers. While low-cost alternatives to these tools exist, they are often complex, tedious, and inaccessible. Even the proprietary tools are tedious and cumbersome for complex analyses since they are designed to perform one study at a time.

Objectives

This thesis aims to develop and demonstrate a free and open-source, cross-platform, novel Python-based tool for panel data analysis that automates the tedious procedure to make it efficient for senior researchers and removes any technical and cost barriers for early-stage researchers and institutions with limited resources, thereby empowering them to use Panel Data Analysis to employ advanced econometric models to generate higher quality insights from the panel data.

Methodology

The tool developed in this thesis can perform panel data analysis in two modes: a) Exploratory and b) Autonomous. In Exploratory mode, the tool can test hundreds of models, exploring all possible permutations of relationships between variables under scrutiny. This mode can identify relevant econometric models to investigate any meaningful relationship between the variables further. In Autonomous mode, the tool can perform descriptive analyses, model diagnostics, endogeneity testing, and selection between fixed-effects, random-effects, instrumental, or dynamic models and provide a step-by-step analysis report.

Theoretical Framework

The theoretical grounding of this thesis can be described in two parts: a) Econometric theories that dictate the workflow or the algorithm of the tool for panel data analysis, and b) Innovation adoption theories that explain the dominance of current tools and conditions required for the acceptance of the novel tool, thereby, also directing the development of the tool. Econometric theories such as panel data modelling, endogeneity correction using instrumental variables, and statistical tests to test the validity of the analysis form the theoretical basis for this tool's functionality. Innovation adoption theories used in this thesis include the Resource-Based View (RBV), the Technology Acceptance Model (TAM), the Unified Theory of Acceptance and Use of Technology (UTAUT), and the Diffusion of Innovation Theory (DIT).

Results

This tool demonstrates the execution of econometric analysis on two real-world panel datasets in a transparent and replicable manner. This tool automates the process of panel data analysis, independent of the platform, at a low cost, with enhanced usability and reduced barriers to use.

Conclusion

The thesis contributes a theoretically informed tool with enhanced functionality that democratizes access to advanced economic analysis of panel data while providing a platform to modify and extend the capabilities of open-source software and cross-platform capabilities for widespread institutional and individual adoption in academic discourse.

Acknowledgment

This thesis was possible due to the significant support of IISER Pune and Symbiosis International University, Pune. I appreciate IISER Pune for accommodating my research interest by allowing this project and SIU Pune for hosting me for the duration of this project, initially at SIBM Pune and later at SIBM Nagpur.

I am grateful to my supervisor, Dr Shailesh Rastogi, under whose guidance I was able to complete this project and who also encouraged me to explore my research interests beyond just the fundamental requirements of this project. I am grateful to Dr Bejoy Thomas (TAC Member, Subject Expert) for his constructive criticism and Dr Pooja Sancheti (TAC Member, HoD) for her academic assistance throughout this project.

I thank Ms Bhakti Agarwal for providing valuable resources and relevant data for my project. I am grateful to Dr Rahul Gautam and Ms Aman Pushp for introducing me to this subject's key theoretical research aspects and gently guiding me through its complexities. I am grateful to Ms Pracheta Tejasmayee, Ms Samiksha Kashyap and Mr Jalaj Gupta for the brainstorming sessions and all the logistics support. I am grateful to Dr. Jagjeevan Kanoujiya for his moral support and nuggets of wisdom.

Lastly, I also appreciate the help of all the non-academic staff for their services and assistance throughout this project.

Chapter 1

Introduction

1.1 Motivation

While exploring different domains for my MS thesis project, I gravitated towards economics because it is central to the routine sustenance of modern societies and nations. A mere inquisitive look at this domain reveals the complexity at every level of interaction among customers, traders, governments, or regulators. During the pandemic, the economy was a point of intense debate and discussion across societies worldwide, whether it was a question of acquiring vaccines (Hale et al. (2022); Sohail (2023)) or providing rations to vulnerable people (GoI (2022)). As the pandemic subsided, the subject of economic recovery took center stage in mainstream debates (Shih (2020)), along with a significant discourse on the diversification of supply chains. Amidst this backdrop, a few banks failed in the Western Hemisphere, leading to an environment of turmoil and fear among citizens, ultimately resulting in dwindling trust in economic systems (Center (2023)). These phenomena evoked human reactions that could have tangible impacts, such as political turmoil, poverty, food shortages and wars. Therefore, I wanted to inspect this domain with academic rigour. Given the complexities of most economic systems, the following quote from Joan Robinson, a 20th-century British economist, underscores the importance of studying economics with scepticism rather than accepting the traditional wisdom at face value.

> "The purpose of studying economics is not to acquire a set of ready-made answers to economic questions, but to learn how to avoid being deceived by economists."

This quote emphasizes the need to critically examine the economic questions in light of adequate context and the limitations of the theory discussed. Over many decades, statistical testing has become a standard tool for testing the validity of

any hypothesis across disciplines of scientific inquiry, including economics and finance Hsiao and Wise (2006). Econometrics applies statistical and mathematical concepts to economic data to test and validate or falsify hypotheses Arellano (2016). Therefore, based on various economic data types, such as time-series data, econometrics employs different techniques and tests to extract adequate insights Das and Das (2019). One such data type is Panel Data, which forms the central basis of this project.

1.2 Panel Data

Panel Data, or longitudinal data, can be described as a structured dataset that records observations associated with multiple entities (such as firms, households, or countries) over regular intervals (such as monthly, quarterly or annually) for an extended period while keeping the set of entities constant throughout the period (Arellano (2003)). For instance, the dataset comprising wages of labourers in different companies over the past 50 years is an example of Panel Data. A few real-world examples can be the World Bank's World Development Indicators (WDI), which tracks many different social, economic and environmental indicators of different over time; the historical data of the National Stock Exchange that has a record of stock prices, volumes traded, profit and loss percentages of the firms listed on NSE; and even the historical meteorological data of different cities in a country or a region. The generic structure of Panel data is shown in Table 1.1.

Sr. No.	Entity	Year	Parameter A	Parameter B	Parameter C	Parameter D
1	AB	2012	12	57	91	23
2	AB	2013	13	61	88	27
3	AB	2014	14	53	95	32
4	CD	2012	16	65	89	37
5	CD	2013	17	69	92	41
6	CD	2014	18	58	87	29

Table 1.1: Example of a Panel Data

Panel data is called *balanced panel data* if all the observations associated with all entities are recorded over all intervals throughout the period of observations. Panel Data that has missing values is considered *unbalanced panel data*. Unbalanced panel data can result from minor inconsistencies in data or when the set of entities or parameters associated with entities change over time (Table 1.3) (Baltagi (2008)). A proper analysis is possible on balanced panel data only; therefore, unbalanced panel data is generally converted to balanced panel data before analy-

sis using different techniques such as insertions, deletions or other modifications (Abdullah et al. (2022)). Henceforth, Panel Data only refers to the balanced panel data in this project unless specified otherwise.

ID	Age	Income	Education Level	City
1	25	45000	Bachelor's	NYC
2	32	52000	Master's	LA
3	41	61000	PhD	Chicago
4	29	48000	Bachelor's	Houston
5	36	53000	Master's	Miami

Table 1.2: Cross-sectional data

ID	Entity	Year	Parameter A	Parameter B
1	AB	2012	12	45
2	AB	2013	13	
3	AB	2015		55
4	CD	2012	16	50
5	CD	2014	17	
6	EF	2013		60
7	EF	2014	21	62

Table 1.3: Unbalanced panel data

In a modern context where most of our economic and financial systems are digitised, it becomes easier to record attributes of firms regularly in a structured format such as Panel Data, which provides aspects of both time-series and cross-sectional data(Sohail (2023)). Therefore, Panel Data Analysis can be extensively used to extract significant insights. Anecdotal discussions would lead you to believe that research papers with at least some quantitative components are prioritised over the ones with solely qualitative components. Therefore, there is a positive bias towards research with quantitative analysis. As time-series-oriented research becomes commonplace and its limitations become apparent, researchers are increasingly turning to panel data analysis to mitigate these limitations and introduce novelty into their work, resulting in more reliable and decisive outcomes Aptech (2018). Additionally, secondary data on financial firms structured as panel data is becoming more accessible through sources such as Bloomberg Terminal Bloomberg (2023), RBI Publications, CMIE Prowess Datasets, stock market data, cryptocurrency price datasets Kaggle (2023), and Thomson Reuters Datastream Staff (2023). These factors, combined with an increasing focus on data analy-

sis, make a compelling case for panel data analysis to become more prevalent in econometric research.

1.3 Challenges of Panel Data Analysis

Despite this, the challenges associated with its adoption are multifold. First, these analyses are tedious and require special software such as STATA and SPSS, which are proprietary and expensive. Table 1.4 records the prices of these software for one user per year and compares it to MATLAB, which is a proprietary computational software that is used by researchers in the natural science domain to depict the disparity. Furthermore, in a developing country like India, where research grants are scarce and the research spending for finance and economics is concentrated in very few top-tier institutes, many under-funded institutes can't afford the premium cost of such software. Therefore, for Indian researchers outside of these few research institutes, such expensive software becomes unaffordable, leading to a significant disparity in the methodologies they can employ in their research studies. While free and open-source software (FOSS) like Gretl exists, it is not widely adopted due to insufficient awareness and adequate institutional support. A lack of user-friendly and cumbersome UI along with fewer advanced features, make it less attractive for the academia at large.

ID	Software	Price per License per annum
1	Stata	$925
2	Eviews	$2370
3	SPSS	$17000
4	MATLAB	$300

Table 1.4: Comparison of Annual Licence Prices of Software per annum per user

Second, Panel Data Analysis can be very repetitive and tedious using traditional statistical software because of the numerous possibilities for simple models. For instance, if a researcher wishes to test a model, she has to perform a calculation for a fixed effects model, a random effects model, a dynamic model, and then with a permutation of control variables for a linear relationship. This process has to be repeated for non-linear relationships as well. These calculations would also consist of tests necessary to comprehensively understand the statistical results. In complex analyses, permutations of such analyses can reach up to hundreds. Therefore, given the time and effort required to perform all the calculations, the researchers prefer to examine a subset of possibilities, which can compromise with the robustness and reliability of the results. While STATA contains an option of do-files which is used for automating these calculations, they require manual setup

for each test, which makes them cumbersome for complex analyses. They also lack the capability to modify the results, making it tedious to identify and differentiate viable models from insignificant ones.

1.4 A Novel Tool for Panel Data Analysis

While adopting programming languages such as Python and R would be a natural solution to most of these problems, the lack of proficiency in programming remains a barrier to their adoption. Programming languages are more flexible, allowing greater automation and reproducibility in econometric analysis. Therefore, R remains a popular choice of computation amongst senior researchers because of their experience with R, the availability of statistical modules and academic acceptance. Despite its advantages, such as the availability of readymade statistical modules, R is considered outdated in many aspects, such as its unstructured syntax, limited integration with other technologies for plotting, lack of scalability for complex analyses and outdated user interface for its compilers.

In contrast, many other scientific disciplines have transitioned to Python and Julia because of their beginner-friendly and structured syntax, more versatile applications, support for newer technologies for extracting and compiling data, more modern user interfaces of their compilers, extensive documentation and community, version control features, and widespread adoption in the industry and academia alike. Python, primarily, is used extensively for automating processes, employing machine learning techniques and data science applications. Python is capable of creating scripts and modules that do not require any knowledge of Python for its usage while harnessing all the advantages listed above. A long-term solution to this problem lies in initiatives such as curriculum reforms, specialised training programs enhancing research funding and providing institutional support for Free and Open-source software for more affordable, accessible and flexible econometric analysis. However, for the specific case of Panel Data Analysis, a novel cross-platform Python-based tool can be created to bridge the gap that I have highlighted above, providing early-stage researchers with a tool that is affordable and accessible while providing senior researchers with a tool that automates the tedious process of panel data analysis to make research investigation more efficient and comprehensive.

1.5 Thesis Structure

This project aims not to create a complete replacement for these existing tools but to democratise and automate the process that is widely considered an optimal

method of panel data analysis. The next section includes a theoretical background for this project which aids in key decisions required for this project. The following section on methodology introduces the panel data analysis techniques formally and describes the workflow that is employed for creating the script and a complete description of the Python code itself. The use case of this tool is demonstrated in the next section using three examples of research questions. The last section discusses the limitations of this tool and opportunities for further research and reflects on my personal learnings from this project.

Chapter 2

Theoretical Framework

The thesis is based on two sets of literature covering the financial and technological aspects of the tool. Literature about the former aspect of the proposed tool, i.e., the financial aspect, determines the flow and algorithm of the tool by grounding it in mainstream theories about econometric modelling, selecting adequate model types (Fixed-effects, Random-effects or Instrumental Variables) and performing diagnostic tests on these models to ensure the academic validity of their insights. Literature about the technology aspect of the tool provides key insights into the dominance of the existing tools despite their limitations, the conditions required for this tool to be considered competitive, and the work that has to be done beyond this thesis to make it accessible and drive its adoption at institutional levels. This literature justifies the development of this tool at this stage and dictates the meaningful choices about human-computer interaction necessary for the design of this tool.

2.1 Financial Literature

2.1.1 Hausman Test: Fixed Effects Vs. Random Effects

Panel Data mainly contains heterogeneity between entities, which implies that there are unobserved factors about specific entities called Individual Effects. Whether these individual effects are uncorrelated with regressors is often the first step after descriptive analysis in panel data analysis. The random effects model assumes that if unobserved heterogeneity is independent of the regressors, then the random effects model is efficient compared to the fixed effects model. The fixed effects model is used when it has been established that unobserved heterogeneity factors are not exogenous. This distinction is formalised in Hausman 1978: The Hausman test compares the estimates of both the fixed effects and the random effects model

to identify if the difference between them is statistically significant. Under the null hypothesis of no correlation between the unobserved heterogeneity and the regressors, if the Hausman statistic is statistically significant, the assumption for the random effects model is violated, and the fixed effects model is used for further analysis. This best practice is codified in the tool's algorithm.

2.1.2 Endogeneity and Instrumental Variables

Yet another fundamental concept in panel data analysis is that of endogeneity in the regressors, i.e. a correlation between regressors and the error term. While the causes of endogeneity might be multiple, as discussed in the next chapter, its presence violates the Gauss-Markov assumptions for the unbiased and consistent OLS regression, which no longer recovers the causal effects in the relationship. Under such circumstances, strong instrumental variables (IV) (discussed in the next chapter) are identified for Two Stage Least Squares (2SLS) regression to purge the bias. Wooldridge explains, "the method of instrumental variables (IV) provides a general solution to the problem of an endogenous explanatory variable." In practice, Staiger and Stock (1997) showed that IV should be sufficiently strong (First stage F-statistic ¿ 10) to avoid severe bias and non-normality by weak IV estimates. The tool automatically tests the strength of Instrumental Variables (IV) and shifts to dynamic modelling if no adequate IV is found. The tool identifies if endogeneity is present by performing a Hausman test under the null hypothesis of exogenous regressors, as discussed in Hausman 1978 and Wooldridge.

2.1.3 Dynamic Panel Models and GMM Estimation

In practice, many financial processes are dynamic, which implies that a dependent variable is a function of the time-lagged variable of itself. However, such a model violates the strict exogenous conditions of the fixed effects or random-effects model as the lagged variable is inherently correlated to the lagged dependent variable (Nickell Bias), which can produce inconsistent estimates. The Arellano-Bond (1991) approach mitigates this by first differencing the two consecutive equations to get rid of the fixed effects and then using a deeper lagged dependent variable as a valid regressor. This approach produces consistent results in the large N, small T panels. Extensions (e.g. Arellano and Bover (1995), Blundell and Bond (1998)) form system GMM using additional moment conditions, further improving efficiency for typical macro and finance panels. This tool can automate the dynamic modelling when the conditions have been met, as described in the next chapter or when hypothesised, following the standard literature

2.1.4 Model Diagnostics and Specification Tests

To check the model's validity in panel data analysis, this tool also includes normal diagnostic procedures. Typically for panel data, serial correlation in the residuals (e.g. by using Wooldridge's (2002) test for serial correlation) and cross-sectional heteroskedasticity are checked, and if violated, robust (clustered) standard errors or GLS-correction are applied. More generally, F-tests, or Wald tests, test joint hypotheses of the coefficients (e.g. whether all the coefficients for a group of variables are zero). Collectively, these tests keep the Gauss–Markov assumptions and the model specification in check, thereby providing unbiased and consistent estimates.

2.1.5 Automating the formalised theories

The tool intends to codify these concepts and theories that are considered economic best practices to follow in a panel data analysis. Most software, including the proprietary ones, leaves these conceptual steps to the user. This makes the task of the user of the software hectic and tedious as they have to keep track of all the steps that they have performed, and missing any one step, more often than not, requires the complete analysis to be performed again. For complex analyses, such executions can become time-consuming and prone to errors, which this tool tries to mitigate. While the tool automates tests, it uses an academically substantiated, rigorous decision tree for the automation.

2.2 Innovation and Adoption Framework

A growing body of evidence shows that academic research has been the root of many open-source projects, such as Apache, R, Python, and so on, that have been adopted by individuals and institutions alike to manage the resource constraints and align with open-science initiatives. Amidst this backdrop, the Python-based cross-platform tool proposed by this thesis provides researchers a user-friendly, efficient, low-cost alternative to proprietary software in alignment with the broader paradigm of the scientific discourse. This section integrates four foundational frameworks—the Resource-Based View (RBV), Diffusion of Innovation Theory, the Technology Acceptance Model (TAM), and the Unified Theory of Acceptance and Use of Technology (UTAUT). Collectively, these perspectives offer a layered understanding of (1) why researchers continue to rely on costly proprietary software, (2) why the current context favors disruption, (3) what design features are essential for a new tool to be viable, and (4) how adoption can be supported and sustained in both individual and institutional contexts.

2.2.1 Resource-Based View (RBV)

In academia, proprietary and expensive software has become the standard for Panel Data Analysis because these softwares are as seen as a resource that is valuable, rare, inimitable and non-substitutable (VRIN) as articulated by Barney (1991). Such resources provide them a competitive edge in terms of research methodology, funding support, marketing opportunities and therefore, institutions get locked in with the vendor despite the alternatives. Well, this works fine for well-funded universities, it creates resource crunch for the under-funded institutions. This resource asymmetry further reinforces the resource-based view of the software making them the standard for the panel data analysis. As West and Gallagher (2006) and McGill et al. (2022) have shown, universities frequently invest heavily in proprietary tools to preserve continuity, despite the availability of capable open-source alternatives. In contrast, the tool proposed in this thesis offers institutions a new class of VRIN-compatible resource—one that is low-cost, open to customization, and conducive to in-house capacity-building.

2.2.2 Diffusion of Innovation Theory

According to Rogers' diffusion of Innovation theory, adoption of innovation can be characterised by a bell-shaped curve. According to Rogers, the adoption initiates with innovators and early adopters, progresses through the early majority and late majority, and finally is adopted by the laggards. The traditional tools, such as STATA, SPSS and Eviews, have been adopted heavily in the market. However, with the cost of licensing, training and compliance so high, it provides little marginal value to the new and/or under-funded institutions. Therefore, the context is viable for an innovator to disrupt the market, restarting the cycle of adoption of the innovation. Crucially, Rogers identifies five attributes that facilitate diffusion: relative advantage, compatibility, complexity, trialability, and observability. The proposed tool provides a relative advantage due to its automated, efficient and user-friendly nature. The tool's compatibility with all platforms through Python makes it compatible while reducing the complexity of setup and increasing the trialability of the tool. What this theory suggests is that a mere thesis is not enough for it to become observable. Events such as workshops and marketing have to be done to make it truly adoptable for the market.

2.2.3 Technology Acceptance Model (TAM)

The Technology Acceptance Model (TAM), developed by Davis (1989), posits that two key factors determine whether users will adopt a technology: Perceived Usefulness (PU) and Perceived Ease of Use (PEOU). PU refers to how much

users believe the tool will improve their performance, while PEOU captures the effort required to learn and use it. In the context of academic research, users often avoid open-source tools not because they lack functionality, but because they are perceived as difficult to use or poorly supported. This tool is designed to directly address both concerns: it automates critical steps in panel data analysis and provides clean, intuitive workflows through Python, improving both PU and PEOU. Additionally, its design philosophy aligns with modern reproducible research practices, increasing its perceived usefulness within open-science-driven academic institutions.

2.2.4 Unified Theory of Acceptance and Use of Technology (UTAUT)

Expanding on TAM, the Unified Theory of Acceptance and Use of Technology (UTAUT) by Venkatesh et al. (2003) introduces four factors that influence adoption: Performance Expectancy, Effort Expectancy, Social Influence, and Facilitating Conditions. While performance and effort expectancy closely mirror TAM, UTAUT emphasizes the role of peers and institutions. Adoption of new tools in academia depends not only on perceived utility but also on whether colleagues endorse the tool, and whether technical support and infrastructure are available. The tool proposed in this thesis has been developed with these adoption enablers in mind: it is free and cross-platform (low effort), it provides measurable performance gains (high performance expectancy), and it is supported by comprehensive documentation and planned training resources (strong facilitating conditions). Peer adoption and visibility through demonstrations, workshops, and collaborative use are essential for triggering the broader diffusion cycle suggested by this framework.

Chapter 3

Research Methodology

As previously discussed, panel data is a much richer dataset when compared to pure cross-sectional data or time-series data and provides better control over heterogeneity and dynamic relationships among the variables under investigation. Nonetheless, Panel data has its own set of econometric and statistical challenges, such as multicollinearity, heteroskedasticity and endogeneity, which must be understood and addressed carefully to get reliable inferences. The methodology section of this project defines and clarifies the significance of essential terms, models, techniques and statistical tests employed in the proposed tool. This section also discusses the workflow that this tool employs to perform the automated panel data analysis the setup required on each operating system, as well as the original Python code that performs the analysis.

3.1 Definitions

3.1.1 Panel Data, Variable and Model

Panel data contains observations associated with multiple entities (e.g., firms, individuals, governments) over regular intervals of an extended time period. Mathematically, panel data can be described as

$$Y_{it}, X_{it} \quad \text{for } i = 1, 2, \ldots, N, \quad t = 1, 2, \ldots, T \tag{3.1}$$

where,

- i implies the i-th entity
- t implies the t-th interval

- X_{it}, Y_{it} is the observation corresponding to i-th entity and t-th time interval for variables (observable attributes of entity) X and Y.

A panel data is considered short panel data if the number of entities N exceeds the time intervals T, i.e. $N > T$ or Long (Tall) panel data if the number of time intervals T exceeds the entities N, i.e. $T > N$.

The columns of the panel data are called variables. These are the attributes associated with the entities that are observed regularly. For instance, in a panel data set that observes GDP, GDP per capita, and PPP of all the countries over 10 years; GDP, GDP per capita and PPP are considered variables.

A model in a panel data analysis is a mathematical and conceptual relationship between two or more variables which are a part of regression analysis. It can be expressed in a general form as:

$$Y_{it} = X_{it}\beta + \alpha_i + \lambda_t + u_{it} \tag{3.2}$$

where:

- Y_{it} = Dependent variable for entity i at time t.
- X_{it} = Vector of independent variables.
- β = Coefficients of explanatory variables.
- α_i = **Unobserved entity-specific effect** (time-invariant).
- λ_t = **Unobserved time-specific effect** (affects all entities equally).
- u_{it} = **Error term** (idiosyncratic disturbance).

3.1.2 Dependent Variable

The variable (or the column in the panel data) being predicted or explained in the model by performing regression on the independent variable(s) (or columns) is called a Dependent Variable (DV), outcome variable, or response variable. In a panel data analysis, the research question tries to quantify the impact experienced by the dependent variable due to the independent variable's variations. For instance, in a research study on stock market data, if a researcher wishes to understand the impact of the volume of stocks on the valuation of stocks, then valuation is the dependent variable (DV). DV can be diagrammatically represented by putting it on the head end of the arrow in a general conceptual diagram, as shown in Figure 3.1.

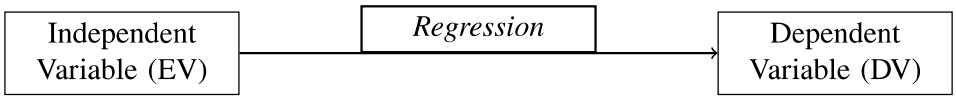


Figure 3.1: Conceptual Diagram of Regression

3.1.3 Independent Variables or Explanatory Variables (EV)

The variable(s) that is being manipulated or observed to quantify its impact on the dependent variable (DV) is called the independent variable, explanatory variable (EV) or regressor. As shown in Figure 3.1, they are represented by putting them on the tail end of the arrow in the concept diagram. In the stock market research study example, the stock volume is the independent variable in the previous example. All explanatory variables in a model must be exogenous, that is, they should be uncorrelated with the error term of the model.

3.1.4 Control Variables

Control variables are the explanatory variables that are not the main focus of the study but are included in the model to improve the overall fit of the regression line to determine the relationship between the dependent variable (DV) and explanatory variable(s) of the model more accurately by reducing the omitted variable bias. Omitted variable bias implies that some variable that impacts both the independent variable and dependent variable is omitted, resulting in biased outcomes.

3.1.5 Instrument Variables

Instrumental variables serve a very important purpose in regression analysis when the primary explanatory variable itself is an endogenous variable, which means that it shows a strong correlation between itself and the error term. In such cases, the primary explanatory variable is replaced with an instrument variable that strongly correlates with the explanatory variable (termed relevance) but is not correlated to the error term (termed exogeneity). Such variables are then used in the two-stage least square estimation (2SLS) technique to produce unbiased and reliable estimates. This technique is called the Instrument Variable estimation technique.

3.1.6 Moderator Variables

Moderator variables are unique variables that are used exclusively in the interaction models for panel data regression analysis. These variables are the context indicators that can influence the direction and strength of the relationship between the

independent variable and the dependent variable of the model. Unlike the control variables that can directly influence the dependent variable, moderator variables may or may not directly impact the dependent variable, instead they act on the independent variables using the interaction term, which can then alter the nature of the impact that the independent variable has on the dependent variable. If the interaction term is determined to be significant in the regression, it implies that the model is context-dependent and the relationship changes based on the value of the moderator variable. Fig. 2.2 shows the diagrammatic way of representing the interaction variable.

3.1.7 Homogeneity and Heterogeneity

In the context of panel data analysis, Homogeneity is the assumption that the relationship under scrutiny is homogenous for all entities and time intervals. If such an assumption is considered valid, pooled OLS can be used to model the relationship using the following mathematical equation.

On the other hand, heterogeneity is the assumption that all entities have distinct characteristics and different relationships between the variables under scrutiny, thereby producing variations that are either entity-specific or time-specific and need to be inculcated in the model in the form of Fixed Effects Model or Random Effects Model.

3.1.8 Exogeneity and Endogeneity

The explanatory variables are considered exogenous variables if the covariance between them and the error term is zero. It is interpreted as explanatory variables not being impacted by the unobserved factors that influence the dependent variable. This allows for unbiased estimates of the coefficients and intercepts of the model.

In contrast, when the explanatory variables are correlated to the error term, i.e. the covariance between them and the error term is not zero, they are termed as endogenous variables. The presence of endogeneity in a model can result in biased and inconsistent estimates. Endogeneity can be attributed to three causes, primarily: first, the Omitted Variable Bias (OVB), which implies that a variable that correlates with both dependent and independent variables is omitted from the model and is thus captured by the error term which leads to a non-zero covariance between the error term and the independent term causing endogeneity. Second, when the independent variable impacts the dependent variable and vice versa, causing simultaneity. Third, the error in measuring the explanatory variable can cause a correlation with the error term.

3.1.9 Homoskedasticity and Heteroskedasticity

If the error term has constant variance across all values of the independent variables, it allows for accurate estimation of standard errors and confidence intervals, leading to reliable hypothesis testing. This is called as homoskedasticity.

However, if the condition mentioned above is not met, that can lead to unreliable confidence intervals and, thus, an unreliable hypothesis testing scenario. This can be mitigated using standard robust errors, weighted Least Squares (WLS) or feasible Generalised Least Squares (FGLS). It can also be remedied using a logarithmic transformation of the dependent variable in cases where these variables have significant variations in their values.

3.1.10 Serial Correlation

When error terms are correlated to lagged error terms, they cause serial correlation. It implies that errors are not identically and independently distributed, rendering models such as pooled OLS, fixed effects and random effects models inefficient and unreliable. This issue can be identified using the Wooldridge test and can be mitigated using the Robust Standard Errors.

3.2 General Workflow of the Panel Data Analysis

This section outlines the process of panel data analysis that is followed in this project for the automation of the tool and analysis using the tool itself.

First, after the research question has been confirmed, relevant data is collected from reliable sources such as CMIE Prowess, Bloomberg Terminal, annual reports and surveys, official websites or any other method necessary. This data is then stored in spreadsheet software such as Google Sheets or Microsoft Office Excel, and it is cleaned by removing unnecessary rows and columns, selecting a balanced sample, filling in missing values using some techniques, and making the data uniform. We must ensure that the first two columns are time intervals and entity identifiers, respectively. It is also necessary to ensure that no column names have any white spaces in them.

Second, the data is checked to ensure no non-numeric values are in the numeric columns used for analysis.

Third, as a part of exploratory data analysis, descriptive analysis is performed, followed by the estimation of a correlation matrix.

Fourth, econometric models are created for hypothesis testing using descriptive statistics and correlation matrix insights. At this stage, the choice of performing Standard panel data regression, quantile panel data regression analysis or even

Difference-in-difference analysis could be made based on the objective of the study. Standard panel data regression is the most common type of analysis that is used to establish and explain the average trend of impact of the independent variable on the dependent variable. Quantile panel data regression is used to establish trends at different quantiles of the explanatory variable. Difference-in-difference establishes a causal effect by comparing the differences between the treatment and control groups before and after an event such as policy implementation.

Fifth, focussing on the standard regression after the selection of the model, endogeneity is checked for the model using the Durbin Chi-Square Test, which assumes homoskedasticity and the Durbin-Wu-Hausman Test, which doesn't assume homoskedasticity. If the p-values of these tests are less than 0.5, then endogeneity is deemed significant. In cases of significant endogeneity, adopting the Two-Stage Least Squares (2SLS) technique is recommended to mitigate. 2SLS works by first regressing an instrument variable with the dependent variable and then regressing the instrument variable with the explanatory variable so that endogeneity doesn't exist in any given model.

Yet another option at this stage is to pivot to dynamic models if the endogeneity issue persists. Dynamic models are time-dependent models with a time-lagged dependent variable as one of the model's independent variables of the equation. They can be estimated using Arellano-Bond GMMs and Blundell-Bond GMMs.

The Fixed Effects and the Random Effects models are recommended if endogeneity is insignificant.

Sixth, if fixed effects and random effects model is used, the Hausman test is performed to choose between the two. If the p-value of the test is less than 0.05, it implies that there is a significant correlation between entity-specific constants and regressors, thus fixed effects model is used to reduce Omitted Variable Bias. On the contrary, if the p-value above 0.05 signifies that there is no meaningful correlation between entity-specific constants and the regressors, thus random effects model is preferred.

After selecting one of the two models or the 2SLS model from the previous step, the Wooldridge test for serial correlation and Wald's test for heteroskedasticity is performed. If any of these two are significant, standard robust errors are used as covariance type while analysing the selected model to make the hypothesis testing reliable by reducing uncertainty due to error-related statistical issues. The summary statistics of these models, such as the coefficients of the model's explanatory variables, are then observed and used as econometric evidence for hypothesis testing of the leading research question formulated in Step 1.

This workflow is demonstrated in figure 2.x

3.3 Python Code's Workflow

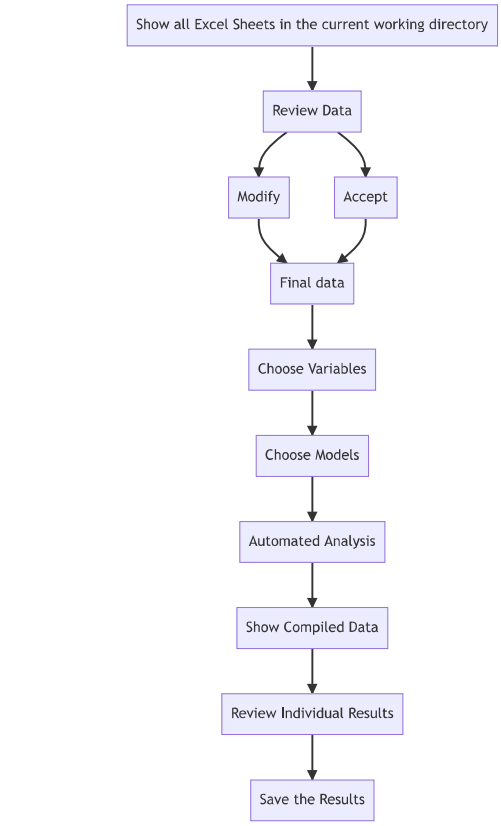


Figure 3.2: Workflow of the Python Code

3.3.1 Environment Setup

The tools were developed in the Python programming language. The environmental setup consisted of a Miniconda on Windows 11 on an Asus Zenbook Duo Pro 14 (2022) with an Intel i9 Processor and 32 GB RAM. After setting up the Miniconda, the following commands were run on the Windows Terminal with Powershell 7.4.5

shell to create a virtual environment. The same can be done on any platform by installing the Miniconda environment on any OS accessing it using any shell, and entering the commands below:

```
conda create -n panelytica   # creates a virtual env
conda activate panelytica    # activates virtual env
conda install pip            # install packages
pip install pandas openpyxl survey datascroller
   linearmodels statsmodels
```

After setting up the environment, create a Python file called *panelytica.py* in any folder and add the code from the following subsection. The file containing the above code can then be executed by running the following command in any shell.

```
python panelytica.py
```

To access this tool in any folder, follow the steps listed below.

1. **For Apple's macOS,** open the Terminal and run the following command:

```
echo "alias panelytica='conda activate panelytica\
&& python /absolute/path/to/panelytica.py'" >> ~/.
   zshrc && \
source ~/.zshrc
```

2. **For Linux OS,** open the Terminal and run the following command:

```
echo "alias panelytica='conda activate panelytica\
&& python /absolute/path/to/panelytica.py'" >> ~/.
   bashrc && \
source ~/.bashrc
```

3. **For Microsoft Windows,** open PowerShell and run the following command:

```
Add-Content -Path $PROFILE -Value "`nfunction
   panelytica { `n  \
conda activate panelytica; `n  \
python 'C:\absolute\path\to\panelytica.py' `n}"
. $PROFILE
```

After restarting the Terminal, this tool can be accessed in any directory by running:

```
panelytica
```

This will locate all the Excel files in the directory, allowing the user to select the relevant file for analysis. The results will be saved in the same folder.

3.3.2 Python Code

The complete code for this tool can be found in the Appendix 1.

Chapter 4

Demonstration of the tool

This section contains two examples to demonstrate two different use cases for the tool. The first example demonstrates the tool's use case for rapid brainstorming on the possible research questions or quick investigation to explore any relationship of interest between any two variables. The second example demonstrates the tool's autonomous panel data analysis algorithm that works as described in the previous section.

4.1 Exploratory Analysis

4.1.1 Problem Statement

What is the impact of Financial Distress on Shareholder Yield? what if using Sustainability and competition as the moderators?

4.1.2 Data

The financial data extracted from Bloomberg Terminal is a panel data of Fortune 500 companies listed on the US Stock Exchange. This contains annual data of the companies between 2010 and 2022.

4.1.3 Variables

The variables of interest are Z-score for financial distress (explanatory variable), SHY for Shareholder yields (Dependent Variable), ESG for sustainability and Lerner's index for competition (moderator variables).

4.1.4 Objective

The main objective here is to identify the range of relationships that are possible between the given variables. we are attempting to have quick glance of the relationship from the econometric perspective.

4.1.5 Literature

The relationship between Financial Distress and Shareholder Yield is complex and can be influenced by many factors. However, It is understood from the literature that financial distress causes the company to focus on its finances by reducing outflow of the cash, reinvesting the profits and therefore, Shareholder's Yield should be effected negatively. The proportion of such an impact can vary from firm to firm based on different aspects. When taking into account the moderating effect of sustainability, firms that focus on sustainability are less likely to face financial distress and even when it does, it is not likely to impact the shareholders as much. In case of financial distress under higher competition, the shareholder yield is likely to impacted adversely.

4.1.6 analysis

The figure 3.1-3.7 shows the interactive user interface. Figure 3.8 showcases the result of permutation and combination of static models that were possible given the problem statement and It shows all the relationship have an R-Squared of zero. This means that there is no empirical evidence of any significant relationship between financial distress and the shareholder yields with or without the moderators. This is corroborated by Purohit et al. (2025).

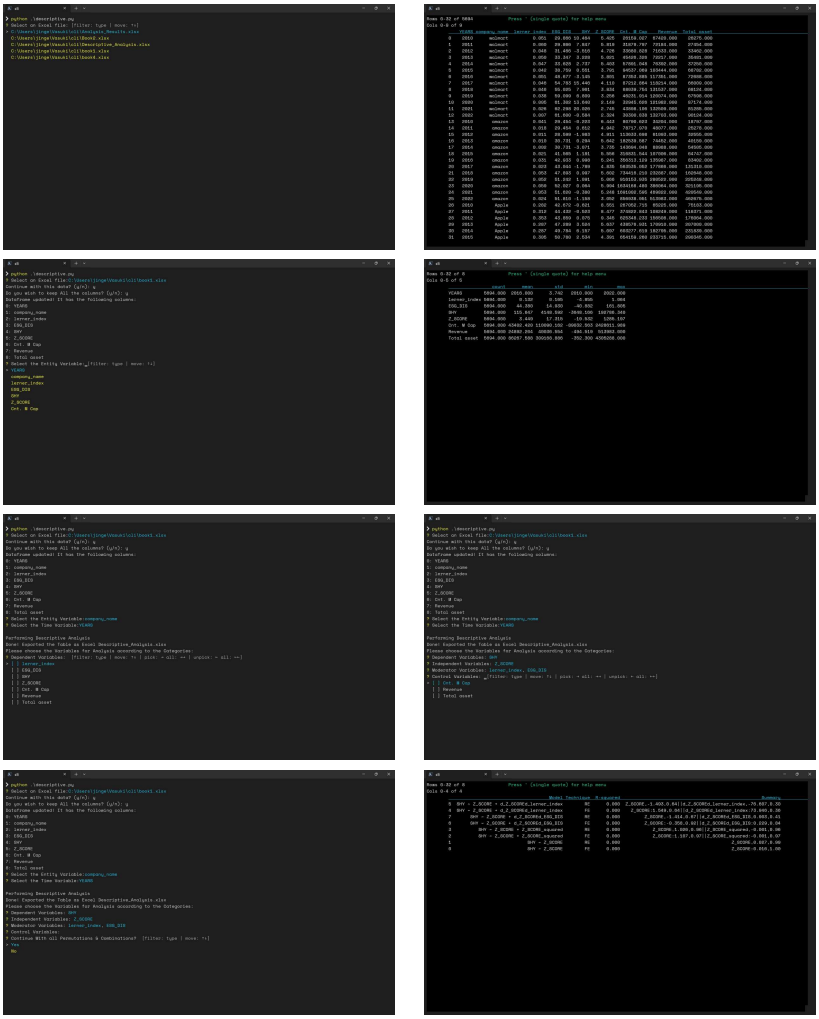


Table 4.1: Interactive UI of the tool

4.2 Autonomous Analysis

4.2.1 Problem Statement

To quantify the impact of the Intangible assets of banks on their valuation.

4.2.2 Data

The data is a panel data of Indian banks that contains annual data between 2010 tO 2024

4.2.3 Variables

The variable of a bank's valuation is MCap (Dependent Variable), for intangible assets is total-intangible-assets (EV) and the control variable is total-debt-to-asset-ratio (CV).

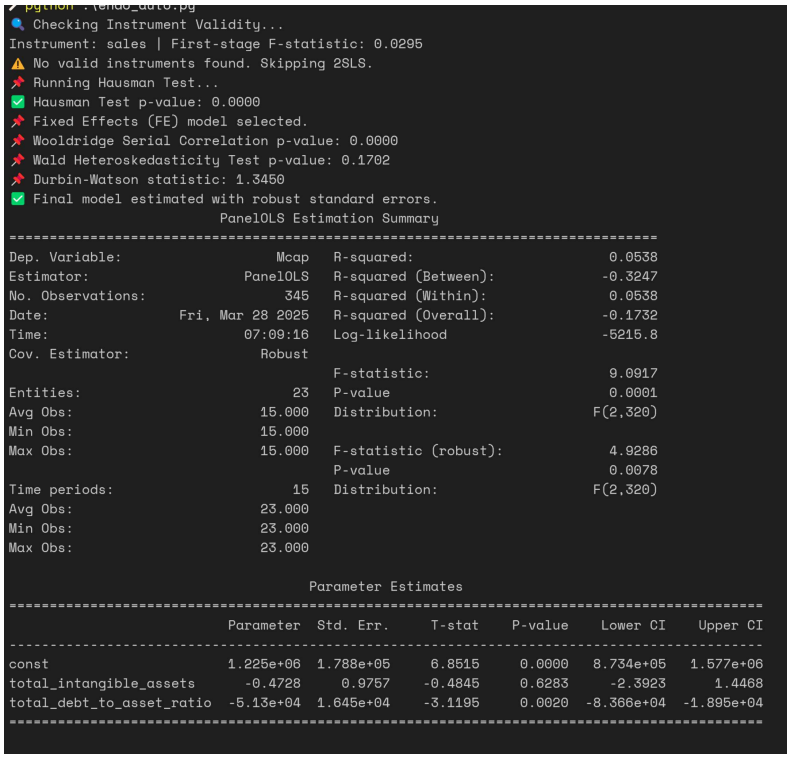


Figure 4.1: Autonomous Test

4.2.4 Analysis

The Figure 3.1 demonstrates how the research question for panel data regression was carried out without any human intervention in an automated and the result of the final model that was best suited for this analysis was presented. In this case, no endogeneity was observed and therefore, it performed the Hausman test which produced significant p-value leading to the selection of Fixed effects model for this analysis. Upon testing for Serial correlation, the results was significant but the heteroskdasticity turned out to be insignificant. To mitigate this, Fixed-effects model was estimated again with Standard Robust Errors and final output is

displayed as in the figure. The results shows that the impact of intangible assets on its valuation is insignificant.

Chapter 5

Conclusion

This project's aim was to create and demonstrate a novel Python-based cross-platform tool for panel data analysis in econometrics. This tool provides an alternative to existing proprietary software by being free, open-source (FOSS), automated, and inexpensive. The previous sections have demonstrated two key capabilities of this tool. First, the ability to test hundreds of models simultaneously, arising from various combinations of suitable variables, allows summarizing results in tabular form. This ensures that comprehensive studies can be conducted efficiently without losing perspective. Since such analyses can be performed within seconds to minutes, they save senior researchers weeks of tedious work. Second, the ability to test targeted models with a high level of autonomy acts as a guardrail for early-stage researchers. The workflow is clearly depicted with relevant decisions printed transparently, ensuring that the tool does not function as a "black box." Instead, it reinforces learning, making it a tool not just for analysis but also for education. The tool offers both verbose output, which explains every step, and a quiet mode that only displays essential information and decisions.

While large language models and generative AI, such as ChatGPT, are increasingly used for statistical analysis, this tool demonstrates the efficacy of a simple algorithm that consumes less power and time while delivering accurate results at lower costs. However, this project has highlighted the difficulty of generalization. Countless tools, libraries, and modules have been tested to ensure cross-platform compatibility and usability across various hardware configurations. For instance, JupyterLab was found unsuitable due to its non-linear code execution which is confusing for non-programmers, clunky user-interface and terrible resource management. A Command Line Interface (CLI) version was explored using at least three different frameworks such as ncurses, Python's Terminal User Interface (TUI) and Go libraries, but they proved unresponsive for large datasets, failing to perform consistently across different operating systems. While a Graphical User Interface (GUI) implementation remains a potential avenue for future development, it poses

the challenge of balancing the responsiveness of user interface and allowing the code to run for longer times on less powerful computers. This implies that the software might hang on the less capable systems. This can be achieved by separating the application's functionality into two different processes that run on two cores of the CPU and handle the GUI interface and the computations separately. But, that is an advanced software architecture that is beyond the scope of this project currently.

Based on the limited testing that was possible in the Research Hub of this project's venue construing to the possibility of plagiarism, the anecdotal response from researchers is positive and they prefer the terminal version over any prospective TUI or GUI due to its ease of use and the speed of computations. This software has been tested on all the operating systems where it has functioned as intended. While a lot of precautions have been taken to make this software bug-free, chances of running into niche and unique bugs are significant. Therefore, this tool needs active development to fix bugs and to integrate newer and more advanced econometric techniques and test to truly democratise the econometric capabilities to enhance the quality of the associated research and paving the way for the discipline to adopt more nuanced analysis.

As I started this project, i made some key observations that were novel for me. Most academicians in humanities and social sciences (HSS) have non-technical backgrounds and are not well-versed in programming. Although they take courses on data analysis, these often fail to match the rigor required by peer-reviewed journals. Numerous workshops and faculty development programs (FDPs) are conducted yearly to teach data analysis techniques and facilitate networking. However, despite these efforts, cutting-edge techniques remain underutilized. Many outdated methods persist, and widely used statistical software is proprietary with long development cycles, limiting access to the latest methodologies. As a result, early-stage research scholars often produce lower-quality research. By making this tool free and open source, this project aims to democratize research methodology, ensuring that the latest techniques are accessible to all users immediately upon release, thereby driving the discussion and debate on them.

Additionally, in Indian academia, a hierarchy exists based on data analysis proficiency. Those skilled in data analysis act as gatekeepers, creating divisions within research groups. This project aims to simplify data analysis, making it accessible to all researchers. By enabling individuals to conduct their own analyses, the tool fosters independence and disrupts hierarchical structures within research teams. This observation has become a source of motivation to carry research that can navigate the complex webs of the academic society and produce positive impacts. A natural fallout of such divisiveness is that many techniques have become outdated and tedious. A paradigm shift is needed to integrate cutting-edge methodologies and cutting-edge technologies to develop comprehensive economic models and pro-

duce novel insightful research. Presently, much of the research conducted using outdated techniques has little practical application despite its academic relevance. Current studies often focus retrospectively rather than incorporating forecasting capabilities.

To substantiate my observations, take the example of ecology and economics which have long been intertwined in their concepts and ideas. Historically, economic principles influenced ecological research, but modern ecology has embraced techniques from physics and planetary sciences, such as chaos theory and complexity science. While complexity science has been explored in econometrics, chaos theory remains underutilized in economic analysis. Developing new tools capable of handling such computations is key to advancing economic theory and analysis. Panel data analysis is primarily used to uncover patterns in retrospective data. The resulting models can form the foundation for new theories or equations, which can then be tested through simulations using chaos theory. These insights could lead to the development of automated financial and economic forecasting tools. Furthermore, these models could serve as a theoretical and mathematical basis for applying chaos theory to complexity economics. While the natural progression of panel data analysis is yet unbeknownst to us, it should bring forth new and exciting insights.

In conclusion, this project demonstrates the potential of open-source, automated tools in econometric research. By streamlining panel data analysis and making it accessible to all researchers, this tool addresses key challenges in Indian academia. Future directions include further refinement of the tool, exploring a GUI-based implementation, and integrating advanced forecasting techniques. Ultimately, this work represents a step toward more democratic, transparent, and efficient economic research methodologies.

Bibliography

Abdullah, H., Sahudin, Z., and Bahrudin, N. Z. (2022). Short guides to static panel data regression model estimator. *Asian Journal of Accounting and Finance*, 4(4):1–6.

Aptech (2018). Introduction to the fundamentals of panel data.

Arellano, M. (1992). Generalized method of moments and optimal instruments.

Arellano, M. (2003). *Panel Data Econometrics*, volume 231. Oxford University Press.

Arellano, M. (2016). Modelling optimal instrumental variables for dynamic panel data models. *Research in Economics*, 70(2):238–261.

Arellano, M. and Honoré, B. (2001). Panel data models: Some recent developments. In *Handbook of Econometrics*, volume 5, pages 3229–3296. Elsevier.

Baltagi, B. H. (2005). *Econometric Analysis of Panel Data*. 3 edition.

Baltagi, B. H. (2008). *Econometric Analysis of Panel Data*. Wiley, Chichester, 4 edition.

Bloomberg (2023). Bloomberg crypto data, research and tools.

Center, P. R. (2023). Most u.s. bank failures have come in a few big waves.

Das, P. and Das, P. (2019). Panel data analysis: Static models. In *Econometrics in Theory and Practice: Analysis of Cross Section, Time Series and Panel Data with Stata 15.1*, pages 457–497.

GoI (2022). Pradhan mantri garib kalyan anna yojana (pmgkay).

Gujarati, D. and Porter, D. (2009). *Basic Econometrics*. McGraw-Hill, New York, NY.

Hale, T., Webster, S., and Petherick, A. (2022). Covid-19 and the global economy: Lessons from lockdown policies. *Oxford Review of Economic Policy*, 38(4):719–735.

Hillard, D. (2023). *Publishing Python Packages: Test, Share, and Automate Your Projects*. Simon and Schuster.

Hsiao, C. (2007). Panel data analysis—advantages and challenges. *Test*, 16(1):1–22.

Hsiao, C. and Wise, C. (2006). Panel data analysis—advantages and challenges.

IndeedEditorialTeam (2023). What is panel data? (with uses, advantages and an example).

Kaggle (2023). Cryptocurrency price and market data.

Min, C. K. (2019). *Applied Econometrics: A Practical Guide*. Routledge.

Purohit, S., Agarwal, B., Kanoujiya, J., and Rastogi, S. (2025). Impact of shareholder yield on financial distress: using competition and firm size as moderators. *Journal of Economic and Administrative Sciences*.

Shih, W. (2020). Global supply chains in a post-pandemic world. *Harvard Business Review*.

Sohail, M. (2023). Vaccine inequities and the prolonged global impact of covid-19. *Clinical and Experimental Immunology*, 212(1):102–116.

Staff, I. (2023). Bloomberg vs. reuters: What's the difference?

WallStreetMojo (2023). Panel data analysis - what it is, examples, advantages, methods.

WHO (2020). Impact of covid-19 on people's livelihoods, their health and our food systems.

Wooldridge, J. M. (2019). Correlated random effects models with unbalanced panels. *Journal of Econometrics*, 211(1):137–150.

Wooldridge, J. M. (2020). *An Introduction to Econometrics*. Cengage Learning, Boston, MA.

.1 The Python Code

```
import os
from pathlib import Path
import survey
import contextlib
from linearmodels.panel import PanelOLS, RandomEffects

def suppress_error(x):
    with contextlib.redirect_stderr(None):
        exec(x)

def select_excel_file():
    # Get the list of Excel files in the current directory
    current_path = Path.cwd()
    file_list = sorted([str(file) for file in current_path.
       glob("*.xls")] + [str(file) for file in current_path
       .glob("*.xlsx")])

    if not file_list:
        print("No Excel files found in the current
           directory.")
        return None
    else:
        # Use the survey library to select a file
        index = survey.routines.select(
            'Select an Excel file:',
            options=file_list,
            focus_mark='> ',
            evade_color=survey.colors.basic('yellow')
        )
        return file_list[index]

import pandas as pd
from datascroller import scroll
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
pd.options.display.float_format = '{:.3f}'.format

def continue_prompt(choice_prompt,loop=None, choice=None):
    if loop == 'while':
        while True:
            user_choice = input(choice_prompt)
```

```
            if user_choice in choice:
                break
            else:
                continue
    else:
        user_choice = input(choice_prompt)
    return user_choice

def select_columns(df):
    choice = continue_prompt("Do you wish to keep All the
       columns? (y/n): ", 'while', ("y", 'n'))
    if choice == 'n':
        columns_to_keep = []
        print("Select the columns to keep from this list.")
        for index, column in enumerate(df.columns):
            print(f"{index}: {column}")
        print("""
You can choose the columns in one of the following three
   ways:
1. By just writing numbers that you wish to keep separated
   by Space. (eg. 1 2 3)
2. By writing a range (eg. 1-3)
3. By a combination of both (eg. 1-3 5-7 10 13 16)
             """)
        string_columns = input("Enter the columns to keep:
           ")
        for part in string_columns.split(' '):
            if '-' in part:
                start,end = map(int, part.split('-'))
                columns_to_keep.extend(range(start,end+1))
            else:
                columns_to_keep.append(int(part))
        columns = df.columns[columns_to_keep]
    else:
        columns = df.columns
    print(f"Dataframe updated! It has the following columns
       : ")
    df = df[columns]
    for index, column in enumerate(df.columns):
        print(f"{index}: {column}")
    return df[columns]


```

```
filename = select_excel_file()
#filename = 'book4.xlsx'
df = pd.read_excel(filename)
scroll(df)
continue_prompt("Continue with this data? (y/n): ", 'while ', ('y', '', 'Y'))
df = select_columns(df)
column_options = df.columns.to_list()
entity_variable = df.columns[survey.routines.select('Select the Entity Variable:', options=column_options, focus_mark='> ', evade_color=survey.colors.basic('yellow'))]
column_options = df.columns.drop(entity_variable)
time_variable = df.columns[survey.routines.select('Select the Time Variable:', options=column_options, focus_mark='> ', evade_color=survey.colors.basic('yellow'))]
column_options = column_options.drop(time_variable)
print("\nPerforming Descriptive Analysis")
descriptive_df = df.describe().drop(['25%', '50%', '75%']).transpose()
scroll(descriptive_df)
descriptive_df.to_excel("Descriptive_Analysis.xlsx")
print("Done! Exported the Table as Excel Descriptive_Analysis.xlsx")
print("Please choose the Variables for Analysis according to the Categories: ")
dependent_variables = column_options[list(survey.routines.basket('Dependent Variables: ', options = column_options, positive_mark = '    '))].tolist()
column_options = column_options.drop(dependent_variables)
independent_variables = column_options[list(survey.routines.basket('Independent Variables: ', options = column_options, positive_mark = '    '))].to_list()
column_options = column_options.drop(independent_variables)
moderator_variables = column_options[list(survey.routines.basket('Moderator Variables: ', options = column_options, positive_mark = '    '))].to_list()
column_options = column_options.drop(moderator_variables)
control_variables = column_options[list(survey.routines.basket('Control Variables: ', options = column_options, positive_mark = '    '))].to_list()


```

```
103
104 continue_choice = ["Yes", "No"]
105 continue_choice_boolean = [True, False]
106
107 automated_analysis = continue_choice_boolean[survey.
    routines.select('Continue With all Permutations &
    Combinations? ', options=continue_choice, focus_mark='>
    ', evade_color=survey.colors.basic('yellow'))]
108 if automated_analysis:
109     linear_model = True
110     quadratic_model = True
111     moderator_model = True
112     static_technique = True
113     dynamic_technique = True
114     linear_models = []
115     for dep_var in dependent_variables:
116         for indep_var in independent_variables:
117             model = f"{dep_var} ~ {indep_var}"
118             linear_models.append(model)
119
120     moderator_models = []
121     for dep_var in dependent_variables:
122         for indep_var in independent_variables:
123             for mod_var in moderator_variables:
124                 model = f"{dep_var} ~ {indep_var} + d_{
                        indep_var}d_{mod_var}"
125                 moderator_models.append(model)
126
127     quadratic_models = []
128     for dep_var in dependent_variables:
129         for indep_var in independent_variables:
130             model = f"{dep_var} ~ {indep_var} + {indep_var}
                    _squared"
131             quadratic_models.append(model)
132
133     static_models = []
134     dyanmic_models = []
135
136 if not automated_analysis:
137     linear_model = True
138     linear_models = []
139     for dep_var in dependent_variables:
140         for indep_var in independent_variables:
```

```
            model = f"{dep_var} ~ {indep_var}"
            linear_models.append(model)


    quadratic_model = continue_choice_boolean[survey.routines.select('Do you wish to employ Quadratic Model? ', options=continue_choice, focus_mark='> ', evade_color=survey.colors.basic('yellow'))]
    if quadratic_model:
        quadratic_models = []
        for dep_var in dependent_variables:
            for indep_var in independent_variables:
                model = f"{dep_var} ~ {indep_var} + {indep_var}_squared"
                quadratic_models.append(model)


    moderator_model = continue_choice_boolean[survey.routines.select('Do you wish to employ Moderator Models? ', options=continue_choice, focus_mark='> ', evade_color=survey.colors.basic('yellow'))]
    if moderator_model:
        moderator_models = []
        for dep_var in dependent_variables:
            for indep_var in independent_variables:
                for mod_var in moderator_variables:
                    model = f"{dep_var} ~ {indep_var} + d_{indep_var}d_{mod_var}"
                    moderator_models.append(model)

    static_technique = continue_choice_boolean[survey.routines.select('Do you wish to employ Static Technique? ', options=continue_choice, focus_mark='> ', evade_color=survey.colors.basic('yellow'))]
    if static_technique:
        static_models = []


    dynamic_technique = continue_choice_boolean[survey.routines.select('Do you wish to employ Dynamic Technique? ', options=continue_choice, focus_mark='> ', evade_color=survey.colors.basic('yellow'))]
    if dynamic_technique:
```

```
        dynamic_models = []

if automated_analysis:
    Models = ""
    Models = Models + "Linear Models: \n"
    for index, model in enumerate(linear_models):
        Models = Models + f"{index}) {model}\n"
    Models = Models + "Quadratic Model \n"
    for index, model in enumerate(quadratic_models):
        Models = Models + f"{index}) {model}\n"
    Models = Models + "Moderator Models \n"
    for index, model in enumerate(moderator_models):
        Models = Models + f"{index}) {model}\n"
    # print(Models)
else:
    Models = ""
    Models = Models + "Linear Models: \n"
    for index, model in enumerate(linear_models):
        Models = Models + f"{index}) {model}\n"
    if quadratic_model:
        Models = Models + "Quadratic Model \n"
        for index, model in enumerate(quadratic_models):
            Models = Models + f"{index}) {model}\n"
    if moderator_model:
        Models = Models + "Moderator Models \n"
        for index, model in enumerate(moderator_models):
            Models = Models + f"{index}) {model}\n"
    # print(Models)



import itertools

def generate_combinations(elements):
    all_combinations = []
    for r in range(1, 6):  # 1 through 5
        for combination in itertools.combinations(elements,
            r):
            combination = ' + '.join(combination)
            all_combinations.append(combination)
    return all_combinations

```

```
control_variables = generate_combinations(control_variables
    )
linear_models_cv = []
for con_var in control_variables:
    for model in linear_models:
        model = model + " + " + con_var
        linear_models_cv.append(model)
linear_models = linear_models + linear_models_cv
quadratic_models_cv = []
for con_var in control_variables:
    for model in quadratic_models:
        model = model + " + " + con_var
        quadratic_models_cv.append(model)
quadratic_models = quadratic_models + quadratic_models_cv
moderator_models_cv = []
for con_var in control_variables:
    for model in moderator_models:
        model = model + " + " + con_var
        moderator_models_cv.append(model)
moderator_models = moderator_models + moderator_models_cv
Models = ""
Models = Models + "Linear Models: \n"
for index, model in enumerate(linear_models):
    Models = Models + f"{index}) {model}\n"
if quadratic_model:
    Models = Models + "Quadratic Model \n"
    for index, model in enumerate(quadratic_models):
        Models = Models + f"{index}) {model}\n"
if moderator_model:
    Models = Models + "Moderator Models \n"
    for index, model in enumerate(moderator_models):
        Models = Models + f"{index}) {model}\n"
print(Models)

panel_data = df.copy()
panel_data = panel_data.set_index([entity_variable,
    time_variable])
results_summary = pd.DataFrame(columns=['Model', 'Technique
    ', 'R-squared', 'Summary'])
results_summary2 = pd.DataFrame(columns=['Model', '
    Technique', 'R-squared', 'Summary'])

if quadratic_model:
```

```
    for indep_var in independent_variables:
        panel_data[f"{indep_var}_squared"] = panel_data[f"{
            indep_var}"]**2
if moderator_model:
    for indep_var in independent_variables:
        for mod_var in moderator_variables:
            panel_data[f"d_{indep_var}d_{mod_var}"] = (
                panel_data[f'{indep_var}'] - panel_data[f"{
                indep_var}"].mean()) * (panel_data[f"{
                mod_var}"] - panel_data[f"{mod_var}"].mean()
                )

if static_technique:
    for model_formula in linear_models + quadratic_models +
        moderator_models:
        fixed_model = PanelOLS.from_formula(model_formula +
            ' + EntityEffects', data=panel_data)
        model_type = "FE"
        results = fixed_model.fit()
        coefficients = results.params
        pvalues = results.pvalues
        combined_results = [f"{term}:{coef:.3f},{pval:.2f}"
            for term, coef, pval in zip(coefficients.index,
            coefficients, pvalues)]
        combined_results_string = "||".join(
            combined_results)
        results_summary.loc[len(results_summary)] = [
            model_formula, model_type, results.rsquared,
            combined_results_string]
        results_summary2.loc[len(results_summary)] = [
            model_formula, model_type, results.rsquared,
            results.summary]

        random_model = RandomEffects.from_formula(
            model_formula, data=panel_data)
        model_type = "RE"
        results = random_model.fit()
        coefficients = results.params
        pvalues = results.pvalues
        combined_results = [f"{term},{coef:.3f},{pval:.2f}"
            for term, coef, pval in zip(coefficients.index,
            coefficients, pvalues)]
```

```
275         combined_results_string = "||".join(
               combined_results)
276         results_summary.loc[len(results_summary)] = [
               model_formula, model_type, results.rsquared,
               combined_results_string]
277         results_summary2.loc[len(results_summary)] = [
               model_formula, model_type, results.rsquared,
               results.summary]
278
279 results_summary2.to_excel("Analysis_Results.xlsx", index=
       False)
280 scroll(results_summary.sort_values(by='R-squared',
       ascending=False))
281
282 import pandas as pd
283 import numpy as np
284 import statsmodels.api as sm
285 from linearmodels.panel import PanelOLS, RandomEffects
286 from linearmodels.iv import IV2SLS
287 from statsmodels.stats.diagnostic import het_breuschpagan
288 from statsmodels.stats.stattools import durbin_watson
289 from scipy.stats import chi2
290
291 def panel_data_analysis(file_path, endog_var, exog_vars,
       instruments=None):
292     """
293     Performs panel data analysis, including 2SLS,
           endogeneity testing, Hausman test,
294     serial correlation test, heteroskedasticity test, and
           robust standard errors.
295     """
296     # Load Data
297     df = pd.read_csv(file_path)
298     df = df.set_index(["id", "year"])
299
300     def first_stage_f_test(df, endog_var, exog_vars,
           instruments):
301         valid_instruments = []
302         for instr in instruments:
303             model = sm.OLS(df[instr], sm.add_constant(df[
                   exog_vars])).fit()
304             f_stat = model.fvalue
305             if f_stat > 10:
```

```
306                 valid_instruments.append(instr)
307         return valid_instruments
308
309     def endogeneity_test(df, endog_var, exog_vars,
        instruments):
310         model_aux = sm.OLS(df[endog_var], sm.add_constant(
            df[exog_vars + instruments])).fit()
311         hausman_stat = model_aux.rsquared * len(df)
312         p_value = 1 - chi2.sf(hausman_stat, len(exog_vars)
            + 1)
313         return hausman_stat, p_value
314
315     def hausman_test(df, endog_var, exog_vars):
316         exog_vars = sm.add_constant(df[exog_vars])
317         fe_model = PanelOLS(df[endog_var], exog_vars,
            entity_effects=True).fit()
318         re_model = RandomEffects(df[endog_var], exog_vars).
            fit()
319         diff = fe_model.params - re_model.params
320         cov_diff = fe_model.cov - re_model.cov
321         hausman_stat = (diff.T @ np.linalg.inv(cov_diff) @
            diff).item()
322         p_value = 1 - chi2.sf(hausman_stat, len(exog_vars))
323         return hausman_stat, p_value, fe_model, re_model
324
325     def wooldridge_serial_corr(df, endog_var, exog_vars):
326         df['lag_resid'] = df[endog_var].shift(1)
327         df.dropna(inplace=True)
328         model = sm.OLS(df[endog_var], sm.add_constant(df[['
            lag_resid'] + exog_vars])).fit()
329         return model.pvalues.iloc[1]
330
331     def wald_heteroskedasticity_test(model):
332         exog_df = pd.DataFrame(model.model.exog.values2d,
            columns=model.model.exog.vars)
333         if "const" not in exog_df.columns:
334             exog_df = sm.add_constant(exog_df)
335         exog_array = exog_df.to_numpy()
336         bp_test = het_breuschpagan(model.resids.to_numpy(),
             exog_array)
337         return bp_test[1]
338
339     def robust_se_model(final_model):
```

```
        return final_model.model.fit(cov_type="robust")

    def iv_2sls(df, endog_var, exog_vars, instruments):
        model = IV2SLS(df[endog_var], sm.add_constant(df[
            exog_vars]), df[instruments], None).fit()
        return model

    if instruments:
        valid_instruments = first_stage_f_test(df,
            endog_var, exog_vars, instruments)
        if not valid_instruments:
            hausman_stat, hausman_pval, fe_model, re_model
                = hausman_test(df, endog_var, exog_vars)
            final_model = fe_model if hausman_pval < 0.05
                else re_model
        else:
            endog_stat, endog_pval = endogeneity_test(df,
                endog_var, exog_vars, valid_instruments)
            if endog_pval < 0.05:
                iv_model = iv_2sls(df, endog_var, exog_vars
                    , valid_instruments)
                return iv_model.summary()
            else:
                hausman_stat, hausman_pval, fe_model,
                    re_model = hausman_test(df, endog_var,
                    exog_vars)
                final_model = fe_model if hausman_pval <
                    0.05 else re_model
    else:
        hausman_stat, hausman_pval, fe_model, re_model =
            hausman_test(df, endog_var, exog_vars)
        final_model = fe_model if hausman_pval < 0.05 else
            re_model

    serial_corr_pval = wooldridge_serial_corr(df, endog_var
        , exog_vars)
    wald_pval = wald_heteroskedasticity_test(final_model)
    dw_stat = durbin_watson(final_model.resids)
    final_model_robust = robust_se_model(final_model)

    return {
        "Final Model Summary": final_model_robust.summary()
            ,
```

```
        "Serial Correlation p-value": serial_corr_pval,
        "Wald Heteroskedasticity p-value": wald_pval,
        "Durbin-Watson Statistic": dw_stat
    }
```