AYBU Business Journal, 2(1), 10-22
https://dergipark.org.tr/tr/pub/abj
https://aybu.edu.tr/isletme/tr/sayfa/5218/AYBU-BUSINESS-JOURNAL

AYBU Business Journal

Submitted: 5 April 2022
Accepted: 27 June 2022
Published Online: 30 June 2022

REVIEW ARTICLES

Exploring a Scientific Research Methodology in Social Sciences: Steps for Analyzing Non-Stationary Heterogeneous Panel Data

Volkan Sezgin[a*]

[a]*Department of Business, Atilim University, Ankara, Turkey.*

Abstract

Accurate and rigorous applications of the econometric analysis is crucial when writing a high-quality analytical research paper in social sciences. This article provides the basic framework on how to construct econometric analysis for heterogenous non-stationary dynamic panel datasets. Panel data econometrics is a very broad field, naturally it will not be possible to include all methods in this study. Thus, we present details of the specific selected highly used standard panel data tests and estimations (Im-Pesaran-Shin unit root testing, Pedroni's cointegration test, panel data ordinary least squares) in this paper and explain why and under which conditions these methods are applied. We employ theoretical formulations of mentioned tests and estimations along with Engle and Granger's error correction mechanism in order to determine the order of integration and long-run relationship between the panel variables. In summary, we aim to explain basic steps for a straightforward empirical panel data research process for new researchers in social sciences.

Keyword

Econometric Analysis, Panel Data, Unit Root, Cointegration, Panel Data Ordinary Least Squares (PDOLS)

1. INTRODUCTION

Using econometrics when drafting research papers is a commonly used methodology in social sciences, particularly in economics and finance. Econometrics research focuses on examining the relationship between variables as it uses various types of statistical and empirical methods over different data sets.

Today, since it is easier to reach datasets through trustable sources like national statistical organizations, central banks, international organizations and/or public institutions, the tendency to do rigorous econometric investigations through powerful econometric and statistical packages have substantially increased. Taking into account that econometrics has close ties with mathematics and statistics, econometric tools have started to be widely applied over model estimations targeting to obtain reliable and quantifiable results in various branches of social sciences.

Since doing econometric research is not always very straightforward, researchers should first know the basic assumptions in order to choose the right methodology and the model for their investigation. Reliable and

[*] Corresponding author.

Contact: Volkan Sezgin ✉ volkan.sezgin@atilim.edu.tr

To cite this article: Sezgin V. (2022). Exploring a Scientific Research Methodology in Social Sciences: Steps for Analyzing Non-Stationary Heterogeneous Panel Data. *AYBU Business Journal*, 2(1), 10-22.

complete results can only be ensured by paying attention to asking right questions and following accurate econometric tools in addition to acknowledging characteristics of the data and details of the modelling. The data should possess specific characteristics to perform an econometric analysis while the hypothesis tests and hypothesis shall be constructed decently.

In this respect, this article aims to provide junior researchers from different disciplines of social sciences a perspective concerning panel data analysis. Taking into account that panel data began to provide more space for the researchers when compared with cross-section or time series, the use of both linear and nonlinear analysis of panel data research has immensely increased particularly after 1980s.

Panel data econometrics is widely applicable in economics as the scholars usually use economic theory in search of the potential variables that clarify or determine the phenomenon under consideration. Logical-deductive economic and finance models of the conduct of agents are often utilized to form fundamental chains of causality that leads in predicting variables to the dependent variable. We normally have a set of independent variables to explain the theory in econometrics; while it is not always empirically possible to elucidate economic matter of interest if the model is not strong enough to validate the empirical relationship.

As noted, panel data is one of the means to test relations between the variables. It basically means measuring the same variable over same set of units of time. Measured units can be individuals, countries, firms, universities, schools etc. Hsiao (2005) indicated that panel data *(longitudinal data)* typically refer to data containing time series observations of a number of individuals, as the data involve at least two dimensions; a cross-sectional dimension, indicated by subscript *i*, and a time series dimension, indicated by subscript *t*.

Time dimension plays an important role in panel data research as it sometimes causes serial correlation to occur. Panel data differs from the cross-sectional data as it sparks off unobserved and systematic differences across units to emerge, correlated with observed elements. In other words, panel data methodology helps to control dependencies of the unobserved over dependent variables. Normally, when constructing traditional linear regressions, these dependencies can lead to biased estimators. Constructing panel data sets can be beneficial to tackle individual unobserved heterogeneity since the data involve points within time.

It can be claimed that three major panel data models are used in the literature, which are *pooled OLS, fixed effects model and random effects model*. Pooled OLS model is similar to applying an OLS in the cross-sectional data as if the data under consideration does has an individual and time dimension in itself (the propositions are just like of ordinary linear regressions.) Fixed effects models are the models based on the differences between individual entities. This modelling approach helps removing the unobserved effects, as the unobserved impacts can be linked with observed covariates.

A random effects models pay regards to individual variations as well as time dependent ones, while the model helps us to discard the biases which change in time and stay unobserved. This model assumes the unobserved impacts are not dependent of the covariates.

In addition, there are some other types of approaches to the analysis of panel data in the literature including change score, graphical chain, pseudo panel data, structural equation models, latent variable models, among others. Although separate in modelling, there are obvious similarities between these approaches. For instance, according to Berrington et al. (2006), there is a small difference when choosing between cross-lagged panel models and graphical chain ones, or between random effects models and latent growth curve models. The authors consider the preference will highly depend on a plenty of factors which include the type of the response variables (continuous, binary, ordinal, etc.), number of variables to be included in the analysis, and weighting issues etc.

The static framework of fixed effects approach is based on estimating the individual parameters, while no assumptions on the distribution of the intercepts is required. Another feature of this approach is that the individual intercepts can be generated through the covariates. The Conditional Maximum Likelihood (CML) and joint maximum likelihood (JML) are the estimation methods used within this approach where the latter includes estimating of individual intercepts one by one for each individual in the sample while the former concentrates on the elimination of the effect of the individual intercept with sufficient statistics by the use of a logistic model which was implemented in 1980 for the first time.

Panel data econometrics is a very broad field, and has various application techniques. For instance, the choice or the condition of an individual in the current period whether influencing his/her future choice (or condition) constitutes an important question in panel data evaluations. The effect is identified as "true state dependence" denoted by Heckman (1981), if the choice in the current period has a direct impact on the future choice while it is identified as indirect (namely spurious state dependence) which works through the presence of unobserved time-constant individual heterogeneity. Within this respect, Heckman's *random effects logit model* controls for initial condition and estimated by maximum likelihood.

In 1990s, economists started to apply panel data methods to examine the economic theories involving particularly micro level individual data. In order to refrain from hypothesizing on heterogeneity and unobserved shocks, researchers usually placed their estimations on expected values which are alluded to rational expectations.

The *fixed-effect logit model* proposed by Honoré & Kyriazidou (2000) was estimated by maximizing a Kernel function and it included discrete covariates estimated on the basis of a weighted conditional log-likelihood. Besides, Hsiao (2005) developed the *dynamic logit model* by including the lagged response variables and individual intercepts to the model further to time varying and/or time-constant individual covariates. Since the fixed effects approach was not operable with huge parameters space and the use of time invariant variables was not possible, the random effects approach was introduced where the individual intercepts are random parameters with a specific distribution of their values.

In the last two decades, panel data analysis sprawled and expanded: production functions have been estimated using semi parametric methods while dynamic censored regression models were proposed in addition to the flexible parametric models. Scholars now focus not only on the parameters but also on the on partial effects in nonlinear model.

The types of panel data are very decisive in determining which methods to apply. For example, while examining *horizontal section dependence* in macro panels, there is no need to test such a phenomenon in micro panels due to the short time dimension. The presence of cross-section dependency will determine the type of unit root tests to be used (1st and 2nd generation unit root tests).

There is a growing usage of panel data in different fields of social sciences, for instance, economics to energy, environment to education. One of the first seminal applications was by Balestra & Nerlove (1966). According to Hsiao (2005), three factors contribute a lot to the exponential rise in the panel data studies recently, which can be depicted as follows: *(i)* availability of the panel data; *(ii)* challenging and competitive methodology, *(iii)* panel data have greater capacity to explain and model the complex human behavior when compared with the time series and cross section data. As noted by Hsiao, today we see that there has been a surge in the use of panel data analysis to estimate dynamic econometric models compared to cross section and time series data modelling.

Bond (2002) noted there are advantages of using panel data over the cross-section: dynamic models cannot be estimated using single point one-time observations while cross-section surveys offer limited information for dynamic relationships from earlier time spans. Panel data can be considered advantageous when compared with the aggregate time series as micro dynamics can be hidden and obscured by aggregation biases, while panel data explores heterogeneity in adjustment dynamics between different types of economic agents. As per Bond (2002), genuine panel data with repeated observations on the same individuals, can be applied to create parameter estimates since it allows more of the variation in the micro data.

Concerning this, Moundigbaye et al. (2018) emphasized that the researchers can choose among a wide variety of possible estimators as one of the key matters is to handle cross sectional dependence. The author discussed three approaches to handle this, as misspecification problems, which reduce the number of parameters to be estimated, may arise if cross-sectional dependencies cannot be reduced to a function of distance (Corrado & Fingleton, 2012). Moundigbaye et al. (2018) perform Monte Carlo experiments to create evidence on the performance of panel data estimators.

Not only choosing among the estimators to be used, but also determining which panel model to employ is crucial for an authentic panel data analysis. To illustrate it, Elhorst (2014) put forward that those different modifications are required to estimate the fixed and the random effects model extended to include endogenous interaction effects or interaction effects among the error terms.

As of today, econometric researchers have the chance to select among many models including commonly used ones like the fixed effects model, the random effects model, the fixed coefficients model, the random coefficients model, and the multilevel model. To illustrate it, in addition to mentioned models, Baltagi et al. (2003) were the first to consider the testing of spatial interaction effects in a spatial panel data model, and they provided a survey of the specification and estimation of spatial panel data models.

In this study, we try to provide basic information on the types of panel data, sorts of tools which should be used, in addition to explaining selected panel data tests and estimations. The study is organized as follows: Section 1 introduces the topic, Section 2 provides econometric formulations and the logic behind formulations, and Section 3 conveys the conclusions.

2. ECONOMETRIC FORMULATION

Linear models are still frequently referred in panel data applications while random sampling is used as a realistic assumption for the populations. When compared with linear models, one can assert that nonlinear models are harder to estimate as the problematic issue is how to tackle unobserved heterogeneity, which is a common issue for panel data econometrics as well.[1]

Social scientists continue to investigate the nature of panel data empirically using numerous methods for exploring the characteristics of these datasets. The development of empirical analysis on panel data requires the applications of several econometric methodologies. Since panel series have long time dimensions, and this allows the researchers to use cointegration and unit root testing to decide the order of integration and relationship between variables.

In this paper, details of the specific selected highly used standard panel data tests and estimations (Im-Pesaran-Shin unit root testing, Pedroni's cointegration test, panel data ordinary least squares) in this paper. First, for the heterogeneous panels Im, Pesaran & Shin (1997) test estimates the t-test for unit roots. Relying on the Dickey Fuller (DF) t-stat of each unit, the test supposes that, *under the null hypothesis*, non-stationarity exists for the series. The test acknowledges common time and individual effects and time trends. This unit root testing is still widely used in the literature.

Pedroni's cointegration test is a 2-step residual-based test based on Engle-Granger methodology. Pedroni extended the cointegration test by Engle-Granger which is predicated on an investigation of residuals through calculating I (1) variables. The residuals are expected to be I (0) in case there is cointegration between variables.

Ordinary least squares in panel data (PDOLS) estimator is an extension to the panel data case of the dynamic ordinary least squares estimator (DOLS) used in time series data, and has the same purpose of estimating a long run equilibrium relationship in non-stationary time series. This is implemented over nonstationary data which show a cointegrating relationship between variables. In short, *(i)* Im-Pesaran-Shin unit root test is employed to determine if the variables are stationary in variance, which is necessary since the variables entering a cointegrated panel must be non-stationary, *(ii)* the statistics proposed in Pedroni (1999) that will make it possible to determine whether the variables define a cointegrated panel and *(iii)* the method of dynamic PDOLS included in Pedroni (2001), with which we estimate the long-term relationship. *(iv)* we also explore error correction model (ECM) which solves the problem on how to integrate into a dynamic model both long-run information contained in variable levels (cointegration) and short-run information contained in variables in differences (stationarity).

[1] If the covariate appears to be persistent for the linear panels with small T, this can be considered as a significant problem for the linear model.

In the following section, we introduce the theoretical and foundational formulations associated with mentioned tests and estimations:

Im-Pesaran-Shin Unit Root Test

After scholars like Breitung & Meyer (1994), Quah (1994), Im et al. (2003), and Levin et al. (2002) formed new approaches to the univariate unit root tests to the panel data, the popularity of cointegration applications to the panel data have increased in the literature. [2] As scholars explored the advantages of panel data, more attention were given to the cointegration tests and estimation with regression models using panel data. Kao (1999), Pedroni (1996, 1997) and Phillips & Moon (1999) led the first works.

The leading studies in the unit root testing field by Breitung & Meyer (1994), Quah (1994), Levin and Lin (1993), Im et al. (1995), and Maddala & Wu (1999) aimed to determine whether economic data are stationary or integrated by implementing classic ways. To start with, Breitung & Meyer (1994) introduced an asymptotic normality of DF test for panel data, which involves a small time series component and a bigger cross-section dimension. Quah (1994) derived a unit root test for panel data with extensive time series and cross-section variations. Levin & Lin (1993) explored the asymptotic distributions for unit roots on panel data while Im et al. (1995) criticized the Levin and Lin (1993) stats and offered some alternative methods. Hadri (1999) claimed that it would be beneficial to apply tests of the null hypothesis of stationarity as well as tests of the null hypothesis of a unit root. To do this, Hadri (1999) extended the tests to panel data with i.i.d. disturbance terms across *i* and over *t*, and displayed how to accommodate the case of heterogeneous disturbance terms across *i*. In addition, Levin et al. (2002) used pooling cross-section time series data for testing the unit root hypothesis. The authors revealed, *thanks to the Monte Carlo simulations*, that for the panels of moderate size asymptotic findings produce a well-performing approximation to the test statistics, while the power of the panel-based unit root test is exceptionally high.

Im-Pesaran-Shin unit root test is one of the regarded unit root test for testing for stationarity in panel data analysis. [3] The reason why we decided to detail Im-Pesaran-Shin test in our research when compared to other tools is this test allows for unbalanced panels while many others presume that the dataset is balanced. The test by Im-Pesaran-Shin predicts *t* test for heterogeneous panels developed by while it allows for time trends and individual effects. The basic assumption is that the series under null are non-stationary. The major difference between Levin & Lin (1993) and Im-Pesaran-Shin is that Levin and Lin assumed all series are stationary under alternative while Im-Pesaran-Shin considers just some are stationary. The theoretical foundation of the test is as follows:

Let's consider a sample of *N* cross sections observed in *T* time periods and y_{it} generated by an autoregressive order process ($p_i + 1$), [4]

$$y_{it} = u_i\phi_i(1) + \sum_{j=1}^{p_i+1} \phi_{ij} y_{it-j} + \varepsilon_{it} i = 1 \ldots \mathrm{N}, \mathrm{t} = 1 \ldots \mathrm{T} \tag{2.1}$$

which can be parameterized as

$$\Delta y_{it} = \alpha_i + \beta_i y_{it-1} + \sum_{j=1}^{p_i} \rho_{ij} \Delta y_{it-j} + \varepsilon_{it} i = 1 \ldots \mathrm{N}, \mathrm{t} = 1 \ldots \mathrm{T} \tag{2.2}$$

where $\phi_i(1) = 1 - \sum_{j=1}^{p_i+1} \phi_{ij},\ \alpha_i = u_i\phi_i(1), \beta_i = -\phi_i(1)$

[2] Kao and Chiang (2001) stress that despite the studies on unit roots and cointegration in time series data, the interest on testing unit roots in panel data analysis were limited.

[3] Since Im-Pesaran-Shin test is widely used in the literatura for heterogenous panels, we decided to deliver information on it, when compared to other unit root testing tools.

[4] We do not provide the details of AR in this paper, as it is not the main scope of the study.

y.$\rho_{ij} = -\sum_{h=j+1}^{p_i+1} \phi_{ih}$

The null and alternative hypotheses will be given by

$H_0\!: \beta_i = 0 \ \ \forall i vs H_1\!: \beta_i < 0$ para i = 1… N$_1$ and $\beta_i = 0$ para i = N$_1$+1,…, N (2.3)

where the alternative hypothesis admits that β_i differs through cross sections. Small negative values will lead to the rejection of the null hypothesis.

It is crucial to note that the test does not provide perfect results in all conditions. Harris (2010) indicated that the researchers should be careful using IPS test when there is uncertainty concerning magnitude of initial conditions. The authors warned that the power of the test decrease when the magnitude of initial conditions is getting larger.

Pedroni's Cointegration Test

Various cointegration techniques have been extensively implemented in the empirical research recently. Orsal (2007) states that scholars started to create unit root and cointegration tests for the panel data because of the problems in finding low power of Augmented Dickey Fuller (ADF) and DF unit root test for the univariate cases and exploring long time series.

Im et al. (1995), Harris & Tzavalis (1996) and Phillips & Moon (1999) helped advances to occur in nonstationary panel analysis. As the number of time observations with large *T* and *N*, non-stationarity has become the matter for the researchers.

There are two major approaches for the panel cointegration tests have come to the forefront in econometrics of late years, which are residual-based (RB) and maximum-likelihood-based (ML). [5]

The scholars who have contributed to the residual-based panel cointegration test statistics were McCoskey & Kao (1998), Kao (1999), and Pedroni (1995, 1997, 1999). McCoskey & Kao (1998) developed tests for the null hypothesis of cointegration, which can be considered as a LM test extension developed in panel data. Kao (1999), on the other hand, derived DF and ADF type of tests and offered four variants of DF test stats within the framework of spurious regression for the panel data.

One of the most important residual-based panel cointegration test was developed by Pedroni (1999). It was in 1995 when Pedroni introduced first version of his residual-based panel cointegration tests. Pedroni (1996) proposed a fully modified estimator for heterogeneous panels. In 1997, he derived asymptotic distributions for residual-based tests of cointegration for both homogeneous and heterogeneous panels, while he extended his tests for the regression equations with two or more independent variables in 1999. Pedroni (1999) used 2 between dimension-based statistics (group-ρ and group-t) and 2 within-dimension-based (panel-ρ and panel-t) null of no cointegration panel. Phillips & Moon (1999) developed both sequential limit and joint limit theories for nonstationary panel data.

The scholars who introduced maximum-likelihood-based panel cointegration test statistics were Larsson & Lyhagen (1999), Larsson et al. (2001) and Groen & Kleibergen (2003). Grounding his research on cross-sectional independence, Larsson et al. (2001) proposed a panel cointegration test statistic. In addition, it was Groen & Kleibergen (2003) who introduced the estimation methodology for homogenous and heterogeneous cointegration vectors through a maximum-likelihood framework using the Generalized Method of Moments (GMM) procedure.

[5] In order to estimate nonlinear panel data models, maximum likelihood estimation is a frequently used method.

Arellano & Bond (1991) introduced GMM estimator relying on the small T panel estimation that is based on random and fixed effects estimators or on the combination of instrumental variable and fixed effects. The individual groups were pooled in this estimation while intercepts differ in the groups. [6]

As indicated, Pedroni (2003) introduced a set of statistics that allow testing the null hypothesis of non-cointegration in heterogeneous panels with multiple non-stationary regressors. Applying seven test statistics, Pedroni tested cointegration among the regressors for nonstationary heterogenous panels which are large as *N* and long as *T.* The test by Pedroni may involve unbalanced panels and common time dummies.

The statistical tests are constructed from the residues of the cointegration regression, as follows:

$$y_{it} = \beta_i + \ \beta_{1i}x_{1it} + \beta_{2i}x_{2it} + \cdots + \beta_{Mi}x_{Mit} + \varepsilon_{it} \tag{2.4}$$

on a sample of N cross sections observed in T time periods. It is admitted that the slopes $\beta_1, \beta_2, \dots, \beta_M$ vary through the cross sections, with β_i the individual effects.

As specified, Pedroni developed seven statistics, four of them grouped according to what the author called within-dimension and the other three grouped according to what he called between-dimension. From the estimated residues of equation (2.4) the auxiliary regression is proposed

$$\hat{e}_{it} = \hat{\delta}_i \hat{e}_{it-1} + \sum_{k=1}^{K_i} \hat{\delta}_{ik} \Delta\, \hat{e}_{it-k} + \hat{u}_{it}^{*} \tag{2.5}$$

For within-dimension statistics the null hypothesis of non-cointegration and the alternative hypothesis are given by:

$$H_0 : \delta_i = 1 \ \ \forall i \quad \text{vs} \quad H_1 : \delta_i = \delta < 1 \ \ \forall i \tag{2.6}$$

on the other hand, for between-dimension statistics the null hypothesis of non-cointegration and the respective alternative hypothesis are given by

$$H_0 : \delta_i = 1 \ \ \forall i \quad \text{vs} \quad H_1 : \delta_i < 1 \ \ \forall i \tag{2.7}$$

this alternative hypothesis is less restrictive as it allows heterogeneity between panels; within-dimension statistics will be referenced as panel cointegration statistics and between-dimension statistics will be referenced as group mean cointegration statistics. We now introduce the explicit definition of each statistics:

1. Panel v-statistic

$$T^2 N^{\frac{3}{2}} Z_{\hat{V}\,N,T} \equiv T^2 N^{\frac{3}{2}} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \hat{L}_{11i}^{-2} \hat{e}_{i,t-1}^{2} \right)^{-1} \tag{2.8}$$

2. Panel ρ- statistic

$$T\sqrt{N} Z_{\hat{p}\,N,T-1} \equiv T\sqrt{N} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \hat{L}_{11i}^{-2} \hat{e}_{i,t-1}^{2} \right)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{L}_{11\,I}^{-2} \left(\hat{e}_{i,t-1} \Delta \hat{e}_{i,t} - \hat{\lambda}_i \right) \tag{2.9}$$

[6] Some other techniques called the mean-group (MG) and pooled mean-group (PMG) estimators were introduced by Pesaran, Shin, and Smith (1997, 1999) in order to estimate nonstationary dynamic panels which have heterogeneous parameters.

3. Panel t- statistic (Non parametric)

$$Z_{tN,T} \equiv \left(\tilde{\sigma}_{N,T}^{2} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{L}_{11i}^{-2} \hat{e}_{i,t-1}^{2}\right)^{\frac{-1}{2}} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{L}_{11\,I}^{-2} \left(\hat{e}_{i,t-1} \Delta \hat{e}_{i,t} - \hat{\lambda}_{i}\right) \qquad (2.10)$$

4. Panel t- statistic (parametric)

$$Z^{*}{}_{t\,N,T} \equiv \left(\tilde{S}_{N,T}^{2} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{L}_{11i}^{-2} \hat{e}_{i,t-1}^{2}\right)^{\frac{-1}{2}} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{L}_{11\,I}^{-2} \left(\widehat{e^{*}}{}_{i,t-1} \Delta \hat{e}_{i,t}{}^{*} - \hat{\lambda}_{i}\right) \qquad (2.11)$$

5. Group p- statistic

$$TN^{\frac{-1}{2}} \tilde{Z}_{\hat{p}N,T^{-1}} \equiv TN^{\frac{-1}{2}} \sum_{i=1}^{N} \left(\sum_{t=1}^{T} \hat{e}_{i,t-1}^{2}\right)^{-1} \sum_{t=1}^{T} \left(\hat{e}_{i,t-1} \Delta \hat{e}_{i,t} - \hat{\lambda}_{i}\right) \qquad (2.12)$$

6. Group t- statistic (Non parametric)

$$N^{\frac{-1}{2}} \tilde{Z}_{tN,T} \equiv N^{\frac{-1}{2}} \sum_{i=1}^{N} \left(\hat{\sigma}_{i}^{2} \sum_{t=1}^{T} \hat{e}_{i,t-1}^{2}\right)^{\frac{-1}{2}} \sum_{t=1}^{T} \left(\hat{e}_{i,t-1} \Delta \hat{e}_{i,t} - \hat{\lambda}_{i}\right) \qquad (2.13)$$

7. Group t- statistic(parametric)

$$N^{\frac{-1}{2}} \widetilde{Z^{*}}{}_{tN,T} \equiv N^{\frac{-1}{2}} \sum_{i=1}^{N} \left(\sum_{t=1}^{T} \hat{S}_{i}^{*2} \hat{e}_{i,t-1}^{*2}\right)^{\frac{-1}{2}} \sum_{t=1}^{T} \hat{e}_{i,t-1}^{*} \Delta \hat{e}_{i,t}^{*} \qquad (2.14)$$

where $\hat{L}_{11i}^{2}$ is the estimation of the long run variance of the residuals and $\hat{\sigma}_{i}^{2}$and $\hat{S}_{i}^{*2}$ are the individual contemporaneous and long run variances respectively of the residuals.

All tests follow N (0,1) distribution. As for the rule of decision of the tests, for panel v-statistical large positive values of the normal distribution lead to the rejection of the null of non-cointegration, for the other six statistic very small negative values of the normal distribution lead to the rejection of the null hypothesis of non-cointegration.

Dynamic Ordinary Least Squares in Panel Data (PDOLS)

Pedroni (2001) developed a PDOLS technique to improve the DOLS method created by Stock & Watson (1993). Although there would be endogenous regressors, PDOLS estimators are normally distributed and asymptotically unbiased. As per Pedroni (2001), the group-mean PDOLS estimator has high robustness even to the omission of the variable.

Before this, a panel fully modified OLS test was introduced by Pedroni in 1996. Pedroni (1996) explored the asymptotic distributions and investigated three types of such estimators: The residual-FM, and the adjusted-FM, pooled the data along the within-dimension. The third one, which is the group-FM, pooled the data along the between-dimension. Kao & Chiang (1997) introduced a parametric DOLS based panel estimator pooled along the within-dimension, and proved that this had the same asymptotic distribution as the Pedroni's panel FMOLS estimator. In 1999, Kao (1999) worked on a spurious regression in panel data together with OLS estimator's asymptotic properties.

Pedroni (2000) identified that group-FM estimator (compared with the other two estimators) displayed minor size distributions in small samples.

Given the panel data model,

$$y_{i,t} = \alpha_i + \beta_i x_{i,t} + u_{it} \tag{2.15}$$

the PDOLS estimator is based on the following DOLS regression for each cross section

$$y_{i,t} = \alpha_i + \beta_i x_{i,t} + \sum_{j=-P}^{P} \gamma_{i,j} \Delta x_{i,t-j} + u_{it}^{*} \tag{2.16}$$

where i = 1, 2,…, N is the number of cross sections, t = 1, 2,…, Tis the number of time periods, P is the number of lags and leads of the DOLS regression, β_i is the slope and $x_{i,t}$ are the explanatory variables.

Estimators of β coefficients and t-statistics are obtained by averaging over the cross sections with Pedroni's group mean method. The estimators are the following:

$$\hat{\beta}_{GM}^{*} = \left[\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=1}^{T} z_{i,t} z_{it}'\right)^{-1}\left\{\sum_{t=1}^{T} z_{i,t}(y_{i,t} - \bar{y}_i)\right\}\right] \tag{2.17}$$

$$t_{\hat{\beta}_{GM}^{*}} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} t_{\hat{\beta}_i^{*}} \tag{2.18}$$

$$t_{\hat{\beta}_i^{*}} = (\hat{\beta}_i^{*} - \beta_0)\left\{\hat{\sigma}_i^{-2}\sum_{t=1}^{T}(x_{i,t} - \bar{x}_i)^2\right\}^{\frac{1}{2}} \tag{2.19}$$

where $z_{i,t}$ is the regressor vector that includes the lags and leads of the explanatory variables in differences and $\hat{\sigma}_i^2$ is the long-term variance of the residues u_{it}^{*}.

PDOLS estimator is averaged along the between-dimension according to the null hypothesis of the Pedroni cointegration test H_0: $\beta_i = \beta_0$vs H_A: $\beta_i \neq \beta_0$, which allows estimating a long run relationship for each cross section.

2.4. Single Equation Error Correction Model

As mentioned, Pedroni’s cointegration test is a 2-step residual-based test based on Engle-Granger methodology, which is predicated on an investigation of residuals through calculating I(1). At this juncture, we think it is important to provide the basics of Engle-Granger’s Error Correction Model (ECM) which has become one of the foundations of modern econometrics in terms of model development, among other things because it solves the problem of how to integrate into a dynamic model both long-run information contained in variable levels (cointegration) and short-run information contained in variables in differences (stationarity).

Engle & Granger (1987) were the formalizers of the error correction model, whose formulation for the simple case of two variables y_t and x_t both with order of integration I (1) and cointegrated would be the following:

$$\Delta y_t = \alpha(y_{t-1} - \beta x_{t-1}) + \sum_{i=1}^{n} \delta_i \, \Delta y_{t-i} + \sum_{i=0}^{m} \gamma_i \, \Delta x_{t-i} + \varepsilon_t \tag{2.20}$$

with Δ denoting the first difference of the variable. Certain considerations are relevant:

Note that since all variables are expressed in first differences and y_t and x_t are I(1) the first difference defines a stationary variable. On the other hand, since y_t and x_t are cointegrated, the linear combination (y_{t-1}-βx_{t-1})

will also be stationary, hence in an error correction model with integrated variables of order one and cointegrated all variables will be stationary and therefore the estimation through OLS is valid[7].

Specifically, the term (y_{t-1}-βx_{t-1}) defines the ECM and picks up the information of the variables in the model that is lost when applying the first difference; on the other hand, it also allows capturing the impact on Δy_t generated by imbalances between y_t and x_t. This type of imbalance can arise due to errors of economic agents in past decisions and the presence in the model of the linear combination (y_{t-1}-βx_{t-1}) reflects the attempts of the agents to correct such errors in the current period through the variable Δy_t, therefore, the α adjustment coefficient must be negative; at the same time, it is the reason why the Equation #2.20 is called "Error Correction Model".

Strictly speaking, the estimation of an error correction model requires that the study of the stationarity of the variables have been previously carried out through some unit root test as well as the cointegration analysis of those variables, because its formulation includes at least one error correction mechanism and variables in differences.

3. CONCLUSION

Although panel data econometrics is a very broad field, and it will not be possible to include all methods in this study; this article focuses on analyzing the basic framework of how to construct econometric analysis based on heterogenous non-stationary dynamic panel datasets. In this study, we aim introduce details of the specific selected of highly used panel data tests and estimations (Im-Pesaran-Shin unit root testing, Pedroni's cointegration test, panel data ordinary least squares) for standard empirical research and explain why and under which conditions these methods are applied. We represent theoretical formulations of these tests and estimations along with Engle and Granger's error correction mechanism in order to determine the order of integration and long-run relationship between the variables. By this, we are hoping to commentate the basic steps for a straightforward empirical panel data research process for junior researchers in different branches of social sciences.

The types of panel data, number of variables included in the analysis, weighting issues are very decisive in determining which methods to apply in panel data research. In this framework, we provide information on the definitions of panel data types, and sorts of investigations that should be applied depending on these types, in addition to foundations of selected tests and estimations for heterogeneous panels.

To be able to perform an established panel data analysis, applying valid unit root tests, checking for cointegration, and following the most appropriate estimation methodology are the keys. In order to choose the best performing tools, scholars first should determine whether panel data analysis may be applied in investigating the economic issue under consideration. The limitations shall be considered to understand if specific econometric tools to analyze panel data might be meaningful, in addition to the assumptions underlying the statistical inference procedures and the data generating process which may not be always fully compatible. The scholars shall also brainstorm on how to surge the efficiency of the estimators.

Despite its various advantages, it should be kept in mind that the strength of the panel data to isolate impacts of particular treatments and actions, is very much related with the data generating process and compatibility of the assumptions of statistical tools.

One major limitation of our research might be strong focus on the basics of the panel data research, which might have caused us to miss the details on panel data econometrics, although the major scope of this paper is not to convey all technicalities regarding panel data research in social sciences.

Acknowledgements This paper is based on the second chapter of my Ph.D. thesis submitted to Department of Economics at the University of Rome "Tor Vergata". I wish to thank Prof. Pasquale Scaramozzino, Assoc. Prof. Barbara Annicchiarico and other anonymous referees for very helpful comments and insightful guidance. The views expressed in this study are solely mine and do not necessarily reflect those of my affiliations. I declare that I have no conflict of interest.

[7] Among other things due to the fact that all stationary variables prevent the possibility of spurious regression.

REFERENCES

Anderson, T.W., & Hsiao C. (1982). Formulation and estimation of dynamic models using panel data. *Journal of Econometrics*, 18, 47-82.

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

Balestra, P., & Nerlove M. (1996). Pooling cross-section and time series data in the estimation of a dynamic model: The demand for natural gas. *Econometrica*, 34, 585–612.

Baltagi, B., & Kao, C. (2000). Nonstationary panels, cointegration in panels and dynamic panels: A survey. *Advances in Econometrics*, 15, 7–51.

Baltagi, B., Song, S., & Koh, W. (2003). Testing panel data regression models with spatial error correlation. *Journal of Econometrics*, 117, 123-150.

Berrington, A., Smith, P., & Sturgis, P. (2006). An overview of methods for the analysis of panel data. *ESRC NCRM Methods Review Papers*, 007.

Bond, S.R. (2002). Dynamic panel data models: A guide to micro data methods and practice. *Portuguese Economic Journal*, 1, 141–162.

Breitung, J., & Meyer, W. (1994). Testing for unit roots in panel data: Are wages on different bargaining levels cointegrated? *Applied Economics*, 26, 353–361.

Corrado, L., & Fingleton, B. (2011). Where is the economics in spatial econometrics? *Scottish Institute for Research in Economics Discussion Papers*, 2011-02,

Dickey, D.A., & Fuller, W.A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica*, 49, 1057–1072.

Elhorst J. P. (2014). *Spatial Econometrics*, Springer Briefs in Regional Science.

Engle, R. & Granger, C. (1987). Cointegration and error correction: Representation, estimation and testing. *Econometrica*, 55, 251-276.

Groen, J. J., & Kleibergen, F. (2003). Likelihood-based cointegration analysis in panels of vector error correction models. *Journal of Business and Economic Statistics*, 21(2), 295–318.

Hadri K. (2000). Testing for stationarity in heterogeneous panel data. *The Econometrics Journal*, 148-161.

Harris, R., & Tzavalis, E. (1996). Inference for unit root in dynamic panels. *Unpublished manuscript.*

Harris, D., Harvey, D. I., Leybourne, S. J., & Sakkas, N. D. (2010). Local asymptotic power of the Im-Pesaran-Shin panel unit root test and the impact of initial observations. *Econometric Theory*, Cambridge University Press, 26(1), 311-324.

Heckman, J.J. (1981). Heterogeneity and state dependence. *National Bureau of Economic Research (NBER) Chapters* in: Studies in Labor Markets, 91-140.

Honoré B. E., & Kyriazidou E. (2000). Panel data discrete choice models with lagged dependent variables. *Econometrica*, 68(4), 839-874.

Hsiao, C. (2005). Why panel data? *The Singapore Economic Review*, 50 (2), 143–154.

Im, K., Pesaran, H., & Shin, Y. (1995). Testing for unit roots in heterogeneous panels. *Manuscript, University of Cambridge*.

Im, K., Pesaran, H., & Shin, Y. (2003). Testing for unit roots in heterogeneous panels. *Journal of Econometrics*, 115(1), 53–74.

Kao, C. (1999). Spurious regression and residual-based tests for cointegration in panel data. *Journal of Econometrics*, 90, 1–44.

Kao, C., & Chiang M-H. (1997). On the estimation and inference of a cointegrated regression in panel data. *Syracuse University manuscript.*

Kao, C., & Chiang, M.-H. (2001). On the estimation and inference of a cointegrated regression in panel data", Baltagi, B.H., Fomby, T.B. and Carter Hill, R. (Ed.) *Nonstationary Panels, Panel Cointegration, and Dynamic Panels* (Advances in Econometrics, Vol. 15*)*, 179-222. Emerald Group Publishing Limited, Bingley.

Larsson, R., & Lyhagen, J. (1999). Likelihood-based inference in multivariate panel cointegration models. *Stockholm School of Economics Working Paper Series in Economics & Finance*, No.331.

Larsson, R., Lyhagen, J., & Løthgren, M. (2001). Likelihood-based cointegration tests in heterogeneous panels. *Econometrics Journal,* 4, 109–142.

Levin, A., & Lin, C. F. (1993). *Unit root tests in panel data: New Results*. San Diego, CA: University of California.

Levin, A., Lin, C. F., & Chu, C.S. (2002). Unit root tests in panel data: Asymptotic and finite-sample properties. *Journal of Econometrics,* 108, 1–24.

Maddala, G. S., & Wu, S. (1999). A comparative study of unit root tests with panel data and a new simple test: Evidence from simulations and the bootstrap. *Oxford Bulletin of Economics and Statistics,* 61, 631–652.

McCoskey, S., & Kao, C. (1998). A residual-based test of the null of cointegration in panel data. *Econometric Reviews,* 17, 57–84.

Moundigbaye, M., Rea W., & Robert R. W. (2018). Which panel data estimator should I use? A corrigendum and extension. *Economics Discussion Papers,* No. 2017-58. Kiel Institute for the World Economy.

Örsal, D. D. K. (2007). Comparison of panel cointegration tests. *SFB 649 Discussion Paper,* No. 2007,029, Humboldt University of Berlin.

Quah, D. (1994). Exploiting cross section variation for unit root inference in dynamic data. *Economics Letters,* 44, 9–19.

Pedroni, P. (2001). Fully modified OLS for heterogeneous cointegrated panels. In B. H. Baltagi, T. B. Fomby, & R. Carter Hill (Eds.), *Nonstationary Panels, Panel Cointegration, and Dynamic Panels* (Advances in Econometrics, Vol. 15), 93-130. Bingley: Emerald Group Publishing Limited.

Pedroni, P. (1995). Panel cointegration; asymptotic and finite sample properties of pooled time series tests, with an application to the PPP Hypothesis. *Indiana University Working Papers in Economics,* No.95-013.

Pedroni, P. (1996). Fully modified OLS for heterogeneous cointegrated panels and the case of purchasing power parity. *Working Paper,* No. 96–20, Indiana University.

Pedroni, P. (1997). On the role of cross-Sectional dependency in panel unit root and panel cointegration exchange rate studies. *Working Paper by Indiana University.*

Pedroni, P. (1999). Critical values for cointegration tests in heterogeneous panels with multiple regressors. *Oxford Bulletin of Economics & Statistics,* 61, 653–670.

Pedroni, P. (2001). Purchasing power parity tests in cointegrated panels. *Review of Economics and Statistics,* 83(4), 727–731.

Pesaran, M. H., Shin Y., & Smith R. P. (1997). Estimating long-run relationships in dynamic heterogeneous panels. *DAE Working Papers,* Amalgamated Series No. 9721.

Pesaran, M. H., Shin Y., & Smith R. P. (1999). Pooled mean group estimation of dynamic heterogeneous panels. *Journal of the American Statistical Association*, 94, 621-634.

Phillips, P. C. B. & Moon, H. R. (1999). Linear regression limit theory for nonstationary panel data. *Econometrica,* 67, 1057–1111.

Stock, J. H., & Watson, M. W. (1993). A simple estimator of cointegrating vectors in higher order integrated systems. *Econometrica: Journal of the Econometric Society*, 61, 783– 820.