Chapter 3

Research Methodology

The purpose of this chapter is to describe the research methodology used in this dissertation. In particular, there are four sections. First, the research design discusses the model that developed from the conceptual framework. Second, the testable hypotheses are provided and next we will present the data and sample selection provides sources of data, the unit of analysis, and the sample size. Finally, data analysis methods are discussed.

3.1 Research Design

From the conceptual framework, this thesis developed the model to analyze factors that determined economic output or gross domestic product (GDP) in developed and developing countries. Moreover, this research can be dividing to two parts. First part is the study the factors affecting economic growth in developing country and second part, we interest in developed country. The details of two parts are explained below.

Part I

To analyze the causal relationship among gross domestic product, macroeconomic, social and political variables in developing country, we first classify data for 95 countries into 11 distinct regions based on continent, climate, and access to sea-lane. We then seek to isolate the intercept and slope shifters of economic growth in four stages. Stage one tests a standard economic model composed only of the interest rate, exchange rate, money supply, tourism, inflation, save, trade, export/import, FDI inflow, capital and labour. Stage two adds 10 regional dummy variables to determine which, if any of the regions are significantly higher or lower than the suppressed base region (Southeast Asia). Stage 3 then adds slope-shifting

interaction terms between each region and the economic variables to determine which macro variables in which regions display significantly different marginal impacts on growth. Step 4 extends the model of stage 3 by adding the socio-political variables which are schooling, political freedom, transparency (i.e. absence of corruption), and criminality. The model of step 4 is inspired by the new institutional economics in general and by the "sufficiency" economy model of the King of Thailand and the gross national happiness paradigm of the King of Bhutan, which posit that true development is inconsistent with an increase in criminality, corruption, and political or educational disenfranchisement. Finally, based on the significant results of each stage of the analysis, we shall draw practical conclusions for development policy by region and for the developing economies as a whole.

This thesis will successively develop and estimate four macroeconomic models, starting from what we shall term the "standard macroeconomic model[2]":

GDP = *f* (money, interest, exchange, inflation, save, trade, exports/imports, FDI inflow, capital, labour, tourism) (3.1)

and ending with the completely specified "sufficiency economy-inspired model":

GDP = *f* (money, interest, exchange, inflation, save, trade, exports/imports, FDI inflow, capital, labour, life, schooling, lack of freedom, transparency, crime, HDI, 11 regions, interaction terms between macro and social/political variables, interaction terms between regions and macro/social/political variables)

(3.2)

[2] There are two main themes in standard macroeconomics: growth and the business cycle. In the developing economies, the issue of, and potential for, growth are of such vital importance that growth must be integrally included in any complete definition of a macro model for policy orientation. At the same time, however, the question of short-term stabilization is equally vital in that a) such a large percentage of the population lives so close to the edge of poverty and starvation, and because of b) the vulnerability of many/most developing economies to swings in trade and the exchange rate. Our model is thus a balanced hybrid of the major aspects of growth and stabilization variables as commonly applied to Western economies.

or the standard macroeconomic model and Sufficiency Economy Inspired Model can be specified as follow:

Standard Macroeconomic Model

$$\ln GDP_{i,t} = \alpha_i + \beta_{1i} \ln Money_{i,t} + \beta_{2i} \ln Interest_{i,t} + \beta_{3i} (\ln Interest_{i,t})^2 + \beta_{4i} \ln Exchange_{i,t} + \beta_{5i} \ln Inflation_{i,t} + \beta_{6i} (\ln Inflation_{i,t})^2 + \beta_{7i} \ln Save + \beta_{8i} \ln Trade_{i,t} + \beta_{9i} \ln Export / import_{it} + \beta_{10i} \ln FDI \inf low_{i,t} + \beta_{11i} \ln Capital_{i,t} + \beta_{12i} \ln Labour + \beta_{13i} \ln Tourism_{i,t} + \varepsilon_{i,t} \quad (3.3)$$

Sufficiency Economy Inspired Model

$$\ln GDP_{i,t} = \alpha_i + \beta_{1i} \ln Money_{i,t} + \beta_{2i} \ln Interest_{i,t} + \beta_{3i} (\ln Interest_{i,t})^2 + \beta_{4i} \ln Exchange_{i,t} + \beta_{5i} \ln Inflation_{i,t} + \beta_{6i} (\ln Inflation_{i,t})^2 + \beta_{7i} \ln Save + \beta_{8i} \ln Trade_{i,t} + \beta_{9i} \ln Export / import_{it} + \beta_{10i} \ln FDI \inf low_{i,t} + \beta_{11i} \ln Capital_{i,t} + \beta_{12i} \ln Labour + \beta_{13i} \ln Tourism_{i,t} + \beta_{14i} \ln Life_{i,t} + \beta_{15i} (\ln Life_{i,t})^2 + \beta_{16i} \ln School_{i,t} + \beta_{17i} \ln Lag\ of\ Freedom_{i,t} + \beta_{18i} \ln Transparency_{i,t} + \beta_{19i} \ln Crime_{i,t} + \beta_{20i} \ln HDI_{i,t} + \varepsilon_{i,t} \quad (3.4)$$

where

GDP = Log of gross domestic product at constant price.

Money = Log of nominal money supply is the sum of currency outside banks and demand deposits other than those of the central government. A positive coefficient is expected, as money supply has been shown to be positively related with GDP.

Interest = Log of nominal interest rate. Since lowering the interest rate encourages investors to invest more frequently and in greater amounts to GDP, a negative coefficient is expected. Moreover, a coefficient of the square of interest is used to capture a possible upward turn in the relationship between interest rate and GDP[3].

Inflation = Log of inflation rate with base 2000. A negative coefficient is expected, as high inflation has been found to negatively affect growth.

[3] The relationship between interest rate and GDP are assumed to be non-linear which consist of Fry (1997) estimates that non-linear functional forms of interest rate and economic growth, allowing higher interest rates to promote growth over some low range of values but then hinder it over a higher range.

Moreover, a coefficient of the square of inflation is used to capture a possible upward turn in the relationship between inflation and GDP, leading to a possible "optimal" inflation rate[4].

Exchange = Log of national currency per US dollar, or the nominal exchange rate. A positive coefficient is expected, as either a high exchange rate or depreciation in the domestic currency has been found to increase exports and hence GDP.

Save = Log of the savings rate as a percent of gross national income. A positive coefficient is expected, as a higher savings rate has been found to positively affect growth.

Trade = Log of the sum of exports and imports of goods and services as a percent of GDP. Assuming that openness to international trade is beneficial to economic growth, a positive coefficient is expected.

Exports/imports = Log of Export-Import ratio. Assuming that more exports compared to imports is beneficial to economic growth, a positive coefficient is expected.

FDI inflow = Log of foreign direct investment as a percent of GDP. Assuming that more foreign direct investment inflow is beneficial to economic growth, a positive coefficient is expected.

Capital = Log of gross capital formation as a percent of GDP. Assuming that increasing the level of gross capital formation is beneficial to economic growth, a positive coefficient is expected.

Labour = Log of the total labour force. A positive coefficient is expected, as an increase in the labour force has been found to positively affect productivity and lead to higher GDP.

Tourism = Log of international tourism expenditures by international inbound visitors from other countries. Tourism expenditure is one kind

[4] The non-linearity is assumed which low level of inflation, the income can increase when level of inflation increase but for high inflation rate, the increase in inflation rate will reduce the country's income. Li (2006) and Hasanov (2011) suggest there exist some threshold that the inflation has a significantly positive effect on growth at below threshold level, but the relationship is negative when inflation rates are above threshold level.

of export which increasing in tourism expenditure can encourage GDP, therefore positive impact is assumed.

Life = Log of life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life. As higher life expectancy can keep workers in the workforce throughout their full potential careers, a positive coefficient is expected. Moreover, a coefficient of the square of life expectancy is used to capture a possible downward turn in the relationship between life expectancy and GDP as the population ages.

School = Log of the ratio of school enrollment in secondary school measured as a share of the gross enrollment ratio. Greater enrollment ratios lead to greater human capital, which should be positively related to GDP. A positive coefficient is expected.

Lack of Freedom = Log of political rights based on a 40-point scale with 0 representing the highest and 40 the lowest level of freedom. A negative coefficient is expected, as a lag in freedom means lower political rights, which can in turn reduce productivity and GDP.

Transparency = Log of Corruption Perceptions Index (CPI) on a scale from 10 (highly clean) to 0 (highly corrupt). Corruption (lower in corruption index) reduces capital productivity which also reduces GDP; we therefore assumed a positive effect.

Crime = Log of intentional homicide rates per 100,000 population. We assumed a negative effect, i.e., that an increasing crime rate leads to greater economic cost and reduced GDP.

HDI = Log of the Human Development Index. As a rising Human Development Index can lead to increasing labour productivity and GDP, we assumed a positive effect.

Part II

This part will present the relationship among gross domestic product, macroeconomic, social and political variables in developed country with deal section sample bias issue. The structure of the sample selection model is a two equation system: the first equation is the outcome equation which can be the same as (3.3) or (3.4) and the second equation is the selection equation which can be written as:

$$s_{i,t} = \gamma_0 + \gamma_1 H\exp ort_{i,t} + \gamma_2 Hfree_{i,t} + \gamma_3 GNIpc_{i,t} + \gamma_4 Health_{i,t} + \upsilon_{i,t} \quad (3.5)$$

where

$s_{i,t}$ = $\begin{cases} 1 & if \quad country\ is\ define\ to\ be\ a\ developed\ country \\ 0 & if \quad country\ is\ define\ to\ be\ a\ developing\ country \end{cases}$

$H\exp ort_{i,t}$ = $\begin{cases} 1 & if \quad country\ have\ a\ ratio\ of\ \exp ort - import\ greater\ than\ 1 \\ 0 & if \quad otherwise \end{cases}$

Assuming that more exports compared to imports is beneficial to income of country which can lead the probability that country to be a developed country, a positive coefficient is expected.

$Hfree_{i,t}$ = $\begin{cases} 1 & if \quad country\ is\ define\ to\ have\ a\ high\ level\ of\ economic\ freedom \\ 0 & if \quad otherwise \end{cases}$

A positive coefficient is expected, as an increase in level of freedom means higher political rights, which can in turn encourage productivity and GDP and country will become to a developed country.

$GNIpc_{i,t}$ = Gross National Income per capita (constant US dollar). As a rising Gross National Income per capita can lead to increasing probability that country to be a developed country, we assumed a positive effect.

$Health_{i,t}$ = Health expenditure (percentage of gross domestic product). A positive coefficient is expected, as higher health expenditure has been found to positively affect probability that country to become a developed country.

3.2 Research Hypotheses

Based on the review literature, conceptual framework and objectives, this research will, across three separate research papers, test the following specific empirical hypotheses:

H1. The macroeconomic variables are important to determine change in output in both of developed and developing countries.

H2. The social and political variable is significant to explain change in economic output in both of developed and developing countries.

H3. The policy implication is different in 11 regions of developing countries and also in developed economies.

H4. Money supply has positively impact on Gross Domestic Product.

H5. The relationship between interest rate and output are non-linear and there exists negative impact of interest rate on output and also coefficient of the square of interest rate is positive.

H6. There exists a positive relation between exchange rate and national income and depreciation in domestic currency will lead to increase in export income and also national income.

H7. There exists non-linear relation between inflation and output which lower inflation can encourage economic growth while high inflation reducing economic growth.

H8. There are positive relationship between saving rate and economic output.

H9. The international trade can increase national income. So the relationship is positive.

H10. Surplus of export compare to import is beneficial to economic growth, hen there are positive impact of export-import ratio on economic growth.

H11. Foreign direct investment (FDI) and economic growth points to a positive FDI-growth relationship.

H12. There is a linkage between capital formation and economic growth and increase in capital formation can lead to increase in economic output.

H13. Labour supply is one of determinants of the feature of national income. The relationship between labour supply and national income is positive and increase in labour force will boost productivity and raise national income.

H14. Tourism expenditure has positive impact on Gross Domestic Product.

H15. Life expectancy can encourage economic growth and increase in life expectancy leads to an increase in health of the population and increase in productivity of labour and hence raise in output.

H16. There are positive effects of enhanced human capital formation or school enrollment to economic growth.

H17. Lack of freedom can reduce the productivity and Gross Domestic Product which there exists negative relationship between lack of freedom and Gross Domestic Product.

H18. Corruption will reduce competition and reduce private investment and, hence, the stock of producible inputs in the long run. The relationship between corruption and economic growth is negative.

H19. Crime increase economic cost and reduce Gross Domestic Product and there exist the negative impact of crime on economic growth.

H20. There exists positive relation between Human Development Index and economic growth.

H21. There exists sample selection bias which indicates the traditional indicator of development is not appropriate.

H22. Increase in Gross National Income per capita, health care expenditure, high of export, and high of economic freedom can help country to become developed country.

3.3 Data and Sample Selection

The study draws upon multiple data sources for annual data spanning the period 1996 to 2008 on a host of macroeconomic, social and political indicators for a sample of 95 developing countries drawn from Central and Eastern Europe, Middle East, Latin America, the Commonwealth of Independent States, Asia and Sub-Saharan Africa and a sample of 22 developed countries (Table 3.1). The countries are divided according to 2008 GNI per capita which are provided by World Bank. The developing countries had income below $12,275 and developed countries had income above $12,276.

To test for non-homogeneity within the sample of developing country, Asia is subdivided into South Asia, Southeast Asia and the socialist emerging economies of China and Vietnam; while Africa is divided into four north-south/coastal-interior groupings.

Some social indicators such as schooling, lack of freedom, transparency, crime and HDI were not available for a uniform period for each country. Consequently, the number of observations varied across our sample countries, leading us to conduct estimations over an incomplete panel data.

All series were obtained from the *International Monetary Fund's International Financial Statistics (IFS)* (IMF,2009), the *World Development Indicators* 2010 (WDI) database and the Central Bank of each country. Level of political freedom data were acquired from *Freedom in the World data* (,2011), the Corruption Perceptions Index from *Transparency International* (2011), Intentional homicide rates per 100,000 population from the *Geneva Declaration on Armed Violence and Development report* (2008) and the Human Development Index from the *Human Development Report* (2010). All variables were converted into natural logarithms prior to the empirical analysis.

Table 3.1: List of countries by region

Type of Country	List of Countries
Developed Country	Australia, Austria, Canada, Cyprus, Czech Republic, Denmark, Finland, Hong Kong , Iceland, Israel, Japan, Korea, Malta, New Zealand, Norway, Singapore, Slovak Republic, Slovenia, Sweden, Switzerland, United Kingdom, United States
Developing Country	**Middle East** Algeria, Egypt, Iran, Mauritania, Morocco, Saudi Arabia, Tunisia, Republic of Yemen **Central and Eastern Europe** Albania, Bulgaria, Croatia, Estonia, Hungary, Latvia, Lithunia, Macedonia, Poland, Romania, Turkey **Latin America (LA)** Argentina, Belize, Bolivia, Brazil, Chile, Colombia, Costa Rica , Dominican Republic, Guatemala, Guyana, Haiti, Honduras, Jamaica, Mexico, Paraguay, Peru, Trinidad and Tobago, Uruguay, Venezuela **Commonwealth of Independent States (CIS)** Armenia, Azerbaijan, Belarus, Kazakhstan, Kyrgyz Republic, Georgia, Russia, Ukraine **South Asia (SA)** Bangladesh, India, Nepal, Pakistan, Sri Lanka **Southeast Asia (SEA)** Cambodia, Fiji, Indonesia, Malaysia, Myanmar, Papua New Guinea, Philippines, Solomon Islands, Thailand, Lao People's Dem.Rep **Socialist emerging Asia (CHVN)** China, Vietnam **Northern coastal Africa (NCA)** Cameroon, Equatorial Guinea, Ghana, Nigeria, Cape Verde, Senegal, Sierra Leone, Côte d'Ivoire, Togo **Southern coastal Africa (SCA)** Gabon, Kenya, Madagascar, Mauritius, Mozambique, Namibia, Seychelles, South Africa, Tanzania **Northern interior Africa (NIA)** Burkina Faso, Central African Rep., Ethiopia, Mali **Southern interior Africa (SIA)** Botswana, Burundi, Republic of Congo, Lesotho, Malawi, Rwanda, Swaziland, Uganda, Zambia, Zimbabwe

Source: Adapted from IMF (2009).

Descriptive statistics of the variables included in the tables of results are shown in Table 3.2 and Table 3.3.

Table 3.2: Descriptive Statistics of 95 developing countries

Variable	Obs.	Mean	St. Dev.	Min.	Max
GDP (billion $ US)	1235	276.67	719.11	1.15	4327.45
Macroeconomic Indicator					
MONEY (billion $ US)	1235	79.116	364.774	0.344	2429.223
INTEREST (percent per Annum)	1235	10.58	6.78	1.91	33.50
EXCHANGE (per US Dollar)	1235	795.16	2161.72	0.47	9825.00
INFLATION (percent per Annum)	1235	195.86	77.83	117.39	448.45
SAVE (percent)	1235	21.87	8.36	1.15	42.62
TRADE (percent)	1235	85.86	37.98	26.97	170.73
EXPORTS/IMPORTS (percent)	1235	0.95	0.42	0.35	2.80
FDI_INFLOW (percent)	1234	6.12	6.67	0.03	40.99
CAP (billion $ US)	1235	47.30	172.00	0.16	1140.00
LABOUR (million)	1235	34.41	118.00	0.13	777.00
TOURISM (billion $ US)	1235	13.41	0.94	0.01	409.87
Socio-political Indicator					
LIFE (years)	1235	67.07	8.77	45.40	78.92
SCHOOL (percent)	935	73.38	24.17	21.93	105.62
LACK_FREEDOM	565	22.68	10.81	2.00	38.00
TRANSPARENCY	809	3.24	1.08	2.00	6.60
CRIME (per 100,000 population)	694	11.42	15.34	0.42	60.92
HDI	459	0.63	0.13	0.28	0.80

Table 3.3 Descriptive Statistics of 22 developed countries

	Obs	Mean	Std. Dev.	Max	Min
GDP (billion $ US)	286	983.086	2,424.344	14,369.080	3.635
Macroeconomic Indicator					
MONEY (billion $ US)	277	272.467	753.678	4,756.582	0.004
INTEREST (percent per annum)	286	6.906	4.195	22.599	0.000
EXCHANGE (per US Dollar)	286	71.118	233.397	1,401.440	0.187
INFLATION (percent per annum)	286	105.814	25.781	215.247	56.687
SAVE (percent)	286	23.810	8.049	55.699	5.198
TRADE (percent)	286	108.268	86.893	438.092	18.969
EXPORTS/IMPORTS (percent)	286	1.039	0.156	1.648	0.644
FDI_INFLOW (percent)	286	5.293	6.329	36.615	-10.140
CAP (billion $ US)	286	193.926	460.011	2,295.612	0.705
LABOUR (million)	286	15.325	32.863	158.000	0.144
TOURISM (billion $ US)	286	14.143	22.028	117.969	0.180
Social and Political Indicator					
LIFE (years)	286	78.534	2.107	82.588	72.566
SCHOOL (percent)	272	104.907	17.088	161.781	76.732
LACK_FREEDOM	126	37.643	4.874	40.000	17.000
TRANSPARENCY	264	7.780	1.744	10.000	3.500
CRIME (per 100,000 population)	168	1.705	1.738	19.000	0.000
HDI	109	0.856	0.046	0.937	0.764

Compare the descriptive statistics between developing countries and developed countries. Table 3.2 and 3.3 show the average GDP of developing countries is 276.67 billion $ US while the average GDP of developed countries is 983.086 billion $ US which higher than average GDP of developing countries about 3.5 times. The average money supply , save , trade , export-import ratio , capital , life expectancy , school enrolment ,lack of freedom, transparency and HDI of developed countries also higher than developing countries.

3.4 Data Analysis Methods

Several methods are used to analyze the data and test the hypotheses. The data analysis methods that are used in this study can be classified to two parts. First, we analyze the relationship among gross domestic product, macroeconomic, social and political variables by using panel cointegration technique. Second, to deal with the sample selection bias problem in indentify level of country's development, this study employed panel sample selection model to take into account the selective nature of the samples.

Part I

The thesis shall conduct tests of the causal relationship among gross domestic product and the other macroeconomic aggregates and social indicators in developing country in four stages:

1) tests for the order of integration in the gross domestic product, money supply, interest rate, exchange rate, inflation rate, saving rate, trade, exports/imports, FDI inflow, capital formation, labor, tourism expenditure, life expectancy, school enrollment, lack of freedom, transparency, Crime rate, HDI series.

2) panel co-integration to examine the long-run relationships among the variables

3) estimate equation using fixed effect and random effect approach and using Hausman test and poolability test to check whether model is Fixed Effect Model (FEM) or Random Effect Model (REM). Moreover, when our model is FEM model then we can apply Least Square Dummy Variable approach to our model. Moreover, this thesis also applies Generalized Method of Moment (GMM) method to corrective for serial correlation and non-exogeneity of the regressors.

4) Panel error correction model to estimate the short-run relation among the variables.

Part II

To investigate the relationship among gross domestic product and the other macroeconomic aggregates and social indicators in developed country, we shall use stages 1) and 2) same as part I but in the stage 3) the dissertation employs panel sample selection model with copula approach. First, this thesis will present the estimated result without controls for selection bias and then this thesis provide economic output model was estimated by sample selection model with copula approach.

The panel unit root test, panel cointegration, the panel estimation methodology, panel error correction model and panel sample selection model with copula approach can be described as follow:

3.4.1. Panel Unit Root Test

Before testing for the presence of co-integrating relationship among macroeconomic and social indicator in developing and developed countries, time series properties of the panel data need to be examined. This thesis employed six panel unit root tests: Levin et al. (2002), or LLC, Breitung (2001), Im et al. (2003), or IPS, Fisher-type tests using ADF (Maddala and Wu, 1999), and Fisher-type tests using PP tests (Choi, 2001), and Hadri (2000) to check for the presence of stationarity around a deterministic trend or mean with a shift against a unit root. The properties of panel-based unit root tests under the assumption that the data is independent and identically distributed (i.i.d.) across individuals.

In general, the type of panel unit root tests is based upon the following regression which include lagged dependent variable to remove autocorrelation;

$$\Delta y_{it} = \rho_i y_{i,t-1} + \sum_{L=1}^{p_i} \phi_{iL} \Delta y_{i,t-L} + z_{it}'\gamma + u_{it} \qquad (3.6)$$

where i=1,2,…,N is the country, t =1,2,…,T is time series observation are available, z_{it} is the deterministic components and u_{it} is iid $(0, \sigma_i^2)$. z_{it} could be zero, one, the fixed effects (μ_i) or fixed effect as well as a time trend (t).

For most of the six tests considered, except Hadri, the null hypothesis is that all series have a unit root, that is $\rho_i = 0 \; \forall i$. Each specific test has a different alternative hypothesis, depending upon different degrees of heterogeneity under the alternative hypothesis. The details of each test can be explained as follow (Baltagi,2008):

1) Levin, Lin and Chu (LLC) Test

In the Levin, Lin and Chu (LLC) (2002) tests, one assumes homogeneous autoregressive coefficients between individuals, i.e. $\rho_i = \rho \; \forall i$ and tests the null hypothesis $H_o : \rho_i = \rho = 0$ against the alternative $H_a : \rho_i = \rho < 0$.

The structure of the *LLC* analysis may be specified as follows:

$$\Delta y_{it} = \rho_i y_{i,t-1} + \sum_{L=1}^{p_i} \phi_{iL} \Delta y_{i,t-L} + z'_{it}\gamma + u_{it} \qquad (3.7)$$

where $i = 1,2,...,N$, $t = 1,2,...,T$

Since the lag order p_i is unknown and can vary across individuals, LLC suggest a three-step procedure to implement their test.

Step 1: Perform separate augmented Dickey-Fuller (ADF) regression for each cross-section

$$\Delta y_{it} = \rho_i y_{i,t-1} + \sum_{L=1}^{p_i} \phi_{iL} \Delta y_{i,t-L} + z'_{it}\gamma + u_{it} \qquad (3.7)$$

For a given T, choose a maximum lag order $p_{\max}$ and then use the t-statistic of $\hat{\phi}_{iL}$ to determine if a smaller lag order is preferred.

Step 2 two auxiliary regressions are run to get orthogonalized residuals:

1) Run Δy_{it} on $\Delta y_{i,t-L} (L = 1,..., p_i)$ and z_{it} to get residual $\hat{e}_{it}$

2) Run y_{it-1} on $\Delta y_{i,t-L} (L = 1,..., p_i)$ and z_{it} to get residual $\hat{v}_{it-1}$

Then standardize the residuals to control for different variance across individual:

$$\tilde{e}_{it} = \hat{e}_{it} / \hat{\sigma}_{ui} \quad (3.8)$$

$$\tilde{v}_{it-1} = \hat{v}_{it-1} / \hat{\sigma}_{ui} \quad (3.9)$$

where $\hat{\sigma}_{ui}$ is standard error for each ADF regression, for i=1,...,N.

Step 3: Run the pooled regression

$$\tilde{e}_{it} = \rho \tilde{v}_{i,t-1} + \tilde{u}_{it} \quad (3.10)$$

base on $N\tilde{T}$ observations where $\tilde{T} = T - \bar{p} - 1$ which is average number of observation per individual in the panel and $\bar{p} = \sum_{i=1}^{N} p_i / N$ which is average lag order of individual ADF regressions.

The conventional t-statistic for $H_0 : \rho = 0$ is $t_\rho = \frac{\hat{\rho}}{\hat{\sigma}(\hat{\rho})}$ where

$$\hat{\rho} = \sum_{i=1}^{N} \sum_{t=2+p_i}^{T} \tilde{v}_{it-1} \tilde{e}_{it} \Bigg/ \sum_{i=1}^{N} \sum_{t=2+p_i}^{T} \tilde{v}_{it-1}^2$$

$$\hat{\sigma}(\hat{\rho}) = \hat{\sigma}_{\tilde{u}} \Bigg/ \left[\sum_{i=1}^{N} \sum_{t=2+p_i}^{T} \tilde{v}_{it-1}^2 \right]^{\frac{1}{2}}$$

and $$\hat{\sigma}_{\tilde{u}}^2 = \frac{1}{N\tilde{T}} \sum_{i=1}^{N} \sum_{t=2+p_i}^{T} \left(\tilde{e}_{it} - \hat{\rho} \tilde{v}_{it-1} \right)^2$$

is the estimated variance of $\tilde{u}_{it}$.

The necessary condition for the Levin-Lin-Chu test is $\sqrt{N_T} / T \to 0$, while sufficient conditions would be $N_T / T \to 0$ and $N_T / T \to k$.(N_T means that the cross-sectional dimension N is a monotonic function of time dimension T.)

However, the LL test has some limitations about the test depends crucially upon the independence assumption across individuals and hence not applicable if cross sectional correlation is presents and the assumption that all cross-sections have or do not have a unit root is restrictive.

2) Im, Pesaran and Shin (2003) tests

Im, Pesaran and Shin (2003) extended the Levin and Lin framework to allow for heterogeneity in the value of the autoregressive coefficient under the alternative hypothesis. Im, Pesaran, and Shin,hereafter IPS, begin by specifying a separate ADF regression for each cross section:

$$\Delta y_{it} = \rho_i y_{i,t-1} + \sum_{L=1}^{p_i} \phi_{iL} \Delta y_{i,t-L} + z'_{it}\gamma + u_{it} \tag{3.11}$$

The null hypothesis is that each series in the panel contains a unit root which can be written as $H_0 = \rho_i = 0$ for all i, while the alternative hypothesis, allows for some (but not all) of the individual series to have unit root, is given by:

$$H_a : \begin{cases} \rho_i < 0, \ for\ i = 1,2,...,N_1 \\ \rho_i = 0, \ for\ i = N_1,...,N \end{cases} \tag{3.12}$$

Formally, it requires non-zero fraction of the individual processes is stationary. This condition is necessary for the consistency of the panel unit root test.

After estimate the separate ADF regressions, IPS compute define their t-bar statistic as a simple average of the individual ADF statistics, t_{iT} , for the null as:

$$\bar{t} = \frac{1}{N} \sum_{i=1}^{N} t_{iT} \tag{3.13}$$

IPS assumes that t_{iT} are *i.i.d.* and have finite mean and variance.

In the case where the lag order is always zero ($p_i = 0$ for all i), IPS provides a simulated critical value for $\bar{t}$ for different numbers of cross sections N,series length T and Dickey-Fuller regressions containing intercepts only or intercepts and linear trends.

In the general case where the lag order p_i may be nonzero for some cross-sections, IPS shows that a property standardized $\bar{t}$ converges to a standard normal variate as $N \to \infty$ under the null hypothesis.

The IPS test statistic is as follow:

$$t_{IPS} = \frac{\sqrt{N}\left(\bar{t} - \frac{1}{N}\sum_{i=1}^{N} E\left[t_{it}|\rho_i = 0\right]\right)}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} Var\left[t_{it}|\rho_i = 0\right]}} \longrightarrow N(0,1) \qquad (3.14)$$

The value of $E\left[t_{it}|\rho_i = 0\right]$ and $Var\left[t_{it}|\rho_i = 0\right]$ have been computed by IPS via simulations for different values of T and $p_i's$.

In practice, however, to use their tables, it is necessary to restrict all the ADF regressions to individual series having the same lag length.

3) Breitung (2001) test

The LLC and IPS test require $N \to \infty$ such that $N/T \to 0$,i.e. N should be small enough relative to T. This means that both tests may not keep nominal size well when either N is small or N is large relative to T. Breitung (2001) studied the local power of LLC and IPS tests statistics versus a sequence of local alternatives. He found that both tests suffer from a dramatic loss of power if individual specific trends are included. This is due to the bias correction that also removes the mean under the sequence of local alternatives.

Breitung (2001) suggested a test statistic that does not employ a bias adjustment whose power is substantially higher than that of LLC or the IPS test using Monte Carlo experiments.

Breitung (2001) followed 3 steps which are;

Step 1 : two auxiliary regressions are run to get orthogonalized residuals:

Run Δy_{it} on $\Delta y_{i,t-L}(L = 1,..., p_i)$ to get residual $\hat{e}_{it}$

Run y_{it-1} on $\Delta y_{i,t-L}(L = 1,..., p_i)$ to get residual $\hat{v}_{it-1}$

The residual are then adjusted to correct for individual-specific variances.

Step 2 : the residual $\tilde{e}_{it}$ are transformed using the forward orthogonalization transformation employed by Arellano and Bond (1991)

$$e_{it}^{*} = \sqrt{\frac{T-t}{T-t+1}}\left(\tilde{e}_{it} - \frac{\tilde{e}_{it+1} + ... + \tilde{e}_{iT}}{T-t}\right) \tag{3.15}$$

Also

v_{it-1}^{*} = $\tilde{v}_{it-1} - \tilde{v}_{i,1} - \frac{t-1}{T}\tilde{v}_{it}$ with intercept and trend

= $\tilde{v}_{it-1} - \tilde{v}_{i,1}$ with intercept and no trend

= $\tilde{v}_{it-1}$ with no intercept and no trend

Step 3 : compute the panel test statistics. Run the pooled regression

$$e_{it}^{*} = \rho v_{it}^{*} + u_{it}^{*} \tag{3.16}$$

and obtain the t-statistic for $H_0 : \rho = 0$ which has in the limit of standard N(0,1) distribution.

4) The Fisher's type test: Maddala and Wu (1999) and Choi (2001) test

A common feature of the LL and IPS tests is that they are designed for balanced panels. While sometimes, like in our case, all individual series have the same length, researchers often have to deal with unbalanced panels. Unlike the LL and IPS tests, the procedure advocated by Maddala and Wu (1999), hereafter MW, and Choi (2001) does not require a balanced panel and it is nonparametric.

Moreover, the null and alternative hypotheses are the same as those of the IPS tests which are

$H_{0,MW}$: $\rho_i = \rho = 0$ for $\forall i$

$H_{A,MW}$: $\rho_i < 0$ for $i = 1,2,...,N_1$ and $\rho_i = 0$ for $i = N_1,...,N$

making the IPS and MW tests directly comparable.

Like the IPS tests, the MW test is based on N independent tests on the N individuals. However, while the LL test combines the test statistics, the MW test,

following Fisher (1932), combined the observed significance levels. It is very simple to use, once the p-values are available. If p_i denotes the p-value from the DF test on the i^{th} time series then, in the case of cross-sectional independence, we have the asymptotic result that

$$P = -2\sum_{i=1}^{N} \ln p_i \tag{3.17}$$

has a χ^2 distribution with $2N$ degrees of freedom as $T_i \to \infty$ for all N.

The Fisher test holds some important advantages: 1) it does not require a balanced panel as in the case of IPS test; 2) it can be carried out for any unit root test derived; 3) it is possible to use different lag lengths in the individual ADF regression.

In addition, when *N* is large, it is necessary to modify the *P* test since in the limit it has a degenerate distribution. Having for the *P* test $E[-2\ln p_i] = 2$ and $Var[-2\ln p_i] = 4$, Choi (2001) demonstrated Z-test that

$$Z = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (-2\ln p_i - 2) \to N(0,1) \tag{3.18}$$

This statistic corresponds to the standardized cross-sectional average of individual *p* values. Under the cross-sectional independence assumption of the p_i's, the Lindeberg-Levy central limit theorem is sufficient to show that under the unit root hypothesis *Z* converges to a standard normal distribution as $(T, N \to \infty)_{seq}$.

5) Hadri (2000) test

Hadri (2000) proposed a test similar to the KPSS unit root test that has a null hypothesis of no unit root in any of the series against the alternative of a unit root in the panel.

More specifically, Hadri adopts the following representation of his model components:

$$y_{it} = z_{it}'\gamma + r_{it} + \varepsilon_{it} \tag{3.19}$$

where z_{it} is the deterministic component, r_{it} is a random walk: $r_{it} = r_{it-1} + u_{it}$. $u_{it} \sim iid\left(0, \sigma_u^2\right)$ and $\varepsilon_{it} \sim iid\left(0, \sigma_\varepsilon^2\right)$ are mutually independent normals that are IID across i and over t.

Using back substitution, model (3.19) becomes

$$y_{it} = z_{it}'\gamma + \sum_{s=1}^{t} u_{is} + \varepsilon_{it} = z_{it}'\gamma + e_{it} \tag{3.20}$$

where $e_{it} = \sum_{s=1}^{t} u_{is} + \varepsilon_{it}$

The stationarity hypothesis is simply $H_0 : \sigma_u^2 = 0$ in which case $e_{it} = \varepsilon_{it}$. The LM statistic can be defined as:

$$LM_1 = \frac{1}{\hat{\sigma}_e^2} \frac{1}{NT^2} \left[\sum_{i=1}^{N} \sum_{t=1}^{T} S_{it}^2 \right] \tag{3.21}$$

which is consistent and has an asymptotically normal distribution as $(T, N \to \infty)_{seq}$ and $S_{it} = \sum_{s=1}^{t} \hat{e}_{is}$ are the partial sum process of the OLS residuals from (3.20) and $\hat{\sigma}_e^2$ be a consistent estimator of σ_e^2 under the null hypothesis H_0.

Hadri (2000) suggested an alternative form of the LM statistic allows for heteroskedasticity across i say $\hat{\sigma}_{ei}^2$, which is:

$$LM_2 = \frac{1}{\hat{\sigma}_{e,i}^2} \frac{1}{NT^2} \left[\sum_{i=1}^{N} \sum_{t=1}^{T} S_{it}^2 \right] \tag{3.22}$$

The Hadri panel unit root tests require only the specification of the form of the OLS regressions: whether to include only individual specific constant terms, or whether to include both constant and trend terms.

3.4.2. Panel Cointegration Test

If the variables appear to be non-stationary, we must proceed to test for cointegration. In this study, we shall employ Pedroni (2004) and Kao et.al. (1999) to test whether there are relationships among gross domestic product, macroeconomic and social variables. Both tests are Residual-based panel cointegration test statistics. However, Kao et.al. (1999) considered the spurious regression for the panel data and introduced the DF and ADF type tests, while Pedroni (2004) suggested a Phillips–Perron-type test for cointegration.

1) Pedroni's (2004) test

Pedroni (2004) proposed a residual-based test for the null of cointegration for dynamic panels with multiple regressors in which the short-run dynamics and the long-run slope coefficients are permitted to be heterogeneous across individuals. The test allows for individual heterogeneous fixed effects and trend terms and no exogeneity requirements are imposed on the regressors of the cointegrating regressions.

The residuals estimation from static cointegrating long-run relation for a time series panel of observables y_{it}

$$y_{it} = \alpha_i + \delta_i t + \beta_{1i} x_{1it} + \beta_{2i} x_{2it} + ... + \beta_{Ki} x_{Kit} + e_{it} \qquad (3.23)$$

where i =1,…,N, t = 1,…,T and k = 1,…, K

The variables y_{it} and x_{it} are assumed to be I(1), for each member i of the panel, and under the null of no cointegration the residual e_{it} will also be I(1). α_i and δ_i are scalar denoting fixed effects and unit-specific linear trend parameters, respectively and β_i are the cointegration slopes which permitted to vary across individuals, so that considerable heterogeneity is allowed by this specification.

Pedroni (2004) provided seven statistics for the test of the null hypothesis of no co-integration in heterogeneous panels. Pedroni (2004) tested can be classified into two categories. One group of such tests are termed "*within dimension*"

(panel tests) and the other "*between dimension*" (group tests). The "within dimension" tests pool the data across the ''within dimension,'' thereby taking into account common time factors and allowing for heterogeneity across members. The ''between dimension'' tests allow for heterogeneity of parameters across members, and are called ''group mean cointegration statistics.''

In fact, even if both sets of test verify the null hypothesis of no cointegration:

$$H_0 : \rho_i = 1 \quad \forall i,$$

where ρ_i is the autoregressive coefficient of estimated residuals under the alternative hypothesis ($\hat{e}_{it} = \rho_i \hat{e}_{it-1} + u_{it}$), alternative hypothesis specification is different:

- the panel cointegration statistics impose a common coefficient under the alternative hypothesis which results:

$$H_a^w : \rho_i = \rho < 1 \quad \forall i,$$

- the panel group mean cointegration statistics allow for heterogeneous coefficients under the alternative hypothesis and it results:

$$H_a^b : \rho_i < 1 \quad \forall i,$$

Seven of Pedroni's tests are based upon the estimated residuals $\hat{e}_{it}$ from the long-run model and test statistics that we employ are as follows:

Within dimension (panel tests):

(a) Panel v-statistic

$$Z_v = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \hat{L}_{11i}^{-2} \hat{e}_{i,t-1}^2 \right)^{-1} \tag{3.24}$$

(b) Panel ρ -statistics.

$$Z_p = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \hat{L}_{11i}^{-2} \hat{e}_{i,t-1}^2 \right)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{L}_{11i}^{-2} \left(\hat{e}_{i,t-1} \Delta \hat{e}_t - \hat{\lambda}_i \right) \tag{3.25}$$

(c) Panel PP-statistic.

$$Z_{pp} = \left(\hat{\sigma}^2 \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{L}_{11i}^{-2} \hat{e}_{i,t-1}^2 \right)^{-1/2} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{L}_{11i}^{-2} \left(\hat{e}_{i,t-1} \Delta \hat{e}_t - \hat{\lambda}_i \right) \tag{3.26}$$

(d) Panel ADF-statistic.

$$Z_t = \left(\hat{S}^{*2} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{L}_{11i}^{-2} \hat{e}^{*2}_{i,t-1} \right)^{-1/2} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\hat{L}_{11i}^{-2} \hat{e}^{*}_{i,t-1} \Delta \hat{e}^{*}_{i,t} \right) \tag{3.27}$$

Between dimension (group tests):

(e) Group ρ -statistics.

$$\tilde{Z}_\rho = \sum_{i=1}^{N} \left[\sum_{t=1}^{T} \hat{e}_{i,t-1}^2 \right]^{-1} \sum_{t=1}^{T} \left(\hat{e}_{i,t-1} \Delta \hat{e}_{i,t} - \hat{\lambda}_i \right) \tag{3.28}$$

(f) Group PP-statistic.

$$\tilde{Z}_{pp} = \sum_{i=1}^{N} \left[\hat{\sigma}^2 \sum_{t=1}^{T} \hat{e}_{i,t-1}^2 \right]^{-1/2} \sum_{t=1}^{T} \left(\hat{e}_{i,t-1} \Delta \hat{e}_{i,t} - \hat{\lambda}_i \right) \tag{3.29}$$

(g) Group ADF -statistic.

$$\tilde{Z}_t = \sum_{i=1}^{N} \left[\sum_{t=1}^{T} \hat{S}^{*2} \hat{e}^{*2}_{i,t-1} \right]^{-1} \sum_{t=1}^{T} \left(\hat{e}^{*}_{i,t-1} \Delta \hat{e}^{*}_{i,t} \right) \tag{3.30}$$

where $\hat{\sigma}^2$ is the pooled long-run variance for non parametric model given as $1/N \sum_{i=1}^{N} \hat{L}_{11} \hat{\sigma}_i^2$ and $\hat{\lambda}_i = 1/2 \left(\hat{\sigma}_i^2 - \hat{S}_i^2 \right)$, where $\hat{L}_i$ is used to adjust for autocorrelation in panel parameter model, $\hat{\sigma}_i^2$ and $\hat{S}_i^2$ are the long-run and contemporaneous variances for individual i,and $\hat{S}_i^2$ is obtained from individual ADF-

test of $e_{i,t} = \rho_i e_{i,t-1} + v_{i,t}$. $\hat{S}_i^{*2}$ is the contemporaneous variances from the parametric model, $\hat{e}_{i,t}$ is the estimated residual from the parametric cointegration,while $\hat{e}^*_{i,t}$ is the estimated residual from parametric model. $\hat{L}_{11i}$ is the estimated long-run covariance matrix for $\Delta\hat{e}_{i,t}$ and L_i is the *i*th component of low triangular Cholesky decomposition of matrix Ω_i for $\Delta\hat{e}_{i,t}$ with the appropriate lag length determined by the New-West method.

It is straightforward to observe that the first category of four statistics includes a type of non-parametric variance ratio statistic, a panel version of a non-parametric Phillips and Perron (1986) v-statistic, a non-parametric form of the average of the Phillips and Perron *t*-statistic and an *ADF* type *t*-statistic.

The second category of panel cointegration statistics is based on a group mean approach and includes a Phillips and Perron type v-statistic, a Phillips and Perron type *t*-statistic and an *ADF* type *t*-statistic. The comparative advantage of each of these statistics will depend on the underlying data-generating process.

The statistics can be compared to appropriate critical values; if critical values are exceeded then the null hypothesis of no cointegration is rejected, implying that a long-run relationship between the variables does exist.

Pedroni (2004) simulation showed that, when T > 100, seven statistics have the same power. For little samples (T < 20), the most powerful test is the ADF test based on the between dimension (group t-statistic).

2) Kao Tests (1999)

Kao et.al (1999) presented two types of cointegration tests in panel data, the DF and ADF types tests. Kao tested the residuals $\hat{\varepsilon}_{i,t}$ of the OLS panel estimation by applying DF type tests:

$$\hat{\varepsilon}_{i,t} = \rho_i \hat{\varepsilon}_{i,t-1} + \mu_{i,t} \qquad (3.31)$$

The null hypothesis of no cointegration, H_0: ρ=1, is tested against the alternative hypothesis of stationary residuals, H_1: $\rho \neq 1$. The OLS estimate of ρ and t-statistic are given as

$$\hat{\rho} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{e}_{it}\hat{e}_{it-1}}{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{e}_{it-1}^2} \tag{3.32}$$

and

$$t_{\hat{\rho}} = \frac{(\hat{\rho}-1)\sqrt{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{e}_{it-1}^2}}{s_e} \tag{3.33}$$

where

$$s_e^2 = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(\hat{e}_{it}-\hat{\rho}\hat{e}_{it-1})^2}{TN} \tag{3.34}$$

Kao et.al (1999) proposed the following four DF type tests

$$DF_\rho = \frac{\sqrt{N}T(\hat{\rho}-1)+3\sqrt{N}}{\sqrt{10.2}}, \tag{3.35}$$

$$DF_t = \sqrt{1.25}t_{\hat{\rho}} + \sqrt{1.875N} \tag{3.36}$$

$$DF_\rho^* = \frac{\sqrt{N}T(\hat{\rho}-1)+\frac{3\sqrt{N}\hat{\sigma}_v^2}{\hat{\sigma}_{0v}^2}}{\sqrt{3+\frac{36\hat{\sigma}_v^4}{5\hat{\sigma}_{0v}^4}}} \tag{3.37}$$

$$DF_t^* = \frac{t_{\hat{\rho}}+\frac{\sqrt{6N}\hat{\sigma}_v}{2\hat{\sigma}_{0v}}}{\sqrt{\frac{\hat{\sigma}_v^2}{2\hat{\sigma}_{0v}}+\frac{3\hat{\sigma}_v^2}{10\hat{\sigma}_{0v}}}} \tag{3.38}$$

where $\hat{\sigma}_v^2 = \hat{\Sigma}_u - \hat{\Sigma}_{u\varepsilon}\hat{\Sigma}_{\varepsilon}^{-1}$ and $\hat{\sigma}_{0v}^2 = \hat{\Omega}_u - \hat{\Omega}_{u\varepsilon}\hat{\Omega}_{\varepsilon}^{-1}$. While DF_ρ and DF_t are besed on the strong exogeneity of the regressors and errors, DF_ρ^* and DF_t^* are for the cointegration with endogeneous relationship between regressors and errors. For the ADF tests, the following regression is considered

$$\hat{\varepsilon}_{i,t} = \rho_i \hat{\varepsilon}_{i,t-1} + \sum_{j=1}^{p} \psi_j \Delta\hat{e}_{i,t-j} + \mu_{i,t} \tag{3.39}$$

With the null hypothesis of no cointegration, the ADF test statistic can be constructed as

$$ADF = \frac{t^{ADF} + \dfrac{\sqrt{6N}\hat{\sigma}_u}{2\sigma_{0v}}}{\sqrt{\dfrac{\hat{\sigma}_{0v}^2}{2\hat{\sigma}_v^2} + \dfrac{2\hat{\sigma}_v^2}{10\hat{\sigma}_{0v}^2}}} \tag{3.40}$$

where t^{ADF} is the t-statistic of ρ in (3.39). The asymptotic distributions of DF_ρ, DF_t, DF_ρ^*, DF_t^* and ADF converge to a standard normal distribution N(0,1).

3.4.3. Estimation Methodology

If we find that all variables are co-integrated, we can then employ ordinary least square (OLS) and the Generalized method of moments (GMM) to estimate equation (3.3) and (3.4).

In this thesis, fixed effects as well as random effects models are considered. The fixed effects model is simpler to conduct and is defined according to the following regression model:

$$y_{it} = \alpha_i + X_{it}'\beta + u_{it} \qquad i = 1,...N \;\; t = 1,...T_i \tag{3.41}$$

y_{it} indicates the dependent variable while X_{it} determines the vector of k explanatory variables. The data is incomplete in the sense that there are N countries observed over varying time period length T_i for $i = 1,...N$. $\alpha_i, i = 1,...N$, are constant

coefficients specific to each country. Their presence assumes that differences across the considered countries appear by means of differences in the constant term. These individual coefficients are estimated together with the vector of coefficients β.

In order to validate the fixed effects specification, the question is to prove, according to the empirical application, that the individual coefficients, $\alpha_i, i = 1,...N$,, are not all equal.

This corresponds to the following joint null hypothesis:

$$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_N = \alpha \qquad (3.42)$$

It is rather the acceptation of the alternative hypothesis which is interesting if we want to differentiate between the situations in each country considered in the sample and confirm the existence of significant heterogeneity across countries. The appropriate statistic of the test is a Fisher distributed one with $\left[N-1, \sum_{i=1}^{N} T_i - N - k \right]$ degrees of freedom under the null hypothesis and is defined as follows:

$$F = \frac{SSR_0 - SSR_1}{SSR_1} \frac{\sum_{i=1}^{N} T_i - N - k}{N-1} \qquad (3.43)$$

where SSR_0 and SSR_1 are, respectively, the sum of squared residuals provided by the estimation of the constrained model (under the null hypothesis that is no individual specific coefficients are considered) and the sum of squared residuals relative to the fixed effects model (equation (3.41)).

In the random effects case, the model is defined as follows:

$$y_{iu} = X'_{it}\beta + \varepsilon_{it} \qquad i = 1,...N \quad t = 1,...T_i \qquad (3.44)$$

where $\varepsilon_{it} = \mu_i + v_{it}$ reflect the error component disturbances which $\mu_i \sim$ IIN$(0, \sigma_\mu^2)$ and independent of $v_{it} \sim$ IIN$(0, \sigma_v^2)$.The estimation of the model is

conducted by the feasible generalized least squares method. First, convergent estimates of the variances σ_μ^2 and σ_v^2 are needed. They are obtained by the following formulae:

$$\hat{\sigma}_v^2 = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T_i}(\hat{v}_{it} - \hat{\bar{v}}_i)^2}{\sum_{i=1}^{N} T_i - N - k} \tag{3.45}$$

$$\hat{\sigma}_\mu^2 = \frac{1}{N-k}\sum_{i=1}^{N}\left[\left[\bar{y}_i - \hat{\beta}_b^{i\prime}\bar{X}_i\right]^2 - \frac{1}{T_i}\hat{\sigma}_v^2\right] \tag{3.46}$$

$\hat{v}_{it}$ are the residuals issued from the estimation of the fixed effects model (equation (3.41)) and $\hat{\bar{v}}_i$ are individual means of these residuals over each time period relative to each country. Next, the first term in equation (3.46) indicates the residuals issued from the estimation of the unit means regression where $\hat{\beta}_b^i$ are called the between estimators.

The second stage consists in the estimation by ordinary least squares of the following transformed regression model:

$$y_{iu} + \left(\sqrt{\hat{\theta}_i} - 1\right)y_i = \beta'\left(X_{it} + \left(\sqrt{\hat{\theta}_i} - 1\right)X_i\right) + \varepsilon_{it} + \left(\sqrt{\hat{\theta}_i} - 1\right)\varepsilon_i \tag{3.47}$$

with:

$$\hat{\theta}_i = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + T_i\hat{\sigma}_u^2} \qquad i = 1,...N \tag{3.48}$$

Finally, a Hausman specification test is conducted in order to compare the two categories of specifications. It may be proven that, under the null hypothesis, the two estimates (equations (3.41) and (3.47)) could not differ systematically since they are both consistent. So, the test can be based on the difference. Under the null hypothesis, the Hausman (1978) statistic is asymptotically distributed as chi-square with k degrees of freedom and is written down as follows:

$$H = \left(\hat{\beta}_{GLS} - \hat{\beta}_F\right)' \left(\hat{V}\left(\hat{\beta}_F\right) - \hat{V}\left(\hat{\beta}_{GLS}\right)\right)^{-1} \left(\hat{\beta}_{GLS} - \hat{\beta}_F\right) \quad (3.49)$$

where $\hat{\beta}_F$ and $\hat{\beta}_{GLS}$ are, respectively, the estimates of the fixed effects and random effects models. $\hat{V}(\cdot)$ are the corresponding variance-covariance matrices of these estimated coefficients.

1) The Least Squares Dummy Variable (LSDV) Estimator

Under this approach of estimation of (3.41), it is assumed that any differences across economic agents can be captured by shifts in the intercept term of a standard OLS regression. This leads to the least square dummy variable (LSDV) estimator of a fixed effects regression model. The LSDV model can be estimated by defining a series of group-specific dummy variables d_{git}=1 (g=i). In terms of (3.44), this gives

$$\begin{aligned} y_{iu} &= \alpha_i + X'_{it}\beta + u_{it} \\ &= \alpha_1 d_{1it} + \alpha_2 d_{2it} + \ldots + \alpha_N d_{Nit} + X'_{it}\beta + u_{it} \end{aligned} \quad (3.50)$$

This model is easily estimated by standard OLS over the full panel to yield the LSDV estimator.

2) Dynamic Ordinary Least Square (DOLS) and Generalized Method of Moment (GMM)

Kao et.al (2000) showed that $\hat{\beta}_{OLS}$ is inconsistent when using this estimator for panel data. As a corrective to OLS for serial correlation and non-exogeneity of the regressors, a panel version of the DOLS estimator can be used, based upon the equation

$$y_{it} = x'_{it}\beta + \sum_{k=-K_i}^{K_i} \gamma_{ik} \Delta x_{it-k} + \varepsilon_{it}, \quad (3.51)$$

where

$$\hat{\beta}_{DOLS} = \left[N^{-1} \sum_{i=1}^{N} \left(\sum_{t=1}^{T} z_{it} z_{it}^{'} \right)^{-1} \left(\sum_{t=1}^{T} z_{it} \tilde{y}_{it} \right) \right]_1 \qquad (3.52)$$

and where

z_{it} = is the 2(K+1) x 1 vector of regressors $z_{it} = \left(x_{it} - \bar{x}_i, \Delta x_{it-k}, \ldots, \Delta x_{it+k} \right)$

$\tilde{y}_{it} = y_{it} - \bar{y}_{it}$, and the subscript 1 outside the brackets indicates the first elements of the vector used to obtain the pooled slope coefficient.

Another method is GMM. Formally, model (3.51) may be transformed into the following difference equation:

$$y_{it} - y_{it-1} = \beta'(X_{it} - X_{it-1}) + \gamma'(z_{it} - z_{it-1}) + (u_{it} - u_{it-1}) \qquad (3.53)$$

i=1,…,n t=2,…,T_i

However, from (3.53) a bias arises: since $y_{it-1} - y_{it-2}$ is correlated with the transform error term ($u_{it} - u_{it-1}$), an OLS on dynamic panel data will be inconsistent. But if there are valid instruments, then GMM can be used to estimate the equation with lags of the dependent variable two periods back as an instrumental variable.

3.4.4. Panel Vector Error Correction model

Once the variables are co-integrated, the causality test will be performed. We shall use a panel-based (VECM) to identify the existence and direction of a long-run equilibrium relationship using the two-step procedure of Engle and Granger (1987). In the first step, we shall estimate the long-run model using Eq. (3.3) or (3.4) to obtain the estimated residual ε (the error correction term; e_{it} hereafter). In the second step, we go on to estimate the panel Granger causality model with dynamic error correction. That model can be estimated using instrumental variables to deal with the correction between the error term and the lagged dependent variables.

For the second step, the equation VECM can be written as follows:

$$\Delta y_{it} = \theta_{1i} + \lambda_1 e_{it-1} + \sum_p \pi_{11ip} \Delta y_{it-p} + \sum_p \pi_{12ip} \Delta x_{it-p} + \varepsilon_{it1} \quad (3.54a)$$

$$\Delta x_{it} = \theta_{2i} + \lambda_2 e_{it-1} + \sum_p \pi_{21ip} \Delta x_{it-p} + \sum_p \pi_{22ip} \Delta y_{it-p} + \varepsilon_{it2} \quad (3.54b)$$

where y_{it} is dependent variable, x_{it} is independence variable, ε_{it1} and ε_{it2} are serially uncorrelated error terms and θ_{1i} and θ_{2i} stand for unobserved fixed effects. This model will include both long run and short run information where π_{12} and π_{22} are the impact multiplier (the short run effect) and λ_1 and λ_2 is the feed back effect (adjustment effect and shows number of disequilibrium being corrected) which are tested using t-statistics.

3.4.5. Panel Sample Selection with Copula Approach

1) Model and Likelihood

The structure of the sample selection model (in its simplest parametric form) is two equations system: the first equation is

The *Selection equation*

$$d_{it} = \begin{cases} 1 & if \quad d_{it}^* = z_{it}'\gamma + \eta_i + u_{it} \geq 0 \\ 0 & if \quad d_{it}^* = z_{it}'\gamma + \eta_i + u_{it} < 0 \end{cases} \quad (3.55)$$

where d_{it}^* is the latent decision variable, z_{it} are the regressors affecting the decision rule, and d_{it}^* determines the observability or not for all the members in the sample of the outcome equation,

The *Outcome equation*

$$y_{it}^* = x_{it}'\beta + \alpha_i + \varepsilon_{it} \quad (3.56)$$

where $y_{it} = y_{it}^* \cdot d_{it}$ and y_{it}^* is the latent variable, which is only observable when the latent decision variable $d_{it}^* \geq 0$ or consequently when $d_{it} = 1$. i

($i = 1,...,N$) denotes the individual and t ($t = 1,...,T$) denotes the time period. β and γ are unknown parameter (column-) vectors, and x_{it}, z_{it} are vectors of strictly exogenous explanatory variables with possible common elements. α_i and η_i are unobservable time-invariant individual-specific effects, which presumably correlated with the regressors. ε_{it} and u_{it} are idiosyncratic errors not necessarily independent of each other.

Without sample selectivity, that is with $d_{it} = 1$ for all i and t, equation (3.56) is the standard panel data linear regression model.

Compute the conditional expectation of y_{it} given x_{it} and the probability that y_{it} is observed:

$$\begin{aligned} E(y_{it} \mid x_{it}, d_{it}^* > 0) &= x_{it}'\beta + E\left(\alpha_i + \varepsilon_{it} \mid x_{it}, d_{it}^* > 0\right) \\ &= x_{it}'\beta + E\left(\alpha_i + \varepsilon_{it} \mid x_{it}, u_{it} > -z_{it}'\gamma - \eta_i\right) \end{aligned} \tag{3.57}$$

This leads to consistent estimates of β under the following condition:

$$\begin{aligned} E(\alpha_i + \varepsilon_{it} \mid x_{it}, u_{it} > -z_{it}'\gamma - \eta_i) &= E(\alpha_i \mid x_{it}, u_{it} > -z_{it}'\gamma - \eta_i) \\ &\quad + E(\varepsilon_{it} \mid x_{it}, u_{it} > -z_{it}'\gamma - \eta_i) = 0 \quad , \quad \forall t \end{aligned} \tag{3.58}$$

or we can say that the estimated β will be unbiased when $E(\alpha_i) = 0$ and ε_{it} is independent of u_{it} (that is, $E(\varepsilon_{it} \mid u_{it}) = 0$),so that the data are missing "randomly," or the selection process is "ignorable."

However, in the case where selection is nonrandom, and/or if a correlation with individual heterogeneity is present, the conditional expectation in (3.58) is unequal to zero then OLS estimates on the selected subsample or equation (3.56) is inconsistent. In another way, assume that ε_{it} and u_{it} are jointly distributed with distribution function $f(\varepsilon_{it}, u_{it}; \theta)$ where θ is a finite set of parameters (for

example, the mean, variance, and correlation of the random variables). Then we can write the expectation of error term (by Bayes rule)

$$E(\varepsilon_{it}|x_{it},u_{it}>-z'_{it}\gamma-\eta_i)=\frac{\int_{-\infty}^{\infty}\int_{-z'_{it}\gamma-\eta_i}^{\infty}\varepsilon_{it}f(\varepsilon_{it},u_{it};\theta)du_{it}d\varepsilon_{it}}{\int_{-\infty}^{\infty}\int_{-z'_{it}\gamma-\eta_i}^{\infty}f(\varepsilon_{it},u_{it};\theta)du_{it}d\varepsilon_{it}}=\lambda(z\gamma;\theta) \tag{3.59}$$

$\lambda(z\gamma;\theta)$ is a (possible) nonlinear function of $z\gamma$ and the parameters θ. That is, in general the conditional expectation of y_{it} given x_{it} and the probability that y_{it} is observed to be equal to the usual regression function $x'_{it}\beta$ plus a nonlinear function of the selection equation regressors z_{it} that has a non-zero mean.

To estimate β and control for sample selection bias, several models have been examined and/or estimated (Hausman and Wise (1979), Kyriazidou(1997), Rochina-barrachina(1999) and Wooldridge(1995)). In general, two main approaches have been followed in the development of panel data sample selection model estimators: two-step estimators follow the idea of Heckman (1979) and maximum likelihood estimators. However, in the applied literature, various more or less suitable methods have been used for the estimation of the panel data version of the sample selection model, and it is not obvious which of the suggested specifications to choose. Jensen et al. (2001) compared the different estimation methods for a panel data sample selection model for various data generating processes, is made, by a Monte Carlo study. The specification chosen for the panel part of the estimations in their paper is parametric panel data random effects model where the two equations are estimated simultaneously by maximum likelihood. Therefore, in this dissertation, we prefer treated unobserved heterogeneity as random effects and estimate the result using the maximum likelihood approach.

Since we estimate (3.55) and (3.56) simultaneously using maximum likelihood, we have to specify the joint distribution of the error components ε_{it} and u_{it}. Assume ε_{it} and u_{it} are jointly distributed with distribution function $f(\varepsilon_{it},u_{it};\theta)$.

Furthermore, we make the following assumptions concerning the random effects and their interactions with the idiosyncratic errors:

$$E[\alpha_i] = E[\eta_i] = 0$$

$$\varepsilon_{it}, u_{it} \perp \alpha_i, \eta_i$$

Thus, the individual-specific components (α_i and η_i) in the selection equation (3.55) and the equation of the interest (3.56) may be correlated, but they are assumed to not be correlated with the idiosyncratic error terms.

The likelihood of a single observation, conditional on the random effects, is then

$$\begin{aligned} L_{it}(\gamma, \beta, \theta | \alpha_i, \eta_i) &= f(\varepsilon_{it}, u_{it}; \theta | \alpha_i, \eta_i) \\ &= \left\{\Pr\left(d_{it}^* \le 0\right)\right\}^{1-d_{it}} x \left\{h_{\varepsilon|d}\left(\varepsilon_{it} \middle| d_{it}^* > 0\right) \times \Pr\left(d_{it}^* > 0\right)\right\}^{d_{it}} \\ &= \left\{\Pr\left(u_{it} \le -z_{it}'\gamma - \eta_i\right)\right\}^{1-d_{it}} \times \\ &\quad \left\{f_{\varepsilon|u}\left(\varepsilon_{it} | u_{it} > -z_{it}'\gamma - \eta_i\right) \times \Pr\left(u_{it} > -z_{it}'\gamma - \eta_i\right)\right\}^{d_{it}} \end{aligned} \tag{3.60}$$

where the first term is the contribution when $d_{it}^* \le 0$, since then $d_{it} = 0$ and the second term is the contribution when $d_{it}^* > 0$.

The presence of the conditional distribution $f_{\varepsilon|u}\left(\varepsilon_{it} | u_{it} > -z_{it}'\gamma - \eta_i\right)$ in the likelihood of presenting complications in estimation and, thus conditional copulas calculated by $C_{u|v}(u,v) = \dfrac{\partial C(u,v)}{\partial v}$ might be useful.

First, observe that the conditional density is $f_{\varepsilon|u}\left(\varepsilon_{it}|u_{it} > -z'_{it}\gamma - \eta_i\right)$. The $f_{\varepsilon|u}$ denotes the probability density function of y^*_{it}, given event $d^*_{it} > 0$. It functional form can be written as follows in terms of marginal density and distribution functions.

$$f_{\varepsilon|u}\left(\varepsilon_{it}|u_{it} > -z'_{it}\gamma - \eta_i\right) = \left(1 - F_u\left(-z'_{it}\gamma - \eta_i\right)\right)^{-1} \frac{\partial}{\partial \varepsilon}\left[F_\varepsilon(\varepsilon_{it}) - F(u_{it}, \varepsilon_{it})\right]$$

$$= \left(1 - F_u\left(-z'_{it}\gamma - \eta_i\right)\right)^{-1}\left[f_\varepsilon(\varepsilon_{it}) - \frac{\partial}{\partial \varepsilon}\left(F(u_{it}, \varepsilon_{it})\right)\right] \tag{3.61}$$

Then substitute (3.61) for (3.60)

$$L_{it}(\gamma, \beta, \theta|\alpha_i, \eta_i) = \left\{F_u\left(-z'_{it}\gamma - \eta_i\right)\right\}^{1-d_{it}} \times$$

$$\left\{\left(1 - F_u\left(-z'_{it}\gamma - \eta_i\right)\right)^{-1}\left[f_\varepsilon(\varepsilon_{it}) - \frac{\partial}{\partial \varepsilon} F(u_{it}, \varepsilon_{it})\right] \times \left(1 - F_u\left(-z'_{it}\gamma - \eta_i\right)\right)\right\}^{d_{it}}$$

$$= \left\{F_u\left(-z'_{it}\gamma - \eta_i\right)\right\}^{1-d_{it}} \times \left[f_\varepsilon(\varepsilon_{it}) - \frac{\partial}{\partial \varepsilon} F(u_{it}, \varepsilon_{it})\right]^{d_{it}} \tag{3.62}$$

where the F and f are the cumulative density function and probability density function respectively for the variables referred by subscripts. F_u will be used from now on to denote $F_u(u_{it}) = \Pr\left(d^*_{it} \le 0\right) = \Pr\left(u_{it} \le -z'_{it}\gamma - \eta_i\right)$ $= F_u(-z'_{it}\gamma - \eta_i)$.Furthermore, from this point F_ε denote $F_\varepsilon(\varepsilon_{it})$ $= \Pr\left(\varepsilon_{it} \le y_{it} - x'_{it}\beta - \alpha_i\right) = F_\varepsilon(y_{it} - x'_{it}\beta - \alpha_i)$, and $f_\varepsilon = f_\varepsilon(\varepsilon_{it})$.

When a distribution is specified for the random effects it is straightforward to integrate them out of the likelihood function. If (α_i, η_i) is distributed according to the distribution function $G(\cdot)$ we have:

$$L_i(\gamma,\beta,\theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[\prod_{t=1}^{T_i} f\left(\varepsilon_{it},u_{it}|\alpha_i,\eta_i\right)\right]dG(\alpha_i,\eta_i)$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[\prod_{t=1}^{T_i} f\left(\varepsilon_{it},u_{it}|\alpha_i,\eta_i\right)\right]g(\alpha_i,\eta_i)d\alpha_i d\eta_i \quad (3.63)$$

T_i is the number of observations for an individual. In the estimations of this dissertation, $\mathrm{G}(\cdot)$ is specified as a bivariate discrete distribution with 2x2 points of support. If the distribution assumptions are satisfied, then this estimator will be consistent for γ and β, but it does not allow for a correlation between the observed and unobserved variables.

The log-likelihood function is

$$LogL = \sum_{i=1}^{N} LogL_i \quad (3.64)$$

However, the random effects formulation can be criticized on the grounds that it neglects the correlation that may exist between the random effects and the explanatory variables. If this correlation is ignored, the estimates of the parameters of interest (here γ and β) are biased. Mundlak (1978) proposed a way to correct for this correlation. Basically, what he does in the linear model, is to approximate $E(\alpha_i|x_i)$ by a linear function and to include the individual means of the explanatory variables in the two equations. In the models of this dissertation, the individual means of the main variables of interest are included. A simple joint F-test of these correction terms then makes it possible to determine whether the correlation is actually present in the random effects model and hence, whether it makes a difference to make the Mundlak correction.

The component of (3.62) that is the most difficult to evaluate is $\frac{\partial}{\partial \varepsilon} F(u_{it}, \varepsilon_{it})$. The next subsection provides the evaluation of $\frac{\partial}{\partial \varepsilon} F(u_{it}, \varepsilon_{it})$ which several specification for the joint distribution $F(u_{it}, \varepsilon_{it})$.

In general the multivariate normal distribution is assumed. However, this assumption can often be seen as excessively restrictive. Smith (2003) suggested applying the copula approach, especially the Archimedean copula to the sample selection model and the result also shows that the copula approach is well suited to apply to a model where the sample selection is biased, using cross-section data. Therefore, this thesis provides the Gaussian copula, Gumbel (1960) copula , Ali-Mikhill-Haq (1978) copula, Clayton (1978) copula, Frank (1979) copula, and Joe (1997) copula to construct the joint distribution $F(u_{it}, \varepsilon_{it})$.

2) Modeling using the copula approach

This subsection presents necessary background on copula and then will present the modeling using the copula approach.

2.1) Background

The notion of copula was introduced by Sklar (1959), when answering a question raised by M. Fréchet about the relationship between a multidimensional probability function and its lower dimensional margins. A copula is a function that links together univariate distribution functions to form a multivariate distribution function. If all of the variables are continuously distributed, then their copula is simply a multivariate distribution function with uniform (0, 1) univariate marginal distributions. The main advantage of copulas consists in representing the joint probability distribution by separating the impact of the marginals from the association structure, explained by the copula functional form. Copulas have been increasingly explored in the literature. Joe (1997) and Nelsen (2006) provided a complete monograph of an introduction to the theory of copulas and a large selection of related

models. Another reviews such as Frees and Valdez (1998) and Cherubini et al. (2004) provided more detail about the application in actuarial and financial settings.

2.2) Copula properties and dependence structure

(1) Copula properties

In this section, we give the general definition of copulas and an equivalent definition for the random variable context. We begin with the definition of copula for the bivariate case (Nelson (2006)).

Definition 1. A *copula* is a function $C: [0,1]^2 \to [0,1]$ which satisfies:

(a) For every u, v in $[0,1], C[u,0] = 0 = C(0,v)$, and $C[u,1] = u$ and $C[1,v] = v$;

(b) For every u_1, u_2, v_1, v_2 in $[0,1]$ such that $u_1 \le u_2$ and $v_1 \le v_2$, $C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \ge 0$.

The importance of copulas in statistics is described in Sklar's Theorem:

Theorem 1 Sklar's Theorem *Let X and Y be random variables with joint distribution function H and marginal distribution functions* $F(x) = u$ *and* $G(y) = v$, *respectively. Then there exists a copula* C *such that*

$$H(x,y) = C(F(x), G(y)) = C(u,v) \quad (3.65)$$

for all x, y in $\overline{\Re}$ *where* $C(u,v)$ *is the copula that captures the dependence structure between X and Y. If F and G are continuous, then* C *is unique. Otherwise, the copula* C *is uniquely determined on Ran(F)×Ran(G). Conversely, if* C *is a copula and F and G are distribution functions, then the function H, defined by (3.65), is a joint distribution function with margins F and G.*

Thus copulas link joint distribution functions to their one-dimensional margins as proven by Nelsen(2006).

The function $C(u,v)$ is known as the copula of $H(x,y)$ and describe how $H(x,y)$ is coupled with the marginal distribution function $F(x)$ and $G(y)$. Copulas themselves can be generated in several different ways, including the method of inversion, geometric methods and algebraic methods (Nelson (2006)). For instant, given a known bivariate distribution $H(x,y)$ with continuous margin $F(x)$ and $G(y)$, the inversion method inverts the relationship in equation (3.66) to obtain a copula:

$$\begin{aligned} C(u,v) &= \Pr(U \le u, V \le v) \\ &= \Pr(X \le F^{-1}(x), Y \le F^{-1}(y)) \\ &= F\left(x = F^{-1}(x), y = F^{-1}(y)\right) \end{aligned} \tag{3.66}$$

As a consequence of Sklar's Theorem, we encounter the Fréchet - Hoeffding bounds for copulas, i.e., for any copula C and for all u,v in [0,1],

$$W(u,v) = \max(u+v-1,0) \le C(u,v) \le \min(u,v) = M(u,v) \tag{3.67}$$

where W is termed the Fréchet lower bound for copulas and M is the Fréchet upper bound for copulas. All other copulas take values in between these bounds on each point of their domain, the unit square. The Fréchet upper bound corresponds to perfect positive dependence and the lower bound corresponds to perfect negative dependence.

(2) Dependence Structure

As copulas are regularly used as tool for modeling and capturing the dependence of two or more random variables, one must specify of how to measure dependence. Traditionally the dependence between two random variables is measured by the linear correlation coefficient. However, the linear correlation is useful only for elliptical distribution and when the dependence is not described by an elliptical distribution it can be quite misleading to use a linear correlation. Therefore, it might be more reasonable to use copula based measures of dependence, which are scale

invariant (see Embrechts (2002) for caveats on using the correlation coefficient for measuring dependence).

One of these more robust copula based measures are *Spearman's rho* and *Kendall's tau*, which provide alternative nonparametric measurements of dependence between variables that, unlike the simple correlation coefficient (Pearson's correlation coefficient), do not require a linear relationship between the variables. *Spearman's rho* and *Kendall's tau* rely on the concept of concordance. Consider two pairs of observations (x_i, y_i) and (x_j, y_j) from the continuous random variables(X,Y). We call these pairs of observations concordant if

$(x_i - x_j)(y_i - y_j) > 0$ and discordant if $(x_i - x_j)(y_i - y_j) < 0$. Hence, two random variables are said to be concordant, when large values of one random variable are associated with large values of the other, and similarly small values tend to be associated with each other.

Using the concept of concordance, we are now able to introduce a measure of association known as *Kendall's tau*. Its sample version is defined as the fraction of concordant pairs of observations in the sample minus the fraction of discordant pairs of observations. The population version of *Kendall's tau* is defined as the difference between the probability of concordance and the probability of discordance.

$$\tau = \tau_{XY} = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0] \tag{3.68}$$

These probabilities can be evaluated by integrating over the distribution of (X_2, Y_2).So that, in terms of copulas, Kendall's τ becomes

$$\tau_C = 4\int_0^1\int_0^1 C(u,v)dC(u,v) - 1 \tag{3.69}$$

where C is the copula associated to (X, Y).

Note that the integral above is the expected value of the random variable $C(U,V)$,

where $U, V \sim U(0,1)$ with joint distribution function C, i.e. $\tau(X,Y) = 4E(C(U,V)) - 1$.

Another measure of association of (X,Y) is the Spearman's rho. Spearman's rho for the random vector (X,Y) is defined as

$$\rho_{XY} = 3(P\{(X_1 - X_2)(Y_1 - Y_3) > 0\} - P\{(X_1 - X_2)(Y_1 - Y_3) < 0\}) \tag{3.70}$$

where $(X_1, Y_1), (X_2, Y_2)$ and (X_3, Y_3) are independent random vectors with a common joint distribution function H.

In terms of the copula C associated to the pair (X, Y) becomes

$$\rho_C = 12\int_0^1\int_0^1 uvdC(u,v) - 3 = 12\int_0^1\int_0^1 C(u,v)dudv - 3 = 12E(C(U,V)) - 3 \tag{3.71}$$

In addition, both assume the value of zero under independence and are not dependent on the margins F(.) and G(.). Hence, these two concordance measures are used to characterize dependence structures in the copula literature, rather than the familiar Pearson's correlation coefficient.

3) Example of copulas

This section will present a few of copulas that were used in this study. For exhaustive lists of copula functions and various methods for constructing copulas books by Joe (1997) and Nelson (2006) may be consulted. Moreover, the copula family studied in this dissertation includes the Gaussian copula, Gumbel (1960) copula , Ali-Mikhill-Haq (1978) copula, Clayton (1978) copula, Frank (1979) copula, and Joe (1997) copula which are shown as follows:

(1) Gaussian copulas

The (bivariate) Gaussian copula proposed by Lee (1983). It can easily be derived from the bivariate normal distribution and has the following distribution function

$$C_{Gaussian}(u,v;\rho) = \Phi_G(\phi^{-1}(u),\phi^{-1}(v);\rho)$$

$$= \int_{-\infty}^{\phi^{-1}(u)}\int_{-\infty}^{\phi^{-1}(v)} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{s^2-2\rho st+t^2}{2(1-\rho^2)}\right) dsdt \quad (3.72)$$

where ρ is the linear correlation coefficient of the corresponding bivariate normal distribution and restricted to the interval (-1, 1) and Φ_G is the standard bivariate normal distribution and ϕ^{-1} denoting the inverse of the univariate Gaussian distribution. Note that it can be shown that the Gaussian copula does not have a tail dependence for ρ < 1. The (bivariate) Gaussian copula satisfies $W(u,v) \le C_{Gaussian}(u,v) \le M(u,v)$, which is said to be comprehensive.

(2) Archimedean Family of Copulas

One of the most important classes of copulas is known as Archimedean copulas. There are a various reasons to apply Archimedean family in practice. These copulas are very easy to construct. Archimedean copulas allow a wide range of possible dependence behavior and all commonly encountered Archimedean copulas have simple closed form expressions. In addition, the Archimedean representation allows us to reduce the study of a multivariate copula to a single univariate function.

An Archimedean copula can be written in the following way:

$$C(u,v) = \varphi^{-1}[\varphi(u)+\varphi(v)] \quad (3.73)$$

for all $u,v \in [0,1]$ and where $\varphi:[0,1]\to[0,\infty]$ is called the generator of the copula which satisfying: $\varphi(0)=0$ and $\varphi(1)=0$; $\forall t\in(0,1), \varphi'(t)<0$, this says that φ is decreasing; $\forall t\in(0,1), \varphi''(t)\ge 0$, this says that φ is convex.

$\varphi^{-1}:[0,\infty]\to[0,1]$ is the inverse of the generator of the copula, which is defined as

$$\varphi^{[-1]}(t)=\begin{cases}\varphi^{-1}(t) & ;0\le t\le\varphi(0)\\ 0 & ;\varphi(0)\le t\le\infty\end{cases}$$

where $\varphi^{-1}(t)=\inf\{t|\varphi(t)\ge t\}$ for $t\in(0,1)$.

The $\varphi(.)$ depends on a single parameter θ that reflect the degree of dependence. Archimedean copula are systematic in the sense of $C(u,v)=C(v,u)$ and associative in the sense of $C\big(C(u,v),w\big)=C\big(u,C(v,w)\big)$.

For φ with a continuous second derivative, we can compute the density $f(u,v)$ of C by taking derivatives:

$$\varphi\big(C(u,v)\big) = \varphi(u)+\varphi(v)$$

$$\varphi'\big(C(u,v)\big)C_u(u,v) = \varphi'(u)$$

$$\varphi''\big(C(u,v)\big)C_v(u,v)C_u(u,v)+\varphi'\big(C(u,v)\big)C_{u,v}(u,v) = 0$$

Therefore,

$$f(u,v)=C_{u,v}(u,v) = -\frac{\varphi''\big(C(u,v)\big)C_v(u,v)C_u(u,v)}{\varphi'\big(C(u,v)\big)} \tag{3.74}$$

where the derivatives do not exists on the boundary $\varphi(u)+\varphi(v)=\varphi(0)$.

The conditional density of the Archimedean copula, which will be used in this study, is

$$\frac{\partial C(u,v)}{\partial v}=\frac{\varphi'(v)}{\varphi'(C(u,v))} \tag{3.75}$$

obtained by differentiating (3.73) with respect to v and rearranging the result.

The measure of dependence is relatively straightforward for Archimedean copulas because Kendall's tau simplifies it to a function of the generator function

$$\tau = 1 + 4\int_0^1 \frac{\varphi(t)}{\varphi'(t)}\,dt \tag{3.76}$$

See Genest and Mackay (1986) for a derivation.

Particular examples are $\varphi(t) = -\ln(t)$ and $\varphi(t) = (-\ln t)^{\theta}$, which are, respectively, generators of product copula $\prod$ and the Gumbel family of copulas $C(u,v;\theta) = \exp\left\{-\left[(-\ln u)^{\theta} + (-\ln v)^{\theta}\right]^{\frac{1}{\theta}}\right\}$ where $\theta \in [1,\infty)$. Example of families of Archimedean copula are listed in Table 3.4

Table 3.4. Examples of families of bivariate Archimedean copulas.

Name	**Copula** $C_{\theta}(u,v)$	**Parameter space**	**Generator** $\varphi(t)$	**Kendall's** τ
Clayton	$(u^{-\theta} + v^{-\theta} - 1)^{\frac{-1}{\theta}}$	$0 \le \theta < \infty$	$\frac{1}{\theta}(t^{-\theta} - 1)$	$0 \le \tau < 1$
Gumbel	$e^{\left\{-\left[(-\ln u)^{\theta} + (-\ln v)^{\theta}\right]^{\frac{1}{\theta}}\right\}}$	$1 \le \theta < \infty$	$(-\ln t)^{\theta}$	$0 \le \tau < 1$
Frank	$-\frac{1}{\theta}\ln\left[1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{(e^{-\theta} - 1)}\right]$	$-\infty < \theta < \infty$	$-\ln\frac{e^{-\theta t} - 1}{e^{-\theta} - 1}$	$-1 < \tau < 1$
Ali-Mikhail-Haq	$\frac{uv}{1 - \theta(1-u)(1-v)}$	$-1 \le \theta < 1$	$\ln[1 - \theta(1-t)]/t$	$-0.1817 \le \tau < \frac{1}{3}$
Joe	$1 - ((1-u)^{\theta} + (1-v)^{\theta} - (1-u)^{\theta}(1-v)^{\theta}$	$1 \le \theta < \infty$	$-\ln(1-(1-t)^{\theta})$	$0 \le \tau < 1$

Source : Nelson (2006)

Frank Copula is "comprehensive" in the sense that it attains all of the Fréchet lower bound, independence, and the Fréchet upper bound. Frank Copula permits negative dependence between the marginals and the dependence is symmetric in both tails, similar to the Gaussian and Student-t copulas. Moreover, for the Clayton copula, the parameter of dependence was restricted on the region $(0,\infty)$ which means

it attains the Fréchet upper bound, but for no value does it attain the Fréchet lower bound. The Clayton copula cannot account for negative dependence, but it exhibits strong left tail dependence and relatively weak right tail dependence. Another Copula family is Gumbel Copula. The dependence parameter of Gumbel Copula is restricted to the interval $[1,\infty)$ which corresponds independence and the Fréchet upper bound, but this copula does not attain the Fréchet lower bound for any value of θ .Similar to the Clayton copula, Gumbel does not allow negative dependence, but in contrasts to Clayton, Gumbel exhibits strong right tail dependence and relatively weak left tail dependence. Moreover, for Ali-Mikhail-Haq , the copula parameter lies on a closed interval between -1 and +1 which dependence parameter shows that it does not contain the Fréchet bounds. The dependence parameters of Joe Copula are the same as Gumbel Copula which is restricted to the interval $[1,\infty)$.

For Archimedean copulas, the following conditional density appearing in the likelihood (3.62) simplifies to

$$\begin{aligned} \frac{\partial}{\partial \varepsilon} F(u_{it}, \varepsilon_{it}) &= \frac{\partial C_\theta(F(u_{it}), v)}{\partial v}\Big|_{v \to F_\varepsilon} \times \frac{\partial F(\varepsilon_{it})}{\partial \varepsilon_{it}} \\ &= \frac{\varphi'(F_\varepsilon)}{\varphi'(C_\theta)} \times f_\varepsilon \end{aligned} \tag{3.77}$$

where C_θ denotes $C_\theta(F_u, F_\varepsilon) = C_\theta(F(u_{it}), F(\varepsilon_{it}))$, which is evaluated as $\varphi^{-1}[\varphi(F_u) + \varphi(F_\varepsilon)]$.

The single observation likelihood of any Archimedean copula families can be written as

$$L_{it}(\gamma, \beta, \theta | \alpha_i, \eta_i) = \{F_u(-z_{it}'\gamma - \eta_i)\}^{1-d_{it}} \times \left[f_\varepsilon(\varepsilon_{it}) - \frac{\varphi'(F_\varepsilon)}{\varphi'(C_\theta)} \times f_\varepsilon(\varepsilon_{it}) \right]^{d_{it}}$$

$$= \quad \{F_u(-z_{it}'\gamma-\eta_i)\}^{1-d_{it}} \times \left[f_\varepsilon(\varepsilon_{it})\left(1-\frac{\varphi'(F_\varepsilon)}{\varphi'(C_\theta)}\right)\right]^{d_{it}} \tag{3.78}$$

where $F_u(u_{it}) = F_u(-z_{it}'\gamma-\eta_i), F_\varepsilon(\varepsilon_{it}) = F_\varepsilon(y_{it}-x_{it}'\beta-\alpha_i).$

As the functional form of $\varphi'(t)$ is generally quite easy to derive, the likelihood is relatively easy to code. For example, under the Clayton family , the likelihood is

$$L_{it}(\gamma,\beta,\Sigma|\alpha_i,\eta_i) \quad = \quad \{F_u(-z_{it}'\gamma-\eta_i)\}^{1-d_{it}} \times \left[f_\varepsilon(\varepsilon_{it})\left(1-\left(\frac{C_\theta}{F_\varepsilon}\right)^{\theta+1}\right)\right]^{d_{it}} \tag{3.79}$$

In Table 3.5, expresses for the component $\left(1-\frac{\varphi'(F_\varepsilon)}{\varphi'(C_\theta)}\right)$ of the likelihood given for the selected families of Archimedean copulas.

Table 3.5 Expressions for $1-\frac{\varphi'(F_\varepsilon)}{\varphi'(C_\theta)}$

Name	Expression
Clayton	$1-\left(\frac{(F_u^{-\theta}+F\varepsilon^{-\theta}-1)^{\frac{-1}{\theta}}}{F_\varepsilon}\right)^{\theta+1}$
Gumbel	$1-\frac{(-\log F_\varepsilon)^{\theta-1}C_\theta(F_u,F_\varepsilon)\left((-\log F_u)^\theta+(-\log F_\varepsilon)^\theta\right)^{-1+\frac{1}{\theta}}}{F_\varepsilon}$
Frank	$\frac{e^{\theta F_\varepsilon}(e^{\theta F_u}-e^\theta)}{e^{\theta(F_\varepsilon+F_u)}+e^\theta(1-e^{\theta F_u}-e^{\theta F_\varepsilon})}$
Ali-Mikhail-Haq	$1-\frac{(1-\theta)F_u+\theta F_u^2}{(1-\theta(1-F_u)(1-F_\varepsilon))^2}$
Joe	$1-\left(1-(1-F_u)^\theta\right)(1-F_\varepsilon)^{\theta+1}\left((1-F_u)^\theta+(1-F_\varepsilon)^\theta-(1-F_u)^\theta(1-F_\varepsilon)^\theta\right)^{-1+1/\theta}$

Source: Smith(2003)

4) Selection of the copula function

The choice of copula can be made using information criteria such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) or the Schwartz information criterion (SIC). Both AIC and BIC penalize the negative maximum log-likelihood of the estimated model by the number of parameters in the model. These criteria are AIC = -2 log (maximum likelihood) + 2 (number of parameters) and BIC = -2 log (maximum likelihood) + (number of parameters)(log of the sample size). A smaller relative AIC or BIC represents a better model fit.