Chapter 2

Research Methodology

In this chapter, Panel data analysis endows regression analysis with both a spatial and temporal dimension. The spatial dimension pertains to a set of cross-sectional units of observation. These could be countries, states, counties, firms, commodities, groups of people, or even individuals. The temporal dimension pertains to periodic observations of a set of variables characterizing these cross-sectional units over a particular time span. The panel unit root test and cointegration techniques used was based on Panel Cointegration while as OLS estimator and DOLS estimator were used to find long-run relationship of the international tourism demand model in Thailand as well as by using fixed and random effects for long run static models , and including short-run relationship estimate dynamic panel data to test tourists different purpose on business and holiday to Thailand while dynamic panel data models adopted the generalized method of moments (GMM) and estimator (panel GMM procedures) and panel GMM of Arellano and Bond.

2.1 Data and Sample Selection

2.1.1 Data

1. According to chapter 3 detect the most significant factors affecting the flow of international tourists demand by country of origin and chapter 4 detect tourists demand with different dependent variables purpose on business and holiday to Thailand both in the long run and short-run relationship estimate by using static and

dynamic panel data, annually data were u sed in 1981-2007 form the top ten number of international tourists to Thailand which include Malaysia ,Japan, Korea ,China ,Singapore ,U.K ,U.S.A, Australia ,Germany and Taiwan .

International tourism demand is usually measured by proxies such as the number of foreign visitors, the volume of earnings generated by foreign visitors, and the number of nights spent by visitors from abroad Consequently, we use the number of foreign visitors, namely international tourist arrivals, to estimate on international tourism demand to Thailand**.** Annual data for 1981 – 2007 period were used for two dependent variables namely the numbers of foreign visitors to Thailand and the numbers of visitors to Thailand from the ten major sending countries in the category of traveling for holiday purpose and in that for business purpose. Annually data for international tourist arrivals are collected from statistical data sets for each country is obtained from World Tourism Organization or Tourism Authority Thailand (TAT).The key independent variables in equations are Real GDP per capita in country of origin or tourism disposable income of individuals coming from origin country)(Y_{it}) This variable is approximated income with origins' per capita GDP at constant prices. Data are taken from GDP per Cap from United States Department of Agriculture, Economic Research Service. International macroeconomic data set. As far as relative prices are concerned, it is common in tourism demand studies to use the CPI of a destination country for relative tourism prices. The inverse of this shows how many "baskets" of goods a tourist has to give up in his home country in order to buy a basket of goods in the destination country (RP_{it} = CPI Thailand / CPI origin country) obtained data from IMF and BOT(Bank Of Thailand). The other independent variables also include nominal exchange rate of original country to which modify the value to Thai Bath per dollar (ER_{it}) obtained from United States Department of Agriculture, Economic Research Service international and macroeconomic data set .Transportation costs from origin country i to Thailand or transport costs to reach Thailand by individuals coming from original country (TC_{it}),since information on bilateral transport costs was unavailable this variable is approximated with Jet Fuel(Dollar)/CPI origin. Data are taken from the United States Energy Information Administration (2007) Rotterdam (ARA) Kerosene-Type Jet Fuel Spot Price FOB.

2. In Chapter 5 ,monthly data from January 1992 to December 2006 were used for the other dependent variable: - the number of visitors from each of the ten major, countries of origin that traveled to four important tourism destinations in Thailand namely Bangkok, Chiang Mai, Cholburi(Pattaya), and Phuket. Monthly data for international tourist arrivals collected from statistical data sets for each country have been obtained from the World Tourism Organization or Tourism Authority of Thailand (TAT). The key independent variables in equations are as follow

Y_{it} = GDP per capita in country of origin. *Disposable tourism income of individuals coming from origin country*. This variable is approximated income with origins' per capita GDP at constant prices. Data are taken from GDP per Cap from United States Department of Agriculture, Economic Research Service, international macroeconomic data set.

RP_{it} = CPI Thailand / CPI origin country. Data from IMF and BOT (Bank of Thailand)

ER_{it} = nominal exchange rate of original country to Thai Baht per dollar . Exchange rate from United States Department of Agriculture, Economic Research Service. International macroeconomic data set.

TC_{it} = transportation costs from origin country i to four important tourism destinations in Thailand namely Bangkok, Chiang Mai, Cholburi(Pattaya), and Phuket.Thailand ***or*** transport costs to reach Thailand by individuals coming from their original country. Since information on bilateral transport costs was unavailable, this variable is approximated with Jet Fuel (Dollar)/CPI origin. Data has been taken from the United States Energy Information Administration (2007) Rotterdam (ARA) Kerosene-Type Jet Fuel Spot and Distance from capital of original country to capital of Thailand Indian Industry Directory of Indian Suppliers air distance calculator. From http:// www. indianindustry.com/travel-tools/air-distance-calulator.html (Sources: United States Energy Information Administration (2007) Rotterdam (ARA) Kerosene-Type Jet Fuel Spot Price FOB. (Note: 1 gallon = 3.785 liters. Total Jet oil per person in Air Bus 380 = 3 liter/100 km/person, TC = (Jet Fuel(Dollar)/CPI origin)/ person * Distance (km) from capital of original country to capital of Thailand)

2.2 Data Analysis Methods

2.2.1 Conceptual framework

Definition provided by the 1963 United Nations conference on international Travel and Tourism, international tourists are generally defined as temporary visitors who spend more than 24 hours in destination other than their normal place of residence and work, and whose journey is for the purpose of leisure, visiting friends and relatives, business, convention or meeting, health, study, religion or sport.

The tourism product is a bundle of goods and services packaged and offered to tourists. This product is of a composite nature and has a number of distinctive characteristics. Firstly, the product is composed of natural resources which are public goods: beaches, waterfalls, mountains, and the general environment. Secondly, safety and infrastructure are also important elements of the product. Studies on the determinants of tourism demand are subjected to specific problems.

The reasons to the problems are the special nature of the demand for tourism, which can be attributed to the complexity of the motivational structure underlying the decision-making process, and the scarcity of relevant data that are fundamental in econometric modeling. However, tourism demand modeling is not easy due to the complexity of the decision-making process, the multiplicity and heterogeneity of the products and services specified, the fact that transportation plays a role in the consumption of tourism, the intertemporal dependence of current demand on its past and future values, and the non reparability between tourism demand and the demand for other goods and services. These difficulties in the specification of a comprehensive and reliable model of demand for tourism are compounded by the existence of unquantifiable factors influencing demand, as well as by the inaccuracy or unavailability of data for those that are, in principle, measurable. Thus, the conceptual and practical problems underlying the empirical studies in this area accounted for the simplifying assumptions that investigators make in their attempts to specify econometric models for explaining the behavior of tourism demand. So long as the assumptions do not distort the estimated results, they perform an important role in facilitating the provision of information which can be useful for policy formation and decision making. When the assumptions are inadequate and questionable, they give rise to models that may embody misspecification bias, leading to accurate and unreliable estimation results that cannot be used for any sound inference and forecasting or policy purposes.

In the last few decades, numerous researchers have studied international tourism demand and a wide range of the available forecasting techniques have been tested. Major focus has been given to econometric studies that involve the use of least squares regression to estimate the quantitative relationship between tourism demand and its determinants.

2.2.2 Tourism Demand

The concepts of tourism demand and forecasting, almost all forecasting involves predicting the tourism demand at some point in the future. In the neoclassical conception of demand, the tourism perspective these include age, education, tastes, previous experience with the product, advertising, product innovation, government policy or new technology. As a luxury good, the demand for tourism tends to be quite elastic while the income elasticity of different tourism products can differ considerably, as some recreation goods may actually show declining consumption with increasing income.

Demand forecasting in tourism research is reviewed from the perspective of which method is most appropriate given the research question, the time period specified and the information needs of managers. Factors which will govern the choice of method include the purpose, the time period being forecasted, the degree of accuracy required, the availability of information, the forecasting environment and the cost of producing the forecast. Inaccuracies in forecasting result may result from five different factors: inappropriate model, incorrect use, error calculation in relationships in model, significant variables omitted, and data used may have been inadequate or inappropriate. A review of quantitative, qualitative and technological forecasting techniques and the factors which influence tourism demand are also included.

The concept of theory uses in international tourist demand since 1950 but the estimation in international tourist demand by econometric method beginning from the first time by Artus(1972). After that a lot of research about international tourist demand function used the econometric method. The researchers studied this research for example Archer (1976), Crouch (1994), Lim (1997), Sinclair (1998), McAleer (2001).

The choice of where to allocate scarce resources among competing choices depends upon an individual's underlying utility function. Tourism is merely one of the many ways consumers can spend surplus money and leisure hours and individuals will engage in it with different propensities. Nonetheless, despite many alternative options for spending or saving disposable income, many consumers choose to direct expenditures at travel services. Once consumers have chosen to travel, they face another decision: whether to travel overseas, and if so, to where. The numbers of substitute destinations are nearly infinite, and will appeal to different individuals for different reasons. Some prefer domestic travel while the more adventurous strive for distant and exotic locales. Certain travelers respond to the call of urban sophistication, to others nature beckons their attention. According to consumer maximization theory, individuals will choose destinations based on an optimization of utility. Faced with income and budgetary limitations, consumers choose between competing destinations. For a given individual, some destinations will be less attractive due to the length of time involved in getting there and the expenses incurred upon arrival. Like all goods, the price of tourism factors into their decision-making process. However, unlike most other goods, tourism must be consumed at the point of supply, further complicating the consumer choice problem. Destinations cannot be packaged attractively and sold at local markets; tourism choices, by definition, account for the willingness of consumers to travel to, and live temporarily in, a given destination. Thus scenery, climate, prejudices, cultural attractions, and many other attributes will affect consumer choices in conjunction with prices. Therefore, the factors influencing consumer maximization theory for tourism will be different than for other goods. To approximate the attributes that influence consumer choices, one can construct a model for tourism demand.

The factors that influenced in tourism demand of foreign tourists (Crouch, 1991)

1) Income level regards the important factor in demand tourism specification of foreign tourist, which found that the elasticity of tourism is the luxuries (Luxury Goods).

2) Relative price of foreign countries, exchange rates, cost of transportation, opportunity cost, and the risk in the foreign travels.

3) Marketing, or budget expenses supports the tourism advertising.

4) Trends and fashion, and special events and Dummy Variables, such as no stability political and social conflict, rebellion, the limitation in the exchange rate, rate change reduces duty-free goods, the economic

condition declines, exhibition arrangement, sport international competition, critical oil price, and national celebration, etc.

5) Other factors, Lag and Lead Effects, Short–Term and Long–Term Effects and the competition between other countries.

International Tourism Demand Model can write as follows, (Lim, 1997)

$$DT_{ij} = f(Y_i, TC_{ij}, RP_{ij}, ER_{ij}, QF_j)$$

DT_{ij} = tourism demand of from country i to country j

Y_{ij} = income level of foreign tourist from country I to country j

TC_{ij} = Cost of Transportation

RP_{ij} = Relative price of foreign countries

ER_{ij} = exchange foreign currency rate ,

QF_j = the quality of side tourism factor in destination country

Thus the variable in demand tourism model of foreign tourist, (Lim, 1997: 838; Lim and McAleer, 2003), compose.

2.2.3 Panel Data Analysis

Panel data analysis is an increasingly popular form of longitudinal data analysis among social and behavioral science researchers. The term "panel data" refers to the pooling of observations on a cross-section of households, countries, firms, etc. over several time periods. This can be achieved by surveying a number of households or individuals and following them over the time.

Panel data analysis is a method of studying a particular subject within multiple sites, periodically observed over a defined time frame. Within the social sciences, panel analysis has enabled researchers to undertake longitudinal analyses in a wide variety of fields. In economics, panel data analysis is used to study the behavior of firms and wages of people over time. With repeated observations of enough cross-sections, panel analysis permits the researcher to study the dynamics of change with short time series. The combination of time series with cross-sections can enhance the quality and quantity of data in ways that would be impossible using only one of these two dimensions (Gujarati, 2003).

Panel analysis can provide a rich and powerful study of a set of people, if one is willing to consider both the space and time dimension of the data. Panel data analysis endows regression analysis with both a spatial and temporal dimension. The spatial dimension pertains to a set of cross-sectional units of observation. These could be countries, states, counties, firms, commodities, groups of people, or even individuals. The temporal dimension pertains to periodic observations of a set of variables characterizing these cross-sectional units over a particular time span. Sayrs (1989) writes that under some circumstances the cross-sections may be nested within time. If there are no missing values, the data set is called a balanced panel, but if there are missing values, the data set is referred to as an unbalanced panel.

Hsiao (1985, 1986), Klevmarken (1989) and Salon (1989) list several benefits from using panel data. These include the following: (Baltagi, 2002)

1. Controlling for individual heterogeneity. Panel data suggest that individuals, firms, states or countries and heterogeneous. Time-series and cross-section studies not controlling for this heterogeneity run the risk of obtaining biased results

2. Panel data give more informative data, more variability, less collinearity amount the variables, more degrees of freedom and more efficiency

3. Panel data are better able to study the dynamics of adjustment. Cross-sectional distributions that look relatively stable hide a multitude of changes. Spells of unemployment, job turnover, residential and income mobility are better studied with panels. Panel data are also well suited to study the duration of economic state like unemployment and poverty, and if these panels are long enough, they can shed light on the speed of adjustments to economic policy changes.

4. Panel data are better able to identify and measure effects that are simply not detectable in pure cross-section or pure time-series data. Ben-Porath (1973) gives an example. Suppose that we have a cross-section of women with a 50% average yearly labor force participation rate. This might be due to (a) each women having a 50% chance of being in the labor force, in any given year, or (b) 50% of the women working all the time and 50% or not at all. Case (a) has high turnover, while case (b) has no turnover. Only panel data could discriminate between these cases.

5. Panel data models allow us to construct and test more complicated behavioral models than purely cross-section or time-series data. For example, technical efficiency is better studied and modeled with panels also, fewer restrictions can be imposed in panels on a distributed lag model than in a purely time-series study (see Hsiao, 1986).

6. Panel data are usually gathered on micro units, like individuals, firms and households. Many variables can be more accurately measured at the micro level, and biases resulting from aggregation over firms or individuals are eliminated

2.2.4 Types of Panel Analytic Models

There are several types of panel data analytic models. There are constant coefficients models, fixed effects models, and random effects models. Among these types of models are dynamic panel, robust, and covariance structure models. (Yaffee . 2003)

1) The Constant Coefficients Model

Constant coefficients, referring to both intercepts and slopes. we could pool all of the data and run an ordinary least squares regression model in the case that there is neither significant country nor significant temporal effects,. There are occasions when either country (for example)or temporal effects are not statistically significant. This model is sometimes called the pooled regression model.

2) The Fixed Effects Model (Least Squares Dummy Variable Model)

Another type of panel model would have constant slopes but intercepts that differ according to the cross-sectional (group) unit—for example, the country. Although there are no significant temporal effects, there are significant differences among countries in this type of model. These models are called fixed effects models. Because *i-1* dummy variables are used to designate the particular country, this same model is sometimes called the Least Squares Dummy Variable model (see Eq. 1).

$$K_{it} = b_1 + b_2 Country_1 + b_2 Country_2 + \alpha_2 DI_{2it} + \alpha_3 IIinc_{3it} + u_{it} \qquad \text{(Eq.1)}$$

Panel data sets generally include sequential blocks or cross-sections of data, within each of which resides a time series. For this example equation

,a typical panel data set, including country, year, personal disposable income(DI) and median household income(IIinc) from 2001 through 2010.

Apart from the variable number, the data structure confers upon the variables two dimensions. They have a cross-sectional unit of observation, which in this case is country *i*. They have a temporal reference, *t*, in this case the year. The error term has two dimensions, one for the country and one for the time period. In this exemplar, assume that there are three countries and ten years of time. Even though time is nested within the cross-section in this example, Lois Sayrs (1989) writes that under some circumstances the cross-sections may be nested within time. If there are no missing values, the data set is called a balanced panel, but if there are missing values, the data set is referred to as an unbalanced panel.

Another type of fixed effects model could have constant slopes but intercepts that differ according to time. In this case, the model would have no significant country differences but might have autocorrelation owing to time-lagged temporal effects. The residuals of this kind of model may have autocorrelation in the process. In this case, the variables are homogenous across the countries. They could be similar in region or area of focus. For example, technological changes or national policies would lead to group specific characteristics that may effect temporal changes in the variables being analyzed. We could acount for the time effect over the *t* years with *t-1* dummy variables on the right-hand side of the equation. In equation 2, the dummy variables are named according to the year they represent.

$$\begin{aligned}K_{it} &= b_1 + \lambda_2 Year\,2001 + \lambda_3 Year\,2002 + \ldots + \lambda_{10} Year\,2009 \\ &+ \alpha_i DI_{it} + \alpha_2 IIinc_{it} + u_{it}\end{aligned} \qquad \text{(Eq.2)}$$

There is another fixed effects panel model where the slope coefficients are constant, but the intercept varies over country as well as time. (see Eq. 3)

$$\begin{aligned}K_{it} &= b_0 + b_1 Country_1 + b_2 Country_2 \\ &+ \lambda_0 + \lambda_1 Year\ 2001 + \lambda_2 Year\ 2002 + \ldots + \lambda_9 Year\ 2009 \\ &+ \alpha DI_{1t} + \alpha_2 IIinc_{2t} + u_{it}\end{aligned} \qquad \text{(Eq.3)}$$

Another type of fixed effects model has differential intercepts and slopes. This kind of model has intercepts and slopes that both vary according to the country. To formulate this model, we would include not only country dummies, but also their interactions with the time-varying covariates (Eq. 4).

$$\begin{aligned}K_{it} &= b_1 + b_2 Country_2 + b_3 Country_3 \\ &+ \alpha_2 DI_{2it} + \alpha_3 IIinc_{3it} + \\ &+ \alpha_4 * Country_2 * DI_{2it} + \alpha_5 * Country_3 * DI_{3it} \\ &+ \alpha_6 * Country_2 * IIinc_{3it} + \alpha_7 * Country_3 * IIinc_{3it} + u_{it}\end{aligned} \qquad \text{(Eq. 4)}$$

In this model, the intercepts and intercepts vary with the country. The intercept for Country1 would be b_1. The intercept for $Country_2$ would also include an additional intercept, b_2, so the intercept for $Country_2$ would be b_1+b_2. The intercept for $Country_3$ would include an additional intercept. Hence, its intercept would be b_1 + b_3. In this way, the intercepts and slopes vary with the country. There is also a fixed effects panel model in which both intercepts and slopes might vary according to country and time. This model specifies *i-1* Country dummies, *t-1* Time Dummies, the variables under consideration and the interactions between them. If all of these are statistically

3) The Random Effects Model

Greene calls the random effects model a regression with a random constant term (Greene, 2003). One way to handle the ignorance or error is to assume that the intercept is a random outcome variable. The random outcome is a function of a mean value plus a random error. But this cross-sectional specific error term v_i, which indicates the deviation from the constant of the cross-sectional unit (in this example, country) must be uncorrelated with the errors of the variables if this is to be modeled. The time series cross-sectional regression model is one with an intercept that is a random effect. (Eq. 5)

$$y_{it} = \beta_{\theta i} + \beta_t x_{it} + \beta_2 x_{it} + e_{it}$$

$$\beta_{\theta i} = \beta_i + v_i$$

$$\therefore y_{it} = \beta_i + \beta_1 x_{it} + \beta_2 x_{it} + e_{it} + v_i \qquad \text{(Eq. 5)}$$

Under these circumstances, the random error v_i is heterogeneity specific to a cross-sectional unit—in this case, country. This random error v_i is constant over time. Therefore, $E[v_i^2 \mid x] = \sigma_i^2$ The random error e_{it} is specific to a particular observation. For v_i to be properly specified, it must be orthogonal to the individual effects. Because of the separate cross-sectional error term, these models are sometimes called one-way random effects models. Owing to this intrapanel variation, the random effects model has the distinct advantage of allowing for time-invariant variables to be included among the repressors.

If, however, the random effects model depends on both the cross-section and the time series within it, the error components (sometimes referred to as variance components) models are referred to as a two-way random effects model. In that case, the error term should be uncorrelated with the time series component and the cross-sectional (group) error. The orthogonally of these components allows the general error to be decomposed into cross-sectional specific, temporal, and individual error components.

$$u_{it} = v_i + e_t + \eta_{it} \qquad \text{(Eq 6)}$$

In equation 6,the component, v_i, is the cross-section specific error. It affects only the observations in that panel. Another, e_t, is the time-specific component. This error component is peculiar to all observations for that time period, *t*. The third η_{it} affects only the particular observation. These models are sometimes referred to as two-way random effects models (SAS Institute, 1999).

In the Hildreth, Houck, and Swamy random coefficient model, the parameters are allowed to vary over the cross-sectional units. This model allows both random intercept and slope parameters that vary around common means. The random parameters can be considered outcomes of a common mean plus an error term, representing a mean deviation for each individual. This model assumes neither heteroskedasticity nor autocorrelation within the panels to avoid complicating the covariance matrix. In multilevel models pertaining to students, schools, and cities, there can be individual student, school, and city random error terms as well. There can also be cross-level interactions within these hierarchical models.

4) Dynamic Panel Models

Garín-Muñoz and Pérez-Amaral (2000) suggested that tourism has a great deal of inertia, so that the dynamic structure of consumer preference should be considered in the tourism demand model (Garín-Muñoz, 2006). In particular, if the impact of previous tourism is neglected, the estimated results of other relevant variables will be overestimated. Furthermore, Song and Witt (2000) noted that the static regressions of tourism demand models might raise some significant problems, such as structural instability, forecasting failures and spurious regression. Hence, including the lagged dependent variable in a dynamic model of tourism demand is one way of sensibly accommodating the dynamic structure of consumer preferences, where changes in tastes might be regarded as endogenous (Garín-Muñoz and Pérez-Amaral, 2000; Garín-Muñoz, 2006; Ledesma-Rodríguez, Navarro-Ibáñez and Pérez-Rodríguez, 2001). In our paper, the lagged dependent variable of tourism demand, which will be interpreted as being based on habit formation or as interdependent preferences, are included as regressors to consider the possibility of a change in consumer preferences over time.

Consider the simple model for dynamic panel data (Eq. 7), Weinhold (1999)

$$D_{it} = \gamma D_{it-1} + \beta X_{it} + \omega_{it} \qquad \text{(Eq. 7)}$$

where $\omega_{it} = \kappa_i + \eta_{it}$ and i = 1, …,N cross section units and t = 1, …,T time periods. There is a clear simultaneity problem as the lagged dependent variable a_{it-1} is correlated with the error term ω_{it} by virtue of its correlation with the time-invariant component of the error term, κ_i. Nickell (1981) has shown that even if the “fixed effects” (FE) or Least Squares Dummy Variable (LSDV) is used, D_{it-1} will still be correlated with the error term and the resulting bias will be of *O*(1/*T*). Andersen and Hsiao (1981) and Hsiao (1986) both provide extensive discussions of this bias.

The usual approach for dealing with this problem is to first first-difference the data to remove the κ_i which yields:

$$D_{it} - D_{it-1} = \gamma\left(D_{it} - D_{it-1}\right) + \beta\left(X_{it} - X_{it-1}\right) + \left(\omega_{it} - \omega_{it-1}\right) \qquad \text{(Eq. 8)}$$

Then, because ΔD_{it-1} is correlated with the first difference error term it is necessary to instrument for it. Andersen and Hsiao (1981) have suggested using ΔD_{it-2} or D_{it-2} as an instrument as these terms are not correlated with $\Delta\omega_{it} = \eta_{it} - \eta_{it-1}$. Arellano (1989) showed that an estimator that uses the levels for instruments has no singularities and displays much smaller variances than does the analogous estimator that uses differences as estimators. Holtz-Eakin et. al.(1988) adopt the approach to panel VAR’s in a framework for testing Granger causality in panels and suggest using a time-varying set of instruments that includes both differences and levels. In addition other instruments have been suggested by a succession of researchers. In practice however it is often difficult to find good instruments for the first-differenced lagged dependent variable, which can itself create problems for the estimation. Kiviet (1995) shows that panel data models that use instrumental variable estimation often lead to poor finite sample efficiency and bias. Using a broad array of Monte Carlo simulations he finds, In particular situations it seems that valid orthogonality restrictions can better not be employed when composing a set of instrumental variables. It is difficult to find clues on when which instrumental variables are better put aside in order to avoid serious small sample bias or relatively large standard deviations, which both entail poor estimator efficiency. As yet, no technique is available that has shown uniform superiority in finite samples over a wide range of relevant situations as far as the true parameter values and the further properties of the data generating mechanism are concerned.

Fortunately, as *T* gets larger this bias becomes less of a problem. Nevertheless it would be useful to have an estimator which did not have such a large bias for small *T* and which did not require instrumental variables estimation. In particular, many cross country data sets have time dimensions of between 15 to 25 years, at which point it is hard to judge whether the Nickell bias or a weak instrument set will do more harm to the estimation. An additional problem of introducing dynamics into a panel data model is the potential bias induced by heterogeneity of the cross section units. Pesaran and Smith (1995) have explored this problem in depth. They show that parameter estimates derived from pooled data are not consistent in dynamic models even for large *N* and *T*. In particular consider a model in which the coefficient on the lagged dependent variable is constrained to be equal across all cross section units so that we have:

$$D_{it} = \alpha_i + \gamma D_{it-1} + \beta X_{it} + \omega_{it} \qquad \text{(Eq. 9)}$$

there could be significant bias introduced if in fact the coefficients on the lagged dependent variable are not constant across the cross section. In this case the difference between the actual value and the estimated coefficient times the dependent variable, $(\gamma_i - \bar{\gamma})D_{it-1}$, will be a component of the error term and this serial correlation induces bias and inconsistency in the estimation. One solution to avoid the bias and inconsistency is to use the first difference transformation, and to treat the lags of the dependent variables as instruments for the lagged dependent variable (Garín-Muñoz, 2006; Ledesma-Rodríguez, Navarro-Ibáñez and Pérez- Rodríguez, 2001).

A generalized method of moments (GMM) approach can be used to unify the estimator and eliminate the disadvantages of reduced sample sizes. As suggested by Arellano and Bond (1991), the list of instruments can be exploited by additional moment conditions and allowing the number to vary with *t,* so that all moment conditions can be estimated by GMM. However, the GMM estimator for *γ* is asymptotically normal, based on the assumptions of homoskedastic and uncorrelated errors term. In this paper, the GMM approach is used to compute the panel GMM and GMM-DIFF estimator. The first difference transformation model, namely GMM-DIFF estimator, as suggested by Arellano and Bond (1991), is based on taking first differences to eliminate the individual effects, and regard the dependent variable lagged two or more periods as instruments for the lagged dependent variable

5) Robust Panel Models

There are a number of problems that plague panel data models. Outliers can bias regression slopes, particularly if they have bad leverage. These outliers can be down weighted with the use of M-estimators in the model. Heteroskedasticity problems arise from GroupWise differences, and often taking group means can remove heteroskedasticity. The use of a White heteroskedasticity consistent covariance estimator with ordinary least squares estimation in fixed effects models can yield standard errors robust to unequal variance along the predicted line (Greene, 2002; Wooldridge, 2002).

Sometimes autocorrelation inheres within the panels from one time period to another. Some problems with dynamic panels that contain autocorrelation in the residuals are handled with a Prais-Winston transformation or a Cochrane-Orcutt transformation that amounts to a first partial differencing to remove the bias from the autocorrelation. Arellano, Bond, and Bover developed one and two step general methods of moments (GMM) estimators for panel data analysis. GMM is usually robust to deviations of the underlying data generation process to violations of heteroskedasticity and normality, insofar as they are asymptotically normal but they are not always the most efficient estimators. If there is autocorrelation in the models, one can obtain a weight-adjusted combination of the White and Newey-West estimator to handle both the heteroskedasticity and the autocorrelation in the model.

The Hausman specification test is the classical test of whether the fixed or random effects model should be used. The research question is whether there is significant correlation between the unobserved person-specific random effects and the repressors. If there is no such correlation, then the random effects model may be more powerful and parsimonious. If there is such a correlation, the random effects model would be inconsistently estimated and the fixed effects model would be the model of choice. The test for this correlation is a comparison of the covariance matrix of the repressors in the LSDV model with those in the random effects model. The null hypothesis is that there is no correlation. If there is no statistically significant difference between the covariance matrices of the two models, then the correlations of the random effects with the repressors are statistically insignificant. The Hausman test is a kind of Wald χ^2 test with *k-1* degrees of freedom (where *k*=number of repressors) on the difference matrix between the variance-covariance of the LSDV with that of the Random Effects model.

2.2.5 Model Estimation

Models have to be estimated by methods that handle the problems afflicting them. A constant coefficients model with residual homogeneity and normality can be estimated with ordinary least squares estimation (OLS). As long as there is no groupwise or other heteroskedastic effects on the dependent variable, OLS may be used for fixed effects model estimation as well (Sayrs, 1989). For OLS to be properly applied, the errors have to be independent and homoskedastic. Those conditions are so rare that is often unrealistic to expect that OLS will suffice for such models (Davidson and MacKinnon, 1993). Heteroskedastic models are usually fitted with estimated or feasible generalized least squares (EGLS or FGLS). Heteroskedasticity can be assessed with a White or a Breusch-Pagan test. For the most part, fixed effects models with groupwise heteroskedasticity cannot be efficiently estimated with OLS. If the sample size is large enough and autocorrelation plagues the errors, FGLS can be used. Random sampling and maximum likelihood iterated by generalized least squares have also been used (Greene, 2002). Beck and Katz (1995) reportedly found that if the sample size is finite or small, the total number of temporal observations must be as large as the number of panels; moreover they reportedly found that OLS with panel corrected errors provided more efficient estimation than FGLS (Greenberg, 2003; STATA, 2003). If the model exhibits autocorrelation and/or moving average errors, first differences (Wooldridge, 2002) or GLS corrected for ARMA errors can be used (Sayrs, 1989).

Hausman and Taylor (1981) have used weighted instrumental variables, based only on the information within the model, for random effects estimation to be used when there are enough instruments for the modeling. The instrumental variables, which are proxy variables uncorrelated with the errors, are based on the group means. The use of these instrumental variables allows researchers to circumvent the inconsistency and inefficiency problems following from correlation of the individual variables with the errors. For dynamic panels with lagged dependent variables, Arellano, Bond, and Bover have used general methods of moments, which are asymptotically normal (Wooldridge, 2002). With greater numbers of moment conditions, they are able to handle some missing data and they can attain gains in efficiency as long as there are three or four periods of data (Greene, 2002). Another estimation procedure was developed by Arnold Zellner, called seemingly unrelated regression (SUR) requires that the number of explanatory variables in each cross-section is the same. In the SUR approach, variables are transformed with a form of Cochrane-Orchutt correction to model the autocorrelation. Feasible generalized least squares is used to estimate a covariance matrix. The parameter estimates are also modeled. The process is iterated until the errors are minimized.

2.2.6 Estimation Procedure

There are different models in panel data estimation and these are pooled, fixed and random effects. The pooled model assumes that countries are homogeneous, while fixed and random effects introduce heterogeneity in the estimation. A decision should be made whether to use random or fixed model because individual effects are included in the regression. A random effects model is appropriate when estimating the model between a country and its randomly selected sample of trading partners from a large group (population).

A fixed effects model is appropriate when estimating the model between a country and predetermined selection of trading partners. The study uses the Hausman test to check whether fixed effects is more appropriate than the random effects model. The fixed effects model will be better than the random effects model if the null hypothesis of no correlation between individual effects and the regressors is rejected. The fixed effects model cannot directly estimate variables that do not change over time because inherent transformation wipes out such variables.

Despite the strengths of fixed and random effects estimators based on panel data, there remains two further shortcomings that needs to be dealt with. These are the potential endogeneity of the **x**j, as well as the loss of dynamic information. If there are persistence/ reputation effects that apply over time in tourist decision on holiday destinations, for example when tourists return to a particular destination when they had a good experience, then this might be a serious omission .To overcome this problem of endogeneity, an instrumental variable needs to be used for two approaches, namely Anderson and Hsiao's (1982) instrumental variable (IV) and Arellano and Bond (1991) two GMM estimators (first-step and second-step, respectively) have been used in this regard.

2.2.7 Panel Unit-Root Tests

Recent literature suggests that panel-based unit root test have higher power than unit root test based on individual time series, see Levin, Lin and Chu (2002), Im, Persaran and Shin (2003), and Breitung (2000) to mention a few of popular test purchasing power parity (PPP) and growth convergence in macro panels using country data over time. This research focus on four type of panel unit root test such as Levin, Lin and Chu (2002), Breitung (2000), Im, Pesaran and Shin (2003), Fisher-Type test using ADF and PP-test (Maddala and Wu (1999) and Choi (2001)).

Testing for unit root is the first step in determining a potentially cointegrated relationship between variables. If all variables do not contain a unit root, the traditional estimation methods can be used to estimate the relationship between variables. If variables are nonstationary, a test for cointegration is required. The literature identifies three types of unit root tests. The first test is Levin, Lin and Chu (2002) and it is referred to as the LLC test. The second test is that of Hadri (2000). These two types of panel unit root test assume that the autoregressive parameters are common across cross-sections. The LLC uses the null hypothesis of a unit root while Hadri uses the null hypothesis of no unit root. Im, Pesaran and Shin (2003) developed a third type of panel unit root test called IPS. This test allows for autoregressive parameters to differ across cross-sections and also for individual unit root processes. It is computed by combining individual cross-section unit root tests in order to come up with a test that is specific to the panel. This test has more power than the single-equation Augmented Dickey-Fuller (ADF) by averaging N independent regressions (Strauss and Yigit, 2003). The ADF specification may include intercept but no trend or may include an intercept and time trend. It uses the null hypothesis that all series have a unit root and the alternative hypothesis is that at least one series in the panel has a unit root. This test is one-tailed or lower tailed based on the normal distribution. This study uses LLC and the IPS to test for unit root**.**

2.2.8 Estimating panel cointegration model

The various (casually single equation) approach for estimating a cointegration vector using panel data such as Pedroni (2000, 2001) approach, Chiang and Kao (2000, 2002) approach and Breitung (2002) approach. For this research we use Chiang and Kao (2000, 2002) approach to estimate panel cointegration. Kao (1999) uses both DF and ADF to test for cointegration in panel as well as this test similar to the standard approach adopted in the EG-step procedures. Also this test start with the panel regression model as set out in equation (10).

$$Y_{it} = X_{it}\,ß_{it} + Z_{it}\,\gamma_0 + \varepsilon_{it} \qquad \text{(Eq.10)}$$

where Y and X are presumed to be non-stationary and: (see equation (11))

$$e^{\wedge}_{it} = \rho\, e^{\wedge}_{it} + \nu_{it} \qquad \text{(Eq. 11)}$$

where $e^{\wedge}_{it} = (Y_{it} - X_{it}\,\overset{\wedge}{ß}_{it} - Z_{it}\,\gamma^{\wedge})$ are the residuals from estimating equation (11). To test the null hypothesis of no cointegrarion amounts to test H0 : $\rho = 1$ in equation 21I against the alternative that Y and X are conitegrated (i, e., H1: $\rho < 1$). Kao (1999) developed both DF-Type test statistics and ADF test statistics were used to test cointegration in panel also both DF-Type (4 Type) test statistics and ADF test statistics can present below that:

$$DF_{\rho} = \frac{\sqrt{N}T(\hat{\rho} - 1) + 3\sqrt{N}}{\sqrt{51/5}},$$

(Eq. 12)

$$DF_{t} = \sqrt{\frac{5t_{\rho}}{4}} + \sqrt{\frac{15N}{8}}.$$

(Eq. 13)

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

(Eq. 14)

$$DF^{*}_{t} = \frac{t_{\rho} + \frac{\sqrt{6N}\hat{\sigma}_{v}}{2\hat{\sigma}_{0v}}}{\sqrt{\frac{\hat{\sigma}^{2}_{0v}}{2\hat{\sigma}^{2}_{v}} + \frac{3\hat{\sigma}^{2}_{v}}{10\hat{\sigma}^{2}_{0v}}}},$$

(Eq. 15)

$$ADF = \frac{t_{ADF} + \sqrt{6N}\hat{\sigma}_{v}/2\hat{\sigma}_{0u}}{\sqrt{\hat{\sigma}^{2}_{0v}/2\hat{\sigma}^{2}_{v} + 3\hat{\sigma}^{2}_{v}/10\hat{\sigma}^{2}_{0v}}}$$

(Eq. 16)

where:

N = cross-section data;

T = time series data;

$\rho^{\wedge}$ = co-efficiencies of (12);

$t\ \rho = [(\rho^{\wedge} -1) \surd (\Sigma^{N}_{i=1} \Sigma^{T}_{t=2}\, e^{\wedge*2}_{i\ t-1})]/Se;$

$Se = (1/NT)\ \Sigma^{N}_{i=1} \Sigma^{T}_{t=2}(e^{\wedge*}_{i\,t} - \rho^{\wedge} e^{\wedge*}_{i\,t-1})^{2}$;

$\sigma u^{\wedge 2}$ = variance of u;

$\sigma v^{\wedge 2}$ = variance of v

$\sigma u^{\wedge}$ = standard deviation of u;

$\sigma v^{\wedge}$ = standard deviation of v;

$t_{ADF} = [(\rho^{\wedge} -1) (\Sigma^{N}_{i=1} (e / Q_{i}\, e_{i}))^{1/2}]/ S_{v}$.

The various estimators available include with-and between-group such as OLS estimators and dynamic OLS estimators. OLS and DOLS are a parametric approach which DOLS estimators include lagged first-differenced term are explicitly estimated as well as consider a simple two variable panel regression model: (see detail calculated of OLS and DOLS in equation 17, 18,19 and 20).

$$Y_{i\,t} = \alpha_i + \beta_i X_{i\,t} + \varepsilon_{i\,t} \qquad \text{(Eq. 17)}$$

A standard panel OLS estimator for the coefficient β_i given by :

$$\hat{\beta}_{i\,,\,OLS} = [\Sigma^{N}_{i=1}\Sigma^{T}_{t=1}(X_{i\,t} - X^{*}_{i})^{2}]^{-1}\,\Sigma^{N}_{i=1}\Sigma^{T}_{t=1}(X_{i\,t} - X^{*}_{i})\,(Y_{i\,t} - Y^{*}_{i}) \quad \text{(Eq. 18)}$$

where

i = cross-section data and N is the number of cross-section

t = time series data and T is the number of time series data

$\hat{\beta}_{i\,OLS}$ = A standard panel OLS estimator

$X_{i\,t}$ = exogenous variable in model

X^{*}_{i} = average of X^{*}_{i}

$Y_{i\,t}$ = endogenous variable in model

Y^{*}_{i} = average of Y^{*}_{i}

Pedroni (2001) has also constructed a between-dimension, group-means panel DOLS estimator that incorporates corrections for endogeneity and serial correlation parametrically. This is done by modifying equation 22I to include lead and lag dynamics: (see equation 19).

$$Y_{i\,t} = \alpha_i + \beta_i X_{i\,t} + \Sigma^{ki}_{j=-k}\,\gamma_{ik}\,\Delta X_{i\,,t-k} + \varepsilon_{i\,t} \qquad \text{(Eq. 19)}$$

where

$$\beta^{\wedge}_{i,\,DOLS} = [N^{-1}\,\Sigma^{N}_{i=1}(\Sigma^{T}_{t=1}\ Z_{it}\,Z^{*}_{it})^{-1}(\ \Sigma^{T}_{t=1}Z_{it}\,Z^{\wedge}_{it})] \qquad (\text{Eq. } 20)$$

and where

i = cross-section data and N is number of cross-section data

t = time series data and T is number of time series data

$\beta^{\wedge}_{i\ DOLS}$ = dynamics OLS estimator

$Z_{i\,t}$ = is the 2(K+1) x 1

$Z^{\wedge}_{it}$ = $(X_{i\,t} - X^{*}_{i})$

X^{*}_{i} = average of X^{*}_{i}

$\Delta X_{i\,,t-k}$ = differential term of X

The above methods were mostly developed by Pedroni (2000,2001). This research also focused on the OLS estimator and the DOLS estimator for estimating panel cointegration for modeling international tourism demand of Thailand.

2.2.9 Estimating panel data

A panel is a set of observations on individuals, collected over time. An observation is the pair $\{y_{it}, \chi_{it}\}$, where the i subscript denotes the individual, and the t subscript denotes time. A panel may be balanced:

$$\{y_{it}, \chi_{it}\}\ : t = 1, \ldots, T; i = 1, \ldots, n,$$

or unbalanced:

$$\{y_{it}, \chi_{it}\}\ : \text{For } I = 1, \ldots, n, \qquad t = \underline{t}_i, \ldots, \bar{t}_i\,.$$

1.) Individual-Effects Model

The standard panel data specification is that there an individual-specific effect which enters linearly in the regression

$$y_{it} = \chi'_{it}\beta + \alpha_i + u_{it} \qquad \text{(Eq.21)}$$

The typical maintained assumptions are that the individuals i are mutually independent, that α_i and u_{it} are independent, that u_{it} is iid across individuals and time, and that u_{it} is uncorrelated with χ_{it}.

OLS of y_{it} on χ_{it} is called pooled estimation. It is consistent if

$$E(\chi_{it}\alpha_i) = 0 \qquad \text{(Eq.22)}$$

If this condition fails, then OLS is inconsistent (individual-specific unobserved effect α_i is correlated with the observed explanatory variables χ_{it}). This is often believed to be plausible if u_{i} is an omitted variable. If equation (1) is true, however, OLS can be improved upon via a GLS technique. In either event, OLS appears a poor estimation choice. Condition equation (22) is called the *random hypothesis*. It is a strong assumption, and most applied researchers try to avoid its use.

2.) Fixed Effects

This is the most common technique for estimation of non-dynamic linear panel regressions.

The motivation is to allow α_i to be arbitrary, and have arbitrary correlated with χ_{it}. The goal is to eliminate α_i from the estimator, and thus achieve invariance.

There are several derivations of the estimators.

First, let

$$d_{ij} = \begin{cases} 1 & \text{if} \quad i = j \\ 0 & \text{eles} \end{cases}, \text{ and } d_i = \begin{pmatrix} d_{i1} \\ \vdots \\ d_{in} \end{pmatrix} \qquad (\text{Eq.23})$$

An n x 1 dummy vector with a "1" in the i'th place. Let $\alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix}$ Then note that

$$\alpha_i = d'_i \, \alpha, \qquad (\text{Eq.24})$$

And

$$y_{it} = \chi'_{it} \, \beta + d'_i \, \alpha + u_{it}. \qquad (\text{Eq.25})$$

Observe that

$$E(u_{it} \mid \chi_{it}, d_{it}) = 0, \qquad (\text{Eq.26})$$

So equation (25) is a valid regression, with d_i as a regressor along with χ_i.

OLS on equation (25) yield estimator $(\hat{\beta}, \alpha)$. Conventional inference applies. Observe that this is generally consistent.

- If χ_i contains an intercept, it will be collinear with d_i, so the intercept is typically omitted from χ_{it}.
- Any regressor in χ_{it} which is constant over time for all individuals (e.g., their gender) will be collinear with d_i, so will have to be omitted.
- There are n + k regression parameters, which is quite large as typically n is very large.

Computationally, you do not want to actually implement conventional OLS estimation, as the parameter space is too large. OLS estimation of β proceeds by the FWL theorem (*Frisch-Waugh-Lovell theorem*). Stacking the observations together:

$$y = \chi\beta + Du + e, \qquad (\text{Eq.27})$$

Then by the FWL theorem,

$$\hat{\beta} = (X'(I - P_D)X)^{-1}(X'(I - P_D)y) \qquad (\text{Eq.28})$$

$$= (X^{*\prime}X^{*})^{-1}\,(X^{*\prime}y^{*}), \qquad (\text{Eq.29})$$

where

$$y^* = y - D\left(D'D\right)^{-1} D'y \qquad \text{(Eq.30)}$$

$$X^* = X - D\left(D'D\right)^{-1} D'X. \qquad \text{(Eq.31)}$$

Since the regression of y_{it} on d_i is a regression onto individual-specific dummies, the predicted value from these regressions is the individual specific mean $\bar{y}_{it}$, and the residual is the dream value

$$y^*_{it} = y_{it} - \bar{y}_i. \qquad \text{(Eq.32)}$$

The fixed effects estimator $\hat{\beta}$ is OLS of y^*_{it} on χ^*_{it} the dependent variable and regressors in deviation-from-mean form.

Another derivation of the estimator is to take the equation

$$y_{it} = \chi'_{it}\,\beta + \alpha_i + u_{it}, \qquad \text{(Eq.33)}$$

and then take individual-specific means by taking the average for the i' th individual:

$$\frac{1}{T_i}\sum_{t=\underline{t}_i}^{\bar{t}_i} y_{it} = \frac{1}{T_i}\sum_{t=\underline{t}_i}^{\bar{t}_i} \chi'_{it}\,\beta + \alpha_i + \frac{1}{T_i}\sum_{t=\underline{t}_i}^{\bar{t}_i} u_{it} \qquad \text{(Eq.34)}$$

or

$$\bar{y}_{it} = \bar{\chi}'_i\,\beta + \alpha_i + \bar{u}_i. \qquad \text{(Eq.35)}$$

Subtracting, we find

$$y^*_{it} = \chi^{*\prime}_{it}\,\beta + u^*_{it}, \qquad \text{(Eq.36)}$$

which is free of the individual-effect u_i.

3.) Dynamic Panel Regression

A dynamic panel regression has a lagged dependent variable

$$y_{it} = \omega y_{it-1} + \chi'_{it}\,\beta + \alpha_i + u_{it}. \qquad \text{(Eq.37)}$$

This is a model suitable for studying dynamic behavior of individual agents.

Unfortunately, the fixed effects estimator is inconsistent, at least of T is held finite as n $\rightarrow \infty$. This is because the sample mean of y_{it-1} is correlated with that of u_{it} The standard approach to estimate a dynamic panel is to combine first-differencing with IV or GMM. Taking first-differences of equation (37) eliminates the individual-specific effect:

$$\Delta y_{it} = \omega \Delta y_{it-1} + \Delta \chi'_{it} \beta + \Delta \alpha_{it}. \qquad \text{(Eq.38)}$$

$$E(\Delta y_{it-1} \Delta \alpha_{it}) = E((y_{it-1} - y_{it-2})(\alpha_{it} - \alpha_{it-1})) = -E(y_{it-1} \alpha_{it-1}) = -\sigma_e^2. \qquad \text{(Eq.39)}$$

However, if u_{it} is iid, then it will be correlated with Δy_{it-1} :

So OLS on equation (38) will be inconsistent.

But if there are valid instruments, then IV or GMM can be used to estimate the equation. Typically, we use lags of the dependent variable, two periods back, as y_{t-2} is uncorrelated with $\Delta \alpha u_{it}$. Thus values of $y_{it-k}, k \geq 2$, are valid instruments.

Hence a valid estimator of α and β is to estimate (Eq.35) by IV using y_{t-2} as an instrument for Δy_{t-1} (which is just identified). Alternatively, GMM using y_{t-2} and y_{t-3} as instruments (which is overidentified, but loses a time-series observation).

A more sophisticated GMM estimator recognizes that for time-periods later in the sample, there are more instruments available, so the instrument list should be different for each equation. This is conveniently organized by the GMM principle, as this enables the moments from the different time-periods to be stacked together to create a list of all the moment conditions. A simple application of GMM yields the parameter estimates and standard errors.