28. Secondary data modeling in tourism and hospitality research

Boopen Seetanah

BUILDING SIMPLE MODELS

Regression analysis is a statistical tool to investigate relationships among variables. Usually, the investigator seeks to ascertain the causal effect of one variable on another; for instance, the effect of changes in tourists' level of income on the demand for international travel and tourism. Thus, to explore such issues, the researcher may collect data on tourism recipients for a specific destination or country, or for a sample of countries on the relevant variables (the variables could be *tourism arrival* as a measure of tourism demand on the *GDP per capita* to measure income capacity of the tourist), and employs regression to estimate the quantitative effect of the causal variable, *income* in our case (known as the explanatory or independent variable), on the variable that they influence, *tourism arrival* (this is known as the explained variable or dependent variable). The investigator also typically assesses the statistical significance of the estimated relationships, that is, the degree of confidence that the true relationship is close to the estimated relationship.

It is noteworthy that regression techniques have long been central to the field of economic statistics (econometrics). Increasingly, they have become important to other disciplines such as tourism. Thus at the outset of a regression study, one formulates some hypothesis about the relationship between the variables of interest. This is based on solid theoretical underpinnings. Coming back to the previous example on the tourism demand, theory and common experience suggest that if the potential tourist earns more, this increased earning capacity will tend make them engage in more leisure, and they are more likely to travel. It further suggests that the causal relation likely runs from income capacity to tourism demand rather than the other way around. Thus, the tentative hypothesis is that higher income levels of tourists cause higher demand for tourism, other things being equal (as, surely, income capacity is not the only factor affecting demand for tourism).

Thus the theoretical model can be written as:

$$TOURISM = f(INCOME) \tag{28.1}$$

where *TOURISM* represents the demand for tourism, measured by number of tourism arrival, and *INCOME* represents the income capacity of tourists (usually proxied as the gross domestic product per capita of the origin country). It should be noted that there exist alternative measurements for the above variables, and the researcher should stand guided by the literature or by rational justifications.

If one gathers data on *TOURISM* and *INCOME* as measured above, these being readily available from various data sources, one can plot this information using a two-dimensional diagram, conventionally termed a scatter diagram,[1] which gives an

overall visual idea of the relationship between the variables. However, to investigate this hypothesis in statistical terms, the above equation (equation 28.1) should be translated to an estimatable form, a regression model which can be written in the form of:

$$TOURISM = \beta_0 + \beta_1 INCOME + \varepsilon \quad (28.2)$$

where β_0 is a constant term (also known as an exogenous term). It represents the expected value for the dependent variable if all of the independent variables are 0. For instance, there may exist some level of tourism which may not depend on any explanatory factors; for instance, a positive constant term may denote a positive perception on the destination.

β_1 is known as the coefficient of determination and measures the exact relationship between the two variables under study (since it is a bivariate equation, it also represents the correlation between the two variables). These coefficients are computed by the regression tool. They are thus values that represent the strength and type of relationship the explanatory variable (*INCOME*) to the explained variable (*TOURISM*). For instance if *INCOME* is increased by US$1 (assuming the income is measured in USD), it would imply that, other things remaining constant, *TOURISM* will increase by β_1 persons. This is quite an interesting statistical relationship indeed.

ε is known as the noise or error term. It is the unexplained portion of the explained variable (*TOURISM*) represented in the regression equation. In fact, in a simple way, it measures the effects of other variables (omitted variables, which have not been taken into account in the simple specification 28.2 above) which may have some potential effects on tourism as well, as it is clear that *TOURISM* does not depend only on *INCOME*. Of course one would not be happy to have an overwhelmingly large error term, which would imply that the researcher has missed out a number of important and influential factors in the specification. Thus large residuals indicate poor model fit.

The above specified model is known as a bivariate model, as only two variables are included. It is also assumed that *INCOME* affects *TOURISM* in a linear fashion. Although this linearity assumption is very common in regression studies, including tourism studies, it is by no means essential to the application of the technique and can be relaxed where the investigator has a theoretical reason to suppose a priori that the relationship in question is non-linear.

Multivariate Regression Analysis

Undoubtedly, *TOURISM* is affected by various other factors in addition to *INCOME*, factors that were aggregated into the residual term in the simple regression model above. Multiple regressions is a technique that allows additional factors to enter the analysis into a single specification so that the effect of each factor can be estimated. It is valuable for quantifying the impact of various simultaneous influences upon a single dependent variable. Further, because of omitted variables bias with simple bivariate regression, multivariate regression is often essential even when the researcher is only interested in the effects of one of the independent variables.

Building from the previous regression model (equation 28.1), an investigator could argue that (in fact, the researcher should stand guided by the theories and past studies) exchange rate (as measured by *EXCHANGE* for instance, measured as the exchange

rate between Mauritian rupees, MUR, and USD) and cost associated for the travel (this may include cost of air fare and cost of living in the destination country, assumed to be measured by an index *COST*, in MUR) are also valid potential explanatory variables. It is noteworthy that these are by no means the only variables which may influence tourism. There are definitely other additional explanatory variables, such as language, borders, level of development of the destination country, among others. However, for ease of explanation, the following relationship is proposed:

$$TOURISM = f(INCOME, EXCHANGE, COST) \tag{28.3}$$

Theoretically, one expects a positive relationship between *INCOME* and *TOURISM*, and also a positive relationship between *EXCHANGE* and *TOURISM* (assuming demand for tourism for Mauritius is modelled and that exchange rate is measured as US$1 = x MUR; higher values of x would imply appreciating USD and this will be at the benefit of the tourist; implying a positive relationship). *COST* is expected to negatively affect tourism.

Figure 28.1 provides a general regression specification (including the dependent and independent variables, the coefficients, and the residual terms) and summarizes the above discussion (on modelling tourism arrivals) in a multivariate linear regression equation, with income, exchange, and cost being the explanatory variables.

Building a regression model is thus an iterative process that involves finding effective independent variables to explain the process the researcher is trying to model and understand (or trying to understand the dependent variable). The model is subsequently run,

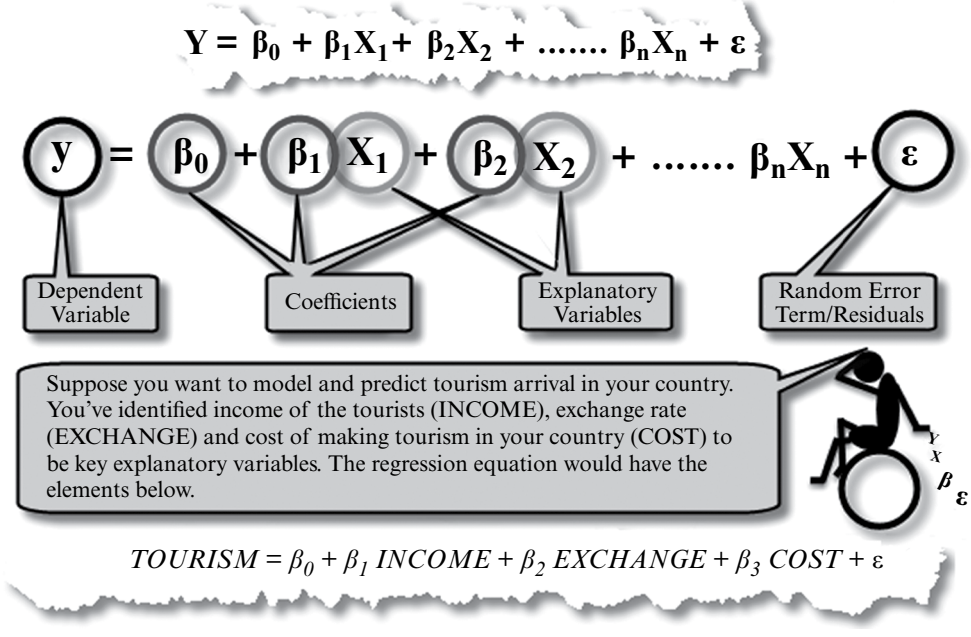

Figure 28.1 A simple multivariate model of tourism demand

using the regression tool, to determine which of these independent variables are effective predictors and to what extent.

Once the model is constructed and the data are collected, the regression is performed using regression statistical software (such as SPSS, STATA, or EViews, amongst others). The regression will generate estimates for both the intercept/constant term β_0 and the respective correlation coefficients, the β_s. Under certain assumptions, these coefficients have the characteristics of unbiasedness, consistency, and efficiency. However, before the data are fed to the model, there is a crucial test to be performed in relation to the data series collected, and this related to that of a unit root test (further discussed below), also known as a stationary test.

EVALUATING THE OVERALL PERFORMANCE OF THE MODEL AND COMMON REGRESSION TECHNIQUES

Some Crucial Statistics

The researcher hopes that the regression model will explain the variation in the dependent variable fairly accurately. If it does, it is said that the model fits the data well. Evaluating the overall fit of the model also helps to compare models that differ with the data set, composition and number of independent variables, and so on.

There are three primary statistics for evaluating overall fit:

1. R^2. The coefficient of determination: multiple R^2 and adjusted R^2 are both statistics derived from the regression equation to quantify the overall model performance. The R^2 ranges from 0 to 1 and can be interpreted as the percentage of the variance in the dependent variable that is explained by the independent variables. If the model fits the observed dependent variable values perfectly, R^2 is 1.0 (which is quite unlikely to happen as no perfect model exists). More likely, lower R^2 values are obtained. A value of 0.65, for instance, would imply that the model explains 65 percent of the variation in the dependent variable.
2. The adjusted R^2 value is the same as the R^2, except that it takes into account the number of independent variables (IVs) in the model.
3. F-statistics. The F-statistic allows the researcher to determine whether the whole model is statistically significant, that is, if the model is overall a good fit.

Hypothesis Testing

Because most data consist of samples from the population, researchers are worried whether the ßs (coefficients) actually matter when explaining variation in the dependent variable. The null hypothesis states that X is not associated with Y, therefore the ß (coefficient/association) is equal to 0; the alternative hypothesis states that X is associated with Y, therefore the ß is not equal to 0. To test this hypothesis, a t-statistic is computed. Technically the t-statistic is equal to the ß divided by the standard error (s.e.) of ß (s.e., a measure of the dispersion of the ß; t = ß/s.e). A (very) rough guide to testing hypotheses might be: "t-statistics above 2 are good."

Most regression methods perform a statistical test to compute a probability, called a p-value, for the coefficients associated with each independent variable, thus easing the interpretation of the hypothesis test (the regression software already report the exact probability vales of the statistics, thus relieving the researcher of the necessity to cross-check in statistical tables). Small p-values reflect small probabilities, and suggest that the coefficient is indeed important to the model with a value that is significantly different from 0 (the coefficient is not 0. A coefficient with a p value of 0.01, for example, is statistically significant at the 99 percent confidence level (the associated variable is an effective predictor). Variables with coefficients near 0 do not help predict or model the dependent variable; they are almost always removed from the regression equation, unless there are strong theoretical reasons to keep them or they are the variables of interest in a study.

Some Problems that Require Attention while Performing a Regression

Specification Bias

Readers are encouraged to read about the problems associated with: (1) omitted variable; (2) irrelevant variable; and (3) functional form in more detail.

Violation of Assumptions

There are several assumptions which must be met for the estimates to be unbiased and have minimum variance. Most commonly violated assumptions that need attention are: multicollinearity, heteroskedasticity, endogeneity bias, and autocorrelation. Readers should be accustomed to the testing of these assumptions, and possible remedies, while engaged in regression analysis. Refer to Gujarati (2004).

Extensions of the Model

Model with dummy

It is worthwhile to note that sometimes some explanatory variables cannot be quantified or proxied, for instance *financial/economic crisis* or *common language/border*, which might be important elements in tourism demand. In regression such qualitative factor can still be taken on board and included in a regression specification through the so-called dummy variables. A dummy variable takes binary values of 1 and 0. For instance, the researcher will give a value of 1 for the years that underwent the crisis (or for countries which have a common language or border) and 0 otherwise. A negative (and significant) dummy coefficient is usually expected for the link between *financial crisis* and tourism.

Limited and qualitative dependent variables: the linear probability model

There are many research studies in tourism for which the dependent variable is qualitative. Researchers often want to predict whether something will happen or not, such as, for example: Will a tourist visit or not? (or: Will the tourist recommend the destination or not?), and this is usually expressed as Yes/No. Linear probability models can be used in such a case and it is a type of regression analysis where the dependent variable is dichotomous and coded 0 or 1.

Endogeneity and simultaneous/VAR models

When modeling tourism, researchers may face issues related to endogeneity, that is, variables in a model may affect each other. For example, in the modeling of the tourism development on economic growth of a country, the number of hotels may be a determinant of tourism but it may also be true that more hotels may be set up due to increasing tourists. So there may be a bi-causal relationship. As such, the income level of the destination country (which is also a measure of the level of development of the destination country) may be a determinant of tourism development (as tourists may be affect by the level of development of a country) and, interestingly, the development level of a country (usually measured by its level of income) has often been argued to be determined by the number tourists visiting the country, bringing positive economic effects. To account for such bi-causal and endogeneous effects, a vector autogressive model (VAR) may be used. The VAR considers several endogenous variables together and resembles a series of equations where each determinant comes as the explained variable in a system, and the system is solved simultaneously. The VAR has proven to be especially useful in describing the dynamic behavior of economic time series, and for forecasting.

Common Estimation Techniques

These include ordinary least squares (OLS), generalized least squares (GLS), maximum likelihood estimation (MLE), instrumental variables regression (IV), generalized methods of moments (GMM), principal component regression (PCR). Interested readers are encouraged to refer to Stock and Watson (2003) and Gujarati (2004) for in-depth treatment of these estimation techniques, and on the other regression models discussed above.

TIME SERIES, CROSS-SECTION AND PANEL REGRESSIONS

Typically there are three types of data sets and analysis which may be useful in tourism studies.

Time Series

Time series analysis is often used for country case studies over a long period of time, for instance in analyzing the determinants of tourist arrival (or international demand for tourism) for the period 1970–2015 for the case of Mauritius. Here the country (also known as the section) is fixed and does not vary, whereas the time dimension varies. Time series data in the above context may be available from the Central Statistics Office of Mauritius or the World Tourism Organization, among others.

Cross-Section

This type of data is usually observed over geographic or demographic groups. For example, the researcher can observe and analyse data on tourism arrival for 50 United States of America states. This would result in 50 observations for each variable in a model (that is, the dependent and independent variables) and over a particular year (or a particular

period of time averaged). Thus, such studies have a unique time dimension (for example, 2010, or average 2010–2015) and a varying level of sections (in the above case, states within a country). Note that the researcher could have easily used a sample of countries to determine the factors affecting tourism, as well as a particular time period. Moreover, cross-section may also often be used to complement survey analysis, where respondents are taken as sections over a fixed time dimension (the date or month of interview, for instance). For instance, following information obtained from a survey of 1500 tourists at the airport while flying away, depending on information captured, one can easily determine the link and magnitude of a number of determining factors that affect the decisions of tourists for that period of time, thus establishing more formal statistical links.

A regression which uses cross-section data sets is called a cross-sectional regression. Cross-sectional regressions usually suffer from the problem of heteroskedasticity. Moreover, the estimates they generate are true for a moment in time, and therefore there is always the lingering question of whether they can adequately represent the unchanging structure that is being researched.

Panel Data

This type of analysis combines the first two types. In this case the researcher has a cross-section, but the cross-section is observed over time. If the same people or states or counties, sampled in the cross-section, are then resampled at a different time, this is referred to as a longitudinal data set, which is a very valuable type of panel data set. For example, if one analyzes tourism arrivals for a sample of 30 countries over a period spanning 1990–2015, thus allowing for both the time and section dimensions to vary, one is then dealing with panel data regression. The challenge remains the availability and collection of data, and also the organization of the panel data sets which surely require much effort. For instance, one has to collect 26 years data for each of the 30 countries for each variable analysed. Thus, in a simple univariate equation in the form of $TR = f(INCOME)$, one need to collect 780 observations (26 years x 30 countries) for both *TR* and *INCOME*. It is noteworthy that the above is a very simple model and that in a more rigorous study the investigator has to make use of multivariate economic models, thus involving much more data collection. In what follows, a more in-depth discussion on time series, cross-sectional, and panel data analysis is provided.

TIME SERIES ANALYSIS

Technically speaking, a time series is a sequence of data points, measured typically at successive time instants spaced at uniform time intervals. Time series analysis comprises of methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series studies can be performed for tourism-related studies for a country, for example in analysing the demand for tourism, investigating the economic impact of tourism development, or studying the impact of air access policies on tourism, amongst others. A time series study will analyze (and quantify), for instance, the effect of tourism development on economic growth for the case of Mauritius over the period 1970–2015.

Stationarity Test

A key idea and a common assumption in many time series techniques is that the data series should be stationary. Roughly speaking, a time series is stationary if its behavior does not change over time. This means, for example, that the values always tend to vary about the same level and that their variability is constant over time. Obviously, not all time series that are encountered are stationary. Indeed, non-stationary series tend to be the rule rather than the exception. If the time series is not stationary, one can often transform it to a stationary form by differencing the series or transforming the series, for example, taking a ratio (tourism marketing expenditure / total government budget), amongst others.[2]

Some tests of stationarity:

- augmented Dickey–Fuller (ADF) test.
- Phillips–Perron (PP) unit root test.
- KPSS unit root test.
- Ng and Perron test.

Refer to Stock and Watson (2003) and Gujarati (2004) for more technical details with respect to the above tests.

It is important to note that when one non-stationary data series is regressed against one or more non-stationary data series, this may result in a spurious or nonsensical relationship. The best way to guard against spurious regressions is to check for cointegration of the variables used in time series modeling. Cointegration is the existence of a long-run equilibrium relationship among time series variables. In non-technical language, it is a property of two or more variables moving together through time, and despite following their own individual trends, these will not drift too far apart since they are linked together in some sense.

Tests for cointegration:

- cointegrating regression Durbin–Watson (CRDW) test.
- augmented Engle–Granger (AEG) test.
- Johansen multivariate cointegration tests, or the Johansen method.
- autoregressive distributed lag (ARDL) cointegration test.

Refer to Stock and Watson (2003) for more technical details with respect to the above tests.

A Simple Time Series Analysis Illustration

Assume a research case where one is assessing the determinants of tourism demand based on a case study for Mauritius for the period 1970–2015, using annual data. Recall equation (28.3) and the related variables explanations. The equation now additionally includes a dummy variable for the financial and economic crisis of 2008–2011 (as such a crisis is suspected to affect the propensity for tourism, as income went down and it also brought uncertainties); it takes the value of 1 for 2008 till 2011 (indicating a financial crisis) and 0 otherwise:

$$TOURISM = f(INCOME, EXCHANGE, COST, CRISIS) \quad (28.5)$$

As discussed previously, theoretically the researcher expects a positive relationship between *INCOME* and *TOURISM*, and the same is expected between *EXCHANGE* and *TOURISM*. However, *COST* and *CRISIS* are expected to negatively affect tourism. The researcher should be cautious, as the above multivariate model may not be the optimally specified one and other interesting explanatory variables may still be required. For illustration, the above relatively simple demand function for international tourism is assumed. Thus the regression model is written as below:

$$TOURISM_t = \beta_0 + \beta_1 INCOME_t + \beta_2 EXCHANGE_t + \beta_3 COST_t + \beta_4 CRISIS_t + \varepsilon_t \quad (28.6)$$

where *t* represents the year.

Data need to be gathered for all the variables over the whole period of study and may be presented as per Table 28.1 (probably in Excel, which can easily be exported to various statistical software). Assume all the data are obtained from the Central Statistical Office (Statistics Mauritius).

Feeding the 46 observations (1970 to 2015) and Table 28.1 into regression software will generate the estimated coefficients (hypothetical) shown in Table 28.2, and other crucial selected information (assuming that all the data series are stationary at level and that the OLS estimation technique is used).

The results obtained after performing the regression are quite interesting. As a matter of interpretation the coefficient of the constant term (15.17) would imply that the destination depicts a positive image, given a reported positive and significant sign. *INCOME* as

Table 28.1 Organization of time series data, 1970-2015, Mauritius

Year	*TOURIST* (000)	*INCOME* (US$) 000	*EXCHANGE* US$1= x Rs	*COST* (USD)	*CRISIS*
1970	27	0.5	7	434	0
1971	76	0.6	8	500	0
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
2007	900	13.5	29	1345	0
2008	900	14	27.5	1354	1
2009	921	15.5	28	1397	1
2010	932	15	28	1450	1
2011	942	15	29	1500	1
2012	950	15.5	30	1550	0
2013	980	15.5	31	1600	0
2014	1010	16	33	1650	0
2015	1100	16	35	1700	0

Note: Figures and data are hypothetical.

Table 28.2 OLS results

Regressor	Coefficient	*t*-ratio
β_0	15.17	3.07
INCOME	158.56	2.23
EXCHANGE	43	0.71
COST	−3.2	−2.94
CRISIS	−50	−2.33
R^2	0.75	

Note: Dependent variable: *TOURISM* (number of tourist arrivals, 000s); F statistics = 6.45.

well has a positive and significant (large t-value) value of 158.56 and this translates to the fact that an increase of $100 in the average income of the tourist resulted in 15 856 additional tourists visiting the country on average. As such, the negative and significant coefficient of *COST* (-4.2) would mean that each US$1 increase in the cost related to undergo tourism in the Mauritius has reduced the number of tourists by 3200. The financial crisis has had a negative impact, as witnessed by the negative and significant correlation between the variable *CRISIS* and tourism arrivals. However, it is noteworthy that even if the variable *EXCHANGE* has the required positive sign as expected, since it is not significant (given the low t value), it would imply that exchange rate changes appear not to have any impact on tourism, at least for the period and for the country under study.

The R^2 of 0.75 means that the model explains 75 percent of the phenomenon under study, that is, the variation in tourism arrivals, and this represents a relatively good fit. The high value of the F-statistics confirms that our model is overall significant.

Log Functions

Since in the above estimation, the scale of measurement of each respective variable is not the same (for instance, *TOURISM* measured in thousand tourists, and *COST* in USD), a direct comparison of estimated coefficients is not meaningful. One way to deal with this and to allow coefficients to be readily and directly comparable is to run the above equation in a log form (that is, all variables are transformed in logarithmic value). In this case a double log function is specified as follows:

$$\ln TOURISM_t = \beta_0 + \beta_1 \ln INCOME_t + \beta_2 \ln EXCHANGE_t + \beta_3 \ln COST_t + \beta_4 \ln CRISIS_t + \varepsilon_t$$

The ultimate aim of running the log function is one of easing interpretation (it will now be interpreted in terms of percentage change), and to allow direct comparison of explanatory variables, which may be crucial for policy implications. Thus if a coefficient of $\beta_1 = 0.8$ is obtained in the above specification, it would imply that a 1 percent increase in *INCOME* will result in a 0.8 percent increase in tourist arrivals. As such, one could generate values of other β_s. The researcher should be very cautious while performing the above regression, as there are a number of tests, both before (test of

stationarity and possibly co integration) and after (in particular autocorrelation, serial correlation and heteroskedasticity tests) the regression that need to be undertaken.

Cross-Section Analysis

As explained earlier, cross-section involves studying varying sections (for instance, countries or states) over a specific period of time (or over an average time period). A simple illustration: assume a researcher wants to analyse the demand for tourism arrivals for a sample of 100 countries for the year 2014 and that the analysis is based on the same equation specified earlier (excluding the financial crisis dummy, for simplicity), that is:

$$TOURISM_i = \beta_0 + \beta_1 INCOME_i + \beta_2 EXCHANGE_i + \beta_3 COST_i + \varepsilon_i \qquad (28.4)$$

where *i* represents the respective country.

Thus instead of varying the time dimension, the sections are varied and researcher can apply the same estimation technique (that is, mainly OLS estimation methods, although GLS is also popular) and similar diagnosis tests (as discussed before) to the above cross-sectional data in Table 28.3 to derive meaningful results.

Illustrative Example

Suppose after regressing equation (28.4) above using cross-section data in Table 28.3, the following coefficients of estimations and related statistics are obtained. Note that the specification is in log terms:

$$\ln TOURISM = 5.34^{*} + 0.87 \ln INCOME^{*} + 0.35 \ln EXCHANGE^{*} - 0.65 \ln COST^{**}$$

where *, **, and *** denote 10 percent, 5 percent, and 1 percent level of significance, respectively.

Table 28.3 Organization of cross-section data, 2014, 100 countries

Year	*TOURIST* (000)	*INCOME* (USD) 000	*EXCHANGE* US$1 = x Rs	*COST* (USD)
Country 1	256	0.5	3	423
Country 2	654	0.6	1	575
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Country 97	345	13.5	6	435
Country 98	34	14	27	565
Country 99	755	15.5	23	397
Country 100	346	15	2	457

Note: The data and the resulting estimations are of a hypothetical nature.

Rsqr = 0.65; F-Stats = 6.75.

From the above regression output, it can be concluded that if income of the tourists rises by 1 percent, this is likely to increase tourist arrivals by 0.87 percent. As such, if the local exchange rate depreciates by 1 percent (increasing the purchasing power of the tourist), tourism is projected to rise by 0.35 percent. And finally, the significant negative coefficient of *COST* implies that an increase in cost of living of the destination country will reduce tourism by 0.65 percent. The R^2 value of 0.65 indicates that 65 percent of the variable under study – that is, *TOURISM* – is explained by the variables in the model, while the significance of the F-test denotes the overall significance of the model.

Pooled Time Series and Panel Data

Pooled time series analysis and panel data combine time series for several cross-sections; that is, they are characterized by having repeated observations (most frequently, years) on fixed units (countries, states, or firms). Thus, extending the analysis from the previous section, one could have a data set which is based on 100 countries, over the time period 2008–2014 (that is, seven years, resulting in 7 x 100 = 700 observations). Having both time and section varying dimensions allow the researcher to benefit from a larger number of observations and to make better inferences over a period of time (instead of having results based on only a year, for example 2014). There is also the possibility to capture not only the variation of what emerges through time or space, but the variation of these two dimensions simultaneously. However, the challenge usually arises in the data availability and collection. There are two types of techniques to estimate varying time and section dimensions: pooled time series and panel data frameworks.

Organizing Pooled Time Series and Panel Data

As discussed briefly earlier, if all the data are pooled together and if no distinction is made between cross-sections, one can run a regression with all the data using ordinary least squares (OLS). This is called a pooled OLS regression. This type of regression is the easiest to run, but is also subject to many types of errors (refer to Baltagi, 2004). Pooled time series is often used as a rough and ready means of analyzing the data. It is a simple and quick way of estimation which is often used to derive preliminary results and analysis. It is based mainly on OLS estimation.

A more sophisticated framework to analyse similar data sets, and which is more popular, is the use of panel data. Panel data are often referred to as longitudinal data. Panels may be balanced if there is an observation for every unit of observation for every time period, and are unbalanced if some observations are missing (not all countries may have the same number of years of observations). The following discussion applies equally to both types.

In estimating panel data coefficients, one has to select the appropriate model, that is, between what are known as fixed effects and random effects models.[3] As compared to the pool time series model, fixed effects and random effects models work to remove omitted variable bias by measuring change within a group. By measuring within a group (across time) one can control for a number of potential omitted variables unique to the group (this is ignored in pooled estimates, and thus represent probably the biggest caveat).

Baltagi (2004) argued that panel data has the following advantages over pooled data: (1) it accounts for heterogeneity across individual units, an element which is ignored by pooled time series data analysis; (2) it deals with time-invariant omitted variables, as one can find in pooled data; and (3) it is less likely to have problems with autocorrelation and multicollinearity, as time series data do.

Fixed effects assumes that the individual specific effect is correlated to the independent variable, while random effects assumes that the individual specific effects are uncorrelated with the independent variables. The fixed effects model thus assumes that individual heterogeneity is captured by the intercept term. The random effects model assumes in some sense that the individual effects are captured by the intercept and a random component.

Fixed or Random Model?

The generally accepted way of choosing between fixed and random effects is to run a Hausman test. Statistically, fixed effects are always a reasonable thing to do with panel data (they always give consistent results), but they may not be the most efficient model to run. Random effects will give better t ratios and p-values as they are a more efficient estimator, so one should run random effects if it is statistically justifiable to do so. The Hausman test checks a more efficient model against a less efficient but consistent model, to make sure that the more efficient model also gives consistent results.

The problem is that the Hausman test rejects the random effects model very often and does not work very well in small samples (Baum, 2006). The Hausman test tests the null hypothesis that the coefficients estimated by the efficient random effects estimator are the same as the ones estimated by the consistent fixed effects estimator. If they are (insignificant P-value), then it is safe to use random effects. If the researcher gets a significant P-value, however, they should use fixed effects.

A word of caution: researchers should look at inherent problems related to panel-level heteroskedasticity and autocorrelation. Read more about this in Stock and Watson (2003) and Wooldridge (2002).

Organizing panel data

Recall the same model specification as above:

$$TOURISM_{it} = \beta_0 + \beta_1 INCOME_{it} + \beta_2 EXCHANGE_{it} + \beta_3 COST_{it} + \varepsilon_{it} \qquad (28.4)$$

where *i* represents the respective country and *t* the time dimensions.

Assume that estimating the panel set given in Table 28.4 results in the following coefficients and accompanying statistics. Assume the Hausman test: $Chi^2 = 3.8$, with a probability value of 0.06 (this need to be checked from statistical tables, but is provided by the software). Since this value is greater than 5 percent, we select the random effect model. The estimates using the random effect model are given below:

$$\ln TOURISM = 5.2^* + 0.75 \ln INCOME^{**} + 0.13 \ln EXCHANGE - 0.56 \ln COST^{***}$$

Table 28.4 Organization of panel data, 2008-2014, 100 countries)

Country	Year	*TOURIST* (000)	*INCOME* (USD) 000	*EXCHANGE* US$1 = x Rs	*COST* (USD)
Country 1	2008	256	0.5	31	343
Country 1	2009	277	0.6	33	349
Country 1	2010	323	0.7	34	412
Country 1	2011	325	0.7	34	425
Country 1	2012	345	0.8	35	430
Country 1	2013	350	0.8	35	435
Country 1	2014	360	0.9	36	440
Country 2	2008	867	1	7	453
Country 2	2009	896	1.1	8	464
Country 2	2010	967	1.2	8	489
Country 2	2011	975	1.2	8	492
Country 2	2012	980	1.2	9	495
Country 2	2013	990	1.2	9	500
Country 2	2014	1000	1.3	9	510
Country 100	2008	1543	0.76	11	278
Country 100	2009	1656	0.79	13	289
Country 100	2010	1754	0.85	14.5	299
Country 100	2011	1780	0.87	15	301
Country 100	2012	1790	0.9	15	311
Country 100	2013	1800	0.9	15.5	325
Country 100	2014	1845	0.95	16	330

Note: The data and the resulting estimations are of a hypothetical nature.

where *, **, and *** denote 10 percent, 5 percent, and 1 percent level of significance, respectively.
R-sqared = 0.45, F-statistics = 5.54.

The interpretation of the above results is similar to the previous sections. That is, for instance, if income of the tourism rises by 1 percent, this is likely to increase tourist arrivals by 0.75 percent. The R^2 value of 0.55 indicates that 55 percent of the variable under study, that is, *TOURISM*, is explained by the variables in the model, while the significance of the F-test denotes the overall significance of the model.

Fixed and random effect estimates are referred to as static panel estimates. However, most relations are dynamic in nature and this may require more appropriate modelling and techniques. It is very often the case that the dependent variable is also explained by itself. For instance, in tourism demand modelling, last year's tourist may influence this year's tourist following a good experience, word of mouth recommendation, or even by repeating their visit. This call for more advance modelling in panel data. Interested readers are encouraged to read more on dynamic panel framework, instrumental variable (IV), and generalized methods of moments (GMM), and even panel vector autoregressive methodologies.

Refer to the research case below, which discusses a study on the impact of tourism on

economic growth for a sample of countries over a certain period of time. This research case is drawn from the author's empirical paper entitled 'Assessing the Economic Impact of Tourism for the Case of Island Economies' (Seetanah, 2011). The original paper employed a dynamic panel data approach, given that the tourism–growth relationship is of a dynamic nature. The research case below deals only with the model building, data measurement, and sources. Readers are encouraged to read the full paper for more insights.

RESEARCH CASE: MODELLING THE IMPACT OF TOURISM ON ECONOMIC GROWTH

A researcher wants to quantitatively assess the role of tourism on growth for a sample of ten countries over the period 1990–2010. This yields a panel data set (10 countries x 21 years) of 210 observations.

Assume that the following economic model is adopted; the choice of the explanatory variables are based from past research and from theories as well. Note that the number of explanatory variables is limited to four, for simplicity:

$$GDP = f(IVT, OPEN, EDU, TOURIST)$$

The regression equation can be written as follows:

$$GDP_{it} = \beta_0 + \beta_1 IVT_{it} + \beta_2 OPEN_{it} + \beta_3 EDU_{it} + \beta_4 TOURIST_{it} + \varepsilon_{it}$$

where: i represents the countries while t represent the time period;
GDP is a measure of output and thus economic growth;
IVT is an indicator of investment level of the country: investment increases the capital level of the countries in terms of equipment and machines, among others, and is believed to be crucial for increased output; this can be measured by the investment ratio of the country (that is, *IVT*/*GDP*);
OPEN is an indicator of the openness level of a country, as it has been a general consensus that more open economies benefit more economically; the most widely used measure is the ratio of export plus imports on GDP;
EDU is a measure of education and human capital level; a better educated and trained labour force is bound to result in enhanced productivity and more output, and many studies have attempted to measure the level of education and human capital by the literacy rate of the country.

Of interest to this study is the variable *TOURIST*, which is added to the above equation in an attempt to capture the effect of tourism on output. It is expected that more tourism can be economically beneficial to the economy, as there will be more domestic production, more tax receipts, higher consumption level, more employment, and foreign receipts, among others. Such a variable can be captured through the total number of tourist arrivals, while other studies have used the total tourism receipts. Researchers may use both measures alternatively to check the robustness of the results.

Sources of Data

Data availability and databases are crucial while performing regressions. This is especially the case while engaging in panel data and cross-country analysis where large amounts of data need to collected. For the above case, international databases are required, for instance the World Development Indicators of the World Bank (http://data.worldbank.org/data-catalog/world-development-indicators), International Financial Statistics of the International Monetary Fund (http://elibrary-data.imf.org/), and the PENN World Tables (https://pwt.sas.upenn.edu/php_site/pwt71/pwt71_form.php) for *EDU*, *IVT*, and *OPEN*, while tourist data (not limited to tourism arrivals and flows, but including other interesting tourism-related data) are available from the Tourism Satellite Accounts of the United Nations World Tourism Organization (http://statistics.unwto.org/).

On another note, if dealing with time series analysis, the countries' respective Central Statistical Offices and digests of tourism provide the most reliable source of data for tourism variables (including tourism arrivals by country of origin and purpose, and so on, tourism receipts, hotel rooms, employment in the tourism sector, contribution of the tourism industry, marketing expenditure, among others)

Coming back to the case study, once the data are successfully collected for all the countries, and ideally for the whole period, panel data analysis can be employed to estimate the coefficients of each potential factor that could explain growth, and in this case the *TOURISM* variable. Interested readers may wish to refer to an article by the author: Seetanah (2011).

CHAPTER SUMMARY

This chapter has introduced the reader to secondary data analysis. It started by familiarizing the researcher with simple regression specifications and models after analyzing multivariate models. Time series, cross-section, and panel data analysis were subsequently discussed, together with related diagnosis tests. After reading this chapter, readers should able to specify multivariate conceptual and regression models related to tourism studies, collect and organize data sets, and engage in the interpretation of regression outputs.

NOTES

1. The reader may wish to read more about the scatter diagram: refer to Gujarati (2004).
2. If all the series are stationary at first difference, then the researcher may regress the equation in its first difference. However, this would imply deriving short-term estimates and potentially losing long-term properties of the estimates. It is noteworthy that a cointegration test (a test of long-term relationship) may still be performed if all series are stationary at the same level, and this may allow the researcher to derive long-term estimates as well. In the eventuality that some series are stationary in levels and others stationary in first difference, there may still be another approach to test for cointegration and estimate the model, namely the autoregressive distributed lag model (ARDL).
3. Interested readers are encouraged to learn more about these models. Refer to http://dss.princeton.edu/online_help/stats_packages/stata/panel.htm for an in-depth treatment.

REFERENCES

Baltagi, B.H. (2005). *Econometric Analysis of Panel Data*, 3rd edition. Chichester: John Wiley & Sons.
Baum, C.F. (2006). *An Introduction to Modern Econometrics Using Stata*. College Station, TX: Stata Press.
Gujarati, D. (2004). *Basic Econometrics*, 4th edition. New York: McGraw-Hill.
Seetanah, B. (2011). Assessing the Dynamic Economic Impact of Tourism for Island Economies. *Annals of Tourism Research* 38 (1), 291–308.
Stock, J.H. and Watson, M.W. (2003). *Introduction to Econometrics*. New York: Prentice Hall.
Wooldridge, J.M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Further Reading

Chatterjee, S. and Hadi, A.S. (2012). *Regression Analysis by Example*, 5th edition. Hoboken, NJ: John Wiley & Sons.
Chatterjee, S. and Simonoff, J.S. (2013). *Handbook of Regression Analysis*. Hoboken, NJ: Wiley.
Harrell, F. (2001). *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer-Verlag.
Lewis-Beck, C. and Lewis-Beck, M. (2016). *Applied Regression: An Introduction*, 2nd edition. Thousand Oaks, CA: SAGE Publications.