Chapter 4: Methodology

4.1 Introduction

This chapter will look into the methodology and statistical tests used in this study. The stationarity test is discussed in order to test if the time series data is stationary. Cointegration analysis is considered a pre–test to avoid spurious regression. Spurious regression occurs when the time series data is not stationary. Spurious regression will invalidate the t and F tests. The OLS method to estimate the regression models will be discussed briefly in this same chapter. Finally, this chapter will also look into the problem of heteroscedasticity and its detection test to determine the problem. The problem of heteroscedasticity in the regression may distort the validity of OLS estimators, hence render the estimation results unreliable.

4.2 A Stationary Process

This empirical study is based on a time series data analysis. An important assumption is that the time series data is *stationary*. Generally, time series data can be thought of as being generated by a *stochastic* or *random process*. A stochastic process is said to be *stationary* if its mean & variance are constant over time and the value of covariance between two time periods depends only on the distance or lag between the two periods and not on the actual time at which the covariance is computed (Gujarati, 1995).

In regressing a time series variable on another time series variable, an obtained high R^2 does not necessarily mean that there is a meaningful relationship between the two variables. It could be a problem of *spurious regression* arising in the time series. Spurious regression is a "false" regression in the sense that it does not show the true relationship between the two variables. Owing to characteristics exhibited by both the time series, the high R^2 observed may be due to the presence of the trend and not to the true relationship between the two variables.

4.2.1 Spurious Regression

As stated above, a spurious regression is not a true regression. A significant feature of a spurious regression is reflected in the results. The results of the regression will appear to be very good, the R^2 extremely high, the t ratios of independent variables with extremely high values, the coefficients high but the Durbin-Watson d will be low. It was suggested, as a good rule of thumb to suspect that the estimated regression suffers from spurious regression when $R^2 > d$ (Granger & Newbold, 1974). A spurious regression occurs when we regress one non-stationary time series on another non-stationary time series (Gujarati, 1995). In the case of non-stationary time series, the estimated t and F values are no longer reliable.

4.2.2 Stationarity Test

Generally, there are two types of stationarity tests. One is based on the so-called auto-correlation function (ACF) and the other is the unit root test which has become popular recently. The purpose of running these two tests is to test if the time series is stationary and to avoid the occurrence of spurious regression in time series data. This study adopts the unit root test for stationarity in the time series data.

4.2.3 Unit Root Test

An alternative test of stationarity is the unit root test. Two types of unit root tests are Dickey –Fuller (DF) test and Augmented Dickey-Fuller (ADF) test. The DF test applies to a model as follows:

$$\Delta Y = \delta Y_{t-1} + u_t \qquad (1)$$

Under the null hypothesis that $\delta = 1$, the conventionally computed t statistic is known as the τ (tau)statistic, whose critical values have been tabulated by Dickey and Fuller(1979) on the basis of Monte Carlo simulations. That is why this tau test is also known as the **Dickey-Fuller (DF) test.**

In its simplest form, we estimate a regression like (1), divide the estimated δ coefficient by its standard error to compute the Dickey-Fuller τ statistic and refer to the

Dickey-Fuller tables to see if the null hypothesis $\rho = 1$ is rejected. If the computed absolute value of the τ statistic (i.e., $|\tau|$) exceeds the DF or Mackinnon DF absolute critical τ values, then we do not reject the hypothesis that the given time series is stationary (Gujarati, 1995).

The ADF test applies as below:

$$\Delta Y_t = \beta_1 + \beta_2 t + \delta Y_{t-1} + u_t$$

The difference between these two regressions lies in the inclusion of the constant (intercept) and the trend term (t). This study will use the ADF test as a tool of stationary test.

4.2.4 The ADF Test

Consider the following equation:

$$Y_t = \rho Y_{t-1} + u_t \qquad (1)$$

Where ρ is a parameter and u_t is assumed to be white noise error term. Y_t has a unit root when $\rho = 1$. Based on econometric theory, a time series that has a unit root is known as a non-stationary time series and in economic term it is known as a **random walk.**

Therefore, the hypothesis of the stationary series can be evaluated by testing the absolute value of ρ. The null hypothesis becomes H_0: $\rho = 1$. Since the series is explosive when $\rho > 1$, the hypothesis is tested against the one-sided alternative H_1 : $\rho < 1$. The test is carried out by estimating the equation with y_{t-1} subtracted from both sides of the equation:

$$\Delta Y_t = (\rho - 1)Y_{t-1} + u_t$$

$$= \delta Y_{t-1} + u_t \qquad (2)$$

Where $\delta = \rho - 1$ and the null and alternative hypothesis are

$H_0 : \delta = 0$ and $H_1 : \delta < 0$

If δ is = 0, we can rewrite equation 2 as:

$$\Delta Y_t = (Y_t - Y_{t-1}) = U_t \qquad (3)$$

What equation (3) says is that the first differences of a random walk time series are stationary time series because by assumption u_t is purely random. If a time series is differenced once and the differenced series is stationary, the original (random walk) series is integrated of order 1, denoted by I(1). Similarly, if the original series has to be differenced twice (i.e., take first difference of the first difference) before it becomes stationary, the original series is integrated of order 2, or I(2). Put it in a general form, if a time series has to be differenced d times, it is integrated of order d or I(d).

4.3 Cointegration

A time series variable is stationary if its mean value and its variance do not vary systematically over time. *Spurious regression,* which results if time series are not stationary[1], should not arise. If there is a problem of spurious regression, the standard t and F test procedures are not valid. However, if all variables are cointegrated, the regressions result may not be spurious, and the usual t and F tests are valid.

One way to know if a time series is cointegrating is to prove that the residual, u_t (error term) obtained from the regression is stationary or integrated of order zero(I(0)). When the u_t is I(0), the "trends" in the variables cancelled out and they will be on the same wavelength if they are integrated of the same order. Implication that can be drawn upon is if a series Y is I(1) and at the same time another series X is also I(1), they are cointegrated. Put it generally, if Y is I(d) and X is also I(d), where d is the same value, these two series are cointegrated (Gujarati, 1995). As Granger notes, " A test for cointegration can be thought of as a pre-test to avoid spurious regression."

Since most of the economic relationships are best expressed in the level form and not as first differences or second differences, such as government expenditure and tax revenues, money supply and prices etc. Thus, a cointegrating regression makes the regression on the levels of the two variables become meaningful (not spurious). At the same time we do not lose any valuable long term information, which would result if we used their first or higher differences instead. Therefore, provided we check that the residuals from the regressions are I(0) or stationary, the traditional regression methodology (including t & F tests) is applicable to the time series data.

4.4 Cointegration Test

Two methods have been proposed for testing of cointegration. The DF /ADF test on the residual (u_t) estimated from the cointegrating regression (Gujarati, 1995) and the cointegrating regression Durbin-Watson (CRWD) test (Gujarati, 1995). In this study, we use the DF/ADF test on the residuals estimated from the cointegrating regressions.

4.4.1 Stationarity Test on Residuals(Augmented Engle-Granger test)

In order to run a stationary test on residuals, we must first obtain the estimated residuals from an estimated regression. Later, we subject the residuals estimated from this regression to the DF or ADF unit root test.

However, since the estimated u is based on the estimated cointegrating parameter β_i, therefore the DF and ADF critical significance values are not suitable.

[1] This problem arises because if both the time series involved exhibit strong trends (sustained upward or downward movements), the high R^2 observed is due to the presence of the trend, not to a true relationship

Engle and Granger (1987) have calculated these values, so it is known as Engle-Granger(EG) test and augmented Engle-Granger (AEG) test. We obtain the estimated u from the regression and subject it to the DF unit root test. The Engle-Granger 1%, 5% and 10% critical values are respectively, -2.5899, -1.9439, -1.6177. Therefore if the computed t statistic is less than any of these value, the hypothesis of a unit root is rejected. If the hypothesis of a residual unit root is rejected, it is proved that the regression is cointegrated. Therefore although the variables may individually exhibit random walks, there seems to be a stable long-run relationship between the variables.

4.5 OLS

This paper will use the OLS (ordinary least squares) method to estimate the regression models. The method of OLS is attributed to Carl Friedrich Gauss, a German Mathematician. Based on the assumptions of **CLRM** (Classical Linear Regression Models), the method of OLS consists of some very attractive statistical properties that have made it one of the powerful and popular methods of regression analysis. The attractive statistical properties is known as the famous **Gauss-Markov** theorem, where actually is the combination of two approaches: the least-squares approach of Gauss and the minimum-variance approach of Markov. The main feature in this theorem is that it advocates the **best linear unbiasedness property** of an estimator. An OLS estimator $\hat{\beta}_2$ is said to be best linear unbiased estimator **(BLUE)** of β_2 if the following conditions hold:

between the two.

1. It is linear function of a random variable.

2. It is unbiased, that is, its average or expected value, $E(\hat{\beta}_2)$, is equal to the true value . β_2.

3. It has minimum variance in the class of all such linear unbiased estimators; an unbiased estimator with the least variance is known as an efficient estimator.

4.6 The Coefficient of Determination, r^2

The regressions will be run using E-views(Micro-TSP) statistical package. The coefficient of determination r^2 is used to measure the goodness of fit of the fitted regression line to a set of data; this indicates how "well" the sample regression line fits the data[2]. The r^2 tells what proportion of the variation in the dependent variable, y is explained by the explanatory variables or regressors, Xs. This r^2 lies between 0 and 1; the closer it is to 1, the better is the fit.

4.7 Testing the Significance of Regression Coefficients: The t-test

Test of significance approach developed by R.A. Fisher and jointly by Neyman and Pearson (1959)[3]. A test of significance is a procedure by which sample results are used to verify the truth or falsity of a null hypothesis. The main idea behind

[2] If all the observations fall on the regression line, we call this a "perfect" fit, but this is rarely the case. What we hope for is that the residuals around the regression line is as small as possible.

[3] Details may be found in E.L Lehman. 1959. *Testing Statistical Hypotheses*. New York: John Wiley & Sons.

tests of significance is that of a test statistic (estimator) and the sampling distribution of such a statistic under the null hypothesis. The decision to accept or reject H_0 is made on the basis of the value of the test statistic obtained from the data[4].

Since we are using the t distribution, the testing procedure is called a t **test**. According to significance test theories, a statistic is said to be statistically significant if the value of the test statistic lies in the critical region. In this case, the null hypothesis is rejected and conversely, a test is said to be statistically insignificant if the value of the test statistic lies in the acceptance region.

4.8 Problem of Estimation

OLS method assumes the regressions are CLRM models. One of the important assumption of the CLRM in the regression model is homoscedastic; that is, for any value of X, they all have the same variance. Put it clearly, the variance of each disturbance term u_i, conditional on the chosen values of the explanatory variables, Xs is some constant number equal to σ^2 :$E(u_i) = \sigma^2$. This is the assumption of homoscesdasticity, or equal (homo) spread (scedasticity), that is equal variance. It may distort the validity of OLS estimators, $\hat{\beta}i$, if this assumption is not fulfilled; that is, the problem of heteroscedasticity exists.

[4] For further explanation, see Gujarati,D. 1995. *Basic Econometrics*. Singapore: McGraw-Hill Book Co.

4.8.1 Heteroscedasticity

Problem of heteroscedasticity occurs when the variances of u_i vary. The OLS estimator,$\hat{\beta}_i$ no longer the " best" or "efficient" estimator or the OLS estimator no longer **BLUE** although it is still linear unbiased. That is we will find that the expectation of all the variances of u_i^2, $E(u_i^2) = \sigma_i^2$. Notice that the subscript of σ^2, which tells that the conditional variances of u_i (= conditional variances of Y_i) are no longer constant. Since homoscedasticity is an important feature in the OLS estimation, therefore the regressions model in this study will be tested against the problem of heteroscedasticity.

4.8.2 The Reasons for the Occurrence of Heteroscedasticity

Generally, there are several reasons why the problem of heteroscedasticity occurs or the variances of u_i may be variable. This may be because the variances of the residual following the error-learning models. That is, as people learn, their errors of behaviour become smaller over time. In this case, σ_i^2 is expected to decrease. As incomes grow, peoples have more discretionary income, this means that they have more choices about the disposition of their income. In such a case, the σ_i^2 is likely to increase with income. As data collecting techniques improve, the variances of σ_i^2 is likely to decrease.

Heteroscedasticity can also arises as a result of the presence of outliers. Outlier is refers to an outlying observation, is an observation that is outstanding among the observations in the sample. The inclusion or exclusion of such an observation especially if the sample size is small can substantially alter the results of regression analysis (Gujarati, 1995). The problem of heteroscedasticity may arise if the model is incorrectly specified. This may be due to the fact that some important variables are omitted from the model.

4.8.3 Consequences of Using OLS in the Presence of Heteroscedasticity

There are two cases of using OLS in the presence of heteroscedasticity. One case is that OLS estimation allowing for heteroscedasticity. This is the situation where β_2 is used, take the variance formula as below:

$$\mathrm{Var}(\hat{\beta}_2) = \frac{\sum x_i^2 \sigma_i^2}{(\sum x_i^2)^2} \quad ; \qquad (1)$$

which takes into account heteroscedasticity explicitly. Using this variance, and assuming σ_i^2 are known, we found that var $(\hat{\beta}_2^*) \leq$ var (β_2). Since that $\hat{\beta}_2^*$ is the best unbiased linear estimator with the smallest variance for β_2, therefore we cannot establish confidence intervals and test hypotheses with the usual t & F test based on the latter in which that confidence intervals will be unnecessary larger.

As a result, the t and F tests are likely to give inaccurate results in that t value is smaller than what is appropriate and contribute to the appearance of statistically insignificant coefficient. The var $(\hat{\beta}_2)$ is overly large and this may cause the confidence intervals to be less than accurate.

Another case is OLS estimation disregarding heteroscesdasticity. This is the case when we are not only use β_2 but also continue to use the usual (homoscedasticity) variance formula :

$$\mathrm{Var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2} \quad ; \qquad (2)$$

even if heteroscedasticity is present or suspected. Given that var$(\hat{\beta}_2)$ in (2) is a biased estimator of var$(\hat{\beta}_2)$ in (1), it overestimates or underestimates the latter, and in general we are not sure whether the bias is positive (overestimation) or negative (underestimation), it depends on the nature of the relationship between σ_i^2 and the value taken by the explanatory variable x. The bias arises from the fact that $\hat{\sigma}^2$, the conventional estimator of σ^2 is no longer an unbiased estimator of the latter when heteroscedasticity is present.

In short, if we persist in using the usual testing procedures despite heteroscedasticity, whatever conclusion or inferences we make may be very misleading (Gujarati, 1995). The results based on these testing procedures are also unreliable.

4.8.4 Detection of Heteroscedasticity

There are several methods to detect the problem of heteroscedasticity in a regression model. This study used the White's general heteroscedasticity test to detect the problem. The Eviews statistical package provides this test. As an example of how to detect the problem using White's test approach, consider the following three variable regression model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ui \qquad (1)$$

First step is to obtain the residuals, u_i, then we run the following (auxiliary) regression:

$$U_i^2 = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X^2_{2i} + \alpha_5 X^2_{3i} + \alpha_6 X_{2i} X_{3i} + v_i \quad (2)$$

That is, the squared residuals from the original regression are regressed on the original X variables or regressors, their squared values, and the cross product (s) of the regressors. Higher powers of regressors can also be introduced. There is a constant term in this equation. Obtain the R^2 from this auxiliary regression. Under the null hypothesis that there is no heteroscedasticity, it can be shown that sample size (n) times the R^2 obtained from the auxiliary regression asymptotically follows the chi-square distribution with df equal to the number of regressors (excluding the constant term) in the auxiliary regression. That is, In our example, there are 5df since there are 5 regressors in the auxiliary regression. If the chi-square value obtained in (3)exceeds the critical chi-square value at the chosen level of significance, the conclusion is that there is heteroscedasticity. If it does not exceed the critical chi-square value, there is no heteroscedasticity, which is to say that in the auxiliary regression (2), $\alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 0$ (Gujarati, 1995).

4.9 Sources of data

The data for FDI in approved industrial projects obtained from MIDA Manufacturing Annual Reports (1978-1998). The data for annual manufacturing output and employment are obtained from the time series data compiled by Department of Statistic. The annual export, import, private investment and public investment and price deflator are obtained from the Malaysia Institute of Economic Research (MIER) and the data for annual manufacturing export, import and GDP compiled from various year of Economic Report (1978-2000).

4.10 Limitations of data

The FDI data compiled for this study holds limitations in some area. Firstly, the data for FDI pertains to that in approved industrial projects and not the FDI in projects which are in operation. Therefore, discrepancies between the results derived from this study and the real situation may emerge. Besides, the FDI data compiled were not able to capture the non-equity arrangements of MNCs, such as franchising, licensing, long-term subcontracting and other non-equity joint ventures. Nonetheless, the data should be fairly representative of the FDI structure in Malaysia.

4.11 Scope of the study

This study assess the impact of FDI on manufacturing sector and not the whole economy. The impact of FDI on export, import and employment generation is also limited to the manufacturing sector.

4.12 Conclusion

This study is a time series data analysis. Therefore the methodology used in this study is basically a method to estimate the regression models as well as various statistical tests to avoid any distortions in the estimation results. A detection test of heteroscedasticity is applied in the study in order to get a more reliable estimation results. A discussion of sources of data, limitations of data and the scope of the study is for the purpose of running the analysis within the specified cope.