CHAPTER 2

DATA AND METHODOLOGY

2.1 Introduction

This chapter describes data used in the study, data source, trading sessions and also subdivision of data group for the events under study. Apart from this, methodology on stock volatility, trading behaviour, speed of price adjustment, contemporaneous relationship, data stationarity and Granger-casuality test will also be discussed in the later section of this chapter.

2.2 Data

In this study, the data used are the intraday 15-minute KLSE CI, volume and value. Number of shares outstanding for KLSE CI are on daily basis. The study covers the period 9 March 1992 - 15 April 1993.

The CI is preferred to other indices such as the KLSE Main board All-Shares (EMAS) Index. The CI is considered to reflect market performance closely since it represents economic contribution of various sectors in the Malaysian economy. For a stock to be selected as a component stock, it has to fulfill certain criteria, one of which is the trading volume as set out by the Index Task Force Committee of the KLSE (see The Kuala Lumpur Stock Exchange, 1986). The problem of thin trading or non-synchronous trading of the component stocks has, therefore, been reduced as compared to other indices. There are 100 component stocks as at end 1995 (see appendix), most of which are actively traded.

Information pertaining to the stock market is obtained mainly from the KLSE. The economic data are sourced from Bank Negara publications. Publication of the Department of Statistics, newspapers and magazines are also used to provide supplementary information.

On 22 July 1992, the KLSE's trading was increased from 4 hours to 5 ½ hours. Prior to this, trading in the morning session was from 10.00 am to 12.30 pm and in the afternoon session from 2.30 pm to 4.00 pm. Since then, the KLSE opens half an hour earlier for the morning session while the afternoon session extends to 5.00 pm. There are five trading days per week, that is, from Monday through Friday.

This study extends Chang et al's (1993a) period to cover both extended trading hours and full automation. The subperiods for the study are as follows:

(a) Superiod 1: 9 March to 21 July 1992

90 trading days after Chang et al's (1993a) period but before extended trading hours.

(b) Subperiod 2: 22 July 1992 to 27 November 1992

86 trading days after extended hours but before full automation.

(c) Subperiod 3: 30 November 1992 to 15 April 1993

84 trading days after full automation.

2.3 Methodology

Return is calculated using the following formula:

$$R_t = \log(I_t/I_{t-1})$$

where I_t is the index at time t and t-1 is the index observed 15 minutes earlier.

The average return (AR_t) is derived based on the return above but across the respective times and periods.

$$AR_t = (1/n)\sum_{i=1}^{n} R_t$$

where n is the number of observations of a particular intraday session, weekday or period.

16 return series are computed for each trading day from 1015 hour to 1600 hour for the period before extended trading hours and 20 series for the period after the extended trading hours from 0945 hour to 1700 hour. Other than the 15-minutes intraday return, overnight non trading return (using morning open and afternoon closing on the previous day), lunch break return (between morning closing and afternoon open), morning return (between morning open and morning closing), afternoon return (between afternoon open and afternoon closing) and daily return (between morning open and afternoon closing) are also calculated.

2.3.1 Volatility

Volatility is measured by standard deviation and variance. The standard deviation (s) is defined as follow:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (R_i - \overline{R}_i)^2}{n-1}}$$

where $\overline{R}$ is the arithmetic mean of return. The skewness of the return is measured by:

$$a_3 = \frac{(1/n) \sum_{i=1}^{n} (R_i - \overline{R})^3}{s^3}$$

where s^3 is the cube of the standard deviation of the return. The skewness of a symmetrical distribution, such as normal distribution, is zero.

The degree of peakedness or kurtosis is given by the formula below:

$$a_4 = \frac{(1/n) \sum_{i=1}^{n} (R_i - \overline{R})^4}{s^4}$$

The kurtosis of a normal distribution is three. If the kurtosis is greater than three, the distribution is called leptokurtic. On the other hand, if the kurtosis is less than three, the distribution is platykurtic.

2.3.2 Trading Behaviour

In the analysis of trading behaviour, researchers do not solely depend on the absolute volume and value. Instead, Jain and Joh (1988) and Cheung et al (1993) defined the volume ratio as the number of shares traded (Volume) divided by the number of shares outstanding (NOS) as given below:

$$\text{Volume ratio (\%)} = \frac{\text{Volume}}{\text{NOS}} * 100$$

The trading volume and number of shares outstanding refer to the total transactions and shares floated, respectively, of all the component stocks in the CI. Volume ratio is calculated on the intraday and weekday basis for the three subperiods, from which the mean for each intraday interval and weekday is determined.

To test the difference between the means, the following F-statistic (or Analysis of variance) is employed:

$$F = \frac{SSB/(k-1)}{SSW/(N-k)}$$

where SSB and SSW represent the sum of squares of between-sample and within-sample, respectively. N is the size of the 'pooled' (enlarged) sample and k= number of samples. The F-statistic has an F-distribution with v_1 (=k-1) and v_2 (=N-k) degrees of freedom.

In regression analysis, the F-statistic is also used to test whether the explanatory variables have any significant influence on the dependent variable (overall significance of explanatory variables).

To capture the extent of the observed increasing trend, the CI is selected. Following Chang et al (1993b), the following simple regressions are run to estimate residual volume and residual value:

$$Volume_t = a + bI_t + \mu_t$$

$$Value_t = c + dI_t + E_t$$

where $Volume_t$ and $Value_t$ refer to the trading transactions (in million units and Ringgit Malaysia, respectively) for the CI component stocks at day t; I_t denotes KLSE CI on day t; μ_t and

E_t are random error terms; a and c are the intercepts while b and d are the regression coefficients. Upon estimation of both residuals, regressions are used to test the equality of trading volume and value in the respective subperiods:

$$Volume^*_t = g_0 + g_1 Dm + \epsilon_t$$

$$Value^*_t = h_0 + h_1 Dm + v_t$$

where $Volume^*_t$ and $Value^*_t$ denote residual volume and residual value, respectively.

To test the equality of volume and value between subperiods 1 and 2, we set Dm=1 for subperiod 1 and 0 for subperiod 2. Similarly, to test the equality between subperiods 2 and 3, Dm=1 for subperiod 2 and 0 for subperiod 3. ϵ_t and v_t represent random errors.

2.3.3 Speed of Price Adjustment

Following Lin and Rozeff (1992) and Chang et al (1993b), the speed of adjustment is estimated using the regression model as given below:

$$\Delta Var_t = \alpha + \beta Var_{t-1} + E_t$$

where $\Delta var_t = Var_t - Var_{t-1}$ and Var_t denotes Parkinson's variance which is defined as:

$$Var_t = [\ln(DH^t/DL^t)]^2 / (4*\ln 2)$$

where DH^t and DL^t are daily high and low of the CI on day t. β represents the speed of adjustment coefficient, and E_t is the error term.

Lin and Rozeff (1992) assume that $0 \leq |\beta| \leq 1$. If $|\beta|=1$, it means all private information released or announced on day t-1 is fully incorporated into stock price on day t. The magnitude of β measures the speed adjustment to new information.

The regression is performed for the periods before and after extended trading hours, as well as for the whole period. To test the equality of the β coefficients in the two subperiods, a regression is utilised for the whole period. Following Chang et al (1993b),

$$\Delta Var_t = \alpha + \beta Var_{t-1} + \gamma(Dm) + \delta(DmVar_{t-1}) + E_t$$

where Dm is the dummy variable with Dm=1 for subperiod 1, and 0 for subperiod 2, for the event of extended trading hours. Likewise, Dm=1 for subperiod 2 and 0 for subperiod 3, for the event of full automation. Test of equality is applied whether there is a difference in the speed of price adjustment for the respective periods (extended trading hours and full automation).

2.3.4 Contemporaneous Relationship

Following Jain and Joh (1988) and Cheung et al (1993), three different specifications of the following model are estimated.

$$V_t = a + b|R_t| + c[D_t \mid R_t|]$$

$$+\sum_{k=1}^{n} e_k DD_{kt} + \sum_{k=1}^{n} f_k [DD_{kt} | R_t|]$$

$$+\sum_{k=1}^{n} g_k [DD_{kt} \mid R_t \mid D_t] + \mu_t$$

where V_t is number of shares traded divided by number of shares outstanding; R_t denotes return for period t; D_t =0 if the period t return is positive and D_t=1 if otherwise; DD_{kt} is a vector of dummy variables, k=1,2...n; n=25 with 21 for periods of the day and 4 for days of the week (due to data inconsistency, period before extended trading hours is excluded); μ_t is a random error term.

The three specifications:

(a) DD_{kt} variables are omitted ($e_k=f_k=g_k=0$)

(b) The seasonal variables affect only the intercept but not the slope of the regression ($f_k=g_k=0$).

(c) Full model in the equation.

2.3.5 Unit Root Test

The usual asymptotic results cannot be expected to apply if any of the variables in a regression model is generated by a non-stationary process. If the two variables are non-stationary, a common trend may cause a causal relationship between them. To keep standard assumptions from being violated when using time series data, the series need to be detrended or differenced.

Volume and return series are checked for stationarity by employing a unit root test. Dickey-Fuller Test is used to test for unit root (Fuller (1976) and Dickey and Fuller (1979)). The test statistic is computed in the same way as a t-statistic but it is normally referred to as τ-statistic. However, the τ-statistic does not have the standard t-distribution. In the presence of serial correlation of unknown form, the Augmented Dickey-Fuller or ADF test is adopted. The τ-statistic is then compared with the critical value provided by MacKinnon (1991) to determine the rejection of the hypothesis.

To demonstrate the ADF test, consider an AR(1) process,

$$y_t = \mu + \rho y_{t-1} + \epsilon_t$$

where μ and ρ are the parameters and the ϵ_t are assumed to be independently and identically distributed with zero mean and equal variance. The AR(1) process is stationary if $-1<\rho<1$. If $\rho=1$, that is, a unit root exists, the above equation represents a random walk with drift and y is said to be non-stationary. When this happens, the series should be differenced once (because it is integrated of order one, I[1]) in order to achieve stationarity.

To test the null hypothesis, consider the equation below:

$$\Delta y_t = \mu + \gamma y_{t-1} + \epsilon_t$$

where Δ is the first-differenced operator and $\gamma=\rho-1$, hence the null hypothesis is Ho: $\gamma = 0$. A large Dickey-Fuller statistic rejects the hypothesis of a unit root and suggests that the series is stationary.

2.3.6 Ljung-Box Q-statistic

Both series of return and volume are tested for white noise using the Ljung-Box Q-statistic. The Ljung-Box Q-statistic is given by,

$$Q_{LB} = T(T+2)\sum_{j=1}^{P} (r_j^2) / (T-j)$$

where r_j is the serial correlation coefficient of lag j and T is the number of observations. The Q-statistic is distributed as χ^2 with degree of freedom equals to p, the number of lagged serial correlations used. It tests the hypothesis that all of the autocorrelations are zero; that is, the series is white noise.

2.3.7 Akaike Information Criterion

The lag lengths of the variables are estimated by using Akaike Information Criterion (AIC) (Akaike, 1969 & 1970) given by:

$$\text{Akaike's AIC} = 2k/T + \log[(1/T) \sum_{t=1}^{T} \mu_t^2]$$

where μ_t represents error term, T is number of observations and k denotes number of regression coefficients, including the constant. Under certain conditions, the lag length is determined by choosing the lowest value of the AIC.

2.3.8 Granger-Causality Test

Return is said to cause volume if volume can be better predicted by using the past value of return than by not doing so and vice versa. Other relevant information, including past value of return and volume series, are being used in either case.

To examine the lead-lag structure of price-volume relationship using the causality test, most researchers argue that sometimes, return does not necessarily cause volume to change. Volume series, instead, may contain information which is useful in predicting the return. The Granger-causality test examines the dynamic relationship between stock return and trading volume. Consider the following unrestricted model for return (R) and volume (V):

$$R_t = \sum_{i=1}^{m} \alpha_i V_{t-i} + \sum_{j=1}^{n} \beta_j R_{t-j} + \mu_{1t} \quad \text{----------------(1)}$$

$$V_t = \sum_{i=1}^{p} \lambda_i V_{t-i} + \sum_{j=1}^{q} \delta_j R_{t-j} + \mu_{2t} \quad \text{----------------(2)}$$

where m, n, p and q are the lag lengths for the respective series. μ_{1t} and μ_{2t} are the error terms in equations 1 and 2, and it is assumed that they are not correlated.

If volume does not cause return, then α_i=0 for i=1,2,....m. Similarly, if return does not cause volume, δ_j should be zero for j=1, 2,....q. Therefore, the respective restricted models are as follow:

$$R_t = \sum_{j=1}^{n} \beta_j R_{t-j} + \mu_{1t} \quad \text{----------------(3)}$$

$$V_t = \sum_{i=1}^{p} \lambda_i V_{t-i} + \mu_{2t} \quad \text{----------------(4)}$$

Unidirectional causality from volume to return is indicated if α_i are jointly significantly different from zero in equation 1 and δ_j in equation 2 are not significantly different from zero. On the other case, unidirectional causality from return to volume is also indicated if α_i are not significantly different from zero and δ_j are significantly different from zero. If both α_i and $_j\delta$ are significantly different from zero in both equations, then there is feedback between return and volume. Finally, if α_i and δ_j are not significantly different from zero, this would suggest that there is no Granger-causality between return and volume.

The Granger-causality is tested by using the standard Wald F-statistic given by:

$$F_c = \frac{(ESSR\text{-}ESSU)/q}{(ESSU/(T\text{-}p\text{-}q))}$$

where T is the number of observations used in the unrestricted model in the above equation, ESSU is the error sum of squares for unrestricted model, and ESSR is the error sum of squares for the restricted model, p and q represent the respective lag lengths in equation 2. To perform Wald F-statistic in equation 1, the lag lengths of p and q would be substituted with m and n.