CHAPTER 2

DATA AND METHODOLOGY

2.1 Data

The data used in this study were the daily closing prices of financial stocks traded on the Kuala Lumpur Stock Exchange (KLSE). 30 stocks chosen for a detailed report were selected randomly from the Finance Sector. The data were from January 1992 to June 1999, which covers a period of seven and a half years. The daily closing prices were collected mainly from the KLSE Daily Dairy, Investors Digest and The Star newspaper.

A list of these 30 stocks with their respective market capitalization is presented in *Appendix I*. Stocks selected were divided into two portfolios according to their market capitalization: small companies and large companies. Stocks with market capitalization above RM800 million are categorized as large companies, while stocks with market capitalization below RM800 million are categorized as small companies. In this study, 15 of the stocks selected were of large companies and 15 were of small companies.

The values of the Kuala Lumpur Stock Exchange Composite Index (KLCI) were also collected for the same period for the purpose of sub-period analysis. An analysis of the sub-periods for each of the individual stock will allow us to know the persistency and consistency of the market anomalies throughout the different sub-periods. The whole period of this study is divided into 3 sub-periods corresponding to the market performance: Up trend market, Stable market and Down trend market.

The trend of the KLCI from January 1992 to June 1999 is presented in *Appendix II.* From the graph, the first sub-period was from 2nd January 1992 to 31 December 1993 where the KLCI increased from 550 to about 1275. The second sub-period of 2nd January 1994 to 31 December 1996 was one of relative stability where the KLCI fluctuated between 950 and 1250 points. The third sub-period was from 2nd January 1997 to 2 October 1998 where the KLSE experienced a decline with the KLCI declining from 1100 to the lowest point at 263. The market crashed during this period as a result of the Asian Financial Crisis in July 1997.

2.2 Methodology

The daily stock return are calculated as follows:

$$R_t = \left(\frac{P_t - P_{t-1}}{P_{t-1}} \right) \times 100$$

whereby, R_t refers to the return for day t

P_t refers to the daily closing price for day t

P_{t-1} refers to the daily closing price for the day (t-1)

Stock returns were calculated after adjustment for capital changes such as stock dividends, bonus issues, right issues and consolidation. A list of capital changes and their adjustment factors is given in *Appendix III.*

A classical linear regression model for daily returns is used in this study, where the stock returns for each stock regress on the dummy variables anomalous effect, such as day-of–the-week, pre-holiday and end-of-the month variables. The Ordinary Least Square (OLS) regression was run to test the significance of the coefficient for each variable. The following shows the model of daily returns for individual stocks:

$$R_t = \mu_1 D_{1t} + \mu_2 D_{2t} + \mu_3 D_{3t} + \mu_4 D_{4t} + \mu_5 D_{5t} + \mu_6 PH_t + \mu_7 EM_t + e_{ij}$$

where R_t refers to the daily returns for the individual stocks; D_{1t} to D_{5t} are dummy variables for Monday to Friday. *PH* is a dummy variable for the day that precedes a holiday, while *EM* is a dummy variable for the day if it is the last trading day of the month. The OLS coefficients μ_1 to μ_5 are the mean returns for Monday through Friday. Coefficients of μ_6 and μ_7 are the mean returns for pre-holiday and end-of-the-month variables. In the classical linear regression model, the error term is assumed to be independently and normally distributed with a constant variance.

Five main null hypotheses were tested in this study, namely:

(1) The mean return of each stock on any day is different across the days of the week. (Day-of-the week effect)

(2) The mean return of each stock between pre-holiday and ordinary days is different. (Pre-holiday effect)

(3) The mean return of each stock between the last trading day of a month and ordinary day is different. (End-of-the month effect)

(4) The mean return of each stock in 'good' news and 'bad' news environment is different across the days of the week.

(5) The existence of the seasonal variation in return volatility across the day-of-the-week effect.

The first hypothesis aims to test whether there is any statistically significant difference among the mean returns across the day of the week. The following null and alternative hypotheses are used to test for each stock:

H_0: $\mu_{d1} = \mu_{d2} = \mu_{d3} = \mu_{d4} = \mu_{d5}$

H_a: at least one pair of $\mu_{di} \neq \mu_{dj}$, $i \neq j = 1, 2, 3, 4, 5$

where $\mu_{d1}, \mu_{d2}, \mu_{d3}, \mu_{d4}, \mu_{d5}$ refer to the mean returns for Monday through Friday, respectively. If the null hypothesis of equality is rejected at 5% level of significance, this means that there is a day-of-the-week effect. A further analysis is carried out in the case of a significant result, such as a multiple comparisons test. The multiple comparisons test is used to identify those trading days that contribute to the rejection of the null hypothesis of equality in mean returns.

The second hypothesis is testing for the pre-holiday effect on each stock. The trading days in the sample period are divided into two subsets: the trading days prior to the holidays (pre-holiday) and the trading days after excluding the pre-holidays (ordinary days). The mean returns for the trading day on pre-holiday are compared with the mean returns for the trading days on ordinary days. The null and alternative hypotheses are used as follows:

H_0: $\mu_{PH} = \mu_d$

H_a: $\mu_{PH} > \mu_d$,

where μ_{PH} refers to the mean return of a stock for the trading day prior to the holiday and μ_d refers to the mean return of a stock for ordinary days. If the null hypothesis is rejected at 5% level of significance, then we conclude that the mean return of the stock for a trading day prior to a holiday is significantly greater than mean returns for ordinary days. To control for the holiday effect, any returns that are preceded by a holiday are omitted. For example, if Monday is a holiday, the return for the succeeding Tuesday is omitted in the sample.

The third hypothesis compares the mean returns of a stock for the last trading day in a month with the mean returns for a trading day on ordinary days. The trading days in the sample are divided into two subsets: the last trading day in a month and the trading days on other days that excluded the last trading day in a month. The mean returns for the last trading day in a month are then compared with the mean returns for the trading days on ordinary days. The null and alternative hypotheses are used as follows:

H_0: $\mu_{EM} = \mu_d$

H_a: $\mu_{EM} > \mu_d$,

where μ_{EM} refers to the mean return of a stock for the last trading day in a month and μ_d refers to the mean return of a stock for ordinary days. If the null hypothesis is rejected at 5% significance level, it can be concluded that the mean returns of the stock for a last trading day of a month is significantly greater than the mean returns for other trading days.

The aim of the fourth hypothesis is to take into consideration the market environment involving 'good' news and 'bad' news where the stocks were traded. Returns of each stock were divided into two categories: positive returns and negative returns. In each category, the mean returns of each stock on any day are compared in a week. If the null hypothesis is rejected at 5% significance level, it means there is a day-of-the-week effect.

In order to investigate whether the day-of-the-week effect in the stock returns may be due to the seasonal variation in return volatility, the ARCH models will be used to estimate simultaneously the conditional mean and variance of the returns in 30 finance stocks. Specifically, the study uses the Generalized Autoregressive Conditional Heteroscedasticity (GARCH) model which was proposed by Bollerslev (1986).

In general, the GARCH (p, q) process is given by,

$$Y_t = X_t\beta + \varepsilon_t$$

$$\varepsilon_t | \varphi_{t-1} \sim N(0, h_t)$$

$$h_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j h_{t-j}$$

where the order of $p \geq 0, q \geq 0$, p is the order of GARCH terms and q is the order of ARCH terms,

$\alpha_0 > 0, \quad \alpha_i > 0, \quad i = 1, ..., p,$

$\beta_j > 0, \quad j = 1, ..., q$

For $p = 0$, the process reduces to ARCH (q) process. For $p = q = 0$, ε_t is simply following a white noise process. Clearly in the ARCH (q) process the conditional variance is specified as a linear function of the past variances. However, in the GARCH (p, q) process it allowing both time-varying conditional heteroscedasticity and conditional variance.

The Akaike Information Criterion (AIC) will be used to identify the order of p and q in GARCH model. The information criterion has been widely used in time series analysis to determine the appropriate length of the distributed lag. This distributed lag can be determined by selecting the model with the smallest information criterion.

The following general GARCH (p, q) – M model will be used to investigate the seasonal variation of return volatility of each stock:

$$R_t = \alpha_0 + \alpha_1 h_t^{1/2} + \alpha_2 R_{t-1} + \sum_{m=1}^{d} \mu_m \delta_t^m + \xi_t$$

$$h_t = \beta_0 + \sum_{i=1}^{q} \beta_i \xi_{t-i}^2 + + \sum_{i=1}^{p} \beta_j h_{t-j} + \sum_{m=1}^{d} \mu_m^* \delta_t^m$$

where R_t is the daily returns of a stock for day *t*, R_{t-1} is the daily returns for day *t – 1*, ξ_t is an error term with zero mean and conditional variance h_t. α_0 and β_0 are constants, $\alpha_1, \alpha_2, \beta_i, \beta_j, \mu_m$ and μ_m^* are coefficients and δ_t^m is a daily dummy variable.

2.2.1 Test for Normality

The Kolmogorov-Smirnov (KS) statistics will be conducted to test if the samples conform to a normal distribution. The KS is a test of goodness-of-fit in which we specify the cumulative frequency distribution that would occur under the theoretical distribution and compare it with the observed cumulative frequency distribution. The null and alternative hypotheses are stated as below:

H_0: $F(x) = F_T(x)$

H_a: $F(x) \neq F_T(x)$

The point of greatest divergence between the observed and the theoretical distribution is identified by the following test statistic:

$$D = \max\left|F_O(x) - F_T(x)\right|$$

whereby

$F_O(x)$ = The observed cumulative frequency distribution of a random sample of *n* observations.

$F_T(x)$ = The theoretical frequency distribution under the null hypothesis.

The *D* value is then compared with the critical value of D_α at the α% significance level. If the calculated *D* value is greater than the critical value, then the rejection of null hypothesis will lead us to conclude that the data is not normally distributed.

2.2.2 Levene Test

The Levene test is used to test the null hypothesis that the populations have equal variances. In other words, it is used to test the homogeneity of variances. Under the classical linear regression model, homoscedasticity (equal variance) is one of the assumptions that must be satisfied when a parametric test is applied. The null and alternative hypotheses are stated as below:

H_0: $\sigma_1^2 = \sigma_2^2 = ...\sigma_n^2$

H_a: $\sigma_i^2 \neq \sigma_j^2$, where $i \neq j = 1,2,......,n$

The test statistic distributed with degrees of freedom $(t-1)$ and $(N-t)$ is given by:

$$F = \frac{\left[\sum_{i=1}^{t} r_i\left(\overline{w}_{i.} - \overline{w}_{..}\right)^2\right] / [t-1]}{\left[\sum_{i=1}^{t}\sum_{j=1}^{r_i}\left(\overline{w}_{ij} - \overline{w}_{i.}\right)^2\right] / [N-t]} \sim F_{t-1,N-t}$$

where $w_{ij} = \left|x_{ij} - \overline{x_i}\right|$ is the absolute difference between the j^{th} observation of the unit receiving i^{th} treatment and the sample mean of the i^{th} treatment.

$$\overline{w}_{i.} = \frac{\sum_{j=1}^{r_i} w_{ij}}{r_i}$$ is the mean of the absolute differences for the i^{th} treatment.

$$\overline{w}_{..} = \frac{\sum_{i=1}^{t}\sum_{j=1}^{r_i} w_{ij}}{N}$$ is the overall mean common to all the absolute differences.

N and r_i refer to the total observations and sample size, respectively.

If the null hypothesis is rejected at the 5% significance level, the populations are assumed to have unequal variances.

2.2.3 Parametric Test

The Parametric tests that will be used are the t-test for one sample, the t-test for two independent samples, One-Way Analysis of Variance (ANOVA) and multiple comparison tests such as Tukey's test.

(i) One sample t- test

One sample t-test is applied to test the null hypothesis that the mean returns from individual stocks are significantly different from zero on a certain day of the week. The null and alternative hypotheses for one sample t-test are stated as below:

H_0: $\mu_{di} = 0$

H_a: $\mu_{di} \neq 0$

The test statistic with a degree of freedom of ($n-1$) is defined as:

$$t = \frac{\mu_{di} - 0}{s/\sqrt{n}} \sim t_{(n-1)}$$

in which μ_{di} refers to the mean return of each stock on given weekday, s refers to the sample standard deviation and n refers to the sample size. The rejection of a null hypothesis shows that the mean return of stock on a certain day of the week is significantly different from zero.

(ii) Two Independent Samples Test

The two independent samples test is conducted by comparing the mean returns of one weekday with those from the rest of the days in a week. The appropriate test statistic to be used is stated below:

$$t = \frac{\left(\overline{R}_i - \overline{R}_j\right) - \left(\mu_i - \mu_j\right)}{\sqrt{S_p^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \sim t_{(n_i + n_j - 2)}$$

in which

S_p^2 is associated with the pooled variance estimate:

$$S_p^2 = \frac{(n_i - 1)S_i^2 + (n_j - 1)S_j^2}{n_i + n_j - 2}$$

$\overline{R}_i$ refers to the mean return on a given weekday or last day of the month or pre-holiday day

$\overline{R}_j$ is the mean return on the rest of the days of the week or ordinary days

S_i^2 is the sample variance for the weekday or last day of the month or pre-holiday day

S_j^2 is the sample variance for the rest of the days of the week or ordinary days

n_i is the sample size for the weekday or last day of the month or pre-holiday day

n_j is the sample size for the rest of the days of the week or ordinary days

If the null hypothesis is rejected at the 5% significance level, then we conclude that there is a statistically significant difference in the mean returns implying seasonality in daily, monthly or the day prior to a holiday.

(iii) One-Way Analysis of Variance (ANOVA)

The statistical method for testing the null hypothesis that the means of several populations are equal is analysis of variance (ANOVA). In one-way ANOVA, we use a single-factor, fixed-effects model to compare the effects of one factor or treatment on a dependent variable. In order to use ANOVA, certain conditions must be met. The samples must be randomly selected from normally distributed populations with equal variances.

The one-way ANOVA is used to test for a difference in mean returns between the days of the week. The test statistic for ANOVA is F ratio, which is a ratio of the mean square for between-groups variance to the mean square for within-group variance.

$$F = \frac{MS_{Treatment}}{MS_E} = \frac{SS_{Treatment}/(k-1)}{SS_E/(n-k)}$$

where $SS_{Treatment}$ is called the sum of squares due to treatments (i.e. between treatment) and SS_E is called the sum of squares due to error (i.e. within treatment). k refers to the number of groups and n is the total number of observations.

The F ratio is then compared with the F distribution, with $k-1$ and $n-k$ degrees of freedom. The null hypothesis will be rejected if the calculated F ratio is greater than the critical value $F_{\alpha,(k-1),(n-k)}$ at α% significance level. The rejection would imply that at least two of the daily mean returns are different, thus exhibiting daily seasonality according to the day-of-the-week effect.

(iv) Tukey's Test

Tukey's test is adopted to test the difference between any pair of mean returns across the day-of-the-week. This test is applied only when the equality of the mean returns is rejected under the one-way ANOVA. It conducts comparison of all the possible pairs of mean returns to determine if there is significant difference. Tukey's test procedure is based on the studentized range statistic, where the difference of two means is only significant if the absolute value of their difference exceeds the critical value, T_α. The critical value T_α is given by:

$$T_\alpha = q_{(\alpha,k,v)}\sqrt{\frac{MS_E}{2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

in which $q_{(\alpha,k,v)}$ = critical value of studentized range statistic

n_i = sample size of group i

n_j = sample size of group j

α = 0.05 significance level

k = number of groups

v = degree of freedom

If the absolute difference between any pair of days $|\mu_i - \mu_j|$, is significantly greater than T_α, it means that the mean returns of this pair of days are significantly different.

2.2.4 Nonparametric Test

If the assumptions of normality and equal variances are not satisfied, a parametric test should not be applied, since the results would not be valid. Under this condition, a nonparametric test should be carried out instead. In this study, nonparametric tests such as Mann-Whitney test and Kruskal-Wallis test are carried out.

(i) Mann-Whitney Test

This test is used to compare the difference of mean returns for two samples. It is an alternative test to the two independent samples t test without the normality assumption. The U statistic is calculated by ranking all observations algebraically from the smallest to the largest. After the ranking, values for each sample are then totaled. The calculated U statistic is then compared with the critical value of U for the Mann-Whitney test at the $\alpha\%$ significance level. The null hypothesis of equal means will be rejected if the calculated value is smaller than the critical value.

(ii) Kruskal-Wallis Test

The Kruskal-Wallis test is a nonparametric alternative to the usual analysis of variance for testing difference in means for more than two groups. To perform a Kruskal-Wallis test, the observations are ranked in ascending order and are replaced by their rank R_i, with the smallest observation having rank 1. In the case of ties, when observations have the same value, an average rank is assigned to each of the tied observations. The test statistic is given by:

$$H = \frac{12}{N(N-1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(N-1)$$

where T_i = sum of ranks in the i^{th} treatment

n_i = number of observations in the i^{th} treatment

N = total number of observations

k = number of samples

When there are a number of ties, a correction factor will be calculated and used to correct the H value. The correction factor is as follows:

$$C = 1 - \left\{ \frac{\sum_{i}^{G} \left(t_i^3 - t_i\right)}{N^2 - N} \right\}$$

where G = number of sets of tied observations

t_i = number of ties in any set i

Then, the new H value is as follows:

$$H = \frac{\frac{12}{N(N-1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(N-1)}{C}$$

H is distributed approximately as χ_{k-1}^2 under the null hypothesis. Therefore, the null hypothesis is rejected if the H value is greater than $\chi_{\alpha,k-1}^2$.

2.3 Statistical Tools

Computer softwares such as Excel, SPSS and Eviews version 3 have been used in this study for editing the data and for producing all the statistical results.