Journal of Applied Finance & Banking, vol.2, no.3, 2012, 271-294
ISSN: 1792-6580 (print version), 1792-6599 (online)
International Scientific Press, 2012

Forecasting SET50 Index with Multiple Regression based on Principal Component Analysis

N. Sopipan[1], W. Kanjanavajee [1] and P. Sattayatham[2]

Abstract

In this paper, we forecast SET50 Index (The stock prices of the top 50 listed companies on SET (Stock Exchange of Thailand)) by using multiple regression. At the same time, we consider the existence of a high correlation (the multicolinearity problem) between the explanatory variables. One of the approaches to avoid this problem is the use of principal component analysis (PCA). In this study, we employ principal component scores (PC) in a multiple regression analysis. As can be seen, 99.4% of variation in SET50 can be explained by all PCAs. Accordingly, we forecast SET50 Index closed price for the period 1/03/2011 through 31/03/2011 by using three models. We compare loss function, the model forecast explained by all PCs have a minimum of all loss function.

---

[1] Program of Mathematics and Applied Statistics, Faculty of Science and Technology, Nakhon Ratchasima Rajabhat University, Nakhon Ratchasima, Thailand.
e-mail: nopsopipan@gmail.com , wanava29@gmail.com
[2] School of Mathematics, Suranaree University of Technology, Nakhon Ratchasima, Thailand, e-mail: pairote@sut.ac.th

Article Info: *Received* : April 21, 2012. *Revised* : May 27, 2012
Published online : June 15, 2012

JEL classification numbers: C53

Keywords: Forecasting, SET50 index, Multiple regression analysis, Principal component analysis

1 Introduction

The characteristic that all stock markets have in common is uncertainty, which is related to their short and long-term future state. This feature is undesirable for the investor, but it is also unavoidable whenever the stock market is selected as an investment tool. The best that one can do is to try to reduce this uncertainty. Stock Market Forecasting (or Prediction) is one of the instruments in this process.

There are two types of forecasting, thequalitative and the quantitative method. Qualitative forecasting techniques are subjective, based on the opinion and judgment of consumers and experts, which is appropriate when past data is not available. It is usually applied to intermediate to long range decisions (e.g. informed opinion and judgment, Delphi method). Quantitative forecasting models are used to estimate future demands as a function of past data, which is appropriate when past data is available. It is usually applied to short to intermediate range decisions (e.g. time series methods, causal / econometric forecasting methods). Time series found the stock market follows a *random walk*, which implies that the best prediction you can have about tomorrow's value is today's value. Another technique is a causal model which establishes a cause-and-effect relationship between independent and dependent variables i.e. regression analysis which includes a large group of methods that can be used to predict future values of a variable using information about other variables. These methods include both parametric (linear or non-linear) and non-parametric techniques.

In this study we consider multiple regression analysis, which is is one of the most widely used methodologies for expressing the dependence of a response variable on several independent (predictor) variables. In spite of its evident success in many applications, however, the regression approach can face serious difficulties when the independent variables are correlated with each other (McAdams et al., (2000)). Multicollinearity, or high correlation between the independent variables in a regression equation, can make it difficult to correctly identify the most important contributors to a physical process. One method for removing such multicollinearity and redundant independent variables is to use multivariate data analysis (MDA) techniques. MDA have been used for analyzing voluminous environmental data (Buhr et al., (1992, 1995); Chang et al., (1988); Sanchez et al., (1986); Statheropoulos et al., (1998)).

One of method is principal component analysis (PCA), which has been employed in air-quality studies (Maenhaut et al., (1989); Statheropoulos et al., (1998); Shi and Harrison, (1997); Tian et al., (1989); Vaidya et al., (2000)) to separate interrelationships into statistically independent basic components. They are equally useful in regression analysis for mitigating the problem of multicollinearity and in exploring the relations among the independent variables, particularly if it is not obvious which of the variables should be the predictors. The new variables from the PCA become ideal to use as predictors in a regression equation since they optimize spatial patterns and remove possible complications caused by multicollinearity.

In this paper, we forecast SET50 Index (The stock prices of the top 50 listed companies on SET(Stock Exchange of Thailand) by using a multiple regression based on PCA. Finally, we compare the performance of some models with their loss function. In the next section, we present multiple a regression model and principal component analysis. The empirical methodology and model estimation are given in section 3 and the conclusion is given in section 4.

2 Models

2.1 Multiple Regression Model

Multiple linear regression (MLR) attempts to model the relationships between two or more explanatory variables and a response variable, by fitting a linear equation to the observed data. The dependent variable (Y) is given by:

$$Y = \widehat{\beta}_0 + \sum_{i=1}^{p} \widehat{\beta}_i X_i + \varepsilon \tag{1}$$

where X_i, $i = 1,..., p$ are the explanatory independent variables, $\widehat{\beta}_i$, $i = 0,1,..., p$ are the regression coefficients, and ε is the error associated with the regression and assumed to be normally distributed with both expectation value zero and constant variance (J.C.M Pires et al., (2007)).

The predicted value given by the regression model ($\widehat{Y}$) is calculated by:

$$\widehat{Y} = \widehat{\beta}_0 + \sum_{i=1}^{p} \widehat{\beta}_i X_i \tag{2}$$

The most common method to estimate the regression parameters $\widehat{\beta}_i$, $i = 0,1,..., p$ is the ordinary least square estimator (OLS).

MLR is one of the most used methods for forecasting. This method is widely used to fit the observed data and to create models that can be used for prediction in many research fields, such as biology, medicine, psychology, economics and the environment. Finance is a research field where developing prediction models (e.g. for the Thai stock market index), where the choice of selection input data is important. Naturally, the Thai stock market has unique characteristics, so the factors influencing the prices of stocks traded in this market are different from the factors influencing other stock markets (Chaigusin et al., 2008a).

Examples of factors that influence the Thai stock market are the foreign stock index, the value of the Thai baht, oil prices, gold prices, the MLR and many others. Some researchers have used these factors to forecast the SET index, including

Tantinakom (1996), who used trading value, trading volume, interbank overnight rates, inflation, the net trading value of investment, the value of the Thai baht, the price-earnings ratio, the Dow Jones index, the Hang Seng index, the Nikkei index, the Straits Times Industrial index and the Kuala Lumpur Stock Exchange Composite index. Khumpoo (2000) used the Dow Jones index, gold prices, the Hang Seng index, the exchange rate for the Japanese yen and Thai baht, the MLR, the Nikkei index, oil prices, the Straits Times Industrial index and the Taiwan weighted index. Chotasiri (2004) used the interest rates for Thailand and the US; the exchange rates for the USD, JPY, HKD and SKD; the stock exchange indices of the US, Japan, Hong Kong and Singapore; the consumer price indexand oil prices. Chaereonkithuttakorn (2005) used US stock indices, including the Nasdaq index, the Dow Jones index and the S&P 500 index. Rimcharoen et al. (2005) used the Dow Jones index, the Nikkei index, the Hang Seng index, gold prices and the MLR. Worasucheep (2007) used MLR, the exchange rate for Thai baht and the USD, daily effective over-night federal fund rates in the US, the Dow Jones index and oil prices. Chaigusin et al. (2008) used the Dow Jones index, the Nikkei index, the Hang Seng index, gold prices, the MLR and the exchange rate for the Thai baht and the USD. Phaisarn S. et al. (2010) used the Dow Jones index, the Nikkei index, the Hang Seng index and the MLR. The common factors that researchers used to predict the SET index are summarised in Table 1.

Table 1: Impact Factor for Stock Exchange of Thailand Index

	Tantinakom (1996)	Khumyoo (2000)	Chotasiri (2004)	Chaereon-Kithuttakorn (2005)	Rimcharoen et al. (2005)	Worasucheep (2007)	Chaigusin et al. (2008)	Phaisarn S. et.al (2010)
Nasdaq index				X				
Down Jones Index	X	X	X	X	X	X	X	X
S&P 500 Index				X				
Nikkei Index	X	X	X		X		X	X
Hang Seng Index	X	X	X		X		X	X
Straits Times industrial Index	X	X	X					
USD		X	X			X	X	
JPY		X	X					
HKD			X					
SKD			X					
Gold prices		X			X		X	
Oil Prices		X	X			X		
MLR		X			X	X	X	X

*X is selected in multiple regression.

2.2 Principal Component Analysis (PCA)

Consider a random variable $X = (X_1, ..., X_p)'$ with mean $\mu = (\mu_1, ..., \mu_p)'$, $(\cdot)'$ denotes transpose, $\mu_i < \infty$ $(i = 1, ..., p)$ and variance $\sum = (\sigma_{ij}), \sigma_{ij} < \infty$ $(i, j = 1, ..., p)$. Assume that the rank of $\sum$ is p and

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0 \tag{3}$$

are the p eigenvalues of $\sum$.

In the PCA we want to find uncorrelated linear function of $X_1,....,X_p$, say, $Z_1,...,Z_m$, $(m \le p)$, such that variances $V(Z_1),...,V(Z_m)$ account for most of the total variances among $X_1,....,X_p$, Also, we require $V(Z_1) > V(Z_2) > ... > V(Z_m)$. Algebraically, principal components are particular linear combinations of $X_1,....,X_p$, Geometrically, the principal component represents a new coordinate system obtained by rotating the original axes $X_1,....,X_p$, The new axes represent the directs with maximum variability.

Let $\alpha_i = (\alpha_{i1},...,\alpha_{ip})'$, $i = 1,...,m$ be a $p \times 1$ vector of weights for the respective components of X .

Consider the linear function

$$Z_1 = \alpha_1' X = \sum_{i=1}^{p} \alpha_{1i} X_i \tag{4}$$

Our aim is to find α_1 such that $V(Z_1)$ is maximum subject to the condition $\alpha_1' \alpha_1 = 1$. It is clear that $V(Z_1)$ can be increased by multiplying α_1 by some constant. To eliminate this arbitrariness we restrict our attention to coefficient vectors of unit lengths.

Now,

$$V(Z_1) = \alpha_1' \sum \alpha_1 .$$

Hence, we are required to find α_1 such that

$$\alpha_1' \sum \alpha_1 \tag{5}$$

is maximum subject condition $\alpha_1' \alpha_1 = 1$.

To maximize $\alpha_1' \sum \alpha_1$ subject to $\alpha_1' \alpha_1 = 1$, the standard approach is to use the technique of Lagrange multipliers. Maximize $\alpha_1' \sum \alpha_1 - \lambda(\alpha_1' \alpha_1 - 1)$, where λ is a Lagrange multiplier.

Differentiation with respect to α_1 gives

$$\sum \alpha_1 - \lambda\alpha_1 = 0, \quad \text{or} \quad (\sum - \lambda I_p)\alpha_1 = 0, \tag{6}$$

where I_p is the ($p \times p$) identity matrix.

Since, $\alpha_1 \neq 0$, there can be a solution only if $\sum - \lambda I_p$ is singular, i.e. if

$$\left| \sum - \lambda I_p \right| = 0$$

such that if λ is a latent root of $\sum$ and α_1 is its corresponding normalized latent vector.

Thus, λ is an eigenvalue of $\sum$ and α_1 is the corresponding eigenvector. To decide which of the p eigenvectors gives $\alpha_1' X$ with maximum variance, note that the quantity to be maximized is

$$\alpha_1' \Sigma \alpha_1 = \alpha_1' \lambda \alpha_1 = \lambda \alpha_1' \alpha_1 = \lambda$$

(by (6)) so λ must be as large as possible. Thus, α_1 is the eigenvector corresponding to the largest eigenvalue of $\sum$, and $Var[\alpha_1' X] = \alpha_1' \sum \alpha_1 = \lambda = \lambda_1$, the largest eigenvalue (by (3)).

In general, the *k*th PC of X is $Z_k = \alpha_k' X$ and $Var[\alpha_k' X] = \lambda_k$, where λ_k is the *k*th largest eigenvalue of $\sum$, and α_k is the corresponding eigenvector. This will now be proved for $k = 2$; the proof for $k \geq 3$ is slightly more complicated, but very similar.

The second PC, $Z_2 = \alpha_2' X$, maximizes $\alpha_2' \sum \alpha_2$ subject to being uncorrelated with $Z_1 = \alpha_1' X$, or equivalently subject to

$$Cov[Z_1, Z_2] = Cov[\alpha_1' X, \alpha_2' X] = 0,$$

where $Cov[x, y]$ denotes the covariance between the random variables *x* and *y*.

But

$$Cov[Z_1, Z_2] = Cov[\alpha_1' X, \alpha_2' X] = \alpha_1' \sum \alpha_2 = \alpha_2' \sum \alpha_1 = \alpha_2' \lambda_1 \alpha_1 = \lambda_1 \alpha_2' \alpha_1 = \lambda_1 \alpha_1' \alpha_2$$

.Thus, any of the equations

$$\alpha_1'\sum\alpha_2 = 0, \qquad \alpha_2'\sum\alpha_1 = 0, \qquad \alpha_1'\alpha_2 = 0, \qquad \alpha_2'\alpha_1 = 0$$

could be used to specify zero correlation between $Z_1 = \alpha_1'X$ and $Z_2 = \alpha_2'X$. Choosing the last of these (an arbitrary choice), and noting that a normalization constraint is again necessary, the quantity to be maximized is

$$\alpha_2'\Sigma\alpha_2 - \lambda(\alpha_2'\alpha_2 - 1) - \phi\alpha_2'\alpha_1$$

where λ, ϕ are Lagrange multipliers. Differentiation with respect to α_2 gives

$$\sum\alpha_2 - \lambda\alpha_2 - \phi\alpha_1 = 0$$

and multiplication of this equation on the left by α_1' gives

$$\alpha_1'\sum\alpha_2 - \lambda\alpha_1'\alpha_2 - \phi\alpha_1'\alpha_1 = 0,$$

which, since the first two terms are zero and $\alpha_1'\alpha_1 = 1$, reduces to $\phi = 0$.

Therefore, $\sum\alpha_2 - \lambda\alpha_2 = 0$, or equivalently $(\sum - \lambda I_p)\alpha_2 = 0$, so λ is once more an eigenvalue of $\sum$, and α_2 the corresponding eigenvector.

Again, $\lambda = \alpha_2'\sum\alpha_2$, so λ is to be as large as possible. Assuming that $\sum$ does not have repeated eigenvalues, λ cannot equal λ_1. If it did, it follows that $\alpha_2 = \alpha_1$, violating the constraint $\alpha_1'\alpha_2 = 0$. Hence λ is the second largest eigenvalue of $\sum$, and α_2 is the corresponding eigenvector.

The second principal component is, therefore,

$$Z_2 = \alpha_2'X \qquad \text{with} \qquad V(Z_2) = \lambda_2.$$

To find the kth principal component, $Z_k = \alpha_k'X$, we are to find α_k such that $V(Z_k)$ is maximum subject to the condition $\alpha_k'\alpha_k = 1$ and $\alpha_k'\alpha_{k'} = 0$, $(k \neq k', k, k' = 1, ..., m)$.

It follows that $Z_k = \alpha_k'X$ with $V(Z_k) = \lambda_k$, $k = 1, ..., m$ where α_k is the normalized eigenvector corresponding corresponding to λ_k. Clearly,

$$Cov(Z_k, Z_{k'}) = Cov(\alpha_k{}' X, \alpha_{k'}{}' X) = \alpha_k{}' \Sigma \alpha_{k'} = \alpha_k{}' \lambda_{k'} \alpha_{k'} = 0. \quad k \neq k'$$

By Spectral Decomposition Theorem, we can write $\Sigma = \mathrm{A}\Lambda\mathrm{A}'$ where $A = (\alpha_1, ..., \alpha_p)$, $\Lambda = Diag.(\lambda_1, ..., \lambda_p)$. Note that some of the λ_i's may be zeros. Therefore, the total population variance among $X_1, ..., X_p$ is

$$\sum_{i=1}^{p} V(X_i) = \mathrm{tr}\,\Sigma = \mathrm{tr}(\mathrm{A}\Lambda\mathrm{A}') = \mathrm{tr}(\Lambda\mathrm{A}\mathrm{A}') = \mathrm{tr}(\Lambda) \quad \text{since } \mathrm{A}\mathrm{A}' = \mathrm{I}$$

$$= \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} V(Z_i).$$

The total population variance among $Z_1, ..., Z_p$ is the same as the total population variance among $X_1, ..., X_p$. The proportion of the total variance accounted for by the *k*th P.C. is $\lambda_k \Big/ \sum_{i=1}^{p} \lambda_i$. The first m P.C.'s with the *m* largest variance account for $\sum_{i=1}^{m} \lambda_i \Big/ \sum_{i=1}^{p} \lambda_i$ proportion of the total variance of X. If, therefore, most (80-90%) of the total variance in X is accounted for by the first m components $Z_1, ..., Z_m$, then for large p ,these components can replace the p original $X_1, ..., X_p$ to explain the variability among the variables and the subsequent components $Z_{m+1}, ..., Z_p$ can be discarded.

2.3 Multiple regressions by principal components

Let $\{X_{it}, t \in T\}$ and $\{Y_t, t \in T\}$ be $p+1$ discrete time stochastic processes defined as $T = \{1, 2, ..., n\}$, $n \in \mathbb{Z}^+$, $i = 1, ..., p$. Let us assume the parallel evolution of processes to be known until a given instant of time. We deal with the problem of forecasting the process $\{Y_t\}$ (output process) by using the additional information of the process $\{X_{it}\}$ (input process).

If $\{X_{it}\}$ process has multicollinearity, the forecasting procedure can be performed by means of the PCA of processes. So, a multiple regression by principal components model states how the output is related to the values of the input through the random variables in the orthogonal decomposition for the output process.

A multiple regression with PCA model consists of expressing the output process *Y*, as a function of the input process, in a similar way to its orthogonal decomposition through the principal components. The predicted value given by the regression model ($\hat{Y}$) is calculated by:

$$\hat{Y} = \widehat{\alpha_0} + \sum_{i=1}^{m} \widehat{\alpha_i} Z_i \tag{7}$$

where $Z = \{Z_1, ..., Z_m\}$, is the PCA matrix of X, $\hat{\alpha}_i$, $i = 0,1,...,m,\ m \le p$ is the regression parameters.

3 Empirical Methodology and Model Estimation Results

3.1 Data

The data sets used in this study are a dependent variable, which is the daily closed prices of SET50 Index at time *t* ($SET50_t$) and the explanatory independent variables are the differences between the daily closed price factors which include:

$SET50_{t-1}$: Stock Exchange of Thailand Index at time $t-1$.

$FTSE$: London Stock Exchange Index at time $t-1$.

DAX : Frankfurt Stock Exchange Index at time $t-1$.

$DJIA$: Dow Jones Index at time $t-1$.

$SP500$: S&P 500 Index at time $t-1$.

NIX : Nikkei Index at time $t-1$.

$HSKI$: Hang Seng Index at time $t-1$.

STI : Straits Times Industrial Index at time $t-1$.

$KLSE$: Kuala Lumpur Stock Exchange Index at time $t-1$.

PSI : Philippine Stock Exchange Index at time $t-1$.

$JKSE$: Jakarta Composite Index at time $t-1$.

$KOPI$: South Korea Stock Exchange (200) Index at time $t-1$.

USD : Currency in Thai Baht to one dollar at time $t-1$.

JPY : Currency in Thai Baht to 100 Yens at time $t-1$.

HKD : Currency in Thai Baht to one dollar of Hong Kong at time $t-1$.

SKD : Currency in Thai Baht to one dollar of Singapore at time $t-1$.

$GOLD$: Gold Price at time $t-1$.

OIL : Oil Price at time $t-1$.

All data is in the period 4/01/2007 through 30/03/2011 ($t=1$,…, 1,038 observations). The data set is obtained from the Stock Exchange of Thailand. The data set is divided into in-sample ($R=1{,}015$ observations) and out-of-sample ($n=23$ observations).

Descriptive statistics and correlations are given in Table 2 and Table 3. As can be seen from Table 3, high correlation coefficients were found between dependent variables (SET50) and explanatory variables with a high significance (p<0.01). Also high correlation coefficients were found between explanatory variables with high significance (p<0.01) which show that there was a multicollinearity problem.

Multiple regression analyses based on raw data also show that there was a multicollinearity problem with the variance inflation factor (VIF) in Table 1 (VIF >=5.0). Once of the approaches to avoid this problem is PCA. Hence, principal component analysis has been completed based on eighteen explanatory variables, and the overall results of the PCA are shown in Tables 3-5, respectively.

Table 2: Descriptive Statistics of SET50 Index and explanatory variables

Index	Mean	Std. Deviation	VIF
SET50	515.7789	114.96460	
$SET50_{(t-1)}$	515.5573	114.85694	54.72502
FTSE	5477.1510	791.24116	53.91208
DAX	6260.8784	1063.66831	60.32773
DJIA	11035.8638	1794.96252	312.1992
SP500	1200.8380	219.90409	438.4295
NIX	12073.7776	3215.51712	119.5085
HSKI	21005.4472	3748.08152	16.0591
STI	2851.2864	551.68098	63.0712
KLSE	1240.6635	184.58338	38.67279
PSI	3051.6809	633.69156	35.81688
JKSE	2374.1069	631.65012	70.35281
KOSPI	1626.9827	258.62063	24.14878
USD	32.7062	1.64552	14.41949
JPY	33.5216	3.28742	36.06898
SGD	23.2101	0.48150	5.880605
HKD	4.2678	0.20110	29.46127
Gold	957.4253	219.30643	34.49864
Oil	78.9187	21.70943	14.59976
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.882	
Bartlett's Test of Sphericity	Approx.	50246.096	
	df	153	
	Sig.	0.000	

Table 3: Correlation matrix of SET50 index and explanatory variables

Index	SET50	$SET50_{(t-1)}$	FTSE	DAX	DJIA	SP500	NIX	HSKI	STI	KLSE	PSI	JKSE	KOSPI	USD	JPY	SGD	HKD	Gold	Oil
SET50	1.0000																		
$SET50_{(t-1)}$	0.9964**	1.0000																	
FTSE	0.7431**	0.7442**	1.0000																
DAX	0.7783**	0.7797**	0.9679**	1.0000															
DJIA	0.6965**	0.6974**	0.9706**	0.9739**	1.0000														
SP500	0.6559**	0.6567**	0.9690**	0.9646**	0.9944**	1.0000													
NIX	0.3407**	0.3424**	0.8394**	0.8139**	0.8740**	0.9095**	1.0000												
HSKI	0.8665**	0.8695**	0.8023**	0.8400**	0.7986**	0.7715**	0.5383**	1.0000											
STI	0.8433**	0.8444**	0.9501**	0.9525**	0.9218**	0.9136**	0.7526**	0.8812**	1.0000										
KLSE	0.9417**	0.9423**	0.7528**	0.7687**	0.6718**	0.6383**	0.3498**	0.8156**	0.8515**	1.0000									
PSI	0.9102**	0.9094**	0.7500**	0.7649**	0.6650**	0.6358**	0.3782**	0.7510**	0.8454**	0.9520**	1.0000								
JKSE	0.8870**	0.8876**	0.4358**	0.4741**	0.3485**	0.2987**	-0.0583**	0.6701**	0.5891**	0.8862**	0.8408**	1.0000							
KOSPI	0.9578**	0.9595**	0.7254**	0.7813**	0.6876**	0.6469**	0.3524**	0.8857**	0.8435**	0.9184**	0.8710**	0.8514**	1.0000						
USD	-0.9042**	-0.9053**	-0.6765**	-0.7305**	-0.6457**	-0.5957**	-0.2978**	-0.7642**	-0.7563**	-0.8953**	-0.8587**	-0.8142**	-0.8639**	1.0000					
JPY	-0.2695**	-0.2727**	-0.7291**	-0.7388**	-0.8144**	-0.8455**	-0.9389**	-0.4826**	-0.6471**	-0.2354**	-0.2465**	0.1321**	-0.2704**	0.2630**	1.0000				
SGD	-0.1444**	-0.1414**	-0.2983**	-0.2925**	-0.3068**	-0.2942**	-0.3610**	-0.0407**	-0.2786**	-0.1863**	-0.2723**	0.0382**	-0.1110**	0.3197**	0.3790**	1.0000			
HKD	-0.7852**	-0.7869**	-0.2857**	-0.3198**	-0.2312**	-0.1683**	0.2117**	-0.5004**	-0.4016**	-0.7281**	-0.6608**	-0.9001**	-0.7123**	0.7738**	-0.2108**	-0.0112**	1.0000		

Gold	0.4163**	0.4155**	-0.1921**	-0.1749**	-0.2942**	-0.3458**	-0.6582**	0.1096**	-0.0457**	0.4234**	0.3705**	0.7516**	0.3749**	-0.3735**	0.6960**	0.2966**	-0.7899**	1.0000	
Oil	0.5539**	0.5617**	0.4101**	0.4691**	0.4822**	0.4604**	0.2300**	0.6399**	0.4576**	0.4012**	0.2647**	0.4425**	0.5436**	-0.4426**	-0.3542**	0.3378**	-0.5036**	0.1670**	1.0000

**Correlation is significant at the 0.01 level (2-tailed).

3.2 Results of Principal Component Analysis

Firstly, the results of Bartlett's sphericity test are shown in Table 2 This test is for all correlations are zero or for testing the null hypothesis where the correlation matrix is an identity matrix (M.Mendes, 2009) which was used to verifying the applicability of PCA. The value of Bartlett's sphericity test SET70 had 50,246.096 which suggests that the PCA is applicable to our data sets ($P < 0.0001$). Overall Kaiser's measure of sampling adequacy was also computed as 0.882 which indicated that sample sizes were enough to apply the PCA (KAISER, 1960).

Table 4: Eigenvalues for PCAs

Component	Initial Eigenvalues		
	Total	% of	Cumulative
1	11.089	61.606	61.606
2	4.340	24.110	85.715
3	1.381	7.670	93.385
4	.536	2.979	96.365
5	.207	1.151	97.515
6	.114	.631	98.147
7	.089	.494	98.641
8	.063	.349	98.990
9	.042	.235	99.225
10	.034	.191	99.416
11	.027	.148	99.564
12	.021	.119	99.682
13	.017	.096	99.778
14	.012	.065	99.843
15	.011	.063	99.906
16	.010	.054	99.960
17	.006	.033	99.993
18	.001	.007	100.000

According to the results of PCA (Table 4), there are three principal components principal components out of eighteen (PCA1-3) with eigenvalues greater than 1 which were selected for multiple regression analysis (Forecast 1).

Because eigenvalues represent variances and a component with an eigenvalue of less than 1 is not significant.

Thus, the first of three principal components provides an adequate summary of the data for most purposes. Only first three principal components, explaining 93.385% of the total variation, should be sufficient for almost any application (Table 4).

According to the results of the correlation matrix of SET50 and PCAs (see Table 5), out of eighteen principal components there are four principal components (PCA1-2, $p<=0.05$, PCA9, PCA13, $p<=0.01$) with correlations between SET50 and PCA not zero which were selected for multiple regression analysis (Forecast 2.). Lastly, we selected all PCAs to forecast SET50 for multiple regression analysis (Forecast 3.).

Table 5: Correlation Matrix of SET50 and PCAs

Component	SET50	PCA1	PCA2	PCA3	PCA4	PCA5	PCA6	PCA7	PCA8	PCA9	PCA10	PCA11	PCA12	PCA13	PCA14	PCA15	PCA16	PCA17	PCA18
SET50	1.0000																		
PCA1	0.9319**	1.000																	
PCA2	0.3246**	.000	1.000																
PCA3	-0.0198	.000	.000	1.000															
PCA4	-0.0223	.000	.000	.000	1.000														
PCA5	0.0276	.000	.000	.000	.000	1.000													
PCA6	0.0198	.000	.000	.000	.000	.000	1.000												
PCA7	0.0428	.000	.000	.000	.000	.000	.000	1.000											
PCA8	-0.0360	.000	.000	.000	.000	.000	.000	.000	1.000										
PCA9	-0.0738*	.000	.000	.000	.000	.000	.000	.000	.000	1.000									
PCA10	-0.0418	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000								
PCA11	-0.0219	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000							
PCA12	-0.0083	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000						
PCA13	0.0717*	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000					
PCA14	0.0236	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000				
PCA15	0.0172	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000			
PCA16	0.0247	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000		
PCA17	-0.0177	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	
PCA18	-0.0070	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000

**, * Correlations significant at the 0.01, 0.05 level (2-tailed), respectively.

3.3 Results of Multiple Regression with Principal Component Analysis

In this study, two approaches were employed using principal component scores in multiple regression analysis. As can be seen from Table 6, 97.4% of variation in SET50 can be explained by the first three PCA (Panel A.: Model Forecast 1.), 98.4% of variation in SET50 can be explained by the PCA1, PCA2, PCA9 and PCA13 (Panel B.: Model Forecast 2) and 99.4% of variation in SET50 can be explained by all PCAs (Panel C.: Model Forecast 3).

For the Forecasts 1-3 predicted SET50 prices were obtained for the following models:

Model Forecast 1.

$$SET50 = 520.073 + 109.243 \cdot PCA1 + 38.057 \cdot PCA2 - 2.32 \cdot PCA3$$

Model Forecast 2.

$$SET50 = 520.073 + 109.243 \cdot PCA1 + 38.057 \cdot PCA2 - 8.648 \cdot PCA9 + 8.404 \cdot PCA13$$

Model Forecast 3.

$$\begin{aligned} SET50 = {} & 520.073 + 109.243 \cdot PCA1 + 38.057 \cdot PCA2 - 2.32 \cdot PCA3 - 2.613 \cdot PCA4 \\ & + 3.240 \cdot PCA5 + 2.316 \cdot PCA6 + 5.016 \cdot PCA7 - 4.216 \cdot PCA8 \\ & - 8.648 \cdot PCA9 - 4.899 \cdot PCA10 - 2.567 \cdot PCA11 - 0.969 \cdot PCA12 \\ & + 8.404 \cdot PCA13 + 2.770 \cdot PCA14 + 2.018 \cdot PCA15 \\ & + 2.891 \cdot PCA16 - 2.073 \cdot PCA17 - 0.821 \cdot PCA18 \end{aligned}$$

In Panel D. we forecast the SET50 Index closed price for the period 1/03/2011 through 31/03/2011 by three models. We compare loss function, loss function for the model forecast 3 which explained by all PCAs have minimum of all MSE, MAE and MAPE. Figure 1 displays the SET50 Index closed prices and three models are used for forecast from the period 1/03/2011 through 31/03/2011.

Table 6: Multiple Regression Model based on PCA

Panel A. Multiple regression model based on first three PCA (Forecast 1)

Model	B	Std. Error	t	Sig.
(Constant)	520.073	.586	887.943	.000
PCA1	109.243	.586	186.425	.000
PCA2	38.057	.586	64.946	.000
PCA3	-2.320	.586	-3.960	.000

RMSE = 2151.207 R^2= 0.974 DW= 0.350

Panel B. Multiple regression model base on correlation PCA with SET50 (Forecast 2)

Model	B	Std. Error	t	Sig.
(Constant)	520.073	.456	1140.427	.000
PCA1	109.243	.456	239.435	.000
PCA2	38.057	.456	83.413	.000
PCA9	-8.648	.456	-18.953	.000
PCA13	8.404	.456	18.419	.000

RMSE = 1872.718 R^2= 0.984 DW= 0.69

Panel C. Multiple regression model based on all PCA with SET50 (Forecast 3)

Model	B	Std. Error	t	Sig.
(Constant)	520.073	0.293	1776.320	.000
PCA1	109.243	0.293	372.942	.000
PCA2	38.057	0.293	129.923	.000
PCA3	-2.320	0.293	-7.922	.000
PCA4	-2.613	0.293	-8.921	.000
PCA5	3.240	0.293	11.061	.000
PCA6	2.316	0.293	7.905	.000
PCA7	5.016	0.293	17.123	.000
PCA8	-4.216	0.293	-14.391	.000
PCA9	-8.648	0.293	-29.522	.000
PCA10	-4.899	0.293	-16.724	.000
PCA11	-2.567	0.293	-8.763	.000
PCA12	-0.969	0.293	-3.308	.001

PCA13	8.404	0.293	28.690	.000
PCA14	2.770	0.293	9.457	.000
PCA15	2.018	0.293	6.890	.000
PCA16	2.891	0.293	9.870	.000
PCA17	-2.073	0.293	-7.076	.000
PCA18	-0.821	0.293	-2.803	.005

RMSE = 886.961 R^2= 0.994 DW=2.047

Panel D. Loss function for a comparison of out of sample SET50 Index closed prices for the period 1/03/2011 through 31/03/2011

Model	**MSE**	**MAE**	**MAPE**
Forecast1	288.7332626	13.9773196	1.9501687
Forecast2	78.7924399	7.2242587	1.0204939
Forecast3	65.7462527	6.4303487	0.9085837

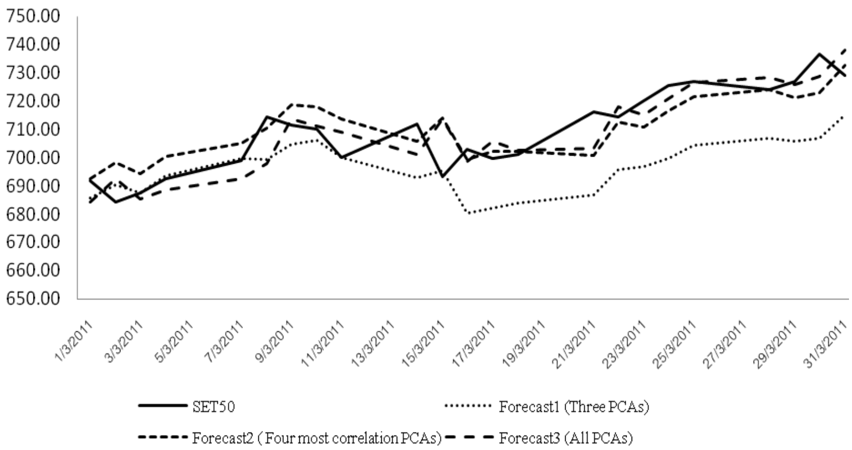

Figure 1: Graph of SET50 Index closed prices , Forecast SET50 with MLR based on first three PCs (Forecast1), four most closely correlated PCs (Forecast2) and all PCs (Forecast 3) for the period 1/03/2011 through 31/03/2011

4 Conclusion

Earlier studies showed that the relationship between SET50 Index and various factors i.e. other stock markets, foreign exchange, gold price, MLR and many others (Phaisarn et.al.,2010). Results of this study showed that regression models estimating SET Index can be used using these factors.

However, the number of significant correlation coefficients between the explanatory variables which were highest affect predictions for SET50 Index. Therefore, the relationships between explanatory variables, the multiple linear regression analysis of the prediction of the multicolinearity problem occurring between the explanatory variables. As for the higher correlations among the variables, some indirect effects on the SET50 Index become inevitable. In this case, it is very difficult to use multiple regression analysis to see and discuss the relationships correctly. In such cases, principal component analysis can be used to both reduce the number of variables and to get rid of the multicolinearity problem as well as to get a meaningful and easy analysis to see the complex relationships.

It has been observed that when the raw data of the study were used for the regression analysis for forecast SET50 Index, a multicolinearity problems existed (VIF >= 5.0). On the other hand, when the PCA analysis was completed on the explanatory variables and the PC scores were included in the multiple regression analysis as predictor variables instead of original predictor values, that problem diminished. Therefore, using the principal component scores in multiple regression analysis for predicting SET50 Index is more appropriate than using the original explanatory variables data.

Results of PCA showed that for, firstly, Bartlett's sphericity test for all correlations is zero or for testing the null hypothesis that the correlation matrix is an identity matrix. It used to verify the applicability of PCA. Overall Kaiser's measure of sampling adequacy indicated that sample sizes are enough to apply the PCA.

According to the results of eighteen principal components there are three principal components with eigenvalue greater than 1 which were selected for multiple regression analysis(Forecast 1.). Thus, the first of three PCs provides an adequate summary of the data for most purposes. If only the first three PCs are selected, this can explain 93.385% of the total variation. According to the results of correlation matrix of SET50 and PCAs, out of eighteen PCs there are four principal components with correlation between SET50 and PCA not zero which was selected for multiple regression analysis(Forecast 2.). Lastly, we selected all PCA to forecast SET50 for multiple regression analysis (Forecast 3.).

In this study, two approaches were employed in using principal component scores in multiple regression analysis. As can be seen 97.4% of variation in SET50 could be explained by the first three PCs, 98.4% of variation in SET50 could be explained by the PCA1, PCA2, PCA9 and PCA13 and 99.4% of variation in SET50 could be explained by all PCAs. Accordingly, we forecast SET50 Index closed prices for the period 1/03/2011 through 31/03/2011 by three models. When we compare loss function, the model forecast 3 is explained by all PCs which have a minimum of all MSE, MAE and MAPE.

References

[1] K. Chaereonkithuttakorn, *The Relationship between the Stock Exchange of Thailand Index and the Stock Indexes in the United States of America*, Master's Thesis in Economics, Chiang Mai University, Chiang Mai, Thailand, 2005.

[2] S. Chaigusin, C. Chirathamjaree and J. Clayden, Soft computing in the forecasting of the stock exchange of Thailand (SET), Management of Innovation and Technology, ICMIT 2008, 4th, (2008).

[3] S. Chotasiri, *The Economic Factors Affecting the Fluctuation of The Stock Exchange of Thailand Index*, Master Thesis in Economics, Chiang Mai University, Chiang Mai, Thailand, 2004.

[4] I.T. Jolliffe, *Principal Component Analysis*, 2nd ed. Springer-Verlag, New York Inc, 2002.

[5] C. Khumyoo, *The Determinants of Securities Price in the Stock Exchange of Thailand*, Master's Thesis in Economics, Ramkhamhaeng University, Bangkok, Thailand, 2000.

[6] M. Mendes, Multiple linear regression models based on principal component scores to predict slaughter weight of broiler, *Arch.Geflugelk*, **73**(2), (2009), 139-144.

[7] P. Mukhopadhyay, *Multivariate Statistical Analysis*, World Scientific Publishing Co. Pte. Ltd., London, 2009.

[8] J. Pires, Selection and validation of parameters in multiple linear and principal component regressions, *Environmental Modelling & Software*, **23**, (2008), 50-55.

[9] S. Rimcharoen, D. Sutivong and P. Chongstitvatana, Prediction of the Stock Exchange of Thailand Using Adaptive Evolution Strategies, Tools with Artificial Intelligence, ICTAI 05, 17th, (2005).

[10] P. Sutheebanjard and W. Premchaiswadi, Analysis of Calendar Effects: Day-of-the-Week Effect on the Stock Exchange of Thailand (SET), *International Journal of Trade, Economics and Finance*, **1**(1), 2010.

[11] P. Sutheebanjard and W. Premchaiswadi, Factors Analysis on Stock Exchange of Thailand (SET) Index Movement, *The 7th International Conference on ICT and Knowledge Engineering*, ICTKE2009, Bangkok, Thailand, (December 1-2, 2009).

[12] T. Tantinakom, *Economic Factors Affecting Stock Exchange of Thailand Index*, Master's Thesis in Economics, Chiang Mai University, Chiang Mai, Thailand, 1996.

[13] C. Worasucheep, *A New Self Adaptive Differential Evolution: Its Application in Forecasting the Index of Stock Exchange of Thailand*, Evolutionary Computation, 2007.