Chapter 2

Theory and Methods

In this chapter we describe the statistical theory and methods used for the analysis of banking shares prices in Thailand. These methods include basic methods for the analysis of data, including analysis of variance, data transformation, and principal components analysis, methods for time series analysis, and methods used in the modeling of market variables with stochastic volatility, including the generalized autoregressive conditional heteroskedacity (GARCH) model.

2.1 One-way Analysis of Variance (anova)

Considering the analysis of data in which the outcome is continuous and the determinant is categorical, this leads to a procedure called the analysis of variance (anova). We test the null hypothesis that the mean outcome is the same for each determinant level by computing a statistic called the *F-statistic* and comparing it with an appropriate distribution to get a *p-value*. Suppose that there are n_j observations in sample j ($j = 1, 2, \ldots, c$) denoted by y_{ij} for $i = 1, 2, \ldots, n_j$. The F-statistic is defined as

$$F = \frac{(S_0 - S_1)/(c-1)}{S_1/(n-c)}$$

where S_0 is the sum of squares of the data after subtracting their overall mean, while S_1 is the sum of squares of the residuals obtained by subtracting the sample mean from each of the c samples. The p-value is obtained as the upper tail area of the F distribution with $c-1$ and $n-c$ degrees of freedom (McNeil, 1996, page 67).

This procedure is based on two important statistical assumptions as follows:

(a) The standard deviations are the same in each of the c groups (also called the variance homogeneity assumption);

(b) The errors are independent and normally distributed.

We use box plots for assessing the assumption in (a) and we use normal scores plots of residuals (see, for example McNeil, 1996, page 45) for checking the normality assumption in (b). We also use Levene's test (see, for example, Brown and Forsythe,

1974) to test the variance homogeneity hypothesis. To check the independence assumption in (b), we use time series analysis, described in the next section.

It is useful to show the differences in the means by plotting confidence intervals. A measure of the difference between the means is given by the root-mean-squared difference, defined as the square root of the average of the squared difference between pairs of sample means, that is

$$rms = \sqrt{\frac{\sum_{j \neq k} (\bar{y}_j - \bar{y}_k)^2}{c(c-1)}}$$

Note that when $c = 2$, *rms* is just the magnitude of the difference between the sample means, $|\bar{y}_1 - \bar{y}_2|$. In this case one-way anova reduces to the two-sample t-test.

2.2 Data Transformation

If the statistical assumptions of variance homogeneity and normality are not satisfied, it might be that the data need to be transformed. The most common data transformation is to take logarithms, such as base 2 or base 10 or natural (base e) logarithms. The base for the logarithmic transformation does not affect the shape of the resulting distribution. It just affects the scale. Other common transformations include square roots, cube roots, and reciprocals.

Making an transformation of the data changes their skewness and kurtosis. The kurtosis is a measure of the extent to which the tails of the distribution are stretched, and should be 0 for a normal distribution.

Making a data transformation can also satisfy the variance homogeneity assumption, by removing the relation between the standard deviations and the means of groups of variables. For time series data, we can plot the standard deviation against the mean as a scatter plot, where each point is based on the data within a period such as a month or a quarter or a year. Then, after transforming the data, the objective is to remove the relation between the standard deviation and the mean.

2.3 Principal Components Analysis

One approach to handling the risk arising from groups of highly correlated market variables is principal components analysis. This takes historical data on related movements in the market variables and attempts to define a set of components or factors that explain the movements.

Principal components analysis is one of the simplest of the methods used in multivariate statistical analysis. The object of the analysis is to take p variables $X_1, X_2, \ldots, X_p$, and find combinations of these to produce indices $Z_1, Z_2, \ldots, Z_p$ that are uncorrelated. The lack of correlation is a useful property because it means that the indices are measuring different 'dimensions' in the data.

A principal components analysis starts with data on p variables for n cases. The first principal component is then the linear combination of the variables $X_1, X_2, \ldots, X_p$

$$Z_1 = a_{11}X_1 + a_{12}X_2 + \ldots, a_{1p}X_p,$$

that varies as much as possible for the individuals, subject to the condition that

$$a_{11}^2 + a_{12}^2 + \ldots + a_{1p}^2 = 1$$

Thus the variance of Z_1 , var(Z_1), is as large as possible given this constraint on the constants a_{1j}. The constraint is introduced because if this is not done then var(Z_1) can be increased by simply increasing any one of the a_{1j} values. The second principal component,

$$Z_2 = a_{21}X_1 + a_{22}X_2 + \ldots, a_{2p}X_p,$$

is such that var(Z_1) is as large as possible subject to the constraint that

$$a_{21}^2 + a_{22}^2 + \ldots + a_{2p}^2 = 1$$

and also to the condition that Z_1 and Z_2 are uncorrelated. The third principal component,

$$Z_3 = a_{31}X_1 + a_{32}X_2 + \ldots, a_{3p}X_p,$$

is such that Var(Z_3) is as large as possible subject to the constraint that

$$a_{31}^2 + a_{32}^2 + ... + a_{3p}^2 = 1$$

and also that Z_3 is uncorrelated with Z_2 and Z_1. Further principal components are defined by continuing in the same way. If there are p variables then there can be up to p principal components.

In order to use the results of a principal component analysis components are derived. However, it is useful to understand the nature of the equations themselves. In fact a principal components analysis just involves finding the eigenvalues of the sample covariance matrix (Manly, 1994: page 78-79)

2.4 Time Series Analysis

A time series is a group of numerical data sequentially in time. A time series is *stationary* if its statistical properties do not change with time. It is unlikely that a stationary time series will repeat itself exactly, but the series is repeatable in a probabilistic sense. Another way of looking at this is to say that the character of the series persists as you move forward or backward in time, and the only aspect that changes is the sampling error, which does not contain useful information. These sampling fluctuations could be relatively large compared to the persistent characteristic. These ideas lead to the sinusoid (the simplest function that repeats itself) and to the idea of measuring the amount of periodicity or repeatability in a time series by finding its covariance or correlation with a sine wave having a given period. A sinusoid is characterized by the property that talking a linear transformation of its argument only shifts its frequency and its phase or position relative to some origin. The cosine function is just a sine function whose argument is shifted by $\pi/2$, that is

$$\cos(x) = \sin(x + \pi/2)$$

Since sinusoidal functions are periodic it is natural to use them as a basis for approximating a stationary time series. This basis comprises sinusoidal waves with different frequencies each defined on the time interval spanned by the data. The first component appears exactly once on this time interval, the second comprises two repeated sinusoids, the third three sinusoids, and so on. These components are also called harmonics. The functional form for the j^{th} harmonic is a cosine wave with some phase ϕ, that is, $\cos\{2\pi j(t-1)/n+\phi\}$, $t = 1, 2, \ldots, n$. Using the mathematical theory of Fourier analysis any function defined at n equispaced points on a finite interval may be

represented exactly by a constant plus $n-1$ harmonics. The number of different frequencies in these components, m, is $(n-1)/2$ or $n/2$ (depending on whether n is odd or even) since there is a sine and a cosine harmonic at each frequency. If n is even this Fourier representation takes the form

$$y(t) = a_0 + \sum[a_j \cos\{2\pi j(t-1)/n\} + b_j \sin\{2\pi j(t-1)/n\}] + a_m \cos\{\pi(t-1)\}$$

where the summation is from $j=1$ to $j=m-1$. (Since $\sin\{\pi(t-1)\}$ is 0 for all integers t, in this case there is no sine harmonic at the highest frequency.) A similar formula applies if n is odd. Using the fact that a linear combination of a sine function and a cosine function at the same frequency may be expressed as a single sinusoid with some phase ϕ, an alternative formula for the Fourier representation is

$$y(t) = a_0 + \sum A_j \cos\{2\pi j(t-1)/n + \phi_j\}$$

where the amplitude $A_j = \sqrt{(a_j^2 + b_j^2)}$ and the summation is from 1 to m.
This Fourier representation is similar to linear regression analysis, where the sinusoidal components play the role of determinants or predictor variables. Since the number of parameters is exactly equal to number of data values, there is no residual error: the regression model provides a perfect fit to the data. Moreover it may be shown that the sum of products of sine and/or cosine harmonics over the range of frequencies is zero, which means that these harmonics are statistically uncorrelated with each other. Consequently each Fourier coefficient (a_j or b_j) is the regression coefficient of the time series y_t on the corresponding harmonic. The formulas for these coefficients (for n even) are as follows.

$$a_0 = \sum y(t)/n,\ a_m = \sum (-1)^{t-1} y(t)/n$$

$$a_j = (2/n)\sum y(t) \cos\{2\pi j(t-1)/n\}$$

$$b_j = (2/n) \sum y(t) \sin\{2\pi j(t-1)/n\}$$

we can see from these formulas that each Fourier coefficient may be interpreted as a covariance between the data and a sinusoid at the given frequency.

The periodogram of a time series $\{I_j, j = 1, 2, \ldots, m\}$ is defined in terms of the amplitudes of the harmonics in the Fourier representation as

$$I_j = (n/2)(a_j^2 + b_j^2)$$

The multiplier $n/2$ ensures that the j^{th} periodogram value is equal to the component of the variance in the data accounted for by a sinusoidal function with frequency j/n. since the sinusoidal terms are uncorrelated with each other, it follows that

$$\Sigma\{y(t) - \Sigma y(t)/n\}^2 = \Sigma I_j$$

This formula is known as Parseval's theorem.

This relation is just an analysis of variance for a time series. So the sum of the periodogram ordinates is equal to the total squared error of the data, and consequently the periodogram shows how much of the squared error of the data is accounted for by the various harmonics. For this reason it is useful to graph the scaled periodogram, obtained by dividing the periodogram by its sum. The scaled periodogram thus shows what proportion of the squared error is associated with each harmonic. Note that the frequency j/n is expressed in terms of the number of cycles per unit time. Since the values of j are 1, 2, …, m, the lowest frequency is $1/n$, corresponding to a period equal to the whole range of the data, and the highest frequency is close to 0.5 (exactly 0.5 if n is even), corresponding to cycles of length 2 with the data oscillating from one value to the next.

Autoregression

The periodogram and its logarithm may be used to investigate the character of a time series. Another useful graphical tool is the correlogram, or sample autocorrelation function, which comprises the set of estimated correlation coefficients between the series and itself at various spacings. Thus the (auto)correlation coefficient at spacing (or lag) s may be estimated from the formula

$$r_s = \sum_{t=1}^{n-s}\{y(t) - \bar{y}\}\{y(t+s) - \bar{y}\} \Big/ \sum_{t=1}^{n}\{y(t) - \bar{y}\}^2$$

and the correlogram is a graph of the series (r_s, $s = 1, 2, \ldots S$) against the spacing s. Since the number of terms used to calculate the correlation coefficient at lag S is $n–s$ where n is the length of the time series, the maximum spacing S should be busstantially less the n. According to statistical theory, when the sample size n is large the standard error of a correlation coefficient is approximately normally distributed with standard deviation $1/\sqrt{n}$, which tends to 0 as n gets large. This means that as the

length of an observed time series increases, the sample autocorrelation function of a stationary time series stabilizes, approaching a smooth curve. For a white noise process the theoretical correlation between observations at different spacing is zero, so you would expect the graph of its sample autocorrelation function to approach the horizontal axis $r = 0$ as n gets large.

Based on the normal distribution, which has 95% of its probability within 1.96 standard deviations of its mean, a 95% confidence interval for the autocorrelation at lag s ranges from $-1.96/\sqrt{(n-s)}$ to $1.96/\sqrt{(n-s)}$. In contrast, the periodogram values of a white noise process, being exponentially distributed with constant standard deviation, do not settle down as the length of the series increases. Instead they become more densely packed. Ljung & Box (1978) suggested using the statistic

$$Q = n(n+2)\sum_{s=1}^{m}\frac{r_s^2}{n-s}$$

where m is a specified integer substantially less than the series length n, to test the hypothesis that a time series is a sample from a white noise process. If it is necessary to fit a linear model involving p parameters to transform the series to a white noise process, where these parameters are estimated from the data, then Q is distributed approximately as a chi-squared distribution with $m-p$ degrees of freedom.

ARMA(1,1) model

A time series model takes the general form

$x_t = s_t + y_t$, where x_t, $t = 1, 2, ..., n$, is the series of observations, s_t is a signal and y_t is the residual (noise) series. The signal is usually modeled as the sum of a small number of sinusoidal components, that is

$$s_t = c + \sum_j A_j \cos(2\pi j\, t/n + \phi_j),$$

where c is a constant. The A_j are called *Fourier* coefficients. A simple specification for the noise, is a first-order autoregressive-moving-average, or ARMA (1,1) process, that is (see, for example, Box and Jenkins, 1976),

$$y_t = a_1 y_{t-1} + z_t + b_1 z_{t-1},$$

where $\{z_t\}$ is a sequence of independent errors with mean 0 and constant standard deviation σ_Z (*white noise*). The model is *stationary*, that is, its statistical properties do not change with time, if $-1 < a_1 < 1$. It may be shown that the standard deviation of y_t is given by

$$\text{st.dev}(y_t) = \sigma_Z \sqrt{(1+b_1^2)/(1-a_1^2)}\,.$$

The Ljung-Box test (Ljung & Box, 1978), may be used to assess the goodness-of-fit of the model. For further details see, for example, Chatfield (1989).

GARCH model

The volatility of a stock, σ, is a measure of our uncertainty about the returns provided by the stock.

Define σ_n as the volatility of a market variable on day n, as estimated at the end of day $n-1$. The square of the volatility on day n, σ_n^2, is the *variance rate.*

The standard approach to estimate σ_n from historical data. Suppose that the value of the market variable at the end of day i is S_i. The variable u_i is defined as the continously compounded return during day i (between the end of day $i-1$ and the end of day i)

$$u_i = \ln \frac{S_i}{S_{i-1}}$$

An unbiased estimate of the variance rate per day, σ_n^2, using the most recent m observations on the u_i is

$$\sigma_n^2 = \frac{1}{m-1} \sum (u_{n-i} - \bar{u})^2$$

where $\bar{u}$ is the mean of the u_i's

$$\bar{u} = \frac{1}{m} \sum_{i=1}^{m} u_{n-1}$$

It is appropriate to give more weight to recent data, so that.

$$\sigma_n^2 = \sum_{i=1}^{m} \alpha_i u_{n-i}^2$$

The exponentially weighted moving average (EWMA) model is particular case of this model where the weights, α_i, decrease exponentially as we move back through time. Specifically, $\alpha_{i+1} = \lambda \alpha_i$ where λ is a constant between zero and one.

It turns out that this weighting scheme leads to a particularly simple formula for updating volatility estimates, for large m so that

$$\sigma_n^2 = \lambda\sigma^2_{n-1} + (1-\lambda)u^2_{n-1}$$

We assume that u_t has mean 0 and follows the model

$$u_t = \sigma_t w_t$$

where $\{w_t\}$ is a series of independent, standardised normal random variables (that is, white noise). The series $\{\sigma_t\}$ is called the *volatility* of the series $\{u_t\}$.

A GARCH (1,1) model can be written in the form

$$\sigma_t^2 = \omega + \alpha u^2_{n-1} + \beta\sigma^2_{t-1}$$

where α, β and ω are parameters to be estimated from the data.

Given a series $\{u_t\}$observed on n consecutive trading days, the likelihood is

$$\prod_{t=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left(-\tfrac{1}{2}u_t^2/\sigma_t^2\right)$$

Taking logarithms, the estimates of the parameters are now obtained by substituting for σ_t and then maximising the expression

$$\sum_{t=1}^{n}\left[-2\ln(\sigma_t) - (u_t/\sigma_t)^2\right]$$

A special case of the GARCH (1,1) model, the exponentially weighted moving average (EWMA) model, arises when $\omega = 0$. In this case only the two parameters α and β need to be estimated.

Engle & Mezrich (1996) suggested simplifying the estimation procedure using variance targeting. In this method, the process u_t is assumed to have long-run variance V, so taking long-run expected values gives $V = \omega + \alpha V + \beta V$, from which $\omega = (1-\alpha-\beta)V$. Again only the two parameters α and β need to be estimated.

In either case the method involves starting with initial estimates, say $\alpha = 0.05$ and $\beta = 0.9$, putting $\omega = 0$ or $(1-\alpha-\beta)V$, and maximising the log-likelihood subject to constraints $0 < \alpha < 1$, $0 < \beta < 1$, and $\alpha + \beta < 1$.