Chapter 2

Methodology

This chapter includes a description of the methods used in the study, namely

1. Study design and conceptual framework
2. Computer programs
3. Methods for statistical analysis

2.1 Study Design and Conceptual Framework

This research is to study consumption of petroleum products in Thailand using time series analysis, by collecting data on consumption, import and production of gasoline, kerosene, diesel, jet petrol, fuel oil, liquid petroleum gas (LPG) by month during 1984-1999, from the Policy and Energy Planning Division of National Energy Policy Office.

2.1.1 Population: Amount of petroleum products consisting of regular gasoline, premium gasoline, kerosene, high speed diesel (HSD), low speed diesel (LSD), jet petrol (JP), fuel oil and liquid petroleum gas (LPG) in Thailand.

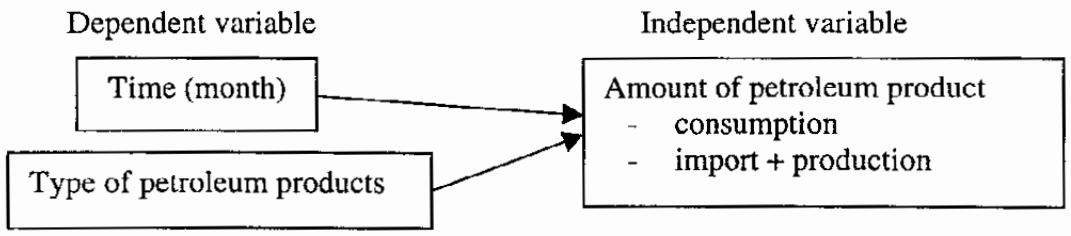
2.1.2 Sample: Amount of consumption, import and production of petroleum products consisting of regular gasoline, premium gasoline, kerosene, high speed diesel (HSD), low speed diesel (LSD), jet petrol (JP), fuel oil and liquid petroleum gas (LPG) in Thailand during 1984-1999.

2.1.3 Data Collection

(1) The data comprise the petroleum consumption, petroleum import, and petroleum production by month from 1984-1999, from the Energy Policy and Planning Department, National Energy Policy Office.

(2) The data are stored in Microsoft Excel.

2.1.4 Conceptual Framework



2.2 Computer Programs

The following computer programs were used for data analysis and thesis preparation.

2.2.1 A Statistical Package (Asp)

Asp is a suite of functions for graphing and analyzing statistical data. These programs run under Matlab Version 5 (Hanselman & Littlefield, 1997). It was mainly used to perform preliminary data and time series analysis.

2.2.2 Microsoft Excel

This program was mainly used to collect the data for this research. Some functions are helpful in finding matrix correlations and plotting graphs.

2.2.3 Microsoft Word

This program was mainly used to write and print the report of this research.

2.2.4 EcStat in Microsoft Excel

EcStat is an add-in to Microsoft Excel 97. It is a suite of routines for graphing and analyzing statistical data using an IBM-compatible PC.

EcStat is being developed because there is a need for both specialists and non-specialists in Statistics to be able to analyze data easily and to display the relevant results in a concise, appropriate way showing the information content of the data.

EcStat was mainly used in 'Comparison' and 'Relation' commands. The result of 'Comparison' is a one way analysis of variance (anova), and the result of 'Relation' is a graph with fitted linear relation shown on the plot.

2.3 Statistical Methods

2.3.1 Univariate and Bivariate Summaries

The mean and standard deviation (SD) are used to summarise the data for a single variable. They are calculated from the formulas

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N} \qquad \text{and} \qquad \text{S.D.} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \overline{X})^2}{N}}$$

Correlation analysis attempts to measure the strength of relationships between two variables by means of a single number called a correlation coefficient. The most widely used measure of linear correlation between two variables is called the Pearson product-moment correlation coefficient or simply the sample correlation coefficient. The measure of linear relationship between two variables X and Y is estimated by the sample correlation coefficient r, defined as

$$r = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 (y_i - \overline{y})^2}}$$

2.3.2 One Way Analysis of Variance

The analysis of variance (anova) is a method for the analysis of data in which the outcome is continuous and the determinant is categorical. This null hypothesis may be tested 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, 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

$$S_0 = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (y_{ij} - \overline{y})^2 \,, \quad S_1 = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (y_{ij} - \overline{y}_j)^2$$

and

$$\overline{y}_j = \frac{1}{n} \sum_{i=1}^{n_j} y_{ij} \,, \quad \overline{y} = \frac{1}{n} \sum_{j=1}^{c} \sum_{i=1}^{n_j} y_{ij} \,, \quad n = \sum_{j=1}^{c} n_j$$

Note that 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 each sample mean. If the population means are the same, the numerator and the denominator in the F-statistic are independent estimates of the square of the population standard deviation (assumed the same for each population). Then the p-value is the area in the tail of the F–distribution with $c-1$ and $n-c$ degrees of freedom. The F-test also requires the further assumption that the adjusted data (that is, the data adjusted by subtracting the population means from their respective samples) should have arisen from a normal distribution. Graphing the residuals against normal scores may check this assumption. If the normal scores plot shows a rough linear trend, the normality assumption might be reasonable for the data.

The standard errors used to compute these confidence intervals are based on an estimate of the common standard deviation given by the formula

$$s = \sqrt{\frac{S_1}{n-c}}$$

(see, for example, McNeil, 1996: page 66).

2.3.3 Two-Way Analysis of Variance

The two-way anova method is an extension of the method for comparing two means based on pair-matched data (using the paired t-test), and may be generalised to the comparison of several means. If the data array has r rows and no missing observations (giving $n = r \times c$ observations altogether), a correct p-value is based on an F-statistic defined as

$$F = \frac{(S_2 - S_{12})/(c-1)}{S_{12}/(n-c-r+1)}$$

where $$S_2 = \sum_{j=1}^{c}\sum_{i=1}^{r}(y_{ij} - \bar{y})^2, S_{12} = \sum_{j=1}^{c}\sum_{i=1}^{r}(y_{ij} - \bar{y}_j + \bar{y})^2$$

and $$\bar{y}_i = \frac{1}{c}\sum_{j=1}^{c} y_{ij}, \bar{y}_j = \frac{1}{r}\sum_{i=1}^{r} y_{ij}, \bar{y} = \frac{1}{rc}\sum_{j=1}^{c}\sum_{i=1}^{r} y_{ij}$$

The p-value is the area in the tail of the F-distribution which has $c-1$ and $n-c-r+1$ degrees of freedom, S_2 is the sum of squares of the data after adjusting for row effects,

S_{12} is the sum of squares after adjusting for both row effects and column effects (McNeil, 1996: page 73).

2.3.4 Regression Analysis

If both the outcome and determinant variables are continuous, a scatter plot may be used to display the data, and then the slope of a fitted straight line is used to represent the association between the determinant and the outcome.

In conventional statistical analysis the line fitted is the *least squares line*, which minimises the distances of the points to the line, measured in the vertical direction. This line is also called the regression line, and may be represented as

$$y = a + bx$$

where a is the *intercept* and b is the *slope* or *regression coefficient*. There is a linear association between a continuous determinant and a continuous outcome if the slope is different from 0. (McNeil, 1996: page 77).

Linear regression analysis rests on three assumptions as follows.

(1) The association is linear.
(2) The variability of the errors (in the outcome variable) is uniform.
(3) These errors are normally distributed.

These assumptions may be assessed by examining the residuals. To assess the first two assumptions, the residuals should be plotted against the *predicted values* given by the linear model. The normality assumption may be assessed by plotting the residuals against their normal scores, and tested using the Shapiro-Wilk test.

If there is a categorical covariate, the regression analysis may be extended to a model comprising a set of parallel straight lines, and this model may be fitted by least squares. The model takes the form

$$y_j = a_j + b\,x$$

where x is the value of the determinant, y_j is the mean outcome for a specified category j of the covariate.

2.3.5 Time Series Methods

A time series is a continuous set of numerical data measured sequentially in time. The measurements are often equispaced in time or nearly so.

There are four important objectives of time series analysis. These are (1) forecasting future values of a series, (2) estimating the trend or overall character of a time series, (3) modelling the dynamic relations between two or more time series, and (4) summarising characteristic features of a time series.

A crucial assumption underlying many of the methods used in time series analysis is stationarity, meaning that the statistical properties of the series do not change with time. If a time series is not stationary, making a transformation may help.

A trend is a general increase or decrease in a time series that persists. Trends are caused by long-term population changes, growth during product and technology introductions, changes in economic conditions, and so on. Trends are not necessarily linear because there are a large number of nonlinear causal influences that yield nonlinear series.

Seasonal series result from events that are periodic and recurrent (e.g., monthly changes recurring each year). Common seasonal influences are climate, human habits, holidays, repeated promotions, new-product announcements, and so on. Seasonality can occur in many different ways, for example, by week of the year, month of the year, day of the month, day of the week. When seasonal influences are present, seasonal forecasting models should be used.

Cyclical patterns, economic and business expansions (increasing demand) and contractions (recessions and depressions) are the most frequent causes of cyclical influences on time series. These influences most often last for two to five years and recur, but with no known period. In the search for explanations of cyclical movements, many theories have been proposed, including sunspots, positions of the planets, stars, long-wave movements in weather conditions, population life cycles, growth and decay of new products and technology (e.g., phonograph records, tape cassettes), product life cycles, and the economy.

Random time series are the result of many influences that act independently to yield nonsystematic and nonrepeating patterns about some average value. Purely random series have a constant mean and no systematic patterns.

Another pattern that is often seen in time series is a concept called *autocorrelation.* Correlation measures the degree of dependence or association between two variables. The term autocorrelation means that the value of a series in one time period is related to the value of itself in previous periods. With autocorrelation, there is an automatic correlation between observations in a series. Highly positive autocorrelated series without trends or seasonality are often random-walk series. When the mean of the series is always changing, the series is called a nonstationary series.

Many time series have a trend. In these situations it may be useful to fit a straight line, or possibly a quadratic function, and use the residuals as a basis for further statistical analysis. Least squares regression may be used to simply fit a linear or quadratic trend to time series data.

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 are moves forward or backward in time, and the only aspect that changes is the sampling error, which does not contain useful information. Of course these sampling fluctuations could be relatively large compared to the persistent characteristic.

These ideas lead to the sinusoid 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 characterised by the property that taking a linear transformation of its argument only shifts its frequency and its phase or position relative to some origin.

Since sinusoidal functions are periodic it is natural to use them as a basis for approximating a stationary time series. This basis comprises sine 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

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

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 or 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_j on the corresponding harmonic. The formulas for these coefficients (for n even) are as follows.

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

$$a_j = (2/n)\sum y_t \cos\{2\pi j(t-1)/n\},\ b_0 = (2/n)\sum y_t \sin\{2\pi j(t-1)/m\}$$

These formulas show 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 thchsinusoidal terms are uncorrelated with each other, it follows that

$$\sum (y_t - \sum y_t / n)^2 = \sum (I_j)^2$$

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 each 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. A function tsplot in the Asp library (McNeil, 1998a) may be used to show a periodogram of a time series.

A time series may be written in the form

$$y_t = p_t + s_t + z_t$$

where p_t is a trend (usually linear or quadratic), s_t is a stationary signal having the Fourier series representation and z_t is the residual, or noise series. In classical time series analysis, it is assumed that z_t has a normal distribution. In the simplest case, the terms in the process z_t are mutually uncorrelated, in which case the noise is called white noise.

Provided the noise is normally distributed, it may be shown that the periodogram coefficients are exponentially distributed. Now an exponential distribution has the property that its standard deviation is equal to its mean. However, the logarithm of an exponential distribution has approximately constant standard deviation. For this reason, it is useful, when analysing time series data, to plot the logarithm of the periodogram.

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

$$r_s = \frac{\sum_{t=1}^{n-s}\{y_t - \bar{y}\}\{y_{t+s} - \bar{y}\}}{\sum_{t=1}^{n}\{y_t - \bar{y}\}^2}$$

where w_t is a white noise process. This process is called a simple Markov process, and is characterised by the fact that the best forecast of its next value, z_{t+1} is based only on the current value, z_t. Note that this process reduces to white noise when the parameter is 0. This leads us to consider introducing a second parameter, extending the simple Markov process to the *second-order autoregressive* model, which takes the form

$$z_t = a_1 z_{t-1} + a_2 z_{t-2} + w_t$$

the general auoregressive process of order p takes the form

$$z_t = \sum_{j=1}^{p} a_j z_{t-j} + w_t$$

where the summation goes from j=1 to j=p. It may be shown that an authoregressive process is stationary if and only if all of the roots of the *characteristic* polynomial

$$P(z) = 1 - \sum_{j=1}^{p} a_j z^j$$

are *outside* the unit circle $|z|=1$ in the plane of complex numbers z. In particular, this means that a simple Markov process is stationary if $|a_1| < 1$. The condition for a

second-order autoregressive process is rather more complicated, but it may be shown that necessary and sufficient conditions are

$$a_1+a_2<1,\ a_2-a_1<1,\ |a_2|<1,$$

The Asp function *tsplot* used to analysis of a univariate time series. It shows the periodogram, the base 10 logarithm of the periodogram, and the autocorrelation function of a time series. It also has the capability of removing a linear or quadratic trend, fitting specified harmonics term, and estimating autoregressive coefficients at specified lags.