Chapter 2

Methodology

This chapter describes the statistical methods used for analyzing and forecasting sparkling beverages sales revenue in Southern Thailand. These methods include multiple linear regression model, observation-driven multiple linear regression model, Lee-Carter model, and Holt-Winters method. In this study, the sales revenue data are log-transformed to expose that statistical assumption of symmetry and variance homogeneously for residually are satisfied.

2.1 Sources of data

Business data used in this thesis was obtained from the sparkling beverages company. All sales revenue data was collected routinely in 14 provinces of Southern Thailand during years 2000 - 2006. Population data in each province were obtained from the 2000 Thai population and housing census by the National Statistical office.

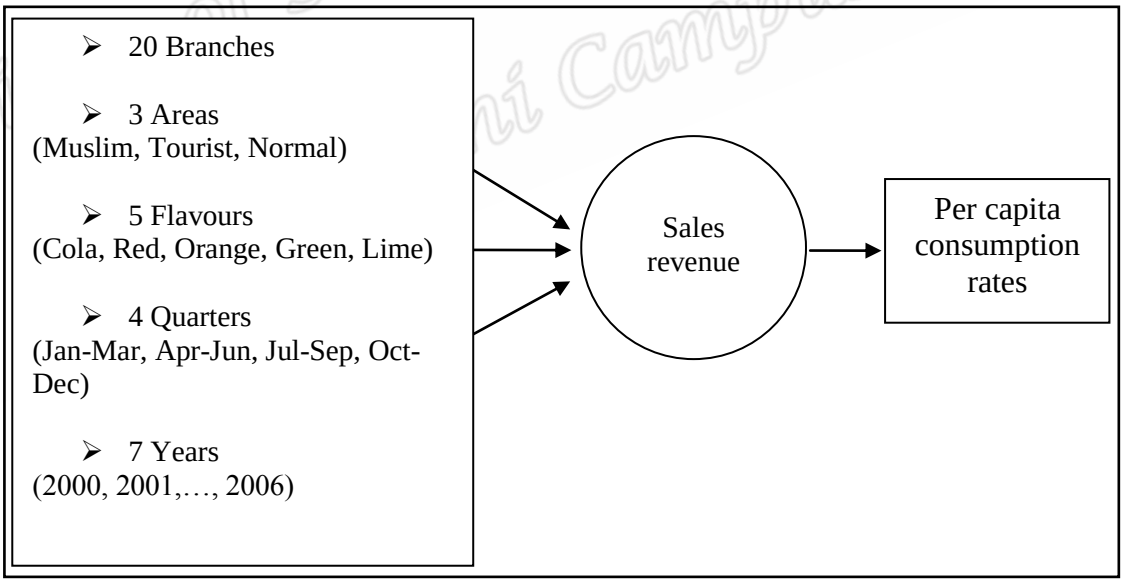
2.2 Data management

Data were available in computer files with records for sales revenue separated by flavour, package type, branch location, month, quarter and year. After correcting or impute data entry errors, records from years 2000 to 2006 were stored in a MySQL database. MySQL and Microsoft Excel programs were used to create sales revenue in Baht by month, quarter, year, flavour and branch location. The socio-demographic data is obtained from the 2000 Population Census of Thailand. All graphical and

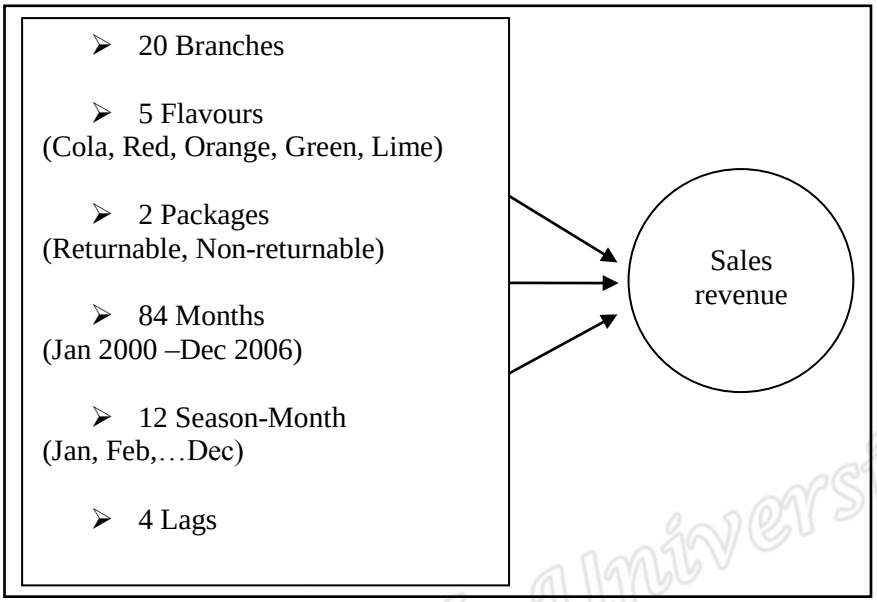
statistical analyses, including the log-transformation were performed using R (R Development Core Team 2008).

In this thesis a three-study path diagrams are followed: (1) sales analysis, (2) short-term sales forecasting, and (3) long-term sales forecasting. The first path diagram reveals the consumption rate analysis based on cells classified by branch location and area (Muslim, tourist and normal), and period of time (quarterly and yearly). The second path diagram reveals short-term sales forecasting based on cells classified by period of time (monthly and season-monthly), flavour, package type, branch location and autoregressive terms. The third path diagram reveals long-term sales forecasting based on cells classified by period of time (monthly) and branch location.

Study 1: Sales analysis



Study 2: Short-term sales forecast



Study 3: Long-term sales forecast

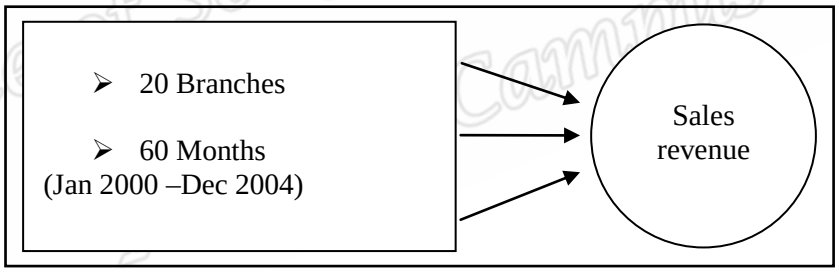


Figure 1: Path diagram for studies

2.3 Statistical methods for sales analysis

The annual per capita consumption rate is computed using sales revenue divided by the number of years and 1,000 populations. Population data is obtained from the 2000 Population and Housing Census of Thailand. Let S is sales revenue in t years and P_i is the number of population (in 1,000s of population) in branch i, flavour j, quarter k. The annual per capita consumption rate (Y_{ijtk}) in branch i is thus

$$Y_{ijtk} = \log\left(1{,}000 \times \frac{S_{ijtk}}{4P_i}\right) \quad (1)$$

2.3.1 Multiple linear regression additive model

The process of developing a statistical model varies depending on whether we follow a classical, hypothesis-driven (confirmatory data analysis) or a more modern, data-driven approach (exploratory data analysis). The goal of either approach is a model which imitates, as closely as possible, in as simple a way as possible, the properties of the objects or phenomena being modeled. Creating a model usually involves the following steps (MathSoft 1997):

1. Determine the variables to observe. In a study involving a classical modeling approach, these variables correspond to the hypothesis being tested. For data-driven modeling, these variables are the link to the phenomena being modeled.
2. Collect and record the data observations.
3. Study graphics and summaries of the collected data to discover and remove mistakes and to reveal low-dimensional relationships between variables.
4. Choose a model describing the important relationships seen or hypothesized in the data.
5. Fit the model using the appropriate modeling technique.
6. Examine the fit using model summaries and diagnostic plots.
7. Repeat steps 4-6 until satisfied with the model.

This study illustrates methods for statistical modeling available to businesses. Such consumption rates have positively skewed distributions so it is conventional to transform them by taking natural logarithms. Since the consumption rates based on

small branch are sometime zero, it is necessary to make some adjustment to avoid taking logarithms of 0**.**

In the modeling process, there is no one answer on how to build good statistical models, so we need to try different modeling technique that will, for example, allow us to fit nonlinear relationships, interactions, or different error structures. By iteratively fitting, plotting, testing, changing sometimes and then refitting, we will arrive at the best fitting model for our data.

At the beginning of the simplest model is based on linear regression. Multiple linear regression models of log-transformed sales revenue per 1,000 population are used to analyze annual per capita consumption and to study the main factors including product preference in each market segment. Three multiple regression analysis models are considered in the present study.

An additive model extends the notion of a linear model by allowing some or all linear functions of the predictors to be replaced by arbitrary smooth functions of the predictors. The standard linear regression model is a simple case of an additive model.

The simplest linear regression model takes the additive form

$$Y_{ijtk} = m + b_i + f_j + a_t + q_k, \quad (2)$$

where Y_{ijtk} is the natural logarithm of the quarterly revenue in 1000s of Baht per 1,000 population for branch i (b_i), flavour j (f_j), and quarter k (q_k) of year t (a_t), whereas m is the overall mean of Y_{ijtk}**.** Model (2) assumes that the patterns of per capita consumption rates vary by branch location, flavour, quarter and year.

2.3.2 Multiple linear regression interaction model

Additive models stumble when there are interactions among the various terms. To allow for possible spatial correlations between observations on different branches at the same time, and also for correlations between different flavours, additional terms allowing for these effects may be included as determinants in the model. Since this additive model does not allow for different flavour preferences in different regions, we also consider a more general model of the form

$$Y_{ijtk} = m + c_{ij} + a_t + q_k \quad (3)$$

To allow for possible spatial correlations between observations on different branches, and also for correlations between different flavours, additional terms allowing for these effects are included as determinants in the model. In this model, c_{ij} is an interaction between branch and flavour. Model (3) assumes that the patterns of per capita consumption rates in each branch location vary each year.

Generalizing further, we also consider the model

$$Y_{ijtk} = m + c_{ij} + d_{it} + q_k \quad (4)$$

Model (4) thus allows for interactions between branch-flavour (c_{ij}) and branch-year (d_{it}). It means that the model allows for possible spatial correlations between observations on different branches, and also for correlations between different flavours and years. The model assumes that the patterns of per capita consumption rates in each branch location vary between years and flavours. After fitting the models, we plotted confidence intervals for parameters after back-transforming so that the parameter estimates were expressed in terms of the original data, that is, in 1,000 Baht per 1,000 population. To do this, it was necessary to incorporate an additional

scale parameter for each factor to ensure that the mean revenue associated with each factor based on the fitted model matched the overall observed mean revenue.

2.3.3 Residuals

Residuals plot are principle tool for assessing how well a model fits the data. For regression models, residuals are used to assess the importance and relationship of a term in the model as well as to search for anomalous values. In *R* software, the function *qqnorm* can produce a normal probability plot, frequently used in analysis of residuals.

2.4 Statistical methods for short-term sales forecasting

2.4.1 Observation-driven multiple linear regression model

Regression analysis is the method for estimating values of one or more response variables from a set of predictor variables. The purpose of regression analysis is to assess the effects of the predictors on the response variable(*s*). The model is used for short-term sales forecasting. Autoregressive terms were included to account for time series and spatial correlations.

We fitted a multiple linear regression model to the data and compared results. Then, the model of log-transformed sales revenue, which contains seasonal effects and time-lagged terms, was applied for 12-month forecasting. The predictor variables compressed (a) the interactions between branch and flavour, (b) month of the year, and (c) the (log-transformed) sales revenues in the previous four months. If Y_t is the sales revenue in branch *i*, flavour *j*, of year *y*, month *t*, *s* is the "season-month"

(January, February,…) and ε_t is a series of independent normally distributed errors with mean 0, we write

$$\log Y_t = \alpha + \beta_{ij} + \gamma_s + \delta_1 \log Y_{t-1} + \delta_2 \log Y_{t-2} + \delta_{11} \log Y_{t-11} + \delta_{12} \log Y_{t-12} + \varepsilon_t \qquad (5)$$

where α, β and $(\delta_1, \delta_2, \delta_{11}, \delta_{12})$ are parameters in the model denoting an initial value, a trend, and two further coefficients denoting the influence of the sales in the previous four months, respectively, and $\gamma_1 = 0, \gamma_2, \gamma_{11}, \gamma_{12}$ is a set of seasonal effects indicating how the sales revenue varied with month of the year. In this case, the high-season months (including February, November and December) are usually affected the sales revenue since customers need to stock more products. Forecasts for $\log Y_{t+h}$ (*h* months in the future) are obtained by substituting the estimated values for the coefficients into the right-hand side of (5), using the forecast values themselves for values of $h > 1$. However, to obtain forecasts for Y_{t+h} (5) must be transformed back by exponentiation and the forecast is then the mean of $Y_{1'}$ which has a log-normal distribution with expected value

$$E[Y_t] = \exp(\mu) \qquad (6)$$

where μ is the mean. So, the forecast of Y_{t+h} is

$$E(Y_{t+h}) = \exp(\alpha + \beta_{ij} + \gamma_s + \delta_1 \log Y_{t+h-1} + \delta_2 \log Y_{t+h-2} + \delta_{11} \log Y_{t+h-11} + \delta_{12} \log Y_{t+h-12}) \qquad (7)$$

We used associative models that used explanatory variables to predict future sales revenue. The model is a multiple regression model since more than one predictor variable is used to predict sales.

2.4.2 Goodness-of-fit and forecasting errors analysis

The goodness-of-fit of the sales forecasting model is checked with such statistics as a r-squared and the standard error of regression relative to the mean and standard deviation of the response variable sales. Later, the partial explanatory power of each predictor variable is checked for expected sign and significance. The error terms are scanned for potential heteroskedasticity (serial autocorrelation of the error term) in order to satisfy the forecasting results**.**

2.5 Statistical methods for long-term sales forecasting

2.5.1 Lee-Carter model

For long-term sales forecasting, we modify the Lee-Carter model. Let Y_{it} be the logarithm of the sales for branch i (where i = 1,.., 20) at month t (where t =1,…, 60), the Lee-Carter model with principal component is

$$\log Y_{it} = a_i + b_i c_t + \varepsilon_{it}, \tag{8}$$

where a_i is the average sales by branch which is constant over time

b_i is the changes in the sales at branch i in response to changes in c_t over time

c_t is the temporal trend of sales changes over time

ε_{it} is a vector of error terms.

To obtain a unique solution, we impose that the branch-specific impact is sum to unity, and that the sum of the time trend index parameters is equal to zero. So, the constraints are:

$$\sum b_i = 1 \quad , \quad \sum c_t = 0 \tag{9}$$

We have extended the Lee-Carter model by including more than one principal component in the model, fitting time-space components, and involving continuous variables as predictors in the model. The modification of the Lee-Carter model with the first few components is also discussed in a study of Australian mortality rates by Booth (2002).

Lee-Carter model with 2 components extension can be written as

$$\log \mathrm{Y}_{it} = a_i + b_{i1}c_{t1} + b_{i2}c_{t2} + \varepsilon_{it} \tag{10}$$

Lee-Carter model with 3 components extension can be written as

$$\log \mathrm{Y}_{it} = a_i + b_{i1}c_{t1} + b_{i2}c_{t2} + b_{i3}c_{t3} + \varepsilon_{it} \tag{11}$$

We estimate the average sales (a_i) by

$$a_i = \prod_{t=t_1}^{t_n} \mathrm{Y}_{it}^{\frac{1}{n}} \tag{12}$$

For b_i and c_t parameters, we estimate using the singular value decomposition (SVD) method that is given in (14). Then we forecast c_t using Holt-Winters exponential smoothing method that is given in (15).

Let h be the 24 months in the future, the monthly sales prediction for Lee-Carter model with 3 components extension is thus obtained from

$$\log \hat{\mathrm{Y}}_{i,t+h} = a_i + b_{i1}\hat{c}_{t1+h} + b_{i2}\hat{c}_{t1+h} + b_{i3}\hat{c}_{t1+h} + \varepsilon_{i,t+h} \tag{13}$$

2.5.2 Least squares method

In order to find a least squares solution to the Lee-Carter equation we use a close approximation, suggested by Lee and Carter (1992), to the singular value decomposition method, assuming that the errors are homoschedastic.

The singular value decomposition takes an $n \times p$ matrix X and decomposes it into two orthogonal matrices and a diagonal matrix. The elements of the diagonal matrix are the singular values of X. The squares of the singular values of X are the eigenvalues of $X^{T}X$. To obtain the singular value decomposition in *R* software, use the *svd* function, which returns a list in which the first component is the vector of singular values, the second component is the orthogonal matrix *V*, and the third component is the orthogonal matrix *U*. The singular value decomposition can be used as a numerically stable way to perform many operations that are used in multivariate statistics. One such operation is estimating the rank of a matrix X (MathSoft 1997).

In this case, the singular value decomposition was applied to the average sales over time *t* for each branch *x* for the estimation of parameters (b_i *and* c_t).

$$Y_{it} - a_i = UDV', \tag{14}$$

where *D* is a diagonal matrix containing singular values and both *U* and *V* are orthogonal matrices. The parameters b_{i1}, b_{i2}, b_{i3} are set equal to the first, second and third column of *U* respectively, and the c_{t1}, c_{t2}, c_{t3} values are set equal to the product of the first, second and third column of *V* and the leading singular value d_1, d_2, d_3 respectively along with the normalizations given in (9). In order to make more accurate in the forecasting results, we adjust b_{i1}, b_{i2}, b_{i3} by comparison with an average of the last 12 months observation data.

2.5.3 Holt-Winters exponential smoothing method

Since the historical data series are seasonal with linear trend, we forecast c_{t1}, c_{t2}, c_{t3} values for up to 24 months ahead as well as their 95% prediction intervals (an estimate of an interval in which future observations will fall, with a certain

probability, given what has already been observed) by using Holt-Winters exponential smoothing with additive seasonality forecasting method. The Holt-Winters prediction function (for time series with period length p) is

$$\hat{C}_{t+h} = a_t + hb_t + s_{t+1+(h-1) \bmod p}, \quad (15)$$

where $\hat{C}_{t+h}$ is the forecast value at month $t+h$.

(a_t, b_t, s_t) are vectors containing the estimated values for the level, trend and seasonal components respectively, given by

$$a_t = \alpha (C_t - s_{t-p}) + (1-\alpha)(a_{t-1} + b_{t-1})$$

$$b_t = \beta (a_t - a_{t-1}) + (1-\beta) b_{t-1}$$

$$s_t = \gamma (C_t - a_t) + (1-\gamma) s_{t-p}$$

In order to study the forecast performance of Lee-Carter model, we compare the forecasting results using Lee-Carter model with the results from separate forecasts. For the separate forecasts, the sales in each branch location are forecasted separately using Holt-Winter exponential smoothing method.

2.5.4 Forecasting performance analysis

As a measure of fitting and forecast accuracy of each model and compare with separate forecasts by branch. MSE, MAD and MAPE can be computed from

$$MSE = \frac{\sum (error)^2}{n}, \quad MAD = \frac{\sum ABS(error)}{n}, \quad MAPE = \frac{\sum ABS(\%error)}{n} \quad (16)$$

MSE is used to measure variance of forecast error. MAD is used to measure average absolute deviation of forecast from actual. MAPE is used to measure absolute error as a percentage of the forecast and measures deviation as a percentage of actual data. The resulting value is multiplied by 100 to obtain the forecasting percentage error.