Chapter 2

Methodology

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

1. Computer programs
2. Methods for statistical analysis

2.1 Computer Programs

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

WebStat is a suite of functions written in HTML and VBScript for graphing and analysing statistical data stored in an SQL database. These programs use a web server. It was mainly used to perform preliminary data analysis and regression modelling for forecasting.

Microsoft Excel was mainly used to manage the data used for this research. Some functions are helpful in plotting graphs. *Microsoft Word* was mainly used to write and print the report of this research.

2.2 Statistical Methods

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 a 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.

One Way Analysis of Variance

One-way 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 the 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

$$F = \frac{(S_0 - S_1)/(c-1)}{S_1/(n-c)} \tag{2.1}$$

where

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

and

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

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*-statistics 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 adjusts 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 the confidence intervals for means are based on an estimate of the common standard deviation given by the formula

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

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 outcome.

In conventional statistical analysis the line fitted is the *least squares line*, which minimizes 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

$$y = a + bx, \tag{2.2}$$

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

After the usual assumption of independent observations is made, 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

$$y_j = a_j + bx, \tag{2.3}$$

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

Multiple linear regressions

If there is more than one determinant, the method generalises to multiple linear regression, in which the regression line extends to the multiple linear relation represented as

$$Y = \beta_0 + \sum \beta_i \chi_i + \varepsilon, \tag{2.4}$$

where Y is the outcome variable, β_0 is a constant, $\{\beta_i\}$ is a set of parameters (i = 1 to p), and $\{\chi_i\}$ is a set of determinants (i = 1 to p) (McNeil, 1998).

The model is fitted to data using least squares, which minimises the sum of squares of the residuals.

Multiple regression models on time series

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

There are four important objectives of time series analysis. There are (1) forecasting future value of 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 time series.

In our study the regression model changes to the form as follows

$$y_t = a_t + bt \tag{2.5}$$

where y_t is the transformed marine fish catch t months after December 1998 and a_t is a monthly seasonal effect.

Forecasting

Predictions of future events and conditions are called forecasts, and the act of making such predictions is called forecasting.

The use of causal forecasting models involves the identification of other variables that are related to the variable to be predicted. Once these related variables have been identified, a statistical model that describes the relationship between these variables and the variable to be forecasted is developed. The statistical relationship derived is then used to forecast the variable of interest.

When making forecasts after the data have been transformed, it is necessary to invert the transformation so that the forecasts are expressed in terms of the original data. For

the marine fish five types, the data were transformed using the square root, so the resulting forecast of this square root at time t is $\mathrm{E}[\sqrt{f_t}\,] = a_t + b\,t$, where a_t and b are estimates of the parameters in the regression model. The square root of the fish catch is assumed to be normal with estimated mean $a_t + b\,t$ and standard deviation s. Thus the forecast of the fish catch itself is

$$\begin{aligned} \mathrm{E}[f_t] &= \left(\mathrm{E}\left[\sqrt{f_t}\right]\right)^2, \\ &= \left(\mathrm{E}\left[\sqrt{f_t}\right]\right)^2 + \mathrm{var}\left[\sqrt{f_t}\right], \\ &= (a_t + bt)^2 + s^2. \end{aligned} \tag{2.6}$$

For the three shellfish types, where the natural logarithm transformation $\ln(f_t)$ is used, the forecast of the fish catch is $\mathrm{E}[\,f_t\,] = \mathrm{E}[\exp\{\ln(f_t)\}]$. In this case the natural logarithm of the catch is assumed to be normal with estimated mean $a_t + b\,t$ and standard deviation s. In this case we can work out the mean of the forecast by using the fact that if X has a normal distribution with mean μ and standard deviation σ, $\exp(X)$ has a lognormal distribution with mean $\exp(\mu + \sigma^2/2)$. Thus

$$\mathrm{E}[\,f_t\,] = \exp(\,a_t + b\,t + s^2/2\,)\,. \tag{2.7}$$