Chapter 3
Statistical Methods

Paul Taylor
University of Hertfordshire, United Kingdom

3.1. Introduction

This chapter describes a collection of statistical methods. The emphasis is upon what the methods are used to do and upon how to interpret the results. There is little on how to calculate the results, because the algorithms required have already been included in statistical packages for computers and this is how the calculations are performed in practice.

Examples, including computer output and my interpretation of what the output means, are given for some of the more widely used techniques. Other techniques are simply described, in terms of what they can be used to do, along with references to more detailed descriptions of these methods. Presenting examples for all the techniques would have made the chapter far too long.

Section 3.2 describes the most widely used statistical technique, namely regression analysis. Regression analysis is widely used, because there are so many statistical problems that can be presented as finding out how to predict the value of a variable from the values of other variables.

The techniques presented in Section 3.3 are regression analysis techniques for use in specific situations, which arise in practice and are not easy extensions of methods in Section 3.2. In fact, the techniques in Section 3.3.2 and Section 3.3.3 are part of an area of current research in statistical methodology.

Despite being around quite a long time (since the late 1970s or earlier, in most cases) the multivariate analysis techniques of Section 3.4 do not seem to be used as much as they might. When they are used, they are often used inappropriately. It seems likely that these techniques will start to be used more and more, because they are useful and the misconceptions about their use will gradually be eliminated. In particular, with the increase in automatic data collection, the

multivariate methods which aim to reduce the number of variables in a data set, discarding uninformative variables, ought to become more important.

3.2. Generalized Linear Models

The fitting of generalized linear models is currently the most frequently applied statistical technique. Generalized linear models are used to describe the relationship between the mean, sometimes called the *trend*, of one variable and the values taken by several other variables. Modelling this type of relationship is sometimes called *regression*. Regression, including alternatives to generalized linear modelling, is described in Section 3.2.1.

Fitting models is not the whole story in modelling; having fitted several plausible models to a set of data, we often want to select one of these models as being the most appropriate. An objective method for choosing between different models is called *analysis of variance*. Analysis of variance is presented in Section 3.2.2.

Within the generalized linear models, there is a subset of models called *linear models*. Sections 3.2.1 and 3.2.2 concentrate on linear models, because these are the most commonly used of the generalized linear models. Log-linear models and logistic regression models are two other heavily used types of generalized linear model. Section 3.2.3 describes log-linear modelling; Section 3.2.4 describes logistic regression.

Section 3.2.5 is about the analysis of survival data. The models used in the analysis of survival data are not generalized linear models. The reason that they are included here is that the techniques used for fitting generalized linear models can also be applied to the fitting of models in the analysis of survival data.

3.2.1 Regression

Regression analysis is the process of determining how a variable, y, is related to one, or more, other variables, $x_1, x_2, \ldots, x_n$. The y variable is usually called the *response* by statisticians; the x_i's are usually called the *regressors* or simply the *explanatory variables*. Some people call the response the *dependent variable* and the regressors the *independent variables*. People with an electrical-engineering background often refer to the response as the *output* and the regressors as the *inputs*. Here, we will use the terms output and inputs. Common reasons for doing a regression analysis include:

- the output is expensive to measure, but the inputs are not, and so cheap predictions of the output are sought;
- the values of the inputs are known earlier than the output is, and a working prediction of the output is required;
- we can control the values of the inputs, we believe there is a causal link between the inputs and the output, and so we want to know what values of the inputs should be chosen to obtain a particular target value for the output;