Multivariate Statistical Analysis: A Brief Introduction

K S Chia,* *FAMS, MD, MSc (OM)*

Introduction

In modern medical research, it is rare to limit data analysis to merely two variables: a single exposure (independent) variable and a single outcome (dependent) variable. Most outcomes are multifactorial and a string of exposure variables are needed to explore their relationships with one another and with the outcome variable. A cursory search of articles in the electronic BMJ from January 1996 to December 1998 yielded 2425 hits for the phrase "multivariate analysis".

Multivariate statistical analysis is designed to handle many independent variables (IVs) and several dependent variables (DVs) all interrelated with one another to a certain degree. For example, several tumour markers were monitored among patients undergoing two different chemotherapy regimes. Several other IVs such as age, gender, ethnicity and stage of disease also affect the tumour marker levels. Multivariate analysis can assist in answering questions like:

1. Which tumour marker show the greatest change after taking into account the various IVs?
2. How are the various tumour markers related to one another after taking into account the various IVs?
3. How do the IVs relate to one another?
4. Which combination of tumour markers best describe "tumour response" to treatment?

Types of Relationships

The simplest case for multivariate analysis consists of a single DV with two IVs. For example, the occurrence of ischaemic heart disease (IHD) is hypothesized to be related to smoking and serum levels of IgA antibodies to *Chlamydia pneumoniae* (IgA*cp*). Theoretically, there can be four different types of relationships:

1. Both IVs and the DV are not related to one another; the occurrence of IHD is not related to smoking or IgA*cp*.
2. Both IVs are *independently* related with the DV; smoking and IgA*cp* are not related to one another but each are related to CHD.
3. Both IVs *interact* with one another and affect the DV; the combined effects of smoking and IgA*cp* on CHD are multiplied compared with their simultaneous individual effect.
4. Both IVs *intermingle* with one another and affect the DV; the effect of smoking on CHD may be due to the effect of IgA*cp* and vice versa. An example of such intermingling is confounding.

It is important to identify all the possible relationships within a given dataset as this will provide a more complete picture of the "truth". However, multivariate statistical analysis will only reveal the probability of the possible relationships. Biological plausibility should be the overriding consideration when exploring relationships.

Strategies in Multivariate Statistical Analysis

The simplest approach to study the relationship between a single DV and two IVs is to construct a two-way cross-tabulation (Table I). This process of *stratification* will bring out the relationships between variables and allows a very clear and intuitive interpretation of the results. However, when the number of variables increases, the table becomes cumbersome and the number of subjects in each column decreases, leading to unstable summary indices (in this case, the proportion with various IgA*cp* levels).

TABLE I: IgA*cp* TITRE BY SMOKING AND IHD STATUS

	CHD patients		Non-CHD patients	
	Smokers	Nonsmokers	Smokers	Nonsmokers
Number	147	83	121	109
Antibody titre				
Zero	39 (26.5%)	17 (20.5%)	50 (41.3%)	47 (43.1%)
Trace	45 (30.6%)	29 (34.9%)	38 (31.4%)	30 (27.5%)
>1 in 16	63 (42.9%)	37 (44.6%)	33 (27.3%)	32 (29.4%)

* *Associate Professor*
Department of Community, Occupational and Family Medicine
National University of Singapore
Address for Reprints: Dr Chia Kee Seng, Department of Community, Occupational and Family Medicine, National University of Singapore, Lower Kent Ridge Road, Singapore 119260.

Traditionally, the stratification approach has been widely used by epidemiologists to adjust for the confounding effect of unwanted study variables. Direct standardisation has been used for rates and the Mantel-Haenszel method for odds ratios. Formulae for calculating the standard errors and consequently tests of statistical significance are easily available.[1]

A more efficient approach is to use mathematical modelling. The rationale is to summarise the relationship between the DV and IVs using a mathematical function (equation). The properties of this mathematical function can then be used to describe the various relationships. The mathematical function is the central figure and like a juggler, it has to balance the various combination of variables. Once the mathematical function is found, several important questions can be addressed:

Which IVs are Significant Predictors of the DV?

Most data sets from large scale studies will have information on several IVs. One of the aims of mathematical modelling is to identify the simplest set of IVs that will explain the variation in the DV. For example, in the study on IHD and IgA*cp*, a host of other dietary, anthropometric and biochemical data may have been collected. The aim is to identify, among these, a set of predictors and to rank them according to their contribution.

Is there Significant Interaction?

Within the mathematical model, meaningful interactions could be explored. The magnitude of such interactions could be evaluated within the model before tests of statistical significance are performed.

Is there Significant Interference, or Confounding?

The effect of confounding can be evaluated by comparing various indices before and after removing the confounding variable from the mathematical model. If the effect is large, its contribution can be reduced or even removed completely by retaining it in the mathematical model.

Standard Mathematical Models

It is clearly not possible to develop a new mathematical model for each combination of DV and IVs. Hence, there are standard mathematical models available whose properties are well known, thus making the computational task much simpler. However, by using standard models, instead of deriving the best mathematical model for the variables in the dataset, the process is reversed. As a result, the variables may not fit into the standard model. To improve the fit, variables are often transformed resulting in summary indices that are not easily interpreted.

In medical research, there are a handful of commonly used mathematical models. The choice depends on the type of DVs and the study design used. For DVs that are measurement variables such as creatinine clearance, the multiple linear regression model is commonly used.[2] In case-control studies where the DV is dichotomous (e.g. presence or absence of IHD), multiple logistic regression model can be used.[3] In an open cohort study, where the DVs are either "survival" time or incidence density rates, Cox's proportional hazards model is often used.[4] Cox's model can be further modified for special situations such as closed cohort studies[5] and cross-sectional studies.[6,7] Less commonly-used models include cumulative logit regression model,[8] nonlinear regression models[9] and the generalised estimating equations.[10]

A Cautionary Note

Mathematical modelling is a very powerful tool for multivariate statistical analysis. However, the choice of mathematical model is crucial and since most researchers will adopt one of the standard models, evaluating how well the data fit the model is a vital step in the multivariate analysis. Finally, with powerful computer programs available, it is very tempting to introduce a long list of IVs derived from a small number of subjects. In the book "Intuitive Biostatistics", Motulsky[11] stated repeatedly the rule of thumb: "there should be 5 to 10 events for every X variable"; in other words, 5 to 10 cases for every IV in case-control studies and 5 to 10 deaths for survival analysis.

REFERENCES

1. Kahn H A. Adjustment of data without the use of multivariate models. In: Kahn H A, editor. An Introduction to Epidemiologic Methods. Chap. 5. NY: Oxford University Press, 1983:63-99.
2. Chia K S, Jeyaratnam J, Lee J, Tan C, Ong H Y, Ong C N, et al. Lead induced nephropathy: relationship between various biological exposure indices and early markers of tubular dysfunction. Am J Ind Med 1995; 27:883-95.
3. Lee H P, Gourley L, Duffy S W, Esteve J, Lee J, Day N E. Dietary effects of breast cancer risks in Singapore. Lancet 1991; 337:1197-200.
4. Cox D R. Regression models and life-tables (with discussion). J R Stat Soc B 1972; 34:926-35.
5. Breslow N E. Covariance analysis of censored survival data. Biometrics 1974; 30:89-99.
6. Lee J, Chia K S. Estimation of prevalence rate ratios for cross-sectional data: an example in occupational epidemiology. Br J Ind Med 1993; 50:861-2.
7. Axelson O, Fredriksson M, Ekberg K. Use of prevalence ratio vs. the prevalence odd's ratio as a measure of risk in cross-sectional studies. Br J Ind Med 1994; 51:574.
8. Lee J. Cumulative logit modeling for ordinal response variables: applications to biomedical research. Comput Appl Biosci 1992; 8:555-62.
9. Gallant A R. Nonlinear regression. Am Stat 1975; 29:73-81.
10. Zeger S L, Liang K Y. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 1986; 42:121-30.
11. Motulsky H. Intuitive Biostatistics. NY: Oxford University Press, 1995.