Review article: Research methods in epidemiology, I

Uses and abuses of multivariate methods in epidemiology

STEPHEN J W EVANS

From the Department of Clinical Epidemiology, The London Hospital Medical College, Turner Street, London E1

Uses

The general medical literature is replete with examples of so called "multivariate methods". These include multiple regression, logistic regression, discriminant analysis, "Cox regression" and log linear models. They are different types of generalisation of the method of simple linear regression, the object being to obtain a description of a biological phenomenon, using an equation. This in its simplest form is $y = a + bx$; y is the outcome (dependent) variable and x is a predictor (independent) or explanatory variable. The complex calculations which are required when using these methods are easily carried out using computers and the computer software for them is widely available.

Multiple regression is the generalisation to more than one predictor variable; eg, $y = a + b_1x_1 + b_2x_2$. This equation should be as simple as possible while still providing an adequate description, so that extra terms are only added if they provide an important gain. The x variables which are not of particular interest in themselves are often referred to as *covariates*. If y is blood pressure and x_1, x_2 are height and weight, then the equation will provide a much better description than either equation for blood pressure involving height or weight alone. Blood pressure is more dependent on how *overweight* (for their height) someone is, than on their weight alone. The relationship of blood pressure with height may be nearly non-existent (b_1 could be positive, negative or zero) but if weight is in the equation the value of b_1 , the regression coefficient for height, will very definitely be negative (that is, for a given weight, the taller someone

is, the *lower* their blood pressure). This fact that the value of the regression coefficient changes when other variables are in the equation illustrates both the major *use* and a major *abuse* of multivariate methods.

All multivariate methods have this use that allows for the effect of a particular variable to be adjusted for the effect of other measured variables. The classical problem of confounding can sometimes have a solution in the use of multivariate methods. For example, to compare the blood pressure in two different racial groups whose height and weight differ, then multiple regression may be used. This will allow the comparison between the groups to be made, uncomplicated by any differences between the groups in height or weight. Any confounding effect of these variables has been removed. In addition, multiple regression may be used to see if the relationship between height, weight and blood pressure is the same for each racial group.¹

The model as described is a linear model in two senses. The first is that the effect on blood pressure of a unit change in, say, height is the same at all heights; a straight line relationship is assumed. A curvilinear relationship could be obtained by including as a further predictor some particular x variable (such as height) as both x and x^2 . Part of the art (for it remains an art) of fitting equations to data is in choosing appropriate types of curve to fit the data. The second sense is that even an equation with both an x and x^2 in it such as

$$y = a + b_1x_1 + b_2x_2 + b_3x_1^2$$

is still described as a *linear* function in that there is addition of the terms rather than multiplication of

them. The word linear is here used in a technical mathematical sense, but this is the sense of the word when used in a general statistical context. Some statisticians reserve the word "multivariate" for analysis of multiple outcome variables (several different y variates) and use "multivariable" for multiple regression and the other methods discussed here, and it may be useful to distinguish them.

Logistic regression is a form of multiple regression where the outcome studied is not a continuous measure like blood pressure, but is instead a binary variable (yes/no, dead/alive, etc). The obvious utility of such a method in medical research is clear. The observations of a binary variable may be coded as 0 and 1—for example 0 = alive, 1 = dead. Although any particular observation will always be 0 or 1, we can talk about P , the probability of death for an individual or a group of people. The range of P is from 0 to 1 while $\frac{P}{1-P}$ —the odds of death—has a range from 0 to

infinity. The logarithm of this ratio—the log odds or logit (P)—has a range from $-\infty$ to $+\infty$, making it suitable for use in a regression equation. Studies of the Framingham type² have used logistic regression to examine many risk factors in predicting death from coronary heart disease. Measures of morbidity such as respiratory symptoms and illnesses (coded as present/absent) can also be used.³ The odds ratio which is obtained from the logistic model provides a convenient and comprehensible summary of the effect of explanatory variables. Recent applications of this method to the analysis of case-control studies have been particularly useful. The outcome variable here is whether an individual is a case or a control. The odds ratios obtained from the logistic regression may, in the situation of a rare disease, be interpreted as relative risks.⁴ Care must be taken in the analysis since different forms of logistic regression should be used depending on whether the study design was matched or not.

Discriminant analysis and logistic regression are closely related and in many instances may be used interchangeably, their differences being in the technical form of the equations used. Discriminant analysis is classically used to derive an equation which will be the best one to predict membership of one of two groups. Group membership may be treated in this case as a binary outcome variable Y (coded as 0 for one group and 1 for the other group). The observed average of Y for a particular combination of x values can be seen as the probability of being in the second group. The discriminant function then becomes essentially identical to ordinary multiple regression. When Y is in the range 0.3 to 0.7, logit (Y) and Y are approximately linearly related and so logistic

regression and discriminant analysis give similar results. It should be noted that using ordinary regression can give *predicted* values of Y outside the range 0 to 1 which is not possible with logistic regression. The technique then allows for discrimination between the groups. There are detailed technical discussions of the issues of which method to use with two groups.⁵ As a general rule, with two groups and where some of the x variables are categorical rather than continuous, or where the probability Y can be outside the range 0.3 to 0.7, logistic regression (sometimes called logistic discrimination in that context) is to be preferred.⁶ Discriminant analysis may also be used to examine a set of equations which use several predictor variables to classify a variable on a nominal scale (unordered categories) having more than two categories. For example, in a study to distinguish three diagnostic groups such as motor neurone disease, stroke and multiple sclerosis, the variables used to distinguish between the groups can themselves be binary or continuous, and a nominal variable with n categories can be re-expressed as $n-1$ binary variables. A scoring system can then be used to classify cases on the basis of symptoms etc.⁷

Cox Regression or proportional hazards regression is a form of logistic regression suitable for survival analysis where the time to occurrence (survival time) of an event (usually death) is of primary importance. It has the advantage of being able to deal with varying amounts of follow up in a natural way and has provided a very important tool in the study of survival time. Since it is a type of regression analysis then all the issues of ordinary multiple regression can apply here. Explanatory variables, whether of primary interest or potential confounders, may be incorporated in the model. They are assumed to have an effect which multiplies the death rate by a certain amount (the equivalent of the slope, or regression coefficient in ordinary multiple regression). This effect may be estimated together with its standard error which allows for calculation of confidence intervals and statistical significance testing (p values) in the usual way. The "Log-Rank" test is a very similar, though less general, technique. The Cox model also has applications in the analysis of case-control studies, though this is a very specialised area. The results from such an analysis may lead to an estimate of relative risk, with standard errors.

Log-linear models are methods for dealing with both unordered and ordered categorical data. They are a categorical analogue of correlation in the way that logistic regression is an analogue of regression. They are mainly used for looking at the inter-relationships of several variables without necessarily concentrating on a single outcome variable. The method is applied to

multi-dimensional tables and ought to be considered in place of standardising mortality rates in many situations.

These five methods detailed above can all be put together as examples of "Generalised Linear Models" and have an overall structure which is similar. There are differences between the methods in the type of equation used but they have an underlying commonality of approach.

For all of the methods their appropriateness may be seen to depend on three things: (1) The variability (called by statisticians the *error* term) which remains in the data when the equation has been used is specified and allowed for correctly; (2) The *form* of the equation relating the variables to each other is correct (this relates strongly to the assumptions made in using the equation); and (3) The *relevant* variables have been measured over a range of medical interest.

Abuses

ERROR

In elementary statistics much emphasis is placed on this factor. In regression (or *t* tests which can be seen as a form of regression) everyone is worried about whether the distribution of the error is Gaussian (Normal). With multivariate methods a rather more restrictive assumption of *multivariate normality* is made and this is rarer to achieve as well as harder to test whether it is present. In practice most methods are reasonably robust to departures from "normality", but occasionally a single, or a few, observation(s) distort the distribution. These "awkward" observations are more likely to be present if many variables are measured and analysed. Recent statistical attention has been paid to methods of detecting *influential* observations which may or may not negate the usual distributional assumptions but which affect the magnitude and statistical significance of the effect of a variable. The better computer programs provide good facilities for testing assumptions and examining the effect of influential observations which increase in their adverse effects in multiple regression. Results from multiple regression analysis should not be published without account being taken of problems in this area.

FORM OF THE EQUATION

Epidemiologists have become aware of the need to allow for different forms of equation being used to describe the same set of data, such as multiplicative or additive models. The fact that a multiplicative model ($y = ax_1^{b_1}x_2^{b_2}$) provides a "statistically significant" equation does not demonstrate unequivocally that an additive model is inappropriate. Similarly (and

equivalently in some cases) to assume that all variables are independent of each other in their effects (that is, there are no *interactions*) may also be naive.

While it may be parsimonious to use a linear additive model with no interactions and sophisticated to use a non-linear model with interactions and a complex error structure, it is more important to be aware of limitations of equations in describing complex reality. The ability of computers to provide good graphical displays has only just begun to be realised and is at a very elementary stage in multivariate analysis. The temptation to use a new method because it is there rather than checking whether it is appropriate has been particularly prevalent in survival analysis. As noted above the "Cox model" is called "proportional hazards": that is, it assumes that the instantaneous death rate (hazard) is multiplied by a certain factor for each variable, but this assumption may not be true and checking it, again using graph plots, is vital.

As yet, no formal mechanism exists for reaching the correct conclusions regarding *error* terms or the *form* of equations. Various attempts to do so automatically (such as stepwise regression) may provide some aid but they can be misleading. The consequence is that analysis requires time and care and is best done by those experienced in handling real data. Some guidelines have been suggested in the context of Cox regression⁸ but, while useful for all multivariable problems, they do not guarantee that the correct answer is obtained.

RELEVANCE

Even experienced statistical analysts may have weaknesses in this area. The story is told of the statistician who presented the results of a multiple regression analysis of a chemical process by saying that water had no effect on yield. The chemists fell about laughing because the concentration of water was controlled to be less than 1 part in a million and varied over a small range below this, water in substantial amounts having caused an explosive effect on yield! Even if a relevant variable is included it must neither be swamped by measurement error nor have an inadequate range of measurement over a biologically relevant "dose" range.

The effect of *unmeasured* variables is unpredictable. Their absence can lead to totally wrong conclusions with regard to the effect of measured variables, and when contradictory results are found in different studies it may well be that none have measured an important variable. The art of epidemiology partly consists of thinking of such variables and of ways to measure them.

Multivariate methods can be very helpful to the epidemiologist and clinical researcher but like any

powerful tool can be most dangerous when used by those of "a little learning".

COMPUTER PROGRAMS

The range of computer programs is very wide indeed and these comments only refer to some of the most well known ones.

Multiple regression—Virtually all statistical computer programs provide for ordinary multiple regression. Many of the better ones, including BMDP, SAS and SPSS (on both mainframe and micro computers) provide useful facilities for checking the assumptions and drawing attention to influential observations. They also provide various methods to build up a model by adding or deleting variables from the equation according to the statistical significance of that variable. These cut down the amount of computer time and number of attempts to find an acceptable model but they do not necessarily obtain the most sensible model automatically.⁸

A particularly useful program in trying different models is BMDP-9R which fits all possible models for a given set of variables and suggests a "best" model according to a statistical criterion. This is a compromise between a model which includes all possible variables which will always be the "best" model in terms of apparently explaining most of the variation, and the simplest model which will include only the single most significant predictor. There is a "penalty" for adding variables so that they are only included if they offer a substantial gain. In spite of the appeal of such a method there is no guarantee of the final "best" model being the "real" model in a scientific sense. BMDP-9R is my best buy program for multiple regression for epidemiologists.

Logistic regression—BMDP-LR and GLIM (on mainframe and micro) have always provided good facilities for logistic regression, with the former having clearer output and a stepping facility, while the latter has great flexibility but output which is only comprehensible to the expert. SAS on mainframe, and in the latest (6.03) version on micro computers, also provides facilities for doing logistic regression with output and flexibility that is intermediate between BMDP and GLIM. SPSS-X on the mainframe (not SPSS-PC) does have a facility for logistic regression but is not as good as the others. A more recent package with good facilities for logistic regression (in both matched and unmatched designs) is EGRET from Seattle. This is my best buy for epidemiologists, though BMDP and SAS are also good.

Discriminant analysis—This technique is less likely to be used by epidemiologists but if it is, BMDP, SAS and SPSS all provide for it with BMDP-7M (best buy) being particularly useful as it allows for a "jack-knife"

classification. In other words, in deciding the success of a model it classifies each case using an equation derived from all the other cases but excluding this one. This leads to a less biased assessment of the success rate of the classification. A better approach, which is applicable to all the multivariate methods (as in fact is jack-knifing), is to use part (half, typically) of the data available as a "training" set to derive the equation and the other part as a "test" set to see how successful the derived equation is for prediction.

Cox regression—BMDP, SAS (not in PC version) and EGRET all provide good facilities for doing proportional hazards regression, with EGRET and BMDP-2L providing joint best buys. SPSS-X does provide some survival analysis on the mainframe but this is weak and non-existent on the PC version at the time of writing (August 1988). GLIM is of course able to do it but only with difficulty!

Log-linear models—All the packages provide a form of log-linear modelling, with BMDP-4F being easiest to use and GLIM giving the greatest flexibility. EGRET provides very good analysis of several 2x2 tables with Mantel-Haenszel statistics, and for that type of analysis is a best buy, but BMDP is more general.

Summary—EGRET and SPSS/PC have the most "user friendly" interface, though EGRET is not at all good at data management and transformations, which are quite good in SPSS/PC. SAS/PC is even better in this area but is slightly more tedious to use, though it is more similar to the mainframe version than is SPSS/PC. BMD-P is virtually identical on micro and mainframe, which is both a strength and also a major disadvantage in that it is very user unfriendly. This is a pity since it provides such a wide range of good facilities. GLIM can do virtually everything but requires quite a bit of work "on the side" to provide usable output.

References

- 1 Silman AJ, Evans SJW, Loysen E. Blood pressure and migration: a study of Bengali immigrants in East London. *J Epidemiol Community Health* 1987; **41**: 152-5.
- 2 Leaverton PE, Sorlie JC, Kleinman JC, et al. Representativeness of the Framingham risk model for coronary heart disease mortality: A comparison with a national cohort study. *J Chron Dis* 1987; **40**: 775-84.
- 3 Somerville SM, Roma RJ, Chinn S. Passive smoking and respiratory conditions in primary school children. *J Epidemiol Community Health* 1988; **42**: 105-10.
- 4 Notani PN. Role of alcohol in cancers of the upper alimentary tract: use of models in risk assessment. *J Epidemiol Community Health* 1988; **42**: 187-92.
- 5 Titterton DM, Murray GD, Murray LS, et al. Comparison of discrimination techniques applied to a complex set of head injured patients. *J Roy Statistical Soc A* 1981; **144**: 145-75.

⁶ Alexander FE, Roberts MM, Huggins A, Muir B. Use of risk factors to allocate schedules for breast cancer screening. *J Epidemiol Community Health* 1988; **42**: 193-9.

⁷ Li TM, Day SJ, Alberman E, Swash M. Differential diagnosis of motorneurone disease from other neurological conditions. *Lancet* 1986; **ii**: 731-3.

⁸ Lew RA, Day CL, Harrist TJ, Wood WC, Mihm MC. Multivariate analysis. Some guidelines for physicians. *JAMA* 1983; **249**: 641-3.

Accepted for publication August 1988

Notice of Meetings

International Conference on Ionising Radiation and Cancer Epidemiology, University of Birmingham, 12-13 July 1989. Further particulars from: Dr Tom Sorahan, Department of Social Medicine, University of Birmingham, Edgbaston, Birmingham B15 2TJ.

There will be an International Conference on Community Nursing organised by the Netherlands Institute of Primary Care on 16-17 March 1989. The conference will take place at the Casino Den Bosch in 's-Hertogenbosch (The Netherlands). Further information from: NIVEL, Mrs A Kerkstra PhD, PO Box 1568, 3500 BN Utrecht, The Netherlands.

The XIIth International Conference on Preventive and Social Medicine will take place on August 13-16,

1986, in Montreal. The theme will be "Nutrition in Health-The Scientific Challenge for the 21st Century". The conference chairman is Dr Josef Vobecky.

For further information write to the Conference Secretariat, The Columbia/Kenness Team, 1010 Ste. Catherine St. West, Suite 645, Montreal, Quebec H3B 1G7, Canada.

The 1st European Conference on Health Economics will take place on September 19-21 1989 in Barcelona. The theme will be "Incentives in Health Systems." Abstracts are required by February 1. For further information write to Dr Guillem López Casasnovas, Asociación de Economía de la Salud, Plaça Catalunya 9, 4^a planta, 08002 Barcelona, Spain.