

*Editorial*

HAZARDS IN THE USE OF THE LOGISTIC FUNCTION  
WITH SPECIAL REFERENCE TO DATA FROM  
PROSPECTIVE CARDIOVASCULAR STUDIES

(Received 9 July 1973)

SUPPOSE we are interested in examining the relation of CHD (coronary heart disease) incidence to the level of systolic blood pressure. The most direct approach is to measure blood pressure in a population of interest, group individuals of nearly similar systolic pressures, and observe the subsequent CHD incidence (within a specified period of time) in each of these groups. If we do this, we are likely to find that groups of persons with higher pressures tend to have higher incidence but that this is not a completely smooth trend. As the number of persons under observation increases, the trend can be expected to become smoother. Where it is not possible to sufficiently increase the number of observations to accomplish the desired amount of smoothing, it is sometimes appropriate to smooth the observations by some form of statistical graduation.

A common method of graduation with data from prospective cardiovascular studies is the logistic function. In the univariate case, each person in the sample of individuals under study is estimated to have a probability of the event expressed as  $y(x) = [1 + e^{-(a+bx)}]^{-1}$ . In the instance cited,  $x$  is a specific systolic pressure at baseline and  $y(x)$  is the probability that a person will develop CHD during a specified duration of follow-up given that he has the pressure. To graduate the data using this function, it is necessary to estimate only two parameters. Very frequently the smoothed results seem quite consistent with the unsmoothed rates obtained from categorical analysis.

There are two methods in use for estimating the logistic parameters. One, proposed by Cornfield [1], yields explicit solutions by a minor but ingenious modification of linear discriminant analysis [2]. The other, proposed by Walker and Duncan [3], is an iterative procedure involving less restrictive assumptions than the discriminant-based estimates. Since it is a least-squares estimate, by that criterion it provides better estimates. However, it is computationally more expensive and sometimes it will not converge to a solution.

Conceptually, the two estimating procedures involve different rationales. Discriminant analysis assumes that there are two different populations, one sick, the other well. As formulated by Cornfield, these two populations are assumed to be normal, with equal variances (or, in the multivariate case a common variance—

covariance matrix [4]) but different means. Even where these assumptions are obviously incorrect, the graduation usually fits the data fairly well. Sometimes, and especially when the independent variable is discontinuous, however, the fit is very unsatisfactory. What is worse, there appears to be no method to predict when a very bad fit will occur [5].

While any procedure leading to a bad fit is inherently suspect, it must be remembered that the analytical use of this function, particularly in the multivariate case, can logically be regarded as a form of discrimination and the robustness of this procedure is well-demonstrated. Moreover, it is our own experience that the discriminant-based estimates and Walker–Duncan ones may yield similar tests of significance for the parametric estimates, even where the estimates themselves are quite different.

The maximum likelihood estimation, on the other hand, is based on the concept of a dose-response. The higher the blood pressure (dose), for instance, the greater the CHD incidence (response). The Walker–Duncan procedure yields satisfactory fits, even where the independent variable is discontinuous. Moreover, the total estimated incidence (the sum of  $y(x)$  for all  $x$  in the sample) is constrained to the total actual incidence. In discriminant-based estimates this is not true, and the resultant estimates are occasionally very bizarre; for example, the estimated number of cases may be far in excess of the actual number of cases. In addition, discriminant-based procedures seem to have a slight tendency to overestimate the higher conditional probabilities. But these are exceptions: discriminant-based estimates generally provide quite satisfactory graduation.

If the logistic function were applied only to the univariate case, it would be of minor interest. So long as the incidence is fairly low (which is usually the case) reasonable graduation over the central range of the independent variable can be achieved by a variety of methods including unweighted linear regression. The chief value of the logistic function in the univariate case lies in providing an estimate of the gradient of incidence on the independent variable and an estimate of the standard error of that estimate. This leads to a test of significance based on the full use of the available detail.

It is the multivariate case where the logistic function has yielded a rich bonus in the analysis of cardiovascular data. The reasons for this are obvious. If continuous variables are partitioned along their scale and several such variables are cross-classified, the number of cells in the resultant table quickly multiply. For example, 5 continuous variables divided into thirds along their range yields a table with 243 cells. Not only is it rare to have enough incidence cases to occupy that many different cells, it is rare to have populations large enough for that purpose. Moreover, where there is a distinct gradient of risk along a variable, a very broad class will include individuals with quite distinct conditional probabilities of the event.

The practical and theoretical limitations to the use of cross-classification in evaluating relationships of this sort has made a multiple logistic model exceedingly attractive. This model is constructed in the same form as the univariate:  $y(x) = [1 + e^{-(a + \sum b_i X_i)}]^{-1}$ , where  $i$  is the index for the  $n$  independent variables entered into the calculation. Instead of calculating the incidence rates for a large number of cells and attempting to see the picture, it becomes possible to graduate the data by estimating the limited number of parameters in the multiple logistic function. If the shape of the  $n$ -dimensional curve has been correctly judged to be logistic, we have achieved not only a good fit of the data, but are in a position to estimate the confidence we can place in this fit.

As in linear regression, a graduation to one set of data cannot be expected to fit another set of data as well. Thus, the ability to predict CHD from a multivariate logistic function (which in the jargon of cardiovascular epidemiology has come to be called a 'risk function') is not usually as good as the fit to the original data would suggest. At the same time, there is a countervailing consideration, namely, the biasing effect of error in the measurement of the independent variable. In the univariate case, using the discriminant-based estimates, technical variation independent of other sources of variation will inevitably lead to an understatement of the expected magnitude of the regression coefficient. The argument for this conclusion is identical with that for linear regression [6].

The synthetic uses of the logistic function, substituting a combination of observable variables (say, blood pressure and serum cholesterol levels) for an unobservable variable (say, the subsequent development of CHD), are fairly clearcut. When such uses are being considered, the statistical manipulations can be quite mechanical and still be useful. Arbitrary step-down or step-up methods may be used to select an efficient subset of variables for identifying 'high risk' individuals (that is, persons whose conditional probabilities are higher than average). Considerations of cost and convenience can be entered into the calculation either formally or informally. The multiple logistic function has proved very useful for cardiovascular screening and will probably be used increasingly for that purpose [7].

However, it is probably its analytical value that has most attracted investigators. Statistical analysis, in the context of the present discussion, can be defined as a process of disentangling the contributions of a set of variables to some outcome. It is imbued with semantic confusion between subject-matter logic and statistical logic. Even its primary statistical aspects have numerous hazards.

The model assumptions themselves are a source of hazard. The logistic model we have used assumes a linear combination of the independent variables. This assumption is not always warranted. For example, the logistic regression of CHD on serum cholesterol or cigarette smoking decreases with age and is different for men and women. In statistical terms, there are interactions among these four independent variables. To enter the independent variables as linear terms into a multiple logistic function, ignoring this interaction, misrepresents the data. The safest way to explore questions of this sort is to revert to cross-classification.

Interaction is, of course, one of the central concerns of statistical analysis. Consider a specific example. Two of the most important precursors to congestive heart failure (CHF) are CHD and hypertension. We might, conceivably, construct a multiple logistic function with CHD (a dichotomous variable) and age and systolic blood pressure (continuous variables) as the independent variables and the probability of CHF as the dependent variable. Suppose, however, that the onset of CHD alters the relation of both blood pressure and age to CHF. The multivariate logistic function proposed may produce a very powerful synthetic 'predictor' but lead to completely misleading analytical conclusions.

The alternative, in this case, is very simple; namely to estimate separate logistic functions for persons with and without CHD. This partly categorical analysis is feasible because nearly half of the CHF incidence is preceded by CHD. Where data are too skimpy for such a categorical approach, we must forbear drawing analytical conclusions based on the apparent general adequacy of the fit. Ironically, then, while the

multiple logistic function is used to repair the inadequacies of analysis by cross-classification, cross-classification is the method of choice for exploring the inadequacies of the multiple logistic function.

Where the interaction can be represented by some simple algebraic device (say, a cross-product term), it may be helpful to enter such a term into the exponent. Of course, when the decision to use a term is based on a prior evaluation of the sample data, it becomes difficult to evaluate the test of significance for this term, but this is really of less concern than the appropriateness of the graduation. However, interactions may be too complex, particularly when the number of variables is fairly large, to be described adequately by a single multiple logistic function. It might also be noted that discriminant-based estimation is not designed to represent non-linear terms. Efforts to use it for that purpose are apt to lead to nonsensical results in cases where the results of Walker–Duncan estimations are quite reasonable.

Another difficult fitting problem arises where the conditional probability is not monotonic in the independent variable. If, for example, incidence is high both for low body weights and high body weights, and low for intermediate weights, the logistic function may not provide a good fit to the data. In general, adding a quadratic term to the exponent is to be avoided and alternative methods of exploring such relationships should be sought.

In brief, the multiple logistic operates satisfactorily if the assumption of a linear exponent is more or less correct. Under these conditions, a modest amount of data can be graduated with considerable assurance. In such a case, however (as in other multivariate analysis), it is practically impossible to test the appropriateness of the assumptions. In effect, we must rely on other evidence (or our hopes) for assurance that the procedures used are relevant. By the same token, where there are relatively few observations of either the independent variable or the dependent, at very high or very low blood pressures, for example, the assumption that the graduation holds has no means of confirmation within the sample data.

Frequently, the analytical interest focuses on a single variable and the concern is to make due allowance for other associated variables in evaluating the one variable of interest. Suppose we are interested in the relation of diabetes to the incidence of congestive heart failure. We know that diabetes is correlated with blood pressure, which in its turn is strongly associated with the incidence of congestive heart failure. Hence, the use of the multivariate technique to control the nuisance variable (in this instance, blood pressure) in order to assess the net effect of diabetes. Whether it is sensible to approach the analysis this way must be based on subject matter considerations. To the extent that blood pressure is controlled by glucose tolerance, the question may not be definable at all.

To take another case in point. Suppose serum cholesterol, blood pressure and relative weight are entered as independent variables in a multiple logistic function with CHD as the dependent variable. It is most likely that the coefficient for relative weight, which may be statistically significant by itself, will approach zero in this multivariate context. What does that mean? It certainly does not *per se* mean that relative weight is unimportant. It may rather mean that its effect is intermediated by blood pressure and serum cholesterol levels. In any case, the question is a subject-matter one, not a statistical one.

A similar difficulty arises when two highly correlated variables are included; for example, both systolic and diastolic pressures. When this is done, the calculations become very unstable and may, in fact, be uninterpretable. If the multivariate function is used synthetically, this is no problem. If the interest is analytic, it is ordinarily best to delete one of these twins before proceeding or to repeat the analysis first with one, and then the other. Similar problems sometimes arise when interaction terms, which are bound to be highly correlated with the parent linear terms, are added to the function.

What is more, the relation of a variable to some outcome depends on the other independent variables included in the set of characteristics under consideration. It is not at all uncommon to have a coefficient which is significant in the univariate case become indistinguishable from zero in the multivariate case. The same thing may happen when we shift from one multivariate set of variables to another. Nor is the effect of adding a variable to the set always to reduce the coefficients assigned to the original variables. Thus, there is never a unique multivariate conclusion. The only general guide is that where a conclusion respecting association of a variable in the multivariate case is not consistent with what is found in the univariate case, great care must be exercised in drawing conclusions.

The specificity of multivariate analysis makes any process of looking for an optimum subset of variables very chancy. If a synthetic function is being sought, this is not a serious problem. If the function is being used analytically to discern which variables are important, this is a most serious reservation.

Persons familiar with multiple linear regression will recognize in slightly altered form many of the same cautions and concerns. That being the case one final caution might be worth making and that is that analogies between the two should not be pushed to an extreme. For example, since the dependent variable is dichotomous, the testing of goodness of fit is not entirely straightforward and can only be approximated by breaking the estimated function along its range and comparing the actual number of cases in each class with the sum of the estimated probabilities. The analog to variance analysis which has been proposed to answer the same question is very tenuous at best.

It must be said, finally, that no method of analysis can redress a shortage of information from prospective studies. It is, of course, this very shortage which provides the most powerful motive for using (and sometimes misusing) the multivariate logistic function. It is the shortage of information that impels analysts, for example, to combine data for all age groups and both sexes into one grand analysis, often without even a preliminary exploration of the problems this may entail. The power and elegance of the logistic function make it an attractive and flexible statistical instrument, but in the end, we cannot push a button and hope that everything will come out all right. Because frequently, it will not.

TAVIA GORDON

*Biometrics Research Branch,  
National Heart and Lung Institute,  
National Institutes of Health,  
Bethesda, Maryland, U.S.A.*

#### REFERENCES

1. Cornfield J, Gordon T, Smith WW: Quantal response curves for experimentally uncontrolled variables. *Bull Int Sta Inst XXXVIII: Part III, 97-115, 1961*

2. Fisher RA: The use of multiple measurements in taxonomic problems. *Ann Eug Lond* 7: 179–188, 1936
3. Walker SH, Duncan DB: Estimation of the probability of an event as a function of several independent variables. *Biometrics* 54: 167–179, 1967
4. Truett J, Cornfield J, Kannel W: A multivariate analysis of the risk of coronary heart disease in Framingham. *J Chron Dis* 20: 511–524, 1967
5. Halperin M, Blackwelder W, Verter J: Estimation of the multivariate risk function: a comparison of the discriminant function and maximum likelihood approaches. *J Chron Dis* 24: 125–158, 1971
6. McNemar Q: *Psychological Statistics*. New York, Wiley, 1949, pp. 134–136
7. Gordon T, Kannel WB: Multiple contributors to coronary risk: Implications for screening and prevention. *J Chron Dis* 25: 561–565, 1972