Logistic regression—explanation and use

ABSTRACT—The aim of this article is to provide a simple explanation of the logistic regression process and a guide of what to look for when assessing a study involving logistic regression.

Medical outcomes are often influenced by many factors, the individual and joint contributions of which need to be evaluated. The statistical techniques of multivariate analysis are increasingly utilised for this purpose and important conclusions have been drawn from data analysed by logistic regression: for example, the routine use of intramuscular vitamin K in neonates has been questioned because of a possible carcinogenic effect thought to be separate from that of other risk factors [1]. The validity of the method is therefore a matter of proper concern for all doctors. A survey in our district general hospital (n = 278; 171 respondents) indicated that only 6% of doctors (95% CI: 2.8–10.6%) claimed to have a reasonable understanding of logistic regression, and only 5% (CI: 2.4–9.1%) gave correct answers to two simple questions about the method. There are many excellent accounts of logistic regression [2–5], but they have a statistical or mathematical emphasis which makes them difficult to comprehend. It is hoped that the following explanation will enable doctors to gain some insight into the basic ideas of logistic regression and assist interpretation of the results.

The logistic regression model

Suppose we are interested in a particular outcome: for example, the development of coronary heart disease. Previous research has shown that many factors appear to influence this outcome, including smoking, serum cholesterol and blood pressure. The effect of any one of these could be determined by comparing the mortality rates for those people with the risk factor to the rates for those without it. The ratio of the two rates is called the *relative risk*. If this is greater than one, it indicates that the people with the risk factor have a higher mortality than those who do not. Ideally, these rates should be obtained from large random samples so that the other factors which may affect the results are evenly distributed within the two groups—otherwise the effect of, say, smoking may be exaggerated or diminished if the proportions of those with high blood pressure are different in the smoking and non-smoking groups (*confounding*). To allow for this, it is possible to split the groups further and consider separately smokers with high blood pressure, smokers with normal blood pressure, and so on (*stratification*). As the number of factors increases, the number of combinations grows rapidly (2^n for n factors), so even with very large data sets the number in each possible subset can become too small to provide statistically significant results.

To avoid this difficulty, the methods of multivariate analysis have been employed [6]. In general, a mathematical model is formulated which tries to describe the data set. Logistic regression is one example of such a model, dealing particularly with yes/no outcomes, such as dead or alive, disease absent or present. As in all models, certain assumptions are made to fit the model to the data. It is suggested that if the effects of each variable could be isolated from those of the others, the overall effect on the outcome could be expressed as an algebraic sum. The chance of the outcome being favourable would depend on the findings. This 'chance' may be represented by probability (lying between 0 and 1) or by the odds. The *odds* are the ratio of the probability of the event occurring to the probability of it not occurring. For example, if a horse has a probability of 0.8 of winning a race, the odds of it winning are 0.8/(1–0.8) = 4:1. In fact, ratios of odds are frequently used in logistic regression presentations, for example, the odds ratio for smokers versus non-smokers in developing lung cancer is the ratio of the odds of smokers getting it to the odds for non-smokers. The advantage of odds ratios is that they approximate relative risks when the disease state is rare and are the same for population or case-control studies.

In the logistic regression model, the relationship between the outcome and the variables thought to affect the outcome is expressed as a simple equation. The relative importance of each variable is determined by weighting factors or coefficients. A constant term must be included because all the factors influencing the outcome may not have been identified. When more than one variable is present, their effects are multiplied together so that, for example, the odds of disease being present in hypertensive smokers is a product of the odds of disease presence in smokers and the odds of disease in hypertensive people, all multiplied by the constant term. The mathematical handling is simplified if this relationship is converted

G H HALL, MD, FRCP
Chief Medical Officer, Medical Sickness Society, Exeter
A P ROUND, MRCP
Registrar in Public Health, Exeter and North Devon Health Authority

to a sum rather than a product, by taking logarithms (usually to the base e). The logistic regression equation then becomes:

logarithm of odds = $a + b_1x_1 + b_2x_{2...} + b_ix_{i...} + b_nx_n$ of outcome,

where a = constant term;
x = the presence or absence (scored 1 or 0, respectively) of a variable or its value; and
b = log weighting coefficient for a variable.

This expresses the likelihood of the outcome as the sum of the effects of each 'explanatory' or 'independent' variable considered. It is helpful to recognise that the b coefficients are in fact the logarithms of the odds ratios for each variable.

As an example, consider the effects of smoking and hypertension (defined as systolic blood pressure > 160 mmHg) on the odds of developing a heart attack, compared with the experience of non-smokers and normotensives. Suppose the odds ratio for smokers developing a heart attack is 3. Then the b coefficient would be log 3, or 1.099. Similarly for the hypertensives, the odds ratio is 2, and the b coefficient 0.693. When the variable is present, $x = 1$, and when absent, $x = 0$, and the constant term, $a = -3.892$.

Then, log odds of heart attack = $-3.892 + (1.099 \times 1) + (0.693 \times 1) = -2.1$
and for non-smokers and normotensives, log odds of heart attack = $-3.892 + (1.099 \times 0) + (0.693 \times 0) = -3.892$

The odds are divided to calculate the odds ratio but, since these are logarithms, the log odds can be subtracted, thus:

Log odds of heart attack for hypertensive smokers compared to normotensive non-smokers = $-2.1 - (-3.892) = 1.792$, and $\text{antilog}_e\ 1.792 = 6$.

Therefore, the odds ratio of a heart attack is 6, and hypertensive smokers in this population are six times more likely to suffer a heart attack than normotensive non-smokers.

For population studies, the probability of an event can be calculated directly from the odds. By putting different factors into the equation, their effects can easily be demonstrated.

When a condition is rare, it may not be practicable to assess the effects of exposure to putative risk factors by random sampling of a population. The *case-control* method circumvents this difficulty by collecting cases with the disease and comparing the rates of exposure in this group with those in a control sample (Table 1). One of the great strengths of the method is that odds ratios can still be calculated from the results of such studies and used in logistic regression equations. Note, however, that individual prediction of outcomes is not necessarily valid.

The logistic regression model provides the following information:

Table 1. Case–control study

Cases		Exposed to risk factor	
		Yes	No
Diseased	Yes	a	b
	No	c	d
		Odds ratio = ad/bc	

The odds ratio will be the same for random samples and case-control studies because it depends on the proportions exposed in the diseased and non-diseased groups (a:b, and d:c) and not on their relative numbers [$(a + b)$:$(c + d)$].

1 The b coefficients in the logistic regression equation provide a measure of the degree of association between each variable and the outcome. This association is represented as the logarithm (to the base e) of an odds ratio. In our example, the odds ratio of a heart attack for smoking versus non-smoking groups is $\text{antilog}_e\ 1.099 = 3$. Thus, smokers would be thrice as likely as non-smokers to develop a heart attack in the population under study. Likewise, the odds ratio for hypertensives versus normotensives is 2. These effects are independent of each other, and the odds ratios are often described as being corrected (or adjusted), for the presence or absence of other variables. It is this property that makes logistic regression so attractive when analysing multifactorial data.

2 The explanatory power of each variable (that is, the difference that addition of a variable to the logistic regression equation makes to the correctness of the prediction) can be estimated. Generally, only those variables that confer a statistically significant contribution (the G statistic) by some chosen amount will be selected for inclusion in the final equation. Unfortunately, the particular selection of variables may make a great difference to the value of the b coefficient. Various procedures have been devised to obtain the 'best' combination, for example, stepwise methods, but some caution should be exercised if variables are selected by rote without reference to a hypothesis or model previously framed by the investigator.

3 The logistic regression model can be used to explore the effect of interactions between variables. Consider the effects of smoking and hypertension on coronary disease. The expected combined effect would be to increase the attack rate sixfold (3×2). If, in fact, the attack rate is greater or less than this, positive or negative interaction has occurred. In the logistic regression equations, an additional explana-

tory variable (combined smoking and hypertension) will allow for this, but in these circumstances the *b* values no longer represent the adjusted odds ratios.

As with most statistical parameters, tests of significance and of appropriate hypotheses can be applied to the *b* values. It is important to remember that these values are estimates of a 'true' value, and confidence limits can be applied to them. The Wald statistic gives some idea of the errors attached to the estimates [5].

In summary, the logistic regression model will evaluate the separate and combined effects of certain chosen predictor (or explanatory) variables on the probability of a given outcome. The predictor variables may be either categorical (yes/no) or continuous (over a range of values) whilst the outcome variable is usually categorical.

Precautions in interpretation of the logistic regression model

Many statistical texts rightly counsel great caution in the use and application of logistic regression models. The widespread and occasionally uncritical use of logistic regression has become possible only by the ready availability of computer progammes. This poses certain problems if the underlying assumptions or limitations are not understood.

Independence of the predictor variables

When estimating the values of the *b* coefficients from the data set, the mathematical procedure needs to assume independence in behaviour of the predictor variables. Thus, although the subsequent model exhibits the desired property of independence of the explanatory variables, this may not reflect the actual state of affairs. Variables in the real world are usually correlated to some extent: for example, body weight and blood pressure, and when the value of one changes, so does the value of the other.

Handling interaction

Logistic regression has been designed to provide a smoothed version of the data when an adequate number of observations is not available on all possible combinations. Unfortunately, gaps in the data may lead to neglect of interactions because of the high standard errors of the estimates. This promotes type II errors: interactions will not be detected when they actually exist. There is a temptation to exclude the interaction terms because the *b* values no longer represent individual odds ratios, which makes interpretation of the equation more difficult. In addition, statistical efficiency dictates the use of a parsimonious model, but this may not necessarily provide the best explanation of biological phenomena.

The multiplicative nature of the model

The multiplicative effect of increasing the 'dose' of a risk factor in the model is sometimes liable to misrepresent reality. Thus, the effect on the odds of developing coronary heart disease as the result of an increase of 1 mmol in serum cholesterol is estimated as 1.36 [7]. With 5 mmol/l as the baseline, the odds would be increased by 1.36 for a rise to 6 mmol/l, but by 2.27 for a rise from 11 to 12 mmol/l. There is no experimental evidence to confirm this. For this reason, it is usually recommended to avoid using continuous variables in the equation and to use only categorical variables.

Selection of variables

Although selection of the best explanatory variables by automatic procedures (eg stepwise analysis) may satisfy statistical criteria, it may not be the best approach. For example, the choice of variables may be guided by examination of the univariate odds ratios in the raw data, selecting first those of the highest values [5]. Different combinations of variables will produce different *b* values, indicating the subjective element in constructing the model. Moreover, the standard errors of the coefficients will change, leading to exclusion of some variables in one combination and not in another. Occasionally, interaction coefficients will be significant even when separate components are not.

Testing the model

Given these reservations, how can the model be tested? Once the *b* values have been estimated, it should be possible to assign a value to the probability of an outcome from a given set of data pertaining to an individual. The techniques involved are designed to minimise erroneous assignments for the population under study. Overall 'goodness-of-fit' estimates of the model tested by the χ^2 statistic may obscure important deviations for data at the extreme ranges of probability of the outcome. Also, tests of goodness-of-fit merely determine whether that model accurately describes the particular data observed. For general application, the models should be tested on a different population. This often leads to disappointing results (eg the seven countries study [98]).

In summary, the behaviour of the variables in the model may well differ greatly from the events in the real world, even though average predictions are reasonable. Accepted scientific method demands the formulation of a hypothesis as the basis for the design of a study, and this also applies to logistic regression design. Once the model is formulated, the effects of the variables are estimated, and an assessment made of the fit of the data to the model. Other possible models should be considered (with perhaps different assumptions and constraints), and conclusions can be drawn.

Finally, further hypotheses can be formed for future testing.

Approach to assessment of a logistic regression analysis

Evaluation of the results of logistic regression would be assisted if answers to the following questions were made available:

1 Are the original question, the hypothesis to be tested and the model utilised clearly described?
2 Is a complete list of variables considered for inclusion at the start?
3 Are all the relevant univariate risk or odds ratios provided?
4 Are reasons given for the inclusion or exclusion of each variable, and are the procedures used to select them described?
5 Is the final logistic regression equation given in full?
6 Have interaction effects been studied? What are the sizes and standard errors of the interaction coefficients?
7 Has the contribution been given of each variable (with confidence limits) to the overall explanatory power of the model (log likelihood contribution)?
8 Have the standard errors of the *b* coefficients been stated?
9 What is the overall goodness-of-fit?
10 What attempts have been made to identify any patterns of variable contributions which are not well fitted?
11 Has the model been tested on a different population (ie what is its general applicability)?
12 Have other models been suggested, tried and rejected?

Discussion

Interpretation of medical and epidemiological phenomena has come to rely heavily on statistical methods: the randomised controlled clinical trial of new treatments and the importance of careful design of such trials are good examples. It may not be obvious that statistics, like medicine, is subjective in the sense that alternative techniques may often be employed which may produce different results. Logistic regression is only one way of analysing multivariate data but even within this technique differing approaches can be used. Hence the importance both of a basic understanding of what the method can and cannot be expected to do, and of close collaboration with statistical colleagues.

Even when a particular technique has been selected as the most suitable, caution must be exercised. The difference between the estimates provided by a model (such as the logistic regression model) and the real world is not always adequately appreciated. It is easy to see that the statement 'an average of 2.4 children' applies only to a model and not to an actual family, but distinctions become less obvious when the statistical processes are complex. For instance, it is often said that a certain variable has been shown to be an 'independent predictor of an outcome', but this is *true only of the model.* Empirical testing of the results of an analysis on a different population is therefore important in the same way as the efficacy of a new drug is tested in various groups of patients.

Any conclusion is only as sound as the observations upon which it is based, and no statistical method can correct for poor data. In this respect, data from observational studies can be misleading, particularly if causal inferences are drawn from associations [9]. Experimental studies, on the whole, provide stronger evidence of causality. Nevertheless, there are other pitfalls in determining the meaning of statistical associations revealed by a study but which are not capable of being interpreted in the light of current medical thinking. Davey Smith *et al* [10] cite an observed association between smoking and suicide. This finding seems implausible on other grounds, so should it be dismissed? To do so would immortalise preconceived ideas, but to accept the association risks acceptance of a statistical artefact. Hence the importance of attempting to reproduce observed findings. It is impossible to reach conclusive proof, but unsuccessful trial of refutation (preferably in different populations) adds to the weight of evidence—Popper's falsification theory, which states that statements that are unprovable remain in principle disprovable, but that a theory holds until it is disproved.

The strength of logistic regression is also its weakness, in that complex matters are presented in simple form. Although this may not mimic the real world, it can afford a useful description and provoke further insights to many medical problems provided that its limitations are realised.

Acknowledgement

We wish to thank Professor J R Ashford for his helpful advice and criticism.

References

1 Golding J, Greenwood R, Birmingham K, Moet M. Childhood cancer: intramuscular vitamin K and pethidine given during labour *Br Med J* 1992;**305**:341–5.
2 Brand R, Keirse MJNC. Using logistic regression in perinatal epidemiology: an introduction for clinical researchers. *Paediatr Perinatal Epidemiol* 1990;**4**:22–38;2211–35.
3 Daley LE, Boursk GJ, McGilroy J. *Interpretation and uses of medical statistics.* London: Blackwell, 1991.
4 Everitt AS. *Statistical methods for medical investigations.* Oxford: Oxford University Press, 1989.
5 Hosmer DW, Lemeshow S. *Applied logistic regression.* New York: J Wiley, 1989.

6 Truett J, Cornfield J, Kannel W. A multivariate analysis of the risk of coronary heart disease in Framingham. *J Chron Dis* 1967;**20**:511–24.
7 Shaper AG, Pocock SJ, Walker M, Phillips AN, *et al.* Risk factors for ischaemic heart disease: the prospective phase of the British regional heart study. *J Epidemiol Community Health* 1985;**39**: 197–209.
8 Keys A (ed). *Seven countries: a multivariate analysis of death and coronary heart disease.* Cambridge, MA: Harvard University Press, 1980:136–60.
9 Bradford Hill A. The environment and disease: association or causation? *Proc R Soc Med* 1965;**58**:296-300.
10 Davey-Smith G, Phillips AN, Neaton JD. Smoking as 'independent' risk factor for suicide. Illustration of an artifact from observational epidemiology. *Lancet* 1992;**340**:709–12.

Address for correspondence: Dr G H Hall, 5A Victoria Park Road, Exeter EX2 4NT.