APPENDIX 2

Choice of explanatory variables in multiple regression

A2.1 Introduction

In Section 2.1 we set out procedures for analysing a general linear logistic regression in which dependence of a probability of success on explanatory variables is expressed via a linear relation on the logistic scale. Implementation of these methods is straightforward once a set of explanatory variables is chosen for inclusion. In applications, however, a crucial aspect is precisely that choice, it being fairly rarely the case that, for example, theoretical considerations indicate unambiguously the equation to be fitted. The points at issue are similar to those arising with 'ordinary normal theory' empirical multiple regression based on the method of least squares and indeed in other forms of empirical linear regression in generalized linear models. In the present Appendix we discuss these issues in fairly broad terms.

One special aspect of the normal-theory case that does affect the strategy of tackling the analysis of highly balanced sets of data is that exact or nearly exact orthogonality in normal theory implies that estimates of certain parameters are unaffected by the inclusion or exclusion of some other parameters. This makes it feasible to begin the analysis of a balanced design by inspection of a 'full' analysis of variance in which possibly large numbers of main effects and interactions are included. In linear logistic regression however a balanced design leads to only approximate orthogonality of the estimated parameters and it is not always possible to see immediately the precise effect of such inclusion or exclusion. For this reason it is commonly sensible to begin with some relatively simple model and then to examine the need to amplify or indeed simplify the initial model. The criteria for the choice of that starting model as well as for modifying the model in the light of the data become of more pressing concern.

Sections A2.2 and A2.3 deal with type and formation of explanatory variables rather than with the strategy for choice of explanatory variables but these ideas are nevertheless important in the analysis and interpretation of regression models. Also much of the material in Sections A2.4 and A2.5 is not specific to binary data but is given here for completeness of discussion. Reference is made to examples discussed in Chapter 2 of this book.

A2.2 Types of explanatory variable

It is convenient to classify potential explanatory variables in several different ways.

First for purposes of interpretation we may classify explanatory variables as in Section 2.8, namely as

1. treatment or quasi-treatment variables representing aspects which can in principle at least be manipulated (Example 2.10);
2. intrinsic variables measuring aspects characterizing an individual under study or the environment in which the study on an individual is carried out, for example age, socio-economic class (Example 2.18);
3. non-specific variables characterizing broad groupings of individuals; often such groupings are described by names such as blocks, strata and so on (Example 2.16).

The object of study is normally the assessment of the effect of treatments and of possible interaction of treatment effects with variables of type (2) or (3) (Example 2.17).

Of course this division into three types depends on the context and may not be clear-cut. Especially in observational studies some intrinsic variables, such as socio-economic class of individuals, may be surrogates for other more specific properties, such as educational background, and wherever possible more specific variables should, of course, be used.

In a randomized experiment, treatment variables are randomized and intrinsic variables are those measurements made on the individuals before randomization.

In an observational study, the treatments are typically aspects that ideally have been investigated via a randomized experiment, but which in fact were determined in a way outside the investigator's control. Thus in a study of the effect of alcohol consumption during pregnancy on some feature of the infant, randomization is obviously not feasible with human subjects. Treatment, perhaps better called quasi-treatment, variables are thus measures of alcohol consumption and other matters, such as diet, necessary to define the treatment effect under study, whereas intrinsic variables are mother's age and parity, socio-economic class, etc. If the study were replicated in a number of centres, centres would form a non-specific variable.

A second classification of explanatory variables, relevant in analytical formulation, is by their mathematical structure, according to whether they take

1. a number of qualitatively different levels, such as one of a number of regions of residence;
2. a number of ordered levels, such as the description of the severity of some condition as slight, moderate, severe and very severe;
3. values specified by a reasonably well-defined quantitative scale.

A third classification is into

1. directly measured variables;
2. derived variables, by which we mean both composite variables obtained by taking combinations of measurements or variables such as squares and products of more directly observed quantities.

A2.3 Formation of explanatory variables

In some situations explanatory variables may be entered into a multiple regression equation either in exactly the form in which they are measured or after rescaling; a simple change of units to make all variables have approximately the same standard deviation in the data, and in some cases a change of origin to produce means that are not too large may help avoid numerical instability (Example 2.11). For essentially positive variables a log transformation may be wise (Examples 2.10, 2.11). Care is, however, needed with variables that have a very wide range, especially where very non-linear effects are likely. Thus if in a clinical study age at entry ranged from 60 to 70 years, direct introduction of, say, age -65 as a quantitative variable would be reasonable, and non-linearity could, if necessary, be checked via a squared term. But if age ranged from 20 to 80 years some grouping of age into a fairly small number of groups and their treatment initially as qualitatively different, as explained below, would protect against strong non-linearity. Again for alcohol consumption in litres per week it would usually be better to work initially with none, slight, moderate, heavy rather than directly with the quantitative measurement; later analysis could refine the initially arbitrary subdivision, if that seemed likely to be fruitful.

For qualitative variables at l levels, the construction of $l-1$ explanatory variables will be needed if the main effect of such a variable is to be represented without prior constraint (Example 2.12). The method of construction is in one sense arbitrary so long as we make the $l-1$ variables linearly independent but the following considerations are helpful.

1. The marginal frequencies of the different levels should be inspected, in particular to avoid giving prominence to levels that occur with very low frequency.
2. If there is a level, say 1, with a very low frequency and its possible merging with another level, say 2, appears possibly sensible, it will be useful to define one variable, say $x_1 = 1$ (level 1), -1 (level 2), 0 (all other levels), so that the resulting estimated parameter provides a test of the reasonableness of the proposed merging. In defining the other $l-2$ variables, the two levels 1 and 2 can then be treated identically.
3. If one level, say 1, is a control, or other natural reference level, or occurs with especially high frequency, it may be sensible to define all the *x*s relative to 1, i.e. to define $x_1, \ldots, x_{l-1}$ by $x_j = 1$ (level $j+1$), -1 (level 1), 0 (otherwise).
4. If the levels are ordered and are of very roughly equal frequency, it may be sensible to define *x*'s via the standard orthogonal polynomials (Pearson and Hartley, 1966, Table 47) for l equally spaced points, using thus for three levels $-1, 0, 1$; $1, -2, 1$ to define respectively x_1, x_2.
5. There should be some rough check that the *x*s are not defined so as to be nearly linearly dependent.
6. Any special arguments indicating contrasts that are likely to be particularly important should, of course, be used in defining the *x*s.

Interactions are normally best studied in this context by defining products of the *x*s defining the main effects in question. In exploratory work, the principle that large main effects are on the whole more likely to generate appreciable interactions than small main effects is often helpful. Thus if two qualitative variables have

l_1 and l_2 levels respectively, leading to the definition of $l_1 - 1$ and $l_2 - 1$ explanatory variables, the set of all products of these variables defines the two-factor interaction with $(l_1 - 1)\ (l_2 - 1)$ degrees of freedom. If it is required to extract a few degrees of freedom from the interaction this can be done via products of component x's with especially strong interpretations or, failing that, via components that happen to be large.

In some applications to calculate *a priori* combinations of the explanatory variables may be useful. For example: near orthogonality may be achieved by replacing diastolic and systolic blood pressure by the sum and difference of the logs of the two measurements; if a particular feature has been measured in several different ways a composite score may initially be tried. In these cases it will often be wise to test from the data whether the indicated combination appears to have sacrificed information about the response variable under study.

A2.4 Small numbers of explanatory variables

In some applications there may be a reasonably small number of explanatory variables, corresponding to say at most five or six parameters. Unless a treatment effect of primary interest is substantially confounded with variables of no direct interest, there seems little point in trying to simplify the resulting equation by omitting explanatory variables merely on grounds of statistical insignificance; it may be a useful quick check on the potential for improving precision of treatment effects to compare the standard error under the full fit with that achieved by omission of *all* other variables.

The model with all explanatory variables in linear form may be augmented by adding non-linear functions, e.g. squares of quantitative variables, and interaction terms (Example 2.11). A simple strategy is to begin by adding such terms one at a time, concentrating on interactions of the treatment effects of primary interest with intrinsic and non-specific explanatory variables and on possible non-linearity of response to important quantitative variables.

In judging statistical significance it is important to make allowance whenever the largest from among many possible contrasts is chosen for interpretation. One way to do this, when a variable is chosen for inclusion out of a block of variables, is to examine the change in 2 log (maximized likelihood) when all the variables in the block are fitted; if this is not statistically significant at an interesting level, there is a danger that the variable selected is an artefact.

A2.5 Large numbers of explanatory variables

A much more difficult situation arises if there are so many potential explanatory variables that some reduction from the full fit is essential either to achieve understandable interpretation or reasonable precision in the primary comparisons. Now many computer packages contain automatic algorithms for variable selection. We strongly recommend against reliance on these algorithms, except occasionally in the very restricted context set out below. This is because the choices they force are often of a very arbitrary character and are often not the most appropriate for the purpose either of prediction or of interpretation; to end with one set of variables when there are other quite different choices having virtually as good a fit to the data invites misinterpretation.

We suggest a procedure along broadly the following lines.

1. List those explanatory variables which it is essential to include either because they are treatment variables of primary concern, or because it is known from previous studies that they are important.
2. Consider whether certain subsets of variables (e.g. measurements of the same feature) should be treated separately or whether some preliminary reduction across subsets, such as by the formation of totals, might be fruitful.
3. Check for the influence of other variables, at first one at a time as in Section A2.4, or by cautious use of a computerized selection algorithm.
4. Iterate this procedure, i.e. repeat both phases with the initial variables, those from the initial list supplemented by and/or replaced by other variables found empirically from the data.
5. When one or more apparently adequate fits have been obtained check for the addition of further variables, interactions and so on as outlined in Section A2.4.

If, as is likely, there are different choices in the later phases that give adequate fits it is important, as far as is feasible, to give *all* fits consistent with the data, making any choice between alternative choices on subject-matter grounds.

A crucial aspect is the behaviour under alternative models of the aspects of primary interest, i.e. parameters representing treatment effects and their potentially important interactions with intrinsic variables. If these are reasonably stable, choice of other aspects of the model is probably not of critical importance.

Note especially that important variables, especially those representing treatments, should not be excluded solely because the corresponding estimates are insignificant statistically; their estimation is likely to be of direct interest and the inclusion of estimates and standard errors is in any case likely to be essential in a final report on the data, if only to allow comparison with subsequent related studies.

Where there are several rather similar sets of data for analysis it will usually be wise to use the same explanatory variables for all sets. Thus if the number of such variables is large a cautious procedure is to aim to choose explanatory variables separately for each set and then to re-analyse using the set of *all* variables so chosen, before possibly attempting some common reduction. Incautious use of automatic selection algorithms is quite likely in these contexts to throw up different choices of variables in the different sets, with consequent dangers of misinterpretation.

A final note of caution concerns the interpretation of significance tests and confidence intervals when a complex sequence of data-dependent choices has been made. If there is really very little or no explanatory power in a block of variables and one or two are selected on the basis of the data as showing the largest apparent effect, there is a clear possibility of considerable exaggeration of significance; this stresses the need for some protection from global tests of blocks of parameters.

We do not think it feasible to specify probability properties of complex sequences of data-dependent choices. Note, however, that if *all* sufficiently simple equations consistent with the data at a specified standard significance level are listed, any 'correct' such specification will be included with the specified confidence coefficient.

We consider, however, that by following the broad guidelines above and concentrating on the treatment effects of primary concern, these puzzling difficulties are to a large extent bypassed. If all that is required is a well-fitting empirical equation and no interest attaches to individual effects, again questions of the significance of individual terms are essentially irrelevant, although we regard such totally empirical prediction equations as of rather limited interest.