Modelling with multiple explanatory variables

Pamela Warner

Centre for Population Health Sciences, University of Edinburgh, Edinburgh, UK

Correspondence to
Dr Pamela Warner, Centre for Population Health Sciences, University of Edinburgh Medical School, Teviot Place, Edinburgh EH8 9AG, UK;
p.warner@ed.ac.uk

This statistical technique has been used in two papers in this issue of the Journal, namely those by Kotb *et al.*[1] and Schembri *et al.*[2] These notes are intended to provide a supplementary explanation of this method. [See Box 1 for a glossary of terms used in this article.]

What is it?

Multivariable modelling is the use of statistical modelling techniques, applied to a dataset for a group of individuals – that dataset comprising some *known* outcome or group membership (usually binary), plus a set of variables that potentially 'explain' that outcome/group membership.[3]

When/why is it useful?

It is often the case that research seeks to understand better the *nature* of the association between some condition/group of interest and a set of potential explanatory variables – which might be demographic, behavioural, exposures, and so on.[4–7] This understanding can be difficult to achieve because of the fact that the outcome is associated with numerous potential explanatory variables, and because, often, these potential explanatory variables are associated among themselves. For example, lack of knowledge about contraception, and difficulty in affording it, might both be associated with low socioeconomic status. Multi-variable modelling can reveal the association of a potential explanatory variable with the outcome, *adjusted for* (or 'independent' of) all the other explanatory variables in the model.

In health care it is sometimes the case that understanding of the precise associations/causes is not of prime importance, whereas what would be useful would be to be able categorise patients into groups, on the basis of known or easily ascertained information about them. In such circumstances, it is often a further requirement for the multi-variable model developed that it is 'parsimonious', including only as many explanatory variables as are needed to give good prediction. Subject to further validation in independent data, and ideally a randomised trial to evaluate the benefit of the implementation of such a prognostic algorithm, a prognostic model/algorithm is derived, for the population from which the data have been drawn.[8]

What precautions are needed?

The number of explanatory variables, and the extent to which they are interrelated, can create difficulties for the mathematics of the modelling. Therefore it is a common precaution to subject the potential explanatory variables to prior screening by means of separate (univariate) analyses of association of outcome with potential explanatory variable, and then include in, or offer to, the multi-variable model, only those variables with some degree of association with the outcome variable. This avoids overloading the initial model with too many variables for the study size, a particular concern when the explanatory variables are highly related one with another.

Example

In the associated research paper by Kotb *et al.*, univariate analyses were reported for 27 potential explanatory variables, and for the multi-variable model this number was reduced to the 18 that satisfied their criterion ($p<0.10$) for consideration in the multi-variable model.[1] (Age was also included in the MV model.) The authors report the odds ratio (ORs) of association from the multivariable model for the five variables most strongly associated with unmet contraceptive need. Table 1 presents these ORs, and the ORs corresponding to the univariate associations reported in Kotb *et al.'s* paper. It can be seen that after adjustment for age and other explanatory variables in the multi-variable model, there was some change in the size of the ORs – one OR has increased, whereas the other ORs are smaller after

Table 1 Comparison of odds ratios summarising association, for univariate and multi-variable analyses*

	Association of unmet contraceptive need with explanatory variable	
Potential explanatory variable	**Univariate OR**	**Multi-variable OR**
Previous side effects from contraception – Yes vs No	5.0	5.7
Previous unwanted pregnancy – Yes vs No	4.6	3.0
Husband opposes contraception use – Yes vs No	4.0	3.0
Religious beliefs		
Not accepted vs Accepted	3.0	2.1
Don't know vs Accepted	3.1	2.2
Desired number of children		
Don't agree vs Agree	2.1	1.5
Don't discuss vs Agree	2.8	2.6

*Multivariate odds ratios (ORs) from Kotb *et al.*[1] and univariate ORs calculated from data reported there.

Box 1 Glossary of statistical terms used in this article

'Adjusted' association	See *Association* and *Logistic regression.*
Association	Relationship between two variables. For binary variables this means that the occurrence of a particular value of one variable, in an individual, is associated with (more likely to be in conjunction with) a particular value of the other variable.
Binary variable	Has only two possible values (e.g. accept screening or not, female or not).
Categorical variable	Has a set of distinct values, such as gender, recruitment setting.
Explanatory variable	A feature potentially associated with outcome.
Logistic regression (LR)	LR estimates and tests association between a binary outcome and one or more explanatory variables, with association being summarised as ORs. In univariate LR there is only *one* explanatory variable and, if that is binary, only one OR. If there is more than one explanatory variable then multivariable LR is needed (MV LR). This estimates the association of outcome with each explanatory variable '*adjusted*' for the joint associations with other explanatory variables in the model.
Multi-variable modelling	A mathematical combination of data for a sample, where there is both an outcome variable, and a number of potential explanatory variables for that outcome. There are various mathematical forms, but where the outcome is a binary variable, not a rare event, then the usual form is multi-variable logistic regression.
Null hypothesis (NH)	A statement, prior to testing, of no effect (in this case, no association). See also *Significance probability.*
Odds ratio (OR)	The OR is the ratio of the odds of outcome (unmet need) in those with the explanatory feature (previous unwanted pregnancy), relative to the odds in those with*out* the explanatory feature (no such pregnancy). When there is no association, the two odds should be approximately equal and the ratio approximately 1, or 'null'. The more extreme the OR (away from 1, i.e. 0.4 vs 0.7, or 3.2 vs 1.8) the greater the degree of association.
Outcome variable	Indicates status with respect to the condition of interest (e.g. unmet contraceptive need).
Significance probability	The probability of obtaining the observed cell counts, or more extreme, if the NH is true.

adjustment. Some of the decreases are likely to be because of shared information across the variables (about unmet contraceptive need). For example, if variable A and B are both associated with 'unmet need', but some of that association is common to them both, then in a multivariable model with both A and B included, and all else being equal, neither will show as strong an *adjusted* association (or, equivalently, as extreme an OR), as was found in univariate analyses. Alternatively, the univariate association of variable C with unmet need might be confounded by other variables. If the confounding variable happens to be one of the other explanatory variables that is included in the multi-variable model, D say, then in the multi-variable model, adjusted association of C with unmet need will be free of confounding by D, which could result in a change in the OR, in either direction. Such might be the case for 'previous side effects' where the univariate OR is 5.0, but the multi-variable OR is 5.7.

Overview

Given the different approaches to multi-variable modelling that can be taken, the strategy used should reflect the research aims. Care is needed with interpretation of the results of analyses of association, in particular the ORs in a multivariable model, where there is adjustment for all variables in that model.

Competing interests None.

Provenance and peer review Commissioned; internally peer reviewed.

References

1 **Kotb MM**, Bakr I, Ismail NA, *et al*. Women in Cairo, Egypt and their risk factors for unmet contraceptive need: a community-based study. *J Fam Plann Reprod Health Care* 2011;**37**:26–31.
2 **Schembri G**, Schober P. Risk factors for chlamydial infection in chlamydia contacts: a questionnaire-based study. *J Fam Plann Reprod Health Care* 2011;**37**:10–16.
3 **Kirkwood BR**, Sterne JAC. *Essential Medical Statistics.* Oxford, UK: Blackwell Science, 2003.

4 **Warner P.** Testing and quantifying association in binary data. *J Fam Plann Reprod Health Care* 2009;**35**:26–27.
5 **Warner P.** Rate and rate ratio. *J Fam Plann Reprod Health Care* 2009;**35**:111–113.
6 **Warner P.** Quantifying association in ordinal data. *J Fam Plann Reprod Health Care* 2010;**36**:83–85.
7 **Warner P.** Ordinal logistic regression. *J Fam Plann Reprod Health Care* 2008;**34**:169–170.
8 **Altman DG,** Vergouwe Y, Royston P, *et al.* Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009;**338**:b605.