Paper 2: Issues arising in the statistical modelling of multivariate health data

Damon Berridge, PhD

Department of Mathematics and Statistics, Lancaster University, Lancaster, UK.

Keywords: multicollinearity, multivariate, socio-economic, parsimonious model

Received: 30/3/2014; Revised: 20/4/2014; Accepted: 25/4/2014

According to Confucius, 'Success depends upon previous preparation, and without such preparation there is sure to be failure'. This philosophy applies equally well in all situations and, when presented with a dataset to analyse, the job of a statistician is no exception. In this article, we address a number of key issues when considering the use of a statistical model as the primary tool for analysing multivariate data, that is, data comprising a single response variable and a set of explanatory variables. These issues include the need to identify and classify the response variable and set of explanatory variables required to define one or more research hypotheses of interest, and the importance of univariate and bivariate exploratory analyses which can inform the more formal analyses, namely the statistical modelling.

The first issue is to use substantive theory to formulate one or more research hypotheses of interest relating to the data. This involves identifying the response variable and explanatory variables of interest. For example, in the first article in this series, we wished to relate the response variable, survival (or death), to a range of biological, psychological and social factors (Shahtahmasebi and Berridge, 2010). Other health-related examples are outlined in Chapter 1 of Berridge and Crouchley (B&C) (2011). These include psychological distress (B&C example 1.1), immunization of Guatemalan children (1.10), respiratory status (1.13), headaches (1.14), epileptic seizures (1.15) and skin cancer deaths (1.16).

Each response variable can be classified into one of two broad groups: either as quantitative/measured or qualitative/categorical. If a response variable is classified as quantitative/measured, then a response variable may be either continuous (e.g. psychological distress score in B&C example 1.1) or discrete, i.e. a count. Examples of count data from Berridge and Crouchley (2011) include number of headaches (1.14), number of epileptic seizures (1.15) and number of male deaths due to malignant melanoma (1.16).

If a response variable is classified as qualitative/categorical, then it may be classified further, depending upon whether the variable consists of two categories: binary, e.g. whether or not a Guatemalan child has been immunized (B&C example 1.10), or the variable comprises more than two categories. If the response variable involves more than two categories, then further classification depends on whether the multiple categories are ordinal/ordered (e.g. respiratory status in B&C example 1.13) or nominal/unordered, for example, the track taken by school pupils at the age of 16 (Penn and Berridge, 2008).

The acid test, when deciding whether a variable is ordinal or nominal, is to swap any two categories. If the swap renders the variable meaningless, then the variable is most likely to be ordinal in nature. The response variable ‘respiratory status’ (B&C example 1.13) comprises the following five

categories: 'terrible', 'poor', 'fair', 'good' and 'excellent'. Suppose we swap two categories, say 'poor' and 'good'. This swap would render the variable meaningless, hence 'respiratory status' is deemed to be an ordinal response. In contrast, the track taken by school pupils at the age of 16 consists of four categories: 'left school at 16', 'vocational education/training', 'lower academic' and 'higher academic' (Penn and Berridge, 2008). Swapping any two of these categories does not affect the meaning of this variable, so this variable is regarded as a nominal response.

Having identified, defined and classified the response variable, we can proceed to perform univariate exploratory analyses, that is, to produce appropriate numerical and graphical representations of that variable. The type of response variable will determine the kinds of numerical and graphical summaries to be used. For example, a histogram is suitable for representing the distribution of a continuous variable, while a barchart or piechart may be more appropriate for a categorical variable.

Univariate numerical and graphical summaries are helpful in helping us to decide how to operationalise the response variable. What do we mean by 'operationalising the response variable'? We explain operationalisation in the context of a specific example. A histogram of a continuous response variable will allow us to check for normality. If the histogram approximates a bell-shaped curve (i.e. symmetric and uni-modal), then we may be safe to assume that the response variable follows a normal distribution. If the histogram indicates positive or negative skewness, then a transformation, for example, taking the natural logarithm of the original response variable, might help to satisfy the assumption of normality. If the histogram displays a bi-modal distribution, then it will be difficult, if not impossible, to assume normality and an alternative strategy must be adopted. For example, the continuous response could be dichotomised to create a binary response, or split into multiple categories to create an ordinal response.

Having decided how to operationalise the response variable, we are now in a position to choose the appropriate type of statistical model. There is a family of regression models, known as the family of generalized linear models (Berridge and Crouchley, 2011; Collett, 2002; Dobson, 1991; McCullagh and Nelder, 1989), that can be fitted. We need to decide which member of this family of models to use. This depends on the nature of the response variable under consideration.

A continuous response may be analysed using a linear regression model which assumes that the response follows a normal (or Gaussian) distribution. The four volumes of Penn and Berridge (2013) include key articles on the statistical analysis of continuous data.

The Poisson distribution is the natural starting point in the modelling of count data. A binary response may be analysed using a range of models, including the binary logit model (or binary logistic regression). Alternatives to the binary logit model include the binary probit and complementary log log models (Collett, 2002). More details on Poisson models and models for binary data are discussed in Berridge and Crouchley (2011).

An ordinal response may be analysed using the proportional odds (or cumulative logit) model which assumes that the ordinal response is symmetric and reversible. Such ordinal responses include Likert (1932) items like the British Household Panel Survey (BHPS) (Taylor *et al*., 2005) item: 'A husband's job is to earn money; a wife's job is to look after the home and family'. This item is coded as 'strongly agree', 'agree', 'neither agree nor disagree', 'disagree' and 'strongly disagree'. Berridge,

Penn and Ganjali (2009) examined changes in attitudes to gender roles in contemporary Britain by applying marginal and conditional cumulative logit models to BHPS data on this item. If an ordinal response is asymmetric and irreversible, such as disease progression, then the continuation ratio model may be more appropriate. Key articles in the modelling of ordinal categorical data are featured in Volume 4 of Penn and Berridge (2010). A nominal response may be analysed using the multinomial logit model, as in the analysis of the track taken by school pupils at the age of 16 (Penn and Berridge, 2008).

Each explanatory variable should be classified in a similar manner to a response variable. An explanatory variable may be classified as a factor (categorical variable) or a continuous covariate. Variables such as gender, social class, ethnicity, religious affiliation and marital status are regarded as factors. In contrast, variables like age, height and weight are considered to be continuous covariates. Appropriate numerical and graphical representations of each explanatory variable should be produced. These univariate exploratory analyses can be used to help us to decide how to operationalise each explanatory variable; for example, whether to combine categories or not.

Exploratory analyses should progress from univariate to bivariate ones. Suitable numerical and graphical summaries of the relationship between the response variable and each explanatory variable should be created. These bivariate summaries help us to identify salient features in the data. For example, we can use scatterplots to help us discern patterns in the data, for example, positive or negative trends and linearity or non-linearity in the relationship between two variables.

If the relationship between two variables is non-linear then often it can be made linear by transforming either or both of the variables. Possible transformations include logarithms, exponentials, square roots and squared values. The idea is to choose a transformation that makes the relationship between the two variables at least approximately linear. Scatterplots may also be useful in the identification of outliers (or 'extreme cases') which may exert undue influence on the fitting of subsequent statistical models.

We may also produce suitable numerical and graphical summaries of the relationship between pairs of explanatory variables, for example, the correlation between height and weight. Exploratory bivariate analyses on pairs of explanatory variables may highlight high correlations/associations between explanatory variables (known as multicollinearity) which may have an impact on the model building process. In general terms, if two factors X_1 and X_2 are highly correlated with each other, then the inclusion of the significant explanatory variable X_1 in a model may make X_2 redundant.

In this article, we have raised a number of issues which arise in the statistical modelling of multivariate data. These issues are discussed in more detail by Shahtahmasebi and Berridge (2010). The next article places these issues in the specific context of an example on survival in old age.

References

Berridge and Crouchley (2011) *Multivariate Generalized Linear Mixed Models Using R*. London: Chapman & Hall

Berridge, D., Penn, R. and Ganjali, M. (2009) Changing attitudes to gender roles: A longitudinal analysis of ordinal response data from the British Household Panel Study, *International Sociology*, **24**(3), 346-367

Collett, D. (2002) Modelling Binary Data (Second Edition). London: Chapman & Hall

Dobson, A.J. (1991) *An Introduction to Generalized Linear Models*. New York: Wiley

Likert, R. (1932) A technique for the measurement of attitudes, *Archive Psychology*, **22**(1), 40-55

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*. London: Chapman & Hall

Penn, R. and Berridge, D. (2010) *Social Statistics Volume 4:The Statistical Modelling of Ordinal Categorical Data.* London: Sage

Penn, R. and Berridge, D. (2013) *The Statistical Analysis of Continuous Data Volumes 1 to 4*. London: Sage

Penn, R. and Berridge, D. (2008) Modelling trajectories through the educational system in North West England, *Education Economics*, **16**(4), 411-431

Shahtahmasebi, S. and Berridge, D. (2010) *Conceptualizing Behaviour in Health and Social Research*. New York: Nova Science Publishing

Taylor, M.F., Brice, J., Buck, N. and Prentice-Lane, E. (2005) *British Household Panel Survey User Manual Volume A: Introduction, Technical Report and Appendices*. Colchester: University of Essex

[1,711 words]