The Use of the Statistical Model as an Effective Tool in the Analysis of Multivariate Data in Health Research

Damon Berridge, PhD
Department of Mathematics and Statistics, Lancaster University, Lancaster, UK.

Correspondence: Damon Berridge – email: d.m.berridge@swansea.ac.uk
Keywords: multicollinearity, multivariate, socio-economic, parsimonious model

In a letter to Jean-Baptiste Leroy in 1789, Benjamin Franklin wrote 'Our new Constitution is now established, and has an appearance that promises permanency; but in this world nothing can be said to be certain, except death and taxes'. Death is an inevitable outcome of old age, yet some people will die soon after retirement while others will survive to a ripe old age. What factors help to explain the variation in death (or survival) post-retirement?

Pfeiffer (1970) concluded that 'there is no single factor which determines longevity but rather a constellation of biological, psychological and social factors amounting to an elite status'. He found that people with high intelligence, sound financial status, good health and an intact marriage may expect to live longer than those without these characteristics. In a study of the 65-69 age group in the UK, (Abrams, 1983) concluded that survival was related to being female and in and good health, as well as having low levels of loneliness and depression, and high levels of social interaction and life satisfaction.

These findings are in general agreement with other studies the past (e.g., Fox & Goldblatt, 1982; Hirdes & Forbes, 1989; Jones, 1987; G. Kaplan, Barell, & Lusky, 1988; G. A. Kaplan, Seeman, Cohen, Knudsen, & Guralnik, 1987; Palmore, 1982), and of more recent studies (e.g. Connolly, Rosato, & O'Reilly, 2011; Grundy & Sloggett, 2003; Houweling & Kunst, 2010; Risnes et al., 2011). Higher levels of occupation, social class, income and education are indicators of socio-economic status and appear to be associated with lower mortality rates. Conversely, bereavement, no supportive network and social isolation are reported to be correlated with higher levels of mortality. Clearly, a large number of factors have been suggested as being associated with survival in old age.

In this article, we define data as being multivariate in the sense that we wish to relate a single response variable (also known in other contexts as an outcome or dependent variable) to more than one explanatory variable (also known as predictor or independent variables). In this article, we promulgate the use of the statistical model as the primary tool for the analysis of such data. Statistical modelling is a comprehensive analytical framework which can be used to analyse data in a highly structured manner. Key stages in the statistical modelling approach involve hypothesis formulation, model specification, model fitting, model criticism and model interpretation (Davies, 1992). In this article, we consider each of these steps in turn.

The initial stage involves the formulation of one or more research hypotheses of interest relating to the data under scrutiny. The formulation of research hypotheses should be guided by substantive theory. Each research hypothesis should be operationalised in terms of the relationship between a response variable and one of more explanatory variables. For example, in our motivatory application on survival in old age, the response variable is death (or survival) which is hypothesised to be highly associated with a range of biological, psychological and social factors, as outlined earlier in this article.

We could begin our exploration of the relationship between survival and a set of explanatory variables with the simplest analysis, commonly known as bivariate analysis, where the association between each explanatory variable and the response variable is explored. For example, we could examine the association between survival and gender by constructing a two-way crosstabulation (or contingency table) of counts cross-classified by survival/death and gender, and by performing a Pearson chi-squared test of association (Penn & Berridge, 2010) and its associated p-value. We will look at some examples of such exploratory analyses in the third article in this series.

A major criticism of bivariate exploratory analyses such as crosstabulations and chi-squared tests is that we would ignore the influence of other factors on survival that may be operating through gender. In other words, we need to allow formally for **control** in the analysis. A multivariate analysis is required to allow us to make conclusions about the effect of gender on survival, having adjusted for other explanatory variables. In examining the net effect of gender on survival, we must 'control' for the effect of other factors.

One way of imposing 'control' is to disaggregate the data, in other words, to construct a series of multi-dimensional crosstabulations of the response variable, survival, with two or more explanatory variables. There are several problems inherent with this approach. First, it will produce numerous tables to examine and interpret. In studies where a large number of explanatory variables is involved, 'the examination and interpretation of a large number of tables becomes cumbersome, painstaking and speculative, and the interpretation of such tables may lead to unsatisfactory conclusions' (Shahtahmasebi & Berridge, 2010). Second, this approach produces sparse multi-way tables with (very) small cell frequencies and associated chi-squared tests which cannot be interpreted in a meaningful way. Finally, this method provides no way of quantifying and of testing for the statistical significance of the effects of each explanatory variable.

In this article, we propose to analyse multivariate data within a statistical modelling framework. With a statistical model, we can achieve the following objectives:

- To explore the relationship between survival and each explanatory variable in the presence of other factors, in other words, to control for other effects
- To explore the relationships between explanatory variables (multicollinearity), for example, the relationships between gender, age and self-assessed health status
- To estimate and quantify the main effects of explanatory variables on survival, thereby enabling conclusions about *ceteris paribus* the effect of each explanatory variable on survival
- To reduce the large number of explanatory variables thought to be related to survival to a smaller number of factors which are easy to manage and interpret
- To estimate and quantify the interactions between explanatory variables, for example, a significant interaction between gender and age would mean that the effect of age on survival varies significantly between males and females.

In the third article in this series, we will illustrate the application of statistical models to real-life data on the factors which determine survival in old age.

A further advantage of analysing multivariate data within a statistical modelling framework is the ability to handle different types of response variable, including continuous and categorical data. Different kinds of statistical model are available. Each type of model is appropriate for a particular type of response. For example, the normal linear model is suitable for analysing

a continuous response. We will discuss the classification of response variables and the specification of appropriate models in more detail in the next article in this series.

The general principle we apply when building a statistical model is the principle of parsimony, also known as 'Occam's Razor'. In other words, we aim to find the **final model** (i.e. the most parsimonious model) which explains as much of the variation in the response variable as possible using the smallest number of explanatory variables.

When dealing with a large number of explanatory variables, it is convenient to automate the variable selection process in order to carry out each step in the model building process. These automated approaches include forward variable selection (FVS) and backward variable elimination (BVE). FVS starts with the **null model** (i.e. the model which includes **no** explanatory variables) and builds up the complexity of the model by adding significant explanatory variables, one variable at a time. In contrast, BVE begins with the **full model** (i.e. the model which includes **all** explanatory variables) and reduces model complexity by removing non-significant explanatory variables, again one at a time.

Having determined the final model, we should assess how well the model has performed on a case by case basis by computing a range of diagnostics which might include residuals and measures of influence. For example, we could predict values of the response variable for each case by plugging values of the explanatory variables into our final model. In its most basic form, the **residual** is defined as the simple difference between predicted (or fitted) value and observed value of the response variable. The residual is computed for each case. We can then plot these residuals against predicted values of the response variable. Such residual plots are useful for checking underlying model assumptions such as normality and homogeneity of variance when fitting a linear model to a continuous response. More details of linear models for continuous responses will be described in the second article in this series.

Once we are content that our final model provides a satisfactory description of our data, it is good practice to present the results for both full and final models. This allows us to see at a glance what effect the inclusion of the non-significant explanatory variables in the full model has on the significant explanatory variables included in the final model. Model results should be summarised in a table. Columns in the table of results should include some, if not all, of the following elements: parameter estimates, standard errors, test statistics, p values and 95% confidence intervals.

Once the final model has been presented clearly and concisely, its results should be interpreted in a similar manner. Interpretation should take place within the context of the original substantive theory which generated the initial research hypotheses. We will see some examples of good practice in presenting and interpreting model results in the third article in this series.

Before that, in the second article, we will see in more detail how we can classify response variables and explanatory variables as either quantitative or categorical, and how these classifications dictate the appropriate types of univariate and bivariate exploratory analyses and more formal statistical modelling we can perform on those data.

References

Abrams, M. (1983). *People in their late sixties: A longitudinal survey of ageing, part I survivors and non-survivors*. Mitcham, Surrey: Age Concern Research Unit.

Connolly, S., Rosato, M., & O'Reilly, D. (2011). The effect of population movement on the spatial distribution of socio-economic and health status: Analysis using the Northern Ireland mortality study. *Health & place*, *17*(4), 1007-1010.

Davies, R. B. (1992). The state of the art in survey analysis. In A. Westlake & et al (Eds.), *Survey and statistical computing*: Elsevier Science Publishers B.V.

Fox, A. J., & Goldblatt, P. O. (1982). *Longitudinal study socio-demographic mortality defferentials*. London: OPCS, HMSO.

Grundy, E., & Sloggett, A. (2003). Health inequalities in the older population: the role of personal capital, social resources and socio-economic circumstances. *Social science & medicine, 56*(5), 935-947.

Hirdes, J., & Forbes, W. (1989). Estimatesof relative risk of mortality based on the Ontario longitudinal study of aging. *Canadian Journal of Aging, 8*(3), 222-237.

Houweling, T. A., & Kunst, A. E. (2010). Socio-economic inequalities in childhood mortality in low-and middle-income countries: a review of the international evidence. *British Medical Bulletin, 93*(1), 7-26.

Jones, D. R. (1987). Heart Disease Morality Following Widowhood: Some Results From the OPCS Longitudinal Study. *Journal of Psychosomatic Research, 313*, 325-333.

Kaplan, G., Barell, V., & Lusky, A. (1988). Subjective state of health and survival in elderly adults. *J of Gerontology, 434*, s114-120.

Kaplan, G. A., Seeman, T. E., Cohen, R. D., Knudsen, L. P., & Guralnik, J. (1987). Mortality Among the Elderly in the Alameda County Study: Behaviourial and Demographic Risk Factors. *American Journal of Public Health, 773*, 307-312.

Palmore, E. (1982). Predictors of the Longevity Difference: A 25-Year Follow-up. *The Gerontologist, 226*, 513-518.

Penn, R., & Berridge, D. (2010). *Social Statistics Volume 1: The Statistical Analysis of Aggregate Categorical Data*. London: Sage Publications.

Pfeiffer, E. (1970). Survival in old age. *Journal of the American Geriatric Society, 184*, 273-285.

Risnes, K. R., Vatten, L. J., Baker, J. L., Jameson, K., Sovio, U., Kajantie, E., . . . Painter, R. C. (2011). Birthweight and mortality in adulthood: a systematic review and meta-analysis. *International journal of epidemiology, 40*(3), 647-661.

Shahtahmasebi, S., & Berridge, D. (2010). *Conceptualising behaviour in health and social research: a practical guide to data analysis*. New York: Nova Sci.