ISSN 2382-1019

Paper 3. The Exploratory Analysis of Data on Survival in Old Age

Damon Berridge, PhD
Farr Institute – CIPHER, College of Medicine, Swansea University, Swansea. UK.
Keywords: multicollinearity, multivariate, socio-economic, parsimonious model
Received: 30/8/2014; Accepted: 25/9/2014 (**Republished**)

Editor's note: by republishing this series I am hoping to draw attention to a widespread lack of consideration paid by researchers to substantive theory guiding methodology and analysis methods and interpretation of results. In an article in the previous issue of DHH (https://journalofhealth.co.nz/?page_id=1692) using teenage smoking data I demonstrated the negative implications of poor research leading to misleading conclusion, and pitfalls of ignoring statistical concepts and substantive theory even when the statistical model is the right choice for the type of data at hand. For more examples see Shahtahmasebi and Berridge (2010). Here is another reminder in the final instalment of the series on statistical modelling.

1. Introduction

In the first paper in this series (Berridge, 2018a), I explained how the statistical modelling framework can be used to analyse multivariate health research data in an efficient manner. I discussed the need to control for sources of variation in search of parsimony. The degree of control that can be applied in any given context depends upon a number of factors, including the study design and the nature of the data.

In the second paper (Berridge, 2018b), I outlined a number of issues which can arise when modelling multivariate health research data. These issues include the use of substantive theory to formulate one or more research hypotheses, which involves the identification and classification of one or more response and explanatory variables. The classification of variables determines the sorts of univariate and bivariate numerical and graphical summaries we use to take a preliminary look at the data. The results of these exploratory analyses will inform the statistical modelling process, including the choice of an appropriate type of model.

In the current paper, the third in the series, I return to the example of survival in old age introduced in Berridge (2018a). Death is an inevitable outcome of old age given natural processes, yet some people die soon after retirement while others may survive to a ripe old age. An obvious question would be what factors may explain the variation in death (or survival) post-retirement. On the other hand, due to a heterogeneous survivor population, defining the outcome as surviving or dying may not be sufficiently sensitive to distinguish random effects from systematic effects of some quality of life variables. We may increase the complexity in the analyses by modelling a multi-categorical outcome variable (for example, an outcome comprising three categories: survivor and in the community; survivor and in care; died) or by modelling the time to death (or duration survived).

In the current paper, I will outline the exploratory analyses of data on survival in old age that was reported in detail in Chapter 2 of Shahtahmasebi and Berridge (2010). The purpose of this paper is not to replicate the results and their interpretation - the interested reader is pointed in the direction of Shahtahmasebi and Berridge (2010) - but rather to explain the rationale behind the decisions made during the analytical process.

This series of papers will conclude with the fourth and final paper, in which I will re-introduce the issue of control in this particular context and will illustrate the benefits gained as a result of incorporating control into the analytical framework through the application of a series of statistical models. I will work through the stages of applying increasingly complex statistical models for the analysis of survival in old age, highlighting the issues of control. I will conclude by comparing and contrasting the benefits and drawbacks of the different models.

Returning to the current paper, in the following section I will describe the exploratory analyses performed, and how these preliminary analyses informed the way in which the outcome variable(s) and the explanatory variables were operationalised as a prelude to the more formal statistical modelling.

2. Exploratory Analyses

2.1 Univariate Analyses

The simplest form of analysis is commonly known as univariate analysis in which we describe the distribution of each variable in the dataset, including the outcome variable(s) and the explanatory variables. In the current context, the initial outcome variable 'survival' is binary (whether an individual survives or not) and the majority of the explanatory variables are categorical, and one-way frequencies and simple barcharts are appropriate numerical and graphical summaries.

These univariate summaries serve a number of purposes. First, they indicate the number of original categories which comprise each variable. For variables with large numbers of (ordinal) categories, for example, age bands, some (adjacent) categories may have to be combined in order to reduce the number of categories, especially in the case of small sample sizes. In this dataset, age comprised three categories: '65 to 74', '75 to 79' and '80 or older'.

The other main reason for producing univariate summaries is to quantify the proportion of missingness in each variable. Most variables in the current dataset had few missing values. However, there were rather more missing values for variables such as income, number in network, morale, and the measures of isolation and loneliness. For the income variable, fieldwork experience indicated that the majority of refusals to answer were from higher income respondents. Therefore, all such cases were allocated to the upper income bracket. A separate non-response category was created in order to handle missingness in the remaining variables.

2.2 Bivariate Analyses

The next step in the analytical process is to perform bivariate analyses in order to explore the relationship between two variables. There are two possible scenarios:

(i) One variable is the outcome variable, for example, survival or not; the other variable is an explanatory variable, for example, grouped age;

(ii) Both variables are explanatory variables, for example, grouped age and gender.

A major concern with bivariate exploratory analyses such as (i) is that we would ignore the influence of other explanatory variables such as gender on the outcome (for example, survival or not) that may be operating through the selected explanatory variable, namely age. In other words, in examining the effect of age on survival, we would be ignoring the potentially high association between age and gender as indicated by analysis (ii). We will need to allow formally for control in the analysis. In the next paper in this series, we will illustrate the process of building in control into the analysis through the use of a series of statistical models. In the current paper, we return to discuss the bivariate analyses (i) and (ii). We consider each scenario in turn.

To explore the relationship between a categorical explanatory variable and a categorical outcome variable, we can construct a two-way contingency table (also known as a cross-tabulation or a cross-classification) and perform a chi-squared test of association on the observed counts in that table. In the current context, Shahtahmasebi and Berridge (2010) produced a series of contingency tables and chi-squared tests in order to explore the association between the categorical explanatory variables and the initial binary outcome variable, survival in old age. I do not reproduce the results of these bivariate analyses here – they are presented in Table 2.1 of Shahtahmasebi and Berridge (2010). Most of the explanatory variables considered were significantly related to survival in old age. Two thirds of the explanatory variables were significantly associated with survival at the 5% level. From the row (or column) percentages in each table, we were able to explore further the nature of the relationship with survival. For example, rate of survival decreased with age, and females were expected to live longer on average than males. Those respondents who classified themselves as

British had a higher survival rate, along with owner occupiers, those individuals in higher income brackets and those in good health.

To investigate how strongly the explanatory variables were interrelated, we computed Cramer's v measure of association for all the explanatory variables. These results are given in Table 2.2 of Shahtahmasebi and Berridge (2010). This table confirms there are complex patterns of inter-relationships between the explanatory variables. Due to the small sample size, a Cramer's v of over 0.2 was considered to be sufficient to indicate a strong association. For example, household composition was associated with a range of variables including marital status, number of children, income, relative seen most often, frequency of contact with family, network type, hours spent alone, and the measures of isolation and loneliness. These variables themselves are interrelated with each other, as well as with other variables. For example, network type is correlated with hours spent alone, isolation, loneliness, morale, health status and visit from the doctor. Therefore, it would be unwise to infer causal effects from Table 2.1. In other words, given the results in Table 2.2, we cannot conclude a direct and independent relationship between each explanatory variable in Table 2.1 and survival. Thus, the question is whether, for example, household composition and network type, which are significant in Table 2.1, reflect social determinants of survival in old age, or whether network type is merely acting as a proxy for age, dependency and/or health effects. A multivariate analysis is therefore required to allow us to make conclusions about the effect of each explanatory variable on survival, having controlled for other explanatory variables.

2.3 Multivariate Analyses

One way of imposing 'control' is through the method of disaggregation, in other words, constructing multi-way cross-tabulations of the outcome variable (in this context, survival) with two or more explanatory variables. There are several problems with this approach. First, it is an extension of the two-way contingency table and will produce numerous tables to examine and interpret. In studies of this type, where a large number of explanatory variables is involved, the examination of a large number of tables will be cumbersome and labour-intensive, and their interpretation may lead to unsatisfactory conclusions. Second, this method may result in (very) sparse tables with a high proportion of cells containing a (very) low frequency. Third, this approach cannot deal satisfactorily with non-categorical outcome variables and explanatory variables. Furthermore, this method provides no way of quantifying and of testing for the statistical significance of the effects of each explanatory variable. An alternative method which allows us to address all of these issues is the statistical modelling approach. In the fourth and final paper in this series, I will outline the statistical modelling strategy to the analysis of data on survival in old age.

References

Berridge, D. (2018a) The Use of the Statistical Model as an Effective Tool in the Analysis of Multivariate Data in Health Research. Dynamics of Human Health (DHH), 1(1): https://journalofhealth.co.nz/?page_id=1586

Berridge, D. (2018b) Paper 2: Issues arising in the statistical modelling of multivariate health data. Dynamics of Human Health (DHH), 1(2): https://journalofhealth.co.nz/?page_id=1697

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