Editors' Introduction: Social Statistics

Roger Penn and Damon Berridge

This collection of articles on social statistics provides a broadly chronological account of the creation and development of statistical methods relevant to the social sciences and in particular, to sociology. Sociological research has posed serious challenges for statistical analysis historically because sociological data are not generally continuous in nature[1]. Statistical methods such as linear regression models and analysis of variance (ANOVA) developed rapidly after Galton's breakthroughs in the 1870s and 1880s which culminated in his book, *Natural Inheritance* (Galton, 1889). However, these methods were based upon assumptions of continuous data, with an underlying normal distribution and were therefore inappropriate for the analysis of much sociological data. New methods had to be created and developed. These four volumes chart these innovations over the last 100 years or so.

Sociological data are often categorical in nature. They may be binary, nominal or ordinal. Binary data (such as 'employed'/'unemployed') have an underlying binomial distribution which can be analysed using logistic and probit models. Nominal data (such as occupational status categories) have an underlying multi-nomial distribution which can be examined by either log-linear modelling or multi-nomial logit modelling. Ordinal responses such as those found in the many Likert[2] items (Likert, 1932) used in survey research can be assessed using proportional odds or cumulative logit models. The generation of these styles of modelling has been a dominant theme during the last 50 years.

The early literature on the analysis of categorical data witnessed a succession of attempts aimed at quantifying the nature of association between two binary variables. The debate between Pearson and Yule became particularly heated. At the heart of their exchanges was a debate about whether categorical variables could be regarded as discrete representations of continuous distributions (Pearson's point of view), or whether binary variables such as ('vaccinated'/'not vaccinated') were inherently discrete (Yule's perspective). Both points of view had some merit. For variables such as religion and ethnicity, the classification is fixed and no underlying continuous distribution is apparent. For other variables, for example social class, the assumption of an underlying continuum is more plausible. Indeed, such an assumption would prove to be integral to the development of later models such as the cumulative logit (see Volume 4).

Pearson (1900) introduced a goodness-of-fit test which constituted the starting point for a subsequent revolution in social statistics. This would

become known as Pearson's chi-squared statistic. Pearson was motivated to develop this test in order to determine whether possible outcomes on a Monte Carlo roulette wheel were equally likely. The chi-squared test statistic had neat mathematical properties[3] which permitted the evaluation of whether probabilities in a multinomial distribution equated to certain prior values. It allowed for the attachment of probabilities to the likelihood of responses occurring by chance (at random). The probability of results occurring less than 1 in 20 times became the standard criterion for measuring statistical significance within the social sciences. This continues today. Indeed, a great deal of routine sociological practice remains locked at this level of descriptive statistical analysis established in the early part of the twentieth century.

These developments closely paralleled the emergence of the notion of hypothesis testing which involved the generation of a 'research hypothesis' and an isomorphic 'null hypothesis' of no association. The research hypothesis itself was often derived from emerging sociological theory. Traditionally, this was seen as an *a priori* starting point for data analysis but later it was also linked to *a posteriori* inductive approaches such as 'grounded theory' (see Glaser and Strauss, 1967). The latter has now become incorporated into contemporary social statistics in the form of 'data mining' (see, for example, Hand, 2001).

Other attempts at measuring association originated in the study of public health (see Macdonell, 1902; 1903) and genetics (see Blanchard, 1902). Their analyses of the nature of statistical association began with the simplest of cases: the 2×2 contingency table (see Macdonell, 1902; Blanchard, 1902). It developed subsequently into the analysis of the more general $2 \times c$ table (see Macdonell, 1903) and culminated in the $r \times c$ case (see Yule 1906a; 1906b)[4].

The presentation of innovative methods in social statistics moved during the inter-war period from biological and medical journals into mainstream statistical journals on both sides of the Atlantic, notably the *Journal of the Royal Statistical Society* (*JRSS*) and the *Journal of the American Statistical Association* (*JASA*). These two journals have remained pivotal for the dissemination of new methods within social statistics and quantitative sociology. The 1930s and 1940s witnessed the further extension of the chi-squared test to handle higher dimensional tables (see Bartlett, 1935; Norton, 1945).

The analysis of complex data structures, such as those encountered in the study of processes of social mobility, required the development of new methods that could permit the estimation of the simultaneous effects of a wide range of explanatory variables. Some of these would be categorical in form (such as social class, gender and ethnicity) but others would be continuous (such as age and income). There was therefore, a pressing need to develop methods that permitted the simultaneous control for categorical and continuous variables, and which also considered likely interactions amongst the explanatory variables themselves. These issues were resolved within a variety of modelling frameworks (see Volumes 2, 3 and 4).

The practical solutions to these statistical problems also relied upon the parallel emergence of computers with sufficient power to fit models. It also

required the development of appropriate software to facilitate their widespread adoption. Initially, social statisticians wrote their own software but the late 1960s saw the emergence of *SPSS* (*Statistical Package for the Social Sciences*) as a standard software package available to social scientists. The combination of *SPSS* software and IBM mainframe computers underpinned the rapid development of the incorporation of new methods of statistical analysis within the social sciences in the 1970s and 1980s, particularly in the United States. Indeed, during this period, key contributions were overwhelmingly published in US statistics and sociology journals.

The main difficulty inherent in much of the traditional analysis of contingency tables was the problem of chronology. Social scientists wanted to be able to examine causal processes such as whether unemployment led to ill health. However, a contingency table that cross-tabulated 'employment'/'unemployment' by 'good health'/'poor health' could not demonstrate such a relationship since a statistically significant correlation between unemployment and illness could equally be the result of illness leading to unemployment. Indeed, a putative correlation could be the result of both processes operating simultaneously.

Longitudinal data offered a solution to this impasse. In the example above, a longitudinal research design producing simultaneous data over time on states of health and on employment status would permit the disentangling of the two processes, by the measurement of the two relative sets of effects. However, longitudinal analysis brought its own problems that also needed new solutions (see Volumes 3 and 4).

Volume 1: The Fundamentals of Descriptive Social Statistics

The earliest articles in this volume laid down the fundamental principles for the exploratory analysis of categorical data. A classic example of such data was social class as measured by occupational groups (see Macdonell, 1903: Table IX, p.138) [1.5]. Social class often comprised several occupational categories, but Macdonell reduced it to a binary variable comprising two categories, 'higher class'/'lower class'.

The relationship between two binary variables could be summarised using a 2×2 contingency table or crosstabulation. The higher and lower classes of fathers were cross-classified by the higher and lower classes of their sons (see Macdonell (1901, 1903) [1.1, 1.5] and Blanchard (1902) [1.3][5]).

The relationship between two binary 'attributes'[6] (or variables) was expressed in terms of a correlation or an association (see Yule, 1903) [1.4]. Association was defined in terms of the presence of dependence, or equivalently, the absence of independence, between two binary variables (see Yule, 1903) [1.4]. The correlation or association displayed in a single 2×2 table was encapsulated in a single summary statistic. Early attempts at developing

such a statistic included the 'coefficient of correlation' (see Macdonell, 1903; Pearson *et al.*, 1903; Yule, 1906a) [1.5, 1.6, 1.7] and the 'coefficient ratio' (see Pearson, 1910) [1.10].

1.1 Data Comprising More Than Two Variables

Even in these early studies, there was an appreciation that processes were rarely bivariate in nature, and could not be summarised adequately using a single 2×2 table. More often, there was a complex series of inter-relationships between more than two binary variables. This was initially resolved by the construction of a series of 2×2 tables (see Macdonell (1901, 1903) [1.1, 1.5]).

The ideas underlying the theory of complete independence were generalised from two variables to more than two (see Yule, 1904) [1.4]. However, modelling approaches, better able than hypothesis testing to handle such multivariate data, would not start to be developed until the second half of the twentieth century (see Volume 2 onwards).

1.2 Variables Comprising More Than Two Categories

As time passed, there was a recognition that two categories were insufficient to capture much social scientific data. In the early research on heredity, 'unmeasured characters' (or categorical variables) such as eye-colour were categorised into more than two categories: these included light/medium/dark hues (see Yule, 1906b: pp.337–8) [1.8] or, with greater resolution, eight tints (see Yule, 1906a: p.330) [1.7].

There was an appreciation that collapsing such multi-categorical data into two categories (see Macdonell, 1903, p.139) [1.5] was problematic for several reasons. The first involved the loss of potentially useful information. The second involved the problem that different ways of dichotomising groupings led to different crosstabulations[7] and resulted in different coefficients of correlation (see Macdonell, 1903) [1.5].

1.3 A Typology of Categorical Data

As a general rule, different types of data require different types of analysis. This rule applies to the analysis of categorical data. Multi-categorical variables can be classified into two types: ordered (or ordinal) and unordered (or nominal).

In the earlier example on heredity, Yule (1906a) [1.7] wished to examine the association between the heights of fathers and their sons. He started his analysis by categorising the fathers' and sons' heights into three categories, thereby creating two new ordered variables, each comprising three categories. These ordered variables were cross-classified to produce a 3×3 table.

However, he subsequently analysed this 3×3 table by means of a series of 2×2 tables. Unfortunately, this served to increase the complexity of the problem rather than to resolve the underlying issue.

1.4 Chi-Squared Test

Several statistics such as Yule's *Q* (see Yule, 1900, 1912) were developed to quantify strength of association. However, no formal test of statistical significance was attached to them. Pearson (1907) concluded: 'I cannot therefore accept Mr Yule's test as very likely to be helpful in measuring deviation from Gaussian…' (pp. 470–1). To rectify this situation, Pearson (1907) [1.9] himself calculated expected frequencies assuming a Gaussian (normal) distribution and applied the chi-squared test to a 3×3 table that examined the transmission of eye colour from father to son (pp. 471–2).

Subsequently, Pearson (1916) [1.11] applied the chi-squared test statistic for testing the null hypothesis of no association. He referred the chi-squared test statistic to the table of critical values for the chi-squared distribution presented earlier by Palin Elderton (1902) [1.2]. In a later article, Fisher (1922) [1.12] clarified the use of the chi-squared test for contingency tables. In particular, he stated that one should take, not the number of cells but the number of degrees of freedom, to produce more accurate p-values.

Yule (1922) [1.13] also applied the chi-squared test to compare observed and expected frequencies. He applied the test to data from a series of physical experiments. Fisher (1923) [1.14] then used the chi-squared test to compare data observed from an experiment with two members of the family of chi-squared distributions (p.145) in order to assess which of the two displayed the better fit. A correction to the chi-squared test for 2×2 tables with small cell counts was proposed subsequently by Yates (1934) [1.16].

Hotelling (1931) [1.15] provided a useful summary of developments at that time, including the chi-squared test and Fisher's 'likelihood'. In the discussion of his article, Shewhart made an observation which still stands today: 'The logic of discovery is beginning to take on a new form, and today, the subject of statistical inference is of interest to all scientists.'

1.5 Interactions

In a 2×2 table comprising two variables A and B, the first order interaction is denoted by 'A $\times$ B'. This indicates how A varies with B. The main effects comprise the marginal distributions of A and of B. Once the two main effects of A and B have been taken into account, the first order interaction can be tested. This principle was extended to higher dimensional tables. As Bartlett (1935) [1.17] explained, 'For a $2 \times 2 \times 2$ table … the only additional problem to consider… is the testing of the second order interaction (A $\times$ B $\times$ C).'

1.6 High Dimensional Tables

Norton (1945) [1.18] presented a scheme of successive approximations for $2^N \times R$ tables[8]. He made an important point which, even today, is sometimes difficult to convey to undergraduates in social statistics and social science: '… a main effect may appear to be nil even though one of the interactions is clearly significant' (p.251)!

Yates (1952) [1.19] provided a concise summary of developments in social statistics by the middle of the twentieth century.

Kastenbaum and Lamphiear (1959) [1.22] developed an iterative procedure to estimate parameters of a multinomial distribution in a $(r-1) \times (s-1) \times (t-1)$ crosstabulation and calculated the corresponding chi-squared test statistic. They emphasised the importance of being able to implement their methods using the latest technology: 'It is the purpose of this paper to demonstrate a technique which… is particularly well suited for modern high-speed computers' (p.107).

Volume 1 concludes with a wide-ranging article by Yates (1984) [1.23], which summarised the developments covered in the volume. He provided an excellent review of early work (Section 3). He also reviewed the chi-squared test, including criticisms of the exact test and continuity correction (Section 11).

Volume 2: The Development of Statistical Modelling

In Volume 1, we examined how the analysis of contingency tables focused on the testing of the null hypothesis of no association using the chi-squared test. In Volume 2, we explore how methods increased in sophistication as researchers started to examine issues such as social mobility (see Goodman, 1969 [2.5]; Hauser, 1978 [2.8]; Duncan, 1979 [2.9]; Penn and Dawkins, 1983 [2.11]; Hout, 1988 [2.15]; Biblarz and Raftery, 1993 [2.19]) and occupational mobility (see Duncan, 1979 [2.9]; Sobel *et al.*, 1985 [2.13]). Social and occupational mobility, together with marital endogamy, were areas in which statisticians and social scientists interacted fruitfully during the 1970s and 1980s.

In Volume 2, we witness the transition from hypothesis testing to statistical modelling. Goodman and Kruskal (1954) [2.1] reviewed traditional measures of association and declared, '… we propose the construction of probabilistic models…' (p. 735). Existing methods for handling two-way contingency tables were adapted to analyse three-way cross-tabulations. Darroch (1962) [2.2] proposed a likelihood ratio test for the hypothesis of no second-order interaction ($A \times B \times C$), and used Kastenbaum and Lamphiear's (1959) [1.22] earlier $2 \times 3 \times 5$ example. Darroch (1962) [2.2] also extended his approach to handle four-way tables ($A \times B \times C \times D$).

Birch (1963) [2.3] showed how interactions could be defined as certain linear combinations of logarithms of expected frequencies which could be estimated using maximum likelihood (ML). These ML estimates were

shown to be identical for a range of different sampling methods (Section 2). He also demonstrated that these ML estimates had three elegant mathematical properties:

(i) likelihood equations had a unique solution at which the likelihood is maximised;
(ii) ML estimates were the same under a wide variety of sampling conditions;
(iii) ML estimates were sufficient statistics[9].

These results were generalised to many-way contingency tables (Section 5). Birch (1965) [2.4] subsequently presented an early form of the log-linear model (see, for example Nelder, 1974 [2.7]; Penn and Dawkins, 1983 [2.11] later in Volume 2) involving three variables I, J and K. He demonstrated that the hypothesis of a common cross-product ratio between variables I and J across all levels of variable K was, in its general form, equivalent to Roy and Kastenbaum's (1956) hypothesis of no three-factor interaction.

2.1 Social Mobility

Goodman (1969) [2.5] presented methods for examining two-way social mobility tables. Social class was classified into three categories: 'upper' (U), 'middle' (M) and 'lower' (L). He divided each 3×3 table into a series of 2×2 sub-tables, for example, father's status (U or M) versus son's status (U or M). He defined the interaction in each 2×2 sub-table in terms of an 'odds ratio'. For example, the odds ratio of having 'destination' M rather than 'destination' U was defined for an individual whose origin is M, relative to an individual whose origin is U. If there were no difference in odds between these two individuals, then the odds ratio would take a value of 1 or, equivalently, the log odds ratio would take a value of 0. Goodman denoted the log odds ratio by G. By deriving a standard error S of G, he used the Z score (G/S) to test whether G was significantly different from zero.

The limitations of such bivariate analyses necessitated the development of multivariate techniques which could combine hypothesis testing and statistical modelling. The advantages of statistical modelling over hypothesis testing were threefold:

(i) It allowed for the examination of the **joint** effects of a set of *n* explanatory variables $\{X_1, X_2, \ldots, X_n\}$ on response variable Y simultaneously.
(ii) It permitted **control** of a set of secondary explanatory variables $\{X_1, X_2, \ldots, X_{n-1}\}$ by including them in the model first, before assessing the relative significance of the factor of primary interest, X_n.
(iii) It enabled the assessment of the statistical significance of **both** main effects and interactions between explanatory variables.

2.2 Major Methodological Developments in the 1970s

2.2.1 Generalised Linear Models

A major advance in data analysis came with the formalisation of the family of generalised linear models by Nelder and Wedderburn (1972). One member of this family, with Poisson error distribution and log link function, defined the log-linear model. Goodman (1972) [2.6] applied log-linear models to the analysis of a 2^4 table of counts. The goodness of fit between nested models was tested using chi-squared and likelihood ratio tests. Nelder (1974) [2.7] subsequently used a log-linear model to relate the log count (or frequency) to a set of explanatory variables. This early work by Nelder would lead to the development of the statistical software package Generalized Linear Interactive Modelling, GLIM (see Francis *et al*., 1993) which has been a cornerstone for the development of complex statistical modelling in the social sciences.

2.2.2 EM Algorithm

In their highly influential article, Dempster *et al*. (1977) [3.4] developed the EM algorithm, an iterative procedure comprising an expectation (E) step followed by a maximisation (M) step. The simplicity and generality of the associated theory would provide the inspiration for much later work (see Volumes 3 and 4). When the underlying data follow the exponential family, whose maximum likelihood estimates are easily computed, each M step is similarly easily computed. Dempster *et al*. (1977) [3.4] provided many examples including 'missing data situations, applications to grouped, censored or truncated data, finite mixture models, variance component estimation, hyperparameter estimation, iteratively reweighted least squares and factor analysis'.

2.3 Occupational Mobility

The exploration of occupational mobility provided the basis for several methodological developments. Hauser (1978) [2.8] proposed a multiplicative model for an observed crosstabulation of respondents' occupations and their fathers' occupations at an earlier time. The model expressed the expected frequencies as the product of four elements: an overall effect, a row effect, a column effect and an interaction effect. The cells in the mobility table were assigned to K mutually exclusive and exhaustive subsets, and each of these subsets shared a common interaction parameter. Thus, aside from total, row and column effects, each expected frequency was determined by only one parameter which reflected the level of mobility or immobility in that cell, *relative* to that in other cells in the table.

Duncan (1979) [2.9] analysed occupational class in terms of eight groupings (p.794) and fitted a series of five different models to the resulting 8×8

occupational mobility table. He used likelihood ratios to compare 'nested' or 'hierarchical' models (see Figure 1, p. 799).

In their own study of occupational mobility, Sobel *et al*. (1985) [2.13] made the distinction between 'reciprocated' and 'unreciprocated' mobility. They matched concepts of structure and exchange to parameters of a quasi-symmetry model. Exchange or 'reciprocated' mobility was defined as the part of the mobility process that resulted from equal flows between pairs of occupational categories. Structural mobility was defined as the effect of marginal heterogeneity operating uniformly on such origins.

Whilst these articles developed the traditional interest in mobility tables, with the emphasis on the distinction between diagonal and off-diagonal cells, they suffered from an increasing propensity to generate complex and arcane explanations. This was resolved by the application of log-linear analysis. Penn and Dawkins (1983) examined matrices of marital endogamy to assess the extent to which men and women from different social classes intermarried in Britain between 1850 and 1964. The concinnity of their analyses exemplified the principle of parsimony and cut through much of the unnecessary complexity within earlier approaches.

Biblarz and Raftery (1993) [2.19] tested the hypothesis that family disruption affected subsequent occupational outcomes. They applied a log-linear model to a four-way table of counts, cross-classified by father's occupation, son's occupation, race ('White'/'Black') and family type ('intact'/'nonintact').

2.4 Canonical Analysis

Canonical analysis has often been employed instead of log-linear modelling to analyse the relationship between two categorical variables. However, until the mid-1980s, canonical analysis had taken place on an *ad hoc* basis. Gilula and Haberman (1986) [2.14] positioned canonical analysis within a more formal theoretical framework. They examined models that placed non-trivial restrictions on the values of canonical parameters so that a parsimonious description of association could be obtained. Parameter estimation was by maximum likelihood, and approximate confidence intervals were derived. Adequacy of models was assessed by using chi-squared tests. The resulting models could be used to determine the appropriateness of latent class analysis. They could also be used to determine whether a set of canonical scores had specified patterns. Comparisons with a log-linear parameterisation of the cell probabilities revealed that canonical analysis, which used interpretations based on regression and correlation, was an effective alternative to log-linear parameterisations based on cross-product ratios.

Gilula and Haberman (1988) [2.16] extended this approach from two-way tables to multivariate scenarios, but these were – somewhat paradoxically – reduced to two dimensions. Responses were treated by them as a single categorical variable, whilst the explanatory variables were likewise reduced to

a single factor. In this way, a multi-way table was reduced to a two-way array to which traditional canonical and association models could be applied. However, this rush to simplification was at the expense of much of the information inherent in the data under scrutiny.

2.5 Latent Variable/Latent Class Analysis

An alternative form of analysis involved the concept of latent classes which underpinned the processes under observation. Clogg and Goodman (1984) [2.12] used four dichotomous response items to measure attitudes towards science and scientific careers. Responses to these items were cross-classified by 'intelligence' (high IQ or low IQ). They proposed a latent structure analysis through the introduction of two (or more) latent classes. Item responses were assumed to be conditionally independent of each other, given such latent class membership. Clogg and Goodman (1984) [2.12] showed how their alternative formulation led directly to the familiar log-linear model.

Lindsay *et al*. (1991) [2.17] fitted a range of models to educational testing data. These models included the Rasch model which was defined as a function of item difficulty and subject ability parameters, and the T-class mixture model in which the subject ability parameters were treated as random effects. In the latter model, respondents were assumed to be drawn from a population with T latent classes. This model was extended by maximising the mixture likelihood over all possible latent class structures which were modelled non-parametrically. The final class of models described by Lindsay *et al*. (1991) [2.17] were conventional models in which, within each latent class, response probabilities were independent with unknown and item-specific success probabilities.

Volume 3: Statistical Modelling of Longitudinal Data

Much of the classical work in social statistics in the first half of the twentieth century focused on the analysis of cross-sectional data. This became institutionalised in the routine use of crosstabulations and associated chi-squared tests to analyse synchronic data. However, social science has had a longstanding interest in diachronic processes of social change. Such concerns permeated the early development of sociological theory as seen in the works of Marx, Durkheim and Weber (see Giddens, 1971).

An early reference to the concept of a 'model' applied to longitudinal data was made by Anderson and Goodman (1957) [3.1] (see Section A: State Dependence below). Heckman (1978) [3.5] subsequently formulated and estimated simultaneous equation models which included both discrete and continuous endogenous variables. His models relied on the idea that discrete

endogenous variables were generated by continuous latent variables crossing thresholds, a concept that dated back to Pearson (1900). Heckman explained how dummy endogenous variables served two distinct roles: firstly, as proxies for unobserved latent variables ('spurious effects') and secondly, as direct shifters of behaviour ('true effects'). He demonstrated that these two roles needed to be carefully distinguished.

In his seminal 1981 article, Heckman (1981a) [3.7] addressed a variety of issues which arose specifically in the analysis of longitudinal categorical data and which set the agenda for subsequent methodological developments in the field over the next 30 years. He formulated a general dynamic model for discrete panel data that could be used to analyse the structure of 'discrete choices' (categorical outcomes or responses) over time. He generated a rich group of stochastic processes based upon time-discrete outcomes, including Markov models (see Section A: State Dependence) and renewal processes. This family of models was flexible enough to accommodate time-varying explanatory variables, general correlation structures for random effects and complex interrelationships among 'decisions' (outcomes) taken at different times.

A. State Dependence

State dependence is present in a dynamic social process if the probability of an individual being in a given state at time [t+1] depends on the state that same individual occupied at time [t]. A classic example of such state dependence was the axiom of cumulative inertia proposed by McGinnis (1968) [3.3]: the longer an individual spends in a given state such as any particular social class, the less likely such an individual would be to leave that state (p.716). Indeed, some individuals would always, or almost always, remain in the same state regardless of duration. The existence of such individuals, known as 'stayers', in a longitudinal dataset presents special problems analytically (see Section C below).

Anderson and Goodman (1957) [3.1] generated maximum likelihood estimates for transition probabilities in a Markov chain. Chi-squared and likelihood ratio tests were derived for testing whether the transition probabilities of a first order chain were constant, and when the transition probabilities were constant, whether they had specified numerical values.

Heckman's (1981a) [3.7] method could be used to address the longstanding problem of distinguishing between true and false contagion (Bates and Neyman, 1951), or in the language of Heckman (1981a) [3.7], distinguishing between 'spurious' and 'true' state dependence. In other words, state dependence may exist through serial correlation in the observables generating the event ('spurious state dependence') or because past experience of the event affects subsequent behaviour ('structural state dependence') (p.167).

Davies *et al*. (1992) [3.10] handled state dependence by adding a dummy variable (previous state) to their model (for more details, see Section B below), rather than conditioning on previous state[10].

B. Residual Heterogeneity

A central problem in the social sciences involves the existence of unmeasured or, indeed, unmeasurable explanatory variables. If such factors are at work, this will be evident in the structure of residuals to any fitted models. Appropriate diagnostic tests to identify such potential variables were initially identified by Heckman (1981a) [3.7], and further developed by Davies and Crouchley (1985) [3.9] and Winship and Mare (1992) [3.11].

Heckman (1979) [3.6] argued that selection bias can also be a contributory factor generating 'omitted variable' bias within a regression analysis. He proposed the incorporation of an explicit individual-specific error term ('random effect') within the modelling framework. This random effect was assumed to follow a normal distribution.

Davies and Crouchley (1985) [3.9] showed that uncontrolled heterogeneity can lead to bias in the estimation of coefficients for any exogenous variables included in the model, and can result in extremely misleading estimates of standard errors. They developed random effects models in order to control for heterogeneity, thereby minimising such bias. These models, which included Markovian structure, were initially applied to British election data (see Davies and Crouchley, 1985). Subsequently, Davies *et al*. (1992) [3.10] used an early version of SABRE to fit a logistic-normal random effects model to employment data in Britain.

C. Movers and Stayers

The mover-stayer model, a generalisation of the Markov chain model, assumed there are two types of individuals in the population under scrutiny: the 'stayer' who had a propensity to remain in the same state (such as 'employed') during the entire period of study, and the 'mover' whose tendency to change state (i.e. become 'unemployed') over time could be described by a Markov chain with a transition probability matrix.

The transition probability matrix for movers and the proportion of stayers were unknown parameters. Goodman (1961) [3.2] presented various estimators of these parameters and compared their accuracy, as well as describing tests of several hypotheses concerning the mover-stayer model.

Davies and Crouchley (1985) [3.9] argued that stayers could be handled within a random effects framework by adding a 'spike' (or concentration) of stayers to the appropriate mixing distribution. They handled stayers in states 1 and 0 by supplementing a normal mixing distribution with mass points at plus and minus infinity.

D. Initial Conditions

Before a statistical model can be fitted to event history data, the social process under scrutiny needs to be appropriately specified. In much social science research, this initial conditions problem is treated casually. Naive assumptions are often made: the initial conditions or relevant pre-sample history of the social process are often assumed to be exogenous. This assumption is valid only if the responses that generate the process are serially independent or if a genuinely new process fortuitously starts at the beginning of the observation period.

Alfo and Aitkin (2000) [3.16] presented a solution to the initial conditions problem. They modelled the initial response separately whilst second and subsequent responses were allowed to depend on previous response(s).

Wooldridge (2005) [3.19] proposed modelling the mixing distribution, conditional on the initial response and any exogenous explanatory variables. For the binary response model with a lagged response, Arellano and Carrasco (2003) proposed a maximum likelihood estimator conditional on the initial response, where the mixing distribution was assumed to be discrete. In the binary case, Wooldridge's approach was more flexible and computationally much simpler than that of Arellano and Carrasco in three respects. Firstly, the response probability could take either the probit or the logit form. Secondly, strictly exogenous explanatory variables could be incorporated easily, along with a lagged response variable. Thirdly, standard random effects software could be used to estimate the resulting parameters.

Specifying a mixing distribution conditional on the initial condition has several advantages. Firstly, the mixing distribution can be chosen flexibly and regarded as an alternative approximation to that of Heckman (1981b) [3.8]. Secondly, in several cases, including the binary probit model, a mixing distribution can be chosen that leads to straightforward estimation using standard software.

E. Dropout

Dropout or 'attrition' occurs at random when each respondent in a panel survey is equally likely to drop out at any given wave. When dropout is correlated with the social process under scrutiny, dropout or attrition is said to be 'informative' about that process. The most efficient and effective way of handling dropout is to model the response and dropout processes simultaneously (see Little, 1995 [3.13]).

Joint models for response and dropout can be classified into two broad types: selection models and pattern-mixture models. The difference between the two lies in the manner in which the joint distribution of the response-dropout mechanism is factored.

A major problem encountered in the development of such joint models was that results were often sensitive to assumptions about the nature of

dropout itself. Winship and Mare (1992) [3.11] provided a development of Heckman's (1979) [3.6] two-stage estimator, and outlined a number of semi- and non-parametric approaches to estimating selection models that relied on weaker assumptions about the nature of the dropout process. Little (1995) [3.13] suggested that sensitivity analyses should be performed in order to assess the effect on inferences of alternative assumptions about the dropout process itself.

Follmann and Wu (1995) [3.12] proposed a separate set of models for continuous responses with binary dropout processes which were linked by a common random parameter. An approximation to the response model was conditioned on dropout, thereby precluding the need to specify an explicit dropout model. This was extended by Ten Have *et al*. (1998) [3.15] to handle longitudinal binary responses subject to informative dropout by the use of a shared parameter model with logistic link. Molenberghs *et al*. (1998) [3.14] developed pseudo-likelihood solutions for the same problem. Albert (2000) [3.17] subsequently developed a transition model to analyse longitudinal binary data subject to non-ignorable missingness. Roy and Lin (2005) [3.18] considered the situation in which dropout was informative and both outcome and time-varying covariates were missing at the time of dropout. Both of these approaches used the EM algorithm developed some 25 years earlier by Dempster *et al*. (1977) [3.4].

Volume 4: Statistical Modelling of Ordinal Categorical Data

In the fourth and final volume, the issues raised in the general context of event history data, and more specifically repeated binary data, are extended to the modelling of ordinal categorical data. This area has been at the forefront of recent developments in the linkage between social scientists and social statisticians.

Ordinal data are widespread in the social sciences but have traditionally been neglected. Conventional statistical analyses have generally involved simplifying ordinal responses as binary outcomes[11]. Such an approach loses much of the resolution inherent within such ordinal data. Another strategy, often adopted by researchers, has been to apply a scoring system to ordered categories. The resulting variable is treated as continuous, and linear regression models which assume the normal distribution are routinely used. However, this strategy violates a range of assumptions about the nature of ordinal data: in particular, it ignores the fact that they are rarely normally distributed.

Agresti (1981) [4.4] presented measures for summarising the strength of association between an ordered categorical variable and a nominal variable. For a $2 \times c$ table, these measures included Somers' d, discrete analogues of the Mann-Whitney test and Goodman and Kruskal's (1954) [2.1] γ. For the $r \times c$ case, Agresti (1981) [4.4] constructed two generalised measures that

were expressed in terms of event probabilities concerning two types of pairs of observations, including an alternative representation of Freeman's (1965) ϑ index.

Agresti (1981) [4.4] also outlined two models: firstly, a log-linear model which assumed that a set of ordered scores could be assigned to the response categories; secondly, a cumulative logit model which did not require such a scoring system.

The cumulative logit model had been described earlier by McCullagh (1980) [4.2]. Within the cumulative logit model, the original ordinal response was represented by a series of binary responses. For example, a Likert item such as 'Strongly Agree' (SA), 'Agree' (A), 'Neither Agree Nor Disagree' (NAND), 'Disagree' (D), 'Strongly Disagree' (SD) would be represented by the four binary responses: (i) SA vs. A to SD; (ii) SA, A vs. NAND to SD; (iii) SA to NAND vs. D, SD; (iv) SA to D vs. SD (see Berridge and Penn, 2009, for an illustration of this approach). The model intercepts were allowed to vary between these four partitions. The effects of explanatory variables were assumed to remain constant across all partitions[12].

McCullagh (1980) [4.2] suggested other models that did not require continuous scores. These included the complementary log-log model, a discrete-data version of Cox's proportional hazards model (see Cox, 1972), and the adjacent-logit model. He also made a distinction between symmetric and asymmetric models. A symmetric model was one in which the results were invariant to a reversal in the order of the categories. The cumulative logit (or proportional odds) model was an example of such a symmetric model, and remains the natural starting point when analysing a Likert item.

Asymmetric (or sequential) responses were appropriately analysed using models such as the continuation ratio. This was originally developed for the analysis of educational data (see Fienberg and Mason, 1979; Mare, 1980, 1981) [4.1, 4.3, 4.5].

The multi-level model for ordinal data, like the continuation ratio model, also originated in the area of educational research, and was closely related to the random effects and variance component models introduced in Volume 3. We have not provided an extensive review of multi-level models here[13]. However, to illustrate their applicability for modelling ordinal data, we have included in Volume 4 a selection of influential articles within which multi-level models have been applied to data on educational attainment (see Aitkin and Longford, 1986; Fielding and Yang, 2005; Penn and Berridge, 2008) [4.7, 4.18, 4.20] and earnings (see Davies *et al.*, 1988) [4.8].

Aitkin and Longford (1986) [4.7], in their general analysis of clustered educational data, argued for the use of variance component or 'random parameter' models. They applied a range of linear models to data on 907 pupils in 18 schools from one Local Education Authority.

Fielding and Yang (2005) [4.18] fitted a series of multi-level cumulative logit models to data on 'A'-level grades[14], cross-classified by student and teaching group within a number of educational institutions.

Their initial models were defined at three levels:

Level 1: a set of explanatory variables to explain 'A'-level grades, lodged within the cross-classification of a particular student and teaching group at level 2, within a particular institution at level 3.
Level 2: independent teaching group and student random effects.
Level 3: an institution-specific random effect.

To incorporate teacher effects, Fielding and Yang examined a subset of data on a small number of further education colleges. They used a fixed effects specification for such colleges by incorporating dummy variables into a set of explanatory factors. The resulting models were defined at two levels:

Level 1: as per the initial model, but without the link to a particular institution at level 3.
Level 2(a): independent teaching group and student random effects, as per the initial model.
Level 2(b): a set of teacher-specific explanatory variables and teacher-specific random effects, both weighted according to each teacher's share of the complete teaching timetable of a teaching group over the two years of its provision.

Penn and Berridge (2008) [4.20] analysed a wide range of school-level and locality-level factors that affect each of the three stages in a young adult's educational trajectory in England: GCSE results, subsequent path taken at age 16 and 'A'-level results. By applying three-level models to data collected as part of the EFFNATIS[15] project, they found no evidence of any locality-level effects. Furthermore, none of the factors conventionally considered to affect educational attainment such as gender, social class and ethnicity had a consistent effect across all three stages. Rather, each factor had a contingent effect at specific points within the overall trajectory of educational outcomes.

Earlier, Davies *et al.* (1988) [4.8] presented a linear model that was also appropriate for examining such hierarchical data. Their model had conventional regression, variance component and random coefficient models as special cases and was fitted using the software VARCL (Longford, 1990). The effectiveness of the model was demonstrated through an analysis of earnings in the British engineering industry. Particular emphasis was placed by them on the interpretation of parameter estimates and residuals.

During the 1980s, extensions to earlier ordinal models were developed. Anderson (1984) [4.6] developed the stereotype model. The choice between stereotype models was made empirically on the basis of comparative model fit. This was deemed particularly important for 'assessed, ordered categorical response variables', where it was not obvious *a priori* whether the ordering was relevant to the regression relationship. Model simplification was assessed in

terms of whether the response categories were distinguishable with respect to the vector of explanatory factors.

Research subsequently progressed from the analysis of cross-sectional ordinal data to the modelling of repeated or clustered ordinal data. Exactly the same issues that arose in the analysis of longitudinal categorical data, outlined in Volume 3, needed to be addressed in the context of repeated or clustered ordinal data.

A. State Dependence

Albert *et al*. (1997) [4.12] proposed a class of latent Markov chain models for analysing repeated ordinal responses with medical diagnostic misclassification. They modelled the underlying monotonic response and the misclassification processes separately, and proposed an EM algorithm (see Dempster *et al.*, 1977 [3.4]) that allowed for ordinal parameterisations, time-dependent covariates and randomly missing data.

Pudney (2008) [4.21] subsequently developed a simulated maximum likelihood approach for the estimation of dynamic linear models, where the continuous response variable was observed through the prism of an ordinal scale. He applied the latent auto-regression (LAR) model to BHPS data on households' perceptions of their financial well-being, and demonstrated the superior fit of the LAR model to the state dependence (SD) model.

B. Residual Heterogeneity

Crouchley (1995) [4.9] developed a random effects model for multivariate and grouped (clustered) ordered categorical data. He preferred the complementary log-log link function to other link functions such as the logit, probit and log-log (McCullagh and Nelder, 1989). The complementary log-log link function had the advantage that it provided a closed-form expression for the likelihood of the model, unconditional on the random effect over a wide range of distributions. This likelihood was computed without recourse to numerical integration or Gaussian quadrature. Crouchley assumed that the distribution for the random effects belonged to the Hougaard (1986) family of distributions[16].

C. Movers and Stayers

Ekholm *et al*. (2003) [4.17] modelled the probabilities of particular sequences of responses ('path probabilities'). In order to specify the association between responses measured for the same 'generic unit' or individual, they used measures of association that were ratios of moments[17] ('dependence ratios'). The two-way dependence ratio (of order two) between any two responses from the same individual was defined as the ratio of the moment of order two

relative to the product of the respective moments of order one. For meaningful association, a strong structure had to be imposed on the dependence ratios by deriving them from a vector of association parameters[18]. In Ekholm *et al.*'s association model, dependence ratios were expressed as explicit functions of association parameters, explanatory variables and time. An explicit expression for an individual's path probability was then derived which permitted the calculation of that individual's contribution to the overall likelihood.

D. Initial Conditions

Pudney (2008) [4.21] indicated that, in the state dependence (SD) model, there are two alternative approaches for dealing with random effects. Heckman (1981b) [3.8] specified an approximation to the distribution of the initial response, conditional on strictly exogenous explanatory variables and on random effects. He then derived the distribution of subsequent responses, conditional on the initial response, explanatory variables and random effects, by using sequential conditioning. The alternative approach, used by Wooldridge (2005) [4.19], was to specify the mixing distribution, conditional on the initial response and exogenous explanatory variables (see Volume 3).

Wooldridge's (2005) [4.19] binary probit model can be extended in a straightforward manner to fit a dynamic ordered probit model. If the response comprises ordered categories, then an ordered probit model can be specified with lagged response indicator variables and strictly exogenous explanatory variables. The observed value of the ordered response variable is determined by a latent variable falling in a particular interval, where the cutpoints must be estimated, as in the cumulative logit model (see McCullagh, 1980 [4.2]). If the mixing distribution, conditional on both the initial response and strictly exogenous explanatory variables, is specified as having a homoscedastic normal distribution, then standard random effects ordered probit software can be used.

The Wooldridge approach has advantages in SD models but is less attractive in LAR models, where the latent response of interest is not observable and cannot be conditioned upon. For this reason, Pudney (2008) [4.21] used the Heckman treatment of initial conditions to model longitudinal panel data.

E. Dropout

In recent years, there has been considerable work examining the issue of repeated ordinal data subject to dropout or 'attrition'. Some approaches have made strong assumptions about the nature of dropout. One of these was the population-averaged approach of Molenberghs *et al.* (1997) [4.10]. They developed a multivariate Dale model combined with logistic regression for dropout. The association between observations was measured by a global odds ratio. Response and dropout processes were modelled as conditionally

independent given complete data, resulting in a likelihood that could be estimated reliably using the EM algorithm. They cautioned that the interpretation of the results of any model depended upon the specific assumptions made and also that a parallel sensitivity analysis should be performed.

There is an important difference between the analyses of Molenberghs *et al*. (1997) [4.10] and those of Ekholm *et al*. (2003) [4.17]. The association parameters used by Molenberghs *et al*. involved global odds ratios which varied over time but which did not discriminate between movers and stayers. Ekholm *et al*. concentrated attention on the characteristics of 'paths travelled' (or the sequences of responses). They examined individual paths by specifying meaningful models in terms of local, rather than global, association measures.

Another method for handling ordinal data subject to dropout involved subject-specific approaches. The random effects model for ordinal data developed by Crouchley (1995) [4.9] has been extended in a variety of ways. Sheiner *et al*. (1997) [4.11] developed a cumulative logit model which was conditional upon subject-specific random effects, and an empirical model for dropout ('censoring') which was conditional upon observed responses and subject-specific random effects. Following the marginal likelihood approach proposed by Davies and Crouchley (1985) [3.9], random effects were assumed to be normally distributed and then integrated out. Hedeker and Mermelstein (2000) [4.14] suggested an alternative random effects model which allowed individual changes over time to be estimated.

Other subject-specific approaches have made fewer assumptions about the nature of dropout. Ten Have *et al*. (2000) [4.15] followed the shared parameter approach of Follmann and Wu (1995) [3.12] by proposing a mixed effects logistic model, in which the random effects included random intercepts and random slopes (see Davies *et al*. (1988) [4.8]). The ordinal outcome and 'survival' (or, conversely, dropout) processes were assumed to share a common random effects structure. Dropout was also assumed to be conditionally independent of outcome, given such underlying random effects. Once again, a parallel sensitivity analysis was suggested.

Agresti and Natarajan (2001) [4.16] presented a review of Bayesian and non-Bayesian methods for clustered (or repeated) ordered categorical data. Their review concentrated on two classes of model. The first involved marginal (or population-averaged) models, including generalised estimating equations. The second entailed cluster-specific (or subject-specific) models which included the maximum likelihood estimation of random effects models as well as a Bayesian approach that used a Markov Chain Monte Carlo (MCMC) approach.

In the context of (ordinal) item response theory (IRT), Bradlow and Zaslavsky (1999) [4.13] developed a Bayesian hierarchical analysis comprising three levels[19]. Inferences were made using samples from the posterior distributions in order to compute point and interval estimates. Samples were obtained from

three different MCMC samplers: data augmentation, a Metropolis step and the Griddy-Gibbs sampler.

The IRT theme was continued in a non-Bayesian context by Liu and Hedeker (2006) [4.19]. They proposed a random effects model that allowed for three-level multivariate ordinal outcomes and accommodated multiple random subject effects. This approach allowed different item factor loadings (item discrimination parameters) to be estimated for multiple outcomes. The explanatory variables in their models were not required to follow the assumption of proportional odds.

Concluding Remarks

In this series, we have demonstrated how the discipline of social statistics has evolved over the last century or more. In recent years, the trend has been towards increasingly rigorous formulation of research hypotheses, larger and more detailed datasets, more complex statistical models to match such data, and a more sophisticated level of statistical analysis in the major international social science journals.

The advent of e-social science and high performance computing has opened up the way to easy replication of results, and has helped to produce standards of scientific rigour within the social sciences which are comparable to those in the natural and physical sciences.

There has been a long and distinguished track record of interdisciplinary collaborative research between statisticians and social scientists. There remain considerable opportunities for continued innovation at the interface between statistics and social science. We hope, therefore, that this present series will inspire a new generation of social statisticians to develop innovative methods of direct use to their social scientific colleagues.

Notes

1. Continuous measures are common in the natural sciences. They include such variables as height, weight and Body Mass Index (BMI). In the social sciences, they feature in psychology (IQ) and economics (income, profits, GNP).
2. Likert items measure attitudes and allow respondents to choose between a range of options ordered in terms of 'strongly agree', 'agree', 'neither agree nor disagree', 'disagree' and 'strongly disagree'. A Likert scale is the sum of responses across several Likert items.
3. Departures of observed frequencies from expected frequencies produce greater chi-squared values for a fixed sample size n. The p-value of the test is the null probability that the chi-squared test statistic takes a value at least as large as the observed value.
4. The relationships in such tables produced a range of indices of association such as Yule's *Q*, Goodman and Kruskal's *tau*, and *lambda* (see Blalock, 1972, Chapter 15).
5. Blanchard's article actually compared the characteristics of horses: specifically 'grandsires' and 'grandchildren'.

6. Whilst much of the conceptual foundation of contemporary social statistics was developed during this period, the terminology used varied amongst authors. This is signalled by our use of quotations. Wherever necessary, we have provided current terms as appropriate.
7. This issue was taken up once again by Penn and Dawkins (see Volume 3) in their log-linear analysis of marital class endogamy. Their analysis revealed that dichotomising their seven occupational groupings into a 'middle class'/'working class' divide produced a gross distortion of the underlying social processes.
8. N refers to the number of variables comprising two categories. R refers to a single variable comprising more than two categories.
9. For example, the distribution of a variable X is characterised by a single unknown parameter, θ. In many of the estimation problems encountered, the information contained within a sample of observations $\{x_1, \ldots, x_n\}$ from this distribution can be summarised. Some function of the sample values (usually the mean) provides as much information about θ as the sample itself. Such a function would be sufficient for estimation purposes (see Mood, Graybill and Boes, 1974, pp. 299–300).
10. This became possible in later versions of the statistical software package, Software for the Analysis of Binary Recurrent Events, SABRE (http://www.sabre/lancs.ac.uk/).
11. See, for example, Scott et al. (1998).
12. This is known as the proportional odds assumption (cumulative logit models fitted under this assumption are known as proportional odds models).
13. See the Sage series on multi-level modelling edited by Skrondal and Rabe-Hesketh (2009).
14. These are assumed to be ordinal.
15. The Effectiveness of Integration Strategies in Europe. This was funded by the European Union's Fourth Framework. The wider results of the project are presented in Penn and Lambert (2009).
16. Three well-known members of this family are the gamma, the inverse Gaussian and the positive stable law distributions (Aalen, 1988).
17. For realisations $Y_1, Y_2, \ldots$ of a variable Y, the first moment is defined as the expectation of Y_1, the second moment is the expectation of $Y_1 \times Y_2$, etc.
18. See the 'manifest classes' or interaction parameters of Hauser (1978) (2.8).
19. Probabilities of individual responses to items were modelled as a function of person- and item-specific parameters θ_1 and general parameters θ_2 (regression coefficients corresponding to the covariates of interest); that is, $[Y|\theta_1, \theta_2]$. The distribution of $[\theta_1|\theta_2]$ was modelled and the prior distributions $[\theta_2]$ were specified.

Additional References

Aalen, O.O. (1988) 'Heterogeneity in survival analysis', *Statistics in Medicine*, 7, 1121–1137.

Agresti, A. (1990) *Categorical Data Analysis*, New York: Wiley.

Arellano, M. and Carrasco, R. (2003) 'Binary choice panel data models with predetermined variables', *Journal of Econometrics*, 115, 125–157.

Bates, G. and Neyman, J. (1951) 'Contributions to the Theory of Accident Proneness II: True or False Contagion', *University of California Publications in Statistics*, 1, 215–253.

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.

Blalock, H.M. (1972) *Social Statistics (2nd edn.)*, New York: McGraw-Hill.

Cox, D.R. (1972) 'Regression models and life-tables', *Journal of the Royal Statistical Society, Series B*, 34, 187–220.

Davies, R.B. and Crouchley, R. (1985) 'The determinants of party loyalty: a disaggregate analysis of panel data from the 1974 and 1979 general elections in England', *Political Geography Quarterly*, 4, 4, 307–320.

Francis, B., Green, M. and Payne, C. (1993) *The GLIM System: Release 4 Manual*, Clarendon Press: Oxford.

Freeman, L.C. (1965) *Elementary Applied Statistics*, New York: Wiley.

Galton, F. (1889) *Natural Inheritance*, London: Macmillan.

Giddens, A. (1971) *Capitalism and Modern Social Theory*, Cambridge: Cambridge University Press.

Glaser, B. and Strauss, A. (1967) *The Discovery of Grounded Theory*, Chicago: Aldine.

Hand, D. (2001) *Principles of Data Mining*, Cambridge, MA: MIT Press.

Hougaard, P. (1986) 'Survival models for heterogeneous populations derived from stable distributions', *Biometrika*, 73, 387–396.

Likert, R. (1932) 'A technique for the measurement of attitudes', *Archives of Psychology*, 140, 1–55.

Longford, N.T. (1990) *VARCL: Software for Variance Component Analysis of Data with Nested Random Effects (Maximum Likelihood)*, Technical Report, Educational Testing Service, Princeton, NJ.

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models (2nd edn.)*, London: Chapman and Hall.

Mood, A.M., Graybill, F.A. and Boes, D.C. (1974) *Introduction to the Theory of Statistics (3rd edn.)*, McGraw-Hill: Singapore.

Nelder, J.A. and Wedderburn, R.W.M. (1972) 'Generalized Linear Models', *Journal of the Royal Statistical Society, Series A*, 135, 370–384.

Pearson, K. (1900) 'On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling', *Philosophical Magazine, Series 5*, 50: 157–175 (Reprinted 1948 in K. Pearson's *Early Statistical Papers*, ed. by E.S. Pearson, Cambridge: Cambridge University Press).

Penn, R. and Berridge, D. (2008) 'Modelling Trajectories through the Educational System in North West England', *Education Economics,* 16, 4, 411–431.

Penn, R. and Lambert, P. (2009) *Children of International Migrants in Europe*, Palgrave Macmillan: London.

Scott, J., Braun, M. and Alwyn, D. (1998) 'Partner, parent, worker: family and gender roles', in R. Jowell, J. Curtice, A. Park, L. Brook, K. Thomson and C. Bryson (eds) *British- and European-Social Attitudes; The 15th Report: How Britain Differs*, Aldershot: Ashgate.

Skrondal, A. and Rabe-Hesketh, S. (2009) *Multilevel Modelling*, Sage: London.

Yule, G.U. (1900) 'On the association of attributes in statistics', *Philosophical Transactions of the Royal Society of London*, A194, 257–319.

Yule, G.U. (1912) 'On the methods of measuring association between two attributes (with discussion)', *Journal of the Royal Statistical Society*, 75, 579–642.