Analysis of categorical data for complex surveys

Analysis of Categorical Data for Complex Surveys

Chris Skinner

Department of Statistics, London School of Economics and Political Science, London WC2A 2AE, United Kingdom

E-mail: C.J.Skinner@lse.ac.uk

Summary

This paper reviews methods for handling complex sampling schemes when analysing categorical survey data. It is generally assumed that the complex sampling scheme does not affect the specification of the parameters of interest, only the methodology for making inference about these parameters. The organisation of the paper is loosely chronological. Contingency table data is emphasized first before moving on to the analysis of unit-level data. Weighted least squares methods, introduced in the mid 1970s along with methods for two-way tables, receive early attention. They are followed by more general methods based on maximum likelihood, particularly pseudo maximum likelihood estimation. Point estimation methods typically involve the use of survey weights in some way. Variance estimation methods are described in broad terms. There is a particular emphasis on methods of testing. The main modelling methods considered are log-linear models, logit models, generalized linear models and latent variable models. There is no coverage of multilevel models.

Key words: pseudo maximum likelihood; Rao-Scott adjustment; score test; survey weight; weighted least squares.

1 Introduction

Categorical variables predominate in social surveys and categorical data analysis has been a major theme in the development of methods of survey data analysis. This paper

will review methods for handling complex sampling schemes when analysing categorical survey data. Such methods were first introduced in a systematic way in the 1970s (e.g. Koch et al., 1975; Rao and Scott, 1979). However, the application of categorical data analysis methods to survey data has a longer history and might be taken to originate after World War II (Alwin and Campbell, 1987).

The two or three decades after 1945 saw important developments in categorical data analysis both in relation to surveys and more generally. Survey data was increasingly analysed in sociology, particularly influenced by the work of Lazarsfeld, who introduced different ways of using multi-way contingency tables to explain relationships between variables in causal contexts (e.g. Lazarsfeld, 1968; Duncan, 1982). Sociological research also motivated a range of modelling developments (e.g. Goodman 1970, 1972, 1979), such as in the use of log-linear models, logit models and latent class models with survey data. This period also played an important part in the development of such modelling methods in categorical data analysis in general, extending methods of regression and multivariate analysis for continuous variables to the case of categorical variables. In addition to Goodman's work at the University of Chicago, Agresti (2013, Ch. 17) highlights developments in maximum likelihood for log-linear and logit models at Harvard University, including the landmark book on log-linear models by Bishop et al. (1975). He also identifies research at the University of North Carolina by Koch and his colleagues and students as particularly influential in the biomedical sciences. To these one can add the introduction of generalized linear models by Nelder and Wedderburn (1972) at Rothamsted Experimental Station, unifying methods for categorical and continuous data.

How to take account of complex sampling schemes in such categorical data analysis methods defined an agenda for research in the 1970s, 1980s and beyond and this is the research which will be reviewed here. In addition, research was still needed to accommodate complex sampling in classical methods which originated much earlier, especially testing in two-way tables. Skinner et al. (1989) distinguish aggregated and disaggregated approaches to taking account of complex sampling schemes. We focus in this paper on the aggregated approach, where the definitions of the parameters (and model) of interest

take no account of complex population features such as stratification or clustering underlying the sampling. These features are only relevant to inference. The disaggregated approach would take the complex population features into account in the specification of the model of interest, for example via fixed or random effect terms. We shall not pursue this approach here and, in particular, we shall not consider multilevel models. See Muthén and Satorra (1995) for further discussion of this distinction in the context of structural equation modeling.

In this paper we suppose that all variables of interest are categorical and distinguish *contingency tables* and *unit-level data*. In the former case, we suppose that sample cell counts in the table (excluding structural zeros) are sufficiently large that central limit theorem arguments will provide a satisfactory approximation to the sampling distribution of the vector of cell-level proportions and that the covariance matrix of this vector can be estimated satisfactorily (e.g. Lehnen and Koch, 1974). In the latter case, we suppose that this is not true, normally because the number of cells in the table is very large relative to the sample size so that the sample cell counts are sparse and it is more natural to treat the data as a microdata file at the level of the underlying unit.

In many applications of categorical data analysis it is common to use Poisson models for *counts*. This may occasionally be appropriate for survey data when count variables arise at the unit level, e.g. the number of children ever born to a women (Bellhouse and Rao, 2002). However, for survey-based contingency tables, it is usually not natural to specify models for sample cell counts because of their arbitrary dependence on sample size aspects of the sampling design. It is more usual to model finite population *proportions* or underlying model *probabilities*. With contingency tables, it is usually straightforward to define finite population cell proportions, for example the proportion of men aged 20-25 in the population who are unemployed and there may be interest in modelling such proportions. With unit-level data, finite population proportions are often not meaningful. For example, if we define a combination of categories of a large number of variables for which there is only one individual in the population, the proportion of such individuals who are unemployed will be either 0 or 1. In such a setting, it will usually be more

interesting scientifically to treat the finite population as drawn stochastically from a super-population and to suppose that the unemployment status of the individual here follows some binary response model, for which the model parameters and associated probabilities are of interest. Of course, model parameters and probabilities may also be of interest in the contingency table case, where model-based analogues of finite population proportions can be defined.

We distinguish between *symmetric models*, where the attribute variables defining the table of interest are treated jointly in a symmetric way and *asymmetric models*, where, in the most common case, we are interested in the relation between one attribute, y say, and a vector of other attributes, $\mathbf{x}$ say.

In the symmetric setting, the cells of the table formed by cross-classifying the attribute variables may be denoted $1, \ldots, K$ and we may denote the corresponding finite population cell proportions or model cell probabilities by $\pi = (\pi_1, \ldots, \pi_K)'$, where $\sum_k \pi_k = 1$. Models for such tables include log-linear models and latent class models and these typically express π in terms of a lower-dimensional parameter θ.

In the asymmetric setting, we may let $\pi_{i|j}$ denote the conditional probability (or corresponding finite population proportion) that y falls into category i given that the combination of categories taken by $\mathbf{x}$ can be labelled j. Supposing that I is the number of categories of y and that J is the number of possible combinations of categories of $\mathbf{x}$, we have IJ possible values of $\pi_{i|j}$, with $\sum_i \pi_{i|j} = 1$ for each j, and these may be collected into a $IJ \times 1$ vector π of interest. Models in such settings include logit models and again typically express π in terms of a lower-dimensional parameter θ.

The organisation of the paper is loosely chronological. The earliest contributions in the 1970s are outlined in section 2 on the general method of weighted least squares and in section 3 on the narrower but important case of two-way tables. The general class of methods based upon maximum likelihood are outlined in section 4. This includes discussion of the analysis of contingency tables, as in the earlier sections, but also introduces the analysis of unit-level data in section 4.3. This methodology is extended to a general estimating equation approach in section 5. The paper concludes with section 6 on latent

variable modelling and some remarks in section 7 on the take-up of the methods in the substantive scientific literature.

2 Weighted Least Squares

A general class of models for contingency tables is defined by

$$F(\pi) = \mathbf{X}\theta, \tag{1}$$

where $F(.)$ is a known smooth function, $\mathbf{X}$ is a known design matrix, with rows that depend on the values of the attribute variables associated with the corresponding elements of π and θ is an unknown vector of parameters. This class includes symmetric models, such as a log-linear model where π consists of a vector of cell probabilities π_k and the elements of $F(\pi)$ consist of the logarithms of these probabilities. It also includes asymmetric models, such as a logit model where π consists of a vector of conditional probabilities $\pi_{i|j}$ and the elements of $F(\pi)$ consist of the logits of these conditional probabilities.

In a seminal paper, Grizzle et al. (1969) proposed weighted least squares (WLS) as a general approach to estimation and testing for such models. Although they made standard multinomial assumptions and did not discuss complex sampling schemes, the framework they introduced lends itself naturally to extensions to complex sampling. For purposes of inference, Grizzle et al. (1969) assumed there is an observed vector of sample proportions $\mathbf{p}$, which is unbiased for π either under multinomial assumptions for a symmetric model or under product multinomial assumptions for the asymmetric model. Grizzle et al. (1969) also assumed that the covariance matrix of $\mathbf{p}$ can be expressed as $V(\pi)$, a function of π.

To apply WLS, the idea is to consider the linear model $F(\mathbf{p}) = \mathbf{X}\theta + \delta$, treating $F(\mathbf{p})$ as the 'dependent variable' and taking δ as the estimation error $\delta = F(\mathbf{p}) - F(\pi)$. An estimated covariance matrix of δ is obtained by linearization as $\mathbf{S} = \mathbf{H}(\mathbf{p})V(\mathbf{p})\mathbf{H}(\mathbf{p})'$, where the matrix $\mathbf{H}(\pi)$ contains the partial derivatives of $F(\pi)$ with respect to π. The WLS estimator of θ is then given by

$$\hat{\theta} = (\mathbf{X}'\mathbf{S}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{S}^{-1}F(\mathbf{p}). \tag{2}$$

The extension to complex surveys is then straightforward in principle (Lehnen and Koch, 1974; Koch et al., 1975; Shuster and Downing, 1976). It is assumed that there exists a consistent estimator $\hat{\pi}_c$ of π and a consistent estimator $\hat{\mathbf{V}}_c$ of the covariance matrix of $\hat{\pi} - \pi$, where c indicates complex design. The estimator $\hat{\pi}_c$ might, for example, involve sample weighting and the estimator $\hat{\mathbf{V}}_c$ might involve survey sampling variance estimation techniques, such as linearization or replication. The WLS estimator then takes the same form as in (2), with $V(\mathbf{p})$ replaced by $\hat{\mathbf{V}}_c$ and $F(\mathbf{p})$ replaced by $F(\hat{\pi}_c)$. We write the estimator as $\hat{\theta}_c$ and write $\mathbf{S}_c = \mathbf{H}(\hat{\pi}_c)\hat{\mathbf{V}}_c\mathbf{H}(\hat{\pi}_c)'$. An estimator of the covariance matrix of $\hat{\theta}_c$ is $\hat{V}(\hat{\theta}_c) = (\mathbf{X}'\mathbf{S}_c^{-1}\mathbf{X})^{-1}$.

The extension of test procedures is also straightforward. Grizzle et al. (1969) proposed testing the goodness of fit of model (1) by referring the Wald test statistic $(F(\mathbf{p}) - \mathbf{X}\hat{\theta})'\mathbf{S}^{-1}(F(\mathbf{p}) - \mathbf{X}\hat{\theta})$ to a χ^2 distribution with appropriate degrees of freedom. In the complex survey case (Koch et al., 1975; Shuster and Downing, 1976), $F(\mathbf{p})$ is replaced by $F(\hat{\pi}_c)$, $\hat{\theta}$ by $\hat{\theta}_c$ and $\mathbf{S}$ by $\mathbf{S}_c$. Nested linear hypotheses $H_0 : \mathbf{C}\theta = \mathbf{0}$ about θ may also be tested, where $\mathbf{C}$ is a matrix of arbitrary constants of full rank. Thus, the Wald statistic $\hat{\theta}_c'\mathbf{C}'[\mathbf{C}\hat{\mathbf{V}}(\hat{\theta}_c)\mathbf{C}']^{-1}\mathbf{C}\hat{\theta}_c$ is referred to a χ^2 distribution with degrees of freedom given by the rank of $\mathbf{C}$.

A complication in the WLS approach is the assumption in (2) that $\mathbf{S}$ (or $\mathbf{S}_c$) is non-singular. In practice, the vector π may contain linear dependencies which induce singularity in $\mathbf{S}$. For example, Grizzle et al. (1969) considered the asymmetric case where the vector π contains IJ conditional probabilities $\pi_{i|j}, i = 1, \ldots, I; j = 1, \ldots, J$. The constraints $\sum_i \pi_{i|j} = 1$ for each j create J linear dependencies in π. To avoid singularity, Grizzle et al. (1969) removed the J linear dependencies in π by defining $F(\pi)$ to be of dimension u, where $u \leq (I-1)J$. Here u is chosen so that the $u \times IJ$ matrix $\mathbf{H}(\pi)$ is of full rank u (and hence that $\mathbf{S}$ is non-singular); u may be strictly less than $(I-1)J$ if there are zero counts in the table. Different ways of redefining $F(\pi)$ to avoid singularity in $\mathbf{S}$ may be needed for different models. See, for example, Grizzle and Williams (1972) for log-linear models. Two alternative approaches to handling the singularity of $\mathbf{S}_c$ were proposed by Scott et al. (1989). One appeals to the optimal theory of linear models

with singular covariance matrices. The other involves modification of $F(\hat{\pi})$ and the use of g-inverses. These approaches are discussed further by Rao et al.(1989).

A basic concern with the WLS approach in complex surveys is that, although the covariance matrix estimator $\hat{\mathbf{V}}_c$ is consistent, in practice it will typically yield 'far less precision than the multinomial analogues' $V(\mathbf{p})$ 'and this reduced precision has a serious effect on the inversion required in the computation of the Wald statistic' (Fay, 1985, p. 148), that is in the inversion of $\mathbf{S}_c$. Fay (1985, p. 148) continued that 'this instability in the estimated inverse in turn inflates the rate of rejection under the null hypothesis, often enough to make the test unusable'. See Fay (1982) for further discussion.

3 Adjustments to Classical Tests in Two-way Tables

An important special case of the nested hypotheses considered in the previous section is that of independence in a two-way table, classically tested with a Pearson or likelihoood ratio test statistic. The distribution of these test statistics under the null hypothesis of independence can be greatly affected by clustering and stratification (Fellegi, 1980; Rao and Scott, 1981) and the use of such tests can give misleading results in practice. A number of approaches have been derived to correct for these effects. Some approaches used models for clustering (Altham, 1976; Cohen, 1976; Brier,1980), although Nathan (1981) drew attention to unrealistic assumptions in these models. The weighted least squares approach is also available, of course (Nathan, 1972, 1975).

The approach proposed by Rao and Scott (1981, 1984) has proved particularly influential. They started with the classical Pearson and likelihood ratio test statistics, redefined as necessary to handle survey weights. Thus, the Pearson test statistic for an $I \times J$ table takes the form

$$X^2 = n \sum_{i=1}^{I} \sum_{j=1}^{J} (\hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j})^2 / (\hat{p}_{i+}\hat{p}_{+j}), \tag{3}$$

where n is the sample size, $\hat{p}_{ij}$ is the survey weighted estimate of the population proportion in cell ij, and $\hat{p}_{i+}$ and $\hat{p}_{+j}$ are the corresponding marginal sums. They showed

that this statistic and the corresponding likelihood ratio test statistic are asymptotically distributed under the hypothesis of independence as a weighted sum, $\sum_1^T \delta_t W_t$, of independent χ_1^2 random variables W_t, where the weights δ_t are related to familiar design effects used by survey samplers and $T = (I-1)(J-1)$. They developed first and second order approximations to this distribution, which may be used to obtain simple adjustments to the standard test statistic. We refer to these as *Rao-Scott adjustments*. For example, the first order Rao-Scott adjustment to X^2 takes the form $X^2_{RS} = X^2/\hat{\delta}$, where $\hat{\delta} = (\sum \hat{\delta}_t)/T$ and the $\hat{\delta}_t$ $(t = 1, \ldots, T)$ are estimated values of the δ_t. These adjustments are discussed further in a more general setting in section 4.1.

Thomas and Rao (1987) undertook a Monte Carlo evaluation of the significance level and power of alternative goodness-of-fit tests under cluster sampling. They found that the Wald test performed poorly with its actual significance level often greatly exceeding the nominal level. The significance levels for the two Rao-Scott adjustments were much closer to their nominal levels. Thomas and Rao (1987) also proposed a further adjustment, whereby the adjusted test statistic is referred to an F distribution to take account of the fact that the δ_t must be estimated, and this adjustment was found to offer further improvements in significance level and generally to improve performance. Fellegi (1980) proposed a slightly different adjustment but Thomas and Rao (1987) found that its significance level performance was similar to that of the first order Rao-Scott adjustment.

Another approach, involving the construction of a test statistic by jackknifing, was proposed by Fay (1985). Thomas and Rao (1987) found this approach to perform similarly to the second order Rao-Scott adjustment. Fay argued that the approach can be applied for any replication method, such as the bootstrap or balanced repeated replication, which provides a consistent estimate of the covariance matrix of the sample estimates. A rather different bootstrap approach was proposed by Beaumont and Bocci (2009) in which not only is a test statistic constructed but also a null distribution for this statistic is simulated by bootstrapping. See also Lumley and Scott (2014).

Tests of independence for two-way tables are also applicable to testing for differences between two groups on a categorical outcome. When the outcome is ordinal a widely used

test which exploits the ordinality is Wilcoxon's rank sum test. Natarajan et al. (2012) showed how this test can be adapted for complex survey data. They used a proportional odds cumulative logistic regression model framework with an ordinal outcome and a single dichotomous covariate. Within this framework the Wilcoxon test was shown to be equivalent to a score test of no effect of the covariate under multinomial sampling. They extended this test to a complex survey setting using the score test based upon weighted estimating equations proposed by Rao et al. (1998), as described in section 5.

4 Approaches based on Maximum Likelihood

Maximum likelihood (ML) is a very widely used inferential framework for modern categorical data analysis and, for various reasons, its use has tended to supersede WLS (Agresti, 2013, sect. 16.7.3).

4.1 Log-linear models for contingency tables

To illustrate an approach based on maximum likelihood, consider first a log-linear model, where the $K \times 1$ vector π contains the cell proportions in a table with K cells and where, in a similar format to (1), we write

$$\log(\pi) = u(\theta)\mathbf{1} + \mathbf{X}\theta, \tag{4}$$

where the $K \times 1$ vector $\log(\pi)$ contains the logarithms of the elements of π, $\mathbf{X}$ is a known matrix of full rank with $\mathbf{X'1} = \mathbf{0}$, $\mathbf{1}$ is the $K \times 1$ vector of 1's and $u(\theta)$ is a normalizing factor chosen so that $\sum_k \pi_k = 1$.

Under multinomial assumptions, the likelihood equations are given by

$$\mathbf{X}'\pi(\theta) = \mathbf{X}\mathbf{p}, \tag{5}$$

where $\pi(\theta)$ is defined by π in (4) viewed as a function of θ and $\mathbf{p}$ is the vector of sample proportions as in section 2.

In the complex survey case, the *pseudo likelihood equations* are given by

$$\mathbf{X}'\pi(\theta) = \mathbf{X}\hat{\pi}_c, \tag{6}$$

where $\hat{\pi}_c$ is a consistent estimator of π as in section 2, and the solution of these equations $\hat{\theta}_{pml}$ is the *pseudo ML estimator* (Imrey et al., 1982; Rao and Scott, 1984). The asymptotic covariance matrix of $\hat{\theta}_{pml}$ (Rao and Scott, 1984) is given by

$$(\mathbf{X}'\Delta(\pi)\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}_c\mathbf{X})(\mathbf{X}'\Delta(\pi)\mathbf{X})^{-1}, \tag{7}$$

where $\mathbf{V}_c$ is the asymptotic covariance matrix of $\hat{\pi}_c$ and $\Delta(\pi) = diag(\pi) - \pi\pi'$ is the multinomial covariance matrix for a single observation.

Turning to testing, the goodness of fit of a specific log-linear model can be assessed by testing this model as a nested hypothesis against the saturated model, so it is sufficient to focus here just on nested tests. It is possible to construct a Wald test of a nested hypothesis, as noted by Rao and Scott (1984). However, the test can be unstable, as illustrated by Rao and Thomas (1988). Fay (1982, 1985) and Lumley and Scott (2014) outline several problem with this test and we do not pursue it further here. We focus instead on classical Pearson or likelihood ratio tests of nested hypotheses with the pseudo MLE replacing the classical MLE. Although these two tests perform similarly, the likelihood ratio test may be preferred since it is invariant to nonlinear transformations of the parameter vector θ. Rao and Scott (1984) showed, as discussed in section 3, that the asymptotic distribution of each of these classical test statistics under the nested hypothesis is a weighted sum, $\sum \delta_t W_t$, of independent χ^2_1 random variables W_t $(t = 1, \ldots, T)$, where the weights δ_t may be viewed as generalized design effects and T is the number of parameters restricted under the nested hypothesis. The classical Pearson and likelihood ratio test statistics may be expressed as

$$X^2 = n\sum_k \frac{(\hat{\pi}_k - \hat{\pi}^*_k)^2}{\hat{\pi}^*_k}, \quad G^2 = 2n\sum_k \hat{\pi}_k \log\left(\frac{\hat{\pi}_k}{\hat{\pi}^*_k}\right), \tag{8}$$

where $\hat{\pi}_k$ and $\hat{\pi}^*_k$ are the elements of $\pi(\theta)$ implied by the pseudo MLE $\hat{\theta}_{pml}$ under the unrestricted and restricted hypotheses, respectively.

The first order Rao-Scott adjusted Pearson and likelihood ratio test statistics are defined by $X^2_{RS} = X^2/\bar{\delta}$ and $G^2_{RS} = G^2/\bar{\delta}$, where $\bar{\delta} = T^{-1}\sum_1^T \hat{\delta}_t$ is the average of the estimated values $\hat{\delta}_t$ of the δ_t. The test statistics X^2_{RS} and G^2_{RS} are referred to a χ^2

distribution with T degrees of freedom. The second order adjustment divides each of X^2_{RS} and G^2_{RS} by $c_1 = \sum \hat{\delta}_t^2/(T\bar{\delta}^2)$ and refers them to a χ^2 distribution with $c_2 = T/c_1$ degrees of freedom. A basic rationale for the Rao-Scott adjustments is that they match the moments of the test statistic under the null hypothesis with the χ^2 reference distribution. Thus, the first order adjustments match the first moments of the asymptotic distributions of X^2_{RS} and G^2_{RS} under the null hypothesis with T, the first moment of χ^2_T. The second order adjustments match the first two moments of these asymptotic distributions scaled by c_1 with the first two moments of $\chi^2_{c_2}$. In both cases, estimation error in the $\hat{\delta}_t$ is ignored.

One reason given originally for preferring the first order to the second order adjustment was that the latter requires an estimate of the full covariance matrix V_c and this was often not available, especially when undertaking secondary analysis from published tables. On the other hand, expressions for $\bar{\delta}$ (and hence the first order adjustment) could often be calculated using more limited information for many models. For example, in the common case of testing independence in a two way table, Rao and Scott (2004) showed that the first order adjustment only requires information on the cell design effects and marginal row and column design effects. Further discussion of the use of simple kinds of design effect for adjustment is given by Holt et al. (1980), Bedrick (1983), Gross (1984) and Rao and Scott (1987). Subsequently, however, it was found that the second order adjustment provides a more stable test when the full covariance matrix is available. As a result both adjustments are now used in some standard software, such as STATA, SUDAAN and SAS(survey software), sometimes as a default option. See section 7 for further comment.

Further adjustments have been proposed to handle error in estimating $\mathbf{V}_c$ and hence the δ_t. A widely used approach is via an F-adjustment. For example, the second order Rao-Scott adjusted test statistic is divided by T and referred to an F distribution with c_2 and $c_2\nu$ degrees of freedom, where ν denotes the degrees of freedom used to estimate $\mathbf{V}_c$ for the complex survey design and is often taken as the number of primary sampling units minus the number of strata. Lumley and Scott (2014) recommendeded taking c_1 as

1 (its minimum value when the $\hat{\delta}_t$ all equal $\bar{\delta}$) and hence c_2 as T when ν is relatively small. They also reported that a saddlepoint approximation worked very well in simulations.

An alternative simple approach to handling survey weights in log-linear models was proposed by Clogg and Eliason (1987). See Skinner and Vallet (2010) for discussion.

4.2 Logit models for contingency tables

An alterative asymmetric class of models may be defined in terms of conditional probabilities $\pi_{i|j}$. We consider here logit models, where i takes only two values 0 or 1 and define π as a $J \times 1$ vector, containing values of $\pi_{1|j}$ for $j = 1, \ldots, J$. We express the model as

$$\text{logit}(\pi) = \mathbf{X}\theta, \tag{9}$$

where $\pi = (\pi_{1|1}, \ldots, \pi_{1|J})'$ and $\text{logit}(\pi)$ contains elements $\log[\pi_{1|j}/(1 - \pi_{1|j})]$ for $j = 1, \ldots, J$. Roberts et al. (1987) defined the $\pi_{1|j}$ as finite population proportions N_{j1}/N_j, where N_j is the size of domain j in the population and N_{j1} is the number of these units with outcome 1. It is, however, also possible to define the $\pi_{1|j}$ as model probabilities.

In a conventional product multinomial setting, we would suppose that frequencies n_j are given in each domain $j = 1, \ldots, J$ and that frequencies n_{1j} with outcome 1 in each of these domains are determined by independent binomial sampling, $n_{1j} \sim Bin(n_j, \pi_{1|j})$. The likelihood equations then become $\mathbf{X}'D(n)\pi(\theta) = \mathbf{X}D(n)\mathbf{p}$, where $D(n) = \text{diag}(n_j)$, $\pi(\theta)$, as a function of θ, is given in (9) and $\mathbf{p}$ is the vector with elements n_{1j}/n_j.

In the complex survey setting (Roberts et al., 1987), the pseudo MLE $\hat{\theta}_{pml}$ is the solution of

$$\mathbf{X}'D(q)\pi(\theta) = \mathbf{X}D(q)\hat{\pi}_c, \tag{10}$$

where $D(q) = diag(q_j)$, $q_j = \hat{N}_j/\hat{N}$ is the estimated relative size of domain j and $\hat{\pi}_c$ is the complex survey point estimator of π. An estimated asymptotic covariance matrix of $\hat{\theta}_{pml}$ is given by

$$(\mathbf{X}'\hat{\Delta}\mathbf{X})^{-1}(\mathbf{X}'D(q)\hat{\mathbf{V}}_c D(q)\mathbf{X})(\mathbf{X}'\hat{\Delta}\mathbf{X})^{-1}, \tag{11}$$

where $\hat{\Delta} = diag(q_j\hat{\pi}_{1|j}(1 - \hat{\pi}_{1|j}))$, the $\hat{\pi}_{1|j}$ are the elements of $\pi(\hat{\theta}_{pml})$ and $\hat{\mathbf{V}}_c$ estimates

the covariance matrix of $\hat{\pi}_c$ under the complex design. Tests of goodness-of-fit and nested hypotheses about θ, extending the kinds of methods described in section 4.1, were discussed by Roberts et al. (1987).

Extensions to polytomous responses and the use of Box-Cox transformations to handle departures from the logit assumption were discussed by Rao et al. (1989).

4.3 Unit-level models

The analysis of categorical data at the level of the unit rather than the table cell has many advantages. It enables the analysis of large numbers of categorical variables, where the contingency table would be sparse, as well as combinations of categorical and continuous variables. It creates natural links between categorical data analysis and regression analysis, two main themes of survey data analysis. It provides the basis for much modern analysis software which allows for complex surveys. One downside is that it removes a natural goodness-of-fit test, but this can generally be resurrected by a nested analysis if the data correspond to a suitable contingency table. Other approaches to goodness of fit testing are also available, e.g. Graubard et al. (1997).

The role of the sampling scheme in the analysis of unit level data (rather than contingency tables) has been more contested. If a unit-level model is specified to represent individual behavior then some, such as Hoem (1989), have argued that the sample can be treated as an ancillary statistic and likelihood-based inference about the model parameters can proceed ignoring the sampling scheme, except in some special cases such as outcome-based sampling. The survey sampling community has generally been more sceptical about ignoring the sampling scheme, in particular because of inadvertent bias that can be induced by informative sampling (e.g. Chambers, 2003; Fuller, 2009, section 6.3.1).

Binder (1983) is a seminal paper, which took a sceptical view of models and provided a framework for unit-level analysis of complex survey data that has influenced much subsequent applied work and survey software. He considered a class of generalized linear models for unit-level data of the form $(y_i, \mathbf{x}_i)$, where y_i is an outcome variable and $\mathbf{x}_i$ is

a vector of covariates for unit i. The probability density function of y_i was taken to be

$$p(y_i;\theta_i,\phi) = exp[\alpha(\phi)\{y_i\theta_i - g(\theta_i) + h(y_i)\} + \gamma(\phi, y_i)], \tag{12}$$

where the mean of y_i is a function $g'(\theta_i)$ of θ_i, denoted $\mu(\theta_i)$, and it is assumed that $\theta_i = f(\mathbf{x}_i'\beta)$, where $f(.)$ is a known differentiable function and β is unknown. If all finite population values $(y_i, \mathbf{x}_i), i = 1, \ldots, N$ were known then the likelihood equations would be

$$\mathbf{S}(\beta) = \sum_{i=1}^{N}[y_i - \mu\{f(\mathbf{x}_i'\beta)\}]f'(\mathbf{x}_i'\beta)\mathbf{x}_i = \mathbf{0}, \tag{13}$$

where $\mathbf{S}(\beta)$ is the population score function of β. The solution of these equations β_N is often referred to as the *census parameter* and Binder (1983) discussed why this may be of interest and how it may be estimated using design-based inference. The framework includes a wide range of models including logistic regression and log-linear models; it also extends naturally to inference about β, assuming the model holds. Design-based inference about β_N is not dependent on the model, which may be viewed as a working model, or as 'a convenient approximation to the real world' (Binder, 1983, p. 279). Point estimation is obtained by treating $\mathbf{S}(\beta)$ as a vector of population totals, estimating it by sample weighting to give $\hat{\mathbf{S}}(\beta) = \sum_{i\in s} w_i[y_i - \mu\{f(\mathbf{x}_i'\beta)\}]f'(\mathbf{x}_i'\beta)\mathbf{x}_i$, where s denotes the sample and w_i is a sample weight (assumed to enable consistent estimation of population totals) and then solving

$$\hat{\mathbf{S}}(\beta) = \mathbf{0} \tag{14}$$

for β to obtain $\hat{\beta}$. The same point estimator is typically natural for both β_N and β and is often called the pseudo maximum likelihood estimator (Skinner, 1989). Binder (1983) proposed a design-based variance estimator for β_N. Binder and Roberts (2003) argued that, for most models, the same variance estimator can be used for β if the sampling fraction is small. See Korn and Graubard (1998) for further discussion of this point.

Wald tests of hypotheses about β or β_N may be constructed in a similar manner to earlier (Skinner, 1989). Rao et al. (1998) proposed quasi score tests as an alternative approach with advantages compared to Wald tests. Both the Wald and score tests depend on estimators of the covariance matrix of the pseudo MLE and the tests can become

unstable if the degrees of freedom used to estimate this covariance matrix are not large. Rao et al. (1998) proposed alternative tests to handle this case. These include Rao-Scott corrections to naive Wald or score tests which ignore the design. See also Rao and Thomas (2003, sect. 7.5). Lumley and Scott (2014) derived the large sample distribution of the naive likelihood ratio test statistic and showed how Rao-Scott adjustments can be applied to this statistic. They also demonstrated the asymptotic equivalence of the score and likelihood ratio approaches.

The use of a design weight w_i in $\hat{\mathbf{S}}(\beta)$ can lead to loss of efficiency. The loss can be particularly important in case-control studies, where sample selection is based upon the binary outcome y_i. In this setting, Scott and Wild (2002, 2003) and Li et al.(2011b) argued that no weighting or the use of an alternative weighting method may be preferable. For a more general discussion of alternatives to standard design-based weighting in the case of generalized linear models see Pfeffermann and Sverchkov (2003).

5 Estimation Equations Approaches for Unit-level Data

The sample-weighted likelihood equations in (14) provide one example of estimating equations, which may be solved to determine a point estimator. They are obtained from the population score function $\mathbf{S}(\beta)$, which is itself derived from the fully specified parametric model in (12). This modelling assumption may be relaxed in a broader estimating equations approach, where the overall model has two parts: a model of interest and a complementary working model. The former defines the parameters of interest and is required for consistent estimation. The latter (along with the model of interest) determines the point estimator but it is not required to hold for consistency and may be viewed more as a 'nuisance' part of the model. One important example was introduced by Rao et al. (1998) as a *quasi-score* approach. The model of interest for $(y_i, \mathbf{x}_i)$ is specified through the mean of y_i, denoted $\mu_i = \mu(\mathbf{x}_i, \beta)$, and this is acccompanied by

a working model under which the y_i are independent with variances $V_{0i} = V_0(\mu_i)$ (see also Molina and Skinner, 1992). For example, for a binary outcome the working variance could be taken as the binomial variance $V_0(\mu_i) = \mu_i(1-\mu_i)$. The estimating equations now become

$$\hat{\mathbf{S}}(\beta) = \sum_{i \in s} w_i \frac{\partial \mu_i}{\partial \beta} V_{0i}^{-1}(y_i - \mu_i) = \mathbf{0}. \tag{15}$$

Rao et al. (1998) noted that $\hat{\beta}$, the point estimator of β obtained by solving these equations, is consistent under more general conditions than the working model. All that is required is that the finite population can be regarded as a self-weighting sample from the superpopulation. They proposed an approach to variance estimation analogous to Binder (1983) and showed how score tests can be used for testing. Rao et al. (2002) showed how these variance estimation methods and test procedures can be extended to handle poststratification.

A more general approach is obtained for a clustered population by allowing for intra-cluster correlation in the working model. Thus, relabel the observations in the clustered population by $(y_{c\ell}, \mathbf{x}_{c\ell})$, for cluster $c = 1, \ldots, C$ and element $\ell = 1, \ldots, n_c$ in cluster c. The model of interest is again defined through the marginal mean of $y_{c\ell}$, denoted $\mu_{c\ell} = \mu(\mathbf{x}_{c\ell}, \beta)$. Write $\mathbf{y}_c$ as the vector of clustered observations $(y_{c1}, \ldots, y_{cn_c})'$, μ_c as the vector of means $\mu_c = (\mu_{c1}, \ldots, \mu_{cn_c})'$ and $\mathbf{V}_{0c}$ as the working covariance matrix of $\mathbf{y}_c$, now allowed to be non-diagonal. The estimating equations now become

$$\hat{\mathbf{S}}(\beta) = \sum_{c \in s_c} w_c \frac{\partial \mu_c}{\partial \beta} \mathbf{V}_{0c}^{-1}(\mathbf{y}_c - \mu_c) = \mathbf{0}, \tag{16}$$

where s_c is the sample of clusters and w_c is the survey weight for cluster c. It is assumed that there is no subsampling within clusters. Various approaches to specifying and/or estimating $\mathbf{V}_{0c}$ are feasible. Rao (1998) proposed a similar inferential approach to that in Rao et al. (1998), considering both Wald and quasi-score tests. He focussed on the working independence model, where $\mathbf{V}_{0c}$ is diagonal, but does also propose a survey-based approach to handle the non-diagonal case. Rotnitzky and Jewell (1990) proposed Wald, score and likelihood ratio procedures for testing hypotheses about β and, in particular, Rao-Scott adjustments to 'working' tests.

An important special case of this set-up is in longitudinal surveys, where the cluster consists of repeated observations across survey waves. Liang and Zeger (1986) discuss this case (in the absence of complex sampling), where the working covariance matrix is expressed as $\mathbf{V}_{0c} = diag(V_{0c\ell}^{0.5})\mathbf{R}(\alpha)diag(V_{0c\ell}^{0.5})$, $\mathbf{R}(\alpha)$ is the working correlation matrix that may depend on a parameter α and $V_{0c\ell}$ is the working variance of observation $y_{c\ell}$. The estimation equations are referred to as *generalized estimating equations* (GEE). Rao (1998) discussed the extension of this approach to a survey setting. Roberts et al. (2009) extended Rao (1998) by considering also a working model which allows for dependence between repeated observations on a binary outcome variable via odds ratios rather than correlations. They referred to the estimating equations as survey-weighted GEE. They proposed a one-step estimating function bootstrap method for variance estimation. Carrillo et al. (2010) provided further theoretical results and simulation studies and adopted a similar approach to Rao (1998) for the estimation of $\mathbf{R}(\alpha)$.

Li et al. (2011a) discussed an application of the general clustered approach to surveys which collect family-based genetic data and the cluster consists of a family. They specified a model for $\mathbf{V}_{0c}$ as a function of α based upon genetic theory. It is perhaps surprising that the general case of clustered survey data with non-diagonal $\mathbf{V}_{0c}$ has not been considered more. Even if clustering is of no scientific interest, it would still be interesting to know more about how much efficiency is lost by making the working independence assumption. One obstacle to considering such questions is that it is very common in surveys for clusters to be subsampled and this raises complications for inference.

6 Latent Variable Models

Options for handling survey weights and complex designs have increasingly appeared in latent variable modeling software since 2000 (see e.g. Oberski, 2014). Latent variable models for categorical outcome variables include latent class models and categorical factor analysis models as well as structural equation models (e.g. Muthén, 1984).

For latent class models, Patterson et al. (2002) proposed a pseudo maximum likeli-

hood approach to inference. This approach can be formulated in the unit-level framework of section 4.3, where the outcome $\mathbf{y}_i$ is now a vector of categorical variables. However, since there is no unit-level covariate $\mathbf{x}_i$, the probability density function of $\mathbf{y}_i$ is the same for all units in the same cell of the table formed by the outcome variables. It follows that, as for the log-linear model in section 4.1, the census log likelihood can be expressed as a sum over the cells k of $N_k \log \pi_k$, where N_k is the population count in cell k, and that the same holds for the pseudo log likelihood, if N_k is replaced by $\hat{N}_k$, its sample weighted estimate. Point estimation can thus be obtained simply by replacing the observed cell proportions $\mathbf{p}$ by the survey estimates $\hat{\pi}_c$ of the population cell proportions, just as in section 4.1, and employing a standard maximisation procedure, such as the EM or Newton-Raphson algorithm, which would be employed under multinomial assumptions. Vermunt and Magidson (2007) proposed an alternative method, extending an approach of Clogg and Eliason (1987) for fitting log-linear models in which the inverse of cell-specific weights are included as an offset term in the model. Skinner and Vallet (2010) raised concerns, however, about the validity of the standard errors and test procedures generated by this approach. Patterson et al. (2002) made some related points in response to Vermunt's discussion of their paper.

Binary factor analysis of an $A \times 1$ vector $\mathbf{y} = (y^{(1)}, \ldots, y^{(A)})'$ of binary observable variables can be formulated via a threshold model, where a continuous latent variable $u^{(a)}$ underlies observed variable $y^{(a)}$ $(a = 1, \ldots, A)$ with $y^{(a)} = 1$ iff $u^{(a)} \geq \tau_a$ for parameters $\tau_1, \ldots, \tau_A$ and where $\mathbf{u} = (u^{(1)}, \ldots, u^{(A)})'$ is multivariate normal, obeying a classical factor analysis model (e.g. Christoffersson, 1975). Maximum likelihood (or pseudo maximum likelihood) estimation tends to be infeasible in practice, unless A is small, because of the need to compute $A-$dimensional integrals for each observation. More common is a three-stage procedure (e.g. Muthén, 1984), where (i) the τ_a are estimated from univariate likelihoods of the $y^{(a)}$, (ii) the covariance matrix of $\mathbf{u}$, subject to constraints to remove identification indeterminacies, is estimated from bivariate likelihoods of the $A(A-1)/2$ pairs $y^{(a)}, y^{(b)}$, $a \neq b$, holding the τ_a fixed at their values estimated at stage (i), and (iii) the parameters of the factor model are estimated by fitting this model to the covariance

matrix estimated at stage (ii) using a WLS approach. Such a three-stage procedure can be extended to a more general structural model involving a vector of observed covariates $\mathbf{x}$ for each unit, as described in Muthén (1984). This point estimation approach and its consistency extends naturally to complex surveys as described by Aparahouhov (2005). This extension is only required at stages (i) and (ii), where the log likelihoods at each of these stages involve sums over observations and survey weights need to be incorporated in these sums as in a pseudo maximum likelihood approach. Aparahouhov (2005) also discussed variance estimation, following Muthén and Satorra (1995). More research seems needed, e.g. to consider the possible role of the complex design in the choice of weight matrix at stage (iii), to consider alternative variance estimation methods for such three-stage procedures and to consider alternative testing methods.

7 Concluding Remarks

We conclude by discussing the take-up of the new methods described in this review in the substantive scientific literature. We follow Scott (2007) by looking at citations of the Rao and Scott (1981, 1984) papers using the Web of Science. We do not gather other systematic evidence and the comments in this section should be taken as personal.

Our first remark is on the variation in take-up between disciplines. Like Scott (2007), we find that the great majority of recent citations of Rao and Scott (1981, 1984) are in medical, health or biometric applications with many fewer in the social sciences. This contrasts with the introduction of this paper, where we noted how sociological research motivated some of the early developments in the methodology of survey data analysis. As Scott notes, this may partially reflect the differential coverage of the Web of Science or the nature of the software used by different disciplines.

This prompts our second remark that take-up is very dependent upon the software used by different researchers and the kinds of complex survey features the software employ. Scott (2007) notes that the trajectories of numbers of citations of Rao and Scott (1981, 1984) stay at a low level until the mid- to late-1990s, when they took off on an

upward trajectory. Inspection of counts from 2007 to 2016 shows that this upward trajectory has continued. Scott (2007) attributes the change to the time when the Rao-Scott methods were first included in major software packages and in Stata and SAS in particular. A contrasting trajectory of number of citations is for Grizzle et al. (1969), which shows a clear increase for around 15 years from publication, in contrast to the static low level of citations of Rao and Scott (1981, 1984) for their first 15 years, and this may be attributable to software being made available much sooner for the implementation of methods in the former paper.

Thirdly, we note that many of the methods being made available now in software are for unit-level data, where the information about complex surveys consists of survey (and perhaps replicate) weights together with identifiers of strata, clusters or replicates. Hence, the methods described in sections 4.3 and 5 are most readily used. It seems very difficult to identify any modern software which can fit a log-linear model as in section 4.1 using as input only the vector $\hat{\pi}_c$ and an associated covariance matrix estimator $\hat{\mathbf{V}}_c$. As a consequence, there seems to be little take-up currently of methods for contingency tables, as described in sections 4.1 and 4.2. There is an exception - one widely analysed type of contingency table is the two-way table. A very common style of paper among those recently citing Rao and Scott (1981, 1984) has a categorical outcome variable and several explanatory variables, mostly categorical. Such papers often begin with one or more two-way tables of the outcome variable versus one of the key categorical explanatory variables, usually accompanied by tests of independence for the complex design as described in section 3. These are then followed by some kind of regression of the outcome variable on a vector of explanatory variables, as described in section 4.3 or 5.

The fact that software usually assumes now that sufficient complex survey information is included in unit-level data to enable the direct calculation of a covariance matrix estimator $\hat{\mathbf{V}}_c$ contrasts with the world described in the early papers in the 1970s and 1980s. Then it was noted that this full matrix was often not available to researchers in practice and thus consideration was given to the possibility of constructing adjustments

from simpler design effect quantities, as discussed in section 4.1. Scott (2007) notes that 'none of this work seems to have had much impact in practice'. Hence, there is usually no strong practical reason now for using, say, a first order rather than a second order Rao-Scott adjustment.

A final remark is based on the observation that a few of the recent substantive scientific papers which cite Rao and Scott (1981, 1984) do not refer to a probability sampling scheme at all. They make use of the kinds of methods described in this paper because these methods can handle unit-level weighting, which might have been constructed to correct for selection bias in a non-probability sample, or complex data structures, such as clustering. Looking to the future, the class of methods which have been developed to handle complex survey features in categorical data analysis may thus find additional kinds of applications, in practice.

Acknowledgements

The author thanks J.N.K. Rao for comments on the first draft. This work was partially supported by the Simons Foundation. The author thanks the Isaac Newton Institute for Mathematical Sciences, University of Cambridge, for support and hospitality during the programme Data Linkage and Anonymisation which was supported by EPSRC grant no EP/K032208/1.

References

Agresti, A. (2013). *Categorical Data Analysis*, 3rd. Ed. Hoboken: Wiley.

Altham, P.M.E. (1976). Discrete variable analysis for individuals grouped into families. *Biometrika*, **63**, 263-9.

Alwin, D.F. & Campbell, R.T. (1987). Continuity and change in methods of survey data analysis. *Public Opinion Quarterly*, **51**, 139-155.

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling*, *12*, 411-34.

Beaumont, J.F. and Bocci, C. (2009). A practical bootstrap method for testing hypotheses from survey data. *Survey Methoodology*, **35**, 25-35.

Bedrick, E.J. (1983). Adjusted chi-squared tests for cross-classified tables of survey data. *Biometrika*, **70**, 591-5.

Bellhouse, D.R. & Rao, J.N.K. (2002). Analysis of domain means in complex surveys. *Journal of Statistical Planning and Inference*, **102**, 47-58.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, **51**, 29-34.

Binder, D.A. & Roberts, G.R. (2003). Design-based and model-based methods for estimating model parameters. In R.L. Chambers and C.J. Skinner eds., *Analysis of Survey Data*, Chichester: Wiley, pp. 29-48.

Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis*. Cambridge MA: MIT Press.

Brier, S.E. (1980). Analysis of contingency table data under cluster sampling. *Biometrika*, **67**, 591-6.

Carillo, I.A., Chen, J. & Wu, C. (2010). The pseudo-GEE approach to the analysis of longitudinal surveys. *Canadian Journal of Statistics*, **38**, 540-54.

Chambers, R.L. (2003). Introduction to Part A. In R.L. Chambers and C.J. Skinner eds., *Analysis of Survey Data*, Chichester: Wiley, pp. 13-28.

Christoffersson, A. (1975). Factor analysis of dichotomized variables, *Psychometrika*, **40**, 5-32.

Clogg, C.C. & Eliason, S.R. (1987). Some common problems in log-linear analysis. *Sociological Methods & Research*, **16**, 8-44.

Cohen, J.E. (1976). The distribution of the chi-squared statistic under cluster sampling from contingency tables. *Journal of the American Statistical Association*, **71**, 665-70.

Duncan, O.D. (1982). Review essay: statistical methods for categorical data. *American Journal of Sociology*, **87**, 957-60.

Fay, R.E. (1982). Contingency tables for complex designs: CPLX. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 44-53.

Fay, R.E. (1985). A jackknifed chi-square test for complex samples. *Journal of the American Statistical Association*, **80**, 148-157.

Fellegi, I.P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. *Journal of the American Statistical Association*, **75**, 261-8.

Fuller, W.A. (2009). *Sampling Statistics*. Hoboken: Wiley.

Goodman, L.A. (1970). The multivariate analysis of qualitative data: interactions among multiple classifications. *Journal of the American Statistical Association*, **65**, 226-56.

Goodman, L.A. (1972). A general model for the analysis of surveys. *American Journal of Sociology*, **77**, 1035-86.

Goodman, L.A. (1979). A brief guide to the causal analysis of data from surveys. *American Journal of Sociology*, **84**, 1078-95.

Graubard, B.I., Korn, E. L. & Midthune, D. (1997). Testing goodness-of-fit for logistic regression with survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 170-4.

Grizzle, J.E., Starmer, C.F. & Koch, G.G. (1969). Analysis of categorical data by linear models. *Biometrics*, **25**, 489-504.

Grizzle, J.E. and Williams, O.D. (1972). Log linear models and tests of independence for contingency tables. *Biometrics*, **28**, 137-156.

Gross, W.F. (1984). A note on "chi-squared tests with survey data". *Journal of the Royal Statistical Society, Series B*, **46**, 270-2.

Hoem, J.M. (1989). The issue of weights in panel surveys of individual behavior. In *Panel Surveys*, D. Kasprzyk, G.Duncan, G.Kalton, M.P.Singh, eds, New York: Wiley, pp. 539-65.

Holt, D., Scott, A.J. & Ewings, P.O. (1980). Chi-squared tests with survey data. *Journal of the Royal Statistical Society, Series A*, **143**, 302-20.

Imrey,P.B., Koch, G.G. & Stokes, M.E. (1982). Categorical data analysis: some reflections on the log linear model and logistic regression. Part II: data analysis. *International Statistical Review*, **50**, 35-63 (in collaboration with J.N. Darroch, D.H. Freeman, Jr. and H.D. Tolley).

Koch, G.G., Freeman, D.J. Jr., & Freeman, J.L. (1975). Strategies in the multivariate analysis of data from complex surveys. *International Statistical Review*, **43**, 59-78.

Korn, E.L. & Graubard, B.I. (1998). Variance estimation for superpopulation parameters. *Statistica Sinica*, **8**, 1131-51.

Lazarsfeld, P.F. (1968). The analysis of attribute data. In D.L. Sills ed. *International Encyclopedia of Social Sciences*, Vol. 15, New York: Macmillan and the Free Press., pp. 419-429.

Lehnen, R.G. & Koch, G.G. (1974). A general linear approach to the analysis of nonmetric data: applications for political science. *American Journal of Political Science*, **18**, 283-313.

Li, Y., Li, Z. & Graubard, B.I. (2011a). Testing for Hardy Weinberg equilibrium in national household surveys that collect family-based genetic data. *Annals of Human Genetics*, **75**, 732-41.

Li, Y., Graubard, B.I. & DiGaetano, R. (2011b). Weighting methods for population-based case-control studies with complex sampling. *Journal of the Royal Statistical Society, Series C*, **60**, 165-85.

Liang, K.-Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22

Lumley, T. and Scott, A. (2014). Tests for regression models fitted to survey data. *Australian and New Zealand Journal of Statistics*, **56**, 1-14.

Molina, E.A. & Skinner, C.J. (1992). Pseudo-likelihood and quasi-likelihood estimation for complex sampling schemes. *Computational Statistics and Data Analysis*, **13**, 395-405.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators, *Psychometrika*, **49**, 115-132.

Muthén, B. & Satorra, A.(1995). Complex sample data in structural equation modeling. *Sociological Methodology*, **25**, 267-316.

Natarajan, S., Lipsitz, S.R., Fitzmaurice, G.M., Sinha, D., Ibrahim, J.G., Haas, J. & Gellad, W. (2012). An extension of the Wilcoxon rank sum test for complex sample survey data. *Journal of the Royal Statistical Society, Series C*, **61**, 653-64.

Nathan, G. (1972). On the asymptotic power of tests for independence in contingency tables from stratified samples. *Journal of the American Statistical Association*, **67**, 917-920.

Nathan, G. (1975). Tests of independence in contingency tables from stratified samples. *Sankhya C*, **37**, 77-87.

Nathan, G. (1981). Notes on inference based on data from complex sample designs. *Survey Methodology*, **7**, 77-87.

Nelder, J. & Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, **135**, 370-84.

Oberski, D. (2014). lavaan.survey: an R package for complex survey analysis of structural equation models. *Journal of Statistical Software*, **57**, 1-27.

Patterson, B.H., Drayton, C.M. & Graubard, B.I. (2002). Latent class analysis of complex sample survey data: application to dietary data. *Journal of the American Statistical Association*, **97**, 721-8..

Pfeffermann, D. & Sverchkov, M.Y. (2003). Fitting generalized linear models under informative sampling. In R.L. Chambers and C.J. Skinner eds., *Analysis of Survey Data*, Chichester: Wiley, pp. 175-95.

Rao, J.N.K. (1998). Marginal models for repeated observations: inference with survey data. *Proceedings of the American Statistical Association*, Section on Survey Research Methods, 76-82.

Rao, J.N.K., Kumar, S. & Roberts, G. (1989). Analysis of sample survey data involving categorical response variables: methods and software. *Survey Methodology*, **15**, 161-86.

Rao, J.N.K. & Scott, A.J. (1979). Chi-squared tests for analysis of categorical data from complex surveys. *Proceedings of the American Statistical Association*, Section on Survey Research Methods, 58-66.

Rao, J.N.K. & Scott, A.J. (1981). The analysis of categorical data from complex surveys: chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, **76**, 221-30.

Rao, J.N.K. & Scott, A.J. (1984). On chi-squared tests for multi-way tables with cell proportions estimated from survey data. *Annals of Statistics*, **12**, 46-60.

Rao, J.N.K. & Scott, A.J. (1987). On simple adjustments to chi-squared tests with survey data. *Annals of Statistics*, **15**, 385-97.

Rao, J.N.K., Scott, A.J. & Skinner, C.J. (1998). Quasi-score tests with survey data. *Statistica Sinica*, **8**, 1059-70.

Rao, J.N.K. & Thomas, D.R. (1988). The analysis of cross-classified categorical data from complex sample surveys. *Sociological Methodology*, **18**, 213-69.

Rao, J.N.K., Yung, W. & Hidiroglou, M.A. (2002). Estimating equations for the analysis of survey data using poststratification information. *Sankya, Series A*, **64**, 364-78.

Roberts, G., Rao, J.N.K & Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, **74**, 1-12.

Roberts, G., Ren, Q. & Rao, J.N.K. (2009). Using marginal mean models for data from longitudinal surveys with a complex design: some advances in methods. In P.Lynn ed. *Methodology of Longitudinal Surveys*, Chichester: Wiley.

Rotnitzky, A. & Jewell, N. P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, **77**, 485-97.

Scott, A.J. (2007). Rao-Scott corrections and their impact. *Proceedings of the American Statistical Association*, Section on Survey Research Methods, 3514-8.

Scott, A.J., Rao, J.N.K. & Thomas, D.R. (1989). Weighted least-squares and quasilikelihood estimation for categorical data under singular models. *Linear Algebra and its Applications*, **127**, 427-47.

Scott, A.J. & Wild, C. J. (2002). On the robustness of weighted methods for fitting models to case-control data. *Journal of the Royal Statistical Society, Series B*, **64**, 207-19.

Scott, A.J. & Wild, C.J. (2003). Fitting logistic regression models in case-control studies with complex sampling. In R.L. Chambers and C.J. Skinner eds., *Analysis of Survey Data*, Chichester: Wiley, pp. 109-21.

Shuster, J.J. & Downing, D.J. (1976). Two-way contingency tables for complex sampling schemes. *Biometrika*, **63**, 271-6.

Skinner, C.J. (1989). Domain means, regression and multivariate analysis. In Skinner, C.J., Holt, D. & Smith, T.M.F. eds. *Analysis of Complex Surveys*. Chichester: Wiley, pp. 59-87.

Skinner, C.J., Holt, D. & Smith, T.M.F. eds. (1989). *Analysis of Complex Surveys*, Chichester: Wiley.

Skinner, C.J. & Vallet, L.-A. (2010). Fitting log-linear models to contingency tables from surveys with complex designs: an investigation of the Clogg-Eliason approach. *Sociological Methods and Research*, **39**, 83-108.

Thomas, D.R. & Rao, J.N.K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, **82**, 630-6.

Vermunt, J.K. & Magidson, J. (2007). Latent class analysis with sampling weights, a maximum likelihood approach *Sociological Methods and Research*, **36**, 87-111.