Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2007, Article ID 56372, 12 pages
doi:10.1155/2007/56372

Research Article

Methods for Stratified Cluster Sampling with Informative Stratification

Alastair Scott and Chris Wild

Received 24 April 2007; Accepted 8 August 2007

Recommended by Paul Cowpertwait

We look at fitting regression models using data from stratified cluster samples when the strata may depend in some way on the observed responses within clusters. One important subclass of examples is that of family studies in genetic epidemiology, where the probability of selecting a family into the study depends on the incidence of disease within the family. We develop the survey-weighted estimating equation approach for this problem, with particular emphasis on the estimation of superpopulation parameters. Full maximum likelihood for this class of problems involves modelling the population distribution of the covariates which is simply not feasible when there are a large number of potential covariates. We discuss efficient semiparametric maximum likelihood methods in which the covariate distribution is left completely unspecified. We further discuss the relative efficiencies of these two approaches.

Copyright © 2007 A. Scott and C. Wild. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. Introduction

In this paper, we look at the problem of fitting models to data from stratified cluster samples. We are particularly interested in situations where the probability that a cluster is in a particular stratum depends on the value of its response. Sometimes this dependence is explicit and obvious. An important special case of this situation is the case-control family study, which is widely used in genetic epidemiology (see Neuhaus et al. [1, 2]). In a simple case-control study, we stratify the population into cases (people with a disease under study, say) and controls (people without the disease), choose independent random samples from each stratum, and record the values of potential covariates for each person selected in the study. In a case-control family study, we record the same information

and, in addition, we identify a set of family members for each person selected in the case-control study and record the disease status and the values of the covariates of each of these family members. For example, Whittemore [3] considers a case-control family study of ovarian cancer. Here, the clusters consist of mother-daughter pairs. The case stratum contains all pairs in which the daughter has been diagnosed with ovarian cancer, and the control stratum contains all the other pairs. Other examples of similar retrospective family studies are given in Zhao et al. [4].

Another example where the strata are determined explicitly by the response is given by Neuhaus and Jewell [5]. They consider data from a stratified cluster sample of individuals enrolled in the Federal Employees Health Benefit Plan in which the response variable indicates whether or not someone used outpatient mental health services during the previous year for each of the years 1979–1981. Here, the clusters consist of the three observations for a single person, and four strata were defined by the total number of times (0, 1, 2, or 3) that the person used the service in the three years of the study.

In all these examples, stratum membership is determined exactly by the value of the (multivariate) response. In most surveys, however, the relationship between the response and stratum membership is less clearcut, with the strata determined by such things as administrative convenience or the availability of a suitable list. This is true even for some case-control family studies. For example, in the study that motivated this work, Wrensch et al. [6] conducted a population-based case-control study of glioma, the most common type of malignant brain tumour, in the San Francisco Bay Area. The investigators gathered all cases of glioma diagnosed in a specified time interval and a population-based sample of comparable controls through random digit dialling. They also gathered the brain tumour status and covariate information from family members of the original case-control sample participants. In this case, the case stratum contains all families with at least one member diagnosed with glioma in the specified time interval. The chance of a family being included in this stratum depends on the number of family members with glioma, but is not completely determined by this.

To cover these more general cases, we consider situations in which we may have to fit a parametric model, $P_h(\mathbf{y}, \mathbf{X}; \boldsymbol{\gamma})$, for the conditional probability of a cluster being included in the hth stratum given values of the response vector, $\mathbf{y}$, and the matrix of covariates, $\mathbf{X}$. Note that there are no problems if this stratum inclusion model only involves $\mathbf{X}$. However, if the model depends on the response, $\mathbf{y}$, as well or, more generally depends on a design variable that is associated with $\mathbf{y}$ but is not included in the model, then the sampling is not ignorable (cf. Rubin [7]) and will affect the likelihood.

A possible strategy that is sometimes suggested for coping with informative stratification is to include the stratum indicator as a covariate in the model. This strategy avoids the technical problems but clearly makes no sense when the stratification depends solely on the response as in many of the examples above. Even in situations where the stratification does not depend directly on the response, it may distort the relationship between $\mathbf{y}$ and $\mathbf{X}$, which is the quantity of interest. For example, Lee et al. [8] consider a secondary analysis of data from a case-control study of Sudden Infant Death Syndrome (SIDS). The response in this analysis was an indicator of immunization, and clearly there would be

little sense in including the stratifying variable (SIDS) as a predictor in the model. Similarly, in our motivating brain cancer example, the researchers did not want to include the date of diagnosis in their predictive model.

The standard survey approach is through weighted estimating equations, with weights inversely proportional to the selection probabilities, as in Binder [9] or Rao et al. [10]. This works well when the weights are reasonably homogeneous but can be inefficient when the weights vary widely as they tend to do in retrospective studies. De Mets and Halperin [11] and Smith [10] looked at a more efficient approach which involved modelling the joint distribution of response, covariates, and design variables used for the stratification. This is efficient but becomes very difficult to implement when there are a large number of potential explanatory variables. In this paper, we look at an efficient intermediate approach based on semiparametric maximum likelihood in which the marginal distribution of the covariates is left unspecified. The general setup is described in Section 2. In Section 3, we examine the survey-weighted approach in some detail and the semi-parametric theory is developed in Section 4. We conclude with a brief discussion.

2. Basic setup

As in the introduction, we let $\mathbf{y}$ denote the vector of responses for the units in a cluster and we let $\mathbf{X}$ be the corresponding matrix of covariate values. In addition, we define a stratum indicator variable Z which takes the value $Z = h$ if the cluster is assigned to the hth stratum ($h = 1,\ldots,L$). We assume that the values in our finite population of N clusters are generated by random sampling from the joint distribution of $(\mathbf{y},\mathbf{X},Z)$. The clusters are then sorted into L strata, $\mathcal{S}_1,\ldots,\mathcal{S}_L$, according to the values of Z, resulting in N_h clusters in $\mathcal{S}_h$ ($\sum_1^L N_h = N$). Finally, we draw independent simple random samples, D_h, of n_h clusters from the N_h clusters in $\mathcal{S}_h$ ($h = 1,\ldots,L$) and observe the corresponding $(\mathbf{y},\mathbf{X})$ values. Let $(\mathbf{y}_{hj},\mathbf{X}_{hj})$ represent observed values for the jth cluster in the hth stratum ($h = 1,\ldots,L$; $j = 1,\ldots,n_h$). Our data are thus of the form $\{(\mathbf{y}_{hj},\mathbf{X}_{hj}, j \in D_h),\ N_h;\ h = 1,\ldots,L\}$. Note that the observed stratum sizes, $N_1,\ldots,N_L$, are random variables in this scenario and contain valuable information.

We are interested in modelling $f(\mathbf{y} \mid \mathbf{X};\boldsymbol{\theta})$, the conditional distribution of the vector of cluster responses $\mathbf{y}$ given $\mathbf{X}$, the matrix of cluster covariates, and, in cases where it is needed, the conditional probabilities that the cluster falls into stratum $\mathcal{S}_h$, $h = 1,\ldots,L$ given $\mathbf{y}$ and $\mathbf{X}$:

$$\operatorname{pr}(\text{cluster} \in \mathcal{S}_h \mid \mathbf{y},\mathbf{X}) = \operatorname{pr}(Z = h \mid \mathbf{y},\mathbf{X}) = P_h(\mathbf{y},\mathbf{X};\boldsymbol{\gamma}). \tag{2.1}$$

Using an argument similar to that given in Scott and Wild [12, Appendix B], the likelihood function can be shown to be given by

$$\begin{aligned} L(\boldsymbol{\theta},\boldsymbol{\gamma},g) &= \prod_{h=1}^{L}\left\{\prod_{j\in D_h}\operatorname{pr}(\mathbf{y}_{hj},\mathbf{X}_{hj} \mid \text{cluster} \in \mathcal{S}_h)\right\}Q_h^{N_h} \\ &= \prod_h\left(\prod_{D_h}\{f(\mathbf{y}_{hj} \mid \mathbf{X}_{hj};\boldsymbol{\theta})g(\mathbf{X}_{hj})\}Q_h^{N_h-n_h}\right), \end{aligned} \tag{2.2}$$

where $g(\mathbf{X})$ denotes the marginal density of X in the population and

$$Q_h = Q_h(\boldsymbol{\theta}, \boldsymbol{\gamma}, g) = \mathrm{pr}(Z = h) = \iint P_h(\mathbf{y}, \mathbf{X}; \boldsymbol{\gamma}) f(\mathbf{y} \mid \mathbf{X}; \boldsymbol{\theta}) g(\mathbf{X}) d\mathbf{y}\, d\mathbf{X} \tag{2.3}$$

denotes the marginal probability that a cluster is in stratum $\mathcal{S}_h$. Strictly speaking, we need to describe how we choose the sample sizes, $n_1, \ldots, n_L$. However, the kernel of the likelihood is as above for any scheme satisfying the condition that $\{n_1, \ldots, n_L\}$ depends only on $\{N_1, \ldots, N_L\}$ and not on the realized values of $(\mathbf{y}, \mathbf{X})$ (see Wild [13] for details).

If we had drawn a simple random sample from the whole population, or if the stratification depended only on $\mathbf{X}$, then the likelihood would factor into two terms, one involving $\boldsymbol{\theta}$ alone and the other involving $\boldsymbol{\gamma}$ and $g(\mathbf{X})$. This means that we could make inferences about $\boldsymbol{\theta}$ conditional on the observed values of $\mathbf{X}$ and not have to bother about terms involving $g(\mathbf{X})$. Unfortunately, we cannot ignore $g(\mathbf{X})$ when $P_h(\mathbf{y}, \mathbf{X}; \boldsymbol{\gamma})$ involves $\mathbf{y}$; just as in a case-control study, we cannot separate $\boldsymbol{\theta}$ from $g(\mathbf{x})$ because both are involved in Q_h. The most common way of coping with this is through weighted estimating equations with weights inversely proportional to the selection probabilities as in Binder [9]. We examine this approach in more detail in Section 3. It is relatively simple to implement and works well in many situations. However, it tends to be very inefficient if the selection probabilities vary widely as they often do in retrospective studies such as the case-control family studies described in the introduction. A more efficient alternative is to build a full parametric model for $g(\mathbf{x})$ and use ordinary maximum likelihood. A good description of this approach is given in Smith and Nathan [14]. It does indeed produce very efficient estimators but, unfortunately, it rapidly becomes impractical when the number of potential covariates increases. This limits its application when we have a large number of potential covariates with a mixture of continuous, categorical, and count variables, as is the case in many surveys.

Ideally, we would like a method that combines the simplicity of the weighted approach with the efficiency of maximum likelihood. In Section 4, we look at a semiparametric approach in which the marginal distribution of $\mathbf{X}$ is treated nonparametrically. In this approach, $g(\mathbf{X})$ becomes a (potentially infinite dimensional) nuisance parameter in the likelihood. The resulting estimators turn out to be very efficient while, perhaps more surprisingly, still reasonably simple to obtain.

3. Weighted estimators

If we had observed the values of $\{\mathbf{y}, \mathbf{X}, Z\}$ for every cluster in the finite population, we would estimate $\boldsymbol{\theta}$ by solving the "census" likelihood equation

$$\mathbf{S}(\boldsymbol{\theta}) = \sum_{h=1}^{L} \sum_{j=1}^{N_h} \mathbf{U}_{hj}(\boldsymbol{\theta}) = \mathbf{0}, \tag{3.1}$$

where $\mathbf{U}_{hj}(\boldsymbol{\theta}) = \partial \log f(\mathbf{y}_{hj} \mid \mathbf{X}_{hj}; \boldsymbol{\theta})/\partial\boldsymbol{\theta}$. We will assume that the standard regularity conditions for likelihood (see, e.g., Lehmann [15, Section 7.3]) are satisfied so that

$$E\{\mathbf{S}(\boldsymbol{\theta})\} = \underline{0}, \qquad \operatorname{Cov}\{\mathbf{S}(\boldsymbol{\theta})\} = -E\left\{\frac{\partial \mathbf{S}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right\} = N\mathcal{J}(\boldsymbol{\theta}), \tag{3.2}$$

at the true value, $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Of course, we do not observe the whole population. However, for any fixed value of $\boldsymbol{\theta}$, $\mathbf{S}(\boldsymbol{\theta})$ is simply a vector of population totals and thus can be estimated from the sample observations by the weighted sample score,

$$\mathbf{S}_W(\boldsymbol{\theta}) = \sum_h \sum_{D_h} \frac{N_h}{n_h} \mathbf{U}_{hj}(\boldsymbol{\theta}). \tag{3.3}$$

The weighted estimator, $\widehat{\boldsymbol{\theta}}_W$, is then defined as the solution to the weighted pseudo-likelihood equation, $\mathbf{S}_W(\boldsymbol{\theta}) = \mathbf{0}$.

Under suitable regularity conditions on $\{\mathbf{U}_{hj}\}$, $\widehat{\boldsymbol{\theta}}_W$ is a consistent estimator of the finite population (or census) regression parameter, $\boldsymbol{\theta}_C$, defined as the solution to (3.1) (see, e.g., Binder [9] or Rao et al. [10]). Our interest here, however, is not in descriptive inferences about a particular finite population, but rather about modelling the underlying processes that lead different units to have different responses y. Thus, in survey sampling terminology, we are interested in estimating the superpopulation parameters. We have to take some care in deriving the properties of $\widehat{\boldsymbol{\theta}}_W$ in this framework since $N_1, \ldots, N_L$ are now random variables rather than fixed constants as in the standard finite population setup.

In sampling terminology, we can think of our situation as being equivalent to two-phase sampling for stratification. In the first phase, the finite population is generated as a random sample of size N from an (infinite) super population and the stratum to which each cluster (i.e., the value of Z) belongs is recorded. At the second phase, we draw a simple random sample of size n_h from the N_h clusters in stratum $\mathcal{S}_h$, with the values of $\{n_1, \ldots, n_L\}$ depending only on $\{N_1, \ldots, N_L\}$, and observe $\{\mathbf{y}_{hj}, \mathbf{X}_{hj}, j \in D_h\}$ for $h = 1, \ldots, L$.

We establish the results by first conditioning on $\mathbf{Z}_\mathbf{N}$, the vector of stratum indicators for the realized finite population and then averaging over the distribution of $\mathbf{Z}_\mathbf{N}$. Given $\mathbf{Z}_\mathbf{N}$, $\{N_1, \ldots, N_L\}$, and hence $\{n_1, \ldots, n_L\}$, are fixed constants and $\mathbf{U}_{hj}(\boldsymbol{\theta})$, $j \in D_h$, are i.i.d. observations from the conditional distribution of $\mathbf{U}(\boldsymbol{\theta}) = (\partial \log f(\mathbf{y} \mid \mathbf{X}; \boldsymbol{\theta}))/\partial\boldsymbol{\theta}$ given $Z = h$. Let $\boldsymbol{\mu}_h(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}_h(\boldsymbol{\theta})$ denote the mean vector and covariance matrix of this conditional distribution, and let $\boldsymbol{\mu}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ denote the corresponding quantities for the unconditional distribution of $\mathbf{U}(\boldsymbol{\theta})$. Recall that $\boldsymbol{\mu}(\boldsymbol{\theta}_0) = \underline{0}$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0) = \mathcal{J}(\boldsymbol{\theta}_0)$ under our regularity conditions. The unconditional distribution of $\{N_1, \ldots, N_L\}$ is multinomial $(N; Q_1, \ldots, Q_L)$ where, as before, Q_h is the marginal probability that $Z = h$ for $h = 1, \ldots, L$.

Note that

$$E\left\{\sum_h N_h \boldsymbol{\mu}_h(\boldsymbol{\theta})\right\} = N\left(\sum_h \boldsymbol{\mu}_h(\boldsymbol{\theta}) Q_h\right) = N\boldsymbol{\mu}(\boldsymbol{\theta}). \tag{3.4}$$

It follows that

$$E\{\mathbf{S}_W(\boldsymbol{\theta})\} = E\left\{\sum_h N_h \boldsymbol{\mu}_h(\boldsymbol{\theta})\right\} = N\boldsymbol{\mu}(\boldsymbol{\theta}). \tag{3.5}$$

Thus $E\{\sum_h N_h \boldsymbol{\mu}_h(\boldsymbol{\theta})\} = \underline{0}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. In addition, using the standard results for conditional variances,

$$\operatorname{Cov}\{\mathbf{S}_W(\boldsymbol{\theta})\} = \operatorname{Cov}\left\{\sum_h N_h \boldsymbol{\mu}_h(\boldsymbol{\theta})\right\} + E\left\{\sum_h N_h^2 \frac{\boldsymbol{\Sigma}_h(\boldsymbol{\theta})}{n_h}\right\}. \tag{3.6}$$

To proceed further, we need to specify how the n_hs are chosen. We will assume that the sampling fractions n_h/N_h are fixed constants f_h with $0 < f_h \leq 1$ for $h = 1, \ldots, L$. (Of course we cannot always achieve this exactly in small samples but if $n_h = [f_h N_h]$ the difference is negligible asymptotically.) Then

$$\begin{aligned}
\operatorname{Cov}\{\mathbf{S}_W(\boldsymbol{\theta})\} &\simeq N\left\{\sum \boldsymbol{\mu}_h \boldsymbol{\mu}_h^T Q_h(1 - Q_h) - 2\sum\sum_{\ell<h} \boldsymbol{\mu}_h \boldsymbol{\mu}_\ell^T Q_h Q_\ell + \sum Q_h \frac{\boldsymbol{\Sigma}_h}{f_h}\right\} \\
&= N\left\{\sum_h Q_h\left(\boldsymbol{\mu}_h \boldsymbol{\mu}_h^T + \frac{\boldsymbol{\Sigma}_h}{f_h}\right) - \left(\sum_h Q_h \boldsymbol{\mu}_h\right)\left(\sum_h Q_h \boldsymbol{\mu}_h\right)^T\right\} \\
&= N\left\{\sum_h Q_h\left(\boldsymbol{\mu}_h \boldsymbol{\mu}_h^T + \frac{\boldsymbol{\Sigma}_h}{f_h}\right) - \boldsymbol{\mu}\boldsymbol{\mu}^T\right\} \\
&= N\sum_h Q_h\left\{\frac{\boldsymbol{\Sigma}_h}{f_h} + (\boldsymbol{\mu}_h - \boldsymbol{\mu})(\boldsymbol{\mu}_h - \boldsymbol{\mu})^T\right\}.
\end{aligned} \tag{3.7}$$

The first term is the covariance matrix that we would get if the weights were known and the second term represents the penalty we pay for incomplete knowledge about the weights. Using the relation

$$\operatorname{Cov}\{\mathbf{U}(\boldsymbol{\theta})\} = E\{\operatorname{Cov}\{\mathbf{U} \mid Z = h\}\} + \operatorname{Cov}\{E\{\mathbf{U} \mid Z = h\}\}, \tag{3.8}$$

we can also rewrite this variance in the form

$$\operatorname{Cov}\{\mathbf{S}_W(\boldsymbol{\theta})\} = N\left\{\Sigma(\boldsymbol{\theta}) + \sum_h Q_h\left(\frac{1}{f_h} - 1\right)\boldsymbol{\Sigma}_h\right\}. \tag{3.9}$$

Now the first term is the covariance matrix that we would have obtained by sampling all clusters in the finite population so that, in this representation, the second term represents the penalty that we pay for incomplete enumeration at the second phase.

Finally, it follows from the results of Chen and Rao [16] that $\mathbf{S}_W(\boldsymbol{\theta})$ is asymptotically multivariate normal as $N \to \infty$ with $n_h/N_h \to f_h$ for $h = 1,\ldots,L$ fixed. Having established the properties of $S_W(\boldsymbol{\theta})$, we use standard results for unbiased estimating equations (see, e.g., Amari and Kawanabe [17]) to invert the equation $\mathbf{S}_W(\hat{\boldsymbol{\theta}}_W) = \mathbf{0}$ and infer results for $\hat{\boldsymbol{\theta}}_W$. In particular, it follows that $\sqrt{N}(\hat{\boldsymbol{\theta}}_W - \boldsymbol{\theta})$ converges in distribution to a multivariate normal random variable with mean vector $\mathbf{0}$ and covariance matrix $N\mathbf{V}(\hat{\boldsymbol{\theta}})$, where

$$N\mathbf{V}(\hat{\boldsymbol{\theta}}) = \mathcal{J}^{-1}(\boldsymbol{\theta}_0)\left(\sum_h Q_h\left[\frac{\boldsymbol{\Sigma}_h}{f_h} + (\boldsymbol{\mu}_h - \boldsymbol{\mu})(\boldsymbol{\mu}_h - \boldsymbol{\mu})^T\right]\right)\mathcal{J}^{-1}(\boldsymbol{\theta}_0), \tag{3.10}$$

with

$$\mathcal{J}(\boldsymbol{\theta}) = E\left\{-\frac{\partial^2 \log\left(f(\mathbf{y} \mid \mathbf{X};\boldsymbol{\theta})\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right\}, \tag{3.11}$$

as $N \to \infty$. Recall that $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathcal{J}(\boldsymbol{\theta})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. This means that we can rewrite $\mathbf{V}(\hat{\boldsymbol{\theta}})$ as

$$\mathbf{V}(\hat{\boldsymbol{\theta}}) = \frac{1}{N}\left[\mathcal{J}^{-1}(\boldsymbol{\theta}_0) + \mathcal{J}^{-1}(\boldsymbol{\theta}_0)\left\{\sum_h Q_h\left(\frac{1}{f_h} - 1\right)\boldsymbol{\Sigma}_h\right\}\mathcal{J}^{-1}(\boldsymbol{\theta}_0)\right]. \tag{3.12}$$

(Here we have used (3.9) to represent $\mathrm{Cov}(\mathbf{S}_W)$). The first term is what we would get if we sampled the whole population and the second term again represents the cost of incomplete enumeration.

We can estimate $\mathbf{V}(\hat{\boldsymbol{\theta}})$ by substituting $\hat{\mathbf{J}} = -(1/N)\partial\mathbf{S}_W(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}^T$ for $\mathcal{J}(\boldsymbol{\theta}_0)$, $\hat{\boldsymbol{\mu}}_h = \sum_j \mathbf{U}_{hj}/n_h$ for $\boldsymbol{\mu}_h$, $W_h = N_h/N$ for Q_h, and the ordinary within-stratum sample variance for $\boldsymbol{\Sigma}_h$. This, in conjunction with (3.10), leads to the estimator

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}}) = \hat{\mathbf{J}}^{-1}\left(\sum_h W_h^2\frac{\hat{\boldsymbol{\Sigma}}_h}{n_h}\right)\hat{\mathbf{J}}^{-1} + \frac{1}{N}\hat{\mathbf{J}}^{-1}\left\{\sum_h W_h\hat{\boldsymbol{\mu}}_h\hat{\boldsymbol{\mu}}_h^T\right\}\hat{\mathbf{J}}^{-1}. \tag{3.13}$$

The first term of (3.13), which is $O(1/n)$, is the variance estimate we would use if we assumed that the N_hs were fixed and the second term, which is $O(1/N)$, measures the effect of not knowing the N_hs in advance. This second term will be negligible in many applications.

The weighted method is relatively straightforward and most large statistical packages now include procedures for implementing it for linear and logistic regression models, although all will assume that the $\{N_h\}$s are fixed constants and thus will produce a slight underestimate of the standard errors. A big advantage over more efficient procedures is that it does not require any modeling of stratum inclusion probabilities. One important consequence of this is that the same procedure can be used for stratified two-stage sampling, where simple random subsamples are chosen from each selected cluster. More complex subsampling schemes can be handled simply by adjusting the weights in the pseudo likelihood (3.1).

In general, weighting works pretty well for standard sampling situations where the sampling fractions do not vary too much among strata. It does not work so well in situations where the sampling fractions vary widely, as they tend to do in retrospective designs like the case-control family studies discussed in the introduction. For example, Lawless et al. [18] report efficiencies of less than 15% (compared to the semiparametric estimators discussed in Section 4) for some unclustered case-control designs and Scott and Wild [19] report similar values for some special clustered designs. An appealing feature of the weighted method is its robustness to departures from the model. When the assumed regression model $f(\mathbf{y} \mid \mathbf{X};\boldsymbol{\theta})$ is not valid, the fitted model produced by the weighted method can still be interpreted as estimating the best fitting model for the whole population; see Scott and Wild [20] for further discussion of this point.

4. Semiparametric estimators

We now return to the likelihood function $L(\boldsymbol{\theta},\boldsymbol{\gamma},g)$ given in (2.2). The semi-parametric maximum likelihood estimators of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are obtained by maximizing $\ell(\boldsymbol{\theta},\boldsymbol{\gamma},g) = \log L(\boldsymbol{\theta},\boldsymbol{\gamma},g)$ over $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$, and g leaving the density function $g(\cdot)$ completely undefined. Essentially, we treat $g(\cdot)$ as a (potentially infinite-dimensional) nuisance parameter. Although it might seem at first glance that this would be formidable task, it turns out that the semi-parametric MLE of $\boldsymbol{\theta}$ (and $\boldsymbol{\gamma}$) can be calculated relatively easily.

We start by reducing the problem to the simpler case in which stratum membership is determined directly by the cluster response. First, we augment the response vector $\mathbf{y}$ with the stratum indicator Z to give modified response variable $\tilde{\mathbf{y}} = \binom{\mathbf{y}}{Z}$. Next, we set $\tilde{\boldsymbol{\theta}} = \binom{\boldsymbol{\theta}}{\boldsymbol{\gamma}}$. Our problem is then reduced to that of fitting the model $f(\tilde{\mathbf{y}} \mid \mathbf{X};\tilde{\boldsymbol{\theta}})$, where

$$f(\tilde{\mathbf{y}} \mid \mathbf{X};\tilde{\boldsymbol{\theta}}) = f(z \mid \mathbf{y},\mathbf{X};\boldsymbol{\gamma})\, f(\mathbf{y} \mid \mathbf{X};\boldsymbol{\theta}) = P_z(\mathbf{y},\mathbf{X};\boldsymbol{\gamma})\, f(\mathbf{y} \mid \mathbf{X};\boldsymbol{\theta}), \tag{4.1}$$

to data from a stratified sample where the strata, $\mathcal{S}_h$ $(h = 1,\ldots,L)$, are determined completely by the response, $\tilde{\mathbf{y}}$. The estimating equations for the semi-parametric maximum likelihood equations in this reduced case are derived in Scott and Wild [12, 19], following earlier work by Cosslett [21]. In a companion paper in this issue, Lee [22] establishes the asymptotic efficiency of this estimator and shows that $\mathcal{J}^*(\hat{\boldsymbol{\phi}})^{-1}$ provides a consistent estimator of the variance. Similar results are obtained in Lee and Hirose [23] using a different approach based on the profile likelihood methods of Newey [24]. In the remainder of this section, we summarize the results of translating these results for the reduced case back into our original notation.

First, we define a pseudo-log-likelihood function

$$\begin{aligned}\ell^*(\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\pi}) = &\sum_h \sum_{D_h} \log f_h^*(\mathbf{y}_{hj} \mid \mathbf{X}_{hj};\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\pi}) \\ &- \sum_h \{(N_h - n_h)\log(1-\pi_h) + n_h \log \pi_h\},\end{aligned} \tag{4.2}$$

where

$$f_h^*(\mathbf{y} \mid \mathbf{X};\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\pi}) \propto \pi_h P_h(\mathbf{y},\mathbf{X};\boldsymbol{\gamma})\, f(\mathbf{y} \mid \mathbf{X};\boldsymbol{\theta}) \tag{4.3}$$

and $\boldsymbol{\pi}$ is an L-dimensional vector of nuisance parameters. Then the semi-parametric maximum likelihood estimators, $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\gamma}}$, of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are the appropriate components of $\hat{\boldsymbol{\phi}}$, the solution of the pseudo score equation

$$\mathbf{S}^*(\boldsymbol{\phi}) = \frac{\partial \ell^*(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = \mathbf{0}, \tag{4.4}$$

where $\boldsymbol{\phi} = (\boldsymbol{\theta}^T, \boldsymbol{\gamma}^T, \boldsymbol{\pi}^T)^T$. This means that, for the purposes of calculating the MLE of $\binom{\boldsymbol{\theta}}{\boldsymbol{\gamma}}$, we can act as if $\ell^*(\boldsymbol{\phi})$ is the log-likelihood; in essence, we have replaced a problem involving an infinite dimensional nuisance parameter, $g(\cdot)$, with one involving an L-dimensional nuisance parameter, $\boldsymbol{\pi}$.

The pseudoscore, $\mathbf{S}^*(\boldsymbol{\phi})$, has many of the properties of a standard score function. In the first place, with appropriate standardization, $\mathbf{S}^*$ is asymptotically normal as $N \to \infty$ provided $n_h/N_h \to f_h$ with $0 < f_h \leq 1$ for $h = 1, \ldots, L$. Secondly, $E\{\mathbf{S}^*(\boldsymbol{\phi})\} = \mathbf{0}$ at the true value, even though the individual terms in $\mathbf{S}^*(\boldsymbol{\phi})$ are neither identically distributed nor have expected value zero under the stratified sampling design. Finally, if we let $\mathcal{J}^*$ denote the observed (pseudo-) information matrix,

$$\mathcal{J}^*(\boldsymbol{\phi}) = -\frac{\partial \mathbf{S}^*(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}^T} = -\frac{\partial^2 \ell^*}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T}, \tag{4.5}$$

and let $\mathcal{J}^*$ denote its expected value, then $\mathbf{S}^*(\boldsymbol{\phi})$ is asymptotically normal with asymptotic covariance matrix equal to

$$\operatorname{Cov}\{\mathbf{S}^*(\boldsymbol{\phi})\} = \mathcal{J}^*(\boldsymbol{\phi}_0) - \mathcal{J}^*(\boldsymbol{\phi}_0) \begin{pmatrix} \mathbf{0} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{K} \end{pmatrix} \mathcal{J}^*(\boldsymbol{\phi}_0), \tag{4.6}$$

where $\mathbf{K}$ is some $L \times L$ symmetric matrix. Properties of $\hat{\boldsymbol{\phi}}$ then follow from standard results for estimating equations (see, e.g., Amari and Kawanabe [17]). In particular, $\hat{\boldsymbol{\phi}}$ is asymptotically normal with mean $\boldsymbol{\phi}_0$ and covariance matrix

$$\mathcal{J}^*(\boldsymbol{\phi}_0)^{-1} - \begin{pmatrix} \mathbf{0} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{K} \end{pmatrix}. \tag{4.7}$$

We are only interested in the block corresponding to the components of $\binom{\boldsymbol{\theta}}{\boldsymbol{\gamma}}$ and this does not involve $\mathbf{K}$. All this means that, for the purpose of estimating $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, we can operate as if $\mathbf{S}^*(\boldsymbol{\phi})$ is a genuine score function. The semiparametric MLE, $\binom{\hat{\boldsymbol{\theta}}}{\hat{\boldsymbol{\gamma}}}$, is obtained by setting $\mathbf{S}^*(\boldsymbol{\phi}) = \mathbf{0}$ and its covariance matrix can be estimated using the appropriate components of the inverse observed information matrix, $\mathcal{J}^*(\hat{\boldsymbol{\phi}})^{-1}$. (Note that some care has to be taken with solving the pseudoscore equations numerically as $\hat{\boldsymbol{\phi}}$ often corresponds to a saddle point of ℓ^* rather than a maximum.)

In principle, we can extend the results to stratified two-stage sampling where subsamples are drawn from the chosen clusters (or primary sampling units). However, to apply (4.3), we need the conditional probability of stratum membership given the observed $(\mathbf{y}, \mathbf{X})$, which requires integration over the values for the unsampled units in the cluster.

This is a nontrivial task in general so that the extension of the semiparametric approach to two-stage sampling is much less straightforward in practice than the weighted approach.

5. Discussion

We have sketched in Section 4 the development of semi-parametric methods for fitting regression models to data from stratified samples as an alternative to the weighted methods of Section 3, which are simple to implement but can be inefficient in retrospective studies, or full maximum likelihood, which is efficient but difficult to implement because it requires fitting a model for the joint distribution of all covariates. The methods are relatively simple to implement and simulations so far (see, e.g., Lawless et al. [18]) suggest that they are much more efficient than weighted methods in situations where the latter perform badly. In the very limited examples that we have looked at so far, they seem to be almost as efficient as full maximum likelihood but much more work needs to be done here.

A number of alternative general approaches to inference from complex surveys have been suggested in the literature and all of these can be specialized to informative stratified sampling. Nathan and Holt [25] and Smith and Nathan [14] suggest alternatives to full maximum likelihood that do not require the fitting of a complete model for the joint distribution of the covariates and design variables. Krieger and Pfeffermann [26] and Pfeffermann and Sverchkov [27] explore methods based on the induced distribution of $\mathbf{y}$ given $\mathbf{X}$ in the sample after the (possible informative) selection mechanism has been taken into account. The general semi-parametric methods for missing data problems that have been developed by Robins and his collaborators (see, e.g., Robins et al.[28, 29]) may also be applicable to our setup here. All of these methods seem to have connections to the methods that we have developed here and we are in the process of exploring these connections.

Acknowledgment

Jeff Hunter has been a colleague and friend for more than 35 years. We would like to take this opportunity to thank him for his advice and moral support throughout that time and to wish him a happy and productive retirement.

References

[1] J. M. Neuhaus, A. Scott, and C. Wild, "The analysis of retrospective family studies," *Biometrika*, vol. 89, no. 1, pp. 23–37, 2002.

[2] J. M. Neuhaus, A. Scott, and C. Wild, "Family-specific approaches to the analysis of case-control family data," *Biometrics*, vol. 62, no. 2, pp. 488–494, 2006.

[3] A. S. Whittemore, "Logistic regression of family data from case-control studies," *Biometrika*, vol. 82, no. 1, pp. 57–67, 1995.

[4] L. P. Zhao, L. Hsu, S. Holte, Y. Chen, F. Quiaoit, and R. L. Prentice, "Combined association and aggregation analysis of data from case-control family studies," *Biometrika*, vol. 85, no. 2, pp. 299–315, 1998.

[5] J. M. Neuhaus and N. P. Jewell, "The effect of retrospective sampling on binary regression models for clustered data," *Biometrics*, vol. 46, no. 4, pp. 977–990, 1990.

[6] M. Wrensch, M. Lee, R. Miike, et al., "Familial and personal medical history of cancer and nervous system conditions among adults with glioma and controls," *American Journal of Epidemiology*, vol. 145, no. 7, pp. 581–593, 1997.

[7] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[8] A. J. Lee, L. McMurchy, and A. Scott, "Re-using data from case-control studies," *Statistics in Medicine*, vol. 16, no. 12, pp. 1377–1389, 1997.

[9] D. A. Binder, "On the variances of asymptotically normal estimators from complex surveys," *International Statistical Review*, vol. 51, no. 3, pp. 279–292, 1983.

[10] J. N. K. Rao, A. Scott, and C. J. Skinner, "Quasi-score tests with survey data," *Statistica Sinica*, vol. 8, no. 4, pp. 1059–1070, 1998.

[11] D. DeMets and M. Halperin, "Estimation of a simple regression coefficient in samples arising from a sub sampling procedure," *Biometrics*, vol. 33, no. 1, pp. 47–56, 1977.

[12] A. Scott and C. Wild, "Maximum likelihood for generalised case-control studies," *Journal of Statistical Planning and Inference*, vol. 96, no. 1, pp. 3–27, 2001.

[13] C. Wild, "Fitting prospective regression models to case-control data," *Biometrika*, vol. 78, no. 4, pp. 705–717, 1991.

[14] T. M. F. Smith and G. Nathan, "The effect of selection on regression analysis," in *Current Analysis of Complex Surveys*, C. J. Skinner, D. Holt, and T. M. F. Smith, Eds., pp. 149–163, Wiley, New York, NY, USA, 1989.

[15] E. L. Lehmann, *Asymptotic Theory*, John Wiley & Sons, New York, NY, USA, 1999.

[16] J. Chen and J. N. K. Rao, "Asymptotic normality under two-phase sampling designs," *Statistica Sinica*, vol. 17, no. 2, pp. 1047–1064, 2007.

[17] S. Amari and M. Kawanabe, "Estimating functions in semiparametric statistical models," in *Selected Proceedings of the Symposium on Estimating Functions (Athens, Ga, 1996)*, I. V. Basawa, V. P. Godambe, and R. L. Taylor, Eds., vol. 32 of *IMS Lecture Notes Monograph Series*, pp. 65–81, Institute of Mathematical Statistics, Hayward, Calif, USA, 1997.

[18] J. F. Lawless, J. D. Kalbfleisch, and C. Wild, "Semiparametric methods for response-selective and missing data problems in regression," *Journal of the Royal Statistical Society. Series B*, vol. 61, no. 2, pp. 413–438, 1999.

[19] A. Scott and C. Wild, "The analysis of clustered case-control studies," *Journal of the Royal Statistical Society C*, vol. 50, pp. 389–401, 2001.

[20] A. Scott and C. Wild, "On the robustness of weighted methods for fitting models to case-control data," *Journal of the Royal Statistical Society. Series B*, vol. 64, no. 2, pp. 207–219, 2002.

[21] S. Cosslett, "Efficient estimation of discrete-choice models," in *Structural Analysis of Discrete Data with Econometric Applications*, C. F. Manski and D. McFadden, Eds., pp. 51–111, Wiley, New York, NY, USA, 1981.

[22] A. J. Lee, "On the semi-parametric efficiency of the Scott-Wild estimator under choice-based and two-phase sampling," to appear in *Journal of Applied Mathematics and Decision Sciences*.

[23] A. J. Lee and Y. Hirose, "Semi-parametric efficiency bounds for regression models under case-control sampling: the profile likelihood approach," to appear in *Annals of the Institute of Statistical Mathematics*.

[24] W. K. Newey, "The asymptotic variance of semiparametric estimators," *Econometrica*, vol. 62, no. 6, pp. 1349–1382, 1994.

[25] G. Nathan and D. Holt, "The effect of survey design on regression analysis," *Journal of the Royal Statistical Society. Series B*, vol. 42, no. 3, pp. 377–386, 1980.

[26] A. M. Krieger and D. Pfeffermann, "Maximum likelihood estimation from complex sample surveys," *Survey Methodology*, vol. 18, pp. 225–239, 1992.

[27] D. Pfeffermann and M. Sverchkov, "Parametric and semi-parametric estimation of regression models fitted to survey data," *Sankhya B*, vol. 61, no. 1, pp. 166–186, 1999.

[28] J. M. Robins, A. Rotnitzky, and L. P. Zhao, "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American Statistical Association*, vol. 89, no. 427, pp. 846–866, 1994.

[29] J. M. Robins, F. S. Hsieh, and W. Newey, "Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates," *Journal of the Royal Statistical Society. Series B*, vol. 57, no. 2, pp. 409–424, 1995.

Alastair Scott: Department of Statistics, University of Auckland, Private Bag,
Auckland 92019, New Zealand
Email address: scott@stat.auckland.ac.nz

Chris Wild: Department of Statistics, University of Auckland, Private Bag,
Auckland 92019, New Zealand
Email address: wild@stat.auckland.ac.nz