Ann Inst Stat Math (2010) 62:1023–1052
DOI 10.1007/s10463-008-0205-1

Semi-parametric efficiency bounds for regression models under response-selective sampling: the profile likelihood approach

Alan Lee · Yuichi Hirose

Received: 5 April 2007 / Revised: 16 June 2008 / Published online: 18 September 2008

Abstract We obtain an information bound for estimates of parameters in general regression models where data are collected under a variety of response-selective sampling schemes, together with a simple formula for the asymptotic variance of the semi-parametric maximum likelihood estimate. This is compared to the bound and the estimate is found to be fully efficient in a variety of settings. A small simulation study is reported to illustrate the small-sample efficiency of the semi-parametric estimator.

Keywords Semi-parametric efficiency · Outcome-dependent sampling · Case-control study · Profile likelihood · Tangent space · Influence function · Efficient score · Information bound

1 Introduction

Suppose we have data (x, y) whose unconditional distribution is given by $f(y|x, \theta)$ $g(x)$, where $f(y|x, \theta)$ is a regression model representing the conditional distribution of y given x, and g is the unconditional density of x, assumed not to involve θ. The goal is the estimation of θ.

If the data are sampled from this joint distribution, no difficulties arise: the function g does not enter the likelihood calculations for the estimation of θ. On the other hand,

A. Lee (✉)
Department of Statistics, University of Auckland, Auckland, New Zealand
e-mail: lee@stat.auckland.ac.nz

Y. Hirose
School of Mathematics, Statistics and Computer Science, Victoria University of Wellington, Wellington, New Zealand

if the probability an individual is selected in the sample depends on y (the *response-selective* case), then things are not so simple and g must be included in the analysis.

In a series of papers, Scott and Wild (Scott and Wild 1997, Scott and Wild 2001, Wild 1991) have developed a methodology to handle this latter case, in which the function g is treated non-parametrically. Their method can be applied to a variety of response-selective sampling methods, including simple and stratified case-control studies. The method also permits the incorporation of supplementary information from a variety of sources, such as prospective samples from the joint distribution of (x, y) or the marginal distribution of x.

In this paper, we present a demonstration that the Scott–Wild method attains full non-parametric efficiency in all these situations. The efficiency of these methods has been demonstrated in special cases by several authors. For example, Breslow et al. (2000) consider case-control sampling, assuming that the data are generated by Bernoulli sampling, where either a case or control is selected by a randomisation device with known selection probabilities, and the covariates of the resulting case or control are measured. In the case of two-phase outcome-dependent sampling, Breslow et al. (2003) apply the missing value theory of Robins et al. (1995) and Robins et al. (1994). Here, individuals in the population are selected at random and their status (e.g. case or control) is determined. Then with a probability depending on their status, the covariates are measured or not. The unobserved covariates are treated as missing data.

In the present paper, we present a unified method that enables us to demonstrate the efficiency of the Scott–Wild approach in a simple way. We first use an adaptation of the profile likelihood method due to Newey (1994) to derive a semi-parametric efficiency bound. We then derive a simple expression for the asymptotic variance of the Scott–Wild estimate. Next, we show that and then show that this asymptotic variance coincides with the efficiency bound, thus demonstrating the efficiency of the estimator. Finally, to illustrate the efficiency of the Scott–Wild method in finite samples, and to assess the accuracy of the asymptotic approximation to the finite-sample variance, we conducted a small simulation study comparing the semiparametric estimator to some other estimators in common use.

The paper is structured as follows. In Sect. 2, we describe the Scott–Wild approach in more detail, discuss some special cases, and give a formula for the asymptotic variance of the Scott–Wild estimator. In Sect. 3, we sketch the theory of semiparametric efficiency that we require, and present an extension of Newey (1994) characterisation of the efficiency bound in terms of a "expected population profile likelihood" to the case of multiple samples. We then use this theory to demonstrate the efficiency of the Scott–Wild estimator by showing that the efficiency bound for this problem coincides with the asymptotic variance. The simulation study is described in Sect. 4, and some further comments on special cases are made in Sect. 5. Proofs and other derivations are in Sect. 6.

2 The Scott–Wild approach to generalised case-control studies

In this section we review the Scott–Wild methodology and give an expression for the asymptotic variance of their estimates.

We assume that the population is divided into K disjoint strata, and that the stratum membership is completely determined by an individual's response and covariate vector (although typically it depends on the response and only some, perhaps even none, of the covariates).

Data are gathered according to the following sampling scheme: In the first phase of sampling, a prospective sample of size N is taken from the whole population, but only the stratum membership is recorded. Suppose N_k of the N sampled in this first stage fall in stratum k, for $k = 1, \ldots, K$. In the second phase, for each stratum k, a simple random sample of size n_k is taken from the N_k individuals sampled in the first phase, and the covariates and responses are measured. In addition, we assume that these data are supplemented by additional observations taken prospectively from the joint distribution of (X, Y), the unconditional distribution of X, together with futher individuals sampled prospectively with only the stratum observed.

Note that the density of x and y conditional on being a member of stratum k is

$$I_k(x, y) f(y|x, \theta) g(x)/Q_k, \quad k = 1, \ldots, K, \tag{1}$$

where $Q_k = \int\int I_k(x, y) f(y|x, \theta) g(x)\, \mathrm{d}x\, \mathrm{d}y$, $f(y|x, \theta)$ is the conditional density of y given x, g is the marginal density of x and I_k is a stratum indicator. It is also convenient to write $Q_k(x, \theta) = \int I_k(x, y) f(y|x, \theta)\, \mathrm{d}y$, so that $Q_k = \int Q_k(x, \theta) g(x)\, \mathrm{d}x$. Thus $Q_k(x, \theta)$ is the probability that an individual with covariate vector x will be in stratum k, and Q_k is the unconditional probability that an individual will be in stratum k.

As explained in Scott and Wild (2001), the log-likelihood for this problem is of the form

$$\sum_A \log f(y|x, \theta) + \sum_B \log g(x) + \sum_{k=1}^{K} m_k \log Q_k \tag{2}$$

where A is the set of individuals who contribute a term $\log f(y|x, \theta)$ to the likelihood (i.e. those in either a prospective sample from the joint distribution, or in one of the second-stage samples), B consists of those in either a prospective sample from the joint distribution, a prospective sample from the marginal distribution, or in one of the second-stage samples, and m_k is a count to which prospectively sampled individuals with only the stratum observed contribute $+1$, and second stage individuals contribute -1.

This general formulation covers a variety of special cases.

1. *The simple case-control study* (Prentice and Pyke 1979). Separate samples of cases and controls are taken from the case and control populations, respectively. Thus there are two strata (cases and controls), no first stage sample (or rather the first stage sample is the whole population) and no supplementary prospective samples.
2. *Two-stage case-control study*. A first stage random sample is taken, and the sampled individuals identified as cases and controls. Then for the second stage of the study, sub-samples are taken from the case and control samples taken at the first stage. No supplementary prospective sampling is done.

3. *Two-stage sampling design* (White 1982). A first stage sample is taken, and divided into a finite number of strata on the basis of the response and certain of the covariates. At the second stage, separate sub-samples are taken from each stratum and futher covariates are measured. Again, no supplementary prospective sampling is done. The two-stage case-control study above is a special case, with strata defined by cases and controls.
4. *Reusing data from case-control studies* (Lee et al. 1997; Jiang et al. 2006). A two-stage case control study is performed. Subsequent to the completion of the study, the data are reanalysed with a discrete covariate measured at the first stage in the first analysis now being used as a discrete response in the second analysis.
5. *Case-augmented sampling* (Lee et al. 2006). Here a prospective sample is taken from the joint distribution of (x, y), where y denotes case or control. In addition, an further sample of cases is taken, and the covariates x measured. A variation is to only measure the covariate in the prospective sample. There is no first stage sample, as the case control status is assumed known for all individuals in the population.
6. *Family studies* (Whittemore 1995; Neuhaus et al. 2002). Here the sampling units are families and a binary response is measured on each family member. A first stage sample is taken, and the families are assigned to strata on the basis of the family responses. Second stage sub-samples are taken from the separate strata. No supplementary prospective samples are taken.
7. *Case control study augmented with population data*. A one- or two-stage case-control study can be augmented with additional prospective data, for example from routinely collected information in hospital records.
8. *Missing data problems* (Whittemore 1995; Lawless et al. 1999). Suppose we have a discrete response variable y and a discrete covariate v. We sample y, v prospectively, and for each unit sampled, with probability $\pi(y, v)$ we measure the value of a more expensive covariate z, which may be continuous or discrete. The goal is to fit a model representing the conditional distribution of y, given v and z. Note that this formulation also covers the case where the covariates z are missing at random (i.e. MAR missingness, where the probability of observing z depends only on y and v).
9. *Analysis of survival and reliability data* (Kalbfleisch and Lawless 1988; Hu and Lawless 1996). Here the strata are formed by censored and non-censored observations. The covariates are available for the all the non-censored observations, but covariate information is available on only some of the censored observations.

The general sampling scheme considered above is equivalent (in the sense of having the same likelihood and asymptotics) to taking $J = K + 3$ independent samples, namely

1. A sample of n_1 individuals sampled unconditionally with only the stratum observed, i.e. from a multinomial distribution with density

$$p_1(x, y, \theta, g) = Q_1^{z_1} \cdots Q_K^{z_K}. \tag{3}$$

Here the z's are stratum indicators with $z_k = I_k(x, y)$ having value 1 if an observation is in stratum k, and zero otherwise. Let $n_1^{(k)}$ be the number falling into stratum k.

2. A sample of n_2 individuals sampled prospectively from the unconditional joint distribution of (X, Y), with density $p_2(x, y, \theta, g) = f(y|x, \theta)g(x)$.
3. A sample of n_3 individuals sampled prospectively from the unconditional distribution of X, with density $p_3(x, y, \theta, g) = g(x)$.
4. For $k = 1, \ldots, K$ we have samples of size $n_4^{(k)}$ from the distribution of (X, Y) conditional on being in stratum k, with densities given by the formula

$$p_{4,k}(x, y, \theta, g) = I_k(x, y) f(y|x, \theta)g(x)/Q_k, \quad k = 1, \ldots, K.$$

The density g is an infinite-dimensional nuisance parameter. We will also assume that $n_1 Q_k \geq n_4^{(k)}$, corresponding to the fact that $N_k \geq n_k$ in the original sampling scheme. Note that under this sampling scheme, we can combine the prospectively sampled individuals for which stratum membership only is observed and the first stage individuals into one group. In the rest of the paper we work with this alternative sampling scheme.

The equivalence of this independent scheme to the two-phase sampling scheme described in the beginning of this section is proved in Lee (2007a) in the case where no prospective samples are included. The proof for the case considered here is similar.

Let $N = n_1 + n_2 + n_3 + \sum_{k=1}^{K} n_4^{(k)}$, let $\rho = (\rho_1, \ldots, \rho_{K-1})^T$ be an arbitrary vector, and let $Q_k(\rho)$, $k = 1, \ldots, K$ be a set of probabilities defined by $\sum_{k=1}^{K} Q_k(\rho) = 1$ and $\log(Q_k/Q_K) = \rho_k$, $k = 1, \ldots, K-1$.

Scott and Wild (2001) show that the the profile likelihood obtained by maximizing (2) over g for fixed θ is of the form $l^*(\theta, \rho_\theta)$, where

$$l^*(\theta, \rho) = \sum_A \log f(y|x, \theta) - \sum_B \log \left\{ \sum_{k=1}^{K} \mu_k^{(N)}(\rho) Q_k(x, \theta) \right\} + \sum_{k=1}^{K} (n_1^{(k)} - n_4^{(k)}) \log Q_k(\rho), \tag{4}$$

$\mu_k^{(N)}(\rho) = N^{-1}\{n_1 + n_2 + n_3 - (n_1^{(k)} - n_4^{(k)})/Q_k(\rho)\}$ and ρ_θ satisfies $\frac{\partial l^*}{\partial \rho} = 0$. It follows that $\hat{\theta}$, the MLE of θ, is the "θ" part of the solution $\hat{\phi}$ of the estimating equation

$$\frac{\partial l^*}{\partial \phi} = 0, \tag{5}$$

where $\phi = (\theta^T, \rho^T)^T$. Thus, for the purposes of estimation, we can treat l^* as if it were an ordinary log-likelihood.

This also extends to the estimation of standard errors: we can estimate the covariance matrix of $\hat{\theta}$ by the $\theta\theta$ block of the "pseudo information matrix"

$$\left(-\frac{\partial^2 l^*}{\partial\phi\partial\phi^T}\right)^{-1}.$$

The consistency of this estimate is demonstrated in the following result, which also expresses the asymptotic variance in terms of the profile information matrix, and gives a formula for the asymptotic variance of $\hat{\phi}$, the solution of (5).

Theorem 1 (i) *The asymptotic variance of $\hat{\phi}$ is given by*

$$\lim_{N\to\infty} N\,Var(\hat{\phi}) = (\mathbf{I}^*)^{-1} - \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\mathbf{A}^-)^T\mathbf{D}\mathbf{A}^- \end{pmatrix},$$

where $\mathbf{A}$ *and* $\mathbf{D}$ *are matrices defined in Sect. 5.1.*

(ii) *Let* $\mathbf{I}^* = -\text{plim}_{N\to\infty} N^{-1}\frac{\partial^2 l^*}{\partial\phi\partial\phi^T}$. *Partition* $\mathbf{I}^*$ *as*

$$\mathbf{I}^* = \begin{bmatrix} \mathbf{I}^*_{\theta\theta} & \mathbf{I}^*_{\theta\rho} \\ \mathbf{I}^*_{\rho\theta} & \mathbf{I}^*_{\rho\rho} \end{bmatrix}.$$

Then

$$\lim_{N\to\infty} N Var(\hat{\theta}) = (\mathbf{I}^*_{\theta\theta} - \mathbf{I}^*_{\theta\rho}\mathbf{I}^*_{\rho\rho}{}^{-1}\mathbf{I}^*_{\rho\theta})^{-1}. \tag{6}$$

(iii) *The "profile information matrix"* $-\frac{\partial^2 l^*(\theta,\rho_\theta)}{\partial\theta\partial\theta^T}$ *satisfies*

$$\text{p}\lim_{N\to\infty} -N^{-1}\frac{\partial^2 l^*(\theta,\rho_\theta)}{\partial\theta\partial\theta^T} = \mathbf{I}^*_{\theta\theta} - \mathbf{I}^*_{\theta\rho}\mathbf{I}^*_{\rho\rho}{}^{-1}\mathbf{I}^*_{\rho\theta} \tag{7}$$

so that the variance of θ is consistently estimated by the inverse of the profile information matrix.

This result is stated in Scott and Wild (2001) but no proof in this general case has appeared in the literature. We sketch a proof in Sect. 6.1. Note that part (i) of the theorem can be used to find a standard error for $\hat{\rho} = \rho_{\hat{\theta}}$, and hence, using the delta method, for a standard error of the estimate $Q_k(\hat{\rho})$ of Q_{k0}, the unconditional probability of being in stratum k.

3 Information bounds via profile likelihood for the multi-sample case

In this section, we first give a short account of the theory of semi-parametric efficiency in the multi-population case and describe how to calculate the efficiency bound. We then apply this theory to prove the efficiency of the Scott–Wild estimator. No proofs are given, but these may be found in Lee (2007b).

3.1 The efficiency bound: general case

Suppose we have J populations. Random sampling from these populations is described by a set of J densities $p_{j0} = p_j(x, \theta_0, \eta_0)$ which are contained in the family of densities

$$\mathscr{P} = \{p_j(x, \theta, \eta) : j = 1, \ldots, J;\ \theta \in \mathscr{B};\ \eta \in \mathscr{N}\}$$

where θ is a k-dimensional parameter belonging to a set $\mathscr{B}$ and η is an infinite dimensional parameter, belonging to a set $\mathscr{N}$. We also assume that we have available a sample of size n_j from population j. All asymptotics are done assuming that $n_j/n \to w_j$, where $n = n_1 + \cdots + n_J$.

Suppose the jth sample is $X_{ij}, i = 1, 2, \ldots, n_j$ and that $\hat{\theta}$ is a regular asymptotically linear (RAL) estimate of θ based on these J samples; see Bickel et al. (1993) for a definition of this concept. Then there exist vector-valued functions ψ_j with

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = n^{-1/2} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \psi_j(X_{ij}) + o_p(1). \tag{8}$$

The functions ψ_j are called the *influence functions* of the estimate and the asymptotic variance of the estimate is

$$\text{Avar}(\hat{\theta}) = \sum_{j=1}^{J} w_j E_j(\psi_j \psi_j^T),$$

where E_j denotes expectation with respect to p_{j0}. Note that the influence functions are assumed to satisfy $E_j[\psi_j] = 0$. The *efficiency bound* for this family of densities is a matrix $\mathbf{B}$ such that $\text{Avar}(\hat{\theta}) \geq \mathbf{B}$ for all RAL estimates of θ. The matrix $\mathbf{B}$ is found as follows: Let $\mathscr{G}$ be a finite-dimensional set of dimension r say, so that

$$\{p_j(x, \theta, \eta(\gamma)) : j = 1, \ldots, J;\ \theta \in \mathscr{B};\ \gamma \in \mathscr{G}\}$$

is a finite-dimensional sub-family of $\mathscr{P}$, assumed to contain the true model p_{j0}. Consider the vector-valued score functions

$$\dot{l}_{j,\eta} = \frac{\partial \log p_j(x, \theta, \eta(\gamma))}{\partial \gamma},$$

whose elements are assumed to be members of $L_2(P_{j0})$, where P_{j0} is the measure corresponding to $p_j(x, \theta_0, \eta_0)$. Consider also the space $L_{2k}(P_{j0})$, the space of all $\Re_k$-valued functions square-integrable with respect to P_{j0}, and the Cartesian product $\mathscr{H}$ of these spaces, equipped with the norm defined by

$$||(f_1, \ldots, f_J)||^2_{\mathscr{H}} = \sum_{j=1}^{J} w_j \int ||f_j||^2 dP_{j0}.$$

The subspace of $\mathscr{H}$ generated by the score functions $(\dot{l}_{1,\eta}, \ldots, \dot{l}_{J,\eta})$ is the set of all vector-valued functions of the form $(\mathbf{A}\dot{l}_{1,\eta}, \ldots, \mathbf{A}\dot{l}_{J,\eta})$ where $\mathbf{A}$ ranges over all k by r matrices. Thus, to each finite-dimensional sub-family of $\mathscr{P}$, there corresponds a score function and subspace of $\mathscr{H}$ generated by the score function. The closure in $\mathscr{H}$ of the union (over all such sub-families) of all these subspaces is called the *nuisance tangent space* and is denoted by $\mathscr{T}_\eta$. This space is fundamental to the definition of the efficiency bound.

Now consider the score functions

$$\dot{l}_{j,\theta} = \frac{\partial \log p_j(x, \theta, \eta)}{\partial \theta}.$$

Note that $\dot{l}_\theta = (\dot{l}_{1,\theta}, \ldots, \dot{l}_{J,\theta})$ is also a member of $\mathscr{H}$. The projection of $\dot{l}_{j\theta}$ onto the orthogonal complement of $\mathscr{T}_\eta$ is called the *efficient score*, and is denoted by $\dot{l}^*_j$. The matrix $\mathbf{B}$ (the efficiency bound) is given by

$$\mathbf{B}^{-1} = \sum_{j=1}^{J} w_j E_j[\dot{l}^*_j \dot{l}^{*T}_j]. \tag{9}$$

The functions $\mathbf{B}\dot{l}^*_j$ are called the *efficient influence functions*, and any multi-sample RAL estimate having these influence functions is asymptotically efficient.

To find the efficient score, we use the following extension of Newey (1994) i.i.d. result characterizing the efficient score in terms of the "population expected log-likelihood".

Theorem 2 *For fixed θ, let $\hat{\eta}(\theta)$ be the maximiser in $\mathscr{N}$ of the "population expected log-likelihood"*

$$\sum_{j=1}^{J} w_j E_j[\log p_j(X, \theta, \eta)]. \tag{10}$$

Then the efficient scores are

$$\dot{l}^*_j = \left. \frac{\partial \log p_j(x, \theta, \hat{\eta}(\theta))}{\partial \theta} \right|_{\theta=\theta_0}.$$

A proof of this theorem is given in Sect. 6.2.

The distributions $p_j(x, \theta, \hat{\eta}(\theta))$ are called the *least favourable distributions* for the problem: they are essentially the distributions having finite dimensional parameters for which the MLE's have the largest possible variance (and attain the information bound). In the case of the response-selective sampling schemes we consider in the rest of the paper, it turns out that the least favorable distributions have a special form that allows the information bound to be calculated very simply.

3.2 The information bound for response-selective studies

In this section we apply the theory of Sect. 3.1 to regression models for data obtained by the various forms of response-selective sampling described in Sect. 2. To calculate the information bound, we first calculate the expected log-likelihood. Denote expectation with respect to the unconditional distributions by E and with respect to the distribution conditional on being in stratum k by E_k, taken at the true values θ_0 and g_0 of θ and g. We also assume that $n_j/N \to w_j$, $j = 1, 2, 3$, and $n_4^{(k)}/N \to w_4^{(k)}$, $k = 1, \ldots, K$ where $N = n_1 + n_2 + n_3 + \sum_{k=1}^{K} n_4^{(k)}$. Writing Z_k for the (random) stratum indicator that takes value 1 if a randomly chosen individual sampled at phase one is in stratum k, the expected log-likelihood (10) takes the form

$$\sum_{k=1}^{K} w_1 E[Z_k \log Q_k] + w_2 E[\log\{f(Y|X,\theta)g(X)\}] + w_3 E[\log g(X)]$$
$$+ \sum_{k=1}^{K} w_4^{(k)} \{E_k[\log I_k(X,Y) f(Y|X,\theta)] + E_k[\log g(X)] - \log Q_k\},$$

which up to a term not involving g can be written

$$\int \log g(x) Q^*(x) g_0(x)\, \mathrm{d}x + \sum_{k=1}^{K} c_k \log Q_k, \tag{11}$$

where $c_k = w_1 Q_{k0} - w_4^{(k)}$, $Q^*(x) = \sum_{k=1}^{K} (w_2 + w_3 + w_4^{(k)}/Q_{k0}) Q_k(x, \theta_0)$ and $Q_{k0} = \int Q_k(x, \theta_0) g_0(x)\, \mathrm{d}x$. We need to maximize (11) over g with θ held fixed.

We first assume that the distribution of X is discrete with finite support $\{x_1, \ldots, x_L\}$, putting mass g_l at x_l. Then we can write (11) as

$$\sum_{l=1}^{L} \log g_l\, Q^*(x_l) g_0(x_l) + \sum_{k=1}^{K} c_k \log \left\{ \sum_{l=1}^{L} g_l Q_k(x_l, \theta) \right\}. \tag{12}$$

Introduce a Lagrange multiplier λ to take account of the constraint $\sum_l g_l = 1$. Then, differentiating with respect to g_l gives

$$\frac{Q^*(x_l) g_0(x_l)}{g_l} + \sum_{k=1}^{K} c_k \left\{ \frac{Q_k(x_l, \theta)}{\sum_{l=1}^{L} g_l Q_k(x_l, \theta)} \right\} + \lambda = 0$$

and multiplying by g_l and adding over l gives $\lambda = -(w_1 + w_2 + w_3)$. Hence the maximizing g is of the form

$$g_l = \frac{Q^*(x_l) g_0(x_l)}{\sum_{k=1}^{K} \mu_k Q_k(x_l, \theta)}, \tag{13}$$

where $\mu_k = w_1 + w_2 + w_3 - c_k/Q_k$ and $Q_k = \sum_{l=1}^{L} g_l Q_k(x_l, \theta)$.

This suggests that in the case of a general g_0, not having finite support, the maximiser of (11) might be of the form

$$g(x, \theta, \rho_\theta) = \frac{Q^*(x) g_0(x)}{\sum_k \mu_k(\rho_\theta) Q_k(x, \theta)}, \tag{14}$$

where $\mu_k(\rho_\theta) = w_1 + w_2 + w_3 - c_k/Q_k(\rho_\theta)$ and $Q_k(\rho_\theta)$ satisfies the equation

$$Q_k(\rho_\theta) = \int g(x, \theta, \rho_\theta) Q_k(x_l, \theta)\, \mathrm{d}x.$$

This turns out to be the case. We give a sketch of the proof in Sect. 6.3.

Our next task is to calculate the efficient scores. Applying Theorem 2, we see that they are

$$i_1^* = \sum_{k=1}^{K} z_k \left. \frac{\partial \log Q_k(\rho_\theta)}{\partial \theta} \right|_{\theta=\theta_0}, \tag{15}$$

$$i_2^* = \left. \frac{\partial \log\{f(x|y, \theta) g(x, \theta, \rho_\theta)\}}{\partial \theta} \right|_{\theta=\theta_0}, \tag{16}$$

$$i_3^* = \left. \frac{\partial \log g(x, \theta, \rho_\theta)}{\partial \theta} \right|_{\theta=\theta_0}, \tag{17}$$

$$i_{4,k}^* = \left. \frac{\partial \log\{f(x|y, \theta) g(x, \theta, \rho_\theta)\} - \log Q_k(\rho_\theta)}{\partial \theta} \right|_{\theta=\theta_0}. \tag{18}$$

Now we can obtain the information bound in terms of the "asymptotic pseudo-information matrix" $\mathbf{I}^*$ introduced in Sect. 2. From (9) and (16)–(18), the inverse of the information bound $\mathbf{B}$ is

$$\begin{aligned}
\mathbf{B}^{-1} = {} & w_1 E\left[\left\{ \sum_{k=1}^{K} Z_k \frac{\partial \log Q_k(\rho_\theta)}{\partial \theta} \right\} \left\{ \sum_{k=1}^{K} Z_k \frac{\partial \log Q_k(\rho_\theta)}{\partial \theta} \right\}^T \right] \\
& + w_2 E\left[\left\{ \frac{\partial \log\{f(x|y, \theta) g(x, \theta, \rho_\theta)\}}{\partial \theta} \right\} \left\{ \frac{\partial \log\{f(x|y, \theta) g(x, \theta, \rho_\theta)\}}{\partial \theta} \right\}^T \right] \\
& + w_3 E\left[\left\{ \frac{\partial \log g(x, \theta, \rho_\theta)}{\partial \theta} \right\} \left\{ \frac{\partial \log g(x, \theta, \rho_\theta)}{\partial \theta} \right\}^T \right] \\
& + \sum_{k=1}^{K} w_4^{(k)} E_k\left[\left\{ \frac{\partial \log\{f(x|y, \theta) g(x, \theta, \rho_\theta)\}}{\partial \theta} - \frac{\partial \log Q_k(\rho_\theta)}{\partial \theta} \right\} \right. \\
& \left. \times \left\{ \frac{\partial \log\{f(x|y, \theta) g(x, \theta, \rho_\theta)\}}{\partial \theta} - \frac{\partial \log Q_k(\rho_\theta)}{\partial \theta} \right\}^T \right].
\end{aligned} \tag{19}$$

Then, using the fact that

$$E_k\left[\frac{\partial \log\{f(x,\theta)g(x,\theta,\rho_\theta)\}}{\partial\phi}\right] = \frac{\partial \log Q_k(\rho_\theta)}{\partial\phi}$$

and the chain rule, we get

$$\mathbf{B}^{-1} = \mathbf{I}^\dagger_{\theta\theta} + \left(\frac{\partial\rho_\theta}{\partial\theta}\right)^T \mathbf{I}^\dagger_{\rho\theta} + \mathbf{I}^\dagger_{\theta\rho}\frac{\partial\rho_\theta}{\partial\theta} + \left(\frac{\partial\rho_\theta}{\partial\theta}\right)^T \mathbf{I}^\dagger_{\rho\rho}\left(\frac{\partial\rho_\theta}{\partial\theta}\right), \tag{20}$$

where $\mathbf{I}^\dagger$ is the matrix

$$\begin{aligned}\mathbf{I}^\dagger = {} & w_2 E\left[\frac{\partial \log\{f(x,\theta)g(x,\theta,\rho)\}}{\partial\phi}\frac{\partial \log\{f(x,\theta)g(x,\theta,\rho)\}}{\partial\phi^T}\right] \\ & + w_3 E\left[\frac{\partial \log g(x,\theta,\rho)}{\partial\phi}\frac{\partial \log g(x,\theta,\rho)}{\partial\phi^T}\right] \\ & + \sum_{k=1}^K w_4^{(k)} E_k\left[\frac{\partial \log\{f(x,\theta)g(x,\theta,\rho)\}}{\partial\phi}\frac{\partial \log\{f(x,\theta)g(x,\theta,\rho)\}}{\partial\phi^T}\right] \\ & + \sum_{k=1}^K c_k \frac{\partial \log Q_k}{\partial\phi}\frac{\partial \log Q_k}{\partial\phi^T}\end{aligned}$$

introduced in Sect. 6.1. We show in Sect. 6.4 that

$$\mathbf{I}^\dagger_{\theta\theta} = \mathbf{I}^*_{\theta\theta}, \tag{21}$$

$$\mathbf{I}^\dagger_{\theta\rho} = 0, \tag{22}$$

$$\mathbf{I}^\dagger_{\rho\rho} = -\mathbf{I}^*_{\rho\rho}, \tag{23}$$

and that

$$\frac{\partial\rho_\theta}{\partial\theta} = -(\mathbf{I}^*_{\rho\rho})^{-1}\mathbf{I}^*_{\rho\theta}. \tag{24}$$

Substituting these results into (20) gives

$$\mathbf{B}^{-1} = \mathbf{I}^*_{\theta\theta} - \mathbf{I}^*_{\theta\rho}(\mathbf{I}^*_{\rho\rho})^{-1}\mathbf{I}^*_{\rho\theta}. \tag{25}$$

Thus, the asymptotic variance of the Scott–Wild estimator coincides with the information bound, and so the estimator is fully efficient.

4 Simulation study

To illustrate the efficiency of the Scott–Wild method in finite samples, we conducted a small sampling experiment to compare the efficiency of the semi-parametric MLE

to two other estimators in common use, namely the conditional pseudo-likelihood estimator (Breslow and Cain (1988)) and the weighted pseudo-likelihood estimator, which employs the Horvitz-Thompson approach used in sample survey problems. This follows a similar experiment in Sect. 6 of Lawless et al. (1999). We consider the case of stratified two-phase sampling, where the population is divided into strata, a first-stage sample is taken with only strata membership observed, followed by a second stage samples from each stratum with further covariates observed.

In our simulation, we took the model $f(y|x,\theta)$ to be logistic, with a binary response y and covariates $x = (x_1, x_2)$, so that logit $f(y|x,\theta) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$. The covariate x_2 was taken to have a standard normal distribution, while the covariate x_1 was a discretised version of another standard normal variate correlated 0.5 with x_2. The resulting covariate x_1 had six values, denoted by $x_1^{(j)}$, $j = 1, \ldots, 6$. The strata were taken to be the twelve possible combinations of x_1 and y, so that y and x_1 are observed for all individuals sampled at the first phase, but x_2 is only observed for second-phase individuals. We denote the stratum with $y = i$ and $x_1 = x_1^{(j)}$ by $\mathscr{S}_{ij}$, using a double index notation in contrast to our previous practice. This is the "expensive covariate" problem, where the "cheap" covariate x_1 is measured for all units, but the "expensive" covariate x_2 is measured for the second-phase individuals only.

A large sample of $N = 75,000$ individuals is taken at the first stage, resulting in N_{ij} individuals falling in $\mathscr{S}_{ij}$. From these individuals, a balanced second phase sub-sample of size $n_{ij} = 75$ is selected. In the language of Lawless et al. (1999), the sampling is basic stratified sampling with fixed second phase sample sizes. We denote the values of the covariate vector for the individuals sampled from $\mathscr{S}_{ij}$ at the second phase by $x_{ijk}, k = 1, \ldots, n_{ij}$.

In addition to the semi-parametric estimator, we consider two other estimators of the regression coefficients, the conditional pseudo-likelihood estimator and the weighted pseudo-likelihood estimator. In the context of two-phase studies, conditional pseudo-likelihood has been studied by Hsieh et al. (1985), Breslow and Cain (1988), Scott and Wild (1997) and Lawless et al. (1999). In our context, the estimate is obtained by maximising the pseudolikelihood

$$\sum_i \sum_j \sum_k \log \left\{ \frac{\hat{\mu}_{ij} Q_{ij}(x_{ijk}, \theta)}{\sum_i \sum_j \hat{\mu}_{ij} Q_{ij}(x_{ijk}, \theta)} \right\},$$

where $\hat{\mu}_{ij} = n_{ij}/N_{ij}$. Note that for the present set of strata,

$$Q_{ij}(x,\theta) = \begin{cases} f(y = i | x = (x_1^{(j)}, x_2), \theta), & \text{if } x_1 = x_1^{(j)} \\ 0, & \text{otherwise.} \end{cases} \tag{26}$$

The weighted pseudo-likelihood estimator has been considered in the response-selective context by Hsieh et al. (1985), Scott and Wild (2002) and Lawless et al. (1999). It merely weights up the terms in the ordinary prospective likelihood to compensate for the differential sampling rates. This results in the estimating equa-

Table 1 Parameter settings for the simulation

Settings	θ_0	θ_1	θ_2
1	−2.95	0.0	0.0
2	−3.05	0.5	0.0
3	−3.25	0.5	0.5
4	−3.35	0.0	1.0
5	−3.30	1.0	0.0
6	−4.00	1.0	1.0

Table 2 Efficiency (%) relative to the semiparametric estimator, computed by simulation (Sim) and by using the asymptotic variance formula (Asym)

Setting	θ_0 (Sim)	θ_0 (Asym)	θ_1 (Sim)	θ_1 (Asym)	θ_2 (Sim)	θ_2 (Asym)
Conditional estimator						
1	99	97	100	98	105	100
2	98	97	102	99	103	100
3	98	99	95	99	96	100
4	96	100	98	100	98	100
5	93	92	95	94	102	100
6	90	97	87	91	95	100
Weighted estimator						
1	100	100	95	95	94	94
2	99	100	96	92	94	91
3	76	75	87	91	74	76
4	73	71	73	72	71	71
5	99	100	82	75	77	71
6	40	38	74	73	52	50

tion

$$\sum_i \sum_j \sum_k \frac{N_{ij}}{n_{ij}} \frac{\partial}{\partial \theta} \log f(y = i | x_{ijk}, \theta) = 0.$$

We tried six combinations of parameter values, with θ_1 and θ_2 set as shown in Table 1, and θ_0 chosen to make the probability of disease (i.e. $P[y = 1]$) have a value around 0.05. For each set of parameter values, we simulated the basic sampling scenario described above 1000 times. We then calculated the efficiency of the conditional and weighted estimates relative to the SPMLE by taking the ratios of the empirical variances from the simulation.

We also calculated the asymptotic relative efficiencies by computing the asymptotic variances using the formula given in the present paper for the SPMLE, and similar formulae for the other methods, which are given in Sect. 6.5. The simulated and asymptotic efficiencies are shown in Table 2. Assuming multivariate normality, and

using the delta method, the standard errors of the simulated efficiencies are in the range 1.5–2.5%, so the efficiencies obtained using the asymptotic formula is in good agreement with the simulated values.

It is clear that the conditional method is very competitive with the SPMLE, at least in this example. This is not surprising, since the two methods differ only in the way that the parameters μ_{ij} are estimated: in the case of the SPMLE a separate estimating equation is required, but for the conditional estimate a simpler estimate $\hat{\mu}_{ij} = n_{ij}/N_{ij}$ is used. Moreover, if a separate parameter is included in the model for each stratum (i.e. if the variable x_1 had entered the model as a factor rather than a numeric variable) then the semi-parametric estimating equation for μ_{ij} would result in the estimate n_{ij}/N_{ij} so that the two methods would coincide. The weighted method loses considerable efficiency for several of the parameter settings, in contrast with the conditional estimate, which can be slightly more efficient than the semi-parametric estimate in finite samples. However, there are situations where the conditional method is inferior to the SPMLE; see the simulation reported in Scott and Wild (1991).

5 Discussion

In this section, we re-examine the special cases of our general sampling scheme and indicate how the general efficiency result applies.

1. *The simple case-control study*. In this situation our general result applies with $K = 2$ and $w_1 = w_2 = w_3 = 0$. The variable y is a binary indicator denoting case or control and $f(1|x, \theta)$ is the conditional probability of being a case, given covariates x.
2. *Two-stage case-control study*. Here the situation is identical to that in 1, except that $w_1 > 0$.
3. *Two-stage sampling design*. Here we have $w_2 = w_3 = 0$. The regression function can be general as long as the number of strata is finite and strata membership depends only on (x, y).
4. *Reusing data from case-control studies*. This situation is similar to 2, except that the regression function is of the form $f(y_1, y_2|x, \theta) = f_1(y_1|y_2, x, \theta) f_2(y_2|x, \theta)$ where y_1 is the response for the first analysis, y_2 is the response for the second analysis, and $f_2(y_2|x, \theta)$ is the regression of interest in the second analysis.
5. *Case-augmented sampling*. In the first case considered, with a prospective sample from the joint distribution, our general result applies with $w_1 = 0$, $w_3 = 0$, and $w_4^{(k)} = 0$ for $k > 2$. In the second case, with a prospective sample from the marginal distribution of x, the general result applies with $w_1 = 0$, $w_2 = 0$, and $w_4^{(k)} = 0$ for $k > 2$. Extensions to discrete responses with more than two values are immediate.
6. *Retrospective family studies*. This is similar to 4, with a multiple response in the regression representing responses on different family members.
7. *Case-control study augmented with population data*. If the case-control study has two stages, and the population data is in the form of additional prospective samples from both the joint and marginal distributions of x and y, the full specification (i.e. none of the w's zero) is required.

8. *Missing data problems*. Provided the covariate v and the response y are discrete, the log-likelihood for the missing value problem can be written in the form (2) (Lawless et al. 1999), and hence our results apply.
9. *Analysis of survival and reliability data*. This falls into the same framework as 8 (Lawless et al. 1999).

Thus, our general result is sufficient to demonstrate the efficiency of the Scott–Wild estimator in all the situations described above.

6 Proofs and derivations

6.1 The asymptotic variance and the proof of Theorem 1

We begin by deriving some expressions for the "pseudo information matrix " $\mathbf{I}^*$ that will be useful in establishing the asymptotic variance of $\hat{\theta}$. To evaluate $\mathbf{I}^*$, we split the terms of (4) into separate sums corresponding to the different samples, differentiate, and apply the law of large numbers to each part. This results in

$$\mathbf{I}^* = w_2 E\left[-\frac{\partial^2 \log\{f(y|x,\theta)g(x,\theta,\rho)\}}{\partial\phi\partial\phi^T}\right] + w_3 E\left[-\frac{\partial^2 \log g(x,\theta,\rho)}{\partial\phi\partial\phi^T}\right] + \sum_{k=1}^{K} w_4^{(k)} E_k\left[-\frac{\partial^2 \log\{f(y|x,\theta)g(x,\theta,\rho)\}}{\partial\phi\partial\phi^T}\right] - \sum_{k=1}^{K} c_k \frac{\partial^2 \log Q_k(\rho)}{\partial\phi\partial\phi^T}, \quad (27)$$

where $c_k = w_1 Q_{k0} - w_4^{(k)}$, and

$$g(x,\theta,\rho) = \frac{Q^*(x)g_0(x)}{\sum_{k=1}^{K}\mu_k(\rho)Q_k(x,\theta)}.$$

In (27), we are using E to denote expectation with respect to the unconditional (prospective) distributions and E_k to denote expectations conditional on being in stratum k.

Using the identity

$$\frac{\partial^2 \log h(\phi)}{\partial\phi\partial\phi} = \frac{1}{h}\frac{\partial^2 h(\phi)}{\partial\phi\partial\phi} - \frac{\partial \log h(\phi)}{\partial\phi}\frac{\partial \log h(\phi)}{\partial\phi^T}$$

and the fact that $g(x,\theta_0,\rho_0) = g_0(x)$, we get

$$\mathbf{I}^* = \left\{ w_2 E\left[\frac{\partial \log\{f(x,\theta)g(x,\theta,\rho)\}}{\partial\phi}\frac{\partial \log\{f(x,\theta)g(x,\theta,\rho)\}}{\partial\phi^T}\right] + w_3 E\left[\frac{\partial \log g(x,\theta,\rho)}{\partial\phi}\frac{\partial \log g(x,\theta,\rho)}{\partial\phi^T}\right]\right.$$

$$+\sum_{k=1}^{K} w_4^{(k)} E_k\left[\frac{\partial \log\{f(x,\theta)g(x,\theta,\rho)\}}{\partial\phi}\frac{\partial \log\{f(x,\theta)g(x,\theta,\rho)\}}{\partial\phi^T}\right]$$
$$+\sum_{k=1}^{K} c_k \frac{\partial \log Q_k}{\partial\phi}\frac{\partial \log Q_k}{\partial\phi^T}\Bigg\}$$
$$-\Bigg\{w_2 E\left[\frac{1}{fg_0}\frac{\partial^2 f(y|x,\theta)g(x,\theta,\rho)}{\partial\phi\partial\phi^T}\right]+ w_3 E\left[\frac{1}{g_0}\frac{\partial^2 g(x,\theta,\rho)}{\partial\phi\partial\phi^T}\right]$$
$$+\sum_{k=1}^{K} w_4^{(k)} E_k\left[\frac{1}{fg_0}\frac{\partial^2 f(y|x,\theta)g(x,\theta,\rho)}{\partial\phi\partial\phi^T}\right]+\sum_{k=1}^{K} c_k \frac{1}{Q_k}\frac{\partial^2 Q_k}{\partial\phi\phi^T}\Bigg\}.$$

Denoting the sum in the first set of braces by $\mathbf{I}^{\dagger}$, and collecting the first three terms in the second set of braces into a single integral, we get

$$\mathbf{I}^* = \mathbf{I}^{\dagger} - \int \frac{\partial^2}{\partial\phi\partial\phi^T}\left\{\frac{\sum_{k=1}^{K}\mu_{k0}Q_k(x,\theta)}{\sum_{k=1}^{K}\mu_k(\rho)Q_k(x,\theta)}\right\} Q^*(x)g_0(x)\,\mathrm{d}x - \sum_{k=1}^{K} c_k \frac{1}{Q_k}\frac{\partial^2 Q_k}{\partial\phi\phi^T}. \tag{28}$$

Moreover, for the $\theta\rho$, $\rho\theta$ and $\rho\rho$ blocks of $\mathbf{I}^*$, note that the function f drops out of (27) and we can write these blocks as

$$-\int \frac{\partial^2 \log g(x,\theta,\rho)}{\partial\phi\partial\phi^T} Q^*(x)g_0(x)\,\mathrm{d}x - \sum_{k=1}^{K} c_k \frac{\partial^2 \log Q_k(\rho)}{\partial\phi\partial\phi^T}.$$

Thus, evaluating these derivatives, we get

$$\mathbf{I}^*_{\rho\theta} = \sum_{k=1}^{K} w_4^{(k)} \frac{\partial \log \mu_k(\rho)}{\partial\rho} E_{k\theta}^T \tag{29}$$

where

$$E_{k\theta} = \frac{1}{w_4^{(k)}}\int \frac{\partial P_k(x,\theta,\rho)}{\partial\theta} Q^*(x)g_0(x)\,\mathrm{d}x.$$

Similarly,

$$\mathbf{I}^*_{\rho\rho} = \sum_{k=1}^{K} w_4^{(k)} \frac{\partial \log \mu_k(\rho)}{\partial\rho}\left(E_{k\rho} - \frac{\partial \log Q_k(\rho)}{\partial\rho}\right)^T \tag{30}$$

where

$$E_{k\rho} = \frac{1}{w_4^{(k)}}\int \frac{\partial P_k(x,\theta,\rho)}{\partial\rho} Q^*(x)g_0(x)\,\mathrm{d}x - \frac{\partial \log \mu_k(\rho)}{\partial\rho},$$

with

$$P_k(x,\theta,\rho) = \frac{\mu_k(\rho)Q_k(x,\theta)}{\sum_{k=1}^{K}\mu_k(\rho)Q_k(x,\theta)}.$$

Proof of Theorem 1 (*i*) A complicating factor in the evaluation of the asymptotic variance is the fact that the quantities $\mu_k^{(N)}(\rho) = \{n_1+n_2+n_3-(n_1^{(k)}-n_4^{(k)})/Q_k(\rho)\}/N$ are random, as they depend on the random quantities $n_1^{(k)}$. To emphasize this, we define $\hat{q}_k = n_1^{(k)}/n_1$ and $\hat{q} = (\hat{q}_1, \ldots, \hat{q}_K)^T$, and write

$$\mu_k^{(N)}(\rho,\hat{q}) = \{n_1+n_2+n_3-(n_1\hat{q}_k-n_4^{(k)})/Q_k(\rho)\}/N$$

and

$$\ell^*(\phi,\hat{q}) = \sum_A \log f(y|x,\theta) - \sum_B \log\left[\sum_{k=1}^{K}\mu_k^{(N)}(\rho,\hat{q})Q_k(x,\theta)\right] + \sum_{k=1}^{K}(n_1\hat{q}_k - n_4^{(k)})\log Q_k(\rho).$$

Let $\mathbf{J}^* = \text{plim}_{N\to\infty} -N^{-1}\frac{\partial^2\ell^*(\phi,\hat{q})}{\partial\rho\partial\hat{q}^T}$, where here and subsequently, all derivatives are evaluated at $\phi=\phi_0$ and $\hat{q}=Q_0$. By expanding $\frac{\partial l^*(\phi,\hat{q})}{\partial\phi}$ about (ϕ_0, Q_0), and using the arguments of Wild (1991), we see that the asymptotic variance of $\hat{\phi}$ is $(\mathbf{I}^*)^{-1}\mathbf{V}(\mathbf{I}^*)^{-1}$, where $\mathbf{V}=\mathbf{V}_1+\mathbf{V}_2$, with

$$\mathbf{V}_1 = \lim_{N\to\infty} N^{-1}\text{Var}\left(\frac{\partial l^*(\phi,\hat{q})}{\partial\phi}\right),$$

and

$$\mathbf{V}_2 = N\mathbf{J}^*\text{Var}(\hat{q})(\mathbf{J}^*)^T.$$

To obtain more explicit versions of these expressions, we first note that, using arguments similar to those used for $\mathbf{I}^*$, we get

$$\text{plim}_{N\to\infty} -N^{-1}\frac{\partial^2\ell^*(\phi,\hat{q})}{\partial\theta\partial\hat{q}_k} = -w_1E_{k\theta},$$

and

$$\text{plim}_{N\to\infty} -N^{-1}\frac{\partial^2\ell^*(\theta,\hat{q}_k)}{\partial\rho\partial\hat{q}_k} = -w_1E_{k\rho}. \tag{31}$$

Next, we evaluate $\mathbf{V}_1$. Using the same partitioning arguments as above, we can write

$$\begin{aligned}\mathbf{V}_1 &= w_2 E\left[\frac{\partial \log f(y|x,\theta)g(x,\theta,\rho)}{\partial\phi}\frac{\partial \log f(y|x,\theta)g(x,\theta,\rho)}{\partial\phi^T}\right]\\ &\quad + w_3 E\left[\frac{\partial \log g(x,\theta,\rho)}{\partial\phi}\frac{\partial \log g(x,\theta,\rho)}{\partial\phi^T}\right]\\ &\quad + \sum_{k=1}^{K} w_4^{(k)} E_k\left[\frac{\partial \log f(y|x,\theta)g(x,\theta,\rho)}{\partial\phi}\frac{\partial \log f(y|x,\theta)g(x,\theta,\rho)}{\partial\phi^T}\right]\\ &\quad - \sum_{k=1}^{K} w_4^{(k)} E_k\left[\frac{\partial \log f(y|x,\theta)g(x,\theta,\rho)}{\partial\phi}\right]\\ &\quad \times E_k\left[\frac{\partial \log f(y|x,\theta)g(x,\theta,\rho)}{\partial\phi^T}\right]. \end{aligned} \tag{32}$$

Using the result (28), this implies that

$$\begin{aligned}\mathbf{V}_1 &= \mathbf{I}^* + \int \frac{\partial^2}{\partial\phi\partial\phi^T}\left\{\frac{\sum_{k=1}^{K}\mu_{k0}Q_k(x,\theta)}{\sum_{k=1}^{K}\mu_k(\rho)Q_k(x,\theta)}\right\} Q^*(x)g_0(x)\,\mathrm{d}x\\ &\quad + \sum_{k=1}^{K} c_k \frac{1}{Q_k}\frac{\partial^2 Q_k}{\partial\phi\phi^T} - \sum_{k=1}^{K} w_4^{(k)} E_k\left[\frac{\partial \log f(y|x,\theta)g(x,\theta,\rho)}{\partial\phi}\right]\\ &\quad \times E_k\left[\frac{\partial \log f(y|x,\theta)g(x,\theta,\rho)}{\partial\phi^T}\right]. \end{aligned} \tag{33}$$

Moreover,

$$E_k\left[\frac{\log f(y|x,\theta)g(x,\theta,\rho)}{\partial\theta}\right] = E_{k\theta}$$

and

$$E_k\left[\frac{\log g(x,\theta,\rho)}{\partial\rho}\right] = E_{k\rho}.$$

Now, for the $\theta\theta$ block, the derivative under the integral sign in (33) is zero, so, using the fact that $n_1\mathrm{Cov}(\hat{q}) \to \mathrm{diag}(Q_0) - Q_0 Q_0^T$, we see that the $\theta\theta$ block of $\mathbf{V} = \mathbf{V}_1 + \mathbf{V}_2$ is given by

$$\begin{aligned}\mathbf{V}_{\theta\theta} &= \mathbf{I}^*_{\theta\theta} - \sum_{k=1}^{K} w_4^{(k)} E_{k\theta} E_{k\theta}^T + w_1 \sum_{k=1}^{K} Q_{k0} E_{k\theta} E_{k\theta}^T - w_1 \sum_{k=1}^{K}\sum_{l=1}^{K} Q_{k0} Q_{l0} E_{k\theta} E_{l\theta}^T\\ &= \mathbf{I}^*_{\theta\theta} - \sum_{k=1}^{K}\sum_{l=1}^{K} d_{kl} E_{k\theta} E_{l\theta}^T \end{aligned} \tag{34}$$

where $d_{kl} = w_1 Q_{k0} Q_{l0} - \delta_{kl} c_k$. We can rewrite (29) as $\mathbf{I}^*_{\rho\theta} = \mathbf{A}\mathbf{E}^T_\theta$, where $\mathbf{E}_\theta$ has columns $E_{1,\theta} \ldots, E_{k,\theta}$, and $\mathbf{A}$ has l, k element $w_4^{(k)} \frac{\partial \log \mu_k}{\partial \rho_l}$. Thus, there is a generalised inverse $\mathbf{A}^-$ with $\mathbf{E}^T_\theta = \mathbf{A}^- \mathbf{I}^*_{\theta\theta}$, so that

$$\mathbf{V}_{\theta\theta} = \mathbf{I}^*_{\theta\theta} - \mathbf{I}^*_{\theta\rho}(\mathbf{A}^-)^T \mathbf{D}\mathbf{A}^- \mathbf{I}^*_{\rho\theta},$$

where $\mathbf{D}$ is the matrix with elements d_{kl}.

Also, for the $\rho\theta$ block, the integral in (33) is equal to

$$-\sum_{k=1}^{K} w_4^{(k)} \frac{\partial \log \mu_k(\rho)}{\partial \rho} E_{k\theta}$$

so that

$$\mathbf{V}_{\rho\theta} = \mathbf{I}^*_{\rho\theta} - \sum_{k=1}^{K} w_4^{(k)} \frac{\partial \log \mu_k(\rho)}{\partial \rho^T} E^T_{k\theta} - \sum_{k=1}^{K}\sum_{l=1}^{K} d_{kl} E_{k\rho} E^T_{l\theta} \tag{35}$$

Since

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{l=1}^{K} d_{kl} \frac{\partial \log Q_k(\rho)}{\partial \rho} E^T_{k\theta} \\
&= w_1 \left(\sum_{k=1}^{K} Q_{k0} \frac{\partial \log Q_k(\rho)}{\partial \rho}\right)\left(\sum_{k=1}^{K} Q_{k0} E^T_{k\theta}\right)^T - \sum_{k=1}^{K} c_k \frac{\partial \log Q_k(\rho)}{\partial \rho} E_{k\theta} \\
&= -\sum_{k=1}^{K} w_4^{(k)} \frac{\partial \log \mu_k(\rho)}{\partial \rho} E^T_{k\theta},
\end{aligned}$$

we can write (35) as

$$\mathbf{V}_{\rho\theta} = \mathbf{I}^*_{\rho\theta} - \sum_{k=1}^{K} d_{kl} \left(E_{k\rho} - \frac{\partial \log Q_k(\rho)}{\partial \rho}\right) E^T_{k\theta}. \tag{36}$$

Using (30), we can write

$$\mathbf{V}_{\rho\theta} = \mathbf{I}^*_{\rho\theta} - \mathbf{I}^*_{\rho\rho}(\mathbf{A}^-)^T \mathbf{D}\mathbf{A}^- \mathbf{I}^*_{\rho\theta}.$$

Similarly, we obtain

$$\mathbf{V}_{\rho\rho} = \mathbf{I}^*_{\rho\rho} - \mathbf{I}^*_{\rho\rho}(\mathbf{A}^-)^T \mathbf{D}\mathbf{A}^- \mathbf{I}^*_{\rho\rho}$$

and hence

$$\mathbf{V} = \mathbf{I}^* - \mathbf{I}^* \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\mathbf{A}^-)^T \mathbf{D} \mathbf{A}^- \end{pmatrix} \mathbf{I}^*.$$

The asymptotic variance is

$$(\mathbf{I}^*)^{-1} \mathbf{V} (\mathbf{I}^*)^{-1} = (\mathbf{I}^*)^{-1} - \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\mathbf{A}^-)^T \mathbf{D} \mathbf{A}^- \end{pmatrix}.$$

(*ii*) Using the partitioned matrix inverse formula, the asymptotic covariance matrix of $\hat{\theta}$ can be written as

$$\mathrm{Avar}(\hat{\theta}) = (\mathbf{I}^*_{\theta\theta} - \mathbf{I}^*_{\theta\rho} (\mathbf{I}^*_{\rho\rho})^{-1} \mathbf{I}^*_{\rho\theta})^{-1}. \tag{37}$$

(*iii*) First, note that, for all θ,

$$\left. \frac{\partial l^*(\theta, \rho)}{\partial \rho} \right|_{\theta = \rho_\theta} = 0,$$

so that, differentiating with respect to θ, we get

$$0 = \left. \frac{\partial^2 l^*(\theta, \rho)}{\partial \rho \partial \theta^T} \right|_{\theta = \rho_\theta} + \left. \frac{\partial^2 l^*(\theta, \rho)}{\partial \rho \partial \rho^T} \right|_{\theta = \rho_\theta} \frac{\partial \rho_\theta}{\partial \theta}$$

and hence

$$\frac{\partial \rho_\theta}{\partial \theta} = - \left(\left. \frac{\partial^2 l^*(\theta, \rho)}{\partial \rho \partial \rho^T} \right|_{\theta = \rho_\theta} \right)^{-1} \left. \frac{\partial^2 l^*(\theta, \rho)}{\partial \rho \partial \theta^T} \right|_{\theta = \rho_\theta}.$$

Thus

$$\begin{aligned} \left. \frac{\partial^2 l^*(\theta, \rho_\theta)}{\partial \theta \partial \theta^T} \right|_{\theta = \rho_\theta} &= \frac{\partial}{\partial \theta} \left[\left. \frac{\partial l^*(\theta, \rho)}{\partial \theta} \right|_{\theta = \rho_\theta} + \left. \frac{\partial l^*(\theta, \rho)}{\partial \rho} \right|_{\theta = \rho_\theta} \frac{\partial \rho_\theta}{\partial \theta} \right] \\ &= \frac{\partial}{\partial \theta} \left[\left. \frac{\partial l^*(\theta, \rho)}{\partial \theta} \right|_{\theta = \rho_\theta} \right] \\ &= \left. \frac{\partial^2 l^*(\theta, \rho)}{\partial \theta \partial \theta^T} \right|_{\theta = \rho_\theta} + \left. \frac{\partial^2 l^*(\theta, \rho)}{\partial \theta \partial \rho^T} \right|_{\theta = \rho_\theta} \frac{\partial \rho_\theta}{\partial \theta} = \left. \frac{\partial^2 l^*(\theta, \rho)}{\partial \theta \partial \theta^T} \right|_{\theta = \rho_\theta} \\ &\quad - \left. \frac{\partial^2 l^*(\theta, \rho)}{\partial \theta \partial \rho^T} \right|_{\theta = \rho_\theta} \left(\left. \frac{\partial^2 l^*(\theta, \rho)}{\partial \rho \partial \rho^T} \right|_{\theta = \rho_\theta} \right)^{-1} \left. \frac{\partial^2 l^*(\theta, \rho)}{\partial \rho \partial \theta^T} \right|_{\theta = \rho_\theta}. \end{aligned}$$

Dividing by N and letting $N \to \infty$ gives (7).

6.2 Proof of Theorem 2

We first show that

$$\left(\frac{\partial \log p_1(x,\theta,\hat{\eta}(\theta))}{\partial\theta}\bigg|_{\theta=\theta_0}, \ldots, \frac{\partial \log p_J(x,\theta,\hat{\eta}(\theta))}{\partial\theta}\bigg|_{\theta=\theta_0} \right) \tag{38}$$

is orthogonal to the nuisance tangent space $\mathscr{T}_\eta$, the subspace of $\mathscr{H}$ defined in Sect. 3.1.

Consider a finite-dimensional submodel $\mathscr{Q}$ of $\mathscr{P}$ of the form

$$\mathscr{Q} = \{p_j(x,\theta,\gamma(t)), \theta \in \mathscr{B}, t \in \mathscr{T}\},$$

where $\gamma(0) = \eta_0$, and define

$$\hat{\eta}(\theta,t) = \text{argmax}_\eta \sum_{j=1}^{J} w_j E_{j,t}[\log p_j(X,\theta,\eta)]$$

where $E_{j,t}$ denotes expectation with respect to $p_j(x,\theta,\gamma(t))$. Then

$$\sum_{j=1}^{J} w_j E_j[\log p_j(X,\theta,\hat{\eta}(\theta,t))]$$

is maximised at $t = 0$, since

$$\sum_{j=1}^{J} w_j E_j[\log p_j(X,\theta,\hat{\eta}(\theta,t))] \le \sum_{j=1}^{J} w_j E_j[\log p_j(X,\theta,\hat{\eta}(\theta))]$$

and $\hat{\eta}(\theta,0) = \hat{\eta}(\theta)$. Hence for every θ,

$$\frac{\partial}{\partial t}\sum_{j=1}^{J} w_j E_j[\log p_j(X,\theta,\hat{\eta}(\theta,t))]\Big|_{t=0} = 0. \tag{39}$$

Differentiating (39) with respect to θ gives

$$\sum_{j=1}^{J} w_j \int \frac{\partial^2 \log p_j(X,\theta,\hat{\eta}(\theta,t))}{\partial\theta\,\partial t}\bigg|_{t=0} p_j(x,\theta_0,\eta_0)\,\mathrm{d}x = 0. \tag{40}$$

Also, differentiating both sides of the identity

$$\sum_{j=1}^{J} w_j \int \frac{\partial \log p_j(X,\theta,\hat{\eta}(\theta,t))}{\partial\theta}\bigg|_{t=0} p_j(x,\theta_0,\hat{\eta}(\theta,t))\,\mathrm{d}x = 0 \tag{41}$$

with respect to t, we get

$$\sum_{j=1}^{J} w_j \int \frac{\partial^2 \log p_j(X,\theta,\hat{\eta}(\theta,t))}{\partial\theta\partial t} p_j(x,\theta_0,\hat{\eta}(\theta_0,t))\,\mathrm{d}x + \sum_{j=1}^{J} w_j \int \frac{\partial \log p_j(X,\theta,\hat{\eta}(\theta,t))}{\partial\theta} \frac{\partial \log p_j(X,\theta,\hat{\eta}(\theta,t))}{\partial t} \times p_j(x,\theta_0,\hat{\eta}(\theta_0,t))\,\mathrm{d}x = 0$$

Setting $\theta = \theta_0,\ t = 0$ and using (40), we get

$$\sum_{j=1}^{J} w_j \int \left.\frac{\partial \log p_j(X,\theta,\hat{\eta}(\theta))}{\partial\theta}\right|_{\theta=\theta_0} \left.\frac{\partial \log p_j(X,\theta,\hat{\eta}(\theta,t))}{\partial t}\right|_{\substack{t=0\\ \theta=\theta_0}} \times p_j(x,\theta_0,\eta_0)\,\mathrm{d}x = 0 \tag{42}$$

so that (38) is orthogonal to

$$\left(\left.\frac{\partial \log p_1(x,\theta_0,\hat{\eta}(\theta_0,t))}{\partial t}\right|_{t=0}, \ldots, \left.\frac{\partial \log p_J(x,\theta_0,\hat{\eta}(\theta_0,t))}{\partial t}\right|_{t=0} \right). \tag{43}$$

But $\hat{\eta}(\theta_0,t) = \gamma(t)$ by the Kullback-Leibler information equality, so that (43) is in fact the score function corresponding to the nuisance parameter $\gamma(t)$. Thus (38) is in the nuisance tangent space of $\mathscr{Q}$, and since $\mathscr{Q}$ was an arbitrary finite-dimensional subfamily of $\mathscr{P}$, (38) must lie in the the nuisance tangent space of $\mathscr{P}$.

Now consider the subfamily of $\mathscr{P}$

$$\{p_j(x,\theta,\hat{\eta}(\theta)), \theta \in \mathscr{B}\}.$$

By the chain rule,

$$\begin{aligned}
\left.\frac{\partial \log p_j(x,\theta,\hat{\eta}(\theta))}{\partial\theta}\right|_{\theta=\theta_0} &= \left.\frac{\partial \log p_j(x,\theta,\hat{\eta}(\theta'))}{\partial\theta}\right|_{\substack{\theta=\theta_0\\ \theta'=\theta_0}} \\
&\quad + \left.\frac{\partial \log p_j(x,\theta,\hat{\eta}(\theta'))}{\partial\theta'}\right|_{\substack{\theta=\theta_0\\ \theta'=\theta_0}} \times \left.\frac{\partial\theta'}{\partial\theta}\right|_{\theta=\theta_0} \\
&= \left.\frac{\partial \log p_j(x,\theta,\eta_0)}{\partial\theta}\right|_{\theta=\theta_0} + h_j \\
&= \dot{l}_{j\theta} + h_j,
\end{aligned}$$

say, where h_j is in the nuisance tangent space. Thus

$$\dot{l}_{j\theta} = h_j + \left.\frac{\partial \log p_j(x, \theta, \hat{\eta}(\theta))}{\partial \theta}\right|_{\theta=\theta_0},$$

so $\dot{l}_{j\theta}$ can be expressed as the sum of an element in the nuisance tangent space plus an element orthogonal to the nuisance tangent space. It follows that (38) is the projection of $\dot{l}_{j\theta}$ onto the orthogonal complement of the nuisance tangent space and so is the efficient score.

6.3 Proof that (14) is the maximizer of (11)

As in Sect. 6.1, define

$$g(x, \theta, \rho) = \frac{Q^*(x)g_0(x)}{\sum_{k=1}^K \mu_k(\rho)Q_k(x, \theta)},$$

where $\mu_k(\rho) = w_1 + w_2 + w_3 - c_k/Q_k(\rho)$. We will show that the function g that maximises (11) is given by $g(x) = g(x, \theta, \rho_\theta)$ where ρ_θ is the solution to the $K - 1$ equations

$$Q_k(\rho) = \int Q_k(x, \theta)g(x, \theta, \rho)\,\mathrm{d}x, \quad k = 1, \ldots, K-1. \tag{44}$$

Note that these equations imply that $Q_K(\rho) = \int Q_K(x, \theta)g(x, \theta, \rho)\,\mathrm{d}x$ and that $g(x, \theta, \rho_\theta)$ is a density, at least in a neighbourhood of θ_0. Let $\tilde{g}$ be an arbitrary density, and write $\tilde{Q}_k(\theta) = \int Q_k(x, \theta)\tilde{g}(x)\,\mathrm{d}x$. We must show that for all θ and $\tilde{g}$,

$$\begin{aligned}&\int \log g(x, \theta, \rho_\theta)Q^*(x)g_0(x)\,\mathrm{d}x + \sum_{k=1}^K c_k \log Q_k(\rho_\theta)\\ &\quad\geq \int \log \tilde{g}(x)Q^*(x)g_0(x)\,\mathrm{d}x + \sum_{k=1}^K c_k \log \tilde{Q}_k(\theta),\end{aligned} \tag{45}$$

or, equivalently, that

$$\int \log\left\{\frac{g(x, \theta, \rho_\theta)}{\tilde{g}(x)}\right\} Q^*(x)g_0(x)\,\mathrm{d}x \geq \sum_{k=1}^K c_k \log\left\{\frac{\tilde{Q}_k(\theta)}{Q_k(\rho_\theta)}\right\}. \tag{46}$$

To prove (46), we set $h_k(x, \theta) = Q_k(x, \theta)\tilde{g}(x)/\tilde{Q}_k(\theta)$, so h_k is a density. Also define

$$H_k(x, \theta) = Q^*(x)g_0(x)P_k^*(x, \theta, \rho_\theta)/(\mu_k(\rho_\theta)Q_k(\rho_\theta)),$$

where

$$P_k^*(x, \theta, \rho) = \frac{\mu_k(\rho)Q_k(x, \theta)}{\sum_{k=1}^K \mu_k(\rho)Q_k(x, \theta)}.$$

The function H_k is also a density for every θ by (44).

The left hand side of (46) can be written as

$$\begin{aligned}
&\int \log\left\{\frac{Q^*(x)g_0(x)P_k^*(x,\theta,\rho_\theta)}{\tilde{Q}_k(\theta)\mu_k(\rho_\theta)h_k(x,\theta)}\right\} Q^*(x)g_0(x)\,\mathrm{d}x \\
&= \int \log\left\{\frac{H_k(x,\theta)}{h_k(x,\theta)}\right\} Q^*(x)g_0(x)\,\mathrm{d}x + (1-w_1)\log\left\{\frac{Q_k(\rho_\theta)}{\tilde{Q}_k(\theta)}\right\} \\
&\geq \mu_k(\rho_\theta)Q_k(\rho_\theta)\int \log\left\{\frac{H_k(x,\theta)}{h_k(x)}\right\} H_k(x,\theta)\,\mathrm{d}x \\
&\quad +(1-w_1)\log\left\{\frac{Q_k(\rho_\theta)}{\tilde{Q}_k(\theta)}\right\}. \qquad (47)
\end{aligned}$$

The last inequality follows because $1 \geq P_k^*(x,\theta,\rho_\theta)$. The integral in (47) is non-negative by the Kullback-Leibler information inequality, so, for each k, we have

$$\int \log\left\{\frac{g(x,\theta,\rho_\theta)}{\tilde{g}(x)}\right\} Q^*(x)g_0(x)\,\mathrm{d}x \geq (1-w_1)\log\left\{\frac{Q_k(\rho_\theta)}{\tilde{Q}_k(\theta)}\right\}.$$

Also, the fact that $0 < \mu_k(\rho_\theta)Q_k(\rho_\theta)$ in a neigbourhood of θ_0 implies that

$$w_4^{(k)} - w_1 Q_{k0} + (w_1+w_2+w_3)Q_k(\rho_\theta) > 0,$$

so multiplying by $\{w_4^{(k)} - w_1 Q_{k0} + (w_1+w_2+w_3)Q_k(\rho_\theta)\}/(1-w_1) > 0$ and summing gives

$$\begin{aligned}
&\int \log\left\{\frac{g(x,\theta,\rho_\theta)}{g(x)}\right\} Q^*(x)g_0(x)\,\mathrm{d}x \\
&\geq \sum_{k=1}^{K}\left\{w_4^{(k)} - w_1 Q_{k0} + (w_1+w_2+w_3)Q_k(\rho_\theta)\right\}\log\frac{Q_k(\rho_\theta)}{\tilde{Q}_k(\theta)} \\
&\geq \sum_{k=1}^{K}(w_4^{(k)} - w_1 Q_{k0})\log\frac{Q_k(\rho_\theta)}{\tilde{Q}_k(\theta)} \\
&= \sum_{k=1}^{K} c_k \log\left\{\frac{\tilde{Q}_k(\theta)}{Q_k(\rho_\theta)}\right\}
\end{aligned}$$

since

$$(w_1+w_2+w_3)\sum_{k=1}^{K} Q_k(\rho_\theta)\log\frac{Q_k(\rho_\theta)}{\tilde{Q}_k(\theta)} \geq 0$$

by the Kullback-Leibler inequality. This implies (45).

6.4 Proof of (21)–(24)

Evaluating the integral in (28), we get

$$\begin{aligned}
\mathbf{I}^{\dagger}_{\theta\theta} &= \mathbf{I}^{*}_{\theta\theta}, \\
\mathbf{I}^{\dagger}_{\rho\theta} &= \mathbf{I}^{*}_{\rho\theta} - \sum_{k=1}^{K} w_4^{(k)} \frac{\partial \log \mu_k(\rho)}{\partial \rho} E_{k\theta}^{T}, \\
\mathbf{I}^{\dagger}_{\rho\rho} &= \mathbf{I}^{*}_{\rho\rho} - 2\sum_{k=1}^{K} w_4^{(k)} \frac{\partial \log \mu_k(\rho)}{\partial \rho} \left(E_{k\rho} - \frac{\partial \log Q_k(\rho)}{\partial \rho} \right)^{T}.
\end{aligned}$$

These results, together with Equations (29) and (30) imply (21)–(23). For (24), note that by (44) we have

$$Q_k(\rho_\theta) = \int Q_k(x,\theta) g(x,\theta,\rho_\theta)\, \mathrm{d}x.$$

Differentiating both sides with respect to θ, and setting $\theta = \theta_0$ we get, after some algebra,

$$\frac{\partial \log Q_k(\rho)}{\partial \rho} \frac{\partial \rho_\theta}{\partial \theta} = E_{k\theta}^{T} + E_{k\rho}^{T} \frac{\partial \rho_\theta}{\partial \theta}.$$

Multiplying both sides by $w_4^{(k)} \frac{\partial \log \mu_k(\rho)}{\partial \rho}$ and summing gives

$$\mathbf{I}^{*}_{\rho\theta} + \mathbf{I}^{*}_{\rho\rho} \frac{\partial \rho_\theta}{\partial \theta} = 0$$

which proves (24).

6.5 Formulae for asymptotic variances

In this section we derive the formulae used to calculate the asymptotic variances in Sect. 4. We use the double index notation for the strata introduced there. In our example, we used a discrete variate x_1 obtained by discretizing a standard normal that was correlated with correlation η with the second variate x_2, also a standard normal. In the formulae below, we require the joint density of x_1 and x_2, so we first record this.

We assume that the discrete variate x_1 has J categories, obtained by dividing the range $(-\infty, \infty)$ into J intervals $(x_{1,j-1}, x_{1,j})$, where $-\infty = x_{1,0} < x_{1,0} < \cdots, x_{1,J} = \infty$. Then $x_1 = x_1^{(j)}$ if the normal variate correlated with x_2 falls in the interval $(x_{1,j-1}, x_{1,j})$. The joint density of x_1 and x_2 is $Pr[x_1 = x^{(j)}, x < x_2 \leq x + \mathrm{d}x] = g_j(x)\,\mathrm{d}x$ where

$$g_j(x) = \phi(x) \left\{ \Phi\left[(x_{1,j} - \eta x)/\sqrt{1-\eta^2} \right] - \Phi\left[(x_{1,j-1} - \eta x)/\sqrt{1-\eta^2} \right] \right\},$$

and ϕ and Φ are the density and distribution function of the standard normal.

The semi-parametric estimator

In the logistic case, the estimating equation for θ becomes

$$S(\theta, \hat{\mu}) = 0,$$

where $\hat{\mu}$ is the estimate of $\mu = (\mu_{01}, \ldots, \mu_{1,J})^T$, derived from the "$\rho$" part of the estimating equation (5), and

$$S(\theta, \mu) = \sum_{i=0}^{1} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} x_{ijk}(y_{ijk} - P_{1j}(x_{ijk}, \theta, \mu))$$

with logit $P_{1j}(x_{ijk}, \theta, \mu) = \log(\mu_{1j}/\mu_{0j}) + \theta^T x$. This follows from the general definition of the P's in Sect. 6.1 and the expression (26) for $Q_{ij}(x)$. Note also that $P_{1j}(x_{ij'k}) = 0$ unless $j \neq j'$.

From Sect. 2, we have

$$\text{Avar}(\hat{\theta}) = (\mathbf{I}^*_{\theta\theta} - \mathbf{I}^*_{\theta\rho}\mathbf{I}^{*-1}_{\rho\rho}\mathbf{I}^*_{\rho\theta})^{-1}.$$

Scott and Wild (1997) show that

$$\mathbf{I}^*_{\theta\rho}\mathbf{I}^{*-1}_{\rho\rho}\mathbf{I}^*_{\rho\theta} = \mathbf{B}^T(\mathbf{W} - \mathbf{A}^{-1})\mathbf{B}$$

where simple computational forms for the matrices **A**, **B**, and **W** are given below. A simple formula for $\mathbf{I}^*_{\theta\theta}$ is

$$\begin{aligned} \mathbf{I}^*_{\theta\theta} &= \text{plim}_{N\to\infty} \frac{1}{N} \sum_{i=0}^{1} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} \frac{\partial x_{ijk}(y_{ijk} - P_{1j}(x_{ijk}, \theta, \mu))}{\partial \theta} \\ &= \sum_{i=0}^{1} \sum_{j=1}^{J} w_{ij} E_{ij}[xx^T P_{0j} P_{1j}] \end{aligned} \tag{48}$$

where E_{ij} denotes expectation conditional on being in stratum $\mathscr{S}_{ij}$ as in Sect. 3.2. Since

$$E_{ij}[xx^T P_{0j} P_{1j}] = \frac{1}{Q_{ij}} \int xx^T f(y = i|x) P_{0j}(x) P_{1j}(x) g_j(x_2)\, \mathrm{d}x_2$$

where $x = (1, x^{(j)}, x_2)^T$, (48) reduces to

$$\int xx^T P_{0j}(x) P_{1j}(x) Q^*(x) \phi(x)\, \mathrm{d}x_2.$$

Define $\mathbf{P}(x) = (P_{01}(x), \ldots, P_{0J}(x), P_{11}(x), \ldots, P_{1,J-1}(x))$. The matrix

$$\mathbf{B} = \text{plim}_{N\to\infty} N^{-1} \sum_{i=0}^{1} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} \frac{\partial \mathbf{P}(x_{ijk})}{\partial \theta}$$

has $2J-1$ rows; the row corresponding to P_{ij} is

$$\begin{aligned}\text{plim}_{N\to\infty} \frac{1}{N} \sum_{i'=0}^{1} \sum_{j'=1}^{J} \sum_{k=1}^{n_{ij}} \frac{\partial P_{ij}(x_{i'j'k})}{\partial \theta} &= \epsilon_i (w_{0j} E_{0j}[x P_{0j} P_{1j}] + w_{1j} E_{1j}[x P_{0j} P_{1j}])^T \\ &= \epsilon_i b_j^T,\end{aligned}$$

where $\epsilon_i = -1$ if $i = 0$ and 1 if $i = 1$, and

$$b_j = \int x P_{0j}(x) P_{1j}(x) Q^*(x) g_j(x_2)\, \mathrm{d}x_2.$$

The matrix

$$\mathbf{W} = \text{plim}_{N\to\infty} N^{-1} \sum_{i=0}^{1} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} \text{diag}(\mathbf{P}(x_{ijk})) - \mathbf{P}(x_{ijk}) \mathbf{P}(x_{ijk})^T$$

has $2J-1$ rows and columns, with diagonal elements

$$\begin{aligned}&\text{plim}_{N\to\infty} \frac{1}{N} \sum_{i'=0}^{1} \sum_{j'=1}^{J} \sum_{k=1}^{n_{ij}} P_{ij}(x_{i'j'k})(1 - P_{ij}(x_{i'j'k})) \\ &= w_{0j} E_{0j}[P_{0j} P_{1j}] + w_{1j} E_{1j}[P_{0j} P_{1j}] \\ &= \int P_{0j}(x) P_{1j}(x) Q^*(x) g_j(x_2)\, \mathrm{d}x_2\end{aligned}$$

and off-diagonal elements zero, except for those corresponding to row P_{0j} and column P_{1j} for j=1, …, J−1. In this case, the element is $-\int P_{0j}(x) P_{1j}(x) Q^*(x) g_j(x_2) \mathrm{d}x_2$. Thus,

$$W = \begin{bmatrix} \text{diag}(d_1, \ldots, d_J) & -\text{diag}(d_1, \ldots, d_{J-1}) \\ -\text{diag}(d_1, \ldots, d_{J-1}) & \text{diag}(d_1, \ldots, d_{J-1}) \end{bmatrix}$$

where $d_j = \int P_{0j}(x) P_{1j}(x) Q^*(x) g_j(x_2)\, \mathrm{d}x_2$.

The conditional estimator

Let $Q = (Q_{01}, \ldots, Q_{1,J})^T$, with corresponding vector $\hat{Q}$ where $\hat{Q}_{ij} = N_{ij}/N$. In the logistic case, the conditional estimator $\hat{\theta}_C$ satisfies the estimating equation

$S(\theta, \hat{Q}) = 0$, where

$$S(\theta, Q) = \sum_{i=0}^{1} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} x_{ijk}(y_{ijk} - P_{1j}(x_{ijk}, \theta, \mu))$$

and P_{1j} is now defined by logit $P_{1j}(x_{ijk}, \theta, \mu) = \log(n_{1j}/n_{0j}) + \log(Q_{0j}/Q_{0j}) + \theta^T x$. Using the argument of Breslow and Cain (1988), we expand S about θ and Q and obtain the asymptotic variance as

$$\text{Avar}(\theta_C) = (\mathbf{I}^*{}_{\theta\theta})^{-1} \left[\lim_{N\to\infty} \text{Var}\, S(\theta, Q) + \mathbf{B}^T(\text{diag}(Q) - QQ^T)\mathbf{B} \right] (\mathbf{I}^*{}_{\theta\theta})^{-1},$$

where $\mathbf{I}^*{}_{\theta\theta} = -\text{plim}_{N\to\infty} N^{-1} \frac{\partial S}{\partial \theta}$, and $\mathbf{B} = \text{plim}_{N\to\infty} N^{-1} \frac{\partial S}{\partial Q}$. Using the same arguments as before, we get

$$\mathbf{B}^T(\text{diag}(Q) - QQ^T)\mathbf{B} = \mathbf{B}_0^T \,\text{diag}(Q_{0j}^{-1} + Q_{1j}^{-1})\mathbf{B}_0$$

where $\mathbf{B}_0$ has J columns with jth element $\int x P_{0j}(x) P_{1j}(x) Q^*(x) g_j(x_2)\, \mathrm{d}x_2$.

Also,

$$\lim_{N\to\infty} \text{Var}\, S(\theta, Q) = \mathbf{I}^*{}_{\theta\theta} - \mathbf{B}_0^T \,\text{diag}(w_{0j}^{-1} + w_{1j}^{-1})\mathbf{B}_0$$

so that

$$\text{Avar}(\hat{\theta}_C) = \mathbf{I}^{*-1}_{\theta\theta} - \mathbf{I}^{*-1}_{\theta\theta} \mathbf{B}_0^T \mathbf{D} \mathbf{B}_0 \mathbf{I}^{*-1}_{\theta\theta}$$

where $\mathbf{D}$ is diagonal with elements $w_{0j}^{-1} + w_{1j}^{-1} - Q_{0j}^{-1} - Q_{1j}^{-1}$.

The weighted estimator

The weighted estimator $\hat{\theta}_W$ satisfies the estimating equation $S(\theta, \hat{Q}) = 0$, where in this case

$$S(\theta, Q) = \sum_{i=0}^{1} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} \frac{N Q_{ij}}{n_{ij}} x_{ijk}(y_{ijk} - f(1|x_{ijk}, \theta)).$$

We apply the Breslow-Cain argument again, with

$$\mathbf{I}^*{}_{\theta\theta} = \int x x^T f(0|x) f(1|x) \phi(x_2)\, \mathrm{d}x_2,$$

and

$$\mathbf{B}^T(\text{diag}(Q) - QQ^T)\mathbf{B} = \mathbf{B}_0^T \,\text{diag}(Q_{0j}^{-1} + Q_{1j}^{-1})\mathbf{B}_0,$$

where now $\mathbf{B}_0$ has jth row $\int x f(0|x) f(1|x) g_j(x_2)\, \mathrm{d}x_2$, and

$$\lim_{N\to\infty} \mathrm{Var}\,(S(\theta, Q) = \mathbf{I}^*_{\theta\theta} + \sum_{j=1}^{J} \int xx^T f(0|x) f(1|x) f_j^*(x) g_j(x_2)\,\mathrm{d}x_2$$
$$-\mathbf{B}_0^T \,\mathrm{diag}(w_{0j}^{-1} + w_{1j}^{-1})\mathbf{B}_0,$$

where

$$f_j^*(x) = \left(\frac{Q_{1j}}{w_{1j}} - 1\right) f(0|x) + \left(\frac{Q_{0j}}{w_{0j}} - 1\right) f(1|x).$$

Thus

$$\mathrm{Avar}(\hat{\theta}_W) = \mathbf{I}^{*-1}_{\theta\theta} - \mathbf{I}^{*-1}_{\theta\theta}(\mathbf{B}_0^T \mathbf{D}\mathbf{B}_0 - \mathbf{G})\mathbf{I}^{*-1}_{\theta\theta}$$

where $\mathbf{D}$ is as for the conditional estimate, and $\mathbf{G} = \sum_{j=1}^{J} \int xx^T f(0|x) f(1|x)$ $f^*(x) g_j(x_2)\,\mathrm{d}x_2$.

References

Bickel, P. J., Klaassen, C. A., Ritov, Y., Wellner, J. A. (1993). *Efficient and adaptive estimation for semi-parametric models.* Baltimore: Johns Hopkins University Press.

Breslow, N. E., Cain, K. C. (1988). Logistic regression for two-stage case-control data. *Biometrika*, *75*, 11–20.

Breslow, N. E., Robins, J. M., Wellner, J. A. (2000). On the semi-parametric efficiency of logistic regression under case-control sampling. *Bernoulli*, *6*, 447–455.

Breslow, N. E., McNeney, B., Wellner, J. A. (2003). Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. *Annals of Statistics*, *31*, 1110–1139.

Hsieh, D. A., Manski, C. F., McFadden, D. (1985). Estimation of response probabilities from augmented retrospective observations. *Journal of the American Statistical Association*, *80*, 651–662.

Hu, X. J., Lawless, J. F. (1996). Estimation from truncated lifetime data with supplementary information on covariates and and censoring times. *Biometrika*, *83*, 747–761.

Jiang, Y., Scott, A. J., Wild, C. J. (2006). Secondary analyses of case-control sampled data. *Statistics in Medicine*, *25*, 1323–1339.

Kalbfleisch, J. D., Lawless, J. F. (1988). Likelihood analysis of multi-state models for disease incidence and mortality. *Statistics in Medicine*, *7*, 149–160.

Lawless, J. F., Kalbfleisch, J. D., Wild, C. J. (1999). Semiparametric methods for response-selective and missing data problems. *Journal of the Royal Statistical Society, Series B*, *61*, 413–438.

Lee, A. (2007a). On the semiparametric efficiency of the Scott–Wild estimator under choice-based and two-phase sampling. *Journal of applied mathematics and decision sciences*, Article ID 86180, vol. 2007.

Lee, A. (2007b). Semi-parametric efficiency bounds for regression models under choice-based sampling. Unpublished manuscript. Available on http://www.stat.auckland.ac.nz/~lee/.

Lee, A. J., McMurchy, L., Scott, A. J. (1997). Re-using data from case-control studies. *Statistics in Medicine*, *16*, 1377–1389.

Lee, A. J., Scott, A. J., Wild, C. J. (2006). Fitting binary regression models with case-augmented samples. *Biometrika*, *93*, 385–397.

Murphy, S. A., Van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association*, *95*, 449–485.

Neuhaus, J., Scott, A. J., Wild, C. J. (2002). The analysis of retrospective family studies. *Biometrika*, *89*, 23–37.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, *62*, 1349–1382.

Prentice, R. L., Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, *66*, 403–11.

Robins, J. M., Rotnitzky, A., Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, *89*, 846–866.

Robins, J. M., Hsieh, F., Newey, W. (1995). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *Journal of the Royal Statistical Society, Series B*, *57*, 409–424.

Scott, A. J., Wild, C. J. (1991). Fitting logistic regression models in stratified case-control studies. *Biometrics*, *47*, 497–510.

Scott, A. J., Wild, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, *84*, 57–71.

Scott, A. J., Wild, C. J. (2001). Maximum likelihood for generalised case-control studies. *Journal of Statistical Planning and Inference*, *96*, 3–27.

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*, 207–219.

White, J. E. (1982). A two-stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology*, *115*, 119–128.

Whittemore, A. S. (1995). Logistic regression of family data from case-control studies. *Biometrika*, *82*, 57–67.

Wild, C. J. (1991). Fitting prospective regression models to case-control data. *Biometrika*, *78*, 705–717.