Semi-parametric efficiency bounds for regression models under generalised case-control sampling: the profile likelihood approach

ALAN LEE* AND YUICHI HIROSE

Department of Statistics, University of Auckland, Auckland, New Zealand, and Victoria University of Wellington, Wellington, New Zealand

We obtain an information bound for estimates of parameters in general regression models where data is collected under a variety of response-selective sampling schemes. The asymptotic variances of the semi-parametric estimates of Scott and Wild (1986, 1997, 2001) are compared to the bound and the estimates are found to be fully efficient.

Keywords and phrases: Semi-parametric efficiency, outcome-dependent sampling, case-control study, profile likelihood, tangent space, influence function, efficient score, information bound.

Running Head: Semi-parametric efficiency bounds

June 13, 2007

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, 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 (1986, 1997, 2001) and 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 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, Robins and Wellner (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, McNeney and Wellner (2003) apply the missing value theory of Robins, Rotnitzky and Zhao (1994) and Robins, Hsieh and Newey (1995). 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 use an adaptation of the the profile likelihood method due to Newey (1994) to derive a semi-parametric efficiency bound, and then show that this bound coincides with the asymptotic variance of the Scott-Wild estimator, hence demonstrating the efficiency of the estimator.

The paper is structured as follows. In Section 2, we describe the Scott-Wild approach in more detail, discuss some special cases, and discuss the asymptotic variance of the Scott-Wild estimator. In Section 3, we sketch the theory of semiparametric efficiency that we require, and present an extension of Newey's (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. Some further comments on special cases are made in Section 5, and proofs and other derivations are in Section 6.

2. The Scott-Wild approach to generalized 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 that the stratum membership is completely determined by an individual's response and covariate vector, (although it typically depends on the response and only some, perhaps even none, of the covariates.)

Data are gathered according to the following two-phase stratified 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. Note that the density of x and y conditional on being a member of stratum k is

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

where $Q_k = \int\int I_k(x, y) f(y|x, \theta) g(x)\, dx\, dy$, $f(y|x, \theta)$ 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 introduce the notation $Q_k(x, \theta) = \int I_k(x, y) f(y|x, \theta)\, dy$, so that $Q_k = \int Q_k(x, \theta) g(x)\, dx$.

Thus $Q_k(x,\theta)$ is the probability 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. 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.

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 conditional 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. These include

1. *The simple case-control study.* 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, Zhao and Lipsitz, 1992). 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, McMurchy and Scott, 1997, Jiang, Scott and Wild, 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, Scott and Wild, 2006). Here a prospective sample is taken from the joint distribution of (x,y), where y denotes case or control. In

addition, an additional 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, Scott and Wild, 2002). Here the sampling units are families and a binary response is measured on family members. A first stage sample is taken, and the families are assiged to strata on the basis of the binary responses. Second stage sub-samples are taken from the seperate 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.* (Robins *et al.* 1995, Lawless, Kalbfleisch and Wild, 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.

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
   $p_{4,k}(x, y, \theta, g) = I_k(x, y)f(y|x, \theta)g_0(x)/Q_k,\ k = 1, \ldots, K.$

The density g is an infinite-dimensional nuisance parameter. We will also assume that $n_1 Q_{k0} \geq n_4^{(k)}$, corresponding to the fact that $N_n \geq n_k$. 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.

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

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

$\mu_k^{(N)}(\rho) = N^{-1}\{n_1 + n_2 + n_3 - (n_1^{(k)} - n_4^{(k)})/Q_k(\rho)\}$ 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 to the estimating equation

$$\frac{\partial l^*}{\partial \phi} = 0, \quad (5)$$

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" $(\mathbf{J}^*)^{-1}$, where

$$\mathbf{J}^* = -\frac{\partial^2 l^*}{\partial \phi \partial \phi^T}.$$

The consistency of this estimate is demonstrated in the following result:

THEOREM 1. *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

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

This result is stated in Scott and Wild (2001) but no proof in this general case seems to have appeared in the literature. We sketch a proof in Section 6.1.

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.

3.1 The efficiency bound - general case Suppose we have J populations. Random sampling from these populations is supposed to be governed by a set of J densities $p_{j0} = p_j(x, \theta_0, \eta_0)$ which are contained in the family of densities

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

where θ is a k-dimensional parameter belonging to a set $\mathcal{B}$ and η is an infinite dimensional parameter, belonging to a set $\mathcal{N}$. 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[1] asymptotically linear estimate (RAL estimate) of θ based on these J samples, so that there are functions ϕ_j with

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = n^{-1/2} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \phi_j(X_{ij}) + o_p(1). \tag{7}$$

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

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

where E_j denotes expectation with respect to p_{j0}. Note that the influence functions are assumed to satisfy $E_j[\phi_j] = 0$. 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 $\mathcal{G}$ be a finite-dimensional set of dimension r say, so that

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

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

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

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 $\mathcal{H}$ of these spaces, equipped with the norm defined by

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

[1] See Bickel *et al.* (1993) p18 for the definition of a regular estimate.

The subspace of $\mathcal{H}$ 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 $\mathcal{P}$, there corresponds a score function and subspace of $\mathcal{H}$ generated by the score function. The closure in $\mathcal{H}$ of the union (over all such sub-families) of all these subspaces is called the *nuisance tangent space* and is denoted by $\mathcal{T}_\eta$. 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 $\mathcal{H}$. The projection of $\dot{l}_{j\theta}$ onto the orthogonal complement of $\mathcal{T}_\eta$ is called the *efficient score*, and is denoted by $\dot{l}_j^*$. The matrix $\mathbf{B}$ (the efficiency bound) is given by

$$\mathbf{B}^{-1} = \sum_{j=1}^{J} w_j E_j[\dot{l}_j^* \dot{l}_j^{*T}]. \tag{8}$$

The functions $\mathbf{B}^{-1}\dot{l}_j^*$ 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's 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 $\mathcal{N}$ of the "population expected log-likelihood"*

$$\sum_{j=1}^{J} w_j E_j[\log p_j(X, \theta, \eta)]. \tag{9}$$

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 Section 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 generalised response-selective studies.

In this section we apply the theory of Section 3.1 to regression models for data obtained by the various forms of response-selective sampling described in Section 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 $n_j/N \to w_j$, $j = 1, \ldots, 3$, and $n_4^{(k)}/N \to w_4^{(k)}$, $k = 1 \ldots, K$ where $N = n_1 + n_2 + n_3 + \sum_{k=1}^{K} n_4^{(k)}$. The expected log-likelihood (9) takes the form

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

which up to a term not involving g can be written

$$\int \log g(x) Q^*(x) g_0(x)\, dx + \sum_{k=1}^{K} c_k \log Q_k, \tag{10}$$

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 $Q_{k0} = \int Q_k(x, \theta_0) g_0\, dx$. We need to maximize (10) 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 (10) as

$$\sum_{l=1}^{L} \log g_l\, Q^*(x_l) g_0(x_l) + \sum_{k=1}^{K} c_k \log \left\{ \sum_{l=1}^{L} g_l Q_k(x_l, \theta) \right\}. \tag{11}$$

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

$$g_l = \frac{Q^*(x_l) g_0(x_l)}{\sum_{k=1}^{K} \mu_k Q_k(x_l, \theta)}, \tag{12}$$

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 least favourable distribution (i.e. the maximiser of (10)) might be of the form

$$g(x, \theta, \rho_\theta) = \frac{Q^*(x) g_0(x)}{\sum_k \mu_k(\rho_\theta) Q_k(x, \theta)}, \tag{13}$$

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

$$Q_k(\rho_\theta) = \int g(x, \theta, \rho_\theta) Q_k(x_l, \theta)\, dx.$$

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

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

$$\begin{aligned}
\dot{l}_1^* &= \sum_{k=1}^{K} z_k \left. \frac{\partial \log Q_k(\rho_\theta)}{\partial \theta} \right|_{\theta=\theta_0}, && (14)\\
\dot{l}_2^* &= \left. \frac{\partial \log\{f(x|y,\theta) g(x,\theta,\rho_\theta)\}}{\partial \theta} \right|_{\theta=\theta_0}, && (15)\\
\dot{l}_3^* &= \left. \frac{\partial \log g(x,\theta,\rho_\theta)}{\partial \theta} \right|_{\theta=\theta_0}, && (16)\\
\dot{l}_{4,k}^* &= \left. \frac{\partial \log\{f(x|y,\theta) g(x,\theta,\rho_\theta)\} - \log Q_k(\rho_\theta)}{\partial \theta} \right|_{\theta=\theta_0}. && (17)
\end{aligned}$$

Now we can obtain the information bound in terms of the "asymptotic pseudo-information matrix" $\mathbf{I}^*$ introduced in Section 6.1. From (8) and (15)–(14), the inverse of the information bound $\mathbf{B}$ is

$$\begin{aligned}
\mathbf{B}^{-1} &= w_1 E\left[\left\{ \sum_{k=1}^{K} Z_k \frac{\partial \log Q_k(\rho_\theta)}{\partial \theta} \right\} \left\{ \sum_{k=1}^{K} Z_k \frac{\partial \log Q_k(\rho_\theta)}{\partial \theta} \right\}^T \right] \\
&\quad + w_2 E\left[\left\{ \frac{\partial \log\{f(x|y,\theta) g(x,\theta,\rho_\theta)\}}{\partial \theta} \right\} \left\{ \frac{\partial \log\{f(x|y,\theta) g(x,\theta,\rho_\theta)\}}{\partial \theta} \right\}^T \right] \\
&\quad + w_3 E\left[\left\{ \frac{\partial \log g(x,\theta,\rho_\theta)}{\partial \theta} \right\} \left\{ \frac{\partial \log g(x,\theta,\rho_\theta)}{\partial \theta} \right\}^T \right] \\
&\quad + \sum_{k=1}^{K} w_4^{(k)} E_k\left[\left\{ \frac{\partial \log\{f(x|y,\theta) g(x,\theta,\rho_\theta)\}}{\partial \theta} - \frac{\partial \log Q_k(\rho_\theta)}{\partial \theta} \right\} \times \right. \\
&\qquad \left. \left\{ \frac{\partial \log\{f(x|y,\theta) g(x,\theta,\rho_\theta)\}}{\partial \theta} - \frac{\partial \log Q_k(\rho_\theta)}{\partial \theta} \right\}^T \right]. \qquad (18)
\end{aligned}$$

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

$$\mathbf{B}^{-1} = \mathbf{I}^\dagger_{\theta\theta} + \left(\frac{\partial \rho_\theta}{\partial \theta} \right)^T \mathbf{I}^\dagger_{\rho\theta} + \mathbf{I}^\dagger_{\theta\rho} \frac{\partial \rho_\theta}{\partial \theta} + \left(\frac{\partial \rho_\theta}{\partial \theta} \right)^T \mathbf{I}^\dagger_{\rho\rho} \left(\frac{\partial \rho_\theta}{\partial \theta} \right), \qquad (19)$$

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 Section 6.1. We show in Section 6.4 that

$$\mathbf{I}^\dagger_{\theta\theta} = \mathbf{I}^*_{\theta\theta}, \tag{20}$$

$$\mathbf{I}^\dagger_{\theta\rho} = 0, \tag{21}$$

$$\mathbf{I}^\dagger_{\rho\rho} = -\mathbf{I}^*_{\rho\rho}, \tag{22}$$

and that

$$\frac{\partial \rho_\theta}{\partial\theta} = -(\mathbf{I}^*_{\rho\rho})^{-1}\mathbf{I}^*_{\rho\theta}. \tag{23}$$

Substituting these results into (19) gives

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

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

5. Discussion. In this section, we reexamine 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.*, 1991), and hence our results apply.

9. *Analysis of survival and reliability data.* This falls into the same framework as 8. (Lawless *et al.*, 1991)

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 6.1. 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

$$\begin{aligned}\mathbf{I}^* &= w_2 E\left[-\frac{\partial^2 \log\{f(y|x,\theta)g(x,\theta,\rho)\}}{\partial\phi\partial\phi^T}\right] + w_3 E\left[-\frac{\partial^2 \log g(x,\theta,\rho)}{\partial\phi\partial\phi^T}\right] \\ &\quad + \sum_{k=1}^{K} w_4^{(k)} E_k\left[-\frac{\partial^2 \log\{f(y|x,\theta)g(x,\theta,\rho)\}}{\partial\phi\partial\phi^T}\right] - \sum_{k=1}^{K} c_k \frac{\partial^2 \log Q_k(\rho)}{\partial\phi\partial\phi^T}. \qquad (25)\end{aligned}$$

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 (25), 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

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

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

$$\mathbf{I}^* = \mathbf{I}^\dagger - \int \frac{\partial^2}{\partial\phi\partial\phi^T}\left\{\frac{\sum_{k=1}^K \mu_{k0} Q_k(x,\theta)}{\sum_{k=1}^K \mu_k(\rho) Q_k(x,\theta)}\right\} Q^*(x) g_0(x)\, dx - \sum_{k=1}^{K} c_k \frac{1}{Q_k}\frac{\partial^2 Q_k}{\partial\phi\phi^T}. \tag{26}$$

Moreover, for the $\theta\rho$, $\rho\theta$ and $\rho\rho$ blocks of $\mathbf{I}^*$, note that the function f drops out of (25) and we can write these blocks as

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

Thus, evaluating these derivatives, we get

$$\mathbf{I}^*_{\rho\theta} = \sum_{k=1}^{K} w_4^{(k)} \frac{\partial \log \mu_k(\rho)}{\partial\rho} E_{k\theta}^T \tag{27}$$

where

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

Similarly,

$$\mathbf{I}^*_{\rho\rho} = \sum_{k=1}^{K} w_4^{(k)} \frac{\partial \log \mu_k(\rho)}{\partial \rho} \left(E_{k\rho} - \frac{\partial \log Q_k(\rho)}{\partial \rho} \right)^T \tag{28}$$

where

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

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 6.1. 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 first stage sample. 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

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

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

$$\text{plim}_{N\to\infty} - N^{-1} \frac{\partial^2 \ell^*(\phi, \hat{q})}{\partial \theta \partial \hat{q}_k} = -w_1 E_{k\theta},$$

and

$$\text{plim}_{N\to\infty} - N^{-1} \frac{\partial^2 \ell^*(\theta, \hat{q}_k)}{\partial \rho \partial \hat{q}_k} = -w_1 E_{k\rho}. \tag{29}$$

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

$$\begin{aligned}
\mathbf{V}_1 &= w_2 E\left[\frac{\partial \log f(y|x,\theta)g(x,\theta,\rho)}{\partial\phi}\frac{\partial \log f(y|x,\theta)g(x,\theta,\rho)}{\partial\phi^T}\right] \\
&+ w_3 E\left[\frac{\partial \log g(x,\theta,\rho)}{\partial\phi}\frac{\partial \log g(x,\theta,\rho)}{\partial\phi^T}\right] \\
&+ \sum_{k=1}^{K} w_4^{(k)} E_k\left[\frac{\partial \log f(y|x,\theta)g(x,\theta,\rho)}{\partial\phi}\frac{\partial \log f(y|x,\theta)g(x,\theta,\rho)}{\partial\phi^T}\right] \\
&- \sum_{k=1}^{K} w_4^{(k)} E_k\left[\frac{\partial \log f(y|x,\theta)g(x,\theta,\rho)}{\partial\phi}\right] E_k\left[\frac{\partial \log f(y|x,\theta)g(x,\theta,\rho)}{\partial\phi^T}\right]. \quad (30)
\end{aligned}$$

Using the result (26), this implies that

$$\begin{aligned}
\mathbf{V}_1 &= \mathbf{I}^* + \int \frac{\partial^2}{\partial\phi\partial\phi^T}\left\{\frac{\sum_{k=1}^K \mu_{k0}Q_k(x,\theta)}{\sum_{k=1}^K \mu_k(\rho)Q_k(x,\theta)}\right\} Q^*(x)g_0(x)\,dx + \sum_{k=1}^{K} c_k \frac{1}{Q_k}\frac{\partial^2 Q_k}{\partial\phi\phi^T} \\
&- \sum_{k=1}^{K} w_4^{(k)} E_k\left[\frac{\partial \log f(y|x,\theta)g(x,\theta,\rho)}{\partial\phi}\right] E_k\left[\frac{\partial \log f(y|x,\theta)g(x,\theta,\rho)}{\partial\phi^T}\right]. \quad (31)
\end{aligned}$$

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 (31)is zero, so, using the fact that $n_1\mathrm{Cov}(\hat{q}) \to \mathrm{diag}(Q_0) - Q_0Q_0^T$, we see that the $\theta\theta$ block of $\mathbf{V} = \mathbf{V}_1 + \mathbf{V}_2$ is given by

$$\begin{aligned}
\mathbf{V}_{\theta\theta} &= \mathbf{I}^*_{\theta\theta} - \sum_{k=1}^{K} w_4^{(k)} E_{k\theta}E_{k\theta}^T + w_1\sum_{k=1}^{K} Q_{k0}E_{k\theta}E_{k\theta}^T - w_1\sum_{k=1}^{K}\sum_{l=1}^{K} Q_{k0}Q_{l0}E_{k\theta}E_{l\theta}^T \\
&= \mathbf{I}^*_{\theta\theta} - \sum_{k=1}^{K}\sum_{l=1}^{K} b_{kl}E_{k\theta}E_{l\theta}^T \quad (32)
\end{aligned}$$

where $b_{kl} = w_1Q_{k0}Q_{l0} - \delta_{kl}c_k$. We can rewrite (27) as $\mathbf{I}^*_{\rho\theta} = \mathbf{A}\mathbf{E}_\theta^T$, where $\mathbf{E}_\theta$ has columns $E_{1,\theta}\ldots, E_{k,\theta}$, and $\mathbf{A}$ has l,k element $w_4^{(k)}\frac{\partial\mu_k}{\partial\rho_l}$. Thus, there is a generalised inverse $\mathbf{A}^-$ with $\mathbf{E}_\theta^T = \mathbf{A}^-\mathbf{I}^*_{\theta\theta}$, so that

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

Also, for the $\rho\theta$ block, the integral in (31) is equal to

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

so that

$$\mathbf{V}_{\rho\theta} = \mathbf{I}^*_{\rho\theta} - \sum_{k=1}^{K} w_4^{(k)} \frac{\partial \log \mu_k(\rho)}{\partial \rho^T} E_{k\theta}^T - \sum_{k=1}^{K}\sum_{l=1}^{K} b_{kl} E_{k\rho} E_{l\theta}^T \tag{33}$$

Since

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

we can write (33) as

$$\mathbf{V}_{\rho\theta} = \mathbf{I}^*_{\rho\theta} - \sum_{k=1}^{K} b_{lk} \left(E_{k\rho} - \frac{\partial \log Q_k(\rho)}{\partial \rho} \right) E_{k\theta}^T. \tag{34}$$

Using (28), we can write

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

Similarly, we obtain

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

and hence

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

The asymptotic variance is

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

so using the partitioned matrix inverse formula, the asymptotic covariance matrix of $\hat{\theta}$ can be written as

$$\mathrm{Avar}(\hat{\theta}) = (\mathbf{I}^*_{\theta\theta} - \mathbf{I}^*_{\theta\rho} (\mathbf{I}^*_{\rho\rho})^{-1} \mathbf{I}^*_{\rho\theta})^{-1}. \tag{35}$$

6.2 Proof of Theorem 2. We first show that

$$\left(\frac{\partial \log p_1(x,\theta,\hat{\eta}(\theta))}{\partial\theta}\bigg|_{\theta=\theta_0}, \ldots, \frac{\partial \log p_J(x,\theta,\hat{\eta}(\theta))}{\partial\theta}\bigg|_{\theta=\theta_0} \right) \tag{36}$$

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

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

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

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 maximized 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,

$$\frac{\partial}{\partial t} \sum_{j=1}^{J} w_j E_j[\log p_j(X,\theta,\hat{\eta}(\theta,t))]\Big|_{t=0} = 0. \tag{37}$$

Differentiating (37) with respect to θ gives

$$\sum_{j=1}^{J} w_j \int \frac{\partial^2 \log p_j(X,\theta,\hat{\eta}(\theta,t))}{\partial\theta\partial t}\bigg|_{t=0} p_j(x,\theta_0,\eta_0)\,dx = 0. \tag{38}$$

Also, differentiating both sides of the identity

$$\sum_{j=1}^{J} w_j \int \frac{\partial \log p_j(X,\theta,\hat{\eta}(\theta,t))}{\partial\theta}\bigg|_{t=0} p_j(x,\theta_0,\hat{\eta}(\theta,t))\,dx = 0 \tag{39}$$

with respect to t, we get

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

Setting $\theta = \theta_0,\ t = 0$ and using (38), we get

$$\sum_{j=1}^{J} w_j \int \left.\frac{\partial \log p_j(X,\theta,\hat{\eta}(\theta))}{\partial\theta}\right|_{\theta=\theta_0} \left.\frac{\partial \log p_j(X,\theta,\hat{\eta}(\theta,t))}{\partial t}\right|_{\substack{t=0\\ \theta=\theta_0}} p_j(x,\theta_0,\eta_0)\,dx = 0 \qquad (40)$$

so that (36) is orthogonal to

$$\left(\left.\frac{\partial \log p_1(x,\theta_0,\hat{\eta}(\theta_0,t))}{\partial t}\right|_{t=0}, \ldots, \left.\frac{\partial \log p_J(x,\theta_0,\hat{\eta}(\theta_0,t))}{\partial t}\right|_{t=0} \right). \qquad (41)$$

But $\hat{\eta}(\theta_0,t) = \gamma(t)$ by the Kullback-Leibler information equality, so that (41) is in fact the score function corresponding to the nuisance parameter $\gamma(t)$. Thus (36) is in the nuisance tangent space of $\mathcal{Q}$, and since $\mathcal{Q}$ was an arbitrary finite-dimensional subfamily of $\mathcal{P}$, (36) must lie in the the nuisance tangent space of $\mathcal{P}$.

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

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

By the chain rule,

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

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 (36) 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 (13) is the least favourable distribution. As in Section 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 (10) is given by $g(x) = g(x,\theta,\rho_\theta)$ where ρ_θ is the solution to the $K-1$ equations

$$Q_k(\rho) = \int Q_k(x,\theta)g(x,\theta,\rho)\,dx, k = 1\ldots., K-1. \tag{42}$$

Note that these equations imply that $Q_K(\rho) = \int Q_K(x,\theta)g(x,\theta,\rho)\,dx$ 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 $\tilde{Q}_k(\theta,g) = \int Q_k(x,\theta)\tilde{g}(x)\,dx$. We must show that for all θ and $\tilde{g}$,

$$\begin{aligned}\int \log g(x,\theta,\rho_\theta)Q^*(x)g_0(x)\,dx + \sum_{k=1}^K c_k \log Q_k(\rho_\theta)& \\ \geq \int \log \tilde{g}(x)Q^*(x)g_0(x)\,dx + \sum_{k=1}^K c_k \log \tilde{Q}_k(\theta,g),&\end{aligned} \tag{43}$$

or, equivalently, that

$$\int \log\left\{\frac{g(x,\theta,\rho_\theta)}{\tilde{g}(x)}\right\} Q^*(x)g_0(x)\,dx \geq \sum_{k=1}^K c_k \log\left\{\frac{\tilde{Q}_k(\theta,g)}{Q_k(\rho_\theta)}\right\}. \tag{44}$$

To prove (44), we set $h_k(x,\theta) = Q_k(x,\theta)\tilde{g}(x)/\tilde{Q}_k(\theta,g)$, so h_k is a density. Also define

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

The function H_k is also a density for every θ by (42).

The left hand side of (44) can be written as

$$\begin{aligned}&\int \log\left\{\frac{Q^*(x)g_0(x)P_k(x,\theta,\rho_\theta)}{\tilde{Q}_k(\theta,g)\mu_k(\rho_\theta)h_k(x,\theta)}\right\} Q^*(x)g_0(x)\,dx \\ =\ &\int \log\left\{\frac{H_k(x,\theta)}{h_k(x,\theta)}\right\} Q^*(x)g_0(x)\,dx + \ (1-w_1)\log\left\{\frac{Q_k(\rho_\theta)}{\tilde{Q}_k(\theta,g)}\right\} \\ \geq\ &\mu_k(\rho_\theta)Q_k(\rho_\theta)\int \log\left\{\frac{H_k(x,\theta)}{h_k(x)}\right\} H_k(x,\theta)\,dx + \ (1-w_1)\log\left\{\frac{Q_k(\rho_\theta)}{\tilde{Q}_k(\theta,g)}\right\}.\end{aligned} \tag{45}$$

The last inequality follows because $1 \geq P_k^*(x, \theta, \rho_\theta)$. The integral in (45) is non-negative by the Kullback-Leibler information inequality, so, for each k, we have

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

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

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

since

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

by the Kullback-Leibler inequality. This implies (43).

6.4 Proof of (20)–(23). Evaluating the integral in (26), 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 (27) and (28) imply (20)–(22). For (23), note that by (42) we have

$$Q_k(\rho_\theta) = \int Q_k(x,\theta) g(x,\theta,\rho_\theta)\, dx.$$

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 (23).

References

Bickel, P.J., Klaassen, C.A., Ritov, Y., and Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins University Press, Baltimore.

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

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

Hu, X.J. and 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. and Wild, C.J. (2006). Secondary analyses of case-control sampled data, *Statist. Med.*, **25**, 1323–1339.

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

Murphy, S.A. and Van der Vaart, A.W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.*, **95**, 449–485.

Lawless, J.F., Kalbfleisch, J.D. and Wild, C.J. (1999). Semiparametric methods for response-selective and missing data problems. *J. Roy. Statist. Soc. B*, **61**, 413–438.

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

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

Neuhaus, J., Scott, A.J., and 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.

Robins, J.M., Hsieh, F., and Newey, W. (1995). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *J. Roy. Statist. Soc. B*, **57**, 409–424.

Robins, J.M., Rotnitzky, A., and Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.*, **89**, 846–866.

Scott, A.J., and Wild, C.J. (1986). Fitting logistic models under case-control or choice-based sampling. *J. Roy. Statist. Soc. B*, **48**, 170–182.

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

Scott, A.J., and Wild, C.J. (2001). Maximum likelihood for generalised case-control studies. *J. Stat. Plan. Inf.*, **96**, 3-27.

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

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

Wild, C.J. (1997). Fitting prospective regression models to case-control data. *Biometrika*, **78**, 705–717.