On the Semiparametric Efficiency of the Scott-Wild Estimator under Choice-Based and Two-Phase Sampling

Alan Lee

Received 30 April 2007; Accepted 8 August 2007

Recommended by Paul Cowpertwait

Using a projection approach, we obtain an asymptotic information bound for estimates of parameters in general regression models under choice-based and two-phase outcome-dependent sampling. The asymptotic variances of the semiparametric estimates of Scott and Wild (1997, 2001) are compared to these bounds and the estimates are found to be fully efficient.

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

1. Introduction

Suppose that for each number of subjects, we measure a response $y$ and a vector of covariates $x$, in order to estimate the parameters $\beta$ of a regression model which describes the conditional distribution of $y$ given $x$. If we have sampled directly from the conditional distribution, or even the joint distribution, we can estimate $\beta$ without knowledge of the distribution of the covariates.

In the case of a discrete response, which takes one of $J$ values $y_1, \ldots, y_J$, say, we often estimate $\beta$ using a case-control sample, where we sample from the conditional distribution of $X$ given $Y = y_j$. This is particularly advantageous if some of the values $y_j$ occur with low probability. In case-control sampling, the likelihood involves the distribution of the covariates, which may be quite complex, and direct parametric modelling of this distribution may be too difficult. To get around this problem, the covariate distribution can be treated nonparametrically. In a series of papers (Scott and Wild [1, 2], Wild [3]) Scott and Wild have developed an estimation technique which yields a semiparametric estimate of $\beta$. They dealt with the unknown distribution of the covariates by profiling it out of the likelihood, and derived a set of estimating equations whose solution is the semiparametric estimator of $\beta$. 
This technique also works well for more general sampling schemes, for example, for two-phase outcome-dependent stratified sampling. Here, the sample space is partitioned into $S$ disjoint strata which are defined completely by the values of the response and possibly some of the covariates. In the first phase of sampling, a prospective sample of size $N$ is taken from the joint distribution of $x$ and $y$, but only the stratum to which the individual belongs is observed. In the second phase, for $s = 1,\ldots,S$, a sample of size $n_{1}^{(s)}$ is selected from the $n_{0}^{(s)}$ individuals in stratum $s$ which were selected in the first phase, and the rest of the covariates are measured. Such a sampling scheme can reduce the cost of studies by confining the measurement of expensive variables to the most informative subjects. It is also an efficient design for elucidating the relationship between a rare disease and a rare exposure, in the presence of confounders.

Another generalized scheme that falls within the Scott-Wild framework is that of case-augmented sampling, where a prospective sample is augmented by a further sample of controls. In the prospective sample, we may observe both disease state and covariates, or covariates alone. Such schemes are discussed in Lee et al. [4].

In this paper, we introduce a general method for demonstrating that the Scott-Wild procedures are fully efficient. We use a (slightly extended) version of the theory of semiparametric efficiency due to Bickel et al. [5] to derive an “information bound” for the asymptotic variance of the estimates. We then compute the asymptotic variances of the Scott-Wild estimators, and demonstrate their efficiency by showing that the asymptotic variance coincides with the information bound in each case.

The efficiency of these estimators has been studied by several authors, who have also addressed this question using semiparametric efficiency theory. This theory assumes an i.i.d. sample, and so various ingenious devices have been used to apply it to the case of choice-based sampling. For example, Breslow et al. [6] consider case-control sampling, that the data are generated by Bernoulli sampling, where either a case or a control is selected by a randomisation device with known selection probabilities, and the covariates of the resulting case or control are measured. The randomisation at the first stage means that the i.i.d. theory can be applied.

The efficiency of regression models under an approximation to the two-phase sampling scheme has been considered by Breslow et al. [7] using missing value theory. In this approach, a single prospective sample is taken. For some individuals, the response and the covariates are both observed. For the rest, only the response is measured and the covariates are regarded as missing values. The efficiency bound is obtained using the missing value theory of Robins et al. [8].

In this paper, we adopt a more direct approach. First, we sketch an extension of Bickel-Klaassen-Ritov-Wellner theory to cover the case of sampling from several populations, which we require in the rest of the paper. Such extensions have also been studied by McNeney and Wellner [9], and Bickel and Kwon [10]. Then information bounds for the regression parameters are derived assuming that separate prospective samples are taken from the case and control populations.

The minor modifications to the standard theory required for the multisample efficiency bounds are sketched in Section 2. This theory is then applied to case-control sampling and an information bound derived in Section 3. We also derive the asymptotic
variance of the Scott-Wild estimator and show that it coincides with the information bound.

In Section 4, we deal with the two-phase sampling scheme. We argue that a sampling scheme, equivalent to the two-phase scheme described above is to regard the data as arising from separate independent sampling from \( S + 1 \) populations. This allows the application of the theory sketched in Section 2. We derive a bound and again show that the asymptotic variance of the Scott-Wild estimator coincides with the bound. Finally, mathematical details are given in Section 5.

In the context of data that are independently and identically distributed, Newey [11] characterises the information bound in terms of a population version of a profile likelihood, rather than a projection. A parallel approach to calculating the information bound for the case-control and two-phase problems, using Newey’s “profile” characterisation, is contained in Lee and Hirose [12].

2. Multisamples, information bounds, and semiparametric efficiency

In this section, we give a brief account of the theory of semiparametric efficiency when the data are not independently and identically distributed, but rather consist of separate independent samples from different populations.

Suppose we have \( J \) populations. From each population, we independently select separate i.i.d. samples so that for \( j = 1, \ldots, J \), we have a sample \( \{ x_{ij}, i = 1, \ldots, n_j \} \) from a distribution with density \( p_j \), say. We call the combined sample a multisample. We will consider asymptotics where \( n_j/n \to w_j \), and \( n = n_1 + \cdots + n_J \).

Suppose that \( p_j \) is a member of the family of densities

\[
P = \{ p_j(x, \beta, \eta), \beta \in \mathcal{B}, \eta \in \mathcal{N} \},
\]

(2.1)

where \( \mathcal{B} \) is a subset of \( \mathbb{R}_k \) and \( \mathcal{N} \) is an infinite-dimensional set. We denote the true values of \( \beta \) and \( \eta \) by \( \beta_0 \) and \( \eta_0 \), and \( p_j(x, \beta_0, \eta_0) \) by \( p_{j0} \). Consider asymptotically linear estimates of \( \beta \) of the form

\[
\sqrt{n}(\hat{\beta} - \beta_0) = \frac{1}{\sqrt{n}} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \phi_j(x_{ij}) + o_p(1),
\]

(2.2)

where \( E_j \phi_j(X) = 0 \), \( E_j \) denoting expectation with respect to \( p_{j0} \). The functions \( \phi_j \) are called the influence functions of the estimate and its asymptotic variance is

\[
\sum_{j=1}^{J} w_j E_j[\phi_j \phi_j^T].
\]

(2.3)

The semiparametric information bound is a matrix \( \mathbf{B} \) that is a lower bound for the asymptotic variance of all asymptotically linear estimates of \( \beta \). We have

\[
\text{Avar} \hat{\beta} = \sum_{j} w_j E_j \left[ \phi_j \phi_j^T \right] \geq \mathbf{B},
\]

(2.4)

where the \( \phi_j \) are the influence functions of \( \hat{\beta} \).
The efficiency bound is found as follows. Let \( T \) be a subset of \( \mathcal{R}_p \) so that \( \mathcal{P}_T = \{ p_j(x, \beta, \eta(t)), \beta \in \mathcal{R}, t \in T \} \) is a \( p \)-dimensional submodel of \( \mathcal{P} \). We also suppose that if \( \eta_0 \) is the true value of \( \eta \), then \( \eta(t_0) = \eta_0 \) for some \( t_0 \in T \). Thus, the submodel includes the true model, having \( \beta = \beta_0 \) and \( \eta = \eta_0 \).

Consider the vector-valued score functions
\[
\dot{l}_{j,\eta} = \frac{\partial \log p_j(x, \beta, \eta(t))}{\partial t},
\] (2.5)
whose elements are assumed to be members of \( L_2(P_{j0}) \), where \( P_{j0} \) is the measure corresponding to \( p_j(x, \beta_0, \eta_0) \). Consider also the space \( L_2^k(P_{j0}) \), the space of all \( \mathcal{R}_k \)-valued square-integrable functions with respect to \( P_{j0} \), and the Cartesian product \( \mathcal{H} \) of these spaces, equipped with the norm defined by
\[
\| (f_1, \ldots, f_J) \|_{\mathcal{H}}^2 = \sum_{j=1}^{J} w_j \int \| f_j \|^2 dP_{j0}.
\] (2.6)

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 \((A\dot{l}_{1,\eta}, \ldots, A\dot{l}_{J,\eta})\), where \( A \) ranges over all \( k \) by \( p \) matrices. Thus, to each finite-dimensional sub-family of \( \mathcal{P} \), there correspond a score function and subspace of \( \mathcal{H} \) generated by the score function. The closure in \( \mathcal{H} \) of the span(over all such subfamilies) of all these subspaces is called the nuisance tangent space and denoted by \( \mathcal{H}_\eta \).

Consider also the score functions
\[
\dot{l}_{\beta,j} = \frac{\partial \log p_j(x, \beta, \eta)}{\partial \beta}.
\] (2.7)

The projection \( \dot{l}^* \) in \( \mathcal{H} \) of \( \dot{l}_\beta = (\dot{l}_{\beta,1}, \ldots, \dot{l}_{\beta,J}) \) onto the orthogonal complement of \( \mathcal{H}_\eta \) is called the efficient score, and its elements (which are members of \( L_2^k(P_{j0}) \)) are denoted by \( \dot{l}^*_j \). The matrix \( B \) (the efficiency bound) is given by
\[
B^{-1} = \sum_{j=1}^{J} w_j E_j \left[ \dot{l}^*_j \dot{l}^*_j^T \right].
\] (2.8)

The functions \( B \dot{l}^*_j \) are called the efficient influence functions, and any multisample asymptotically linear estimate of \( \beta \) having these influence functions is asymptotically efficient.

### 3. The efficiency of the Scott-Wild estimator in case-control studies

In this section, we apply the theory sketched in Section 2 to regression models, where the data are obtained by case-control sampling. Suppose that we have a response \( Y \) (assumed as discrete with possible values \( y_1, \ldots, y_J \)) and a vector \( X \) of covariates, and we want to model the conditional distribution of \( Y \) given \( X \) using a regression function
\[
f_j(x, \beta) = P(Y = y_j \mid X = x),
\] (3.1)
say, where \( \beta \) is a \( k \)-vector of parameters. If the distribution of the covariates \( X \) is specified by a density \( g \), then the joint distribution of \( X \) and \( Y \) is

\[
f_j(x, \beta)g(x)
\]

and the conditional distribution of \( x \) given \( Y = y_j \) is

\[
p_j(x, \beta, \eta) = \frac{f_j(x, \beta)g(x)}{\pi_j},
\]

where

\[
\pi_j = \int f_j(x, \beta)g(x)dx.
\]

In case-control sampling, the data are not sampled from the joint distribution, but rather from the conditional distributions of \( X \) given \( Y = y_j \). We are thus in the situation of Section 2 with \( g \) playing the role of \( \eta \) and

\[
p_j(x, \beta, g) = \frac{f_j(x, \beta)g(x)}{\pi_j}.
\]

### 3.1. The information bound in case-control studies

To apply the theory of Section 2, we must identify the nuisance tangent space \( T_\eta \) and calculate the projection of \( \dot{l}_\beta \) on this space. Direct calculation shows that

\[
\dot{l}_{\beta,j} = \frac{\partial \log f_j(x, \beta)}{\partial \beta} - \mathcal{E}_j \left[ \frac{\partial \log f_j(x, \beta)}{\partial \beta} \right],
\]

where \( \mathcal{E}_j \) denotes expectation with respect to the true density \( p_{j0} \), given by \( p_{j0}(x) = p_j(x, \beta_0, g_0) \), where \( \beta_0 \) and \( g_0 \) are the true values of \( \beta \) and \( g \). Here, and in what follows, all derivatives are evaluated at the true values of parameters.

Also, for any finite-dimensional family \( \{g(x, t)\} \) of densities with \( g(x, t_0) = g_0(x) \), we have

\[
\dot{l}_{\eta,j} = \frac{\partial \log g(x, t)}{\partial t} - \mathcal{E}_j \left[ \frac{\partial \log g(x, t)}{\partial t} \right].
\]

It follows by the arguments of Bickel et al. [5, page 52] that the nuisance tangent space is of the form

\[
T_\eta = \{ (h - \mathcal{E}_1[h], \ldots, h - \mathcal{E}_J[h]) : h \in L_{2,k}(G_0) \},
\]

where \( dG_0 = g_0dx \), and \( L_{2,k}(G_0) \) is the space of all \( k \)-dimensional functions \( f \) satisfying the condition \( \int \| f \|^2 dG_0(x) < \infty \).
The efficient score, the projection of \( l_\beta \) on the orthogonal complement of \( T_\eta \), is described in our first theorem. In the theorem, we use the notations
\[
\pi_j 0 = \int f_j(x, \beta_0) \, dG_0(x),
\]
\[
f^*(x) = \sum_{j=1}^{J} \frac{w_j}{\pi_j} f_j(x),
\]
\[
l_{\beta,j} = (l_{\beta,j1}, \ldots, l_{\beta,jk})^T,
\]
\[
\phi_l(x) = \sum_{j=1}^{J} \frac{w_j}{\pi_{j0}} l_{\beta,jl} f_j(x, \beta_0).
\]

(3.9)

Then we have the following result.

**Theorem 3.1.** Let \( A \) be the operator \( L_2(G_0) \to L_2(G_0) \) defined by
\[
(Ah)(x) = f^*(x)h(x) - \sum_{j=1}^{J} \frac{w_j}{\pi_j} f_j(x) \left( \frac{f_j}{\pi_j}, h \right)_2,
\]
(3.10)

where \((\cdot, \cdot)_2\) is the inner product in \( L_2(G_0) \). Then the efficient score has \( j, l \) element
\[
l_{\beta,jl} - h^*_l + E_j[h^*_l],
\]
(3.11)

where \( h^*_l \) is any solution in \( L_2(G_0) \) of the operator equation
\[
Ah^*_l = \phi_l.
\]
(3.12)

A proof is given in Section 5.1.

It remains to identify a solution to (3.12). Define \( P_j(x) = (w_j/\pi_{j0}) f_j(x, \beta_0)/f^*(x) \) and \( v_{jj} = \int P_j P_j^* f^* dG_0 \). Let \( V = (v_{jj}) \), \( W = \text{diag}(w_1, \ldots, w_J) \), and \( M = W - V \). Note that the row and column sums of \( M \) are zero since
\[
w_j - \sum_{j'=1}^{J} \int P_j P_{j'} f^* dG_0 = w_j - \frac{w_j}{\pi_j} \int f_j dG_0 = 0.
\]
(3.13)

Using these definitions and (3.10), we get
\[
Ah_l = h_l f^* - \sum_{j=1}^{J} \left( h_l, \frac{f_j}{\pi_j} \right)_2 P_j f^*
\]
(3.14)

so that \( Ah_l = \phi_l \) if and only if
\[
h_l = \frac{\phi_l}{f^*} + \sum_{j=1}^{J} \left( h_l, \frac{f_j}{\pi_j} \right)_2 P_j.
\]
(3.15)

This suggests that \( h^*_l \) will be of the form
\[
h^*_l = \frac{\phi_l}{f^*} + \sum_{j=1}^{J} c_j P_j
\]
(3.16)
for some constants $c_1, \ldots, c_J$. In order that $h_i^*$ satisfy (3.12), we must have
\[ c_j - \sum_{j'=1}^J c_{j'} \left( P_j', \frac{f_j}{\pi_j} \right)_2 - w_j^{-1}(\phi_l, P_j)_2 = 0, \quad j = 1, \ldots, J. \tag{3.17} \]

Now,
\[ \left( P_j', \frac{f_j}{\pi_j} \right)_2 = \pi_j^{-1} \int P_j' f_j dG_0 = w_j^{-1} \int P_j' P_j f^* dG_0 = (W^{-1}V)_{jj} \]
so that (3.17) will be satisfied if the vector $c = (c_1, \ldots, c_J)^T$ satisfies
\[ M c = d_{(l)}, \tag{3.19} \]
where $d_l = (d_{1l}, \ldots, d_{Jl})^T$ with $d_{jl} = (\phi_l, P_j)_2$. Thus, we require that $c = M^{-1} d_{(l)}$, where $M^{-1}$ is a generalised inverse of $M$.

Our next result gives the information bound.

**Theorem 3.2.** Let $D = (d_1, \ldots, d_k)$ and $\phi = (\phi_1, \ldots, \phi_k)^T$. The inverse of the information bound $B$ is given by
\[ B^{-1} = \sum_{j=1}^I w_j \mathcal{E}_j [l_{\beta, j} l_{\beta, j}^T] - \int \frac{\phi \phi^T}{f^*} dG_0 - D^T M^{-1} D. \tag{3.20} \]

See Section 5.2 for a proof.

### 3.2. Efficiency of the Scott-Wild estimator in case-control studies.

Suppose that we have $J$ disease states (typically $J = 2$, with disease-state case and control), and we choose $n_j$ individuals at random from disease population $j$, $j = 1, \ldots, J$, observing covariates $x_{1j}, \ldots, x_{nj}$ for the individuals sampled from population $j$. Also suppose that we have a regression function $f_j(x, \beta)$, $j = 1, \ldots, J$, giving the conditional probability that an individual with covariates $x$ has disease state $j$. The unconditional density $g$ of the covariates is unspecified. The true values of $\beta$ and $g$ are denoted by $\beta_0$ and $g_0$, and the true probability of being in disease state $j$ is $\pi_{j0} = \int f(x, \beta_0)g_0(x)dx$.

Under the case-control sampling scheme, the log-likelihood (Scott and Wild [2]) is
\[ \sum_{j=1}^J \sum_{i=1}^{n_j} \log f_j(x_{ij}, \beta) + \sum_{j=1}^J \sum_{i=1}^{n_j} \log g(x_{ij}) - \sum_{j=1}^J n_j \log \pi_j. \tag{3.21} \]

Scott and Wild show that the nonparametric MLE of $\beta$ is the “beta” part of the solution of the estimating equation
\[ \sum_{j=1}^J \sum_{i=1}^{n_j} \frac{\partial \log P_j^* (x_{ij}, \beta, \rho)}{\partial \theta} = 0, \tag{3.22} \]
where \( \theta = (\beta, \rho) \), \( \rho = (\rho_1, \ldots, \rho_{I-1}) \),

\[
P_j^* (x, \beta, \rho) = \frac{e^{\phi_j} f_j(x, \beta)}{\sum_{i=1}^{I-1} e^{\phi_i} f_i(x, \beta) + f_j(x, \beta)}, \quad j = 1, \ldots, I-1,
\]

\[
P_j^* (x, \beta, \rho) = \frac{f_j(x, \beta)}{\sum_{i=1}^{I-1} e^{\phi_i} f_i(x, \beta) + f_j(x, \beta)}.
\]

(3.23)

A Taylor series argument shows that the solution of (3.22) is an asymptotically linear estimate.

Thus, to estimate \( \beta \), we are treating the function \( l^* (\theta) = \sum_{j=1}^{I} \sum_{i=1}^{n_j} \log P_j^* (x_{ij}, \beta, \rho) \) as though it were a log-likelihood. Moreover, Scott and Wild indicate that we can obtain a consistent estimate of the standard error by using the second derivative \(-\partial^2 l^* (\theta)/\partial \theta \partial \theta^T\), which they call the “pseudo-information matrix.”

Now let \( n = n_1 + \cdots + n_J \), let the \( n_j \)'s converge to infinity with \( n_j/n = w_j \), \( j = 1, \ldots, J \), and let \( \rho_0 = (\rho_{01}, \ldots, \rho_{0J-1})^T \), where \( \exp(\rho_{0j}) = (w_j/\pi_{0j})/(w_j/\pi_{0j}) \). It follows from the law of large numbers and the results of Scott and Wild that the asymptotic variance of the estimate of \( \beta \) is the \( \beta \beta \) block of the inverse of the matrix

\[
I^* = -\sum_{j=1}^{J} w_j E_j \left[ \frac{\partial^2 \log P_j^* (x_{ij}, \beta, \rho)}{\partial \theta \partial \theta^T} \right],
\]

(3.24)

where all derivatives are evaluated at \((\beta_0, \rho_0)\). Using the partitioned matrix inverse formula, the \( \beta \beta \) block of \((I^*)^{-1}\) is

\[
\left( I_{\beta \beta}^* - I_{\beta \rho}^* \left( I_{\rho \rho}^* \right)^{-1} I_{\rho \beta}^* \right)^{-1},
\]

(3.25)

where \( I^* \) is partitioned as

\[
I^* = \begin{bmatrix}
I_{\beta \beta}^* & I_{\beta \rho}^* \\
I_{\rho \beta}^* & I_{\rho \rho}^*
\end{bmatrix}.
\]

(3.26)

To prove the efficiency of the estimator, we show that the information bound (3.20) coincides with the asymptotic variance (3.25). To prove this, the following representation of the matrix \( I^* \) will be useful. Let \( S \) be the \( J \times k \) matrix with \( j, l \) element \( S_{jl} = (\partial \log f_j(x, \beta)/\partial \beta_l) |_{\beta = \beta_0} \) and \( j \)th row \( S_j \), and let \( E \) be the \( J \times k \) matrix with \( j, l \) element \( E_j[S_{jl}] \). Also note that \( P_j(x) = P_j^* (x, \beta_0, \rho_0) \) and write \( P = (P_1, \ldots, P_3)^T \). Then we have the following theorem.

**Theorem 3.3.**

1. \( I_{\beta \beta}^* = \sum_{j=1}^{J} w_j E_j [S_j S_j^T] - \int S^T P P^T S f^* dG_0. \)
2. Let \( U = WE - \int P P^T S f^* dG_0 \). Then \( I_{\beta \beta}^* \) consists of the first \( J - 1 \) rows of \( U \).
3. \( I_{\rho \rho}^* \) consists of the first \( J - 1 \) rows and columns of \( M = W - V \).

A proof is given in Section 5.3.

Now we show that the information bound coincides with the asymptotic variance. Using the definition \( f_i(x) = \sum_{j=1}^{J} (w_j/\pi_{0j}) f_i(x, \beta_0) \), we can write \( \phi = (S - E)^T P f^* \),
and substituting this and the relationship $l_\beta = S - E$ into (3.20), we get

$$B^{-1} = \sum_{j=1}^{J} w_j E_j \left[ S_j S_j^T \right] - E^T W E - \int (S - E)^T P P^T (S - E) f^* dG_0(x) - D^T M^- D. \quad (3.27)$$

Moreover,

$$D = \int P \phi^T dG_0(x) = \int P P^T (S - E) f^* dG_0(x) = WE - U - VE = ME - U. \quad (3.28)$$

Substituting this into (3.27) and using the relationships described in Theorem 3.3, we get

$$B^{-1} = I^*_{\beta\beta} - U^T M^- U - E^T (I - MM^-) U - U^T (I - M^- M) E. \quad (3.29)$$

By Theorem 3.3, the matrix

$$\begin{bmatrix} I^*_{\beta\beta}^{-1} & 0 \\ 0^T & 0 \end{bmatrix} \quad (3.30)$$

is a generalised inverse of $M$, so $U^T M^- U = I^*_{\beta\beta} I^*_{\beta\beta}^{-1} I^*_{\beta\beta}$. Also,

$$(I - MM^-) U = (I - MM^-) (ME - D) = (I - MM^-) ME - (I - MM^-) MC = 0 \quad (3.31)$$

by the properties of a generalised inverse. Thus, $B^{-1} = I^*_{\beta\beta} - I^*_{\beta\beta} I^*_{\beta\beta}^{-1} I^*_{\beta\beta}$ and the Scott-Wild estimate is efficient.

4. Efficiency of the Scott-Wild estimator under two-stage sampling

In this section, we use the same techniques to show that the Scott-Wild nonparametric MLE is also efficient under two-stage sampling.

4.1. Two stage sampling. In this sampling scheme, the population is divided into $S$ strata, where stratum membership is completely determined by an individual’s response $y$ and possibly some of the covariates $x$—typically those that are cheap to measure. In the first sampling stage, a random sample of size $n_0$ is taken from the population, and the stratum to which the sampled individuals belong is recorded. For the $i$th individual, let $Z_{is} = 1$ if the individual is in stratum $s$, and zero otherwise. Then $n_0^{(s)} = \sum_{i=1}^{n_1^{(s)}} Z_{is}$ is the number of individuals in stratum $s$. In the second sampling stage, for each stratum $s$, a simple random sample of size $n_1^{(s)}$ is taken from the $n_0^{(s)}$ individuals in the stratum. Let $x_{is}$, $i = 1, \ldots, n_1^{(s)}$ and $y_{is}$, $i = 1, \ldots, n_1^{(s)}$ be the covariates and responses for those individuals. Note that $n_1^{(s)}$ depends on $n_0^{(s)}$ and must be regarded as random since $n_0^{(s)} \geq n_1^{(s)}$ for $s = 1, \ldots, S$. We assume that the distribution of $n_1^{(s)}$ depends only on $n_0^{(s)}$, and that, conditional on the $n_0^{(s)}$’s, the $n_1^{(s)}$’s are independent.

As in Section 3, let $f(y \mid x, \beta)$ be the conditional density of $y$ given $x$, which depends on a finite number of parameters $\beta$, which are the parameters of interest. Let $g$ denote the density of the covariates. We will regard $g$ as an infinite-dimensional nuisance parameter.
The conditional density of \((x, y)\), conditional on being in stratum \(s\), using Bayes theorem, is

\[
I_s(x, y) f(y \mid x, \beta) g(x) \int I_s(x, y) f(y \mid x, \beta) g(x) \, dx \, dy',
\]

(4.1)

where \(I_s(x, y)\) is the stratum indicator, having value 1 if an individual having covariates \(x\) and response \(y\) is in stratum \(s\), and zero otherwise. The unconditional probability of being in stratum \(s\) in the first phase is

\[
Q_s = \int I_s(x, y) f(y \mid x, \beta) g(x) \, dx \, dy.
\]

(4.2)

Introduce the function \(Q_s(x, \beta) = \int I_s(x, y) f(y \mid x, \beta) \, dy\). Then,

\[
Q_s = \int Q_s(x, \beta) g(x) \, dx.
\]

(4.3)

Under two-phase sampling, the log-likelihood (Wild [3], Scott and Wild [2]) is

\[
\sum_{s=1}^{S} \sum_{i=1}^{n_{i1}^{(s)}} \log f(y_{is} \mid x_{is}, \beta) + \sum_{s=1}^{S} \sum_{i=1}^{n_{i1}^{(s)}} \log g(x_{is}) + \sum_{s=1}^{S} m_s \log Q_s,
\]

(4.4)

where \(m_s = n_{i1}^{(s)} - n_{i1}^{(s)}\). Scott and Wild show that the semiparametric MLE \(\hat{\beta}\) (i.e., the “\(\beta\)” part of the maximiser \((\hat{\beta}, \hat{g})\) of (4.4)) is equal to the “\(\beta\)” part of the solution of the estimating equations

\[
\frac{\partial \ell^*}{\partial \beta} = 0, \quad \frac{\partial \ell^*}{\partial \rho} = 0.
\]

(4.5)

The function \(\ell^*\) is given by

\[
\ell^*(\beta, \rho) = \sum_{s=1}^{S} \sum_{i=1}^{n_{i1}^{(s)}} \log f(y_{is} \mid x_{is}, \beta) - \sum_{s=1}^{S} \sum_{i=1}^{n_{i1}^{(s)}} \log \left[ \sum_{r} \mu_r(\rho) Q_r(x_{is}, \beta) \right] + \sum_{s=1}^{S} m_s \log Q_s(\rho),
\]

(4.6)
where \( Q_1(\rho), \ldots, Q_S(\rho) \) are probabilities defined by \( \sum_{s=1}^{S} Q_s(\rho) = 1 \) and \( \log Q_s/Q_S = \rho_s, s = 1, \ldots, S \), and \( \mu_s(\rho) = c(n_0 - m_1/Q_s(\rho)) \). The \( \mu_s \)'s depend on the quantity \( c \) and the \( m_1 \)'s, and for fixed values of these quantities, they are completely determined by the \( S - 1 \) quantities \( \rho_s \). Note that the estimating equations (4.5) are invariant under choice of \( c \). It will be convenient to take \( c = N^{-1} \), where \( N = n_0 + n_1 \), where \( n_1 = \sum_{s=1}^{S} n_1(s) \).

In order to apply the theory of Section 2 to two-phase sampling, we will prove that the asymptotics under two-phase sampling are the same as those under the following multi-sample sampling scheme.

1. As in the first scheme, take a random sample of \( n_0 \) individuals and record the stratum in which they fall. This amounts to taking an i.i.d. sample \( \{ (Z_1, \ldots, Z_s), \ldots, (Z_{n_0}) \} \). For fixed values of these quantities, they are completely determined by the \( S - 1 \) quantities \( \rho_s \).

2. For \( s = 1, \ldots, S \), take independent i.i.d. samples \( \{(x_{is}, y_{is}), i = 1, \ldots, n_1(s)\} \) of size \( n_1(s) \) from the conditional distribution of \((x, y)\) given \( s \), having density \( p_s(x, y, \beta, g) = I_s(x, y) f(y | x, \beta) g(x)/Q_s \).

We note that the likelihood under this modified sampling scheme is the same as before, and we show in Theorem 4.1 that the asymptotic distribution of the parameter estimates is also the same. It follows that if an estimate is efficient under the multisampling scheme, it must also be efficient under two-phase sampling.

**Theorem 4.1.** Let \( N = n_0 + n_1 \), where \( n_1 = \sum_{s=1}^{S} n_1(s) \), and suppose that \( \sqrt{N}(n_0/N - w_0) \xrightarrow{p} 0 \) and \( \sqrt{N}(n_1(s)/N - w_s) \xrightarrow{p} 0, s = 1, \ldots, S \).

Let \( \hat{\theta} \) be the solution of the estimating equation (4.5), and let \( \theta_0 \) be the solution to the equation

\[
\log E_s[\psi_s(Z_1, \ldots, Z_s, \theta)] + \sum_{s=1}^{S} \frac{1}{w_s} E_s[\psi_s(x, y, \theta)] = 0,
\]

where \( E_s \) denotes expectation with respect to \( p_s \),

\[
\psi_s(Z_1, \ldots, Z_s, \theta) = \frac{\partial}{\partial \theta} \sum_{s=1}^{S} Z_s \log Q_s,
\]

\[
\psi_s(x, y, \theta) = \frac{\partial}{\partial \theta} \left\{ \log f(y | x, \beta) - \log \left[ \sum_{s} \mu_s Q_s(x, \beta) \right] - \log Q_s \right\}, \ s = 1, \ldots, S.
\]

Then \( \sqrt{N}(\hat{\theta} - \theta_0) \) is asymptotically \( N(0,(I^*)_1V(I^*)_1) \) under both sampling schemes, where \( V = \sum_{s=0}^{S} w_s E_s[(\psi_s - E_s[\psi_s])(\psi_s - E_s[\psi_s])]^T \) and \( I^* = -\sum_{s=0}^{S} w_s E_s[\partial \psi_s / \partial \theta] \).

A proof is given in Section 5.4.

**4.2. The information bound.** Now we derive the information bound for two-stage sampling. By the arguments of Section 4.1, the information bound for two-phase sampling is the same as that for the case of independent sampling from the \( S + 1 \) densities \( p_s(x, y, \beta, g) \),
where
\[
p_s(x, y, \beta, g) = \frac{I_s(x, y)f(y \mid x, \beta)g(x)}{Q_s}, \quad s = 1, \ldots, S,
\]
\[
p_0(x, y, \beta, g) = Q_1^{Z_1} \cdots Q_T^{Z_T},
\]
where \(Z_s = I_s(x, y)\) is the \(s\)th stratum indicator.

First, we identify the form of the nuisance tangent space (NTS) for this problem. As in Section 3, we see that the score functions for this problem are
\[
\dot{l}_0 = \frac{\partial \log p_0(x, y, \beta, g)}{\partial \beta} = \sum_{s=1}^{S} Z_s \mathcal{E}_s \{J\},
\]
\[
\dot{l}_s = \frac{\partial \log p_s(x, y, \beta, g)}{\partial \beta} = J - \mathcal{E}_s \{\mathcal{J}\}, \quad s = 1, \ldots, S,
\]
where \(J = \partial \log f(y \mid x, \beta)/\partial \beta\) and \(\mathcal{E}_s\) denotes expectation with respect to the true density \(p_s(x, y, \beta_0, g_0)\). Similarly, if \(g(x, t)\) is a finite-dimensional subfamily of densities, then
\[
\frac{\partial \log p_s(x, y, \beta, g(x, t))}{\partial t} = h - \mathcal{E}_s \{h\}, \quad s = 1, \ldots, S,
\]
and
\[
\sum_{s=1}^{S} Z_s \mathcal{E}_s \{h\}, \quad (4.11)
\]
where \(h = \partial \log g(x, t)/\partial t\). Arguing as in Section 3, we see that the NTS consists of all elements of the form
\[
T(h) = \left(\sum_{s=1}^{S} Z_s \left(\mathcal{E}_s \{h\} - \mathcal{E} \{h\}\right), h - \mathcal{E}_1 \{h\}, \ldots, h - \mathcal{E}_s \{h\}\right), \quad (4.12)
\]
where \(\mathcal{E}\) denotes expectation with respect to \(G_0\).

As before, the efficient score is \(\dot{l}^* = \dot{l} - T(h^*)\), where \(h^*\) is the element of \(L_{2k}(G_0)\) which minimises \(||\dot{l} - T(h)||^2_{\mathcal{J}}\). An explicit expression for this squared distance is
\[
\sum_{j=1}^{k} \left\{ w_0 \sum_{s=1}^{S} \mathcal{E} \left[ Z_s \left(\mathcal{E}_s \{\mathcal{J}_j\} - \mathcal{E}_s \{h_j\} + \mathcal{E} \{h_j\}\right) \right] + \sum_{s=1}^{S} w_s \mathcal{E}_s \left[ \left\{ \mathcal{E}_s \{\mathcal{J}_j\} - h_j + \mathcal{E}_s \{h_j\}\right\}^2 \right] \right\}, \quad (4.13)
\]
where \(h_j\) and \(\mathcal{J}_j\) are the \(j\)th elements of \(h\) and \(\mathcal{J}\), respectively. To obtain the projection, we must choose \(h_j\) to minimise the term in the braces in (4.13). Some algebra shows that this term may be written as
\[
(h_j, Ah_j)_2 - 2(h_j, \phi_j)_2 + \sum_{s=1}^{S} (w_0 Q_{s0} - w_s) \mathcal{E}_s \{\mathcal{J}_j\}^2 + \sum_{s=1}^{S} w_s \mathcal{E}_s \{\mathcal{J}_j^2\}, \quad (4.14)
\]
where $Q_{s0} = \int Q(x, \beta_0) g_0(x) \, dx$ is the true value of $Q_s$, $(\cdot, \cdot)_2$ is the inner product on $L_2(G_0)$, and $A$ is a selfadjoint nonnegative definite operator on $L_2(G_0)$ defined by

$$Ah = Q^* \left\{ h - \sum_{r=1}^{S} \sum_{s=1}^{S} (\delta_{rs} - y_{rs}) \frac{\int h(x) Q_r(x, \beta_0) g_0(x) \, dx}{Q_{r0}} P_s \right\},$$

$$Q^*(x) = \sum_{s=1}^{S} \frac{w_s}{Q_{s0}} Q_s(x, \beta_0),$$

$$P_s(x) = \frac{(w_s/Q_{s0}) Q_j(x, \beta_0)}{Q^*(x)},$$

$$y_{rs} = \begin{cases} w_0 Q_r(1 - Q_r), & r = s, \\ w_r, & r \neq s, \end{cases}$$

$$\phi_j(x) = \sum_{s=1}^{S} \frac{w_s}{Q_{s0}} Q_s(x, \beta_0) \frac{\partial \log Q_s(x, \beta_0)}{\partial \beta_j} |_{\beta = \beta_0} - \sum_{s=1}^{S} \sum_{r=1}^{S} Q^*_s(x) P_t(x) (\delta_{rs} - y_{rs}) \delta_{j}, (4.15)$$

As in Section 3, (4.14) is minimised when $h_j = h^*_j$, where $h_j$ is a solution of $Ah_j = \phi_j$, which must be of the form

$$h^*_j = \frac{\phi_j}{f^*} + \sum_{r=1}^{S} c_{rj} P_r$$

(4.16)

for constants $c_{rj}$ which satisfy the equation

$$c_{rj} - \sum_{s=1}^{S} \sum_{t=1}^{S} \frac{(\delta_{rs} - y_{rs})}{w_s} v_{st} c_{ij} = \sum_{s=1}^{S} \frac{(\delta_{rs} - y_{rs})}{w_s} d_{sj},$$

(4.17)

where $v_{rs} = \int P_r P_s Q^* \, dG_0$ and $d_{sj} = (P_s, \phi_j)$. Writing $\Gamma = (y_{rs}), \mathbf{C} = (c_{rj}), \mathbf{D} = (d_{rj}), \mathbf{W} = \text{diag}(w_1, \ldots, w_S)$, and $\mathbf{V} = (v_{rs})$, (4.17) can be expressed in matrix terms as

$$\mathbf{MC} = \mathbf{D},$$

(4.18)

where $\mathbf{M} = \mathbf{W}(\mathbf{I} - \Gamma)^{-1} - \mathbf{V}$. These results allow us to find the efficient score and hence the information bound, which is described in the following theorem.

**Theorem 4.2.** The information bound $\mathbf{B}$ is given by

$$\mathbf{B}^{-1} = \sum_{s=1}^{S} w_s \mathbf{E}_s [\phi \phi^T] + \sum_{s=1}^{S} (w_0 Q_{s0} - w_s) \mathbf{E}_s [\phi] [\mathbf{E}_s [\phi]^T] - \int \frac{\phi \phi^T}{Q^*} \, dG_0(x) - \mathbf{D}^T \mathbf{M}^{-1} \mathbf{D}.$$

(4.19)

The proof is similar to that of Theorem 3.2 and hence omitted.
4.3. **Efficiency of the Scott-Wild estimator.** Let \( \hat{\theta} = (\hat{\beta}, \hat{\rho}) \) be the solutions of the estimating equations (4.5). By Theorem 4.1, under suitable regularity conditions, \( \hat{\theta} \) is asymptotically normal with asymptotic variance

\[
I^*-1VI^*-1, \quad (4.20)
\]

where \( I^* \) and \( V \) are as in Theorem 4.1. It turns out that the matrix \( V \) is of the form

\[
V = I^* - I^* \begin{pmatrix} 0 & 0 \\ 0^T & A \end{pmatrix} I^*, \quad (4.21)
\]

for some matrix \( A \). Thus, the asymptotic variance of \( \hat{\theta} \) is

\[
I^*-1 - \begin{pmatrix} 0 & 0 \\ 0^T & A \end{pmatrix}, \quad (4.22)
\]

and it follows from the partitioned matrix inverse formula that the asymptotic variance matrix of \( \hat{\beta} \) is the inverse of

\[
\begin{pmatrix} I^*_\beta\beta & I^*_\beta\rho \\ I^*_\rho\beta & I^*_\rho\rho \end{pmatrix}, \quad (4.23)
\]

where \( I^* \) is partitioned as

\[
I^* = \begin{bmatrix} I^*_\beta\beta & I^*_\beta\rho \\ I^*_\rho\beta & I^*_\rho\rho \end{bmatrix}. \quad (4.24)
\]

To demonstrate the efficiency of \( \hat{\beta} \), we must show that (4.23) and (4.19) coincide. To do this, we need a more explicit formula for \( I^* \). Let \( S \) be the \( S \times k \) matrix with \( s, j \) element \( (\partial \log Q_s(x, \beta_j)|_{\beta = \beta_0}) \), and let \( E \) be the \( S \times k \) matrix with \( l \)th row \( E_s = \mathcal{E}_s[\mathcal{F}] \), where \( \mathcal{F} = (\partial \log f(y | x, \beta)/\partial \beta)|_{\beta = \beta_0} \). Also define

\[
P_s^*(x, \beta, \rho) = \frac{\mu_s(\rho)Q_s(x, \beta)}{\sum_{r=1}^S \mu_r(\rho)Q_r(x, \beta)} \quad (4.25)
\]

and note that \( P_s(x) = P_s^*(x, \beta_0, \rho_0) \), where \( \rho_0 \) satisfies \( Q_s(\rho_0) = Q_{s0}, s = 1, \ldots, S \). Finally, write \( P = (P_1, \ldots, P_S)^T \). Then we have the following theorem.

**Theorem 4.3.**

1. \( I^*_\beta\beta = \sum_{s=1}^S w_s \mathcal{E}_s[\mathcal{F}^T] - \int S^T PP^T SQ^* dG_0(x) \).
2. Let \( U = WE - \int PP^T SQ^* dG_0(x) \). Then \( I^*_\beta\rho = A^T U_0 \), where \( U_0 \) consists of the first \( S - 1 \) rows of \( U \) and \( A \) is a nonsingular \((S - 1) \times (S - 1)\) matrix.
3. \( I^*_\rho\rho = A^T M_0 A \), where \( M_0 \) consists of the first \( S - 1 \) rows and columns of \( M = W(I - \Gamma)^{-1} - V \).

The proof is given in Section 5.5.
We now use Theorems 4.1 and 4.2 to show that the efficiency bound (4.19) equals the asymptotic variance (4.23). Arguing as in Section 3, we get
\[
B^{-1} = I_{\beta}^* - I_{\beta}^* I_{\beta}^{-1} I_{\beta}^* + \left\{ \sum_{s=1}^{S} (w_0 Q_{s,0} - w_s) E_s E_s^T + E^T W (I - \Gamma) E \right\}. \tag{4.26}
\]
We complete the argument by showing that the term in the braces in (4.26) is zero. We have
\[
E^T W (I - \Gamma) E = \sum_{s=1}^{S} (w_s - w_0 Q_{s,0}) E_s E_s^T + w_0 \left( \sum_{s=1}^{S} Q_{s,0} E_s \right) \left( \sum_{s=1}^{S} Q_{s,0} E_s \right)^T
\]
\[
= \sum_{s=1}^{S} (w_s - w_0 Q_{s,0}) E_s E_s^T
\]
since \( \sum_{s=1}^{S} Q_{s,0} E_s = 0 \). Since the term in the braces in (4.26) is zero, the asymptotic variance coincides with the information bound and so the Scott-Wild estimator has full semiparametric efficiency.

5. Proofs

5.1. Proof of Theorem 3.1. The efficient score is the projection of \( \hat{l}_\beta \) onto \( \mathcal{T}_\eta \), and so it is of the form \( \hat{l}_\beta - g \), where \( g \) is the unique minimiser of \( \| \hat{l}_\beta - g \|^2_{\mathcal{H}} \) in \( \mathcal{T}_\eta \). By (3.8), this is \( \hat{l}_\beta - T(h^*) \), where \( h^* \) is the (unique) minimiser of \( \| \hat{l}_\beta - T(h) \|^2_{\mathcal{H}} \) in \( L_{2,k}(G_0) \). Write \( h^* = (h_1^*, \ldots, h_k^*) \). Then,
\[
\| \hat{l}_\beta - T(h^*) \|^2_{\mathcal{H}} = \sum_{i=1}^{k} \sum_{j=1}^{J} \frac{W_{ij}}{\pi_j} \int (\hat{l}_{\beta,ij} - h_i^* - E_j[h_i^*])^2 f_j dG_0 \tag{5.1}
\]
so that we must choose \( h_i^* \) to minimise
\[
\sum_{j=1}^{J} \frac{W_{ij}}{\pi_j} \int (\hat{l}_{\beta,ij} - h_i^* - E_j[h_i^*])^2 f_j dG_0 = \sum_{j=1}^{J} w_j E_j \hat{l}_{\beta,ij}^2 + (A h_i^*, h_i^*)_2 - 2(\phi_i, h_i^*)_2. \tag{5.2}
\]
Now let \( h_i^* \) be any solution in \( L_2(G_0) \) to (3.12). Then for any \( h \) in \( L_2(G_0) \), using the fact that \( A \) is selfadjoint and positive-definite, we get
\[
\sum_{j=1}^{J} w_j E_j \hat{l}_{\beta,ij}^2 + (A h, h)_2 - 2(\phi_i, h)_2 = \sum_{j=1}^{J} w_j E_j \hat{l}_{\beta,ij}^2 - (A h_i^*, h_i^*)_2 + (h - h_i^*, A (h - h_i^*))_2
\]
\[
\geq \sum_{j=1}^{J} w_j E_j (\hat{l}_{\beta,ij}^2) - (A h_i^*, h_i^*)_2 \tag{5.3}
\]
with equality if \( h = h_i^* \) so that the efficient score has \( j, l \) element \( S_{\beta,ij} - h_i^* + E_j[h_i^*] \) as asserted.
5.2. Proof of Theorem 3.2. The $l, l'$ element of $B^{-1}$ is

$$
\sum_{j=1}^{J} w_j E_j \left[ \dot{l}_j^* \dot{l}_{j'}^* \right] = \sum_{j=1}^{J} \frac{w_j}{\pi_j} \left( \dot{l}_{j, l}^* - h_{l}^* - E_j \left( h_{l'}^* \right) \right) \left( \dot{l}_{j, l'}^* - h_{l'}^* - E_j \left( h_{l}^* \right) \right) f_j \ dG_0
$$

$$= \sum_{j=1}^{J} w_j E_j \left[ \dot{l}_{j, l}^* \dot{l}_{j, l'}^* \right] + \left( A h_{l}^* , h_{l'}^* \right)_2 - \left( \phi_{l} , h_{l}^* \right)_2
$$

$$= \sum_{j=1}^{J} w_j E_j \left[ \dot{l}_{j, l}^* \dot{l}_{j, l'}^* \right] - \left( \phi_{l} , h_{l}^* \right)_2
$$

$$= \sum_{j=1}^{J} w_j E_j \left[ \dot{l}_{j, l}^* \dot{l}_{j, l'}^* \right] - \int \phi_{l} \phi_{l'} f^* \ dG_0 - d_{i i} M - d_{i i'} . \tag{5.4}
$$

5.3. Proof of Theorem 3.3. First, we note the formula

$$\frac{\partial^2 \log P_j^*}{\partial \theta \partial \theta^T} = \frac{\partial^2 P_j^*}{\partial \theta \partial \theta^T} \frac{1}{P_j^*} - \frac{\partial \log P_j^*}{\partial \theta} \frac{\partial \log P_j^*}{\partial \theta^T} \tag{5.5}
$$

and the fact that

$$\sum_{j=1}^{J} w_j E_j \left[ \frac{\partial^2 P_j^*}{\partial \theta \partial \theta^T} \frac{1}{P_j^*} \right] = \sum_{j=1}^{J} \frac{w_j}{\pi_j} \int \frac{\partial^2 P_j^*}{\partial \theta \partial \theta^T} \frac{1}{P_j^*} f_j \ dG_0 (x)
$$

$$= \sum_{j=1}^{J} \int \frac{\partial^2 P_j^*}{\partial \theta \partial \theta^T} f^* \ dG_0 (x)
$$

$$= \frac{\partial^2}{\partial \theta \partial \theta^T} \int f^* \ dG_0 (x)
$$

$$= 0 \tag{5.6}
$$

since $\sum_{j=1}^{J} P_j^* = 1$. Hence

$$I^* = - \sum_{j=1}^{J} w_j E_j \left[ \frac{\partial^2 \log P_j^*}{\partial \theta \partial \theta^T} \right] = \sum_{j=1}^{J} w_j E_j \left[ \frac{\partial \log P_j^*}{\partial \theta} \frac{\partial \log P_j^*}{\partial \theta^T} \right]. \tag{5.7}
$$

Next, we note the derivatives

$$\frac{\partial \log P_j^* (x, \beta, \rho)}{\partial \beta} = S_j - \sum_{s=1}^{l} S_s P_s,
$$

$$\frac{\partial \log P_j^* (x, \beta, \rho)}{\partial \rho_r} = \delta_{j,r} - P_r . \tag{5.8}$$
when the derivatives are evaluated at \((\beta_0, \rho_0)\). Thus

\[
I_{\beta\beta}^* = \sum_{j=1}^I w_j E_j \left[ \frac{\partial \log P_j^*}{\partial \beta} \frac{\partial \log P_j^*}{\partial \beta^T} \right]
\]

\[
= \sum_{j=1}^I \frac{w_j}{\pi_j} \int \left( S_j - \sum_{s=1}^J S_s P_s \right) \left( S_j - \sum_{s=1}^J S_s P_s \right)^T f_j(x) dG_0(x)
\]

(5.9)

\[
= \sum_{j=1}^I w_j E_j \left[ S_j S_j^T \right] - \int \left( \sum_{s=1}^J S_s P_s \right) \left( \sum_{s=1}^J S_s P_s \right)^T f^*(x) dG_0(x)
\]

which proves part 1. Also

\[
I_{\rho\beta}^* = \sum_{j=1}^I w_j E_j \left[ \frac{\partial \log P_j^*}{\partial \beta} \frac{\partial \log P_j^*}{\partial \rho r} \right]
\]

\[
= \sum_{j=1}^I \frac{w_j}{\pi_j} \int \left( \delta_{r,s} - P_r \right) \left( S_j - \sum_{s=1}^J S_s P_s \right) f_j(x) dG_0(x)
\]

(5.10)

\[
= w_r E_r \left[ S_{ri} \right] - \int \left( \sum_{j=1}^I S_j P_j \right) P_r f^*(x) dG_0(x)
\]

which proves part 2. Finally,

\[
I_{\rho\rho}^* = \sum_{j=1}^I w_j E_j \left[ \frac{\partial \log P_j^*}{\partial \rho r} \frac{\partial \log P_j^*}{\partial \rho s} \right]
\]

\[
= \sum_{j=1}^I \frac{w_j}{\pi_j} \int \left( \delta_{r,s} - P_r \right) \left( \delta_{j,s} - P_s \right) f_j(x) dG_0(x)
\]

(5.11)

\[
= \int \left( \delta_{rs} - P_s \right) P_r f^*(x) dG_0(x)
\]

\[
= \delta_{rs} w_r - v_{rs}
\]

\[
= M_{rs}.
\]

5.4. Proof of Theorem 4.1. Under the two-stage sampling scheme, the joint distribution of \(\{n_0^{(s)}\}, \{n_1^{(s)}\}\) and \(\{(x_{is}, y_{is}), i = 1, \ldots, n_1^{(s)}, s = 1, \ldots, S\}\) (Wild [3]) is

\[
\prod_{s=1}^S P \left[ n_1^{(s)} | n_0^{(s)} \right] \times \frac{n_0!}{n_0^{(1)}! \cdots n_0^{(S)}!} Q_1^{n_0^{(1)}} \cdots Q_S^{n_0^{(S)}}
\]

\[
\times \prod_{s=1}^S \left\{ \prod_{i=1}^{n_1^{(s)}} I_s(x_{is}, y_{is}) f(y_{is} | x_{is}, \beta) g(x_{is}) \right\}
\]

(5.12)

\[
\times \prod_{s=1}^S \frac{Q_s^{n_0^{(p)}}}{Q_s^{n_0^{(p)}}}.
\]
Thus, conditional on the \( \{ n_0^{(s)} \} \) and \( \{ n_1^{(s)} \} \), the random variables \( \{ (x_{is}, y_{is}), i = 1, \ldots, n_1^{(s)} \}, \) \( s = 1, \ldots, S \) are independent, with \( \{ (x_{is}, y_{is}), i = 1, \ldots, n_1^{(s)} \} \) being an i.i.d. sample from the conditional distribution of \( (x, y) \), conditional on being in stratum \( s \), having density

\[
p_s(x, y, \beta, g) = \frac{I_s(x, y) f(y \mid x, \beta) g(x)}{Q_s}.
\] (5.13)

Define

\[
\psi_s^{(N)}(x, y, \theta) = \frac{\partial}{\partial \theta} \left\{ \log f(y \mid x, \beta) - \log \left[ \sum_s \mu_s Q_s(x, \beta) \right] - \log Q_s \right\}, \quad s = 1, \ldots, S,
\]

\[
\psi_0^{(N)}(Z_1, \ldots, Z_S, \theta) = \frac{\partial}{\partial \theta} \sum_{s=1}^S Z_s \log Q_s.
\] (5.14)

Then the estimating equations (4.5) can be written in the form

\[
\sum_{i=1}^{n_0} \psi_0^{(m)}(Z_{i1}, \ldots, Z_{iS}, \theta) + \sum_{s=1}^S \sum_{i=1}^{n_1^{(s)}} \psi_s^{(m)}(x_{is}, y_{is}, \theta) = 0.
\] (5.15)

Note that the functions \( \psi_s^{(N)} \) depend on \( N \), the \( n_1^{(s)} \)'s and the \( n_0^{(s)} \)'s through the \( \mu_s \)'s, and the \( Q_s \)'s. As \( N \rightarrow \infty \), the functions converge to

\[
\psi_s(x, y, \theta) = \frac{\partial}{\partial \theta} \left\{ \log f(y \mid x, \beta) - \log \left[ \sum_s \mu_s Q_s(x, \beta) \right] - \log Q_s \right\}, \quad s = 1, \ldots, S,
\]

\[
\psi_0(x, y, \theta) = \frac{\partial}{\partial \theta} \sum_{s=1}^S Z_s \log Q_s,
\] (5.16)

where \( \mu_s = w_0 - (w_0 Q_{s0} - w_s)/Q_s \).

Put

\[
S_N(\theta) = \sum_{i=1}^{n_0} \psi_0^{(N)}(Z_{i1}, \ldots, Z_{iS}, \theta) + \sum_{s=1}^S \sum_{i=1}^{n_1^{(s)}} \psi_s^{(N)}(x_{is}, y_{is}, \theta).
\] (5.17)

A standard Taylor expansion argument gives

\[
\sqrt{N}(\hat{\theta} - \theta_0) = \left( - \frac{1}{N} \frac{\partial S_N}{\partial \theta} \bigg|_{\theta=\theta_0} \right)^{-1} \frac{1}{\sqrt{N}} S(\theta_0) + \frac{1}{\sqrt{N}} \left( - \frac{1}{N} \frac{\partial S_N}{\partial \theta} \bigg|_{\theta=\theta_0} \right)^{-1} R,
\] (5.18)

where the jth element of \( R \) is

\[
R_j = \frac{1}{2} \left( \hat{\theta} - \theta_0 \right)^T \frac{\partial^2 S_N}{\partial \theta \partial \theta^T} \bigg|_{\theta=\hat{\theta}} \left( \hat{\theta} - \theta_0 \right),
\] (5.19)

\( S_{Nj} \) is the jth element of \( S_N \) and \( \| \hat{\theta} - \theta_0 \| \leq \| \hat{\theta} - \theta_0 \|. \)
Consider first $S_N(\theta_0)/\sqrt{N}$. We have

$$\frac{S_N(\theta_0)}{\sqrt{N}} = \sqrt{\frac{n_0}{N}} \frac{1}{\sqrt{n_0}} \sum_{i=1}^{n_0} \left\{ \psi_0^{(N)}(Z_{i1}, \ldots, Z_{iS}, \theta_0) - E[\psi_0] \right\}$$

$$+ \sum_{s=1}^{S} \sqrt{\frac{n_1^{(s)}}{N}} \frac{1}{\sqrt{n_1^{(s)}}} \sum_{i=1}^{n_1^{(s)}} \left\{ \psi_s^{(N)}(x_{is}, y_{is}, \theta) - E_s[\psi_s] \right\}$$

$$+ \sqrt{N} \sum_{s=1}^{S} \left( \frac{n_0}{N} - w_0 \right) E[\psi_0] + \sqrt{N} \sum_{s=1}^{S} \left( \frac{n_1^{(s)}}{N} - w_s \right) E_s[\psi_s].$$

(5.20)

Since $\sqrt{N}(n_0/N - w_0)$ and $\sqrt{N}(n_1^{(s)}/N - w_s)$ converge to zero in probability, we see that

$$\frac{S(\theta_0)}{\sqrt{N}} = \sqrt{w_0} \frac{1}{\sqrt{n_0}} \sum_{i=1}^{n_0} \sum_{s=1}^{S} \left\{ \psi_0^{(N)}(Z_{i0}, \theta_0) - E[\psi_0] \right\}$$

$$+ \sum_{s=1}^{S} \sqrt{w_s} \frac{1}{\sqrt{n_1^{(s)}}} \sum_{i=1}^{n_1^{(s)}} \left\{ \psi_s^{(N)}(x_{is}, y_{is}, \theta) - E_s[\psi_s] \right\} + o_p(1).$$

(5.21)

So it suffices to consider $S_N = S_N^{(1)} + S_N^{(2)}$, where $S_N^{(1)}$ and $S_N^{(2)}$ are the first and second terms above.

Under the alternative multisampling scheme, $S_N^{(1)}$ and $S_N^{(1)}$ are independent, as are the $S$ summands of $S_N^{(2)}$. Thus, by the CLT, provided $\psi_s^{(N)}$ converges to $\psi_s$ sufficiently quickly, we see that $S_N$ is asymptotically normal with zero mean and asymptotic variance $V = \sum_{s=0}^{S} w_s \text{Var}[\psi_s]$.

Conversely, under two-phase sampling, the characteristic function of $S_N$ is

$$E[e^{itS_N}] = \sum_0 E[e^{itS_N} | \{n_0^{(s)}\}, \{n_1^{(s)}\}] P[\{n_0^{(s)}\}, \{n_1^{(s)}\}],$$

(5.22)

where $\Sigma_\{0\}$ denotes summation over all possible values of the $\{n_0^{(s)}\}$ and $\{n_1^{(s)}\}$. Since $S_N^{(2)}$ depends on $\{n_0^{(s)}\}$ only through $\{n_1^{(s)}\}$, (5.22) equals

$$E[e^{itS_N}] = \sum_0 E[e^{itS_N^{(1)}} e^{itS_N^{(2)}} | \{n_1^{(s)}\}] \cdot \text{P}[\{n_0^{(s)}\}, \{n_1^{(s)}\}].$$

(5.23)

Let $V_2 = \sum_{s=1}^{S} w_s \text{Var}[\psi_s]$. Assuming that the $\psi_s^{(N)}$ converge sufficiently quickly to the $\psi_s$, it follows that $E[e^{itS_N^{(2)}} | \{n_1^{(s)}\}] \rightarrow \exp\{-(1/2)t^T V_2 t\}$ since the distribution of $S_N^{(2)}$, conditional on $\{n_0^{(s)}\}$ and $\{n_1^{(s)}\}$, is the same as that (unconditionally) under multisampling.

Now let $\epsilon$ be arbitrary and let $N_0$ be such that

$$\left| E[e^{itS_N^{(2)}} | \{n_1^{(s)}\}] - \exp\left\{-\frac{1}{2} t^T V_2 t\right\} \right| < \frac{\epsilon}{2},$$

(5.24)
whenever \( n_i^{(s)} \geq N_0 \) for \( s = 1, \ldots, S \). Also, assume that the (random) sample sizes ultimately get large, in the sense that there exists \( N^* \) such that

\[
P[n_1^{(1)} \geq N_0, \ldots, n_S^{(1)} \geq N_0] \geq 1 - \frac{c}{4}.
\]  

(5.25)

whenever \( N > N^* \). Denote by \( \sum_{(1)} \) summation over all values of \( \{n_0^{(s)}\} \) and \( \{n_1^{(s)}\} \) for which \( n_1^{(s)} \geq N_0 \) for \( s = 1, \ldots, S \), and let \( \sum_{(2)} \) denote summation over all remaining values. Then,

\[
E[e^{itS_N}] = E[e^{itS_N^{(1)}}] \exp \left\{ -\frac{1}{2} t^TV_2t \right\} + \sum_{(1)} E\left[ e^{itS_N^{(1)}} \left( E\left[ e^{itS_N^{(2)}} \mid \{n_1^{(s)}\} \right] - \exp \left\{ -\frac{1}{2} t^TV_2t \right\} \right) \right] \\
+ \sum_{(2)} E\left[ e^{itS_N^{(1)}} \left( E\left[ e^{itS_N^{(2)}} \mid \{n_1^{(s)}\} \right] - \exp \left\{ -\frac{1}{2} t^TV_2t \right\} \right) \right].
\]  

(5.26)

If \( n_0 > N^* \), the sum of the second two terms is less than \( \epsilon \) in absolute value. So

\[
E[e^{itS_N}] = E[e^{itS_N^{(1)}}] \exp \left\{ -\frac{1}{2} t^TV_2t \right\} + o(1).
\]  

(5.27)

Again by the same arguments as above, \( e^{itS_N^{(1)}} \) converges to \( \exp\{-\frac{1}{2}t^TV_1t\} \), where \( V_1 \) is \( w_0 \text{Var}[\psi_0(Z_1, \ldots, Z_S, \theta_0)] \) so that \( E[e^{itS_N}] \) converges to \( \exp\{-\frac{1}{2}t^TV_1t\} \), and hence \( S_N \) converges in distribution to a multivariate normal with variance \( V = V_1 + V_2 \).

Assuming that \( \hat{\theta} \) is \( \sqrt{N} \)-consistent, similar arguments show that \( -(1/N)(\partial S/\partial \theta)|_{\theta=\theta_0} \) converges in probability to \( \mathbf{I}^* \) under both sampling schemes, and that \( R/\sqrt{N} \) is \( o_p(1) \). Thus, as asserted, in both cases, \( \sqrt{N}(\hat{\theta} - \theta_0) \) converges to a multivariate normal with variance \( (\mathbf{I}^*)^{-1}V(\mathbf{I}^*)^{-1} \).

### 5.5. Proof of Theorem 4.3.

Let

\[
P_s^\dagger(x, y, \beta, \rho) = \frac{\mu_s(\rho)I_s(x, y) f(y \mid x, \beta)}{\sum_r \mu_r(\rho)Q_r(x, \beta)}.
\]  

(5.28)

From the definition of \( \mathbf{I}^* \) in Theorem 4.1 and the law of large numbers, we get

\[
\mathbf{I}^* = -w_0 \mathbf{E}_s \left[ \sum_{s=1}^S Z_s \frac{\partial^2 \log Q_s}{\partial \theta \partial \theta^T} \right] - \sum_{s=1}^S w_s \mathbf{E}_s \left[ \frac{\partial^2 \log P_s^\dagger}{\partial \theta \partial \theta^T} - \frac{\partial^2 \log Q_s \mu_s}{\partial \theta \partial \theta^T} \right] \\
= \sum_{s=1}^S w_s \mathbf{E}_s \left[ \frac{\partial \log P_s^\dagger}{\partial \theta} \frac{\partial \log P_s^\dagger}{\partial \theta^T} \right] - \sum_{s=1}^S w_s \mathbf{E}_s \left[ \frac{1}{P_s^\dagger} \frac{\partial^2 P_s^\dagger}{\partial \theta \partial \theta^T} \right] \\
+ \sum_{s=1}^S w_s \frac{\partial^2 \log Q_s \mu_s}{\partial \theta \partial \theta^T} - \sum_{s=1}^S w_0 Q_0 \frac{\partial^2 \log Q_s}{\partial \theta \partial \theta^T}.
\]  

(5.29)
The second term of this expression is zero since

\[
\sum_{s=1}^{S} w_s \mathcal{E}_s \left[ \frac{1}{P^\dagger_s} \frac{\partial^2 P^\dagger_s}{\partial \theta \partial \theta^T} \right] = \sum_{s=1}^{S} \int \frac{\partial^2}{\partial \theta \partial \theta^T} \int P^\dagger_s dy Q^* dG_0(x) \\
= \sum_{s=1}^{S} \int P_s Q^* dG_0(x) \\
= \frac{\partial^2}{\partial \theta \partial \theta^T} \int Q^* dG_0(x) \\
= 0.
\]  

(5.30)

Now, we evaluate \( \mathbf{I}_{\beta \beta}^* \). For the \( \beta \beta \) submatrix, the third and fourth terms of (5.29) are zero. Thus, using the derivative 

\[
\frac{\partial P^\dagger_s}{\partial \beta} = \mathcal{F} - S^T P,
\]  

(5.31)

we get

\[
\mathbf{I}_{\beta \beta}^* = \sum_{s=1}^{S} w_s \mathcal{E}_s \left[ \frac{\partial \log P^\dagger_s}{\partial \beta} \frac{\partial \log P^\dagger_s}{\partial \beta^T} \right] \\
= \sum_{s=1}^{S} \frac{w_s}{Q_s} \left[ \int (\mathcal{F} - S^T P)(\mathcal{F} - S^T P)^T I_s(x,y) f(y|x,\beta_0) dy dG_0(x) \right] \\
= \sum_{s=1}^{S} \frac{w_s}{Q_s} \left[ \int \mathcal{F} \mathcal{F}^T I_s(x,y) f(y|x,\beta_0) dy dG_0(x) - \int S^T P (S^T P)^T Q^*(x) dG_0(x) \right] \\
= \sum_{s=1}^{S} w_s \mathcal{E}_s [\mathcal{F} \mathcal{F}^T] - \int S^T PP^T S Q^* dG_0(x),
\]  

(5.32)

which proves part 1.

Now, consider \( \mathbf{I}_{\rho \beta, rj}^* \). Again, the third and fourth terms of (5.29) are zero. Introduce the parameters \( \lambda_1, \ldots, \lambda_{S-1} \) defined by

\[
\lambda_r = \log \left( \frac{\mu_r(\rho)}{\mu_S(\rho)} \right), \quad r = 1, \ldots, S-1.
\]  

(5.33)

Then,

\[
\frac{\partial P^\dagger_s}{\partial \rho_r} = \sum_{p=1}^{S-1} \frac{\partial \lambda_p}{\partial \rho_r} \frac{\partial P^\dagger_s}{\partial \lambda_p} = \sum_{p=1}^{S-1} \frac{\partial \lambda_p}{\partial \rho_r} (\delta_{sp} - P_p).
\]  

(5.34)
Thus,

\[
I_{p\beta,rj}^* = \sum_{s=1}^{S} w_s \mathcal{E}_s \left[ \frac{\partial \log P^\dagger_s}{\partial \rho} \frac{\partial \log P^\dagger_s}{\partial \beta_j} \right]
\]

\[
= \sum_{s=1}^{S} \frac{w_s}{Q_{0s}} \int \left[ \sum_{p=1}^{S-1} \frac{\partial \lambda_p}{\partial \rho} (\delta_{sp} - P_p) \right] \left( (\mathcal{F} - SP)I_s(x,y) f(y | x, \beta_0) dy dG_0(x) \right)
\]

\[
= \sum_{p=1}^{S-1} \frac{\partial \lambda_p}{\partial \rho} u_{pj},
\]

where

\[
u_{pj} = \sum_{s=1}^{S} \frac{w_s}{Q_{0s}} \int \left( \delta_{ps} - P_p \right) (\mathcal{F} - SP)I_s(x,y) f(y | x, \beta_0) dy dG_0(x). \tag{5.36}\]

Then, as in Theorem 3.3, we see that \(u_{pj}\) is the \(p, j\) element of \(U\), and so part 2 of the theorem is true with \(A_{pr} = \frac{\partial \lambda_p}{\partial \rho_r}\).

The \(\rho \rho\) submatrix is

\[
I_{\rho \rho}^* = \sum_{s=1}^{S} w_s E_s \left[ \frac{\partial \log P^\dagger_s}{\partial \rho} \frac{\partial \log P^\dagger_s}{\partial \rho} \right] - \sum_{s=1}^{S} w_0 Q_{0s} \frac{\partial^2 \log Q_s}{\partial \rho \partial \rho^T} + \sum_{s=1}^{S} \frac{\partial^2 \log Q_s \mu_s}{\partial \rho \partial \rho^T}
\]

\[
= \sum_{s=1}^{S} w_s E_s \left[ \frac{\partial \log P^\dagger_s}{\partial \rho} \frac{\partial \log P^\dagger_s}{\partial \rho} \right] - w_0 \sum_{s=1}^{S} \frac{1}{\kappa_s} \frac{\partial Q_s}{\partial \rho} \frac{\partial Q_s}{\partial \rho^T}, \tag{5.37}\]

where \(\kappa_s = Q_{0s} w_s / \epsilon_s\). It follows from (5.34) that \(I_{\rho \rho}^* = \Lambda^T M_0 A\), where \(M_0\) has \(p, q\) element

\[
\sum_{s=1}^{S} w_s E_s \left[ \frac{\partial \log P^\dagger_s}{\partial \lambda_p} \frac{\partial \log P^\dagger_s}{\partial \lambda_q} \right] - w_0 \sum_{s=1}^{S} \frac{1}{\kappa_s} \frac{\partial Q_s}{\partial \lambda_p} \frac{\partial Q_s}{\partial \lambda_q}. \tag{5.38}\]

As in Section 5.3, the first term of this expression is \(\delta_{pq} w_p - v_{pq}\). Routine calculations using the relationships \(\lambda_p = \log(\mu_p / \mu_S)\) and \(\mu_p = w_0 - \epsilon_p / Q_p\) give

\[
\frac{\partial Q_p}{\partial \lambda_q} = \delta_{pq} \kappa_p - \frac{\kappa_p \kappa_q}{\kappa^*}, \tag{5.39}\]

where \(\kappa^* = \sum_{p=1}^{S} \kappa_p\). This representation implies that

\[
\sum_{s=1}^{S} \frac{1}{\kappa_s} \frac{\partial Q_s}{\partial \lambda_p} \frac{\partial Q_s}{\partial \lambda_q} = \frac{\partial Q_p}{\partial \lambda_q} \tag{5.40}\]

so that the \(p, q\) element of \(M_0\) is \(\delta_{pq} w_p - v_{pq} - w_0 (\partial Q_p / \partial \lambda_q)\).

By the Sherman-Morrison formula, the \(p, q\) element of the matrix \(W(I - \Gamma)^{-1} - W\) is

\[-w_0 (\partial Q_p / \partial \lambda_q).\]

So the matrix \(M_0\) consists of the first \(S - 1\) rows and columns of \(W - V + W(I - \Gamma)^{-1} - W = W(I - \Gamma)^{-1} - V = M\).
References


Alan Lee: Department of Statistics, University of Auckland, Auckland 1142, New Zealand
Email address: aj.lee@auckland.ac.nz
Submit your manuscripts at http://www.hindawi.com