On the semi-parametric efficiency of the Scott-Wild estimator under choice-based and two-phase sampling

ALAN LEE

Department of Statistics, University of Auckland, Auckland, New Zealand.

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 semi-parametric estimates of Scott and Wild (1997, 2001) are compared to these bounds and the estimates are found to be fully efficient.

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

April 30, 2007

1. Introduction

Suppose that for each of a number of subjects, we measure a response y and a vector of covariates x, in order to estimate the parameters β 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 β without knowledge of the distribution of the covariates.

In the case of a discrete response, which takes one of J values $y_1, \dots, y_J$, say, we often estimate β 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 non-parametrically. In a series of papers, (Scott and Wild 1986, 1997, 2001, Wild 1991) Scott and Wild have developed an estimation technique which yields a semi-parametric estimate of β. 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 semi-parametric estimator of β.

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 the individual belongs to is observed. In the second phase, for $s = 1, \dots, S$, a sample of size $n_1^{(s)}$ is selected from the $n_0^{(s)}$ individuals in stratum s who 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 augmenmted 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, Scott and Wild (2006).

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 semi-parametric efficiency due to Bickel *et al.* (1993) 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 semi-parametric efficiency theory. This theory assumes an i.i.d. sample, so various ingenious devices have been used to apply it to the case of choice-based sampling. For example, Breslow, Robins and Wellner (2000) consider case-control sampling, 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. 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, McNeney and Wellner (2003) 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, the covariates being regarded as missing values. The efficiency bound is obtained using the missing value theory of Robins, Hsieh and Newey (1995).

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 (2000) and Bickel and Kwon (2001). 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 multi-sample 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 (1994) characterizes 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" characterization, is contained in Lee and Hirose (2007).

2. Multi-samples, information bounds and semi-parametric efficiency

In this section, we give a brief account of the theory of semi-parametric 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 multi-sample. 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

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

where $\mathcal{B}$ is a subset of $\Re_k$ and $\mathcal{N}$ is infinite-dimensional. We denote the true values of β and η by β_0 and η_0, and $p_j(x, \beta_0, \eta_0)$ by p_{j0}. Consider *asymptotically linear* estimates of β of the form

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

where $E_j\phi_j(X) = 0$, E_j denoting expectation with respect to p_{j0}. The functions ϕ_j are called the *influence functions* of the estimate and its asymptotic variance is $\sum_{j=1}^{J} w_j E_j[\phi_j \phi_j^T]$.

The *semi-parametric information bound* is a matrix $\mathbf{B}$ that is a lower bound for the asymptotic variance of all asymptotically linear estimates of β: we have

$$\text{Avar}\hat{\beta} = \sum_j E_j[\phi_j \phi_j^T] \geq \mathbf{B}$$

where the ϕ_j are the influence functions of $\hat{\beta}$.

The efficiency bound is found as follows. Let T be a subset of of $\Re_p$, so that $\mathcal{P}_T = \{p_j(x, \beta, \eta(t)),$
$\beta \in \mathcal{B}, t \in T\}$ is a p-dimensional submodel of $\mathcal{P}$. We also suppose that if η_0 is the true value of η, 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

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

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 $L_{2k}(P_{j0})$, 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

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

The subspace of $\mathcal{H}$ generated by the score functions $(\dot{l}_{1,\eta}, \dots, \dot{l}_{J,\eta})$ is the set of all vector-valued functions of the form $(\mathbf{A}\dot{l}_{1,\eta}, \dots, \mathbf{A}\dot{l}_{J,\eta})$ where $\mathbf{A}$ ranges over all k by p 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 span(over all such sub-families) of all these subspaces is called the *nuisance tangent space* and is denoted by $\mathcal{T}_\eta$.

Consider also the score functions

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

The projection $\dot{l}^*$ in $\mathcal{H}$ of $\dot{l}_\beta = (\dot{l}_{1,\beta}, \dots \dot{l}_{J,\beta})$ onto the orthogonal complement of $\mathcal{T}_\eta$ is called the *efficient score*, and its elements (which are members of $L_{2,k}(G_0)$) are denoted by $\dot{l}^*_j$. The matrix $\mathbf{B}$ (the efficiency bound) is given by

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

The functions $\mathbf{B}\,\dot{l}^*_j$ are called the *efficient influence functions*, and any multi-sample asymptotically linear estimate of β 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 above in Section 2 to regression models where the data are obtained by case-control sampling. Suppose that we have a response Y (assumed discrete with possible values $y_1, \dots, y_J$) and a vector X of covariates, and we want to model the conditional distribution of Y given X using a regression function

$$f_j(x, \beta) = P(Y = y_j | X = x),$$

say, where β 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

$$p_j(x, \beta, \eta) = f_j(x, \beta) g(x)/\pi_j$$

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 are sampled 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 η and

$$p_j(x,\beta,g) = f_j(x,\beta)g(x)/\pi_j.$$

3.1 The information bound in case-control studies. To apply the theory of Section 2, we must identify the nuisance tangent space $\mathcal{T}_\eta$ 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 β_0 and g_0 are the true values of β and g. Here, and in what follows, all derivatives are evalueted 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. (1993, p52) that the nuisance tangent space is of the form

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

where $dG_0 = g_0 dx$, 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 $\dot{l}_\beta$ on the orthogonal complement of $\mathcal{T}_\eta$, is described in our first theorem. In the theorem, we use the notations $\pi_{j0} = \int f_j(x,\beta_0)\,dG_0(x)$,

$$f^*(x) = \sum_{j=1}^J \frac{w_j}{\pi_j} f_j(x),$$

$$\dot{l}_{\beta,j} = (\dot{l}_{\beta,j1}, \ldots, \dot{l}_{\beta,jk})^T$$

and

$$\phi_l(x) = \sum_{j=1}^J \frac{w_j}{\pi_{j0}} \dot{l}_{\beta,jl} f_j(x,\beta_0).$$

Then we have the following result:

Theorem 1. *Let A be the operator $L_2(G_0) \to L_2(G_0)$ defined by*

$$(Ah)(x) = f^*(x)h(x) - \sum_{j=1}^{J} \frac{w_j}{\pi_j} f_j(x)(f_j/\pi_j, h)_2 \tag{3}$$

Then the efficient score has j, l element

$$\dot{l}_{\beta,jl} - h_l^* + E_j[h_l^*]$$

where h_l^* is any solution in $L_2(G_0)$ of the operator equation

$$Ah_l^* = \phi_l. \tag{4}$$

A proof is given in Section 5.1.

It remains to identify a solution to (4). Define $P_j(x) = \frac{w_j}{\pi_{j0}} f_j(x, \beta_0)/f^*(x)$ and $v_{jj'} = \int P_j P_{j'} f^* \, dG_0$. Let $\mathbf{V} = (v_{jj'})$, $\mathbf{W} = \text{diag}(w_1, \ldots, w_J)$ and $\mathbf{M} = \mathbf{W} - \mathbf{V}$. Note that the row and column sums of $\mathbf{M}$ are zero, since

$$w_j - \sum_{j'=1}^{J} \int P_j P_{j'} f^* \, dG_0 = w_j - \frac{w_j}{\pi_j} \int f_j \, dG_0 = 0.$$

Using these definitions and (3), we get

$$Ah_l = h_l f^* - \sum_{j=1}^{J} (h_l, f_j/\pi_j)_2 P_j f^*$$

so that $Ah_l = \phi_l$ if and only if

$$h_l = \frac{\phi_l}{f^*} + \sum_{j=1}^{J} (h_l, f_j/\pi_j)_2 P_j.$$

This suggests that h_l^* will be of the form

$$h_l^* = \frac{\phi_l}{f^*} + \sum_{j=1}^{J} c_j P_j$$

for some constants $c_1, \ldots, c_J$. In order that h_l^* satisfy (4), we must have

$$c_j - \sum_{j'=1}^{J} c_{j'} (P_{j'}, f_j \pi_j)_2 - w_j^{-1} (\phi_l, P_j)_2 = 0, \; j = 1, \ldots, J. \tag{5}$$

Now

$$\begin{aligned}(P_{j'}, f_j\pi_j)_2 &= \int P_{j'}, f_j/\pi_j \, dG_0 \\ &= w_j^{-1} \int P_{j'}, P_j f^* \, dG_0 \\ &= (\mathbf{W}^{-1}\mathbf{V})_{jj'}\end{aligned}$$

so that (5) will be satisfied if the vector $c = (c_1, \ldots, c_J)^T$ satisfies

$$\mathbf{M}c = d_{(l)} \tag{6}$$

where $d_l = (d_{1l}, \ldots, d_{Jl})^T$ with $d_{jl} = (\phi_l, P_j)_2$. Thus, we require that $c = \mathbf{M}^- d_{(l)}$ where $\mathbf{M}^-$ is a generalised inverse of $\mathbf{M}$.

Our next result gives the information bound.

Theorem 2. *Let* $\mathbf{D} = (d_1, \ldots, d_k)$ *and* $\phi = (\phi_1, \ldots, \phi_k)^T$*. The inverse of the information bound* $\mathbf{B}$ *is given by*

$$\mathbf{B}^{-1} = \sum_{j=1}^{J} w_j \mathcal{E}_j[\dot{l}_{\beta,j}\dot{l}_{\beta,j}^T] - \int \frac{\phi\phi^T}{f^*} \, dG_0 - \mathbf{D}^T\mathbf{M}^-\mathbf{D}. \tag{7}$$

See Section 5.2 for a proof.

3.2. Efficiency of the Scott-Wild estimator in case-control studies. Suppose we have J disease states (typically J=2, with disease states case and control), and we choose n_j individuals at random from disease population j, $j = 1, \ldots, J$, observing covariates $x_{1,j}, \ldots, x_{n_j,j}$ 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 β and g are denoted by β_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 is (Scott and Wild 2001)

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j} \log f_j(x_{ij}, \beta) + \sum_{j=1}^{J}\sum_{i=1}^{n_j} \log g(x_{ij}) - \sum_{j=1}^{J} n_j \log \pi_j. \tag{8}$$

Scott and Wild show that the non-parametric MLE of β is the "beta" part of the solution of the estimating equation

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j} \frac{\partial \log P_j^*(x_{ij}, \beta, \rho)}{\partial\theta} = 0, \tag{9}$$

where $\theta = (\beta, \rho)$, $\rho = (\rho_1, \ldots, \rho_{J-1})$,

$$P_j^*(x, \beta, \rho) = \frac{e^{\rho_j} f_j(x, \beta)}{\sum_{l=1}^{J-1} e^{\rho_l} f_l(x, \beta) + f_J(x, \beta)}, \quad j = 1, \ldots, J-1 \tag{10}$$

and

$$P_J^*(x,\beta,\rho) = \frac{f_J(x,\beta)}{\sum_{l=1}^{J-1} e^{\rho_l} f_l(x,\beta) + f_J(x,\beta)}. \tag{11}$$

A Taylor series argument shows that the solution of (9) is an asyptotically linear estimate.

Thus, to estimate β, we are treating the function $l^*(\theta) = \sum_{j=1}^{J}\sum_{i=1}^{n_j} \log P_j^*(x_{ij},\beta,\rho)$ 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 $-\frac{\partial^2 l^*(\theta)}{\partial\theta\partial\theta^T}$, which they call the "pseudo-information matrix".

Now let $n = n_1 + \cdots + n_J$ and let the n_j's converge to infinity with $n_j/n \to w_j$, $j = 1,\ldots,J$, and let $\rho_0 = (\rho_{01},\ldots,\rho_{0,J-1})^T$ 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 β is the $\beta\beta$ block of the inverse of the matrix

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

where all derivatives are evaluated at (β_0,ρ_0). Using the partitioned matrix inverse formula, the the β,β block of $(\mathbf{I}^*)^{-1}$ is

$$(\mathbf{I}^*_{\beta\beta} - \mathbf{I}^*_{\beta\rho}(\mathbf{I}^*_{\rho\rho})^{-1}\mathbf{I}^*_{\rho\beta})^{-1}, \tag{12}$$

where $\mathbf{I}^*$ is partitioned as

$$\mathbf{I}^* = \left[\begin{array}{cc} \mathbf{I}^*_{\beta\beta} & \mathbf{I}^*_{\beta\rho} \\ \mathbf{I}^*_{\rho\beta} & \mathbf{I}^*_{\rho\rho} \end{array}\right].$$

To prove the efficiency of the estimator, we show that the information bound (7) coincides with the asymptotic variance (12). To prove this, the following representation of the matrix $\mathbf{I}^*$ will be useful. Let $\mathbf{S}$ be the $J \times k$ matrix with j,l element $S_{jl} = \frac{\partial \log f_j(x,\beta)}{\partial \beta_l}|_{\beta=\beta_0}$ and jth row S_j, and let $\mathbf{E}$ be the $J \times k$ matrix with j,l element $\mathcal{E}_j[S_{jl}]$. Also note that $P_j(x) = P_j^*(x,\beta_0,\rho_0)$ and write $P = (P_1,\ldots,P_S)^T$. Then we have the following theorem:

Theorem 3.

1. $\mathbf{I}^*_{\beta\beta} = \sum_{j=1}^{J} w_j \mathcal{E}_j[S_j S_j^T] - \int \mathbf{S}^T P P^T \mathbf{S} f^*\, dG_0$,
2. *Let* $\mathbf{U} = \mathbf{WE} - \int PP^T\mathbf{S}f^*\, dG_0$. *Then* $\mathbf{I}^*_{\rho\beta}$ *consists of the first* $J-1$ *rows of* $\mathbf{U}$,
3. $\mathbf{I}^*_{\rho\rho}$ *consists of the first* $J-1$ *rows and columns of* $\mathbf{M} = \mathbf{W} - \mathbf{V}$.

A proof is given in Section 5.

Now, we show that the information bound coincides with the asymptotic variance. Using the definition $\phi_l(x) = \sum_{j=1}^{J} \frac{w_j}{\pi_{j0}} \dot{l}_{\beta,jl} f_j(x,\beta_0)$, we can write $\phi = (\mathbf{S}-\mathbf{E})^T P f^*$, and substituting this and the relationship $\dot{l}_\beta = \mathbf{S} - \mathbf{E}$ into (7), we get

$$\mathbf{B}^{-1} = \sum_{j=1}^{J} w_j E_j[S_j S_j^T] - \mathbf{E}^T\mathbf{W}\mathbf{E} - \int (\mathbf{S}-\mathbf{E})^T P P^T (\mathbf{S}-\mathbf{E}) f^*\, dG_0(x) - \mathbf{D}^T\mathbf{M}^-\mathbf{D} \tag{13}$$

Moreover,

$$\begin{aligned}
\mathbf{D} &= \int P\phi^T \, dG_0(x) \\
&= \int PP^T(\mathbf{S} - \mathbf{E}) f^* \, dG_0(x) \\
&= \mathbf{WE} - \mathbf{U} - \mathbf{VE} \\
&= \mathbf{ME} - \mathbf{U}.
\end{aligned}$$

Substituting this into (13) and using the relationships described in Theorem 3, we get

$$\mathbf{B}^{-1} = \mathbf{I}^*_{\beta\beta} - \mathbf{U}^T\mathbf{M}^-\mathbf{U} - \mathbf{E}^T(\mathbf{I} - \mathbf{MM}^-)\mathbf{U} - \mathbf{U}^T(\mathbf{I} - \mathbf{M}^-\mathbf{M})\mathbf{E}. \tag{14}$$

By Theorem 3, the matrix

$$\begin{bmatrix} \mathbf{I}^*_{\rho\rho}{}^{-1} & \mathbf{0} \\ \mathbf{0}^T & 0 \end{bmatrix}$$

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

$$\begin{aligned}
(\mathbf{I} - \mathbf{MM}^-)\mathbf{U} &= (\mathbf{I} - \mathbf{MM}^-)(\mathbf{ME} - \mathbf{D}) \\
&= (\mathbf{I} - \mathbf{MM}^-)\mathbf{ME} - (\mathbf{I} - \mathbf{MM}^-)\mathbf{MC} \\
&= \mathbf{0}
\end{aligned}$$

by the properties of a generalized inverse. Thus, $\mathbf{B}^{-1} = \mathbf{I}^*_{\beta\beta} - \mathbf{I}^*_{\beta\rho}\mathbf{I}^*_{\rho\rho}{}^{-1}\mathbf{I}^*_{\rho\beta}$ 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 non-parametric 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 ith individual, let $Z_{is} = 1$ if the individual is in stratum s, and zero otherwise. Then $n_0^{(s)} = \sum_{i=1}^{n_1} Z_{is}$ 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 $f(y|\, x, \beta)$ be the conditional density of y given x, which depends on a finite number of parameters β, 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 is, using Bayes theorem,

$$\frac{I_s(x,y)f(y|x)g(x)}{\int\int I_s(x,y)f(y|x,\beta)g(x)\,dx\,dy}$$

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

$$Q_s = \int\int I_s(x,y)f(y|x,\beta)g(x)\,dx\,dy.$$

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

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

Under two-phase sampling, the log likelihood is (Wild 1991, Scott and Wild, 2001)

$$\sum_{s=1}^{S}\sum_{i=1}^{n_1^{(s)}}\log f(y_{is}|x_{is},\beta) + \sum_{s=1}^{S}\sum_{i=1}^{n_1^{(s)}}\log g(x_{is}) + \sum_{s=1}^{S} m_s \log Q_s \tag{15}$$

where $m_s = n_0^{(s)} - n_1^{(s)}$. Scott and Wild show that the semi-parametric MLE $\hat{\beta}$ (i.e. the "β" part of the maximiser $(\hat{\beta}, \hat{g})$ of (15)) is equal to the "β" part of the solution of the estimating equations

$$\frac{\partial \ell^*}{\partial \beta} = 0, \quad \frac{\partial \ell^*}{\partial \rho} = 0. \tag{16}$$

The function ℓ^* is given by

$$\ell^*(\beta,\rho) = \sum_{s=1}^{S}\sum_{i=1}^{n_1^{(s)}}\log f(y_{is}|x_{is},\beta) - \sum_{s=1}^{S}\sum_{i=1}^{n_1^{(s)}}\log\Big[\sum_r \mu_r(\rho)Q_r(x_{is},\beta)\Big] + \sum_{s=1}^{S} m_s\log Q_s(\rho),$$

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 $\mu_s(\rho) = c(n_0 - m_s/Q_s(\rho))$. The μ_s's depend on the quantity c and the m_s's, and for fixed values of these quantities are completely determined by the $S - 1$ quantities ρ_s. Note that the estimating equations (16) are invariant under choice of c. It will be convenient to take c as 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 $\{(Z_{i1}, \ldots, Z_{iS}), i = 1, \ldots, n_0\}$ of size n_0 from $MULT(1, Q_1, \ldots, Q_S)$,

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 $p_s(x, y, \beta, g) = I_s(x, y)f(y|x)g(x)/Q_s$.

We note that the likelihood under this modified sampling scheme is the same as before, and we show in Theorem 4 below that the asymptotic distribution of the parameter estimates is also the same. It follows that if an estimate is efficient under the multi-sampling scheme, it must also be efficient under two-phase sampling.

Theorem 4. *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 (16), and let* θ_0 *be the solution to the equation*

$$w_0 \mathcal{E}[\psi_0(Z_1, \ldots, Z_s, \theta)] + \sum_{s=1}^S \mathcal{E}_s[\psi_s(x, y, \theta)] = 0, \tag{17}$$

where $\mathcal{E}_s$ *denotes expectation with respect to* p_s*,*

$$\psi_0(Z_1, \ldots, Z_s, \theta) = \frac{\partial}{\partial \theta} \sum_{s=1}^S Z_s \log Q_s$$

and

$$\psi_s(x, y, \theta) = \frac{\partial}{\partial \theta}\Big\{ \log f(y|x, \beta) - \log \Big[\sum_s \mu_s Q_s(x, \beta) \Big] - \log Q_s \Big\}, \qquad s = 1, \ldots, S.$$

Then $\sqrt{N}(\hat{\theta} - \theta_0)$ *is asymptotically* $N(0, (\mathbf{I}^*)^{-1}\mathbf{V}(\mathbf{I}^*)^{-1})$ *under both sampling schemes, where* $\mathbf{V} = \sum_{s=0}^S w_s E_s[(\psi_s - E_s[\psi_s])[(\psi_s - E_s[\psi_s])^T]$ *and* $\mathbf{I}^* = -\sum_{s=0}^S w_s E_s[\partial \psi_s/\partial \theta]$*.*

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

$$p_s(x, y, \beta, g) = \frac{I_s(x, y)f(y|x, \beta)g(x)}{Q_s}, s = 1, \ldots, S$$

and

$$p_0(x, y, \beta, g) = Q_1^{Z_1} \cdots Q_J^{Z_J}$$

where $Z_s = I_s(x, y)$ is the sth 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

$$\dot{l}_0 = \frac{\partial \log p_0(x, y, \beta, g)}{\partial \beta} = \sum_{s=1}^{S} Z_s \mathcal{E}_s[\mathcal{S}],$$

and

$$\dot{l}_s = \frac{\partial \log p_s(x, y, \beta, g)}{\partial \beta} = \mathcal{S} - \mathcal{E}_s[\mathcal{S}], s = 1, \ldots, S$$

where $\mathcal{S} = \frac{\partial \log f(y|x,\beta)}{\partial \beta}$ 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 $\frac{\partial \log p_s(x,y,\beta,g(x,t))}{\partial t} = h - \mathcal{E}_s[h], s = 1, \ldots, S$ and $\frac{\partial \log p_0(x,y,\beta,g(x,t))}{\partial t} = \sum_{s=1}^{S} Z_s \mathcal{E}_s[h]$, where $h = \frac{\partial \log g(x,t)}{\partial t}$. Arguing as in Section 3, we see that the NTS consists of all elements of the form

$$T(h) = \Big(\sum_{s=1}^{S} Z_s(\mathcal{E}_s[h] - \mathcal{E}[h]), h - \mathcal{E}_1[h], \ldots, h - \mathcal{E}_s[h]\Big)$$

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 $||\dot{l} - T(h)||^2_{\mathcal{H}}$. An explicit expression for this squared distance is

$$\sum_{j=1}^{k} \left\{ w_0 \sum_{s=1}^{S} \mathcal{E}\left[Z_s\{\mathcal{E}_s[\mathcal{S}_j] - \mathcal{E}_s[h_j] + \mathcal{E}[h_j]\}^2 \right] + \sum_{s=1}^{S} w_s \mathcal{E}_s \left[\{\mathcal{S}_j - \mathcal{E}_s[\mathcal{S}_j] - h_j + \mathcal{E}_s[h_j]\}^2 \right] \right\} \tag{18}$$

where h_j and $\mathcal{S}_j$ are the jth elements of h and $\mathcal{S}$ respectively. To obtain the projection, we must choose h_j to minimise the term in the braces in (18). Some algebra shows that this term may be written as

$$(h_j, Ah_j)_2 - 2(h_j, \phi_j)_2 + \sum_{s=1}^{S} (w_0 Q_{s0} - w_s) \mathcal{E}_s[\mathcal{S}_j]^2 + \sum_{s=1}^{S} w_s \mathcal{E}_s[\mathcal{S}_j^2] \tag{19}$$

where $Q_{s0} = \int Q(x, \beta_0) g_0(x)\, dx$ is the true value of Q_s, $(., .)_2$ is the inner product on $L_2(G_0)$, A is a self-adjoint nonnegative-definite operator on $L_2(G_0)$ defined by

$$Ah = Q^* \left\{ h - \sum_{r=1}^{S} \sum_{s=1}^{S} (\delta_{rs} - \gamma_{rs}) \frac{\int h(x) Q_r(x, \beta_0) g_0(x)\, dx}{Q_{r0}} P_s \right\}, \tag{20}$$

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

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

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

and

$$\phi_j(x) = \sum_{s=1}^{S} \frac{w_s}{Q_{s0}} Q_s(x, \beta_0) \frac{\partial \log Q_s(x, \beta)}{\partial \beta_j}|_{\beta=\beta_0} - \sum_{s=1}^{S} \sum_{r=1}^{S} Q^*(x) P_r(x)(\delta_{rs} - \gamma_{rs}) \mathcal{E}_s(\mathcal{S}_j). \tag{21}$$

As in Section 3, (19) is minimised when $h_j = h_j^*$, where h_j is a solution of $Ah_j = \phi_j$, which must be of the form

$$h_j^* = \frac{\phi_j}{f^*} + \sum_{r=1}^{S} c_{rj} P_r \tag{22}$$

for constants c_{rj} which satisfy the equations

$$c_{rj} - \sum_{s=1}^{S} \sum_{t=1}^{S} \frac{(\delta_{rs} - \gamma_{rs})}{w_s} v_{st} c_{tj} = \sum_{s=1}^{S} \frac{(\delta_{rs} - \gamma_{rs})}{w_s} d_{sj} \tag{23}$$

where $v_{rs} = \int P_r P_s Q^* dG_0$ and $d_{sj} = (P_s, \phi_j)_2$. Writing $\mathbf{\Gamma} = (\gamma_{rs})$, $\mathbf{C} = (c_{rj})$, $\mathbf{D} = (d_{rj})$, $\mathbf{W} = \text{diag}(w_1, \ldots, w_S)$ and $\mathbf{V} = (v_{rs})$, (23) can be expressed in matrix terms as

$$\mathbf{MC} = \mathbf{D} \tag{24}$$

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

Theorem 5. *The information bound* $\mathbf{B}$ *is given by*

$$\mathbf{B}^{-1} = \sum_{s=1}^{S} w_s \mathcal{E}_s[\mathcal{S}\mathcal{S}^T] + \sum_{s=1}^{S} (w_0 Q_{s0} - w_s) \mathcal{E}_s[\mathcal{S}] \mathcal{E}_s[\mathcal{S}]^T - \int \frac{\phi \phi^T}{Q^*} \, dG_0(x) - \mathbf{D}^T \mathbf{M}^- \mathbf{D}. \tag{25}$$

The proof is similar to that of Theorem 2 and is omitted.

4.3 Efficiency of the Scott-Wild estimator

Let $\hat{\theta} = (\hat{\beta}, \hat{\rho})$ be the solutions of the estimating equations (16). By Theorem 4, under suitable regularity conditions, $\hat{\theta}$ is asymptotically normal with asymptotic variance

$$\mathbf{I^*}^{-1} \mathbf{V} \mathbf{I^*}^{-1}$$

where $\mathbf{I}^*$ and $\mathbf{V}$ are as in Theorem 4. It turns out that the matrix $\mathbf{V}$ is of the form

$$\mathbf{V} = \mathbf{I}^* - \mathbf{I}^* \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0^T} & \mathbf{A} \end{pmatrix} \mathbf{I}^* \tag{26}$$

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

$$\mathbf{I}^{*-1} - \left(\begin{array}{cc} \mathbf{0} & \mathbf{0} \\ \mathbf{0^T} & \mathbf{A} \end{array} \right),$$

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

$$\mathbf{I}^*_{\beta\beta} - \mathbf{I}^*_{\beta\rho}(\mathbf{I}^*_{\rho\rho})^{-1}\mathbf{I}^*_{\rho\beta} \tag{27}$$

where $\mathbf{I}^*$ is partitioned as

$$\mathbf{I}^* = \left[\begin{array}{cc} \mathbf{I}^*_{\beta\beta} & \mathbf{I}^*_{\beta\rho} \\ \mathbf{I}^*_{\rho\beta} & \mathbf{I}^*_{\rho\rho} \end{array} \right].$$

To demonstrate the efficiency of $\hat{\beta}$, we must show that (27) and (25) coincide. To do this, we need a more explicit formula for $\mathbf{I}^*$. Let $\mathbf{S}$ be the $S \times k$ matrix with s, j element $\frac{\partial \log Q_s(x,\beta_j)}{\partial \beta}|_{\beta=\beta_0}$, let $\mathbf{E}$ be the $S \times k$ matrix with lth row $E_s = \mathcal{E}_s[\mathcal{S}]$, where $\mathcal{S} = \frac{\partial \log f(y|x,\beta)}{\partial \beta}|_{\beta=\beta_0}$. Also define

$$P^*_s(x, \beta, \rho) = \frac{\mu_s(\rho)Q_s(x,\beta)}{\sum_{r=1}^S \mu_r(\rho)Q_r(x,\beta)} \tag{28}$$

and note that $P_s(x) = P^*_s(x, \beta_0, \rho_0)$, where ρ_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 6.

1. *$\mathbf{I}^*_{\beta\beta} = \sum_{s=1}^S w_s\mathcal{E}_s\left[\mathcal{S}\mathcal{S}^T\right] - \int \mathbf{S}^T PP^T\mathbf{S}Q^*\, dG_0(x)$,*

2. *Let $\mathbf{U} = \mathbf{W}\mathbf{E} - \int PP^T\mathbf{S}Q^*\, dG_0(x)$. Then $\mathbf{I}^*_{\rho\beta} = \mathbf{A}^T\mathbf{U}_0$, where $\mathbf{U}_0$ consists of the first $S-1$ rows of $\mathbf{U}$ and $\mathbf{A}$ is a non-singular $(S-1) \times (S-1)$ matrix.*

3. *$\mathbf{I}^*_{\rho\rho} = \mathbf{A}^T\mathbf{M}_0\mathbf{A}$ where $\mathbf{M}_0$ consists of the first $S-1$ rows and columns of $\mathbf{M} = \mathbf{W}(\mathbf{I} - \mathbf{\Gamma})^{-1} - \mathbf{V}$.*

The proof is given in Section 5.5.

We now use theorems 4 and 5 to show that the efficiency bound (25) equals the asymptotic variance (27). Arguing as in Section 3, we get

$$\mathbf{B}^{-1} = \mathbf{I}^*_{\beta\beta} - \mathbf{I}^*_{\beta\rho}\mathbf{I}^{*}_{\rho\rho}{}^{-1}\mathbf{I}^*_{\rho\beta} + \Big\{ \sum_{s=1}^{S}(w_0Q_{s0} - w_s)E_sE_s^T + \mathbf{E}^T\mathbf{W}(\mathbf{I} - \mathbf{\Gamma})\mathbf{E} \Big\}. \tag{29}$$

We complete the argument by showing that the term in the braces in (29) is zero. We have

$$\begin{aligned} \mathbf{E}^T\mathbf{W}(\mathbf{I} - \mathbf{\Gamma})\mathbf{E}^T &= \sum_{s=1}^{S}(w_s - w_0Q_{s0})E_sE_s^T + w_0\left(\sum_{s=1}^{S} Q_{s0}E_s\right)\left(\sum_{s=1}^{S} Q_{s0}E_s\right)^T \\ &= \sum_{s=1}^{S}(w_s - w_0Q_{s0})E_sE_s^T \end{aligned}$$

since $\sum_{s=1}^{S} Q_{s0}E_s = 0$. Hence the term in the braces in (29) is zero, the asymptotic variance coincides with the information bound and so the Scott-Wild estimator has full semi-parametric efficiency.

5. Proofs

5.1 Proof of Theorem 1

The efficient score is the projection of $\dot{l}_\beta$ onto $\mathcal{T}_\eta^\perp$, so is of the form $\dot{l}_\beta - g$, where g is the unique minimizer of $||\dot{l}_\beta - g||_{\mathcal{H}}^2$ in $\mathcal{T}_\eta$. By (2), this is $\dot{l}_\beta - T(h^*)$, where h^* is the (unique) minimizer of $||\dot{l}_\beta - T(h)||_{\mathcal{H}}^2$ in $L_{2,k}(G_0)$. Write $h^* = (h_1^*, \ldots, h_k^*)$. Then

$$||\dot{l}_\beta - T(h^*)||_{\mathcal{H}}^2 = \sum_{l=1}^{k}\sum_{j=1}^{J} \frac{w_j}{\pi_j}\int(\dot{l}_{\beta,jl} - h_l^* - E_j[h_l^*])^2 f_j\, dG_0 \qquad (30)$$

so that we must choose h_l^* to minimize

$$\sum_{j=1}^{J} \frac{w_j}{\pi_j}\int(\dot{l}_{\beta,jl} - h_l^* - E_j[h_l^*])^2 f_j\, dG_0 = \sum_{j=1}^{J} w_j E_j[\dot{l}_{\beta,jl}^2] + (Ah_l^*, h_l^*)_2 - 2(\phi_l, h_l^*)_2. \qquad (31)$$

Now let h_l^* be any solution in $L_2(G_0)$ to (4). Then for any h in $L_2(G_0)$, using the fact that A is self-adjoint and positive definite, we get

$$\begin{aligned}\sum_{j=1}^{J} w_j E_j[\dot{l}_{\beta,jl}^2] + (Ah,h)_2 - 2(\phi_l,h)_2 &= \sum_{j=1}^{J} w_j E_j[\dot{l}_{\beta,jl}^2] - (Ah_l^*, h_l^*)_2 + (h - h_l^*, A(h-h_l^*))_2 \\ &\geq \sum_{j=1}^{J} w_j E_j(\dot{l}_{\beta,jl}^2) - (Ah_l, h_l^*)_2\end{aligned}$$

with equality if $h = h_l^*$, so that the efficient score has j,l element $S_{\beta,jl} - h_l^* + E_j[h_l^*]$ as asserted.

5.2 Proof of Theorem 2

The l, l' element of $\mathbf{B}^{-1}$ is

$$\begin{aligned}\sum_{j=1}^{J} w_j E_j[\dot{l}_{\beta,jl}^*\dot{l}_{\beta,jl'}^*] &= \sum_{j=1}^{J}\frac{w_j}{\pi_j}\int(\dot{l}_{\beta,jl} - h_l^* - E_j(h_l^*))(\dot{l}_{\beta,jl'} - h_{l'}^* - E_j(h_{l'}^*))f_j\, dG_0 \\ &= \sum_{j=1}^{J} w_j E_j[\dot{l}_{\beta,jl}\dot{l}_{\beta,jl'}] + (Ah_l^*, h_{l'}^*)_2 - (\phi_l, h_{l'}^*)_2 - (\phi_{l'}, h_l^*)_2 \\ &= \sum_{j=1}^{J} w_j E_j[\dot{l}_{\beta,jl}\dot{l}_{\beta,jl'}] - (\phi_l, h_{l'}^*)_2 \\ &= \sum_{j=1}^{J} w_j E_j[\dot{l}_{\beta,jl}\dot{l}_{\beta,jl'}] - \int\frac{\phi_l\phi_{l'}}{f^*}\, dG_0 - d_{(l)}^T\mathbf{M}^- d_{(l')}.\end{aligned}$$

5.3 Proof of Theorem 3

First, we note the formula

$$\frac{\partial^2 \log P_j^*}{\partial\theta\partial\theta^T} = \frac{\partial^2 P_j^*}{\partial\theta\partial\theta^T}\frac{1}{P_j^*} - \frac{\partial \log P_j^*}{\partial\theta}\frac{\partial \log P_j^*}{\partial\theta^T} \tag{32}$$

and the fact that

$$\begin{aligned}\sum_{j=1}^J w_j E_j\left[\frac{\partial^2 P_j^*}{\partial\theta\partial\theta^T}\frac{1}{P_j^*}\right] &= \sum_{j=1}^J \frac{w_j}{\pi_j}\int \frac{\partial^2 P_j^*}{\partial\theta\partial\theta^T}\frac{1}{P_j^*} f_j\, dG_0(x) \\ &= \sum_{j=1}^J \int \frac{\partial^2 P_j^*}{\partial\theta\partial\theta^T} f^*\, dG_0(x) \\ &= \frac{\partial^2}{\partial\theta\partial\theta^T}\int f^*\, dG_0(x) \\ &= 0,\end{aligned}$$

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

$$\begin{aligned}\mathbf{I}^* &= -\sum_{j=1}^J w_j E_j\left[\frac{\partial^2 P_j^*}{\partial\theta\partial\theta^T}\right] \\ &= \sum_{j=1}^J w_j E_j\left[\frac{\partial \log P_j^*}{\partial\theta}\frac{\partial \log P_j^*}{\partial\theta^T}\right].\end{aligned}$$

Next, we note the derivatives

$$\begin{aligned}\frac{\partial \log P_j^*(x,\beta,\rho)}{\partial\beta} &= S_j - \sum_{s=1}^J S_s P_s, \\ \frac{\partial \log P_j^*(x,\beta,\rho)}{\partial\rho_r} &= \delta_{j,r} - P_r,\end{aligned}$$

when the derivatives are evaluated at (β_0, ρ_0). Thus

$$\begin{aligned}\mathbf{I}^*_{\beta\beta} &= \sum_{j=1}^J w_j E_j\left[\frac{\partial \log P_j^*}{\partial\beta}\frac{\partial \log P_j^*}{\partial\beta^T}\right] \\ &= \sum_{j=1}^J \frac{w_j}{\pi_j}\int \left(S_j - \textstyle\sum_{s=1}^J S_s P_s\right)\left(S_j - \textstyle\sum_{s=1}^J S_s P_s\right)^T f_j(x)\, dG_0(x) \\ &= \sum_{j=1}^J w_j E_j[S_j S_j^T] - \int \left(\textstyle\sum_{s=1}^J S_s P_s\right)\left(\textstyle\sum_{s=1}^J S_s P_s\right)^T f^*(x)\, dG_0(x)\end{aligned}$$

$$= \sum_{j=1}^{J} w_j E_j[S_j S_j^T] - \int \mathbf{S}^T P P^T \mathbf{S} f^* \, dG_0(x)$$

which proves 1. Also

$$\begin{aligned}
\mathbf{I}^*_{\rho\beta,r} &= \sum_{j=1}^{J} w_j E_j \left[\frac{\partial \log P_j^*}{\partial \rho_r} \frac{\partial \log P_j^*}{\partial \beta} \right] \\
&= \sum_{j=1}^{J} \frac{w_j}{\pi_j} \int (\delta_{r,s} - P_r) \left(S_j - \textstyle\sum_{s=1}^{J} S_s P_s \right) f_j(x) \, dG_0(x) \\
&= w_r E_r[S_{rl}] - \int \left(\textstyle\sum_{j=1}^{J} S_j P_j \right) P_r f^*(x) \, dG_0(x)
\end{aligned}$$

which proves 2. Finally,

$$\begin{aligned}
\mathbf{I}^*_{\rho\rho,rs} &= \sum_{j=1}^{J} w_j E_j \left[\frac{\partial \log P_j^*}{\partial \rho_r} \frac{\partial \log P_j^*}{\partial \rho_s} \right] \\
&= \sum_{j=1}^{J} \frac{w_j}{\pi_j} \int (\delta_{jr} - P_r)(\delta_{js} - P_s) f_j(x) \, dG_0(x) \\
&= \int (\delta_{rs} - P_s) P_r f^*(x) \, dG_0(x) \\
&= \delta_{rs} w_r - v_{rs} \\
&= M_{rs}.
\end{aligned}$$

5.4 Proof of Theorem 4

Under the two-stage sampling scheme, the joint distribution of $\left\{ n_0^{(s)} \right\}$, $\left\{ n_1^{(s)} \right\}$ and $\{(x_{is}, y_{is}), i = 1, \ldots, n_1^{(s)}, s = 1, \ldots, S\}$ is (Wild 1991),

$$\prod_{s=1}^{S} P[n_1^{(s)} | n_0^{(s)}] \times \frac{n_0!}{n_0^{(1)}! \cdots n_0^{(S)}!} Q_1^{n_0^{(1)}} \cdots Q_S^{n_0^{(S)}} \times \prod_{s=1}^{S} \left\{ \prod_{i=1}^{n_1^{(s)}} I_s(x_{is}, y_{is}) f(y_{is} | x_{is}, \beta) g(x_{is}) \right\} / Q_s^{n_1^{(s)}}.$$

Thus, conditional on the $\left\{ n_0^{(s)} \right\}$ and $\left\{ n_1^{(s)} \right\}$, 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

$$p_s(x, y, \beta, g) = I_s(x, y) f(y | x, \beta) g(x) / Q_s.$$

Define

$$\psi_s^{(N)}(x,y,\theta) = \frac{\partial}{\partial\theta}\Big\{ \log f(y|x,\beta) - \log\Big[\sum_s \mu_s Q_s(x,\beta)\Big] - \log Q_s\Big\}, \qquad s = 1,\ldots,S,$$

and

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

Then the estimating equations (16) can be written in the form

$$\sum_{i=1}^{n_0}\psi_0^{(n_0)}(Z_{i1},\ldots,Z_{is},\theta) + \sum_{s=1}^{S}\sum_{i=1}^{n_0^{(s)}}\psi_s^{(n_0)}(x_{is},y_{is},\theta) = 0, \tag{33}$$

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

$$\psi_s(x,y,\theta) = \frac{\partial}{\partial\theta}\Big\{ \log f(y|x,\beta) - \log\Big[\sum_s \mu_s Q_s(x,\beta)\Big] - \log Q_s\Big\}, \qquad s = 1,\ldots,S,$$

and

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

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

Put

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

A standard Taylor expansion argument gives

$$\sqrt{N}(\hat\theta - \theta_0) = \left(-\frac{1}{N}\left.\frac{\partial S_N}{\partial\theta}\right|_{\theta=\theta_0}\right)^{-1}\frac{1}{\sqrt{N}}S(\theta_0) + \frac{1}{\sqrt{N}}\left(-\frac{1}{N}\left.\frac{\partial S_N}{\partial\theta}\right|_{\theta=\theta_0}\right)^{-1} R$$

where the jth element of R is

$$R_j = \frac{1}{2}(\hat\theta - \theta_0)^T \left.\frac{\partial^2_{Nj}}{\partial\theta\partial\theta^T}\right|_{\theta=\tilde\theta}(\hat\theta - \theta_0),$$

S_{Nj} is the jth element of S_N and $||\tilde\theta - \theta_0|| \le ||\hat\theta - \theta_0||$.

Consider first $S_N(\theta_0)/\sqrt{N}$. We have

$$\begin{aligned}\frac{S_N(\theta_0)}{\sqrt{N}} &= \sqrt{\frac{n_0}{N}}\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0}\{\psi_0^{(N)}(Z_{i1},\ldots,Z_{iS},\theta_0)-\mathcal{E}[\psi_0]\}\\ &\quad+\sum_{s=1}^{S}\sqrt{\frac{n_1^{(s)}}{N}}\frac{1}{\sqrt{n_1^{(s)}}}\sum_{i=1}^{n_1^{(s)}}\{\psi_s^{(N)}(x_{is},y_{is},\theta)-\mathcal{E}_s[\psi_s]\}\\ &\quad+\sqrt{N}\sum_{s=1}^{S}\left(\frac{n_0}{N}-w_0\right)\mathcal{E}[\psi_0]+\sqrt{N}\sum_{s=1}^{S}\left(\frac{n_1^{(s)}}{N}-w_s\right)\mathcal{E}_s[\psi_s].\end{aligned}$$

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

$$\begin{aligned}\frac{S(\theta_0)}{\sqrt{N}} &= \sqrt{w_0}\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0}\sum_{s=1}^{S}\{\psi_0^{(N)}(Z_{is},\theta_0)-\mathcal{E}[\psi_0]\}\\ &\quad+\sum_{s=1}^{S}\sqrt{w_s}\frac{1}{\sqrt{n_0^{(s)}}}\sum_{i=1}^{n_0^{(s)}}\{\psi_s^{(N)}(x_{is},y_{is},\theta)-\mathcal{E}_s[\psi_s]\}+o_p(1)\end{aligned}$$

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 ψ_s sufficiently quickly, we see that S_N is asymtotically normal with zero mean and asymptotic variance $\mathbf{V} = \sum_{s=0}^{S} w_s \mathrm{Var}\,\psi_s$.

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

$$E[e^{itS_N}] = \textstyle\sum_{(0)} E[e^{itS_N}|\{n_0^{(s)}\},\{n_1^{(s)}\}]P[\{n_0^{(s)}\},\{n_1^{(s)}\}] \tag{34}$$

where $\sum_{(0)}$ 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 $\{n_1^{(s)}\}$, (34) equals

$$E[e^{itS_N}] = \textstyle\sum_{(0)} E[e^{itS_N^{(1)}}E[e^{itS_N^{(2)}}|\{n_1^{(s)}\}]]P[\{n_0^{(s)}\},\{n_1^{(s)}\}]. \tag{35}$$

Let $\mathbf{V}_2 = \sum_{s=1}^{S} w_s Var[\psi_s]$. Assuming that the $\psi_s^{(N)}$ converge sufficiently quickly to the ψ_s , it follows that $E[e^{itS_N^{(2)}}|\{n_1^{(s)}\}] \to \exp\{-\frac{1}{2}t^T\mathbf{V}_2 t\}$, since the distribution of $S_N^{(2)}$ conditional on $\{n_0^{(s)}\}$ and $\{n_1^{(s)}\}$ is the same as that (unconditionally) under multi-sampling.

Now let ϵ be arbitrary, and let N_0 be such that

$$\left|E[e^{itS_N^{(2)}}|\{n_1^{(s)}\}] - \exp\{-\tfrac{1}{2}t^T\mathbf{V}_2 t\}\right| < \frac{\epsilon}{2}$$

whenever $n_1^{(s)} \geq N_0$ for $s = 1, \ldots, S$. Also, assume that the (random) sample sizes ultimately get large, in the sense that there exists N^* such that

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

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

$$\begin{aligned} E[e^{itS_N}] &= E[e^{itS_N^{(1)}}]\exp\{-\tfrac{1}{2}t^T\mathbf{V}_2 t\} + \textstyle\sum_{(1)} E[e^{itS_N^{(1)}}(E[e^{itS_N^{(2)}}|\{n_1^{(s)}\}] - \exp\{-\tfrac{1}{2}t^T\mathbf{V}_2 t\})] \\ &\quad + \textstyle\sum_{(2)} E[e^{itS_N^{(1)}}(E[e^{itS_N^{(2)}}|\{n_1^{(s)}\}] - \exp\{-\tfrac{1}{2}t^T\mathbf{V}_2 t\})] \end{aligned}$$

If $n_0 > N^*$, the sum of the second two terms is less than ϵ in absolute value, so

$$E[e^{itS_N}] = E[e^{itS_N^{(1)}}]\exp\{-\tfrac{1}{2}t^T\mathbf{V}_2 t\} + o(1).$$

Again by the same arguments as above, $[e^{itS_N^{(1)}}]$ converges to $\exp\{-\frac{1}{2}t^T\mathbf{V}_1 t\}$ where $\mathbf{V}_1$ is $w_0\mathrm{Var}[\psi_0(Z_1, \ldots, Z_S, \theta_0)]$ so that $E[e^{itS_N}]$ converges to $\exp\{-\frac{1}{2}t^T\mathbf{V}t\})]$, and hence S_N converges in distribution to a multivariate normal with variance $\mathbf{V} = \mathbf{V}_1 + \mathbf{V}_2$.

Assuming that $\hat{\theta}$ is $\sqrt{N}$-consistent, similar arguments show that $-\frac{1}{N}\frac{\partial S}{\partial\theta}|_{\theta=\theta_0}$ 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 $(\mathbf{I}^*)^{-1}\mathbf{V}(\mathbf{I}^*)^{-1}$.

5.5 Proof of Theorem 6

Let

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

From the definition of $\mathbf{I}^*$ in Theorem 4 and the law of large numbers, we get

$$\begin{aligned} \mathbf{I}^* &= -w_0\mathcal{E}\left[\sum_{s=1}^S Z_s \frac{\partial^2 \log Q_s}{\partial\theta\partial\theta^T}\right] - \sum_{s=1}^S w_s\mathcal{E}_s\left[\frac{\partial^2 \log P_s^\dagger}{\partial\theta\partial\theta^T} - \frac{\partial^2 \log Q_s\mu_s}{\partial\theta\partial\theta^T}\right] \\ &= \sum_{s=1}^S w_s\mathcal{E}_s\left[\frac{\partial \log P_s^\dagger}{\partial\theta}\frac{\partial \log P_s^\dagger}{\partial\theta^T}\right] - \sum_{s=1}^S w_s\mathcal{E}_s\left[\frac{1}{P_s^\dagger}\frac{\partial^2 P_s^\dagger}{\partial\theta\partial\theta^T}\right] \\ &\quad + \sum_{s=1}^S w_s\frac{\partial^2 \log Q_s\mu_s}{\partial\theta\partial\theta^T} - \sum_{s=1}^S w_0 Q_{s0}\frac{\partial^2 \log Q_s}{\partial\theta\partial\theta^T}. \end{aligned} \tag{36}$$

The second term of this expression is zero, since

$$\sum_{s=1}^S w_s\mathcal{E}_s\left[\frac{1}{P_s^\dagger}\frac{\partial^2 P_s^\dagger}{\partial\theta\partial\theta^T}\right] = \sum_{s=1}^S \int \frac{\partial^2}{\partial\theta\partial\theta^T}\int P_s^\dagger\, dyQ^*\, dG_0(x)$$

$$
\begin{aligned}
&= \sum_{s=1}^{S} \frac{\partial^2}{\partial\theta\partial\theta^T} \int P_s Q^* \, dG_0(x) \\
&= \frac{\partial^2}{\partial\theta\partial\theta^T} \int Q^* \, dG_0(x) \\
&= 0.
\end{aligned}
$$

Now we evaluate $\mathbf{I}^*_{\beta\beta}$. For the $\beta\beta$ submatrix, the third and fourth terms of (36) are zero. Thus, using the derivative

$$
\frac{\partial P_s^\dagger}{\partial\beta} = \mathcal{S} - \mathbf{S}^T P,
$$

we get

$$
\begin{aligned}
\mathbf{I}^*_{\beta\beta} &= \sum_{s=1}^{S} w_s \mathcal{E}_s \left[\frac{\partial \log P_s^\dagger}{\partial\beta} \frac{\partial \log P_s^\dagger}{\partial\beta^T} \right] \\
&= \sum_{s=1}^{S} \frac{w_s}{Q_{s0}} \int \int (\mathcal{S} - \mathbf{S}^T P)(\mathcal{S} - \mathbf{S}^T P)^T I_s(x,y) f(y|x,\beta_0) dy \, dG_0(x) \\
&= \sum_{s=1}^{S} \frac{w_s}{Q_{s0}} \int \int \mathcal{S}\mathcal{S}^T I_s(x,y) f(y|x,\beta_0) dy \, dG_0(x) - \int \mathbf{S}^T P (\mathbf{S}^T P)^T Q^*(x) \, dG_0(x) \\
&= \sum_{s=1}^{S} w_s \mathcal{E}_s [\mathcal{S}\mathcal{S}^T] - \int \mathbf{S}^T P P^T \mathbf{S} Q^* \, dG_0(x).
\end{aligned}
$$

which proves part 1.

Now, consider $\mathbf{I}^*_{\rho\beta,rj}$. Again, the third and fourth terms of (36) are zero. Introduce the parameters $\lambda_1, \ldots, \lambda_{S-1}$ defined by

$$
\lambda_r = \log(\mu_r(\rho)/\mu_S(\rho)), \; r = 1, \ldots, S-1.
$$

Then

$$
\begin{aligned}
\frac{\partial P_s^\dagger}{\partial\rho_r} &= \sum_{p=1}^{S-1} \frac{\partial\lambda_p}{\partial\rho_r} \frac{\partial P_s^\dagger}{\partial\lambda_p} \\
&= \sum_{p=1}^{S-1} \frac{\partial\lambda_p}{\partial\rho_r} (\delta_{sp} - P_p). \qquad (37)
\end{aligned}
$$

Thus

$$
\mathbf{I}^*_{\rho\beta,rj} = \sum_{s=1}^{S} w_s \mathcal{E}_s \left[\frac{\partial \log P_s^\dagger}{\partial\rho_r} \frac{\partial \log P_s^\dagger}{\partial\beta_j} \right]
$$

$$= \sum_{s=1}^{S} \frac{w_s}{Q_{s0}} \int\int [\sum_{p=1}^{S-1} \frac{\partial \lambda_p}{\partial \rho_r}(\delta_{sp} - P_p)](\mathcal{S} - \mathbf{S}P)_j I_s(x,y) f(y|x,\beta_0)\, dy\, dG_0(x)$$
$$= \sum_{p=1}^{S-1} \frac{\partial \lambda_p}{\partial \rho_r} u_{pj}$$

where

$$u_{pj} = \sum_{s=1}^{S} \frac{w_s}{Q_{s0}} \int\int (\delta_{ps} - P_p)(\mathcal{S} - \mathbf{S}P)_j I_s(x,y) f(y|x,\beta_0)\, dy\, dG_0(x).$$

Then, as in Theorem 3, we see that u_{pj} is the p,j element of $\mathbf{U}$, and so part 2 of the theorem is true with $\mathbf{A}_{pr} = \frac{\partial \lambda_p}{\partial \rho_r}$.

The $\rho\rho$ submatrix is

$$\mathbf{I}^*_{\rho\rho} = \sum_{s=1}^{S} w_s E_s \left[\frac{\partial \log P_s^\dagger}{\partial \rho} \frac{\partial \log P_s^\dagger}{\partial \rho^T} \right] - \sum_{s=1}^{S} w_0 Q_{s0} \frac{\partial^2 \log Q_s}{\partial \rho \partial \rho^T} + \sum_{s=1}^{S} w_s \frac{\partial^2 \log Q_s \mu_s}{\partial \rho \partial \rho^T}$$
$$= \sum_{s=1}^{S} w_s E_s \left[\frac{\partial \log P_s^\dagger}{\partial \rho} \frac{\partial \log P_s^\dagger}{\partial \rho^T} \right] - w_0 \sum_{s=1}^{S} \frac{1}{\kappa_s} \frac{\partial Q_s}{\partial \rho} \frac{\partial Q_s}{\partial \rho^T}$$

where $\kappa_s = Q_{s0} w_s / c_s$. It follows from (37) that $\mathbf{I}^*_{\rho\rho} = \mathbf{A}^T \mathbf{M}_0 \mathbf{A}$ where $\mathbf{M}_0$ has p,q element

$$\sum_{s=1}^{S} w_s E_s \left[\frac{\partial \log P_s^\dagger}{\partial \lambda_p} \frac{\partial \log P_s^\dagger}{\partial \lambda_q} \right] - w_0 \sum_{s=1}^{S} \frac{1}{\kappa_s} \frac{\partial Q_s}{\partial \lambda_p} \frac{\partial Q_s}{\partial \lambda_q}$$

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 $\mu_p = w_0 - c_p/Q_p$ give

$$\frac{\partial Q_p}{\partial \lambda_q} = \delta_{pq} \kappa_p - \frac{\kappa_p \kappa_q}{\kappa^*}$$

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

$$\sum_{s=1}^{s} \frac{1}{\kappa_s} \frac{\partial Q_s}{\partial \lambda_p} \frac{\partial Q_s}{\partial \lambda_q} = \frac{\partial Q_p}{\partial \lambda_q},$$

so that the p,q element of $\mathbf{M}_0$ is $\delta_{pq} w_p - v_{pq} - w_0 \frac{\partial Q_p}{\partial \lambda_q}$.

By the Sherman-Morrison formula, the p,q element of the matrix $\mathbf{W}(\mathbf{I}-\mathbf{\Gamma})^{-1} - \mathbf{W}$ is $-w_0 \frac{\partial Q_p}{\partial \lambda_q}$, so the matrix $\mathbf{M}_0$ consists of the first $S-1$ rows and columns of $\mathbf{W} - \mathbf{V} + \mathbf{W}(\mathbf{I}-\mathbf{\Gamma})^{-1} - \mathbf{W} = \mathbf{W}(\mathbf{I}-\mathbf{\Gamma})^{-1} - \mathbf{V} = \mathbf{M}$.

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.

Bickel, P.J., and Kwon, J. (2001) Inference for semi-parametric models: some questions and an answer. *Statistica Sinica*, **11**, 863–960.

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.

Lee, A. and Hirose, Y. (2007). Semi-parametric efficiency bounds for regression models under case-control sampling: the profile likelihood approach. Unpublished manuscript.

McNeney, B., and Wellner, J.A. (2003). Application of convolution theorems to semiparametric models with non-i.i.d. data. *J. Stat. Plan. Inf.*, **91**, 441–480.

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.

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.

Wild, C.J. (1991). Fitting prospective regression models to case-control data. *Biometrika*, **78**, 705–817.