Semi-parametric efficiency bounds for regression models under choice-based sampling

Alan Lee

Department of Statistics, University of Auckland

Private Bag 92019, Auckland, New Zealand

email: `lee@stat.auckland.ac.nz`

SUMMARY. We extend the Bickel–Klaassen–Ritov–Wellner theory of semi-parametric efficiency bounds to the case of sampling from several populations, and discuss the form of the efficient score and efficient influence function in this situation. The theory is applied to obtain an information bound for estimates of parameters in general regression models under case-control sampling. The variances of the semi-parametric estimates of Scott and Wild (1991, 1997, 2001) are compared to the bound and the estimates are found to be fully efficient.

KEY WORDS: Semi-parametric efficiency, choice-based sampling, case-control study, tangent space, influence function, efficient score, information bound.

January 8, 2004

1. INTRODUCTION

In this paper, we present a semi-parametric efficiency bound for the parameters of regression models fitted to data obtained by choice-based sampling. Previous authors have addressed this question using the theory of semi-parametric efficiency developed by Bickel *et al.* (1993). 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, assuming that the data are generated by Bernoulli sampling, where either a case or control is selected by a randomisation device with known selection probabilities, and the covariates of the resulting case or control are measured. The randomisation at the first stage means that the i.i.d. theory can be applied.

Breslow, McNeney and Wellner (2003) apply the missing value theory of Robins, Rotnitzky and Zhao (1994) and Robins, Hsieh and Newey (1995) to render the i.i.d. theory applicable. Here, individuals in the population are selected at random and their status (case or control) is determined. Then with a probability depending on their status, the covariates are measured or not. The unobserved covariates are treated as missing data.

We adopt a more direct approach. First, the Bickel–Klaassen–Ritov–Wellner theory is extended to the case of sampling from several populations. 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. The approach to estimation based on profile likelihood outlined by Scott and Wild (1991, 1997, 2001) is considered in Section 4 and found to be fully efficient.

2. INFORMATION BOUNDS FOR THE MULTI-SAMPLE CASE

2.1 Preliminaries

We first consider a direct sum of Hilbert spaces that will play an important role in the derivation of the information bound. Suppose $\mathcal{H}_1, \ldots, \mathcal{H}_J$ are Hilbert spaces with inner product $(\cdot,\cdot)_j$ on $\mathcal{H}_j$, and that $w_1, \ldots, w_J$ are positive constants. Then we may define an inner product $(\cdot,\cdot)$ on the direct sum $\mathcal{H} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_J$ by

$$\big((g_1, \ldots, g_J), (h_1, \ldots, h_J)\big) = \sum_{j=1}^{J} w_j (g_j, h_j)_j. \tag{1}$$

Specifically, the spaces we will consider will be of the form $L_{2,k}(P_1), \ldots, L_{2,k}(P_J)$, where $L_{2,k}(P_j)$ is the set of all k-dimensional functions that are square integrable with respect to a probability measure P_j. The inner product on $L_{2,k}(P_j)$ is

$$\begin{aligned} \big((a_1, \ldots, a_J), (b_1, \ldots, b_J)\big) &= \sum_{l=1}^{k} \int a_l(x) b_l(x)\, dP_j(x) & (2)\\ &= \sum_{l=1}^{k} (a_l, b_l)_{lj}, & (3) \end{aligned}$$

say. The following result will be useful:

Theorem 1 *Let $\mathcal{H} = L_{2,k}(P_1) \times \cdots \times L_{2,k}(P_J)$, and for $h = (h_1, \ldots, h_J) \in \mathcal{H}$, let $[h]$ denote the subspace of all functions of the form $(Ah_1, \ldots, Ah_J)$ for some constant $k \times k$ matrix A. If $g = (g_1, \ldots, g_J) \in \mathcal{H}$, then $g \perp [h]$ if and only if*

$$\sum_{j=1}^{J} w_j E_j(g_j h_j^T) = 0,$$

where E_j denotes expectation with respect to P_j.

Proof. Let $A = (a_{ij})$ be an arbitrary $k \times k$ matrix. Then

$$\begin{aligned}
(g, Ah) &= \sum_{j=1}^{J} w_j (g_j, Ah_j)_j \\
&= \sum_{j=1}^{J} w_j \sum_{i=1}^{k} (g_{ij}, \sum_{l=1}^{k} a_{il} h_{lj})_{lj} \\
&= \sum_{i=1}^{k} \sum_{l=1}^{k} a_{il} \sum_{j=1}^{J} w_j (g_{ij}, h_{lj})_{lj} \\
&= \sum_{i=1}^{k} \sum_{l=1}^{k} a_{il} \left\{ \sum_{j=1}^{J} w_j E_j (g_j h_j^T)_{il} \right\}
\end{aligned}$$

so that $g \perp [h]$ if and only if $\sum_{j=1}^{J} w_j E_j (g_j h_j^T) = 0$. □

2.2 The multi-sample model

Suppose for $j = 1, \ldots, J$ we observe independent random variables $X_{ij},\ i = 1, \ldots, n_j$, which for fixed j are identically distributed with density p_{0j}. The densities p_{0j} are members of classes of densities $\mathcal{P}_j$ of the form

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

where B is a finite dimensional set and N is infinite dimensional. We regard $\mathcal{P} = \mathcal{P}_1 \times \cdots \times \mathcal{P}_J$ as a model for our data. We will need to consider parametric submodels of $\mathcal{P}$; these are models of the form $\mathcal{Q} = \mathcal{Q}_1 \times \cdots \times \mathcal{Q}_J$ where

$$\mathcal{Q}_j = \{p_j(x; \beta, \gamma) : \beta \in B_0, \gamma \in \Gamma\},$$

B_0 is a subset of B and Γ has finite dimension.

We suppose that the family of densities $\mathcal{P}$ is regular, in the sense that for every finite-dimensional subfamily $\mathcal{Q}$ with $p_{j0} \in \mathcal{Q}_j$, the mapping from $B_0 \times \Gamma$ to $L_2(P_{0j})$ defined by

$$(\beta, \gamma) \to 2 \left(\sqrt{\frac{p_j(\cdot, \beta, \gamma)}{p_{0j}(\cdot)}} - 1 \right) I_{\{p_{0j} > 0\}}$$

is Fréchet differentiable for every p_j in $\mathcal{Q}_j$, $j = 1, \ldots, J$. This is sufficient to guarantee the existence of a square-integrable score function; see Bickel *et al.* (1993, Section 2.1) for details.

2.3 Tangent spaces

The tangent space $\mathcal{T}$ of the family $\mathcal{P}$ is the subspace of $\mathcal{H} = L_{2,k}(P_1) \times \cdots \times L_{2,k}(P_J)$ formed by taking the closure of the linear space of all elements of the form $(AS_1, \ldots, AS_J)$, where A is a constant matrix with k rows and $S = (S_1, \ldots, S_J)$ is the score function of a finite dimensional submodel of $\mathcal{P}$. We can also define the tangent spaces $\mathcal{T}_\beta$ and $\mathcal{T}_\eta$ corresponding to the families $\mathcal{P}_\beta$ and $\mathcal{P}_\eta$ defined by

$$\mathcal{P}_\beta = \{\big(p_1(\cdot, \beta, \eta_0), \ldots, p_J(\cdot, \beta, \eta_0)\big) : \beta \in B\}$$

and

$$\mathcal{P}_\eta = \{\big(p_1(\cdot, \beta_0, \eta), \ldots, p_J(\cdot, \beta_0, \eta)\big) : \eta \in N\}.$$

The space $\mathcal{T}_\eta$ is called the *nuisance tangent space*. We will require that $\mathcal{T} = \mathcal{T}_\beta + \mathcal{T}_\eta$; this will have to be established for the examples we consider.

2.4 RAL estimators and influence functions.

Let $n = \sum_{j=1}^J n_j$ and suppose that for each j, $n_j/n \to w_j$. An estimator $\hat\beta_n$ based on our data X_{ij} is *asymptotically linear* if

$$\sqrt{n}(\hat\beta_n - \beta_0) = n^{-1/2} \sum_{j=1}^J \sum_{i=1}^{n_j} \sqrt{w_j}\psi_j(X_{ij}) + o_p(1). \tag{4}$$

where $\psi = (\psi_1, \ldots, \psi_J)$ is in $\mathcal{H}$. The function ψ is called the *influence function* of the estimator.

As in the i.i.d. case, for a finite dimensional family we will say that an estimate is *regular* if $\sqrt{n}(\hat\beta_n - \beta_n)$ converges to the same distribution whenever $\sqrt{n}\big((\beta_n, \gamma_n) - (\beta_0, \gamma_0)\big)$ converges to a constant. Here the convergence is under the assumption that for a given n, the data on which $\hat\beta_n$ is based are distributed as $p_j(\cdot, \beta_n, \gamma_n)$.

An estimate is regular for an infinite dimensional family if it is regular for every finite-dimensional subfamily. We shall be concerned with estimates that are both regular and asymptotically linear, or RAL.

A key part of the theory of efficiency bounds in the i.i.d. case is a theorem that relates the influence function of a RAL estimate to the scores. Versions of this theorem may be found for example in Bickel *et al.* (1993, p 39, p 65) and Newey (1990, Theorem 2.2). We now extend this theorem to the multisample case.

Theorem 2 *Let $\mathcal{Q}$ be a finite-dimensional parametric family of densities with*

$$\mathcal{Q}_j = \{p_j(x; \beta, \gamma) : \beta \in B, \gamma \in \Gamma\},$$

and score function $S = (S_1, \ldots, S_J)$, where $S_j = (S_{\beta,j}, S_{\gamma,j})$. Suppose that $\hat\beta_n$ is a RAL estimator with influence function ψ. Then

$$\sum_{j=1}^J w_j E_j(\psi_j S_{\beta,j}^T) = I^{k\times k} \tag{5}$$

and

$$\sum_{j=1}^{J} w_j E_j(\psi_j S_{\gamma,j}^T) = O^{k\times r} \tag{6}$$

where r is the dimension of Γ.

Proof. For brevity, we write $\theta = (\beta, \gamma)$. We assume that the densities p_j are sufficiently well behaved so that the log-likelihood

$$l(\theta) = \sum_{j=1}^{J}\sum_{i=1}^{n_j} \log p_j(x_{ij}, \theta)$$

satisfies

$$l(\theta_0 + n^{-1/2}t) - l(\theta_0) = t^T \frac{\partial l}{\partial \theta} - \frac{1}{2} t^T I(\theta) t + o_p(1)$$

where

$$I(\theta) = \sum_{j=1}^{J} w_j E_j(S_j S_j^T).$$

Then

$$\left[\begin{array}{c} \sqrt{n}(\hat{\beta}_n - \beta_0) \\ l(\theta + n^{-1/2}t) - l(\theta) \end{array}\right] = \left[\begin{array}{c} 0 \\ -\frac{1}{2}t^T I(\theta) t \end{array}\right] + \sum_{j=1}^{J} n_j^{-1/2} \sum_{i=1}^{n_j} w_j^{1/2} \left[\begin{array}{c} \psi_j(x_{ij}) \\ t^T S_j \end{array}\right] + o_p(1)$$

which converges to

$$N\left(\left[\begin{array}{c} 0 \\ -\frac{1}{2}t^T I(\theta) t \end{array}\right], \left[\begin{array}{cc} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{array}\right]\right)$$

where

$$\begin{aligned} \Sigma_{11} &= \sum_{j=1}^{J} w_j E_j(\psi_j \psi_j^T), \\ \Sigma_{12} &= \sum_{j=1}^{J} w_j E_j(\psi_j t^T S_j), \\ \Sigma_{21} &= \Sigma_{12}^T, \\ \Sigma_{21} &= t^T I(\theta) t. \end{aligned}$$

Hence, by standard contiguity arguments,

$$\sqrt{n}(\hat{\beta}_n - \beta_n) \overset{\theta_n = \theta_0 + t/\sqrt{n}}{\longrightarrow} N(\Sigma_{12}, \Sigma_{11}). \tag{7}$$

Now write $t = (t_\beta, t_\gamma)$. Since $\hat{\beta}_n$ is regular, $\sqrt{n}(\hat{\beta}_n - \beta_0 - n^{-1/2}t_\beta)$ converges to the same limit no matter what the value of t When $t = 0$, the limit is $N(0, \Sigma_{11})$ by the asymptotic linearity, so that we must have

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \overset{\theta_n = \theta_0 + t/\sqrt{n}}{\longrightarrow} N(t_\beta, \Sigma_{11}). \tag{8}$$

Comparing (7) and (8), we see that $t_\beta = \Sigma_{12}$, or

$$t_\beta = \sum_{j=1}^{J} w_j E_j(\psi_j S_{\beta,j}^T) t_\beta + \sum_{j=1}^{J} w_j E_j(\psi_j S_{\gamma,j}^T) t_\gamma.$$

Since this is true for all t_β, t_γ, we must have (5) and (6). $\square$

Our next result extends this to infinite-dimensional models. We first need the concept of the efficient influence function.

2.5 Efficient influence functions and the information bound.

Now we return to the case where N may be infinite dimensional. For $j = 1, \ldots, J$, let $S_{j,\beta}$ denote the score $\frac{\partial p_j(\cdot,\beta,\eta)}{\partial \beta} / p_j(\cdot,\beta,\eta) I_{\{p_j>0\}}$, and let $S_\beta = (S_{1,\beta}, \ldots, S_{J,\beta})$. The *efficient score* is the element of $\mathcal{H}$ defined by

$$S^{eff} = S_\beta - \Pi(S_\beta | \mathcal{T}_\eta) = \Pi(S_\beta | \mathcal{T}_\eta^\perp)$$

where $\Pi(\cdot|\mathcal{T}_\eta)$ denotes the orthogonal projection in $\mathcal{H}$ onto the nuisance tangent space $\mathcal{T}_\eta$.

Let I_{eff} denote the matrix $\sum_{j=1}^{J} w_j E_j(S_j^{eff} {S_j^{eff}}^T)$. The element $(I_{eff}^{-1} S_1^{eff}, \ldots, I_{eff}^{-1} S_J^{eff})$ of $\mathcal{H}$ is called the *efficient influence function* and is denoted by ψ^{eff}. Our next result establishes the information bound.

Theorem 3 *Let $\mathcal{P}$ be as in Section 2.2 with N infinite dimensional, and suppose that $\mathcal{T} = \mathcal{T}_\beta + \mathcal{T}_\eta$, and that $\hat\beta_n$ is a RAL estimate with influence function ψ. Then $\psi - \psi^{eff} \perp \mathcal{T}$ and hence the matrix $n\,Var(\hat\beta_n) - I_{eff}^{-1}$ is positive definite.*

Proof. Let h be the score function for a finite-dimensional submodel $\mathcal{Q}$ of $\mathcal{P}$. Since $\mathcal{T} = \mathcal{T}_\beta + \mathcal{T}_\eta$, we can write $h = h_\beta + h_\gamma$, where $h_\beta \in \mathcal{T}_\beta$ and $h_\gamma \in \mathcal{T}_\eta$. Since $\mathcal{Q}$ is finite dimensional, h_γ must be the score function of a model of the form $\{\big((p_1(\beta_0,\gamma), \ldots, p_J(\beta_0,\gamma)\big) : \gamma \in \Gamma\}$ where Γ is finite-dimensional. In the rest of the proof we will make use of the finite dimensional model $\mathcal{Q}^* = \{\big((p_1(\beta,\gamma), \ldots, p_J(\beta,\gamma)\big) : \beta \in B, \gamma \in \Gamma\}$.

We first prove that $(\psi - \psi^{eff}, h_\beta) = 0$. Since $h_\beta \in \mathcal{T}_\beta = [S_\beta]$, by Theorem 1 it is enough to prove that

$$\sum_{j=1}^{J} w_j E_j[(\psi_j - \psi_j^{eff}) {S_{j,\beta}}^T] = 0. \tag{9}$$

By Theorem 2 applied to the submodel $\mathcal{Q}^*$, we get

$$\sum_{j=1}^{J} w_j E_j(\psi_j {S_{j,\beta}}^T) = I^{k\times k}. \tag{10}$$

Also,

$$\begin{aligned}
\sum_{j=1}^{J} w_j E_j(\psi_j^{eff} {S_{j,\beta}}^T) &= \sum_{j=1}^{J} w_j E_j[\psi_j^{eff} (I_{eff} \psi_j^{eff} + \Pi(S_\beta|\mathcal{T}_\eta))^T] \\
&= \big[\sum_{j=1}^{J} w_j E_j(\psi_j^{eff} {\psi_j^{eff}}^T)\big] I_{eff} + \sum_{j=1}^{J} w_j E_j[\psi_j^{eff} \Pi(S_\beta|\mathcal{T}_\eta)_j^T] \\
&= I^{k\times k}. \qquad (11)
\end{aligned}$$

The last line follows by Theorem 1 since $\Pi(S_\beta|\mathcal{T}_\eta)$ is perpendicular to $[S^{eff}]$. Combining (10) and (11) we get (9).

Very similar arguments applied to h_γ show that $(\psi - \psi^{eff}, h_\eta) = 0$, so that $(\psi - \psi^{eff}, h) = 0$. Since $\mathcal{T}$ is the closed linear span of elements such as h, it follows that $\psi - \psi^{eff} \perp \mathcal{T}$.

To prove the second part of the theorem, note that

$$\begin{aligned}
n \mathrm{Var}\, \hat{\beta}_n &= \sum_{j=1}^{J} w_j E_j(\psi_j \psi_j^T) \\
&= \sum_{j=1}^{J} w_j E_j(\psi_j^{eff} {\psi_j^{eff}}^T) + \sum_{j=1}^{J} w_j E_j[(\psi_j - \psi_j^{eff})(\psi_j - \psi_j^{eff})^T] \\
&= I_{eff}^{-1} + \sum_{j=1}^{J} w_j E_j[(\psi_j - \psi_j^{eff})(\psi_j - \psi_j^{eff})^T]
\end{aligned}$$

The cross product terms vanish by Theorem 1, since $\psi - \psi^{eff} \perp \mathcal{T}$ and $[\psi^{eff}] \subseteq \mathcal{T}$.

□

3. THE INFORMATION BOUND FOR 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 $y_1, \ldots, y_J$) and a vector X of covariates, and we want to model the conditional distribution of Y given X using a regression function

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

say, where β is a k-vector of parameters. If the distribution of the covariates X is specified by a density η, assumed to be absolutely continuous with respect to a measure μ, then the joint distribution of X and

Y is

$$f_j(x)\eta(x)$$

and the conditional distribution of x given $Y = y_j$ is

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

where

$$\pi_j = \int f_j(x, \beta)\eta(x)\, d\mu(x).$$

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

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

To apply the theory of Section 2, we must identify the spaces $\mathcal{T}$, $\mathcal{T}_\beta$ and $\mathcal{T}_\eta$.

Theorem 4 *Let* $\mathcal{P} = (\mathcal{P}_1 \times \cdots \times \mathcal{P}_J)$ *with* $\mathcal{P}_j = \{f_j(x, \beta)\eta(x)/\pi_j : \beta \in B, \eta \text{ a density}\}$*. Then*

(i) $\mathcal{T}_\beta = [S_\beta]$ *where* $S_\beta = (S_{\beta,1}, \ldots, S_{\beta,J})$*, and* $S_{\beta,j} = \frac{\partial \log f_j(x,\beta)}{\partial\beta} - E_j\left[\frac{\partial \log f_j(x,\beta)}{\partial\beta}\right]$.

(ii) The nuisance tangent space is $\mathcal{T}_\eta = \{\big(h - E_1(h), \ldots, h - E_J(h)\big) : h \in L^0_{2,k}(G_0)\}$*, where* $dG_0 = \eta_0 d\mu$*, and* $L^0_{2,k}(G_0)$ *is the space of all* k*-dimensional functions* f *satisfying* $\int ||f||^2\eta_0(x)\, d\mu(x) < \infty$ *and* $\int f(x)\eta_0(x)\, d\mu(x) = 0$.

(iii) The tangent space is $\mathcal{T} = \mathcal{T}_\beta + \mathcal{T}_\eta$.

Proof. Consider a finite dimensional submodel $\mathcal{Q}$ of $\mathcal{P}$, of the form

$$\mathcal{Q}_j = \{p_j(x, \gamma) = f_j(x, \beta(\gamma))\eta(x, \gamma)/\pi_j : \gamma \in \Gamma\}$$

where Γ has dimension r, say. The score function for $\mathcal{Q}$ is

$$\left[\frac{\partial p_j(x, \gamma)}{\partial\gamma} p_j\right] I_{\{p_j > 0\}}$$

which for simplicity we write as

$$\frac{\partial \log p_j(x, \gamma)}{\partial\gamma}.$$

Direct calculation gives

$$\frac{\partial \log p_j(x, \gamma)}{\partial\gamma} = \frac{\partial\beta}{\partial\gamma}\left[\frac{\partial \log f_j(x, \beta)}{\partial\beta} - E_j\left(\frac{\partial \log f_j(x, \beta)}{\partial\beta}\right)\right] + \frac{\partial \log \eta(x, \gamma)}{\partial\gamma} - E_j\left(\frac{\partial \log \eta(x, \gamma)}{\partial\gamma}\right)$$

so that for a constant matrix A with k rows,

$$A\frac{\partial \log p_j}{\partial \gamma} = A_1 S_{\beta,j} + A_2\left[\frac{\partial \log \eta(x,\gamma)}{\partial \gamma} - E_j\left(\frac{\partial \log \eta(x,\gamma)}{\partial \gamma}\right)\right]. \tag{12}$$

Now consider the spaces $\mathcal{T}_\beta$ and $\mathcal{T}_\eta$. To prove (i), note that $\mathcal{T}_\beta$ is the closure of the linear span of scores of finite-dimensional submodels of

$$\mathcal{P}_\beta = \left\{\big(p_1(x,\beta,\eta_0)\ldots,p_J(x,\beta,\eta_0)\big) :\ \beta \in B\right\},$$

so a calculation similar to (12) shows that $\mathcal{T}_\beta = [S_\beta]$.

Now define the operator $T_r : L^0_{2,r}(G_0) \to \mathcal{H}$ by

$$T_r(h) = \big(h_1 - E_1(h_1),\ldots,h_J - E_J(h_J)\big).$$

Again calculating as in (12), we see that $\mathcal{T}_\eta$ is the closure of the linear span, $\mathcal{T}^0_\eta$ say, of

$$\bigcup_r \{T_r(h) : h \text{ is the score of an } r\text{−dimensional submodel of } \mathcal{G}\}$$

where $\mathcal{G} = \{\eta : \eta \text{ a density for } X\}$. To prove (ii), we must show that

$$\overline{\mathcal{T}^0_\eta} = \{T_k(h) : h \in L^0_{2,k}(G_0)\}. \tag{13}$$

Let h be a score of a submodel of $\mathcal{G}$ having dimension r, say, and A an $k\times r$ matrix. Then $AT_r(h) = T_k(Ah)$ and Ah is in $L^0_{2,k}(G_0)$, so that

$$\mathcal{T}^0_\eta \subseteq \{T_k(h) : h \in L^0_{2,k}(G_0)\}. \tag{14}$$

Conversely, let h be in $L^0_{2,k}(G_0)$. Then using the arguments of Bickel *et al.* (1993, p 52), it follows that h is the score function of the k-dimensional submodel

$$\{\eta_0(x)(1 + \exp(-2\gamma^T h(x))^{-1} : \gamma \in \Re_k\}$$

so that the reverse inclusion in (14) is also true.

To complete the proof of (13), we show that $\{T_k(h) : h \in L^0_{2,k}(G_0)\}$ is closed. Let h_n be a sequence in $L^0_{2,k}(G_0)$ such that $T_k(h_n) \to g$ in $\mathcal{H}$. Note that by definition of the norm in $\mathcal{H}$, we have

$$||T_k(h_n) - g||^2_{\mathcal{H}} \geq \frac{w_j}{\pi_j}\int |h_n - E_j(h_n) - g_j|^2 f_j\, dG_0$$

so that $(h_n - E_j(h_n))f_j^{1/2} \to g_j f_j^{1/2}$ in $L^0_{2,k}(G_0)$, and hence, since the f_j's are bounded, $(h_n - E_j(h_n))b \to g_j b$ in $L^0_{2,k}(G_0)$, where $b = \prod_{j=1}^J f_j^{1/2}$. Subtracting, we have $(E_j(h_n) - E_J(h_n))b \to (g_j - g_J)b$ in $L^0_{2,k}(G_0)$, which implies that $E_j(h_n) - E_J(h_n)$ converges to a constant c_j say, and that $(g_j - g_J)b = c_j b$ a.e. G_0. Since $b \neq 0$, $g_j = g_J + c_j$. Moreover, $E_j(g_j) = 0,\ j = 1, \ldots J$, so that $c_j = -E_j(g_J)$, and $g = T_k(g_J)$. Finally, we show that g_j is in $L^0_{2,k}(G_0)$. The limit $g_j f_j^{1/2}$ and hence (since the f_j's are bounded) $g_j f_j$ is in $L^0_{2,k}(G_0)$, so

$$\begin{aligned}\sum_j g_j f_j &= \sum (g_J - c_j) f_j \\ &= g_J - \sum_j c_j f_j \quad (\text{since } \sum_j f_j = 1)\end{aligned}$$

is also in $L^0_{2,k}(G_0)$. Since $\sum_j c_j f_j$ is bounded and hence in $L^0_{2,k}(G_0)$, so must be g_J. Thus g is of the form $T_k(h)$ with $h \in L^0_{2,k}(G_0)$, which proves (ii).

To prove (iii), let $\mathcal{T}^0$ be the linear span of scores of finite-dimensional submodels of $\mathcal{P}$. From (12), we have $\mathcal{T}^0 \subseteq [S_\beta] + \mathcal{T}^0_\eta$, and the reverse inclusion is also true since $[S_\beta] \subseteq \mathcal{T}^0$ and $\mathcal{T}^0_\eta \subseteq \mathcal{T}^0$. Hence

$$\begin{aligned}\mathcal{T} &= \overline{\mathcal{T}^0} \\ &= \overline{[S_\beta] + \mathcal{T}^0_\eta} \\ &= [S_\beta] + \overline{\mathcal{T}^0_\eta} \quad (\text{since } [S_\beta] \text{ is finite-dimensional}) \\ &= \mathcal{T}_\beta + \mathcal{T}_\eta.\end{aligned}$$

□

Our next result derives the efficient score.

Theorem 5 *Let A be the operator $L_2(G_0) \to L_2(G_0)$ defined by*

$$(Ah)(x) = f^*(x)h(x) - \sum_{j=1}^J \frac{w_j}{\pi_j} f_j(x)(f_j/\pi_j, h)_2 \tag{15}$$

where

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

and $(\cdot,\cdot)_2$ is the inner product in $L_2(G_0)$. Let $S_{\beta,j} = (S_{\beta,j1}, \ldots, S_{\beta,jk})^T$ where $S_{\beta,jl} \in L_2(G_0)$, and define $\phi_l = \sum_{j=1}^J \frac{w_j}{\pi_j} S_{\beta,jl} f_j(x)$. Then the efficient score has j,l element

$$S_{\beta,jl} - h^*_l + E_j[h^*_l]$$

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

$$Ah^*_l = \phi_l. \tag{16}$$

Proof. The efficient score is the projection of S_β onto $\mathcal{T}_\eta^\perp$, so is of the form $S_\beta - g$, where g is the unique minimizer of $||S_\beta - g||^2_{\mathcal{H}}$ in $\mathcal{T}_\eta$. By Theorem 4, this is $S_\beta - T_k(h^*)$, where h^* is the (unique) minimizer of $||S_\beta - T_k(h)||^2_{\mathcal{H}}$ in $L^0_{2,k}(G_0)$. Write $h^* = (h^*_1, \ldots, h^*_k)$. Then

$$||S_\beta - T_k(h^*)||^2_{\mathcal{H}} = \sum_{l=1}^{k}\sum_{j=1}^{J} \frac{w_j}{\pi_j} \int (S_{\beta,jl} - h^*_l - E_j(h^*_l))^2 f_j \, dG_0 \tag{17}$$

so that we must choose h^*_l to minimize

$$\begin{aligned} \sum_{j=1}^{J} \frac{w_j}{\pi_j} \int (S_{\beta,jl} - h^*_l - E_j(h^*_l))^2 f_j \, dG_0 \quad &= \quad \sum_{j=1}^{J} w_j E_j(S^2_{\beta,jl}) + (Ah^*_l, h^*_l)_2 - 2(\phi_l, h^*_l)_2 \\ &= \quad I_{\beta\beta,ll} + (Ah^*_l, h^*_l)_2 - 2(\phi_l, h^*_l)_2 \end{aligned} \tag{18}$$

where $I_{\beta\beta}$ is the information matrix for the parametric part of the model.

Note that (18) is unaltered if we change h^*_l by constant function. Hence, minimising (18) over $L^0_2(G_0)$ is the same as minimising over $L_2(G_0)$. If h^*_l in $L_2(G_0)$ minimises (18), so does $h^*_l - E_0(h^*_l)$ in $L^0_2(G_0)$, where E_0 denotes expectation with respect to G_0.

Now we show that the operator A is self-adjoint and positive semi-definite, in the sense that $(Ah, h)_2 \geq 0$. First, for any h_1, h_2 in $L_2(G_0)$, we have

$$(h_1, Ah_2)_2 = \int h_1 h_2 f^* \, dG_0 - \sum_{j=1}^{J} w_j (h_1, f_j/\pi_j)_2 (h_2, f_j/\pi_j)_2$$

which is symmetric in h_1 and h_2. Thus A is self-adjoint.

To demonstrate that A is positive semi-definite, put $S_{\beta,jl} = 0$ in (18). Then

$$(Ah, h) = \sum_{j=1}^{J} \frac{w_j}{\pi_j} \int (h - E_j(h))^2 f_j \, dG_0 \geq 0.$$

Now let h^*_l be any solution in $L_2(G_0)$ to (16). Then for any h in $L_2(G_0)$, using the fact that A is self-adjoint,

$$\begin{aligned} I_{\beta\beta,ll} + (Ah, h)_2 - 2(\phi_l, h)_2 \quad &= \quad I_{\beta\beta,ll} - (Ah^*_l, h^*_l)_2 + \big(h - h^*_l, A(h - h^*_l)\big)_2 \\ &\geq \quad I_{\beta\beta,ll} - (Ah^*_l, h^*_l)_2 \end{aligned}$$

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. □

It remains to identify a solution to (16). Define $p_j = \frac{w_j}{\pi_j} f_j / f^*$ and $v_{jj'} = \int p_j p_{j'} f^* \, dG_0$. Let $V = (v_{jj'})$, $W = \mathrm{diag}(w_1, \ldots, w_J)$ and $M = W - V$. Note that the row sums of M are zero, since

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

Using these definitions and (15), 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 (16), we must have

$$f^* \left(\frac{\phi_l}{f^*} + \sum_{j=1}^{J} c_j p_j \right) - \sum_{j=1}^{J} \left(\frac{\phi_l}{f^*} + \sum_{j=1}^{J} c_j p_j, f_j/\pi_j \right)_2 p_j f^* = \phi_l,$$

or, equivalently,

$$\sum_{j=1}^{J} \left\{ c_j - \sum_{j'=1}^{J} c_{j'} (p_{j'}, f_j \pi_j)_2 - w_j^{-1} (\phi_l, p_j)_2 \right\} p_j f^* = 0. \tag{19}$$

Now

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

so that (19) will be satisfied if the vector $c = (c_1, \ldots, c_J)^T$ satisfies

$$Mc = d_{(l)} \tag{20}$$

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

Our final result in this section gives the information bound.

Theorem 6 *The variance - covariance matrix of the efficient score is I_{eff} where*

$$I_{eff,ll'} = I_{\beta\beta,ll'} - \int \frac{\phi_l \phi_{l'}}{f^*} \, dG_0 - d_{(l)}^T M^- d_{(l')}. \tag{21}$$

Proof.

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

□

4. EFFICIENCY OF THE SCOTT-WILD ESTIMATOR

In a famous paper, Prentice and Pyke (1979) showed that it is possible to estimate odds-ratio parameters from simple case-control studies by using an ordinary prospective regression program. In a series of papers (Scott and Wild 1991, 1997, 2001) Scott and Wild have generalised the classic Prentice-Pyke result to general regression models for a variety of choice-based sampling situations. In this section, we focus on general regression models for case control studies, and show that the Scott-Wild estimators are fully efficient. More general sampling situations will be considered in a forthcoming publication.

Suppose we sample prospectively n_1 cases and n_2 controls from their respective populations, and observe covariates $x_{1,1}, \ldots, x_{n_1,1}$ for the cases and $x_{1,2}, \ldots, x_{n_2,2}$ for the controls. Suppose, as in Section 3, that we have regression functions $f_j(x, \beta)$, $j = 1, 2$, giving the conditional probability that an individual with covariates x is a case ($j = 1$) or a control ($j = 2$). The unconditional distribution G_0 of the covariates is unspecified. As in Section 3, let π_1 and π_2 be the unconditional probabilities of being a case or control respectively.

Now let n_1 and n_2 converge to infinity with $n_j/(n_1+n_2) \to w_j$, $j = 1, 2$, and let $\kappa = (w_1/\pi_1)/(w_2/\pi_2)$. Put $\theta = (\beta, \kappa)^T$ and define $P_j^*(x, \theta)$ by

$$\operatorname{logit} P_j^*(x, \theta) = \operatorname{logit} f_j(x, \beta) + \log \kappa. \tag{22}$$

Then the Scott-Wild estimator of θ is the solution of the "pseudo-score" equations

$$\sum_{j=1}^{2} \sum_{i=1}^{n_j} \frac{\partial \log P_j^*(x_{ij}, \theta)}{\partial \theta} = 0.$$

Scott and Wild show that asymptotic variance of the estimate of β is the appropriate block of the inverse of the "pseudo" information matrix

$$I^*(\theta) = \sum_{j=1}^{2} w_j E_j \left(\Big(\frac{\partial \log P_j^*(x_{ij}}{\partial \theta} \Big) \Big(\frac{\partial \log P_j^*(x_{ij}}{\partial \theta} \Big)^T \right).$$

We now demonstrate that the inverse of this block coincides with the information bound in Theorem 6, thus showing that the Scott-Wild estimate is fully efficient. Using the partitioned matrix inverse formula, the inverse of the block is

$$I^{(1)} - I^{(2)} {I^{(2)}}^T / I^{(3)} \tag{23}$$

where

$$I^* = \begin{bmatrix} I^{(1)} & \kappa^{-1} I^{(2)} \\ \kappa^{-1} {I^{(2)}}^T & \kappa^{-2} I^{(3)} \end{bmatrix}.$$

Let S_j^0 denote the vector $\frac{\partial \log f_j}{\partial \beta}$. Then routine calculations give $P_j^*(x, \theta) = p_j$ and

$$\begin{aligned} I^{(1)} &= \int (S_1^0 - S_2^0)(S_1^0 - S_2^0)^T p_1 p_2 f^* \, dG_0, \\ I^{(2)} &= \int (S_1^0 - S_2^0) p_1 p_2 f^* \, dG_0, \\ I^{(3)} &= \int p_1 p_2 f^* \, dG_0. \end{aligned}$$

Now we evaluate the information bound I_{eff} using (21) in Section 3. We have

$$\begin{aligned} I_{\beta\beta,ll'} &= \sum_{j=1}^{2} \frac{w_j}{\pi_j} \int S_{\beta,jl} S_{\beta,jl'} f_j \, dG_0 \\ &= \sum_{j=1}^{2} \int S_{\beta,jl} S_{\beta,jl'} p_j f^* \, dG_0 \end{aligned} \tag{24}$$

and

$$\frac{\phi_l}{f^*} = S_{\beta,jl} p_1 + S_{\beta,2l} p_2$$

so that after some algebra we get

$$I_{\beta\beta,ll'} - \int \frac{\phi_l \phi_{l'}}{f^*} \, dG_0 = \int (S_{\beta,1l} - S_{\beta,2l})(S_{\beta,1l'} - S_{\beta,2l'}) p_1 p_2 f^* \, dG_0. \tag{25}$$

Now consider the matrix M. Since the row sums of M are zero, we can write M as

$$M = \begin{bmatrix} I^{(3)} & -I^{(3)} \\ -I^{(3)} & I^{(3)} \end{bmatrix}$$

so that a generalised inverse of M is

$$M^- = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} / I^{(3)}$$

and $d_l^T M^- d_{l'} = d_{1l} d_{1l'} / I^{(3)}$. Finally, we have $S_{\beta,1l} - S_{\beta,2l} = (S_1^0 - S_2^0)_l - K_l$ where $K_l = \big(E_1(S_1^0) - E_2(S_2^0)\big)_l$, so that

$$\begin{aligned} I_{\beta\beta,ll'} - \int \frac{\phi_l \phi_{l'}}{f^*} \, dG_0 &= \int \big((S_1^0 - S_2^0)_l - K_l\big)\big((S_1^0 - S_2^0)_{l'} - K_{l'}\big) p_1 p_2 f^* \, dG_0 \\ &= I_{ll'}^{(1)} - K_l I_{l'}^{(2)} - K_{l'} I_l^{(2)} + K_l K_{l'} I^{(3)}. \end{aligned} \tag{26}$$

Moreover, $d_{1l} = (\phi_l, p_1)_2 = \int (S_{\beta,1l} - S_{\beta,1l}) p_1 p_2 f^* \, dG_0$ so that

$$\begin{aligned} d_l^T M^- d_{l'} &= d_{1l} d_{1l'} / I_{(3)} \\ &= \int \big((S_1^0 - S2^0)_l - K_l \big) p_1 p_2 f^* \, dG_0 \times \int \big((S_1^0 - S2^0)_{l'} - K_{l'} \big) p_1 p_2 f^* \, dG_0 / I^{(3)} \\ &= (I_l^{(2)} - K_l I^{(3)})(I_{l'}^{(2)} - K_{l'} I^{(3)}) / I^{(3)}. \end{aligned} \tag{27}$$

Substituting (26) and (27) into (21) we see that

$$I_{eff,ll'} = I_{ll'}^{(1)} - I_l^{(2)} I_{l'}^{(2)} / I^{(3)}$$

so that the Scott-Wild estimator is fully efficient.

REFERENCES

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

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

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

Newey, W.K. (1990). Semiparametric efficiency bounds. *J. Appl. Econ.*, **5**, 99–135.

Prentice, R.L., and Pyke, R. (1997). Logistic disease incidence models and case-control studies. *Biometrika*, **66**, 403–411.

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

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

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

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.