M-ESTIMATORS IN SEMI-PARAMETRIC MULTI-SAMPLE MODELS

Yuichi Hirose

School of Mathematics, Statistics and Computer Science,

Victoria University of Wellington, New Zealand

April 18, 2007

Multi-sample data arises in response-selective sampling in Epidemiology and Econometrics. The standard M-estimation theory for independent and identically distributed (i.i.d.) data no longer applies since the data is independent but not identically distributed. This paper develops the theory of the M-estimator for multi-sample data. We present conditions for asymptotic linearity and show the efficiency bounds for M-estimators in multi-sample data.

Key words: Multi-sample; Semi-parametric model; Case-control study; Response-selective sampling; Two-phase, outcome-dependent sampling; Efficiency; M-estimator; Maximum likelihood estimator; Efficient information bound.

E-mail: Yuichi.Hirose@mcs.vuw.ac.nz

Telephone: 64 - 4 - 4636421

Fax: 64 - 4 - 4635045

1 Introduction

Suppose we have s-populations and, for each population $i = 1, \ldots, s$, we take random samples of size n_i. The resulting data, called multi-sample data, is independent but not identically distributed observations.

This type of data arises in many important situations in Epidemiology and Econometrics. For example, epidemiologists study the association between rare diseases and risk factors using a case-control study which is a response-selective sampling design. In such a case-control study, we collect two independent samples of predetermined sizes from the case and control populations. The risk factors or covariates are then measured for the sampled individuals. This type of response-selective sampling is more efficient for the study of rare outcomes than the more commonly used prospective sampling such as cohort or cross-sectional designs.

Many authors (e.g., Breslow and Chatterjee (1999), Breslow, Robins and Wellner (2000), Breslow, McNeney and Wellner (2003)) have used the single-sample version of large sample theory to establish asymptotic properties of the estimator in variations of case-control sampling which are multi-sample in nature. It is important to develop a multi-sample version of large sample theory for the further development of estimation in response-selective sampling.

McNeney and Wellner (2000) discussed the asymptotic theory of non-i.i.d. data. They used convolution theorems to establish the differentiability of the parameter of interest in semiparametric models. Their approach is very general and the theory includes many cases which we will not cover. In discussion, they concluded that the sufficient conditions for efficiency of M-estimators remains to be established.

The idea of multi-sample models which generate multi-sample data was introduced by Lee (2004). Lee, also, used convolution theorems to establish the efficiency bounds for estimators that are regular and asymptotic linear (RAL) in semiparametric multi-sample models.

This paper aims to present the conditions for asymtotic linearity of M-estimators in semiparametric models which generate multi-sample data. Under the conditions, we give a simple proof of asymptotic lower bounds for the variance of M-estimators in multi-sample models.

The structure of this paper is as follows. In section 2, we introduce a semiparametric model which generates multi-sample data. Examples of semiparametric models are given here. In section 3, we present conditions for an estimating function for the parameter of interest in multi-sample models. Then, we prove the asymptotic linearity of the M-estimator that corresponds to an estimating function. In section 4, we prove the asymptotic efficiency of the MLE among the class of M-estimators in multi-sample data. In Appendix B, we show that, for each multi-sample model, there is a corresponding i.i.d. model which shares important features relating to M-estimation: they share the log-likelihood function, the score function, the tangent space, estimating functions and influence functions.

2 Multi-sample models

The idea of multi-sample data is familiar from elementary statistics, for example, the well known two-sample t-test and one-way ANOVA for comparing several means involve multiple samples. These examples are familiar and the calculations involve similar reasoning and justification to the theory of multi-sample M-estimators which we will present in this paper.

Now, we define a multi-sample model. We consider an s-vector of semi-parametric models $(\mathcal{P}_1, \ldots, \mathcal{P}_s)$ where, for each $i = 1, \ldots, s$,

$$\mathcal{P}_i = \{p_i(x; \beta, \eta) : \ \beta \in \Theta_\beta \subset \mathbb{R}^m, \ \eta \in \Theta_\eta\}$$

is a probability model on the sample space $\mathcal{X}_i$ with an m-dimensional parameter of interest β and nuisance parameter η, which may be an infinite-dimensional parameter. Let (β_0, η_0) be the true value of (β, η). We assume Θ_β is a compact set containing an open neighborhood of β_0 in $\mathbb{R}^m$, and Θ_η is a convex set containing η_0 in a Banach space $\mathcal{B}$.

We observe s independent samples

$$X_{i1}, \ldots, X_{in_i}, \quad i = 1, \ldots, s,$$

where $X_{i1}, \ldots, X_{in_i}$ are independent and identically distributed (i.i.d.) according to the model $\mathcal{P}_i$. Let $n = \sum_{i=1}^s n_i$. We assume the sample size proportions $(\frac{n_1}{n}, \ldots, \frac{n_s}{n})$ converge to weight probabilities $(\lambda_1, \ldots, \lambda_s)$:

$$(\frac{n_1}{n}, \ldots, \frac{n_s}{n}) \to (\lambda_1, \ldots, \lambda_s) \tag{1}$$

where $\lambda_i > 0$ and $\sum_{i=1}^s \lambda_i = 1$.

The data $X_{i1}, \ldots, X_{in_i}$, $i = 1, \ldots, s$ are called *multi-sample data*, and the s-vector of models and the s-vector of weight probabilities $((\mathcal{P}_1, \ldots, \mathcal{P}_s), (\lambda_1, \ldots, \lambda_s))$ are called a *multi-sample model*. For ease of notation, we often omit an indication of the weight probabilities $(\lambda_1, \ldots, \lambda_s)$.

The expectation of a function $f(i, x)$ in the model $\mathcal{P}_i$ is denoted by

$$E_{i,\beta,\eta} f(i, X) = \int f(i, x)\, p_i(x; \beta, \eta) dx, \quad i = 1, \ldots, s.$$

2.1 Examples

Example 1. (Biased sampling model) Vardi (1985) developed the method of estimation in the s-sample biased sampling model with known selection bias weight functions. The following setup and notation are from Gill, Vardi and Wellner (1988).

Suppose that nonnegative weight functions $w_1(x), \ldots, w_s(x)$ are given and let $G(x)$ be an unknown distribution function on a sample space $\mathcal{X}$. Define the corresponding biased sampling model by

$$p_i(x; G) = \frac{w_i(x) g(x)}{W_i(G)} \quad i = 1, \ldots, s$$

where $g(x) = dG(x)/d\mu$ with respect to Lebesgue measure μ and $W_i(G) = \int_{\mathcal{X}} w_i(x)dG(x)$. The s-sample biased sampling model generates s independent samples

$$X_{i1}, \ldots, X_{in_i} \text{ iid } p_i(x; G), \quad i = 1, \ldots, s.$$

Gilbert, Lele, and Vardi (1999) considered an extension of this model which allows the weight function to depend on an unknown finite-dimensional parameter θ.

Suppose a set of nonnegative weight functions $w_1(x, \theta), \ldots, w_s(x, \theta)$ depend on θ. The semi-parametric biased sampling model is defined by

$$p_i(x; \theta, G) = \frac{w_i(x, \theta)g(x)}{W_i(\theta, G)} \quad i = 1, \ldots, s$$

where $W_i(\theta, G) = \int_{\mathcal{X}} w_i(x, \theta)dG(x)$. Gilbert (2000) provides a large sample theory of this example.

The following examples are semi-parametric multi-sample models which all have the same underlying data generating process on the sample space $\mathcal{Y} \times \mathcal{X}$, called the full data model,

$$\mathcal{Q} = \{p(y, x; \theta, G) = f(y|x; \theta)g(x) : \ \theta \in \Theta, \ G \in \mathcal{G}\}$$

where $f(y|x; \theta)$ is a conditional density of Y given X which depends on a finite dimensional parameter θ, $G(x)$ is an unspecified distribution function of X which is an infinite-dimensional nuisance parameter ($g(x)$ is the density of $G(x)$). We assume the set Θ is a compact set containing a neighborhood of the true value θ_0 and $\mathcal{G}$ is the set of all distribution functions of x. Unless stated otherwise Y may be a discrete or continuous variable.

Example 2. (Case-control study) We assume that Y takes values in $\{1, \ldots, s\}$. In a case-control study, due to the design, we do not observe a random sample from the full data model $\mathcal{Q}$. Instead, for each $i = 1, \ldots, s$, we observe n_i-samples from the conditional distribution $P(X|Y = i)$. By Bayes theorem, the density of $P(X|Y = i)$ is

$$\frac{f(i|x; \theta)g(x)}{\int f(i|x; \theta)dG(x)}.$$

The case-control study is a special case of the semiparametic biased sampling model of Example 1 with weight functions $w_i(x, \theta) = f(i|x; \theta)$, $i = 1, \ldots, s$.

Example 3. (Missing data) Instead of observing full data (Y, X) from the full data model $\mathcal{Q}$ for all individuals, we observe (Y, X) for n_0-samples and observe Y for n_1-samples. The result is the multi-sample data

$$(x_{01}, y_{01}), \ldots, (x_{0n_0}, y_{0n_0}), y_{11}, \ldots, y_{1n_1}$$

from a multi-sample model with

$$p_0(y, x; \theta, g) = f(y|x; \theta)g(x)$$

and

$$p_1(y;\theta,g) = \int f(y|x;\theta)g(x)dx.$$

This example is not a special case of Example 1.

Example 4. (Standard stratified sampling and two-phase, outcome-dependent sampling) For a partition of the sample space $\mathcal{Y} \times \mathcal{X} = \cup_{i=1}^{s} \mathcal{S}_i$, let

$$Q_i(\theta, G) = \int f(y|x;\theta)\, 1_{(y,x)\in\mathcal{S}_i}\, dy\, dG(x)$$

be the probability of (Y, X) belonging to stratum $\mathcal{S}_i$.

In standard stratified sampling, for each $i = 1, \ldots, s$, a random sample of size n_i is taken from the conditional distribution

$$p_i(y, x; \theta, G) = \frac{f(y|x;\theta)g(x)1_{(y,x)\in\mathcal{S}_i}}{Q_i(\theta, G)}$$

of (X, Y) given stratum $\mathcal{S}_i$. This is a slightly general version of the semiparametic biased sampling model of Example 1 with weight functions $w_i(y, x, \theta) = f(y|x;\theta)1_{(y,x)\in\mathcal{S}_i}$, $i = 1, \ldots, s$.

Lawless, Kalbfleish and Wild (1999) discussed variations of the two-phase, outcome-dependent sampling design (the variable probability sampling designs (VPS1,VPS2), the basic stratified sampling design (BSS)). For all sampling schemes (VPS1,VPS2, and BSS), we have m_i fully observed units and $n_i - m_i$ subjects where the only information retained is the identity of the stratum, $i = 1, \ldots, s$. The corresponding likelihood is

$$L(\theta, G) = \left\{ \prod_{i=1}^{s} \prod_{j=1}^{m_i} f(y_{ij}|x_{ij};\theta)g(x_{ij}) \right\} \left\{ \prod_{i=1}^{s} Q_i(\theta, G)^{n_i - m_i} \right\} \tag{2}$$

We interpret the observed data from two-phase, outcome-dependent sampling as data from multi-sample model with densities

$$p_1(y, x; \theta, G) = f(y|x;\theta)g(x)$$

and

$$p_2(i; \theta, G) = Q_i(\theta, G).$$

This example is, also, not a special case of Example 1.

3 Estimating functions

Motivation: The method of maximum likelihood estimation motivates defining a general estimating function in a multi-sample model.

Given multi-sample data $X_{i1}, \ldots, X_{in_i}$, $i = 1, \ldots, s$, the maximum likelihood estimator $\hat{\beta}_n$ is the maximizer of the likelihood

$$\prod_{i=1}^{s} \prod_{j=1}^{n_i} p_i(X_{ij}; \beta, \eta_0)$$

where we assume the nuisance parameter $\eta = \eta_0$ is known. The log-likelihood for the multi-sample data is

$$\ell_n(\beta, \eta) = \sum_{i=1}^{s} \sum_{j=1}^{n_i} \log\, p_i(X_{ij}; \beta, \eta).$$

We algebraically define the *log-likelihood function* for a single observation in the multi-sample model as

$$\ell(i, x, \beta, \eta) = \log\, p_i(x; \beta, \eta). \tag{3}$$

Usually, the maximum likelihood estimator is obtained by solving the maximum likelihood equation

$$\sum_{i=1}^{s} \sum_{j=1}^{n_i} \frac{\partial}{\partial \beta} \ell(i, X_{ij}, \beta, \eta_0) = 0.$$

In this case, the function

$$\psi(i, x, \beta, \eta_0) = \frac{\partial}{\partial \beta} \ell(i, x, \beta, \eta_0)$$

is an estimating function that corresponds to the maximum likelihood estimator.

Motivated by this example, in the following, we give the definition of an estimating function in multi-sample models.

Path-wise differentiability: A *path* in a convex subset $\mathcal{C}$ of a Banach space $\mathcal{B}$ is a continuously differentiable map $\eta(t) : \Theta_t \to \mathcal{C}$ where Θ_t is a closed interval in $\mathbb{R}$. The derivative of a path $\eta(t)$ is denoted by $\dot{\eta}(t)$. A map $f(\eta) : \mathcal{C} \to \mathbb{R}^m$ is *path-wise differentiable* with respect to η if there exists a bounded linear operator $d_\eta f(\eta)$, called the *derivative* of $f(\eta)$, such that, for each path $\eta(t)$ and $t \in \Theta_t$,

$$\frac{\partial}{\partial t} f(\eta(t)) = d_\eta f(\eta(t)) \dot{\eta}(t).$$

A norm of the derivative $d_\eta f(\eta)$ at η_0 is defined by

$$\|d_\eta f(\eta_0)\| = \sup \left\{ \frac{\|d_\eta f(\eta_0)\dot{\eta}(0)\|}{\|\dot{\eta}(0)\|} : \ \eta(t) \text{ path with } \eta(0) = \eta_0 \text{ and } \dot{\eta}(0) \neq 0 \right\}.$$

A map $f(\eta)$ is *continuously path-wise differentiable* with respect to η if the derivative $d_\eta f(\eta)$ is a continuous function of η.

Estimating function: A function $\psi : \{1, \ldots, s\} \times \mathcal{X} \times \Theta_\beta \times \Theta_\eta \to \mathbb{R}^m$ is an *estimating function* for β_0 in the presence of the nuisance parameter η in the multi-sample model $(\mathcal{P}_1, \ldots, \mathcal{P}_s)$ if:

(E1) $\psi(i, x, \beta, \eta)$ is continuously differentiable with respect to β for all $\beta \in \Theta_\beta$ and continuously path-wise differentiable with respect to η for all $\eta \in \Theta_\eta$;

(E2) $\sum_{i=1}^{s} \lambda_i E_{i,\beta,\eta}[\psi(i, X, \beta, \eta)] = 0$ for all $(\beta, \eta) \in \Theta_\beta \times \Theta_\eta$;

(E3) for $i = 1, \ldots, s$, $E_{i,\beta_0,\eta_0}[\psi(i, X, \beta_0, \eta_0)] = 0$;

(E4) $\sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0}[\psi^T \psi(i, X, \beta_0, \eta_0)] < \infty$;

(E5) $\sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0}[\frac{\partial}{\partial \beta} \psi(i, X, \beta_0, \eta_0)]$ is nonsingular with bounded inverse;

(E6) there is a compact and convex subset $\mathcal{C} \subset \Theta_\eta$ containing η_0 such that, for each $i = 1, \ldots, s$, the class of functions $\{\psi(i, x, \beta, \eta) : (\beta, \eta) \in \Theta_\beta \times \mathcal{C}\}$ is $P_{i,\beta_0\eta_0}$-Donsker with square integrable function and the class of functions $\{\frac{\partial}{\partial\beta}\psi(i, x, \beta, \eta) : (\beta, \eta) \in \Theta_\beta \times \mathcal{C}\}$ is P_{i,β_0,η_0}-Glivenko-Cantelli with integrable envelope function;

(E7) ψ and $\dot{\ell}_\eta$ are uncorrelated at (β_0, η_0):

$$\sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0} \psi \dot{\ell}_\eta(i, X, \beta_0, \eta_0) = 0$$

where $\dot{\ell}_\eta(i, x, \beta, \eta) = d_\eta \ell(i, x, \beta, \eta)$ is the derivative of Equation 3.

Remark 3.1: Suppose a function $\psi(i, x, \beta, \eta)$ satisfies conditions (E1)–(E7) except condition (E3). When $\sqrt{n}(\frac{n_i}{n} - \lambda_i) = O_P(1)$ (i.e. $\frac{n_i}{n}$ is a $\sqrt{n}$-consistent estimator of λ_i), $i = 1, \ldots, s$, then the function $\psi^c(i, x, \beta, \eta) = \psi(i, x, \beta, \eta) - E_{i,\beta_0,\eta_0}\psi(i, x, \beta_0, \eta_0)$ satisfies condition $(E3)$ and can be used for an estimating function.

In the next theorem, we prove the asymptotic linearity of an M-estimator.

THEOREM 3.1. *[An M-estimator is asymptotically linear] Suppose $\psi(s, x, \beta, \eta)$ satisfies conditions (E1)–(E7). Then, for any $\sqrt{n}$-consistent estimator $\hat{\eta}_n$ of η_0 (i.e. $\sqrt{n}(\hat{\eta}_n - \eta_0) = O_P(1)$), a consistent solution $\hat{\beta}_n$ to the estimating equation*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \psi(i, X_{ij}, \hat{\beta}_n, \hat{\eta}_n) = o_P(1)$$

is an asymptotically linear estimator with influence function

$$\tilde{\psi}(i, x, \beta, \eta) = \left[-\sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0} \left(\frac{\partial}{\partial\beta} \psi \right) \right]^{-1} \psi(i, x, \beta, \eta),$$

so that

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \tilde{\psi}(i, X_{ij}, \beta_0, \eta_0) + o_P(1)$$

and

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N\left(0, \sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0}(\tilde{\psi}\tilde{\psi}^T)\right).$$

PROOF.

By conditions (E2) and (E7), we have

$$\sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0} d_\eta \psi(i, X, \beta_0, \eta_0) = -\sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0} \psi \dot{\ell}_\eta(i, X, \beta_0, \eta_0) = 0 \quad \text{(zero operator)}.$$

Then, by the mean value theorem for vector valued function (cf. Hall and Newell (1979)) and Equation (1), for any $\sqrt{n}$-consistent estimator $\hat{\eta}_n$ of η_0,

$$\left\| \sqrt{n} \sum_{i=1}^{s} \frac{n_i}{n} E_{i,\beta_0,\eta_0} \psi(i, X, \beta_0, \hat{\eta}_n) \right\|$$

$$
\begin{aligned}
&\leq \sup_{t\in[0,1]} \left\| \sum_{i=1}^{s} \frac{n_i}{n} E_{i,\beta_0,\eta_0} d_\eta \psi(i, X, \beta_0, \eta_0 + t(\hat{\eta}_n - \eta_0)) \right\| \sqrt{n} \|\hat{\eta}_n - \eta_0\| \\
&= o_P(1) O_P(1) = o_P(1). \qquad (4)
\end{aligned}
$$

Suppose $(\beta_n^*, \eta_n^*) \xrightarrow{P} (\beta_0, \eta_0)$. Since the functions $\psi(i, x, \beta, \eta)$ and $\frac{\partial}{\partial\beta}\psi(i, x, \beta, \eta)$ are continuous at (β_0, η_0), and they are dominated by the square integrable function and the integrable function, respectively, by the dominated convergence theorem, we have, for $i = 1, \ldots, s$,

$$
E_{i,\beta_0,\eta_0} \|\psi(i, X, \beta_n^*, \eta_n^*) - \psi(i, X, \beta_0, \eta_0)\|^2 \xrightarrow{P} 0
$$

and

$$
E_{i,\beta_0,\eta_0} \left\| \frac{\partial}{\partial\beta}\psi(i, X, \beta_n^*, \eta_n^*) - \frac{\partial}{\partial\beta}\psi(i, X, \beta_0, \eta_0) \right\| \xrightarrow{P} 0.
$$

Together with condition (E6), these imply that

$$
\begin{aligned}
&\frac{1}{\sqrt{n_i}} \sum_{j=1}^{n_i} \{\psi(i, X_{ij}, \beta_n^*, \eta_n^*) - \psi(i, X_{ij}, \beta_0, \eta_0)\} \\
&= \sqrt{n_i} E_{i,\beta_0,\eta_0} \{\psi(i, X, \beta_n^*, \eta_n^*) - \psi(i, X, \beta_0, \eta_0)\} + o_P(1), \qquad (5)
\end{aligned}
$$

$$
\frac{1}{n_i} \sum_{j=1}^{n_i} \frac{\partial}{\partial\beta}\phi(i, X_{ij}, \beta_n^*, \eta_n^*) \xrightarrow{P} E_{i,\beta_0,\eta_0} \left(\frac{\partial}{\partial\beta}\psi(i, X, \beta_0, \eta_0) \right). \qquad (6)
$$

By combining Equations (5), (4) and condition (E3), we get

$$
\frac{1}{\sqrt{n}} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \psi(i, X_{ij}, \beta_n^*, \eta_n^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \psi(i, X_{ij}, \beta_0, \eta_0) + o_P(1) \qquad (7)
$$

Equations (1) and (6) imply

$$
\frac{1}{n} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \frac{\partial}{\partial\beta}\phi(i, X_{ij}, \beta_n^*, \eta_n^*) \xrightarrow{P} \sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0} \left(\frac{\partial}{\partial\beta}\psi(i, X, \beta_0, \eta_0) \right). \qquad (8)
$$

By the central limit theorem (CLT) in the model $\mathcal{P}_i$, $\frac{1}{\sqrt{n_i}} \sum_{j=1}^{n_i} \psi(i, X_{ij}, \beta_0, \eta_0) \xrightarrow{d} \Phi_i \sim N(0, E_{i,\beta_0,\eta_0}(\psi\psi^T))$. By Equation (1) and by independence of samples, we have a multi-sample version of the CLT

$$
\begin{aligned}
\frac{1}{\sqrt{n}} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \psi(i, X_{ij}, \beta_0, \eta_0) &= \sum_{i=1}^{s} \frac{\sqrt{n_i}}{\sqrt{n}} \frac{1}{\sqrt{n_i}} \sum_{j=1}^{n_i} \psi(i, X_{ij}, \beta_0, \eta_0) \\
&\xrightarrow{d} \sum_{i=1}^{s} \sqrt{\lambda_i}\Phi_i \sim N\left(0, \sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0}(\psi\psi^T)\right).
\end{aligned}
$$

This implies

$$
\frac{1}{\sqrt{n}} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \psi(i, X_{ij}, \beta_0, \eta_0) = O_P(1). \qquad (9)
$$

For $k = 1, \ldots, m$, let ψ_k be the kth coordinate function of $\psi = (\psi_1, \ldots, \psi_m)^T$. By the usual mean value theorem and Equations (7) and (8), for some β_n^* with $\|\beta_n^* - \beta_0\| \leq \|\hat{\beta}_n - \beta_0\| \xrightarrow{P} 0$,

$$\begin{aligned}
o_P(1) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \psi_k(i, X_{ij}, \hat{\beta}_n, \hat{\eta}_n) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \psi_k(i, X_{ij}, \beta_0, \hat{\eta}_n) + \frac{1}{n} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \frac{\partial}{\partial \beta} \psi_k(i, X_{ij}, \beta_n^*, \hat{\eta}_n) \sqrt{n}(\hat{\beta}_n - \beta_0) \\
&= \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \psi_k(i, X_{ij}, \beta_0, \eta_0) + o_P(1) \right\} \\
&\quad + \left\{ \sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0} \left(\frac{\partial}{\partial \beta} \psi_k(i, X, \beta_0, \eta_0) \right) + o_P(1) \right\} \sqrt{n}(\hat{\beta}_n - \beta_0), \ k = 1, \ldots, m.
\end{aligned}$$

By condition (E5), these equations implies

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_n - \beta_0) &= \left\{ \sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0} \left(\frac{\partial}{\partial \beta} \psi(i, X, \beta_0, \eta_0) \right) \right\}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \psi(i, X_{ij}, \beta_0, \eta_0) \\
&\quad + o_P(1)\{1 + \sqrt{n}(\hat{\beta}_n - \beta_0)\}. \qquad (10)
\end{aligned}$$

Equations (9) and (10) imply that

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = O_P(1) + o_P(1)\{1 + \sqrt{n}(\hat{\beta}_n - \beta_0)\}.$$

By rearranging this equation we have a $\sqrt{n}$-consistency: $\sqrt{n}(\hat{\beta}_n - \beta_0) = O_P(1)$. Finally, Equation (10) gives the result. □

4 Efficiency and the geometry of influence functions

By adapting a geometric argument of the space of influence functions in Chapter 3, Tsiatis (2006), we prove the efficiency of the MLE for a semiparametric multi-sample model. The efficiency of the best linear unbiased estimator (BLUE) in the linear model can be proved using the same geometric argument (cf. THEOREM 3.2 in Seber and Lee (2003)).

4.1 The efficient score function $\dot{\ell}_\beta^*$

Here, we give the definition of the efficient score function by using an uncorrelated projection.

We assume the log-likelihood function for a single observation $\ell(i, x, \beta, \eta)$ (defined by Equation 3) is twice continuously differentiable with respect to β for all $\beta \in \Theta_\beta$ and continuously path-wise differentiable with respect to η for all $\eta \in \Theta_\eta$. The *score function* $\dot{\ell}_\beta(i, x, \beta, \eta)$ for β and the *score operator* $\dot{\ell}_\eta(i, x, \beta, \eta)$ for η in the multi-sample model are the derivative of the log-likelihood function with respect to β and η, respectively, i.e., $\dot{\ell}_\beta(i, x, \beta, \eta) = \frac{\partial}{\partial \beta} \ell(i, x, \beta, \eta)$ and $\dot{\ell}_\eta(i, x, \beta, \eta) = d_\eta \ell(i, x, \beta, \eta)$.

The *tangent space* $\dot{\mathcal{P}}_\eta$ for η is defined by the closed linear span of the set of derivatives

$$\bigcup \left\{ \frac{\partial}{\partial t}\Big|_{t=0} \ell(i, x, \beta_0, \eta(t)) \right\}$$

where the union is taken over all paths $\eta(t)$ with $\eta(0) = \eta_0$. The *product tangent space* $\dot{\mathcal{P}}_\eta^m$ for η in the estimation of m-dimensional parameter β_0 is the m-Cartesian product of the tangent space $\dot{\mathcal{P}}_\eta$, i.e., $\dot{\mathcal{P}}_\eta^m = \dot{\mathcal{P}}_\eta \times \ldots \times \dot{\mathcal{P}}_\eta$. Similarly, the tangent space $\dot{\mathcal{P}}_\beta$ for β is a closed linear span of the score function $\dot{\ell}_\beta(i, x, \beta, \eta)$ at (β_0, η_0), i.e., $\dot{\mathcal{P}}_\beta = [\dot{\ell}_\beta] = \{a^T \dot{\ell}_\beta : \ a \in \mathbb{R}^m\}$. The product tangent space $\dot{\mathcal{P}}_\beta^m$ for β is the m-Cartesian product of the tangent space $\dot{\mathcal{P}}_\beta$ for β, i.e., $\dot{\mathcal{P}}_\beta^m = \{A\dot{\ell}_\beta : \ A \in \mathbb{R}^{m\times m}\}$.

The uncorrelated complement of the score function $\dot{\ell}_\beta$ with respect to $\dot{\mathcal{P}}_\eta^m$,

$$\dot{\ell}_\beta^* = \dot{\ell}_\beta - \Pi(\dot{\ell}_\beta | \dot{\mathcal{P}}_\eta^m)$$

is called the *efficient score function* in the multi-sample model $(\mathcal{P}_1, \ldots, \mathcal{P}_S)$ (cf. see Appendix A for the definition of uncorrelated projection). We assume the efficient score function $\dot{\ell}_\beta^*$ satisfies conditions (E1)–(E7). By THEOREM 3.1, for any $\sqrt{n}$-consistent estimator $\hat{\eta}_n$ of η_0, if the solution $\hat{\beta}_n$ to the estimating equation

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \dot{\ell}_\beta^*(i, X_{ij}, \hat{\beta}_n, \hat{\eta}_n) = o_P(1)$$

is consistent, then it is an asymptotically linear estimator with the influence function

$$\tilde{\ell}_\beta^* = \left[-\sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0} \left(\frac{\partial}{\partial \beta} \dot{\ell}_\beta^* \right) \right]^{-1} \dot{\ell}_\beta^* = \left[\sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0} (\dot{\ell}_\beta^* \dot{\ell}_\beta^{*T}) \right]^{-1} \dot{\ell}_\beta^*$$

(called the *efficient influence function*) with its variance

$$\sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0} (\tilde{\ell}_\beta^* \tilde{\ell}_\beta^{*T}) = \left[\sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0} (\dot{\ell}_\beta^* \dot{\ell}_\beta^{*T}) \right]^{-1}$$

(called the *efficient information bound*).

4.2 Efficiency in the multi-sample semi-parametric model

We establish the efficient information bound in the multi-sample model.

Let Ψ_β be the set of all estimating functions for β_0 which satisfy conditions E(1)–E(7) in Section 3. Then, as a consequence of THEOREM 3.1, the set of influence functions corresponding to the set of estimating functions is

$$\tilde{\Psi}_\beta = \left\{ \tilde{\psi}(i, x, \beta_0, \eta_0) = \left[-\sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0} \left(\frac{\partial}{\partial \beta} \psi \right) \right]^{-1} \psi(i, x, \beta_0, \eta_0) : \ \psi \in \Psi_\beta \right\}.$$

We show that $\tilde{\ell}_\beta^*$ is the element with the smallest variance in the set $\tilde{\Psi}_\beta$ of influence functions. This demonstrates that the corresponding M-estimator (MLE) is the most efficient with the

asymptotic variance $\sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0}(\dot{\ell}^*_\beta \dot{\ell}^{*T}_\beta)^{-1}$ called the *efficient information bound* for β_0 in the multi-sample model.

Proof: Suppose $\psi(i, x, \beta, \eta)$ is an estimating function which satisfies conditions (E1)–(E7) and

$$\tilde{\psi}(i, x, \beta_0, \eta_0) = \left[-\sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0} \left(\frac{\partial \psi}{\partial \beta} \right) \right]^{-1} \psi(i, x, \beta_0, \eta_0)$$

is the corresponding influence function.

By differentiating the equation in condition (E2) with respect to β,

$$\begin{aligned} 0 &= \frac{\partial}{\partial \beta} \sum_{i=1}^{s} \lambda_i E_{i,\beta,\eta}[\psi(i, X, \beta, \eta)] \\ &= \frac{\partial}{\partial \beta} \sum_{i=1}^{s} \lambda_i \int \psi(i, x, \beta, \eta) p_i(x; \beta, \eta) dx \\ &= \sum_{i=1}^{s} \lambda_i E_{i,\beta,\eta} \left(\frac{\partial \psi}{\partial \beta} \right) + \sum_{i=1}^{s} \lambda_i E_{i,\beta,\eta}(\psi \dot{\ell}^T_\beta). \end{aligned}$$

This implies, at (β_0, η_0),

$$\sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0}(\psi \dot{\ell}^T_\beta) = -\sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0} \left(\frac{\partial \psi}{\partial \beta} \right).$$

As a result,

$$\begin{aligned} \sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0}(\tilde{\psi} \dot{\ell}^T_\beta) &= \left[-\sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0} \left(\frac{\partial \psi}{\partial \beta} \right) \right]^{-1} \sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0}(\psi \dot{\ell}^T_\beta) \\ &= \left[-\sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0} \left(\frac{\partial \psi}{\partial \beta} \right) \right]^{-1} \left[-\sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0} \left(\frac{\partial \psi}{\partial \beta} \right) \right] \\ &= I^{m \times m} \quad (m \times m \text{ identity matrix}). \end{aligned}$$

Similarly, since we assumed the efficient score function $\dot{\ell}^*_\beta$ satisfies conditions (E1)–(E7), we have

$$\sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0}(\tilde{\ell}^*_\beta \dot{\ell}^T_\beta) = I^{m \times m}.$$

Therefore, $\sum_{s=1}^{s} \lambda_i E_{i,\beta_0,\eta_0}(\tilde{\psi} \dot{\ell}^T_\beta) = \sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0}(\tilde{\ell}^*_\beta \dot{\ell}^T_\beta)$. This implies

$$\begin{aligned} 0 &= \sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0}([\tilde{\psi} - \tilde{\ell}^*_\beta] \dot{\ell}^T_\beta) \\ &= \sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0}([\tilde{\psi} - \tilde{\ell}^*_\beta][\dot{\ell}^*_\beta + \Pi(\dot{\ell}_\beta | \dot{\mathcal{P}}^m_\eta)]^T) \\ &= \sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0}([\tilde{\psi} - \tilde{\ell}^*_\beta] \dot{\ell}^{*T}_\beta) \quad (\text{ by condition (E7)}) \\ &= \sum_{i=1}^{s} \lambda_i E_{i,\beta_0,\eta_0}([\tilde{\psi} - \tilde{\ell}^*_\beta] \tilde{\ell}^{*T}_\beta). \end{aligned}$$

Then, for all $\tilde{\psi}$,

$$\begin{aligned}\sum_{i=1}^{s}\lambda_i E_{i,\beta_0,\eta_0}(\tilde{\psi}\tilde{\psi}^T) &= \sum_{i=1}^{s}\lambda_i E_{i,\beta_0,\eta_0}(\{\tilde{\ell}^*_\beta+[\tilde{\psi}-\tilde{\ell}^*_\beta]\}\{\tilde{\ell}^*_\beta+[\tilde{\psi}-\tilde{\ell}^*_\beta]\}^T)\\ &= \sum_{i=1}^{s}\lambda_i E_{i,\beta_0,\eta_0}(\dot{\ell}^*_\beta\dot{\ell}^{*T}_\beta)+\sum_{i=1}^{s}\lambda_i E_{i,\beta_0,\eta_0}([\tilde{\psi}-\tilde{\ell}^*_\beta][\tilde{\psi}-\tilde{\ell}^*_\beta]^T)\\ &\geq \sum_{i=1}^{s}\lambda_i E_{i,\beta_0,\eta_0}(\dot{\ell}^*_\beta\dot{\ell}^{*T}_\beta)\end{aligned}$$

since $\sum_{i=1}^{s}\lambda_i E_{i,\beta_0,\eta_0}[\tilde{\psi}-\tilde{\ell}^*_\beta][\tilde{\psi}-\tilde{\ell}^*_\beta]^T$ is a positive definite matrix (for $m\times m$ matrices A and B, $B\geq A$ if and only if $B-A$ is a nonnegative definite matrix).

5 Conclusion

We have presented conditions for asymptotic linearity of M-estimators in semiparametric multi-sample models and proved asymptotic efficiency bound for M-estimators under these conditions. This is a natural extension of the theory of M-estimators for i.i.d. models and is a special case of the situation in McNeney and Wellner (2000). It remains to establish efficiency theory for M-estimators in the general situation of McNeney and Wellner (2000).

Appendix A: Product Hilbert space and uncorrelated projection

In Appendix A, we define the product Hilbert space, covariance operators and an uncorrelated projection.

Product Hilbert space and covariance operator

Let $\mathcal{H}$ be the Hilbert space of measurable functions with zero mean and finite variance:

$$\mathcal{H}=\left\{f(i,x):\ E_{i,\beta_0,\eta_0}f=0, i=1,\ldots,s,\ \sum_{i=1}^{s}\lambda_i E_{i,\beta_0,\eta_0}f^2<\infty\right\}.$$

The inner-product of $f,g\in\mathcal{H}$ is defined by $\langle f,g\rangle=\sum_{i=1}^{s}\lambda_i E_{i,\beta_0,\eta_0}(fg)$. The *product Hilbert space* for the m-dimensional parameter β_0 is an m-Cartesian product of the Hilbert space $\mathcal{H}$ of measurable functions, i.e., $\mathcal{H}^m=\mathcal{H}\times\cdots\times\mathcal{H}$. The product Hilbert space can be considered as a space of all estimating functions $\psi(i,x,\beta_0,\eta_0)$ for β_0 at the true value (β_0,η_0).

For an arbitrary Banach space $\mathcal{B}$, let $\mathcal{B}^*$ be its dual. Let $A:\mathcal{B}\to\mathcal{H}$ be a bounded linear operator and $h\in\mathcal{H}$. The *adjoint operator* $A^T:\mathcal{H}\to\mathcal{B}^*$ of $A:\mathcal{B}\to\mathcal{H}$ is defined by the map

$$(A^Th)a=\langle Aa,h\rangle=\sum_{i=1}^{s}\lambda_i E_{i,\beta_0,\eta_0}((Aa)h),\ a\in\mathcal{B}.$$

Let $\mathcal{A}$ and $\mathcal{B}$ be two arbitrary Banach spaces. For two bounded linear operators $A:\mathcal{A}\to\mathcal{H}$ and $B:\mathcal{B}\to\mathcal{H}$, the *covariance operator* $\mathrm{Cov}(A,B):\mathcal{A}\to\mathcal{B}^*$ is the map defined by $\mathrm{Cov}(A,B)=$

$B^T A$,i.e.,

$$\mathrm{Cov}(A,B)ab = (B^T Aa)(b) = \langle Aa, Bb\rangle,\ a \in \mathcal{A},\ b \in \mathcal{B}.$$

Two operators A and B are *uncorrelated*, denoted by $A \perp B$, if $\mathrm{Cov}(A,B) = 0$. Two sets $\mathcal{F}, \mathcal{G}$ of bounded linear operators are uncorrelated, denoted by $\mathcal{F} \perp \mathcal{G}$, if $A \perp B$ for all $A \in \mathcal{F}$ and all $B \in \mathcal{G}$. Let $\mathcal{F}$ be a closed subspace of $\mathcal{H}^m$. The *uncorrelated compliment* of $\mathcal{F}$, denoted by $\mathcal{F}^\perp$, in the space $\mathcal{H}^m$ is the smallest closed subspace of $\mathcal{H}^m$ containing all elements that are uncorrelated with $\mathcal{F}$.

The uncorrelated projection

Let $A : \mathcal{B} \to \mathcal{H}$ be a bounded linear operator from an arbitrary Banach space $\mathcal{B}$ to the Hilbert space $\mathcal{H}$. The closure $\overline{A(\mathcal{B})}$ of the range of A is a closed subspace in $\mathcal{H}$. The closed subspace of $\mathcal{H}^m$ generated by A, denoted by $[A]^m$, is an m-Cartesian product of the closed subspace $\overline{A(\mathcal{B})}$ of $\mathcal{H}$, i.e., $[A]^m = \overline{A(\mathcal{B})} \times \cdots \times \overline{A(\mathcal{B})}$. The *uncorrelated projection* of $\psi \in \mathcal{H}^m$ onto $[A]^m$ is the element $\pi(\psi|[A]^m) \in \mathcal{H}^m$ such that

$$\pi(\psi|[A]^m) \in [A]^m \text{ and } \psi - \pi(\psi|[A]^m) \in [A]^{m\perp}$$

where $[A]^{m\perp}$ is the uncorrelated compliment of $[A]^m$ in the space $\mathcal{H}^m$. Suppose that $(A^T A)^{-1}$ exists and let $h \in \mathcal{H}$. By the the projection theorem for an operator equation,

$$A(A^T A)^{-1} A^T h$$

is a projection of h onto the closure $\overline{A(\mathcal{B})}$ of range of A. This implies that, for $f = (f_1, \ldots, f_m)^T \in \mathcal{H}^m$, the uncorrelated projection of f onto the closed subspace $[A]^m$ generated by A is given by

$$\Pi(f|[A]^m) = A(A^T A)^{-1} A^T f = \begin{pmatrix} A(A^T A)^{-1} A^T f_1 \\ \vdots \\ A(A^T A)^{-1} A^T f_m \end{pmatrix}. \tag{11}$$

Appendix B: The weighted sampling model

In this Appendix, we establish that there is an i.i.d. model which is equivalent to a multi-sample model in the sense that their M-estimators have the same limiting distribution, corresponding estimating function and influence function.

[Weighted sampling model] For a multi-sample model $(\mathcal{P}_1, \ldots, \mathcal{P}_s)$ with weight probabilities $(\lambda_1, \ldots, \lambda_s)$, define a corresponding model ${}_\lambda\mathcal{P}$ by

$${}_\lambda\mathcal{P} = \{p(i,x;\beta,\eta) = \lambda_i p_i(x;\beta,\eta) :\ \beta \in \Theta_\beta,\ \eta \in \Theta_\eta\}.$$

Then ${}_\lambda\mathcal{P}$ is a probability model on the space $\{1, \ldots, s\} \times \mathcal{X}$. The model ${}_\lambda\mathcal{P}$ is called the *weighted sampling model* corresponding to the multi-sample model $(\mathcal{P}_1, \ldots, \mathcal{P}_s)$.

Since the weighted sampling model ${}_\lambda\mathcal{P}$ is an i.i.d. model, ordinary definitions of the score function and the expectation of a function can be applied. The score function for β and η are

$$\dot{\ell}_\beta(i,x,\beta,\eta) = \frac{\partial}{\partial\beta}\log p_i(x;\beta,\eta) \text{ and } \dot{\ell}_\eta(i,x,\beta,\eta) = d_\eta \log p_i(x;\beta,\eta),$$

the expectation for a function $f(i,x)$ is

$$E_{\beta,\eta}(a(I,X)) = \sum_{i=1}^{s} \lambda_i E_{i,\beta,\eta}(a(i,X)). \tag{12}$$

The score functions for β and η and the expectation in the weighted sampling model ${}_\lambda\mathcal{P}$ are equal to the ones in the multi-sample model $(\mathcal{P}_1,\ldots,\mathcal{P}_s)$.

THEOREM 3.1 with Equation 12 imply that, for an estimating function $\psi(i,x,\beta,\eta)$ for β_0 and a $\sqrt{n}$-consistent estimator $\hat{\eta}_n$ of η_0, the solution $\hat{\beta}_1$ to an estimating equation in the multi-sample model

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{s}\sum_{j=1}^{n_i}\psi(i,X_{ij},\hat{\beta}_1,\hat{\eta}_n) = o_P(1)$$

and the solution $\hat{\beta}_2$ to an estimating equation in the weighted sampling model

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\psi(I_j,X_j,\hat{\beta}_2,\hat{\eta}_n) = o_P(1)$$

have the same influence function

$$\tilde{\psi}(i,x,\beta_0,\eta_0) = \left[-\sum_{i=1}^{s}\lambda_i E_{i,\beta_0,\eta_0}\left(\frac{\partial}{\partial\beta}\psi\right)\right]^{-1}\psi(i,x,\beta_0,\eta_0)$$

so that the estimator $\hat{\beta}_1$ and $\hat{\beta}_2$ have the same limiting distribution: i.e.

$$\sqrt{n}(\hat{\beta}_k-\beta_0) \xrightarrow{d} N\left(0,\sum_{i=1}^{s}\lambda_i E_{i,\beta_0,\eta_0}(\tilde{\psi}\tilde{\psi}^T)\right),\ k=1,2.$$

As a consequence, the spaces of influence functions in both models are the same.

Moreover, since the score functions $\dot{\ell}_\beta$ and $\dot{\ell}_\eta$ are the same in both models, so are the efficient score function $\dot{\ell}^*_\beta$, the efficient influence function $\tilde{\ell}^*_\beta$, and the product tangent spaces $\mathcal{P}^m_\beta$ and $\mathcal{P}^m_\eta$. The same logic can be used to show the MLE $\hat{\beta}_n$, whose influence function is the efficient influence function $\tilde{\ell}^*_\beta$, is the most efficient among multi-sample M-estimators (in the multi-sample model) and M-estimators (in the corresponding weighted sampling model) with the same efficient information bound. This establishes the equivalence of two models in terms of M-estimation.

References

Begun, J. M., Hall, W. J., Huang, W. M. and Wellner, J. A. (1983). Information and asymptotic efficiency in parametric–nonparametric models. *Ann. Statist.* **11** 432–452.

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

Breslow, N.E. and Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumor prognosis. *Appl. Statist.* **48** 457–468.

Breslow, N.E. and Holubkov, R. (1997). Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *J. Roy. Statist. Soc. ser. B* **59** 447–461.

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.

Cosslett, S. (1981). Maximum likelihood estimator for choice-based samples. *Econometrica* **49** 1289–1316.

Godambe, V.P. (1991). Orthogonality of estimating functions and nuisance parameters. *Biometrika* **78** 143–151.

Gilbert, P.G., Lele, S.R. and Vardi, Y. (1999). Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials *Biometrika* **86** 27–43.

Gilbert, P.G. (2000). Large sample theory of maximum likelihood estimation in semiparametric selection biased sampling models *Ann. Statist.* **28** 151–194.

Gill, R.D., Vardi, Y. and Wellner, J.A. (1988). Large sample theory of empirical distribution in biased sampling models *Ann. Statist.* **3** 1069–1112.

Hall, W.S. and Newell, M.L. (1979). The mean value theorem for vector valued functions: a simple proof. *Mathematics Magazine* **52** 157–158.

Hirose, Y. (2005). Efficiency of the semi-parametric maximum likelihood estimator in generalized case-control studies. Ph.D. dissertation, Univ. Auckland.

Imbens, G. (1992). An efficient methods of moments estimator for discrete choice models with choice-based sampling. *Econometrica* **60** 1187–1214.

Imbens, G., and Lancaster, T. (1996). Efficient estimation and stratified sampling. *Journal of Econometrics* **74** 289–318.

Lawless, J.L., Kalbfleish, J.D. and Wild, C.J. (1999). Estimation for response-selective and missing data problems in regression. *J. Roy. Statist. Soc. Ser. B* **61** 413–438.

Lee, A.J. (2004). Semi-parametric efficiency bounds for regression models under choice-based sampling. Unpublished manuscript, Univ. Auckland.

Lee, A.J. and Hirose, Y. (2005). Semi-parametric efficiency bounds for regression models under case-control sampling: the profile likelihood approach. Unpublished manuscript, Univ. Auckland.

McLeish, D. L. and Small, C. G. (1992). A projected likelihood function for semiparametric models. *Biometrika* **79** 93–102.

McNeney, B. and Wellner, J.A. (2000). Application of convolution theorems in semiparametric models with non-i.i.d. data. *J. Statist. Plan. and Inf.* **91** 441–480.

Nan, B., Emond, M. and Wellner, J.A. (2004). Information bounds for cox regression models with missing data. *Ann. Statist.* **32** 723–753.

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

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

Prentice, R.L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66** 403–411.

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

Seber, G.A.F. and Lee, A.J. (2003). *Linear Regression Analysis, Second Edition*. Wiley, New York.

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. Plann. Inference* **96** 3–27.

Tsiatis, A.B. (2006). *Semiparametric Theory and Missing Data*. Springer, New York.

van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge Univ. Press, Cambridge.

Vardi, Y. (1985). Empirical distributions in selection bisa models *Ann. Statist.* **13** 178–203.