2-22-2002

Large Sample Theory for Semiparametric Regression Models with Two-Phase, Outcome Dependent Sampling

Norm Breslow
University of Washington, norm@u.washington.edu

Brad McNeney
Simon Fraser University, mcneney@stat.sfu.ca

Jon A. Wellner
University of Washington, jaw@stat.washington.edu

1 Introduction

Outcome-dependent, two-phase stratified sampling designs can dramatically reduce the costs of observational studies by selecting the most informative subjects for detailed covariate measurement. Although ad-hoc, inefficient estimation methods often have been used with these designs, recent work has focused on maximum likelihood estimation for semiparametric regression models. Scott and Wild (1997), who considered simple random samples at the first phase of sampling, and Breslow and Holubkov (1997), who considered case-control sampling at phase one and worked exclusively with the logistic model, developed maximum likelihood estimators for binary response models. This work extended the classical theory of Prentice and Pyke (1979) to samples that were jointly stratified by outcomes and covariates. Lawless, Kalbfleisch, and Wild (1999) (LKW) and Scott and Wild (2000) generalized the approach of Scott and Wild (1997) and demonstrated that computation of maximum likelihood estimators is feasible for a wide range of parametric regression models and two-phase designs provided that the phase one data are discrete. For an example of a two-phase design, see Breslow and Chatterjee (1999).

Robins, Hsieh, and Newey (1995) (RHN) derived the semiparametric efficient scores for a more general problem in which a portion of the covariate vector is missing for some subjects, but the outcome variable and the other covariates are fully known for everyone. In the general case of continuous data, calculation of the optimal estimator involves numerical solution of an (infinite dimensional) integral equation. When the outcomes and covariates observed for everyone are discrete and used to define the samping strata, in which case the problems considered by LKW and RHN are identical, RHN calculated an optimal estimator by solving finite dimensional linear equations to obtain the scores.

LKW remarked that the RHN methods "appear to be asymptotically equivalent" to theirs for the case of discrete phase one data, but offered no proof. They also remarked that they had "no theoretical justification" for their empirical observation that inferences based on the observed information for the profile likelihood "performed excellently" even when the covariates were continuous. Our goal is to provide asymptotic theory that resolves these outstanding issues. We first establish asymptotic lower bounds using the methods of Bickel, Klaassen, Ritov, and Wellner (1993) to compute the efficient score functions, (efficient) information, and efficient influence functions for the problem considered by LKW. For at least the i.i.d. special case of variable probability (Bernoulli) sampling, these results also follow from the more general information calculations of RHN. The models considered here yield sufficiently explicit formulas, however, that they deserve special consideration. Although we do not go beyond the i.i.d. Bernoulli sampling framework here, McNeney (1998) shows that the information bounds for the more realistic basic stratified sampling model (*cf.* LKW) agree with those for Bernoulli sampling under mild conditions. We intend to give a complete treatment of the stratified sampling model and other designs elsewhere.

In Section 3 we identify a least favorable parametric submodel and verify that it satisfies the key hypothesis of Theorem 1 of Murphy and Van der Vaart (2000). In Section 4, we use the least favorable parametric submodel to justify the asymptotic expansion of the profile likelihood in terms of the efficient score and information, which allows it to be treated as an ordinary likelihood for purposes of statistical inference. A corollary of this development is asymptotic normality and

efficiency of the maximum likelihood estimator $\widehat{\theta}_n$ of θ. The final task, undertaken in Section 5, is to prove joint asymptotic normality and efficiency of the ML estimators. Results given for both the parametric and nonparametric components of the model are apparently new. Our approach, which requires only modest regularity assumptions, is via Theorem 1 of Murphy and Van der Vaart (2000) and a verification of their hypotheses for our particular class of models.

In Section 6 we discuss other designs and further problems. The more lengthy arguments, including derivation of the semiparametric likelihood under several sampling designs, direct computation of the information bounds using operator theory, verification of regularity conditions for a least favorable parametric submodel, a statement of the infinite-dimensional $Z-$theorem, connections with the formulas of RHN, and a derivation of the information formula for the important special case of logistic regression, are spelled out in complete detail in the companion technical report Breslow, McNeney, and Wellner (2000) (BMW).

2 Information Bounds, Bernoulli sampling.

In this section we derive information bounds for estimation of the regression parameters assuming that the sampling design yields i.i.d. data. Suppose that (Y, X) has density $f(y|x;\theta)g(x)$ with respect to a dominating measure $\nu \times \mu$ on $\mathcal{Y} \times \mathcal{X}$ for some $\theta \in \Theta \subset R^m$ and some $G \in \mathcal{G}$, where

$$\mathcal{G} \equiv \{G : G \text{ is a distribution on } \mathcal{X} \text{ with density } g \text{ with respect to } \mu\},$$

and let $Q_{\theta,G}$ denote the corresponding probability measure. Both X and Y may be multivariate. Let $\mathcal{Y} \times \mathcal{X} = \cup_{j=1}^{J} \mathcal{S}_j$ for a partition $\{\mathcal{S}_j\}$ into J mutually exclusive strata. Following LKW, we set

$$Q_j(\theta, G) = \Pr[(Y, X) \in \mathcal{S}_j], \qquad \text{and} \qquad Q_j^*(x, \theta) \equiv \Pr[(Y, x) \in \mathcal{S}_j | X = x] 1_{\mathcal{S}_j^*}(x),$$

for $j = 1, \ldots, J$ where $\mathcal{S}_j^* = \{x \in \mathcal{X} : \text{ for some } y,\ (y, x) \in \mathcal{S}_j\}$. Thus $Q_j(\theta, G) = \int Q_j^*(x, \theta) dG(x)$. Note that the $\mathcal{S}_j^*$'s do *not* form a partition of $\mathcal{X}$, and may in fact intersect in quite arbitrary ways.

Suppose that $(Y_1, X_1), \ldots, (Y_n, X_n)$ are i.i.d. as (Y, X) with density

$$p(y, x; \theta_0, g_0) = f(y|x;\theta_0) g_0(x). \tag{2.1}$$

We assume throughout that the true distribution governing the underlying data is given by (2.1) corresponding to $(\theta_0, G_0) \in \Theta \times \mathcal{G}$. We also assume that

$$Q_j(\theta_0, G_0) > 0, \qquad j \in \{1, \ldots, J\}. \tag{2.2}$$

At the first phase of sampling we do not observe the complete (Y_i, X_i) pairs, but only observe *stratum indicators*

$$\delta_{ij} = 1\{(Y_i, X_i) \in \mathcal{S}_j\}, \qquad i = 1, \ldots, n, \quad j = 1, \ldots, J.$$

Thus

$$\underline{\delta}_i = (\delta_{i1}, \ldots, \delta_{iJ}) \sim \text{Mult}_J(1, \underline{Q} = (Q_1, \ldots, Q_J)^T)$$

where $Q_j \equiv Q_j(\theta, G)$, $j = 1, \ldots, J$. We will sometimes use the alternative and completely equivalent stratum variables S_i, defined by $S_i = s$ if and only if $\delta_{is} = 1$ for $i = 1, \ldots, n$. Now suppose that selection of subjects for complete response and covariate ascertainment at the second phase of sampling is defined by the indicators

$$R_i = \begin{cases} 1 & \text{if } (Y_i, X_i) \text{ is fully observed} \\ 0 & \text{if only } S_i \text{ is observed} \end{cases} .$$

We set $D_j = \{i : \ \delta_{ij} = 1, R_i = 1\}$, $N_j = \sum_{i=1}^n \delta_{ij} = \#\{i : \ (Y_i, X_i) \in \mathcal{S}_j\}$, and $n_j = \#(D_j)$, for $j = 1, \ldots, J$ so that $\underline{N} = (N_1, \ldots, N_J)^T \sim \text{Mult}_J(n, \underline{Q})$.

We confine our attention in this paper to *Variable Probability Sampling (VPS):* units are inspected sequentially as they arise from the density (2.1). When $(Y_i, X_i) \in \mathcal{S}_j$, the ith unit is selected for full observation ($R_i = 1$) with specified probability p_j; thus

$$\Pr(R_i = 1 | Y_i, X_i) = \sum_{j=1}^J p_j 1\{(Y_i, X_i) \in \mathcal{S}_j\} = \sum_{j=1}^J p_j \delta_{ij} = p_{S_i}.$$

Two variants of this plan depend on how the sampling is terminated:
VPS1: Inspect a pre-specified number n of units (Bernoulli sampling).
VPS2: Inspect units until a total of k have been selected (Negative Binomial sampling).
As shown by Scott and Wild (1997) or Appendix 1A of BMW, VPS1 (Bernoulli) sampling results in the following density for the observed data $(R, Z) \equiv (R, (Y, X)1_{[R=1]} + \underline{\delta} 1_{[R=0]}) \equiv (R, (Y, X)1_{[R=1]} + S1_{[R=0]})$: with $q_j \equiv 1 - p_j$, $j = 1, \ldots, J$,

$$\begin{aligned} p(r, z; \theta, g) &\equiv \left\{ f(y|x;\theta) g(x) \left(\sum p_j \delta_j \right) \right\}^r \left\{ \prod_{j=1}^J Q_j^{\delta_j} \right\}^{1-r} \left\{ \sum q_j \delta_j \right\}^{1-r} \\ &= \prod_{j=1}^J \left\{ [f(y|x;\theta) g(x)]^{\delta_j r} Q_j^{\delta_j (1-r)} \right\} \left\{ \sum p_j \delta_j \right\}^r \left\{ \sum q_j \delta_j \right\}^{1-r} . \end{aligned} \tag{2.3}$$

This is our starting point for information calculations in the i.i.d. version of the model. Let $\mathcal{P}$ be the collection of all probability distributions $P_{\theta,G}$ with densities given by (2.3) for $\theta \in \Theta$, $G \in \mathcal{G}$.

Proposition 2.1. (Scores for the i.i.d. model). Suppose that (R, Z) has the density (2.3), that (2.2) holds, and that for a fixed $G_0 \in \mathcal{G}$

$$\mathcal{Q}_{G_0} \equiv \{ Q_{\theta, G_0} : \frac{dQ_{\theta,G_0}}{d(\nu \times \mu)}(y, x) = f(y|x;\theta) g_0(x), \ \ \theta \in \Theta \}$$

is a regular parametric model. Suppose $\theta_0 \in \Theta^0$, the *interior* of Θ, and write P_0 for $P_{(\theta_0, G_0)}$ and E_0 for expectation under P_0, respectively. Then the score for θ and the score operator for g at P_0 in the VPS1 model are given by

$$\begin{aligned} \dot{\mathbf{l}}_\theta(r, z) &= r \dot{\mathbf{l}}_\theta(y|x) + (1 - r) \sum_{j=1}^J \delta_j \dot{\underline{Q}}_j(\theta_0, G_0) / Q_j(\theta_0, G_0) \\ &= r \dot{\mathbf{l}}_\theta(y|x) + (1 - r) E_0 \{ \dot{\mathbf{l}}_\theta(Y|X) \big| S = s \} \end{aligned} \tag{2.4}$$

where

$$\dot{\mathbf{l}}_\theta(y|x;\theta_0) \equiv \dot{\mathbf{l}}_\theta(y|x) \equiv \frac{\partial}{\partial\theta}\log f(y|x;\theta)|_{\theta=\theta_0} \equiv \nabla_\theta \log f(y|x;\theta_0),$$

$$\underline{\dot{Q}}_j(\theta_0, G_0) \equiv \nabla_\theta Q_j(\theta_0, G_0) \equiv \int \underline{\dot{Q}}_j^*(x,\theta_0) dG_0(x),$$

and, for $h \in L_2^0(G_0) \equiv \{h \in L_2(G_0) : \int h dG_0 = 0\}$,

$$\begin{aligned} \dot{\mathbf{l}}_g h(r,z) &\equiv A_{\theta_0,G_0} h(r,z) \equiv A_{\theta_0,G_0} h(r, r(y,x) + (1-r)\underline{\delta}) \\ &= rh(x) + (1-r)E_0\{h(X)|S=s\} \\ &= rh(x) + (1-r)\underline{\delta}^T \mathrm{diag}(1/\underline{Q})\langle \underline{Q}^*, h\rangle \end{aligned} \tag{2.5}$$

where $\langle h_1, h_2\rangle = \int h_1 h_2 dG_0$ denotes the inner product in $L_2(G_0)$.

Computation of the scores in Proposition 2.1 and inversion of the information operator $A^T_{\theta_0,G_0} A_{\theta_0,G_0}$, which calculations are carried out explicitly in Section 2 of BMW, lead directly to the information bounds for θ given in the following proposition. Since the derivation is rather lengthy, however, the proof here relies instead on results of RHN.

Proposition 2.2. (Efficient scores and Information bounds for the i.i.d. model). Suppose that the assumptions of Proposition 2.1 hold at P_0 and that $0 < p_j < 1$ for each $j \in \{1, \ldots, J\}$. Define

$$\begin{aligned} \psi(y,x) &\equiv \dot{\mathbf{l}}_\theta(y|x) - \frac{\dot{\mathbf{Q}}^*}{\pi^*}(x)\underline{p} - (\dot{\mathbf{Q}} - \mathbf{C})\mathbf{M}^{-1}\frac{\underline{Q}^*}{\pi^*}(x) \\ &= \dot{\mathbf{l}}_\theta(y|x) - E_0\{\dot{\mathbf{l}}_\theta(Y|X)|R=1, X=x\} \\ &\quad - E_0\{(\dot{\mathbf{Q}} - \mathbf{C})\mathbf{M}^{-1}\mathrm{diag}(1/\underline{p})\underline{\delta}|R=1, X=x\}, \end{aligned} \tag{2.6}$$

where

$$\pi^*(x) \equiv \sum_{j=1}^J p_j Q_j^*(x,\theta_0) = E_0(p_S|X=x)\,, \tag{2.7}$$

$$\mathbf{M} \equiv \mathrm{diag}(\underline{Q}/\underline{q}) + \langle \underline{Q}^*, \frac{1}{\pi^*}\underline{Q}^{*T}\rangle\,, \tag{2.8}$$

is always nonsingular, $\dot{\mathbf{Q}} \equiv (\underline{\dot{Q}}_1, \ldots, \underline{\dot{Q}}_J)$ is an $m \times J$ matrix, $\dot{\mathbf{Q}}^* \equiv (\underline{\dot{Q}}_1^*, \ldots, \underline{\dot{Q}}_J^*)$ is an $m \times J$ matrix of functions, and

$$\mathbf{C} \equiv \langle \dot{\mathbf{Q}}^* \underline{p}, \frac{\underline{Q}^{*T}}{\pi^*}\rangle \qquad (\text{ an } \ m \times J \text{ matrix})\,. \tag{2.9}$$

Then the efficient score function for θ is given by

$$\mathbf{l}_\theta^*(r,z) = r\psi(y,x) + (1-r)E_0\{\psi(Y,X)|\underline{\delta}\}\,, \tag{2.10}$$

the information for θ at (θ_0, G_0) is

$$\begin{aligned} I(\theta_0) &= E_0\left\{ R\left[\dot{\mathbf{l}}_\theta(Y|X) - \frac{\dot{\mathbf{Q}}^*}{\pi^*}(X)\underline{p}\right]^{\otimes 2}\right\} + (\dot{\mathbf{Q}} - \mathbf{C})\mathbf{M}^{-1}(\dot{\mathbf{Q}} - \mathbf{C})^T \qquad (2.11) \\ &= E_0\left\{ R\left[\dot{\mathbf{l}}_\theta(Y|X) - E_0\{\dot{\mathbf{l}}_\theta(Y|X)|R=1,X\}\right]^{\otimes 2}\right\} + (\dot{\mathbf{Q}} - \mathbf{C})\mathbf{M}^{-1}(\dot{\mathbf{Q}} - \mathbf{C})^T\,, \end{aligned}$$

and the efficient influence function is

$$\widetilde{\mathbf{l}}_\theta(r,z) = I(\theta_0)^{-1}\mathbf{l}^*_\theta(r,z)\,. \qquad (2.12)$$

Remark 2.1. The efficient score function for θ given by (2.10) agrees with the calculations of RHN after making the following minor correction to their Proposition 1. According to their equations (11), (17) and (23) and the arguments on page 421, the expressions for the optimal $U^{(2)}(\phi_{op})$ on pages 413 and 414 should read

$$U^{(2)}(\phi_{op}) = -\Delta E\left[\frac{1-\pi(W)}{\pi(W)}\phi_{op}(W)\middle|\Delta = 1, X, V\right] + (1-\Delta)\phi_{op}(W).$$

Then, with $\underline{q} = \underline{1} - \underline{p}$,

$$\xi(\underline{\delta}) \equiv E_0\{\psi(Y,X)|\underline{\delta}\} = (\dot{\mathbf{Q}} - \mathbf{C})\mathbf{M}^{-1}\mathrm{diag}(1/\underline{q})\underline{\delta} \qquad (2.13)$$

plays the role of RHN's $\phi_{op}(W)$ and satisfies the finite dimensional, integral (linear) equation

$$\xi(\underline{\delta}) = E_0\{\dot{\mathbf{l}}_\theta - E_0(\dot{\mathbf{l}}_\theta|R=1,X)|\underline{\delta}\} - E_0\{E_0[\xi(\underline{\delta})\underline{q}^T\mathrm{diag}(1/\underline{p})\underline{\delta}|R=1,X]|\underline{\delta}\} \qquad (2.14)$$

that corresponds to their equation (8). See Section 3 and Appendix 1C of BMW. For an independent recent derivation of the more general integral equation of RHN, see Nan, Emond, and Wellner (2000).

Remark 2.2. Calculations based on the score operator (2.5) also lead easily to an information bound for estimation of the distribution G as in Begun, Hall, Huang, and Wellner (1983), Theorem 4.1, page 441, or Bickel, Klaassen, Ritov, and Wellner (1993), Corollary 3, page 215. See the statement of Theorem 4.1 and Section 4 of BMW.

Remark 2.3. The hypothesis that $0 < p_j < 1$ for all $j = 1, \ldots, J$ in Proposition 2.2 can be weakened to $0 < p_j \le 1$ for all $j = 1, \ldots, J$. This is important in practice since often the p_j's in strata with relatively small Q_j, and hence small counts N_j, will be taken to be 1. Note that the second term in (2.11) can be rewritten as

$$[(\dot{\mathbf{Q}} - \mathbf{C})\mathrm{diag}(\sqrt{\underline{q}})]\widetilde{\mathbf{M}}^{-1}[\mathrm{diag}(\sqrt{\underline{q}})(\dot{\mathbf{Q}} - \mathbf{C})^T] \qquad (2.15)$$

where

$$\widetilde{\mathbf{M}} \equiv \mathrm{diag}(\underline{Q}) + \mathrm{diag}(\sqrt{\underline{q}})\langle \underline{Q}^*, \frac{1}{\pi^*}\underline{Q}^{*T}\rangle\mathrm{diag}(\sqrt{\underline{q}})$$

is always invertible (even if some $p_j = 1$, $q_j = 0$) by virtue of (2.2). Also note that if all $p_j = 1$ so that all $q_j = 0$, then the second term (as rewritten in (2.15)) vanishes, $R = 1$ identically, and the first term becomes

$$E_0\left\{\left(\dot{\mathbf{l}}_\theta(Y|X) - E_0(\dot{\mathbf{l}}_\theta(Y|X)|X)\right)^{\otimes 2}\right\} = E_0\left\{\dot{\mathbf{l}}_\theta(Y|X)^{\otimes 2}\right\}, \tag{2.16}$$

the information for θ with complete data from (2.1).

Remark 2.4. Any (locally regular) estimator of θ in the i.i.d. two-phase sampling model has an influence function of the form

$$\phi(r,z) = \frac{r}{\pi(y,x)}\dot{\chi}(y,x) - \frac{r-\pi(y,x)}{\pi(y,x)}c(s) \tag{2.17}$$

for some function $c : \{1,\ldots,J\} \to R^m$ where $\dot{\chi}$ is an influence function for some estimator of θ in the complete data model $\mathcal{Q}$, with true element Q_0, in which all the (Y_i, X_i)'s are observed; i.e.

$$\dot{\chi}(y,x) = I_{11}^{-1}(\theta_0)\dot{\mathbf{l}}_\theta(y|x) + h(y,x)$$

where $I_{11}(\theta_0) = E_{Q_0}\dot{\mathbf{l}}_\theta^{\otimes 2}$, $h = (h_1,\ldots,h_m)$, and $h_i \perp L_2^0(G)$ in $L_2(Q_0)$ for $i = 1,\ldots,m$. For a proof of (2.17), see Van der Vaart (1998), pages 379 - 383. In particular, all of the inefficient estimators considered in LKW have influence functions of this form for some h and c.

Example 2.1. (Logistic regression for stratified case-control studies.) Suppose that

$$f(y|x) = f(y|x;\theta) = \left(\frac{e^{\theta^T x}}{1+e^{\theta^T x}}\right)^y \left(\frac{1}{1+e^{\theta^T x}}\right)^{1-y}, \qquad y \in \{0,1\},\ x \in R^m, \theta \in R^m. \tag{2.18}$$

Then, since the logit is the canonical link function for the Bernoulli distribution (see e.g. McCullagh and Nelder (1989), pages 28-31), $\dot{\mathbf{l}}_\theta(y|x) = x[y - f(1|x)]$. For stratified case-control sampling, as discussed by Scott and Wild (1997) or Breslow and Holubkov (1997), the partition of $\mathcal{Y} \times \mathcal{X}$ is formed by intersecting a partition of $\mathcal{X}$ into J sets $\{\mathcal{X}_j\}$ with the sets $1\{y = 0\}$ and $1\{y = 1\}$. This leads to $2J$ strata $\mathcal{S}_{yj} = 1\{Y = y, X \in \mathcal{X}_j\}$ for $y = 0, 1;\ j = 1,\ldots,J$. Continuing this double subscript system, let p_{yj} denote the corresponding sampling probabilities for selection at phase two.

Corollary 2.1. (Information for θ, logistic regression special case). Suppose that the model is given by (2.18). Then (2.11) yields

$$\begin{aligned} I(\theta) &= \sum_{j=1}^{J}\int_{\mathcal{X}_j} \underline{x}\,\underline{x}^T \frac{p_{0j}p_{1j}f(0|x)f(1|x)}{\pi_j^*(x)}dG(x) \\ &\quad + \sum_{j=1}^{J} \frac{p_{0j}p_{1j}\left(\int_{\mathcal{X}_j}\underline{x}\frac{f(0|x)f(1|x)}{\pi_j^*(x)}dG(x)\right)^{\otimes 2}}{\frac{Q_{0j}Q_{1j}}{p_{0j}Q_{0j}+p_{1j}Q_{1j}-p_{0j}p_{1j}G(\mathcal{X}_j)} - \int_{\mathcal{X}_j}\frac{f(0|x)f(1|x)}{\pi_j^*(x)}dG(x)} \end{aligned} \tag{2.19}$$

where $\pi_j^*(x) \equiv p_{0j} f(0|x) + p_{1j} f(1|x)$ for $x \in \mathcal{X}_j$.

Proof. See Appendix 1D of BMW. The same expression may be derived by using the linearization discussed in Section 4 of Breslow and Holubkov (1997), and additional Taylor series expansions, in a direct computation of the influence function for the maximum likelihood estimator. □

Remark 2.5. Consider the special case where $J = 1$, so that sampling depends only on the binary outcome, and drop the j subscript in what follows. Suppose the linear predictor contains an intercept: $f(0|x) = (1 + e^{\theta_1 + \theta_2^T x})^{-1}$. Let $f^\pi(y|x) = p_y e^{y(\theta_1+\theta_2^T x)}/(p_0 + p_1 e^{\theta_1+\theta_2^T x}) = \Pr(Y = y|X = x, R = 1)$ denote the logistic regression probabilities of the "biased sampling model" $Q = Q_{\theta,G}$ induced by the condition $R = 1$. Then the information matrix may be written

$$I(\theta) = E_Q[\mathrm{Var}_Q(Y|X)] \left\{ \begin{bmatrix} 1 & \mu^T \\ \mu & \mu_2 \end{bmatrix} + c \begin{bmatrix} 1 & \mu^T \\ \mu & \mu\mu^T \end{bmatrix} \right\}$$

where

$$c = \frac{\pi E_Q[\mathrm{Var}_Q(Y|X)](p_0 Q_0 + p_1 Q_1 - p_0 p_1)}{p_0 p_1 Q_0 Q_1 - \pi E_Q[\mathrm{Var}_Q(Y|X)](p_0 Q_0 + p_1 Q_1 - p_0 p_1)},$$

$$\mu = \frac{E_Q[X \mathrm{Var}_Q(Y|X)]}{E_Q[\mathrm{Var}_Q(Y|X)]} = \frac{\int x f^\pi(0|x) f^\pi(1|x) \pi^*(x) dG(x)}{\int f^\pi(0|x) f^\pi(1|x) \pi^*(x) dG(x)}$$

and

$$\mu_2 = \frac{E_Q[XX^T \mathrm{Var}_Q(Y|X)]}{E_Q[\mathrm{Var}_Q(Y|X)]} = \frac{\int x x^T f^\pi(0|x) f^\pi(1|x) \pi^*(x) dG(x)}{\int f^\pi(0|x) f^\pi(1|x) \pi^*(x) dG(x)},$$

and where $\pi = \pi(\theta, G) = \Pr(R = 1)$. Now the information for θ_2 is

$$\begin{aligned} I(P|\theta_2, \mathcal{P}) &= I_{22}(\theta) - I_{21}(\theta) I_{11}^{-1}(\theta) I_{12}(\theta) \\ &= E_Q[\mathrm{Var}_Q(Y|X)](\mu_2 - \mu\mu^T) \\ &= E_Q \left\{ \left(X - \frac{E_Q[X \mathrm{Var}_Q(Y|X)]}{E_Q[\mathrm{Var}_Q(Y|X)]} \right)^{\otimes 2} \mathrm{Var}_Q(Y|X) \right\}. \end{aligned}$$

This expression, which agrees with formulas (4) and (9) of Breslow, Robins, and Wellner (2000), is precisely the information about θ_2 obtained by fitting an ordinary logistic regression model to the second phase data alone. It confirms once again that, for simple case-control sampling, "prospective" logistic regression analyis of the "retrospectively" sampled data yields efficient estimates of the odds ratio parameters in logistic regression models (Prentice and Pyke (1979)).

Proof of Proposition 2.2. This follows from Proposition 1 of RHN after the corrections noted in Remark 2.1. We rewrite (2.14) (or equation (4.2) of Nan, Emond, and Wellner (2000)) as a matrix equation, and express the solution in terms of the inverse of a certain matrix. First,

$$E_0\{\dot{\mathbf{l}}_\theta | \underline{\delta}\} = \dot{\mathbf{Q}} \mathrm{diag}(1/\underline{Q}) \underline{\delta}$$

and

$$E_0\{E_0(\dot{\mathbf{l}}_\theta|R=1,X)|\underline{\delta}\} = \mathbf{C}\text{diag}(1/\underline{Q})\underline{\delta}$$

where $\mathbf{C}$ is as defined in (2.9). Thus the first term on the right side of (2.14) is

$$E_0\{\dot{\mathbf{l}}_\theta - E_0(\dot{\mathbf{l}}_\theta|R=1,X)|\underline{\delta}\} = (\dot{\mathbf{Q}} - \mathbf{C})\text{diag}(1/\underline{Q})\underline{\delta}\,. \tag{2.20}$$

Furthermore, writing $\xi(\underline{\delta}) = \sum_{j=1}^{J} \underline{\xi}_j \delta_j = \boldsymbol{\xi}\underline{\delta}$ for an $m \times J$ matrix $\boldsymbol{\xi} = (\underline{\xi}_1, \ldots, \underline{\xi}_J)$, we can rewrite the second term on the right side of (2.14) as

$$E_0\{E_0[\xi(\underline{\delta})\underline{q}^T\text{diag}(1/\underline{p})\underline{\delta}|R=1,X]|\underline{\delta}\} = \boldsymbol{\xi}\text{diag}(\underline{q})\mathbf{D}\text{diag}(1/\underline{Q})\underline{\delta} \tag{2.21}$$

where $\mathbf{D} = \langle \underline{Q}^*, \frac{1}{\pi^*}\underline{Q}^{*T}\rangle$. Substitution of (2.20) and (2.21) into (2.14) and rearranging yields

$$\boldsymbol{\xi}(\mathbf{I} + \text{diag}(\underline{q})\mathbf{D}\text{diag}(1/\underline{Q})) = (\dot{\mathbf{Q}} - \mathbf{C})\text{diag}(1/\underline{Q})\,,$$

or, with $\mathbf{M} \equiv \text{diag}(\underline{Q}/\underline{q}) + \mathbf{D}$ as in (2.8),

$$\boldsymbol{\xi}\text{diag}(\underline{q})\,\mathbf{M} = (\dot{\mathbf{Q}} - \mathbf{C}).$$

Note that

$$\underline{a}^T\mathbf{M}\underline{a} = \sum_{j=1}^{J}(Q_j/q_j)a_j^2 + \|\underline{a}^T\underline{Q}^*/\sqrt{\pi^*}\|^2 > 0$$

for all $\underline{a} \neq \underline{0}$. Therefore the matrix $\mathbf{M}$ is nonsingular, $\mathbf{M}^{-1}$ exists and

$$\boldsymbol{\xi} = (\dot{\mathbf{Q}} - \mathbf{C})\mathbf{M}^{-1}\,\text{diag}(1/\underline{q})\,. \tag{2.22}$$

Using (2.22) in the (corrected) formula for $U^{(2)}(\phi_{op})$ in Remark 2.1, together with $U^{(1)}$ in RHN's Proposition 1, yields the claimed efficient score given in (2.10). □

3 A least favorable parametric submodel

An alternative approach to understanding of the efficient scores and influence function for θ in the semiparametric model (2.3) is to determine a least favorable submodel for G as in Murphy and Van der Vaart (2000)(MvdV). We initially determine a candidate least favorable submodel by partial maximization of the expected log-likelihood, assuming that G is discrete. Subsequent calculations show that our submodel satisfies MvdV's key conditions for a least favorable parametric submodel; we give regularity conditions under which the remaining hypotheses of their Theorem 1 hold. This provides theoretical confirmation for LKW's simulation studies, which showed that inferences based on the observed information matrix of the profile likelihood function had appropriate frequency properties.

Suppose then that X takes K values $\{x_k\}$ with probabilities g_k, $\sum_k g_k = 1$. Define

$$\pi_j(Q) \equiv 1 - \frac{Q_j(1-p_j)}{Q} \text{ for } Q \in (0,1], \tag{3.1}$$

where $Q_j = Q_j(\theta, G)$ is defined in Section 2, and note that $\pi_j(Q_j) = p_j$. With $E \equiv E_{\theta,G}$ denoting expectation with respect to (θ, G), we also define

$$g_k^* \equiv g_k^*(\theta, G) \equiv E(R1_{\{X=x_k\}}) = \sum_{j=1}^{J} p_j Q_j^*(x_k;\theta) g_k.$$

For t in a neighborhood of θ and H ranging over the discrete distributions for X, our goal is to find the distribution

$$G_t \equiv G_t(\theta, G) \equiv \text{argmax}_H \; E[\log p(R, Z; t, H)]$$

that maximizes the expected log-likelihood $\ell(t, H) \equiv \ell(t, H; \theta, G)$ given by

$$\ell(t,H) \equiv E[R \log f(Y|X;t)] + E[R \log h(X)] + E[(1-R)\log Q_S(t,H)]. \tag{3.2}$$

Towards this end we fix t and maximize (3.2) as a function of $H = \{h_k\}$ subject to $\sum_k h_k = 1$. Following the arguments in Scott and Wild (1997) and LKW, introduce the Lagrange multiplier λ for the side condition $(\sum_k h_k - 1) = 0$ and jointly solve the $K+1$ equations

$$\frac{\partial[\ell(t,H) + \lambda(\sum h - 1)]}{\partial h_k} = \frac{g_k^*}{h_k} + \sum_j (1-p_j) Q_j \frac{Q_j^*(x_k,t)}{\sum_\ell Q_j^*(x_\ell,t) h_\ell} + \lambda = 0 \tag{3.3}$$

for $k = 1, \ldots, K$, and

$$\sum_k h_k - 1 = 0.$$

Multiplying (3.3) by h_k and summing over k gives $ER + (1 - ER) + \lambda = 0$ or $\lambda = -1$. This allows (3.3) to be re-expressed

$$h_k = \frac{g_k^*}{\sum_j \pi_j[Q_j(t,H)] Q_j^*(x_k,t)}. \tag{3.4}$$

Substituting for h_k in (3.2) using (3.4) yields the profile expected log-likelihood

$$\begin{aligned} E \log p(R, Z; t, G_t) \;\; &= \;\; E[R \log f(Y|X;t)] \;\; - \;\; \sum_k g_k \log \sum_j \pi_j[Q_j(t,G_t)] Q_j^*(x_k,t) \\ &\quad + \sum_j (1-p_j) Q_j \log Q_j(t, G_t) + \text{constant}. \end{aligned} \tag{3.5}$$

This depends on $G_t = \{g_t(x_k)\}$ only through the values of

$$Q_j^\dagger(t) \equiv Q_j(t, G_t) = \sum_k Q_j^*(x_k, t) g_t(x_k). \tag{3.6}$$

By substitution of $g_t(x_k) = h_k$ from (3.4) into (3.6), the $Q_j^\dagger(t)$ are determined for each t from the equations

$$Q_j^\dagger(t) = \sum_k \frac{Q_j^*(x_k; t) g_k^*}{\sum_\ell \pi_\ell [Q_\ell^\dagger(t)] Q_\ell^*(x_k; t)}, \qquad j = 1, \ldots, J. \tag{3.7}$$

It follows that G_t has point masses $g_t(x_k)$ that arise by substitution of $Q_j^\dagger(t)$ for the $Q_j(t, G_t)$ in (3.4). Thus for $x = x_k$, $k = 1, \ldots, K$,

$$g_t(x) = \frac{\sum_{j=1}^J p_j Q_j^*(x, \theta)}{\sum_{j=1}^J \pi_j (Q_j^\dagger(t)) Q_j^*(x, t)} \, g(x) \, . \tag{3.8}$$

Generalizing (3.7) and (3.8), suppose now that $G_t \equiv G_t(\theta, G)$ has density g_t with respect to G given by

$$g_t(x) \equiv \frac{dG_t}{dG}(x) = \frac{\sum_{j=1}^J p_j Q_j^*(x, \theta)}{\sum_{j=1}^J \pi_j [Q_j^\dagger(t)] Q_j^*(x, t)} \tag{3.9}$$

where the $Q_j^\dagger(t) \equiv Q_j(t, G_t)$ satisfy

$$Q_j^\dagger(t) = \int \frac{\sum_{\ell=1}^J p_\ell Q_\ell^*(x, \theta)}{\sum_{\ell=1}^J \pi_\ell [Q_\ell^\dagger(t)] Q_\ell^*(x, t)} \, Q_j^*(x, t) dG(x), \qquad j = 1, \ldots, J \, . \tag{3.10}$$

(In the next section we will also use the notation $Q_j^\dagger(t, \theta, G) = Q_j(t, G_t(\theta, G))$.) The log-likelihood for one observation for our proposed least favorable submodel is given by

$$\begin{aligned} l(t, \theta, G)(r, z) &\equiv l(t, G_t(\theta, G))(r, z) \\ &= r \left(\log f(y|x; t) + \log \frac{dG_t}{dG}(x, \theta, G) \right) + (1 - r) \sum_{j=1}^J \delta_j \log Q_j^\dagger(t) \, . \end{aligned} \tag{3.11}$$

Since the Q_j themselves satisfy (3.10) when $t = \theta$, it follows that $G_\theta = G$ and thus that the submodel passes through (θ, G) as required by MvdV's equation (8). To calculate $\dot{\ell}(\theta, \theta, G)$, the t derivative of (3.11), let $\mathbf{\Delta}_{\cdot j} \equiv \partial Q_j^\dagger(t)/\partial t$ evaluated at $t = \theta$. The corresponding $m \times J$ matrix $\mathbf{\Delta}$ has elements $\mathbf{\Delta}_{kj}$, rows $\mathbf{\Delta}_{k\cdot}$ and columns $\mathbf{\Delta}_{\cdot j}$. Differentiating both sides of (3.10) with respect to t shows that, for each k, the $1 \times J$ gradient vector $\mathbf{\Delta}_{k\cdot}$ solves a system of linear equations. In fact, with $\mathbf{M}$ and $\mathbf{C}$ as defined in equations (2.8) and (2.9), it is given by

$$\mathbf{\Delta}_{k\cdot} = (\dot{\mathbf{Q}} - \mathbf{C})_{k\cdot} \mathbf{M}^{-1} \mathrm{diag}(\underline{Q}/\underline{q}) \tag{3.12}$$

where $\dot{\mathbf{Q}} - \mathbf{C}$ has components

$$(\dot{\mathbf{Q}} - \mathbf{C})_{kj} = Q_j \left(E[\dot{\mathbf{l}}_{\theta_k}(Y|X)|S=j] - E\{E[\dot{\mathbf{l}}_{\theta_k}(Y|X)|R=1,X]|S=j\} \right). \tag{3.13}$$

One interpretation of equations (3.12) and (3.13) is that the finite dimensional random variable $\xi(S) = \mathbf{\Delta}_{\cdot S}/Q_S$ satisfies the linear equation (compare equation (2.14))

$$\xi(S) = E[\dot{\mathbf{l}}_\theta - E(\dot{\mathbf{l}}_\theta|R=1,X)|S] - E\{E[q_S p_S^{-1}\xi(S)|R=1,X]|S\}. \tag{3.14}$$

From (3.9) we have

$$\begin{aligned} \nabla_t \log g_t(x) \mid_{t=\theta} &= -\frac{\sum_j \dot{\pi}_j \mathbf{\Delta}_{\cdot j} Q_j^*(x,\theta) + \sum_j p_j \dot{Q}_{\cdot j}^*(x,\theta)}{\pi^*(x)} \\ &= -E\left[\frac{q_S \mathbf{\Delta}_{\cdot S}}{p_S Q_S} \middle| R=1, X=x \right] - E[\dot{\mathbf{l}}_\theta | R=1, X=x], \end{aligned} \tag{3.15}$$

where $\dot{\pi}_j \equiv \partial \pi_j / \partial Q|_{Q=Q_j} = (1-p_j)/Q_j$. Similarly,

$$\begin{aligned} \nabla_t \log Q_j(\theta, G_t) \mid_{t=\theta} &= -\frac{1}{Q_j}\sum_\ell \dot{\pi}_\ell \mathbf{\Delta}_{\cdot \ell} \left\langle \frac{Q_\ell^*}{\pi^*}, Q_j^* \right\rangle - \frac{1}{Q_j}\sum_\ell p_\ell \left\langle \frac{\dot{\mathbf{Q}}_{\cdot \ell}^*}{\pi^*}, Q_j^* \right\rangle \\ &= -E\left\{ E\left[\frac{q_S \mathbf{\Delta}_{\cdot S}}{p_S Q_S} + \dot{\mathbf{l}}_\theta(Y|X) \middle| R=1, X \right] \middle| S=j \right\}. \end{aligned} \tag{3.16}$$

Combining equations (3.11) - (3.16), we find

$$\dot{\ell}(\theta,\theta,G)(r,z) = r\psi(y,x) + (1-r)E[\psi(Y,X)|S=s] \tag{3.17}$$

where

$$\psi(y,x) = \dot{\mathbf{l}}_\theta(y|x) - E[\dot{\mathbf{l}}_\theta|R=1,X=x] - E\left[\frac{q_S \mathbf{\Delta}_{\cdot S}}{p_S Q_S} \middle| R=1, X=x \right]. \tag{3.18}$$

In view of (3.14), furthermore,

$$E(\psi|S) = \xi(S) = \frac{\mathbf{\Delta}_{\cdot S}}{Q_S} = \mathbf{\Delta}\mathrm{diag}(1/\underline{Q})\underline{\delta} = (\dot{\mathbf{Q}} - \mathbf{C})\mathbf{M}^{-1}\ \mathrm{diag}(1/\underline{q})\underline{\delta}.$$

Consequently, we see that

$$\dot{\ell}(\theta,\theta,G)(r,z) = \mathbf{l}_\theta^*(r,z) = r\psi(y,x) + (1-r)\xi(s) \tag{3.19}$$

is the efficient score given by (2.10). Equation (3.19), corresponding to MvdV's equation (9), is the key condition for a least favorable submodel.

4 Asymptotic theory via the least favorable submodel

The main goal here is to give hypotheses which imply the conditions, and hence also the conclusions, of Theorem 1 of Murphy and Van der Vaart (2000). Then we will prove a theorem giving joint asymptotic normality and efficiency of the estimators $(\widehat{\theta}_n, \widehat{G}_n)$.

The first issue is consistency. Although the models we are considering are quite closely related to those treated by Van der Vaart and Wellner (1992) (they are exactly the same if θ is known), the sufficient conditions for consistency given there fail in the present situation. In particular, (3.3) on page 138 of Van der Vaart and Wellner (1992) fails in our current setting. However, a slightly different approach yields consistency in our case. Van der Vaart and Wellner (2001) have established consistency of $(\widehat{\theta}_n, \widehat{G}_n)$. For completeness we give a brief statement of their results.

A1. $p_j > 0$ for $j = 1, \ldots, J$.

A2. The pair of parameters (θ, G) is identifiable in the model

$$\mathcal{Q} = \{Q_{\theta,G} : dQ_{\theta,G}/d(\nu \times \mu) = q(\cdot; \theta, G), \theta \in \Theta,\ G \in \mathcal{G}\},$$

where $q(y, x; \theta, G) = f(y|x; \theta)g(x)$ as in (2.1).

A3. $Q_j(\theta_0, G_0) \in (0, 1)$ for $j = 1, \ldots, J$; this holds without loss of generality.

C1. $\mathcal{X}$ is a semi-metric space that has a completion that is compact and contains $\mathcal{X}$ as a Borel set.

C2. The maps $(\theta, x) \mapsto Q_j^*(x, \theta)$ are uniformly continuous, and $\theta \mapsto f(y|x; \theta)$ are upper semicontinuous for all $(y, x) \in \mathcal{Y} \times \mathcal{X}$.

C3. Θ is a compact metric space.

C4. $P_0 \left(\sup_{\theta \in \Theta} \log(f(Y|X; \theta)/f(Y|X; \theta_0)\right) < \infty$.

Proposition 4.1. (Consistency of $(\widehat{\theta}_n, \widehat{G}_n)$). Suppose that A1-A3 and C1 - C4 hold. Then $\widehat{\theta}_n \to_{a.s.} \theta_0$ and $\sup_{h \in \mathcal{H}} |(\widehat{G}_n - G_0)h| \to_{a.s.} 0$ for every GC-class $\mathcal{H}$ that is bounded in $L_1(G_0)$.

The proof of Proposition 4.1 is given in Van der Vaart and Wellner (2001).

With consistency established, we now turn to a study of the asymptotic distributions of the profile likelihood and the the maximum likelihood estimators for the special case of VPS1 (Bernoulli) sampling. We will rely on the results of Murphy and Van der Vaart (2000) (see also Murphy and Van der Vaart (1997), Murphy and Van der Vaart (1999)). We have already verified the key condition (9) of Theorem 1 of Murphy and Van der Vaart (2000) in (3.19) of the previous section.

Now let the log-profile likelihood $\ell_n^P(\theta)$ be defined by

$$\ell_n^P(\theta) = \log L_n(\theta, \widehat{G}_n(\cdot, \theta))$$

where $\widehat{G}_n(\cdot,\theta)$ is the maximizer of $\log L_n(\theta,G)$ over distributions G concentrated at the the observed X_i's as in Lawless, Kalbfleisch, and Wild (1999). Thus for a Borel subset A of $\mathcal{X}$

$$\widehat{G}_n(A,\theta) = \mathbb{P}_n\left(1_A(X)\frac{R}{\widehat{s}_n(X,\underline{\widehat{Q}}_n(\theta),\theta)}\right) \tag{4.1}$$

where, with $N_j = n\mathbb{P}_n 1_{[S=j]}$ and $n_j = n\mathbb{P}_n(R1_{[S=j]})$ as defined in Section 2,

$$\widehat{s}_n(x,\underline{Q},\theta) = \sum_{j=1}^{J}\left(1-\frac{N_j-n_j}{nQ_j}\right)Q_j^*(x,\theta)\,, \tag{4.2}$$

and $\underline{\widehat{Q}}_n = \underline{\widehat{Q}}_n(\theta)$ satisfies

$$\underline{\widehat{Q}}_n(\theta) = \mathbb{P}_n\left(\frac{R}{\widehat{s}_n(\cdot,\underline{\widehat{Q}}_n(\theta),\theta)}\underline{Q}^*(\cdot,\theta)\right)\,. \tag{4.3}$$

Then $\widehat{\theta}_n = \text{argmax}_\theta \ell_n^P(\theta)$, and $\widehat{G}_n = \widehat{G}_n(\cdot,\widehat{\theta}_n)$.

In order to establish the remaining conditions of MvdV's Theorem 1 we assume the following:

L0. Assumptions A1-A3 and C1-C4 hold.

L1. The maps $\theta \mapsto \{\dot{l}_\theta(y|x) : y\in\mathcal{Y}, x\in\mathcal{X}\}$, $\theta\mapsto Q^*(\cdot,\theta)$ and $\theta\mapsto \dot{Q}^*(\cdot,\theta)$ are all Lipschitz in the sense that, for all t,s in a neighborhood of θ_0 and all $(y,x)\in\mathcal{Y}\times\mathcal{X}$:

$$|\dot{l}_t(y|x)-\dot{l}_s(y|x)| \le M(y,x)|t-s| \qquad \text{where}$$

$$P_0M^2 = \int M^2(y,x)f(y|x,\theta_0)d\nu(y)dG_0(x) < \infty\ ;$$

$$|Q^*(x,t)-Q^*(x,s)| \le M(x)|t-s| \qquad \text{where}$$

$$G_0|M|^2 = \int M^2(x)dG_0(x) < \infty; \qquad \text{and}$$

$$|\dot{Q}^*(x,t)-\dot{Q}^*(x,s)| \le \dot{M}(x)|t-s| \qquad \text{where}$$

$$G_0|\dot{M}|^2 = \int \dot{M}^2(x)dG_0(x) < \infty.$$

L2. For some $\delta_0>0$ the collections of functions

$$\{r\ddot{l}_{\theta,k,l}(y|x) :\ |t-\theta_0|\le\delta_0, k,l=1,\ldots,m\}$$

and

$$\{r\ddot{Q}^*_{j,k,l}(x,t) :\ |t-\theta_0|\le\delta_0, j=1,\ldots,J,\ k,l=1,\ldots,m\}$$

are P_0-Glivenko-Cantelli classes of functions.

L3. There is no m-vector a such that $a^T \dot{\mathbf{l}}_\theta(Y|X)$ is constant in Y for $G-$ a.e X. (Equivalently, the information matrix for θ with no missing data given in (2.16) is nonsingular.)

Theorem 4.1. Suppose that **L0 - L3** hold. Then for any random sequence $\tilde{\theta}_n \to_p \theta_0$ it follows that

$$\begin{aligned} \ell_n^P(\tilde{\theta}_n) &= \ell_n^P(\theta_0) + (\tilde{\theta}_n - \theta_0)^T \sum_{i=1}^n \mathbf{l}_\theta^*(R_i, Z_i) - \frac{1}{2} n(\tilde{\theta}_n - \theta_0)^T I(\theta_0)(\tilde{\theta}_n - \theta_0) \\ &\quad + o_p(\sqrt{n}\|\tilde{\theta}_n - \theta_0\| + 1)^2 \end{aligned} \tag{4.4}$$

for any random sequence $\tilde{\theta}_n \to_p \theta_0$ where $\mathbf{l}_\theta^*$ is given by (2.10) and $I(\theta_0)$ is given by (2.11).

As shown by Murphy and Van der Vaart (2000) in their Corollaries 1 and 2, the expansion (4.4) together with invertibility of $I(\theta_0)$ implies $\widehat{\theta}_n$ is asymptotically linear with efficient influence function $\widetilde{\mathbf{l}}_\theta = I(\theta_0)^{-1}\mathbf{l}_\theta^*$ given by (2.12):

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n I(\theta_0)^{-1} \mathbf{l}_\theta^*(R_i, Z_i) + o_p(1)\,. \tag{4.5}$$

Moreover the expansion

$$\begin{aligned} \ell_n^P(\tilde{\theta}_n) &= \ell_n^P(\widehat{\theta}_n) + (\tilde{\theta}_n - \widehat{\theta}_n)^T \sum_{i=1}^n \mathbf{l}_\theta^*(R_i, Z_i) - \frac{1}{2} n(\tilde{\theta}_n - \widehat{\theta}_n)^T I(\theta_0)(\tilde{\theta}_n - \widehat{\theta}_n) \\ &\quad + o_p(\sqrt{n}\|\tilde{\theta}_n - \theta_0\| + 1)^2\,, \end{aligned} \tag{4.6}$$

also holds, and the likelihood ratio statistic based on the profile likelihood is asymptotically χ^2_m:

$$2\{\ell_n^P(\widehat{\theta}_n) - \ell_n^P(\theta_0)\} \to_d \chi^2_m\,. \tag{4.7}$$

Proof of Theorem 4.1: We begin by verifying conditions (8)-(10) of Murphy and Van der Vaart (2000). Condition (8) of Murphy and Van der Vaart (2000) is indeed satisfied by the submodel $(t, G_t(\theta, G))$ given by (3.9): $(t, G_t(\theta, G))$ passes through (θ, G) at $t = \theta$. We have already seen in (3.19) that condition (9) of Murphy and Van der Vaart (2000) holds. Condition (10) of Murphy and Van der Vaart (2000) requires that $\widehat{G}_n(\theta)$, the maximizer of the log-likelihood over G for fixed θ, satisfies

$$\widehat{G}_n(\tilde{\theta}_n) \to_p G_0 \tag{4.8}$$

for every random sequence $\tilde{\theta}_n$ with $\tilde{\theta}_n \to_p \theta_0$. This holds by virtue of the arguments in Van der Vaart and Wellner (2001), pages 281 and 282.

We will postpone verification of condition (11) of Murphy and Van der Vaart (2000); in fact, verification of this condition will occupy most of our proof.

To verify the Donsker and Glivenko-Cantelli hypotheses of Theorem 1 of MURPHY AND VAN DER VAART (2000), we need first to compute the functions $\dot{l}_t(t,\theta,G)$ and $\ddot{l}_t(t,\theta,G)$. The log-likelihood for one observation for the least favorable submodel is given by $l(t,\theta,G)$ in (3.11) where $Q_j^\dagger(t) \equiv Q_j(t, G_t(\theta,G))$, $j = 1,\ldots,J$. We need to calculate the first and second derivatives of this function with respect to t. To do this, we first calculate the first and second derivatives of $\log(dG_t/dG)$:

$$\begin{aligned}
&\nabla_t \log\left(\frac{dG_t}{dG}(x;\theta,G)\right) \\
&= \frac{-\sum_{j=1}^J \nabla_t \left\{\left(1-\frac{Q_j(1-p_j)}{Q_j^\dagger(t)}\right) Q_j^*(x,t)\right\}}{\sum_{j=1}^J \left(1-\frac{Q_j(1-p_j)}{Q_j^\dagger(t)}\right) Q_j^*(x,t)} \\
&= -\frac{\sum_{j=1}^J \left(1-\frac{Q_j(1-p_j)}{Q_j^\dagger(t)}\right) \dot{Q}_j^*(x,t) + \sum_{j=1}^J Q_j^*(x,t)\frac{Q_j(1-p_j)}{[Q_j^\dagger(t)]^2}\nabla_t Q_j^\dagger(t)}{\sum_{j=1}^J \left(1-\frac{Q_j(1-p_j)}{Q_j^\dagger(t)}\right) Q_j^*(x,t)}\,. \qquad (4.9)
\end{aligned}$$

Here the derivative vector $\nabla_t Q_j^\dagger(t)$ satisfies a linear equation which can be derived by differentiating across (3.10); see (7.3). Note that this is basically a ratio of a (family of) linear combination(s) of the functions $Q_j^*(\cdot,t)$, $\dot{Q}_j^*(\cdot,t)$ and the family of functions s_t given by

$$s_t(x) \equiv \sum_{j=1}^J \left(1-\frac{Q_j(1-p_j)}{Q_j^\dagger(t)}\right) Q_j^*(x,t)\,. \qquad (4.10)$$

We also define

$$s_0(x,\underline{Q},\theta) = \sum_{j=1}^J \left(1-\frac{Q_j^0(1-p_j)}{Q_j}\right) Q_j^*(x,\theta) \qquad (4.11)$$

where $Q_j^0 = Q_j(\theta_0, G_0)$. Note that $s_{\theta_0}(x;\theta_0,G_0) = \pi^*(x) = s_0(x,\underline{Q}^0,\theta_0)$ for all $x \in \mathcal{X}$ with π^* as defined in (2.7). Thus we also write (in a slight abuse of notation) s_0 instead of π^*. Calculation of the second derivatives yields

$$\begin{aligned}
&\nabla_t\left(\nabla_t \log\left(\frac{dG_t}{dG}(x;\theta,G)\right)\right) \\
&= \left\{\nabla_t \log\left(\frac{dG_t}{dG}(x;\theta,G)\right)\right\}^{\otimes 2} - \sum_{j=1}^J \left\{\left(1-\frac{Q_j(1-p_j)}{Q_j^\dagger(t)}\right)\ddot{Q}_j^*(x,t) \qquad (4.12)\right. \\
&\quad + \frac{Q_j(1-p_j)}{[Q_j^\dagger(t)]^2}(\dot{Q}_j^*(x,t))^{\otimes 2} + \dot{Q}_j^*(x,t)\frac{Q_j(1-p_j)}{[Q_j^\dagger(t)]^2}\nabla_t Q_j^\dagger(t) \\
&\quad \left. - 2Q_j^*(x,t)\frac{Q_j(1-p_j)}{[Q_j^\dagger(t)]^3}(\nabla_t Q_j^\dagger(t))^{\otimes 2} + Q_j^*(x,t)\frac{Q_j(1-p_j)}{[Q_j^\dagger(t)]^2}\ddot{Q}_j^\dagger(t)\right\}/s_t(x)
\end{aligned}$$

where s_t is defined in (4.10). Thus we find that

$$\begin{aligned}\dot{l}(t,\theta,G)(r,z) &\equiv \nabla_t l(t,\theta,G)(r,z) \\ &= r\left(\dot{l}_t(y|x) + \nabla_t \log\left(\frac{dG_t}{dG}(x;\theta,G)\right)\right) + (1-r)\sum_{j=1}^{J}\delta_j \frac{\nabla_t Q_j(t,G_t)}{Q_j(t,G_t)} \quad (4.13)\end{aligned}$$

and

$$\begin{aligned}\ddot{l}(t,\theta,G)(r,z) &\equiv \nabla_t \dot{l}(t,\theta,G)(r,z) \quad (4.14) \\ &= r\left(\ddot{l}_t(y|x) + \nabla_t\left(\nabla_t \log\left(\frac{dG_t}{dG}(x;\theta,G)\right)\right)\right) \\ &\quad + (1-r)\sum_{j=1}^{J}\delta_j\left\{\frac{\ddot{Q}_j^{\dagger}(t)}{Q_j^{\dagger}(t)} - \frac{[\dot{Q}_j^{\dagger}(t)]^2}{[Q_j^{\dagger}(t)]^2}\right\}. \quad (4.15)\end{aligned}$$

We now show that there is a neighborhood V of (θ_0,θ_0,G_0) such that the classes of functions

$$\{\dot{l}_k(t,\theta,G) : (t,\theta,G)\in V,\ k=1,\ldots,m\}$$

with $\dot{l}(t,\theta,G)$ as given by (4.13) and (4.9) are P_0-Donsker with square integrable envelope function. First note that by **L1** the collections

$$\{r\dot{l}_{k,t} : (t,\theta,G)\in V,\ k=1,\ldots,m\},$$

$$\{rQ_j^*(\cdot,t) : (t,\theta,G)\in V, j=1,\ldots,J\}$$

and

$$\{r\dot{Q}_{jk}^*(\cdot,t) : (t,\theta,G)\in V, j=1,\ldots,J, k=1,\ldots,m\}$$

are P_0-Donsker by virtue of the Jain-Marcus CLT (see Example 2.11.13, page 213, Van der Vaart and Wellner (1996)). Then, since products of these functions with bounded families of constants are also P_0-Donsker by an application of Corollary 2.10.13, page 193, Van der Vaart and Wellner (1996), the individual terms appearing in the numerator of (4.9) are also P_0-Donsker, and hence also their sum by application of Example 2.10.7, page 192, Van der Vaart and Wellner (1996). Then, since $s_0(x)$ is bounded uniformly away from zero by A1 and $\sum_j Q_j^*(x,\theta_0)=1$, $s_t(x)$ is also bounded away from zero uniformly in x and t in a sufficiently small neighborhood of θ_0. Hence the ratio appearing in (4.9) is also P_0-Donsker by virtue of Example 2.10.9, page 192, Van der Vaart and Wellner (1996). Furthermore, the neighborhood V of (θ_0,θ_0,G_0) can be chosen so that the class of functions

$$\{\ddot{l}(t,\theta,G) : (t,\theta,G)\in V\}$$

with $\ddot{l}(t,\theta,G)$ given by (4.15) and (4.12) is P_0-Glivenko-Cantelli with integrable envelope function. This follows from **L2** (to handle the terms involving $\ddot{l}_t(y|x)$ and $\ddot{Q}_{j,k,l}^*(x,t)$), Lemma 2.10.14, page 194, Van der Vaart and Wellner (1996), and the Glivenko-Cantelli preservation theorem of Van der Vaart and Wellner (2000) to handle the remaining terms.

We now turn our attention to verification of the remaining condition (11) of MURPHY AND VAN DER VAART (2000). The discussion leading to MdvV's (16) applies so that, in place of their condition (11), it suffices to verify that for convergent sequences $\tilde{\theta}_n$

$$P_0\dot{\ell}(\theta_0, \theta_0, \widehat{G}_n(\tilde{\theta}_n)) = o_p(\|\tilde{\theta}_n - \theta_0\| + n^{-1/2})\,. \tag{4.16}$$

As argued by MvdV, page 458, (4.16) can be shown to hold if their display (18) holds; in our context their display (18) becomes

$$\|\widehat{G}_n(\tilde{\theta}_n) - G_0\|_{\mathcal{H}} = O_p(\|\tilde{\theta}_n - \theta_0\|) + O_p(n^{-1/2}) \tag{4.17}$$

where $\mathcal{H}$ is a universal-Donsker class of real-valued (measurable) functions on $\mathcal{X}$. Once again the key observation comes from LKW: to find $\widehat{G}_n$ one does not need to estimate all of G, but just the quantities $Q_j(\theta, G)$ which, for fixed θ, are determined by the equations (3.10). We view this system of equations as processes in θ in a neighborhood of θ_0 and show that the convergence of the corresponding $\widehat{Q}_j(\theta)$ processes is uniform in θ. Towards this end, consider

$$\Psi_n(\underline{Q})(\theta) \equiv \underline{Q} - \mathbb{P}_n\left(\frac{R}{\widehat{s}_n(\cdot, \underline{Q}, \theta)}\underline{Q}^*(\cdot, \theta)\right)\,,$$

and

$$\Psi(\underline{Q})(\theta) \equiv \underline{Q} - P_0\left(\frac{R}{s_0(\cdot, \underline{Q}, \theta)}\underline{Q}^*(\cdot, \theta)\right)\,;$$

here $\widehat{s}_n$ is given by (4.2) and $s_0(\cdot, \underline{Q}, \theta)$ is given by (4.11). Note that $\Psi_n(\underline{\widehat{Q}}_n(\theta))(\theta) = \underline{0}$ defines $\underline{\widehat{Q}}_n(\theta)$, while $\Psi(\underline{Q}_0(\theta))(\theta) = \underline{0}$ defines $\underline{Q}_0(\theta)$. Also note that $\underline{Q}_0(\theta_0) = \underline{Q}^0 = \underline{Q}(\theta_0, G_0)$.

Proposition 4.2. Suppose that **L0 - L3** hold. Then for some (sufficiently small) closed ball $B(\theta_0)$ in R^m centered at θ_0,

$$\sqrt{n}(\underline{\widehat{Q}}_n(\theta) - \underline{Q}_0(\theta)) \Rightarrow \mathbb{Q}(\theta) \qquad \text{in } \; C[B(\theta_0)]^J \tag{4.18}$$

where $\mathbb{Q}(\theta)$ is a zero mean Gaussian process.

Once we have proved (4.18), the next step is to show that (4.17) holds for any random sequence $\tilde{\theta}_n \to_p \theta_0$. In other words, we want to show that

$$\|\sqrt{n}(\widehat{G}_n(h; \tilde{\theta}_n) - G_0(h))\|_{\mathcal{H}} = O_p(1) + \sqrt{n}(\tilde{\theta}_n - \theta_0)\,. \tag{4.19}$$

To this end, we first abbreviate notation slightly: $\widehat{s}_n(x, \underline{\widehat{Q}}_n(\tilde{\theta}_n), \tilde{\theta}_n) \equiv \widehat{s}_n(x, \tilde{\theta}_n)$. Then we have

$$\begin{aligned}
\sqrt{n}(\widehat{G}_n(h; \tilde{\theta}_n) - G_0(h)) &= \sqrt{n}\left(\mathbb{P}_n\left(\frac{R}{\widehat{s}_n(\cdot, \tilde{\theta}_n)}h\right) - P_0\left(\frac{R}{s_0}h\right)\right) \\
&= \sqrt{n}\left(\mathbb{P}_n\left(\frac{R}{s_0}h\right) - P_0\left(\frac{R}{s_0}h\right)\right) + \sqrt{n}\mathbb{P}_n\left(Rh\left(\frac{1}{\widehat{s}_n(\cdot, \tilde{\theta}_n)} - \frac{1}{s_0}\right)\right) \\
&= \mathbb{G}_n\left(\frac{R}{s_0}h\right) - \mathbb{P}_n\left(Rh\frac{\sqrt{n}(\widehat{s}_n(\cdot, \tilde{\theta}_n) - s_0)}{\widehat{s}_n(\cdot, \tilde{\theta}_n)s_0}\right) \\
&\equiv I_n(h) - II_n(h)\,;
\end{aligned} \tag{4.20}$$

Here $\mathbb{G}_n \equiv \sqrt{n}(\mathbb{P}_n - P_0)$ is the empirical process, and hence $\|I_n\|_{\mathcal{H}} = O_p(1)$ easily via standard theory. To understand the term $II_n(h)$, we write

$$\begin{aligned}
II_n(h) &= -\mathbb{P}_n\left(\frac{Rh}{s_0\widehat{s}_n(\cdot,\tilde{\theta}_n)}\sqrt{n}\sum_{j=1}^{J}\left(\frac{N_j-n_j}{n}\frac{1}{\widehat{Q}_j(\tilde{\theta}_n)}Q_j^*(\cdot,\tilde{\theta}_n) - \frac{Q_j^0(1-p_j)}{Q_j^0}Q_j^*(\cdot,\theta_0)\right)\right) \\
&\quad + \mathbb{P}_n\left(\frac{Rh}{s_0\hat{s}_n(\cdot,\tilde{\theta}_n)}\sum_{j=1}^{J}\sqrt{n}(Q_j^*(\cdot,\tilde{\theta}_n) - Q_j^*(\cdot,\theta_0))\right) \\
&= -\sum_{j=1}^{J}\sqrt{n}\left(\frac{N_j-n_j}{n} - Q_j^0(1-p_j)\right)\frac{1}{\widehat{Q}_j(\tilde{\theta}_n)}\mathbb{P}_n\left(\frac{Rh}{s_0\widehat{s}_n(\cdot,\tilde{\theta}_n)}Q_j^*(\cdot,\tilde{\theta}_n)\right) \\
&\quad + \sum_{j=1}^{J}\sqrt{n}\left\{\frac{\widehat{Q}_j(\tilde{\theta}_n) - Q_j(\tilde{\theta}_n)}{\widehat{Q}_j(\tilde{\theta}_n)Q_j^0}\right\}Q_j^0(1-p_j)\mathbb{P}_n\left(\frac{Rh}{s_0\widehat{s}_n(\cdot,\tilde{\theta}_n)}Q_j^*(\cdot,\tilde{\theta}_n)\right) \\
&\quad + \sum_{j=1}^{J}\sqrt{n}\left\{\frac{Q_j(\tilde{\theta}_n) - Q_j^0}{\widehat{Q}_j(\tilde{\theta}_n)Q_j^0}\right\}Q_j^0(1-p_j)\mathbb{P}_n\left(\frac{Rh}{s_0\widehat{s}_n(\cdot,\tilde{\theta}_n)}Q_j^*(\cdot,\tilde{\theta}_n)\right) \\
&\quad + \mathbb{P}_n\left(\frac{Rh}{s_0\widehat{s}_n(\cdot,\tilde{\theta}_n)}\right)\sum_{j=1}^{J}p_j\dot{Q}_j^*(\cdot,\theta_n^\sharp)\sqrt{n}(\tilde{\theta}_n-\theta_0) \qquad (4.21) \\
&\equiv A_n(h) + B_n(h) + C_n(h) + D_n(h)\,.
\end{aligned}$$

Here $Q_j(\tilde{\theta}_n) \equiv Q_{j0}(\tilde{\theta}_n)$ satisfies $\Psi(\underline{Q}_0(\tilde{\theta}_n))(\tilde{\theta}_n) = 0$. Now $\|A_n\|_{\mathcal{H}} = O_p(1)$ by $\sqrt{n}((N_j-n_j)/n - Q_j^0(1-p_j)) = O_p(1)$, consistency of $\widehat{G}_n$, and uniform (in θ) convergence of $\underline{\widehat{Q}}_n(\theta)$ in a neighborhood of θ_0; $\|B_n\|_{\mathcal{H}} = O_p(1)$ by Proposition 4.2 and consistency of $\widehat{G}_n$; $\|C_n\|_{\mathcal{H}} = O_p(\sqrt{n}(\tilde{\theta}_n - \theta_0))$ by differentiability of the maps $\theta \mapsto Q_{0j}(\theta)$ and consistency of $\widehat{G}_n$; and $\|D_n\|_{\mathcal{H}} = O_p(\sqrt{n}(\tilde{\theta}_n - \theta_0))$ easily by consistency of $\widehat{G}_n$.

But in view of the differentiability of $\underline{Q}^\dagger(\theta_0,\theta,G_0) \equiv \underline{Q}_0(\theta)$ with respect to θ proved in (7.2) (using the definition of $Q_j^\dagger(t,\theta,G)$ following (3.10), we have (by the mean-value theorem),

$$\sqrt{n}(\underline{Q}_0(\tilde{\theta}_n) - \underline{Q}_0(\theta_0)) = \nabla_\theta \underline{Q}_0(\theta)\big|_{\theta=\theta_n^\sharp}\cdot\sqrt{n}(\tilde{\theta}_n-\theta_0)\,,$$

and thus we see, by combining I_n and II_n together with a bit more Glivenko-Cantelli that (4.19) holds. □

Proof of nonsingularity of $I(\theta_0)$ and (4.5)-(4.7): We now prove that $I(\theta_0)$ is non-singular and hence, via Corollaries 1 and 2 of Murphy and Van der Vaart (2000), that (4.5)- (4.7) hold. Recall the formula for $I(\theta_0)$ given in (2.11):

$$I(\theta_0) = E_0\left\{R\left(\dot{\mathrm{l}}_\theta(Y|X) - \frac{\dot{\mathbf{Q}}^*}{\pi^*}(X)\underline{p}\right)^{\otimes 2}\right\} + (\dot{\mathbf{Q}}-\mathbf{C})\mathbf{M}^{-1}(\dot{\mathbf{Q}}-\mathbf{C})^T\,. \qquad (4.22)$$

It is clear from the form of the two terms in (4.22) that each is non-negative definite. Thus, to be invertible, at least one term must be positive definite.

In the first term in (4.22) we have

$$\dot{\mathbf{Q}}^*(X)\underline{p} = \sum_{j=1}^{J} p_j \underline{\dot{Q}}_j^*(X) = E_0\left(R\dot{\mathbf{i}}_\theta(Y|X) \mid X\right) \quad \text{since} \quad \underline{\dot{Q}}_j^*(X) = E_0(\delta_j \dot{\mathbf{i}}_\theta(Y|X) \mid X)$$

and we recall that

$$\frac{\dot{\mathbf{Q}}^*}{\pi^*}(X)\underline{p} = E_0(\dot{\mathbf{i}}_\theta \mid X, R=1).$$

Consider quadratic forms of this matrix with an arbitrary m-vector a. We have

$$\begin{aligned}
& a^T E_0\left\{ R\left(\dot{\mathbf{i}}_\theta(Y|X) - \frac{E_0(R\dot{\mathbf{i}}_\theta(Y|X) \mid X)}{\pi^*(X)}\right)^{\otimes 2}\right\} a \\
&= E_0\left\{\sum_{j=1}^{J} p_j \delta_j \, E_0\left(\left(a^T\dot{\mathbf{i}}_\theta(Y|X) - \frac{E_0(Ra^T\dot{\mathbf{i}}_\theta(Y|X) \mid X)}{\pi^*(X)}\right)^2 \middle| \underline{\delta}\right)\right\} \\
&\geq \min_j p_j E_0\left\{\left(a^T\dot{\mathbf{i}}_\theta(Y|X) - \frac{E_0(Ra^T\dot{\mathbf{i}}_\theta(Y|X) \mid X)}{\pi^*(X)}\right)^2\right\}
\end{aligned}$$

which is 0 if and only if

$$a^T\dot{\mathbf{i}}_\theta(Y|X) = \frac{E_0(Ra^T\dot{\mathbf{i}}_\theta(Y|X) \mid X)}{\pi^*(X)} \quad P_0\text{-a.s.}$$

This would require $a^T\dot{\mathbf{i}}_\theta(Y|X)$ to be constant in Y in which case the equality follows from the fact that $E_0(R \mid X) = \pi^*(X)$. Thus **L3** implies that the first term in (4.22) is positive definite, and hence $I(\theta_0)$ is invertible. □

Proof of Proposition 4.2: We will apply Van der Vaart's $Z-$theorem; see Van der Vaart (1995) and Van der Vaart and Wellner (1996), Theorem 3.3.1, page 310. To this end, note that

$$\begin{aligned}
\mathbb{Z}_n(\underline{Q})(\theta) &\equiv \sqrt{n}(\Psi_n(\underline{Q})(\theta) - \Psi(\underline{Q})(\theta)) \\
&= -\mathbb{G}_n\left(\frac{R}{s_0(\cdot,\underline{Q},\theta)}\underline{Q}^*(\cdot,\theta)\right) \\
&\quad - \mathbb{P}_n\left(R\frac{\sum_{j=1}^{J}\sqrt{n}\left(\frac{N_j-n_j}{n} - Q_j^0(1-p_j)\right)Q_j^*(\cdot,\theta)/Q_j}{\widehat{s}_n(\cdot,\underline{Q},\theta)s_0(\cdot,\underline{Q},\theta)}\underline{Q}^*(\cdot,\theta)\right) \\
&= -\mathbb{G}_n\left(\frac{R}{s_0(\cdot,\underline{Q},\theta)}\underline{Q}^*(\cdot,\theta)\right) \\
&\quad - \sum_{j=1}^{J}\sqrt{n}\left(\frac{N_j-n_j}{n} - Q_j^0(1-p_j)\right)\mathbb{P}_n\left(\frac{RQ_j^*(\cdot,\theta)/Q_j}{\widehat{s}_n(\cdot,\underline{Q},\theta)s_0(\cdot,\underline{Q},\theta)}\underline{Q}^*(\cdot,\theta)\right).
\end{aligned}$$

Now it follows easily from **L1** and the Jain-Marcus CLT that

$$\mathbb{Z}_n(\underline{Q}_0(\theta))(\theta) \Rightarrow \mathbb{Z}_0(\theta)$$

as a (vector of) process(es) indexed by $\theta \in B(\theta_0)$, and that

$$\sup_{\underline{Q} \in Lip[B(\theta_0)]^J : \|\underline{Q}(\theta) - \underline{Q}_0(\theta)\| \le \delta_n} \|\mathbb{Z}_n(\underline{Q}) - \mathbb{Z}_n(\underline{Q}_0)\|_{B(\theta_0)} = o_p(1)$$

for every sequence $\delta_n \to 0$. Furthermore,

$$\begin{aligned} \nabla_Q \Psi(Q) &= I + P_0 \left(\frac{R}{s_0(\cdot, Q, \theta)^2} \underline{Q}^*(\cdot, \theta) \nabla_Q s_0(\cdot, Q, \theta) \right) \\ &= I + P_0 \left(\frac{R}{s_0(\cdot, Q, \theta)^2} \underline{Q}^*(\cdot, \theta) \mathrm{diag}(Q_j^0(1 - p_j)/Q_j^2) \underline{Q}^*(\cdot, \theta)^T \right) . \end{aligned}$$

is always nonsingular, and hence via the chain rule we see that the derivative map $\dot{\Psi} : Lip(B(\theta_0))^J \mapsto Lip(B(\theta_0))^J$ exists and has a bounded inverse at $\underline{Q}_0 = \{\underline{Q}_0(\theta) : \theta \in B(\theta_0)\}$. □

5 Joint Asymptotic Normality and Efficiency of $(\widehat{\theta}_n, \widehat{G}_n)$, Bernoulli sampling

We now turn to a study of the joint asymptotic distributions of the maximum likelihood estimators $(\widehat{\theta}_n, \widehat{G}_n)$ for the special case of VPS1 (Bernoulli) sampling. The density for the data under VPS1 given in (2.3) is our starting point. We will use the infinite-dimensional $Z-$theorem given in Van der Vaart (1995), Van der Vaart and Wellner (1996), pages 314 - 319, and Van der Vaart (1998), section 25.12, primarily as a way to organize the statement of the theorem. In fact the proof will use the development of Section 4.

Our first job is to calculate the score functions Ψ_n (using the notation of Van der Vaart (1998)). It follows from (2.3) that

$$\begin{aligned} \log L_n(\theta, G) &= \sum_{j=1}^{J} \left\{ \sum_{i \in D_j} \Big[\log f(Y_i | X_i; \theta) + \log g(X_i) \Big] + (N_j - n_j) \log Q_j(\theta, G) \right\} \\ &= \sum_{i=1}^{n} \left\{ R_i \left(\log f(Y_i | X_i; \theta) + \log g(X_i) \right) + (1 - R_i) \sum_{j=1}^{J} \delta_{ij} \log Q_j(\theta, G) \right\} . \quad (5.1) \end{aligned}$$

With the notation as in Proposition 2.1, this yields

$$\begin{aligned}\Psi_{n1}(\theta, G) &\equiv \frac{1}{n}\dot{\mathbf{l}}_{n\theta}(\theta, G) \equiv \frac{1}{n}\nabla_\theta \log L_n(\theta, G) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{ R_i \dot{\mathbf{l}}_\theta(Y_i|X_i) + (1-R_i)\sum_{j=1}^{J}\delta_{ij}\frac{\dot{Q}_j(\theta, G)}{Q_j(\theta, G)}\right\} \\ &= \mathbb{P}_n\left(R\dot{\mathbf{l}}_\theta(Y|X) + (1-R)\frac{\dot{Q}_S(\theta, G)}{Q_S(\theta, G)}\right),\end{aligned}$$

so that the MLE $(\widehat{\theta}, \widehat{G})$ of (θ, G) satisfies $\Psi_{n1}(\widehat{\theta}, \widehat{G}) = 0$. Now let $\widehat{G}$ be the MLE of G and, for any bounded real-valued function h on $\mathcal{X}$, let

$$d\widehat{G}_t \equiv (1 + t(h - \int h d\widehat{G}))d\widehat{G}\,.$$

Then, with

$$\Psi_{n2}(\theta, G)(h) = \mathbb{P}_n A_{\theta,G}h - P_{\theta,G}A_{\theta,G}h\,, \tag{5.2}$$

where $A_{\theta,G}$ is given by (2.5), we find that the MLE $(\widehat{\theta}, \widehat{G})$ of (θ, G) also satisfies

$$\begin{aligned}0 = \frac{1}{n}\dot{l}_t(\widehat{\theta}, \widehat{G}) &= \frac{1}{n}\sum_{i=1}^{n}\left\{ R_i(h(X_i) - \int h d\widehat{G}) + (1-R_i)\sum_{j=1}^{J}\delta_{ij}\frac{\int Q_j^*(x, \widehat{\theta})(h(x) - \int h d\widehat{G})d\widehat{G}}{Q_j(\widehat{\theta}, \widehat{G})}\right\} \\ &= \mathbb{P}_n A_{\widehat{\theta},\widehat{G}}h - P_{\widehat{\theta},\widehat{G}}A_{\widehat{\theta},\widehat{G}}h = \Psi_{n2}(\widehat{\theta}, \widehat{G})(h).\end{aligned} \tag{5.3}$$

The population version of the score for θ is

$$\begin{aligned}\Psi_1(\theta, G) &= P_0(\dot{\mathbf{l}}_\theta(\theta, G)) = P_0(R\dot{\mathbf{l}}_\theta(Y|X) + (1-R)\sum_{j=1}^{J}\delta_j\frac{\dot{Q}_j(\theta, G)}{Q_j(\theta, G)}) \\ &= P_0\left(R\dot{\mathbf{l}}_\theta(Y|X) + \dot{\mathbf{Q}}(\theta, G)\mathrm{diag}(Q^{-1})\underline{\delta}(1-R)\right).\end{aligned} \tag{5.4}$$

Similarly, the population version of the score for G is given by

$$\Psi_2(\theta, G)(h) = P_0 A_{\theta,G}h - P_{\theta,G}A_{\theta,G}h\,. \tag{5.5}$$

Under the hypotheses of the following theorem, $\Psi = (\Psi_1, \Psi_2)$ is differentiable in a suitably strong sense with derivative $\dot{\Psi} : (R^m \times \ell^\infty(\mathcal{H})) \to (R^m \times \ell^\infty(\mathcal{H}))$ at $(\theta_0, G_0) \in \Theta \times \mathcal{G}$ given by

$$\dot{\Psi}\begin{pmatrix}\theta - \theta_0 \\ G - G_0\end{pmatrix} = \begin{pmatrix}\dot{\Psi}_{11} & \dot{\Psi}_{12} \\ \dot{\Psi}_{21} & \dot{\Psi}_{22}\end{pmatrix}\begin{pmatrix}\theta - \theta_0 \\ G - G_0\end{pmatrix},$$

where

$$\begin{aligned}
\dot\Psi_{11} : R^m \to R^m \qquad &\text{is given by} \qquad \dot\Psi_{11}(\theta-\theta_0) = -I_{11}(\theta_0)(\theta-\theta_0)\,, \\
\dot\Psi_{12} : \ell^\infty(\mathcal{H}) \to R^m \qquad &\text{is given by} \qquad \dot\Psi_{12}(G-G_0) = -\int A_0^T \dot{\mathbf{l}}_\theta d(G-G_0)\,, \\
\dot\Psi_{21} : R^m \to \ell^\infty(\mathcal{H}) \qquad &\text{is given by} \qquad \dot\Psi_{21}(\theta-\theta_0)h = -(\theta-\theta_0)^T \int A_0 h \dot{\mathbf{l}}_\theta dP_0\,, \\
\dot\Psi_{22} : \ell^\infty(\mathcal{H}) \to \ell^\infty(\mathcal{H}) \qquad &\text{is given by} \qquad \dot\Psi_{22}(G-G_0)h = -\int A_0^T A_0 h d(G-G_0)\,.
\end{aligned} \tag{5.6}$$

Here

$$\begin{aligned}
I_{11}(\theta_0) &= P_0(\dot{\mathbf{l}}_\theta^{\otimes 2}(R,Z;\theta_0,G_0)) \\
&= E_0\left\{\pi(Y,X)\dot{\mathbf{l}}_\theta^{\otimes 2}(Y|X)\right\} + \dot{\mathbf{Q}}\mathrm{diag}(q/Q_0)\dot{\mathbf{Q}}^T
\end{aligned} \tag{5.7}$$

with $\dot{\mathbf{l}}_\theta(R,Z;\theta_0,G_0)$ given by (2.4), is the information for θ when G is known.

It is shown in Section 2 of BMW that the information operator $A_0^T A_0 = \dot{\mathbf{l}}_g^T \dot{\mathbf{l}}_g$ given by

$$A_0^T A_0 h(x) = \pi^*(x)h(x) + \underline{Q}^{*T}(x)\mathrm{diag}(q/Q)\langle \underline{Q}^*, h\rangle \tag{5.8}$$

is invertible with

$$(A_0^T A_0)^{-1}h(x) = \frac{1}{\pi^*(x)}h(x) - \left\langle h, \frac{\underline{Q}^{*T}}{\pi^*}\right\rangle M^{-1}\frac{\underline{Q}^*(x)}{\pi^*(x)}\,.$$

From this it follows, using standard formulae for inverses of operators defined in blocks as above (the same as for block-matrices), that the inverse $\dot\Psi_0^{-1} : (R^m \times \ell^\infty(\mathcal{H})) \to (R^m \times \ell^\infty(\mathcal{H}))$ exists, is continuous, and is given by

$$\dot\Psi^{-1} = \begin{pmatrix} \dot V^{-1} & -\dot V^{-1}\dot\Psi_{12}\dot\Psi_{22}^{-1} \\ -\dot\Psi_{22}^{-1}\dot\Psi_{21}\dot V^{-1} & \dot\Psi_{22}^{-1}\left(\dot\Psi_{22} + \dot\Psi_{21}\dot V^{-1}\dot\Psi_{12}\right)\dot\Psi_{22}^{-1} \end{pmatrix} \tag{5.9}$$

where $\dot V = \dot\Psi_{11} - \dot\Psi_{12}\dot\Psi_{22}^{-1}\dot\Psi_{21}$ and

$$\begin{aligned}
\dot\Psi_{12}\dot\Psi_{22}^{-1}\dot\Psi_{21}(\theta-\theta_0) &= -\int A_0(A_0^TA_0)^{-1}A_0^T\dot{\mathbf{l}}_\theta \;\; \dot{\mathbf{l}}_\theta^T(\theta-\theta_0)dP_0 \\
&= -\int (A_0^TA_0)^{-1}A_0^T\dot{\mathbf{l}}_\theta \;\; A_0^T\dot{\mathbf{l}}_\theta^T(\theta-\theta_0)dG_0 \\
&= -\int (A_0^TA_0)^{-1}A_0^T\dot{\mathbf{l}}_\theta \;\; (A_0^TA_0)(A_0^TA_0)^{-1}A_0^T\dot{\mathbf{l}}_\theta^T(\theta-\theta_0)dG_0 \\
&= -\int A_0(A_0^TA_0)^{-1}A_0^T\dot{\mathbf{l}}_\theta \;\; A_0(A_0^TA_0)^{-1}A_0^T\dot{\mathbf{l}}_\theta^T(\theta-\theta_0)dP_0 \\
&= -E_0\left(A_0(A_0^TA_0)^{-1}A_0^T\dot{\mathbf{l}}_\theta^{\otimes 2}\right)(\theta-\theta_0).
\end{aligned}$$

Note that (5.9) is *not* the same as the block inverse form in VAN DER VAART (1998), page 422. Thus

$$\dot{V} = -I(\theta_0) = -I_{11} + E_0\left(A_0(A_0^T A_0)^{-1}A_0^T \dot{\mathbf{l}}_\theta^{\otimes 2}\right) = -E_0\left(\mathbf{l}_{\theta_0}^{*\otimes 2}\right)\,,$$

which equals minus one times the efficient information matrix given in (2.11).

Here are the additional assumptions we will impose to establish joint asymptotic normality of $(\widehat{\theta}, \widehat{G})$.

L4. $\mathcal{X}$ is a bounded convex subset of R^d with nonempty interior and $\mathcal{H}$ is a universal Donsker class of real-valued measurable functions defined on $(\mathcal{X}, \mathcal{B})$.

Let $\underline{h}^* \in (L_2^0(G_0))^m$ given by

$$\underline{h}^*(x) \equiv (A_0^T A_0)^{-1}(A_0^T(\dot{\mathbf{l}}_\theta))(x) = \frac{\dot{\mathbf{Q}}^*}{\pi^*}(x)\underline{p} + \ (\dot{\mathbf{Q}} - \mathbf{C})\mathbf{M}^{-1}\frac{\underline{Q}^*}{\pi^*}(x) \tag{5.10}$$

denote the least favorable direction. Then we have:

Theorem 5.1. (Joint asymptotic normality and efficiency of the MLE, i.i.d. sampling). Suppose that conditions **L0** - **L4** hold. Then it follows that

$$\sqrt{n}\begin{pmatrix} \widehat{\theta}_n - \theta_0 \\ \widehat{G}_n - G_0 \end{pmatrix} \Rightarrow -\dot{\Psi}_0^{-1}(\mathbb{Z}) \equiv \mathbb{W} \equiv \begin{pmatrix} \mathbb{W}_1 \\ \mathbb{W}_2 \end{pmatrix} \qquad \text{in} \qquad R^d \times l^\infty(\mathcal{H})$$

where $\mathbb{W}_1 \sim N_d(0, I(\theta_0)^{-1})$ and $\mathbb{W}_2$ is a mean zero Gaussian process indexed by $\mathcal{H}$ with

$$\begin{aligned} Cov(\mathbb{W}_2(h_1), \mathbb{W}_2(h_2)) \ &= \ \langle h_1 - G(h_1), (A_0^T A_0)^{-1}(h_2 - G(h_2))\rangle \\ &\quad + \langle h_1, \underline{h}^*\rangle^T I(\theta_0)^{-1}\langle h_2, \underline{h}^*\rangle\,, \qquad h_1, h_2 \in \mathcal{H}, \end{aligned}$$

where the inner products are in $L_2(G_0)$. Moreover,

$$Cov(\mathbb{W}_1, \mathbb{W}_2(h)) = -\ I(\theta_0)^{-1}\langle \underline{h}^*, h\rangle, \qquad h \in \mathcal{H}\,.$$

Further, $(\widehat{\theta}_n, \widehat{G}_n)$ is asymptotically efficient; in particular $\widehat{\theta}_n$ has influence function $\widetilde{\mathbf{l}}_\theta$ given by (2.12).

Proof. Replacing $\tilde{\theta}_n$ by $\widehat{\theta}_n$ in (4.20) and (4.21) yields

$$\begin{aligned} \sqrt{n}(\widehat{G}_n(h) - G_0(h)) \ &= \ \sqrt{n}\left(\mathbb{P}_n\left(\frac{R}{\widehat{s}_n(\cdot, \widehat{\theta}_n)}h\right) - P_0\left(\frac{R}{s_0}h\right)\right) \\ &= \ \mathbb{G}_n\left(\frac{R}{s_0}h\right) - \mathbb{P}_n\left(Rh\frac{\sqrt{n}(\widehat{s}_n(\cdot, \widehat{\theta}_n) - s_0)}{\widehat{s}_n(\cdot, \widehat{\theta}_n)s_0}\right) \\ &\equiv \ I_n(h) - II_n(h)\,, \end{aligned} \tag{5.11}$$

where the term $II_n(h)$ can be written as

$$\begin{aligned} II_n(h) &= -\sum_{j=1}^{J}\sqrt{n}\left(\frac{N_j-n_j}{n}-Q_j^0(1-p_j)\right)\frac{1}{\widehat{Q}_j(\widehat{\theta}_n)}\mathbb{P}_n\left(\frac{Rh}{s_0\widehat{s}_n(\cdot,\widehat{\theta}_n)}Q_j^*(\cdot,\widehat{\theta}_n)\right) \\ &+ \sum_{j=1}^{J}\sqrt{n}\left\{\frac{\widehat{Q}_j(\widehat{\theta}_n)-Q_j(\widehat{\theta}_n)}{\widehat{Q}_j(\widehat{\theta}_n)Q_j^0}\right\}Q_j^0(1-p_j)\mathbb{P}_n\left(\frac{Rh}{s_0\widehat{s}_n(\cdot,\widehat{\theta}_n)}Q_j^*(\cdot,\widehat{\theta}_n)\right) \\ &+ \sum_{j=1}^{J}\sqrt{n}\left\{\frac{Q_j(\widehat{\theta}_n)-Q_j^0}{\widehat{Q}_j(\widehat{\theta}_n)Q_j^0}\right\}Q_j^0(1-p_j)\mathbb{P}_n\left(\frac{Rh}{s_0\widehat{s}_n(\cdot,\widehat{\theta}_n)}Q_j^*(\cdot,\widehat{\theta}_n)\right) \\ &+ \mathbb{P}_n\left(\frac{Rh}{s_0\widehat{s}_n(\cdot,\widehat{\theta}_n)}\right)\sum_{j=1}^{J}p_j\dot{Q}_j^*(\cdot,\theta_n^\sharp)\sqrt{n}(\widehat{\theta}_n-\theta_0)\,. \end{aligned} \tag{5.12}$$

Upon use of the differentiability arguments in the Appendix, careful grouping of terms, and using Proposition 4.1, (4.5), (5.11), and (4.21), we find that

$$\sqrt{n}(\widehat{G}_n(h)-G_0(h)) = \mathbb{G}_n(A_0(A_0^TA_0)^{-1}h)-\mathbb{G}_n(\widetilde{\mathbf{l}}_\theta^T)\langle\underline{h}^*,h\rangle+R_n(h) \tag{5.13}$$

where $\widetilde{\mathbf{l}}_\theta = I(\theta_0)^{-1}\mathbf{l}_\theta^*$ is given by (2.12) and $\|R_n\|_{\mathcal{H}} = o_p(1)$. Theorem 5.1 follows immediately from (4.5), (5.13), and standard arguments. □

6 Discussion: Other Designs and Further Problems

The proof of Theorem 5.1 given in section 5 differs from our first attempts which are given in Breslow, McNeney, and Wellner (2000). Our present hypotheses **L0** - **L4** are apparently weaker (and easier to verify) than the hypotheses imposed there in **D0** - **D5** and especially **D5**. Another advantage is that the present Theorem 5.1 allows for many more classes of functions $\mathcal{H}$. We do not yet know how to use the Z-theorem approach of Breslow, McNeney, and Wellner (2000) to prove Theorem 5.1 under **L0** - **L4**.

A. Other designs: In this paper we have treated the variable probability sampling (VPS1) or Bernoulli (i.i.d.) version of the two-phase designs. We expect similar results to hold when the sampling is carried out without replacement within strata. Indeed, McNeney (1998) shows that the information bounds calculated here carry over to this. Proofs of asymptotic efficiency remain to be established for these versions of the designs.

B. Choice of p's to maximize information: It would be of interest to study the optimal choice of p_j's to minimize the asymptotic variance of some particular function of θ. It is intuitively clear that an efficient choice of the p_j's will often entail choosing $p_j = 1$ for the strata with small (rare events) $Q_j(\theta, G)$.

C. Other models: If the basic model $f(y|x;\theta)$ is not just parametric, but semiparametric as in the case-cohort sampling designs studied by Self and Prentice (1988), then the methods of

the present paper do not apply. Although some work on information bounds has been carried out by Robins, Rotnitzky, and Zhao (1994) and RHN, we do not know of any easily implementable efficient estimators in these models.

Although the LKW approach accomodates continuous outcomes, its key feature is that the phase one data, those available for all subjects, are reduced to counts of subjects in a finite number of strata. This implies a loss of information if in fact continuous outcome data are available. Chatterjee, Chen, and Breslow (2002) developed a semiparametric "pseudo-score" estimator that only requires discretization of the phase one covariates. They demonstrated in simulations that its efficiency was sometimes substantially superior to that of the LKW profile likelihood estimator, even when 6 categories were used for discretization of the continuous outcomes. The information loss for the pseudo-score estimator in comparison with the semiparametric information bound for the general RHN problem has not yet been investigated.

D. Asymptotic distribution of the estimators off the model: It would be of interest to apply the $Z-$ theorem when the parametric model does not hold to better understand what parameters are being estimated, and how we should estimate variances robustly.

E. Validity of the bootstrap: If asymptotic normality of the estimators could be proved via the Z-theorem, then it would be straightforward to verify that the nonparametric bootstrap (and many other weighted bootstraps) is asymptotically valid via the results of Wellner and Zhan (1997).

7 Appendix. Differentiability Arguments.

At several points we need to understand how $Q_j^\dagger(t,\theta,G)$ changes with t and θ. Recall that $Q_j^\dagger(t,\theta,G) = Q_j(t, G_t(\cdot;\theta,G))$, $j = 1,\ldots,J$, satisfy the following system of equations:

$$Q_j^\dagger(t,\theta,G) = \int \frac{s(x,\theta)}{\sum_{l=1}^J \left(1 - \frac{Q_l(\theta,G)(1-p_l)}{Q_l^\dagger(t,\theta,G)}\right) Q_l^*(x,t)} Q_j^*(x,t) dG(x)\,, \qquad j = 1,\ldots,J\,. \tag{7.1}$$

Differentiation across (7.1) with respect to θ yields

$$\begin{aligned}
&\nabla_\theta Q_j^\dagger(t,\theta,G) \\
&= \int \frac{\sum_{l=1}^J p_l \dot{Q}_l^*(x,\theta)}{s_t(x;\theta,G)} Q_j^*(x,t) dG(x) \\
&\quad + \sum_{l=1}^J \frac{\nabla_\theta Q_l(\theta,G)(1-p_l)}{Q_l^\dagger(t,\theta,G)} \int \frac{s(x,\theta)}{[s_t(x;\theta,G)]^2} Q_l^*(x,t)\, Q_j^*(x,t) dG(x) \\
&\quad - \sum_{l=1}^J \frac{Q_l(\theta,G)(1-p_l)}{[Q_l^\dagger(t,\theta,G)]^2} \int \frac{s(x,\theta)}{[s_t(x;\theta,G)]^2} Q_l^*(x,t) Q_j^*(x,t) dG(x) \cdot \nabla_\theta Q_l^\dagger(t,\theta,G)\,.
\end{aligned}$$

Putting this in matrix form, we see that

$$\begin{aligned}
&\left(I + \text{diag}\left(\frac{Q_l(1-p_l)}{Q_l^\dagger(t,\theta,G)}\right)\int \frac{s(x,\theta)}{[s_t(x,\theta,G)]^2}\underline{Q}^*(x,t)\underline{Q}^*(x,t)^T\,dG(x)\right)\nabla_\theta \underline{Q}^\dagger(t,\theta,G) \\
&= \int \frac{\sum_{l=1}^J p_l \dot{Q}_l^*(x,\theta)}{s_t(x,\theta,G)}\underline{Q}^*(x,t)dG(x) \\
&\quad + \dot{\mathbf{Q}}\text{diag}\left(\frac{(1-p_l)}{Q_l^\dagger(t,\theta,G)}\right)\int \frac{s(x,\theta)}{[s_t(x,\theta,G)]^2}\underline{Q}^*(x,t)\underline{Q}^*(x,t)^T\,dG(x)\,. \qquad (7.2)
\end{aligned}$$

Similarly, differentiating across (7.1) with respect to t yields

$$\begin{aligned}
&\nabla_t Q_j^\dagger(t,\theta,G) \\
&= \int \frac{s(x,\theta)}{s_t(x;\theta,G)}\dot{Q}_j^*(x,t)dG(x) \\
&\quad - \int \frac{s(x,\theta)}{[s_t(x;\theta,G)]^2}\sum_{l=1}^J\left(1 - \frac{Q_l(\theta,G)(1-p_l)}{Q_l^\dagger(t,\theta,G)}\right)\dot{Q}_l^*(x,t)\,Q_j^*(x,t)dG(x) \\
&\quad - \int \frac{s(x,\theta)}{[s_t(x;\theta,G)]^2}\sum_{l=1}^J\frac{Q_l(\theta,G)(1-p_l)}{[Q_l^\dagger(t,\theta,G)]^2}Q_l^*(x,t)\nabla Q_l(t,\theta,G)Q_j^*(x,t)dG(x)\,,
\end{aligned}$$

or, in matrix form

$$\begin{aligned}
&\left(I + \int \frac{s(x,\theta)}{[s_t(x;\theta,G)]^2}\underline{Q}^*(x,t)\text{diag}\left(\frac{Q_l(\theta,G)(1-p_l)}{[Q_l^\dagger(t,\theta,G)]^2}\right)\underline{Q}^{*T}(x,t)dG(x)\right)\nabla_t \underline{Q}^\dagger(t,\theta,G) \\
&= \int \frac{s(x,\theta)}{s_t(x;\theta,G)}\dot{\mathbf{Q}}^*(x,t)dG(x) \qquad (7.3) \\
&\quad - \int \frac{s(x,\theta)}{[s_t(x;\theta,G)]^2}\sum_{l=1}^J\left(1 - \frac{Q_l(\theta,G)(1-p_l)}{Q_l^\dagger(t,\theta,G)}\right)\dot{Q}_l^*(x,t)\,Q_j^*(x,t)dG(x)\,.
\end{aligned}$$

ACKNOWLEDGEMENTS: We owe thanks to Nilanjan Chatterjee, Mary Emond, and the other participants in the *Missing Data Working Group* at the University of Washington during Winter Quarter, 1998, for many useful discussions about missing data and the subject of this paper.

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 University Press, Baltimore.

Breslow, N. E., and Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcomes applied to Wilms tumour prognosis. *Applied Statistics* **48**, 457 - 468.

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

Breslow, N. E. , McNeney, B., and Wellner, J. A. (2000). Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. *Technical Report* ***381***, Department of Statistics, University of Washington.

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.

Chatterjee, N.,Chen, Y.H., and Breslow, N.E. (2002). A pseudo-score estimator for regression problems with two-phase sampling. *J. Amer. Statist. Assoc.*, under revision.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Symp. Math. Statist. Prob.* **1**, 221 - 233. Univ. California Press.

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

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Second Edition. Chapman and Hall, London.

McNeney, B. (1998). Asymptotic Efficiency in Semiparametric Models with non-i.i.d. Data. Ph.D. dissertation, University of Washington.

Murphy, S. and Van der Vaart, A. W. (1997). Semiparametric likelihood ratio inference. *Ann. Statist.* **25**, 1471 - 1509.

Murphy, S. and Van der Vaart, A. W. (1999). Observed information in semi-parametric models. *Bernoulli* **5**, 381–412.

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

Nan, B., Emond, M., and Wellner, J. A. (2000). Information bounds for regression models with missing data. *Technical Report* **378**, Department of Statistics, University of Washington.

Pollard, D. (1985). New ways to prove central limit theorems. *Econometric Theory* **1**, 295 - 314.

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. (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. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika* **84**, 57 - 71.

Scott, A. J. and Wild, C. J. (2000). Maximum likelihood for generalised case-control studies. *J. Plan. Statist. Inf.*, **96**, 3 - 27.

Self, S.G. and Prentice, R.L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *Ann. Statist.* **16**, 64-81.

Van der Vaart, A. W. (1995). Efficiency of infinite-dimensional M-estimators. *Statistica Neerl.* **49**, 9 - 30.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Van der Vaart, A. W. and Wellner, J. A. (1992). Existence and consistency of maximum likelihood in upgraded mixture models. *J. Mult. Anal.* **43**, 133 - 146.

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

Van der Vaart, A. W. and Wellner, J. A. (2000). Preservation theorems for Glivenko-Cantelli and uniform Glivenko-Cantelli classes. In *High Dimensional Probability II* (E. Giné, D. M. Mason, and J. A. Wellner, eds.), Birkäuser, Boston, pp. 113-132.

Van der Vaart, A. W. and Wellner, J. A. (2001). Consistency of semiparametric maximum likelihood estimators for two-phase sampling. *Canad. J. Statist.* **29**, 269 - 288.

Wellner, J. A. and Zhan, Y. (1997). Bootstrapping $Z-$estimators. *Technical Report* **308**, Department of Statistics, University of Washington, Seattle.

University of Washington
Biostatistics
Box 357232
Seattle, Washington 98195-7232
U.S.A.
e-mail: norm@u.washington.edu

Department of Statistics and Actuarial Science
Simon Fraser University
8888 University Drive
Burnaby, British Columbia
Canada V5A 1S6
e-mail: mcneney@stat.sfu.ca

University of Washington
Statistics
Box 354322
Seattle, Washington 98195-4322
U.S.A.
e-mail: jaw@stat.washington.edu