On the Breslow-Holubkov estimator

by

A.J. Lee[1], A.J. Scott[2] and C.J. Wild[3]

Abstract

Breslow and Holubkov (1997) developed semiparametric maximum likelihood estimation for two-phase studies with a case-control first phase under a logistic regression model and noted that, apart for the overall intercept term, it was the same as the semiparametric estimator for two-phase studies with a prospective first phase developed in Scott and Wild (1997) . In this paper we extend the Breslow-Holubkov result to general binary regression models and show that it has a very simple relationship with its prospective first-phase counterpart. We also explore why the design of the first phase only affects the intercept of a logistic model, simplify the calculation of standard errors, establish the semiparametric efficiency of the Breslow-Holubkov estimator and derive its asymptotic distribution in the general case.

Keywords Binary regression; Case-control sampling; Logistic regression; Two-phase sampling; Response-selective sampling; Estimating equations; Semi-parametric efficiency.

1 Introduction

As Breslow and Holubkov (1997) noted, outcome dependent sampling can increase the efficiency of studies with rare outcomes substantially. The effect is at its simplest and starkest with the case-control study investigating risk factors for a binary response variable. The ubiquity of these designs in epidemiology is such that Breslow and Day (1980) called the case-control study "the backbone of epidemiology". They are also used in other many fields, often by other names such as choice-based sampling designs in econometrics. Although case-control studies and related designs are just one small part of Norm Breslow's wide statistical interests, being a central concern of less that 16% of his publications, we believe that they are a very important part of his work. The fact that he chose case-control studies as the topic of his seminal 1995 Fisher Lecture suggests that they are important to him too. The Fisher lecture (Breslow, 1996) dealt with the history and development of case-control studies, together with generalizations and associated methods of analysis. In this paper we will concentrate on just one generalization, the two-phase (or two-stage) case-control design. Norm Breslow's work on two-phase designs dates back to before 1988 when he published three papers on the topic (Breslow and Cain, 1988; Cain

[1]Department of Statistics, University of Auckland, Private Bag 92019, Auckland, NZ
e-mail: a.lee@auckland.ac.nz
[2]Department of Statistics, University of Auckland
[3]Department of Statistics, University of Auckland

and Breslow, 1988; Breslow and Zhao, 1988) and has continued with a sequence of papers right up to the present day.

Let Y denote a binary response variable which can take value $Y = 1$ (corresponding to a case) or $Y = 0$ (corresponding to a control) and let $\boldsymbol{x}$ be a p-dimensional vector of explanatory variables or covariates. Our purpose is to fit a general parametric regression model, $P(Y = 1 \mid \boldsymbol{x}) = p_1(\boldsymbol{x}; \boldsymbol{\beta})$ say. The usual model of choice in applications is the logistic regression model, $p_1(\boldsymbol{x}; \boldsymbol{\beta}) = \exp(\boldsymbol{x}^T\boldsymbol{\beta}) / \left\{1 + \exp(\boldsymbol{x}^T\boldsymbol{\beta})\right\}$. It is sometimes convenient to write $\boldsymbol{x}^T\boldsymbol{\beta} = \beta_0 + \boldsymbol{x}^{*T}\boldsymbol{\beta}_1$ to emphasize the role of an overall intercept β_0.

To put the current work in context, we start with the simple (unmatched, single-phase) case control design. In a simple case-control study we take a sample of size n_1 cases (often all available cases) and a sample of size n_0 from the available controls. When fitting a logistic regression model including an intercept term it is well known, following the landmark papers of Anderson (1972) for the discrete-$\boldsymbol{x}$ case, and Prentice and Pyke (1979) for general $\boldsymbol{x}$, that the semiparametric maximum likelihood estimate of $\boldsymbol{\beta}_1$ and its asymptotic covariance matrix can be obtained by fitting a logistic regression model using standard software as if it had been obtained prospectively. The intercept β_0 is completely confounded with the relative sampling rates of cases and controls but can be recovered using additional information such as finite population totals of cases and controls (see Scott and Wild, 1986). The semiparametric efficiency of the standard logistic analysis was established, and the underlying asymptotic theory made rigorous, by Breslow, Robbins and Wellner (2000) – albeit for random rather than fixed n_1 and n_0 thus enabling i.i.d. theory to be used. McNeney (1998) demonstrated that the efficiency properties extended to the fixed n_1, n_0 case.

The two-phase (two-stage) case-control design was introduced by White (1982) as a design for studying an association between a binary response Y and a binary exposure variable V (our notation) adjusted for discrete covariates. Motivated by considerations of cost-effectiveness, she proposed taking separate samples at phase two from the individuals in each of the 4 cells of the 2×2 cross-classification of Y and V, and determining covariate information only for the subsampled individuals. She proposed over-sampling small cells, e.g. by taking equal sized subsamples from each of the four cells. She noted that the first-phase $Y \times V$-data could itself come either from case-control sampling, or be from a cohort or cross-sectional study. We combine the latter two situations and refer to them as being "prospective". The distinction between a prospective or case-control first phase in a two-phase study is pivotal to this paper and mirrors the distinction between a simple (single-phase) case-control or prospective study.

By the end of the 1980s, following the work of Fears and Brown (1986), Breslow and Cain (1988), Cain and Breslow (1988) and Breslow and Zhao (1988), theory was available to handle cases when $\boldsymbol{x}$ included continuous covariates and V took J values and was included as a linear term, as opposed to a set of categories, in the regression model. Indeed, provided all the constituent variables were discrete, we could have a vector $\boldsymbol{V}$ of variables defining the V-strata. These generalisations allowed the following uses of the resulting methodology, all recognised by Cain and Breslow (1988).

"Cost" savings can be obtained by using a genuine two-phase design (e.g. Engels et al. 2005) and only measuring covariates that are particularly expensive or particularly invasive on comparatively small subsamples. Such studies are becoming increasingly useful, particularly as expensive new techniques for extracting genetic information become more and more widely available.
Secondary analysis: Second-phase sampling provides a cost-effective way of making an after-the-fact adjustment for a confounder that was not considered in the original single-phase study.
Incorporating "whole population" information: There may be administrative or other population $Y \times V$-data available for all individuals in the finite population(s) from which the cases and controls in single-phase study were drawn. Efficiency can be increased by considering the finite-population data as the first-phase and the study data as the second phase.
Missing data: If, in a single-stage study, there are substantial numbers of missing values in the covariates and we are willing to assume that they are missing at random given Y and the V-variables, then we can treat the data as coming from a two-phase study with those for which $\boldsymbol{x}$ is observed forming the $Y \times V$-subsamples. This is more defensible than a complete-case analysis, especially when the missingness rates differ appreciably between $Y \times V$ cells.

By making proper use of stratum-specific offsets, prospective logistic-regression programs can be used to obtain valid estimates of the parameters of a logistic regression (Fears and Brown, 1986) fitted to data from a two-phase study in the full generality described above. Substantial work is needed to correct the standard errors, however, and the procedure is not in general either maximum likelihood (Breslow and Cain, 1988; Breslow and Zhao, 1988) or efficient (Scott and Wild, 1991). Semiparametric maximum likelihood estimation for two-phase studies with a prospective first phase was developed for general models by Scott and Wild (1991, 1997, 2001), whereas Breslow and Holubkov (1997) worked with logistic models and developed semiparametric maximum likelihood for a case-control first phase. They made the interesting observation that the resulting estimator was the same as the Scott and Wild (1997) estimator. Just as for simple studies, for logistic models fitted to two-phase data, whether the first phase is prospective or case-control only affects the overall intercept β_0. Semiparametric efficiency was established by Breslow, McNeney and Wellner (2003) for a prospective first phase and random sample-size sub-sampling mechanism.

In this paper we explore why the design of the first phase only affects the intercept of a logistic model, develop the Breslow-Holubkov estimator for general binary regression models, and present a very simple relationship to its prospective first-phase counterpart. We also simplify the calculation of standard errors, establish the semiparametric efficiency of the Breslow-Holubkov estimator, and derive its asymptotic distribution.

2 General Framework and Results

2.1 Framework

We wish to fit an arbitrary binary regression model

$$P(Y = 1 \mid \boldsymbol{x}) = p_1(\boldsymbol{x}; \boldsymbol{\beta}), \tag{1}$$

with $p_0(\boldsymbol{x}; \boldsymbol{\beta}) = 1 - p_1(\boldsymbol{x}; \boldsymbol{\beta})$. Logistic regression is the special case in which

$$p_1(\boldsymbol{x}; \boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}^T \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}^T \boldsymbol{\beta})}. \tag{2}$$

Sometimes we write $\boldsymbol{x}^T \boldsymbol{\beta} = \beta_0 + \boldsymbol{x}^{*T} \boldsymbol{\beta}_1$ to emphasize the role of an overall intercept β_0.

The data are obtained as follows. At Phase 1, we draw independent samples of N_1 cases and N_0 controls and obtain values of $\boldsymbol{V}$ for all individuals in the sample. Some or all of the components of $\boldsymbol{V}$ may be included in the covariate vector $\boldsymbol{X}$. We shall assume that $\boldsymbol{V}$ has finite support with possible values $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_J$. Let N_{ij} denote the number of sampled individuals with $Y = i$ and $\boldsymbol{V} = \boldsymbol{v}_j$. At Phase 2, we draw a simple random sample of size n_{ij} from these N_{ij} individuals, and observe the values of the remaining covariates, say $\boldsymbol{W}$, which can be discrete or continuous. This results in samples $\{\boldsymbol{x}_{ij1}, \ldots, \boldsymbol{x}_{ijn_{ij}}\}$ for $i = 0, 1$ and $j = 1, \ldots, J$.

Note that the n_{ij}s are random variables since we must have $n_{ij} \leq N_{ij}$ for $i = 1, \ldots, I$ and $j = 1, \ldots, J$. We assume that the distribution of $\{n_{ij}\}$ depends only on $\{N_{ij}\}$. Then the resulting likelihood has the form

$$L = \prod_{i=0}^{1} \left[\prod_{j=1}^{J} \left\{ P(\boldsymbol{V} = \boldsymbol{v}_j \mid Y = i)^{N_{ij}} \prod_{k=1}^{n_{ij}} f(\boldsymbol{x}_{ijk} \mid Y = i, \boldsymbol{V} = \boldsymbol{v}_j) \right\} \right] \tag{3}$$

(see Wild 1991 for more details).

2.2 The Breslow-Holubkov estimator

In this section we develop the Breslow-Holubkov estimator for general binary regression models. Since the full likelihood depends on the distribution of the covariates and there is no interest in modelling this distribution for its own sake, we follow Breslow and Holubkov (1997) in adopting a semi-parametric approach in which the covariate distribution is left completely unspecified. Using Bayes Theorem and the model specified in (1), we can write the log-likelihood from (3) in the form

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\rho}) = \sum_i \sum_j \sum_k \log p_i(\boldsymbol{x}_{ijk}; \boldsymbol{\beta}) + \sum_i \sum_j (N_{ij} - n_{ij}) \log Q_{ij} - \sum_i N_i \log q_i \\ + \sum_j N_{+j} \log \rho_j + \sum_i \sum_j \sum_k \log \gamma_j(\boldsymbol{w}_{ijk}), \end{aligned} \tag{4}$$

where

$$\begin{aligned} Q_{ij} &= Q_{ij}(\boldsymbol{\beta}, \Gamma_j) = P(Y = i \mid \boldsymbol{V} = \boldsymbol{v}_j) = \int p_i(\boldsymbol{v}_j, \boldsymbol{w}; \boldsymbol{\beta}) d\Gamma_j(\boldsymbol{w}), \\ q_i &= q_i(\boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\rho}) = P(Y = i) = \sum_j Q_{ij}(\boldsymbol{\beta}, \Gamma_j)\rho_j, \\ \rho_j &= P(\boldsymbol{V} = \boldsymbol{v}_j) \end{aligned}$$

and $\Gamma_j(\boldsymbol{w})$, $\gamma_j(\boldsymbol{w})$ are the conditional distribution and density functions of $\boldsymbol{W}$ given $\boldsymbol{V} = \boldsymbol{v}_j$.

We want to find to estimate $\boldsymbol{\beta}$ without having to think about the unknown distribution of the covariates at all. Ideally, we would like to maximize the log-likelihood (4) over $(\boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\rho})$ without making any assumptions about $\boldsymbol{\Gamma}$ or $\boldsymbol{\rho}$. Because each Γ_j is a distribution that may be continuous, and thus potentially infinite dimensional, this maximization looks difficult. However, it turns out that we can obtain an efficient semi-parametric estimator, $\widehat{\boldsymbol{\beta}}$, relatively simply by adopting an indirect approach and working with a much simpler "loglikelihood" $\ell^*(\ )$, which involves only $(J+1)$ nuisance parameters, as if it were an ordinary loglikelihood. The parameters of ℓ^* are $\boldsymbol{\phi} = (\boldsymbol{\beta}, \boldsymbol{\pi}, \alpha)$ where the nuisance parameters are $\boldsymbol{\pi}$ which has dimension J and α which is 1-dimensional. Although these parameters are to be treated simply as formal parameters, they do have meaningful interpretations which arise in the derivation of ℓ^* and are given below. ℓ^* is defined as follows.

$$\ell^*(\boldsymbol{\phi}) = \sum_i \sum_j \sum_k \log\{p^*_{ij}(\boldsymbol{x}_{ijk}; \boldsymbol{\phi})\} + \sum_i \sum_j N_{ij} \log \pi_{ij} - \sum_i \sum_j n_{ij} log\left(N_{+j}\pi_{ij} - \widetilde{N}_{ij}\right), \quad (5)$$

where $\pi_{1j} = \pi_j$, $\pi_{0j} = 1 - \pi_j$, $\widetilde{N}_{ij} = N_{ij} - n_{ij}$ and $p^*_{ij}(x; \boldsymbol{\phi})$ is defined by setting

$$\text{logit} p^*_{1j}(\boldsymbol{x}; \boldsymbol{\phi}) = \text{logit} p_1(\boldsymbol{x}; \boldsymbol{\beta}) + \alpha + \sigma_j(\pi_j), \quad (6)$$

with $p^*_{0j}(\boldsymbol{x}; \boldsymbol{\phi}) = 1 - p^*_{1j}(\boldsymbol{x}; \boldsymbol{\phi})$ and

$$\sigma_j(\pi_j) = \log\left(N_{+j} - \frac{\widetilde{N}_{1j}}{\pi_j}\right) - \log\left(N_{+j} - \frac{\widetilde{N}_{0j}}{1 - \pi_j}\right).$$

As previously stated, efficient semiparametric inferences about $\boldsymbol{\beta}$ can be obtained by acting as if the pseudo-likelihood $\ell^*(\boldsymbol{\phi})$ is the true likelihood. This means that we can obtain $\widehat{\boldsymbol{\beta}}$ by solving the pseudo-likelihood equations,

$$U^*(\boldsymbol{\phi}) = \frac{\partial \ell^*}{\partial \boldsymbol{\phi}} = \mathbf{0},$$

and we can estimate $\text{Cov}\{\widehat{\boldsymbol{\beta}}\}$ with the appropriate submatrix of $\mathbf{J}^{*-1}$, where $\mathbf{J}^*$ is the observed pseudo-information matrix,

$$\mathbf{J}^* = -\frac{\partial^2 \ell^*}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T} = -\frac{\partial U^*}{\partial \boldsymbol{\phi}^T}.$$

(We note that $\mathbf{J}^*$ is a direct byproduct of a Newton-Raphson maximization of $\ell^*(\boldsymbol{\phi})$.) Further, we can treat the appropriate differences in $-2\ell^*$ as chi-squared random variables to test hypotheses about $\boldsymbol{\beta}$.

In Section 3, we consider the case when $\boldsymbol{W}$ has finite support and show that the procedure outlined above gives the semiparametric maximum likelihood estimator. We consider the general case in Section 4, and show that the procedure produces a semiparametric efficient estimator of $\boldsymbol{\beta}$ for any $\boldsymbol{W}$.

2.3 Interpretation of terms and nuisance parameters

From the derivation in Section 3, the nuisance parameters of ℓ^* in (5) are

$$\alpha = \log\left(\frac{N_1}{P(Y=1)}\Big/\frac{N_0}{P(Y=0)}\right), \pi_j = \frac{N_1 P(Y=1 \mid \mathbf{V}=\boldsymbol{v}_j)}{N_1 P(Y=1 \mid \mathbf{V}=\boldsymbol{v}_j) + N_0 P(Y=0 \mid \mathbf{V}=\boldsymbol{v}_j)}.$$

The quantities making up ℓ^* can be interpreted most easily in terms of a random equivalent of the two-phase model in which Y is chosen with $Y=i$ having probability N_i/N_+, then V-stratum is chosen with $\boldsymbol{v}_j$ having probability N_{ij}/N_i (Phase 1), after which n_{ij} values are sampled from $\text{pr}(\boldsymbol{x} \mid Y=i, \boldsymbol{V}=\boldsymbol{v}_j)$ (Phase 2). The nuisance parameter α is the log of the relative phase-one sampling rates for cases and controls and the nuisance parameter π_j represents the conditional probability that a unit is a case, given that $\mathbf{V}=\boldsymbol{v}_j$ and the unit is sampled. Additionally, $p^*_{1j}(\boldsymbol{x};\boldsymbol{\phi})$ is essentially $\text{pr}(\text{Y} = 1 \mid \boldsymbol{x}, \boldsymbol{v}_\text{j}, \text{sampled})$. This follows from several applications of Bayes Theorem and noting that if we replace π_{ij} by N_{ij}/N_{+j}, then $\sigma_j(\pi_j)$ reduces to $\log\{(n_{1j}/N_{1j})/(n_{0j}/N_{0j})\}$.

We conclude this subsection with some comments about notation. Our parameter π_{ij} corresponds to P_{ij} in the notation of Breslow & Holubkov and our α, the log of the relative phase-one sampling rates for cases and controls, to Breslow & Holubkov's $\log\left(\frac{N_1}{N_0}\right) - \alpha$.

2.4 Relationship to the case of a prospective first-phase

Scott and Wild (1997) worked through the maximisation above with the likelihood appropriate for a prospective first stage (page 65 of that paper). The only difference between (5) and (6) and their prospective equivalents is that in the prospective case α does not appear in (6), i.e., $\alpha = 0$. Setting the log of the relative sampling rates for cases and controls, α, to zero makes intuitive sense because cases and controls are sampled at the same rate with a prospective first phase.

The correspondence between the inferences for the two schemes is even closer in the special case where $p_1(\boldsymbol{x};\boldsymbol{\beta})$ takes the logistic form (2). In this case, the pseudo-model p^*_{ij} in (6) is also logistic with the same slope coefficients $\boldsymbol{\beta}_1$ as the original model but intercept $\beta_0 + \alpha + \sigma_j = \beta^*_0 + \sigma_j$ say. Clearly, we can only estimate the sum $\beta^*_0 = \beta_0 + \alpha$ and not the individual components, β_0 and α, with case-control sampling unless we have further information about sampling rates. Then, if we rewrite (5) and (6) in terms of β^*_0 rather than β_0 and α, the pseudo-likelihoods for the two schemes become identical.

The only difference is in the interpretation of β_0^*. With prospective sampling at Phase 1, $\beta_0^* = \beta_0$. With case-control sampling at Phase 1, $\beta_0^* = \beta_0 + \alpha$, with β_0 and α individually indeterminate. As Breslow & Holubkov point out, this provides a direct analogue to the usual result for ordinary (single phase) case-control sampling: if we have a logistic regression model with an intercept term, then we obtain valid (and efficient) inferences for all coefficients except the intercept by proceeding as if we had a prospective sample.

Note that setting $\frac{\partial \ell^*}{\partial \pi_j} = 0$ in (5) is equivalent to setting

$$N_{+j}\pi_j = N_{1j} - n_{1j} + \sum_k n_{+jk} p_{1j}^*(\boldsymbol{x}_{1jk}). \tag{7}$$

Scott & Wild (1997) suggested an iterative procedure in which ℓ^* is maximized with $\boldsymbol{\pi}$ held fixed and then $\boldsymbol{\pi}$ is updated using (7) . With a logistic model, the maximization step can be carried out simply by including a fixed offset in a standard logistic regression program and updating the offset at each iteration. A similar procedure could be used here. However, the procedure has turned out to be rather slow to converge in practice, and applying something like the Newton-Raphson procedure directly to ℓ^* is usually much more efficient. Moreover, the variance estimates obtained from the final iteration of the logistic regression are not the required components of $\mathbf{J}^{*-1}$, whereas this is produced automatically by the Newton-Raphson procedure.

3 Derivation of the Breslow-Holubkov Estimator

We want to maximize $l(\boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\rho})$ given in (4) with respect to $\Gamma_k(\boldsymbol{w})$ and $\boldsymbol{\rho}$ to obtain the profile likelihood of $\boldsymbol{\beta}$. The derivation is very similar to that given in Scott & Wild (1997) for the situation where we have a prospective first stage. Recall that $\boldsymbol{W}$ consists of the elements of $\boldsymbol{X}$ that are not in $\boldsymbol{V}$. We establish the result under the assumption that $\boldsymbol{W}$ has finite support, taking values $\boldsymbol{w}_k$, for $k = 1, \ldots, K$ with $P(\boldsymbol{W} = \boldsymbol{w}_k \mid \boldsymbol{V} = \boldsymbol{v}_j) = \gamma_{jk}$, say. Let n_{ijk} be the number of times that $\boldsymbol{w}_k$ is observed in the ijth sample, and put $p_{ijk}(\boldsymbol{\beta}) = p_i(\boldsymbol{v}_j, \boldsymbol{w}_k, \boldsymbol{\beta})$. Then the log-likelihood in (4) becomes

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\gamma}, \rho) = &\sum_i \sum_j \sum_k n_{ijk} \log p_{ijk}(\boldsymbol{\beta}) + \sum_i \sum_j \widetilde{N}_{ij} \log Q_{ij} \\ &- \sum_i N_i \log q_i + \sum_j N_{+j} \log \rho_j + \sum_j \sum_k n_{+jk} \log \gamma_{jk}, \end{aligned} \tag{8}$$

where $Q_{ij} = Q_{ij}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_k p_{ijk}(\boldsymbol{\beta})\gamma_{jk}$ and $q_i = q_i(\boldsymbol{\beta}, \boldsymbol{\gamma}, \rho) = \sum_j Q_{ij}\rho_j$. To find the profile likelihood for $\boldsymbol{\beta}$, we need to replace $\boldsymbol{\gamma}$ and ρ by $\widehat{\boldsymbol{\gamma}}(\boldsymbol{\beta})$ and $\widehat{\rho}(\boldsymbol{\beta})$, the values obtained by maximizing the log likelihood over $\boldsymbol{\gamma}$ and ρ for fixed $\boldsymbol{\beta}$. We introduce Lagrange multipliers η_0 and $\{\eta_j; j = 1, \ldots, J\}$ to take care of the constraints $\sum \rho_j = 1$ and $\{\sum_k \gamma_{jk} = 1,\ j = 1, \ldots, J\}$. Differentiating (8) with respect to ρ_j and setting the result equal to η_0 leads to

$$\frac{\partial l}{\partial \rho_j} = \frac{N_{+j}}{\rho_j} - \sum_i N_i \frac{Q_{ij}}{q_i} = \eta_0 \,. \tag{9}$$

Similarly, differentiating (8) with respect to γ_{jk} and setting the result equal to η_j leads to

$$\frac{\partial l}{\partial \gamma_{jk}} = \frac{n_{+jk}}{\gamma_{jk}} + \sum_i \widetilde{N}_i \frac{p_{ijk}(\boldsymbol{\beta})}{Q_{ij}} - \rho_j \sum_i N_i \frac{p_{ijk}(\boldsymbol{\beta})}{q_i} = \eta_j \,. \tag{10}$$

Multiplying (9) through by ρ_j and summing over j gives $\eta_0 = 0$, and multiplying (10) through by γ_{jk} and summing over k then gives $\eta_j = \rho_j \eta_0 = 0$. Thus

$$\widehat{\rho}_j = \frac{N_{+j}}{\sum_i N_i \frac{Q_{ij}}{q_i}} \quad \text{and} \quad \widehat{\gamma}_{jk} = \frac{n_{+jk}}{\sum_i \mu_{ij} p_{ijk}(\boldsymbol{\beta})},$$

where

$$\mu_{ij} = \frac{\widehat{\rho}_j N_i}{q_i} - \frac{\widetilde{N}_{ij}}{Q_{ij}} = \left(N_{+j}\pi_{ij} - \widetilde{N}_{ij}\right)/Q_{ij},$$

with

$$\pi_{ij} = \frac{N_i Q_{ij}/q_i}{N_1 Q_{1j}/q_1 + N_0 Q_{0j}/q_0}.$$

Note that Q_{ij} and q_i must satisfy the conditions

$$Q_{ij} = \sum_k p_{ijk}(\boldsymbol{\beta})\widehat{\gamma}_{jk} = \frac{1}{\mu_{ij}} \sum_k n_{+jk} p^*_{ijk} \quad \text{and} \quad q_i = \sum_j Q_{ij}\widehat{\rho}_j,$$

where $p^*_{ijk} = \frac{\mu_{ij} p_{ijk}}{\sum_t \mu_{tj} p_{tjk}}$. We can rewrite these conditions in the form

$$\sum_k n_{+jk} p^*_{ijk} + \widetilde{N}_{ij} - N_{+j}\pi_{ij} = 0 \tag{11}$$

and

$$\sum_j N_{+j}\pi_{ij} = N_i. \tag{12}$$

Substituting the expressions for $\widehat{\rho}_j$ and $\widehat{\gamma}_{jk}$ into (8), we obtain

$$l^* = \sum_i \sum_j \sum_k n_{ijk} \log p^*_{ijk} + \sum_i \sum_j N_{ij} \log \pi_{ij} - \sum_i \sum_j n_{ij} \log\left(N_{+j}\pi_{ij} - \widetilde{N}_{ij}\right). \tag{13}$$

We note that $p^*_{ijk} = p^*_{ijk}(\boldsymbol{\beta}, \alpha, \boldsymbol{\pi})$ can be re-written in the form given in (6), i.e.

$$\text{logit } p^*_{1jk} = \text{logit } p_{1jk} + \alpha + \sigma_j(\pi_{1j}),$$

with

$$\alpha = \log\{(N_1 q_0)/(N_0 q_1)\} \quad \text{and} \quad \sigma_j(\pi_{1j}) = \log\left\{(N_{+j} - \widetilde{N}_{1j}/\pi_{1j})/(N_{+j} - \widetilde{N}_{0j}/\pi_{0j})\right\}.$$

Thus the profile loglikelihood function, $l_P(\boldsymbol{\beta}) = \sup_{\boldsymbol{\gamma},\boldsymbol{\rho}} l(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\rho})$, can be expressed in the alternative form $l_P(\boldsymbol{\beta}) = l^*(\boldsymbol{\beta}, \boldsymbol{\pi}(\boldsymbol{\beta}), \alpha(\boldsymbol{\beta}))$, where $\boldsymbol{\pi}(\boldsymbol{\beta})$ and $\alpha(\boldsymbol{\beta})$ satisfy (11) and (12). Note that

$$\frac{\partial \ell^*}{\partial \pi_j} = A_j \left(N_{+j}\pi_{1j} - \widetilde{N}_{1j} - \sum_k n_{+jk} p^*_{1jk} \right), \tag{14}$$

where $A_j = \sum_i \frac{\widetilde{N}_{ij}}{\pi_{ij}\{N_{+j}\pi_{ij} - \widetilde{N}_{ij}\})}$, and

$$\frac{\partial \ell^*}{\partial \alpha} = n_1 - \sum_j \sum_k n_{+jk} p^*_{1jk}. \tag{15}$$

Setting these two expressions equal to 0, and recalling that $n_1 + \sum_j \widetilde{N}_{1j} = N_1$, leads to (11) and (12). Thus these conditions are equivalent to $\frac{\partial l^*}{\partial \boldsymbol{\pi}} = \mathbf{0}$ and $\frac{\partial l^*}{\partial \alpha} = 0$. (This also follows directly from the construction of $\boldsymbol{\pi}$ and α.)

This establishes the result that, when $\boldsymbol{W}$ has finite support, we can obtain the semiparametric MLE, $\widehat{\boldsymbol{\beta}}$, by treating $l^*(\boldsymbol{\beta}, \boldsymbol{\pi}, \alpha)$ as if it were a likelihood involving the $(p+J+1)$-dimensional parameter $\boldsymbol{\phi} = (\boldsymbol{\beta}, \boldsymbol{\pi}, \alpha)$ and setting $U^*(\boldsymbol{\phi}) = \frac{\partial l^*}{\partial \boldsymbol{\phi}} = 0$. In the next section, we show that the estimator $\widehat{\boldsymbol{\beta}}$ produced by this procedure has full semiparametric efficiency even when the distribution of $\boldsymbol{W}$ does not have finite support. (Whether or not $\widehat{\boldsymbol{\beta}}$ is actually the semiparametric MLE in this more general case is an open question - see Gill, Vardi & Wellner 1988 and van der Vaart & Wellner 2001 for a discussion of related problems). The consistency of the variance estimate based on $\mathbf{J}^* = -\frac{\partial^2 \ell^*}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T}$ is also established in the next section. The validity of testing hypotheses about components of $\boldsymbol{\beta}$ using appropriate differences in $-2\log \ell^*$ follows almost exactly as in Scott & Wild (1989).

4 Efficiency of the estimator

To demonstrate the semi-parametric efficiency of the methods described above, we will adopt the following strategy. We first derive an expression for the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$. Then, we compute the semi-parametric efficiency bound for our problem (i.e a lower bound on the covariance matrix of all estimates of $\boldsymbol{\beta}$). Finally, we show that the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$ coincides with this bound.

4.1 The asymptotic variance of the estimate

Our asymptotics are carried out under the assumption that the first phase sample sizes grow without bound at the same rate, and that the sampling fractions n_{ij}/N_i converge to numbers between zero and one i.e. we assume that $N_i/(N_0 + N_1) \to w_i$ and $n_{ij}/(N_0 + N_1) \to w_{ij}$.

First, we define

$$\mathbf{I}^* = -\text{plim}_{N_0,N_1\to\infty}(N_0+N_1)^{-1}\frac{\partial^2}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^T}l^*(\boldsymbol{\phi}),$$

the limit in probability of the pseudo-information matrix $\mathbf{J}^*$ introduced in Section 3. It turns out (see Scott & Wild, 2001, Lee, Scott & Wild, 2005) that the asymptotic variance matrix of $\hat{\boldsymbol{\beta}}$ is the inverse of the matrix

$$\mathbf{I}^*_{\beta\beta} - \mathbf{I}^*_{\beta\eta}(\mathbf{I}^*_{\eta\eta})^{-1}\mathbf{I}^*_{\eta\beta}, \tag{16}$$

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

$$\mathbf{I}^* = \begin{bmatrix} \mathbf{I}^*_{\beta\beta} & \mathbf{I}^*_{\beta\eta} \\ \mathbf{I}^*_{\eta\beta} & \mathbf{I}^*_{\eta\eta} \end{bmatrix}$$

and $\boldsymbol{\eta} = (\boldsymbol{\pi}^T, \alpha)^T$. This follows from the fact that, under suitable regularity conditions, the solution $\hat{\boldsymbol{\phi}}$ of

$$\frac{\partial l^*}{\partial\boldsymbol{\phi}} = 0$$

is asymptotically normal with asymptotic variance

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

where the matrix $\mathbf{V}$ is of the form

$$\mathbf{V} = \mathbf{I}^* - \mathbf{I}^*\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0^T} & \mathbf{A} \end{pmatrix}\mathbf{I}^* \tag{17}$$

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

$$\mathbf{I}^{*-1} - \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0^T} & \mathbf{A} \end{pmatrix},$$

and it follows from the partitioned matrix inverse formula that the asymptotic variance matrix of $\hat{\boldsymbol{\beta}}$ is given by (16).

We now derive an expression for $\mathbf{I}^*$ under a different but equivalent sampling scheme, which is convenient for demonstrating the efficiency of the estimator $\hat{\boldsymbol{\beta}}$ in Section 4. The new sampling scheme consists of

1. For $i = 0, 1$, we sample N_i individuals from the conditional distribution of $\boldsymbol{V}$, given $Y = i$. Let N_{ij} be the number of these having $\boldsymbol{V} = \boldsymbol{v}_j$. Then $(N_{i1}, \ldots, N_{iJ})$ have a multinomial distribution with probabilities Δ_{ij}. (Note that Δ_{ij} corresponds to the quantity denoted by Q^i_j in Breslow and Holubkov.)

2. For $i = 0, 1,\ j = 1, \ldots, J$, we sample n_{ij} individuals from the conditional distribution of $\boldsymbol{W}$, given $Y = i, \boldsymbol{V} = \boldsymbol{v}_j$, with density

$$p_i(\boldsymbol{v}_j, \boldsymbol{w}, \boldsymbol{\beta})\gamma_j(\boldsymbol{w})/Q_{ij},\ j = 1, \ldots, J,$$

where γ_j is the conditional density of $\boldsymbol{W}$ given $\boldsymbol{V} = \boldsymbol{v}_j$.

This scheme results in the same likelihood, and hence gives rise to the same asymptotics. The densities $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)$ constitute an infinite-dimensional nuisance parameter. As before, we assume that $N_i/(N_0 + N_1) \to w_i$ and $n_{ij}/(N_0 + N_1) \to w_{ij}$ and, in addition, that $w_i\Delta_{ij} > w_{ij}$, corresponding to the fact that in the Breslow sampling scheme, we must have $N_{ij} > n_{ij}$.

Assuming this sampling scheme, by the weak law of large numbers, we obtain

$$\begin{aligned}\mathbf{I}^* &= -\sum_i \sum_j w_{ij} E_{ij}\left[\frac{\partial^2 \log p^*_{ij}(\boldsymbol{W}, \boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^T}\right] - \sum_i \sum_j w_i \Delta_{ij0} \frac{\partial^2 \log \pi_{ij}}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^T} \\ &\quad + \sum_i \sum_j w_{ij} \frac{\partial^2 \log(w^{(j)}\pi_{ij} - w_i\Delta_{ij0} + w_{ij})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^T}. \end{aligned} \tag{18}$$

The symbol $w^{(j)}$ is defined by

$$w^{(j)} = \sum_i w_i \Delta_{ij0}$$

where we write Δ_{ij0} for the true values of the multinomial probabilities Δ_{ij}, and E_{ij} denotes expectation with respect to the true conditional distribution of $\boldsymbol{W}$, given $Y = i, \boldsymbol{V} = \boldsymbol{v}_j$. The function $p^*_{ij}(\boldsymbol{w}, \boldsymbol{\phi})$ now has a limiting form as $N_0,\ N_1 \to \infty$, defined by

$$\text{logit } p^*_{1j}(\boldsymbol{w}, \boldsymbol{\phi}) = \text{logit } p_i(v_j, \boldsymbol{w}, \boldsymbol{\beta}) + \alpha + \sigma_j(\pi_{1j}),$$

where now $\sigma_j(\pi_{1j}) = \log(w^{(j)} - (w_1\Delta_{1j0} - w_{1j})/\pi_{1j}) - \log(w^{(j)} - (w_0\Delta_{0j0} - w_{0j})/\pi_{0j})$. Using the identity

$$\frac{\partial^2 \log h}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^T} = \frac{1}{h}\frac{\partial^2 h}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^T} - \frac{\partial \log h}{\partial\boldsymbol{\phi}}\frac{\partial \log h}{\partial\boldsymbol{\phi}^T}$$

we finally obtain the representation

$$\begin{aligned}\mathbf{I}^* &= \sum_i \sum_j w_{ij} E_{ij}\left[\frac{\partial \log p^*_{ij}(\boldsymbol{W}, \boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\frac{\partial \log p^*_{ij}(\boldsymbol{W}, \boldsymbol{\phi})}{\partial\boldsymbol{\phi}^T}\right] - \sum_i \sum_j w_i \Delta_{ij0}\frac{\partial \log \pi_{ij}}{\partial\boldsymbol{\phi}}\frac{\partial \log \pi_{ij}}{\partial\boldsymbol{\phi}^T} \\ &\quad + \sum_i \sum_j w_{ij}\frac{\partial \log(w^{(j)}\pi_{ij} - w_i\Delta_{ij0} + w_{ij})}{\partial\boldsymbol{\phi}}\frac{\partial \log(w^{(j)}\pi_{ij} - w_i\Delta_{ij0} + w_{ij})}{\partial\boldsymbol{\phi}^T}. \end{aligned} \tag{19}$$

4.2 Establishing the efficiency bound

We first describe a general result that shows how the efficiency bound is calculated. Suppose we have observations z from J different populations, with where population j has density $f_j(z, \boldsymbol{\beta}, \gamma)$. The parameter $\boldsymbol{\beta}$ has finite dimension, and the parameter $\boldsymbol{\gamma}$ is infinite-dimensional. Let $\hat{\boldsymbol{\beta}}$ be regular asymptotically linear semi-parametric estimate of $\boldsymbol{\beta}$, based on samples of sizes $n_1, \ldots, n_J$ from the J populations, where $\lim n_j/n_+ \to \nu_j$. Then, the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ must satisfy

$$\text{Var}\,\hat{\boldsymbol{\beta}} \geq \mathbf{B}$$

where $\mathbf{B}$ is the semi-parametric efficiency bound.

The matrix $\mathbf{B}$ may be found as follows (see Lee 2007). Consider the "expected population log likelihood"

$$\sum_j \nu_j E_j[\log f_j(z, \boldsymbol{\beta}, \boldsymbol{\gamma})]. \tag{20}$$

For fixed $\boldsymbol{\beta}$, let $\hat{\boldsymbol{\gamma}}(\boldsymbol{\beta})$ be the maximiser of (20), assuming it exists. The *efficient scores* S_j^* are given by

$$S_j^* = \left.\frac{\partial \log f_j(z, \boldsymbol{\beta}, \hat{\boldsymbol{\gamma}}(\boldsymbol{\beta}))}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}, \quad j = 1, \ldots, J,$$

where $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$. The distributions $f_j(z, \boldsymbol{\beta}, \hat{\boldsymbol{\gamma}}(\boldsymbol{\beta}))$ are called the *least favourable distributions*. The efficiency bound $\mathbf{B}$ is given by

$$\mathbf{B}^{-1} = \sum_j \nu_j E_j[S_j^* S_j^{*T}]. \tag{21}$$

Thus, to establish the efficiency of the procedures we discuss, we need only show that the asymptotic variance of our estimate coincides with $\mathbf{B}$.

Now we apply this theory to regression models for data obtained by the modified sampling scheme described in the previous section. The results obtained will also apply to the case of two-phase response-selective sampling, where the data are obtained by case-control sampling at the first stage.

To calculate the information bound under the modified sampling scheme, we first calculate the expected log-likelihood. Write $p_{ij}(\boldsymbol{w}, \boldsymbol{\beta}) = p_i(\boldsymbol{v}_j, \boldsymbol{w}; \boldsymbol{\beta})$, and denote the true values of parameters by a zero subscript, as in $\boldsymbol{\beta}_0$, γ_{j0} and Δ_{ij0}. The expected log-likelihood is

$$\sum_i \sum_j w_i \Delta_{ij0} \log \Delta_{ij} + \sum_i \sum_j w_{ij} E_{ij} \left[\log p_{ij}(\boldsymbol{W}, \boldsymbol{\beta}) \gamma_j(\boldsymbol{W}) / Q_{ij}\right] \tag{22}$$

where $\gamma_j(\boldsymbol{w})$ is the conditional density of $\boldsymbol{W}$ given $\boldsymbol{V} = \boldsymbol{v}_j$, and

$$\begin{aligned} Q_{ij} &= \int p_{ij}(\boldsymbol{w}, \boldsymbol{\beta}) \gamma_j(\boldsymbol{w})\, d\boldsymbol{w}, \\ q_i &= \sum_j Q_{ij} \rho_j, \\ \Delta_{ij} &= \frac{Q_{ij} \rho_j}{q_i}, \end{aligned}$$

as before. To calculate the least favourable distributions and the efficient scores, we must minimise (22) over the γ_j's and ρ_j's, for $\boldsymbol{\beta}$ held fixed. If $\boldsymbol{W}$ has finite support, an argument analogous to that in Section 3 shows that for fixed $\boldsymbol{\beta}$, the maximising values $\hat{\gamma}_j(\boldsymbol{w}, \boldsymbol{\beta})$ and $\hat{\rho}_j(\boldsymbol{\beta})$ of these parameters satisfy the equations

$$\hat{\gamma}_j(\boldsymbol{w}, \boldsymbol{\beta}) = \frac{p_j^*(\boldsymbol{w}) \gamma_{j0}(\boldsymbol{w})}{\sum_i \mu_{ij} p_{ij}(\boldsymbol{w}, \boldsymbol{\beta})} \tag{23}$$

and

$$\hat{\rho}_j(\boldsymbol{\beta}) = \frac{\sum w_i \Delta_{ij0}}{\sum_i w_i Q_{ij}/q_i}, \tag{24}$$

where

$$p_j^*(\boldsymbol{w}) = \sum_i \frac{w_i}{Q_{ij0}} p_{ij}(\boldsymbol{w}, \boldsymbol{\beta}_0).$$

In (23) and (24), μ_{ij} and q_i are given by

$$\mu_{ij} = (w_i \Delta_{ij} - w_i \Delta_{ij0} + w_{ij})/Q_{ij} \tag{25}$$

and

$$q_i = \sum_j Q_{ij} \hat{\rho}_j,$$

where Q_{ij} satisfies

$$Q_{ij} = \int p_{ij}(\boldsymbol{w}, \boldsymbol{\beta}) \hat{\gamma}_j(\boldsymbol{w}, \boldsymbol{\beta})\, d\boldsymbol{w}. \tag{26}$$

Thus, the solution Q_{ij}, and hence q_i and μ_{ij} are all functions of $\boldsymbol{\beta}$.

To any set of Q_{ij}'s and q_i's, there corresponds a unique set of π_{ij}'s and a unique value of α, through the equations

$$\pi_{ij} = \frac{w_i Q_{ij}/q_i}{w_0 Q_{0j}/q_0 + w_1 Q_{1j}/q_1} \tag{27}$$

and

$$\alpha = \log \frac{w_0}{q_0} \Big/ \frac{w_1}{q_1}. \tag{28}$$

Thus, to the solutions Q_{ij} of (26), and the corresponding q_i's, there is a corresponding set of π's and a corresponding α, which we denote by $\boldsymbol{\eta}(\boldsymbol{\beta}) = (\boldsymbol{\pi}(\boldsymbol{\beta}), \alpha(\boldsymbol{\beta}))$. We also write $\boldsymbol{\phi}(\boldsymbol{\beta}) = (\boldsymbol{\beta}, \boldsymbol{\eta}(\boldsymbol{\beta}))$. Also, because of the relationships between the Q_{ij}'s and q_i's and the π_{ij}'s and α, the definition of μ_{ij} in (25) implies that

$$\mu_{ij} = (w^{(j)} \pi_{ij} - w_i \Delta_{ij0} + w_{ij})/Q_{ij}. \tag{29}$$

The functions p_{ij}^* can also be written in terms of μ_{ij} as

$$p_{ij}^*(\boldsymbol{w}, \boldsymbol{\phi}) = \frac{\mu_{ij} p_{ij}(\boldsymbol{w}, \beta)}{\mu_{0j} p_{0j}(\boldsymbol{w}, \beta) + \mu_{1j} p_{1j}(\boldsymbol{w}, \beta)}. \tag{30}$$

In fact, as we prove in Appendix 1, these formulae remain true (at least in a neighbourhood of $\boldsymbol{\beta}_0$) when the support of $\boldsymbol{W}$ is not necesarily finite. It follows that the efficient scores are

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log \Delta_i \boldsymbol{v}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \log \pi_i \boldsymbol{v}(\boldsymbol{\beta})$$

and

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log \{p_{ij}(\boldsymbol{w}, \boldsymbol{\beta}) \hat{\gamma}_j(\boldsymbol{w}, \boldsymbol{\beta})/Q_{ij}(\boldsymbol{\beta})\} = \frac{\partial}{\partial \boldsymbol{\beta}} \log p_{ij}^*(\boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{\beta})) - \frac{\partial}{\partial \boldsymbol{\beta}} \log \mu_{ij}(\boldsymbol{\beta}) Q_{ij}(\boldsymbol{\beta}),$$

where the first formula follows from the relationship

$$\Delta_{ij}(\boldsymbol{\beta}) = \frac{Q_{ij}(\boldsymbol{\beta})\hat{\rho}_j(\boldsymbol{\beta})}{q_i(\boldsymbol{\beta})}.$$

The inverse of the information bound is (dropping the argument $\boldsymbol{\beta}$)

$$\begin{aligned}
\mathbf{B}^{-1} &= \sum_i \sum_j w_i \Delta_{ij0} \frac{\partial}{\partial \boldsymbol{\beta}} \log \pi_{ij} \frac{\partial}{\partial \boldsymbol{\beta}^T} \log \pi_{ij} \\
&+ \sum_i \sum_j w_{ij} E_{ij} \left\{ \left[\frac{\partial}{\partial \boldsymbol{\beta}} \log p^*_{ij}(\boldsymbol{W}, \boldsymbol{\phi}) - \frac{\partial}{\partial \boldsymbol{\beta}} \log \mu_{ij} Q_{ij} \right] \right. \\
&\left. \times \left[\frac{\partial}{\partial \boldsymbol{\beta}^T} \log p^*_{ij}(\boldsymbol{W}, \boldsymbol{\phi}) - \frac{\partial}{\partial \boldsymbol{\beta}^T} \log \mu_{ij} Q_{ij} \right] \right\}.
\end{aligned}$$

Since

$$E_{ij} \left[\frac{\partial}{\partial \boldsymbol{\beta}} \log p^*_{ij}(\boldsymbol{W}, \boldsymbol{\phi}) \right] = \frac{\partial}{\partial \boldsymbol{\beta}} \log \mu_{ij} Q_{ij},$$

$\mathbf{B}^{-1}$ can be written

$$\begin{aligned}
\mathbf{B}^{-1} &= \sum_i \sum_j w_{ij} E_{ij} \left[\frac{\partial}{\partial \boldsymbol{\beta}} \log p^*_{ij}(\boldsymbol{W}, \boldsymbol{\phi}) \right] \left[\frac{\partial}{\partial \boldsymbol{\beta}^T} \log p^*_{ij}(\boldsymbol{W}, \boldsymbol{\phi}) \right] \\
&+ \sum_i \sum_j w_i \Delta_{ij0} \frac{\partial}{\partial \boldsymbol{\beta}} \log \pi_{ij} \frac{\partial}{\partial \boldsymbol{\beta}^T} \log \pi_{ij} \\
&- \sum_i \sum_j w_{ij} \frac{\partial}{\partial \boldsymbol{\beta}} \log \mu_{ij} Q_{ij} \frac{\partial}{\partial \boldsymbol{\beta}^T} \log \mu_{ij} Q_{ij}.
\end{aligned}$$

Our final task is to prove that this expression coincides with (16). Comparing it to (19), and applying the chain rule we obtain the representation

$$\mathbf{B}^{-1} = \mathbf{I}^*_{\boldsymbol{\beta}\boldsymbol{\beta}} + \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}} \right)^T \mathbf{I}^*_{\boldsymbol{\eta}\boldsymbol{\beta}} + \mathbf{I}^*_{\boldsymbol{\beta}\boldsymbol{\eta}} \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}} \right) + \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}} \right)^T \mathbf{I}^*_{\boldsymbol{\eta}\boldsymbol{\eta}} \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}} \right). \tag{31}$$

In Appendix 2, we show that

$$\left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}} \right) = -(\mathbf{I}^*_{\boldsymbol{\eta}\boldsymbol{\eta}})^{-1} \mathbf{I}^*_{\boldsymbol{\eta}\boldsymbol{\beta}},$$

so from (31) we get

$$\mathbf{B}^{-1} = \mathbf{I}^*_{\boldsymbol{\beta}\boldsymbol{\beta}} - \mathbf{I}^*_{\boldsymbol{\beta}\boldsymbol{\eta}} (\mathbf{I}^*_{\boldsymbol{\eta}\boldsymbol{\eta}})^{-1} \mathbf{I}^*_{\boldsymbol{\eta}\boldsymbol{\beta}}.$$

Thus, the asymptotic variance of the estimate of $\boldsymbol{\beta}$ coincides with the information bound, proving the efficiency of the estimate.

Appendix

A1. The least favourable distribution

We must show that for $\hat{\gamma}_j$ and $\hat{\rho}_j$ as defined in (23) and (24), and for arbitrary densities γ_j and probabilities ρ_j, we have

$$\begin{aligned}
&\sum_i \sum_j w_i \Delta_{ij0} \log \Delta_{ij}(\beta) + \sum_i \sum_j \frac{w_i}{Q_{ij0}} \int \log \hat{\gamma}_j(\boldsymbol{w}) p_{ij}(\boldsymbol{w}, \beta_0) \gamma_{j0}(\boldsymbol{w})\, d\boldsymbol{w} \\
&\qquad - \sum_i \sum_j w_{ij} \log Q_{ij}(\beta) \\
\geq\ &\sum_i \sum_j w_i \Delta_{ij0} \log \Delta_{ij} + \sum_i \sum_j \frac{w_i}{Q_{ij0}} \int \log \gamma_j(\boldsymbol{w}) p_{ij}(\boldsymbol{w}, \beta_0) \gamma_{j0}(\boldsymbol{w})\, d\boldsymbol{w} \\
&\qquad - \sum_i \sum_j w_{ij} \log Q_{ij} \qquad (32)
\end{aligned}$$

where $\pi_{ij}(\beta)$ and $Q_{ij}(\beta)$ are as defined in (26) and the quantity π_{ij} is given by $\pi_{ij} = Q_{ij}\rho_j / (\sum_j Q_{jk}\rho_j)$, where $Q_{ij} = \int \gamma_j(\boldsymbol{w}) p^*_{ij}(\boldsymbol{w}, \beta)\, d\boldsymbol{w}$. The inequality (32) is equivalent to

$$\sum_i \sum_j w_i \Delta_{ij0} \log \frac{\Delta_{ij}(\beta)}{\Delta_{ij}} - \sum_i \sum_j w_{ij} \log \frac{Q_{ij}(\beta)}{Q_{ij}} + \sum_j \int \log \frac{\hat{\gamma}_j(\boldsymbol{w})}{\gamma_j(\boldsymbol{w})} p^*_j(\boldsymbol{w}) \gamma_{j0}(\boldsymbol{w})\, d\boldsymbol{w} \geq 0. \qquad (33)$$

When $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, (23) and (24) show that $\hat{\gamma}_j(\boldsymbol{w}, \boldsymbol{\beta}) = \gamma_{j0}(\boldsymbol{w})$ and that $\rho_j(\boldsymbol{\beta}) = \rho_{j0}$. When $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, (33) becomes

$$\sum_i \sum_j w_j \Delta_{ij0} \log \frac{\Delta_{ij0}}{\Delta_{ij}} - \sum_i \sum_j w_{ij} \log \frac{Q_{ij0}}{Q_{ij}} + \sum_j \int \log \frac{\gamma_{j0}(\boldsymbol{w})}{\gamma_j(\boldsymbol{w})} p^*_j(\boldsymbol{w}) \gamma_{j0}(\boldsymbol{w})\, d\boldsymbol{w} \geq 0. \qquad (34)$$

An argument based on the Kullback-Leibler information inequality shows that the integral in (34) is strictly greater than $w_{+j} \log \frac{Q_{ij0}}{Q_{ij}}$, provided $\gamma_j \neq \gamma_{j0}$. Thus, multiplying the integral by w_{ij} and summing first over i and then over j gives

$$\sum_j \int \log \frac{\gamma_{j0}(\boldsymbol{w})}{\gamma_j(\boldsymbol{w})} p^*_j(\boldsymbol{w}) \gamma_{j0}(\boldsymbol{w})\, d\boldsymbol{w} > \sum_i \sum_j w_{ij} \log \frac{Q_{ij0}}{Q_{ij}}.$$

Moreover, the Kullback-Leibler inequality implies that

$$\sum_i \sum_j w_j \pi_{ij0} \log \frac{\pi_{ij0}}{\pi_{ij}} \geq 0.$$

Hence, the right-hand side of (33) is strictly positive at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, and by a continuity argument is non-negative for all $\boldsymbol{\beta}$ in some neighbourhood of $\boldsymbol{\beta}_0$.

A2: Evaluation of $\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}}$.

Let $\mathcal{E}$ be the expected log-liklihood (22) , and recall that $\hat{\boldsymbol{\gamma}}(\boldsymbol{\beta})$ is the maximiser of $\mathcal{E}$ over all densities $\boldsymbol{\gamma}$ for each fixed $\boldsymbol{\beta}$. Then, arguing as in the data case in Section 3, for each fixed $\boldsymbol{\beta}$, we have

$$\mathcal{E}(\beta, \hat{\boldsymbol{\gamma}}(\boldsymbol{\beta})) = \mathcal{E}^*(\boldsymbol{\beta}, \boldsymbol{\eta}(\boldsymbol{\beta}))$$

where $\boldsymbol{\eta} = (\boldsymbol{\pi}, \alpha)$ and, up to a constant,

$$\begin{aligned} \mathcal{E}^*(\boldsymbol{\beta}, \boldsymbol{\eta}) &= -\sum_i \sum_j w_{ij} E_{ij} \left[\log p^*_{ij}(\boldsymbol{W}, \boldsymbol{\beta}, \boldsymbol{\eta})\right] \\ &\quad -\sum_i \sum_j w_i \Delta_{ij0} \log \pi_{ij} + \sum_i \sum_j w_{ij} \log(w^{(j)} \pi_{ij} - w_i \Delta_{ij0} + w_{ij}). \end{aligned}$$

It follows that $\boldsymbol{\eta}(\boldsymbol{\beta})$ maximises $\mathcal{E}^*(\boldsymbol{\beta}, \boldsymbol{\eta})$ over $\boldsymbol{\eta}$ for each fixed $\boldsymbol{\beta}$, so that

$$\left.\frac{\partial \mathcal{E}^*(\boldsymbol{\beta}, \boldsymbol{\eta})}{\partial \boldsymbol{\eta}}\right|_{\boldsymbol{\eta}=\boldsymbol{\eta}(\boldsymbol{\beta})} = 0$$

for each $\boldsymbol{\beta}$. Differentiating again by the chain rule, we get

$$\left.\frac{\partial^2 \mathcal{E}^*(\boldsymbol{\beta}, \boldsymbol{\eta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\eta}^T}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0, \boldsymbol{\eta}=\boldsymbol{\eta}(\boldsymbol{\beta}_0)} + \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}}\right)^T \left.\frac{\partial^2 \mathcal{E}^*(\boldsymbol{\beta}, \boldsymbol{\eta})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0, \boldsymbol{\eta}=\boldsymbol{\eta}(\boldsymbol{\beta}_0)} = 0.$$

Thus, by (18), we get

$$\mathbf{I}^*_{\boldsymbol{\beta}\boldsymbol{\eta}} + \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}}\right)^T \mathbf{I}^*_{\boldsymbol{\eta}\boldsymbol{\eta}} = 0$$

so that

$$\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}} = -\left(\mathbf{I}^*_{\boldsymbol{\eta}\boldsymbol{\eta}}\right)^{-1} \mathbf{I}^*_{\boldsymbol{\eta}\boldsymbol{\beta}}.$$

References

Anderson, J.A. (1972). Separate sample logistic discrimination. *Biometrika*, **59**, 19-35.

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

Breslow, N.E. (1996). Statistics in epidemiology: the case-control study. *J. Amer. Statist. Soc.*, **91**, 14-28.

Breslow, N.E. and Cain, K.C. (1988). Logistic regression for two-stage case-control data. *Biometrika*, **75**, 11-20.

Breslow, N.E. and Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *Appl. Statist.*, **48**, 4:457-468.

Breslow, N. E. and Day, N. E. (1980). *The Analysis of Case-control Studies*. International Agency for Research on Cancer, Lyon.

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

Breslow, N.E. and Holubkov, R. (1997). Weighted likelihood, pseudolikelihood and maximum likelihood methods for logistic regression analysis of two-stage case-control data. *Statistics in Medicine*, **16**, 103-116.

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**, 1110-39.

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.

Breslow, N.E. and Zhao, L.P. (1988). Logistic regression for stratified case-control studies. *Biometrics*, **44**, 891-899.

Cain, K.C. and Breslow, N.E. (1988). Logistic regression analysis and efficient design for two-stage studies. *Am. J. Epidemiol.*, **128**, 1198-1206.

Engels, E.A., Chen, J., Hartge, P., Cerhan, J.R., Davis, S., Severson, R.K., Cozen, W. and Viscidi, R.P. (2005). Antibody responses to Simian Virus 40 T antigen: a case-control study of Non-Hogkin Lymphoma. *Cancer Epidemiology, Biomarkers and prevention*, **14**, 521-244

Fears, T.R. and Brown, C.C. (1986). Logistic regression methods for retrospective case-control studies using complex sampling procedures. *Biometrics*, **42**, 955-60.

Gill, R.D., Vardi, Y., and Wellner, J.A. (1988). Large sample theory of empirical distributions in biased sampling models. *Ann. Statist.*, **16**, 1069–112.

Heyde, C.C. (1997). *Quasi-likelihood and its application: a general approach to optimal parameter estimation*. New York: Springer.

Jiang, Y. (2004) *Semiparametric maximum likelihood for multi-phase response-selective sampling and missing data problems*. Doctoral dissertation, the University of Auckland.

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

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

Lee, A.J., Scott, A.J. and Wild, C.J. (2005) . Fitting binary regression models with case-augmented samples. *Biometrika* **93**, 385–397.

McNeney, W.B. (1998). *Asymptotic efficiency in semiparametric models with non-i.i.d. data*. Ph.D. Thesis, University of Washington.

Prentice, R.L., and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–11.

Scott, A.J. and Wild, C.J. (1989). Likelihood ratio tests in case/control studies. *Biometrika*, **76**, 806–9.

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

Scott, A.J. and Wild, C.J. (1997). Fitting regression models to case-control data by

maximum likelihood. *Biometrika*, **84**, 57–71.

Scott, A.J. and Wild, C.J. (2001) Maximum likelihood for generalised case-control studies. *Journal of Statistical Planning and Inference*, **96**, 3-27.

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

White, J. E. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease. *Am. J. Epidemiol.*, **115**, 119-128.

Wild, C. J. (1991). Fitting prospective regression models to case-control data. *Biometrika*, **78**, 705-717.