Efficient estimation in multi-phase case-control studies

A.J. LEE, A.J. SCOTT and C.J. WILD
Department of Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand,
aj.lee@auckland.ac.nz

Abstract

In this paper we discuss the analysis of multi-phase, or multi-stage, case-control studies and present an efficient semiparametric maximum-likelihood approach that unifies and extends earlier work, including the seminal case-control paper by Prentice & Pyke (1979) as well as work by Breslow & Cain (1988), Scott & Wild (1997), Breslow & Holubkov (1997), and others. The theoretical derivations apply to arbitrary binary regression models but we present results for logistic regression and show that the approach can be implemented by including additional intercept terms in the logistic model and then making some simple corrections to the score and information equations from the prospective loglikelihood.

Keywords: Logistic regression; maximum likelihood; multi-stage sampling; response-selective sampling; semiparametric efficiency; two and three-phase sampling.

1 Introduction

In a two-phase, or stratified, case-control study, a prospective cohort is stratified according to some variables known for the whole cohort. Separate random samples of cases, i.e. units with some characteristic of interest, and controls, i.e. units without the characteristic, are then drawn from each stratum and values of other covariates are obtained for each of the sampled units. In a three-phase study, some of the more expensive, invasive or difficult covariates are not measured on all the units sampled at the second phase, but only on a subsample drawn from them. This can result in considerable savings. Chatterjee & Chen (2007) point to the increasing importance of such sampling designs in genetic epidemiology, where they can reduce the cost of studies by limiting expensive ascertainments of genetic and environmental exposure to an efficiently selected subsample of the main study.

The process can be continued indefinitely. Whittemore & Halpern (1997) discuss several studies with three or more phases of sampling. For example, in a study to investigate the relationship of prostrate cancer risk to diet and other lifestyle characteristics, the cases were men with a history of prostrate cancer and controls were men without such a history. Case-control status was identified in the initial phase. Then, at the second phase, all the cases and a sample of controls were asked whether or not they had a father or brother with the disease. This information was then used to draw the third phase sample in which more detailed information on family size and structure, age at prostrate cancer occurrence or censoring, and place and date of prostrate cancer diagnosis was collected. Subjects who had three or more family members with prostrate cancer were asked to participate in phase four, in which family members provided blood and/or tissue samples for DNA analysis.

Multi-phase designs have other uses besides reducing the cost of sampling expensive covariates. For example, adding an extra phase of sampling can provide an efficient way of making an after-the-fact adjustment for a confounder that was overlooked and not measured in the original study. In fact, this was the motivation for White (1982) when she introduced the idea of two-phase sampling. Similarly, if we have administrative or other population information available on some variables for all individuals in the finite population from which the study data has been sampled, then efficiency can often be increased by considering the finite-population data as the first phase and the study data as coming from one or more subsequent phases. The methods of this paper may also be useful in some missing-data situations when, under missing-at-random assumptions, the observed/missing mechanism can be thought of as corresponding to an additional phase of sampling. The missing data example of Arbogast et al. (2002), for example, has exactly the same structure as a three-phase case-control sample.

What we are calling multi-*phase* studies have been more commonly called multi-*stage* studies in the biostatistics literature; see Whittemore & Halpern (1997), for example. Multi-phase sampling is the term used in the survey sampling literature where multi-stage sampling already has another well-established meaning; see Cochran (1977), for example. We follow Breslow & Holubkov (1997) in using the survey terminology. We note in passing that the "two-phase" designs of Breslow & Holubkov, in which the initial phase is a case-control sample, are actually three-phase designs in our terminology.

In this paper we present an efficient semiparametric maximum-likelihood solution for multi-phase population-based case-control studies that unifies and extends previous work by Prentice & Pyke (1979), Breslow & Cain (1988), Scott & Wild (1997), Breslow & Holubkov (1997), and others. In the main body of the paper we present the results in a way that is intended to give the reader an appreciation of the nature of the problem, the nature of the solution and how it can be implemented. We will see that for logistic regression the approach can be implemented by including additional intercept terms in the logistic model and then making some simple corrections to the score and information equations from the prospective loglikelihood. The theoretical derivations and justifications are given in the Appendices.

2 Results

2.1 Review of two-phase results

Suppose that we have a binary response variable, Y, with units with $Y = 1$ being the cases and units with $Y = 0$ being the controls, and a vector of potential explanatory variables, X. We want to fit a logistic regression model with

$$p_1(x;\beta) = \text{pr}(Y = 1 \mid x;\beta) = \frac{e^{x^T\beta}}{1+e^{x^T\beta}} \tag{1}$$

for the probability that a unit with covariate values $X = x$ is a case and $p_0(x;\beta) = 1 - p_1(x;\beta)$ for the probability of a control.

We start with a prospectively drawn cohort of N units. In the first phase, we measure the case-control status, Y, and the values of variables $X^{(1)}$, some or all of which may be included in X, for all units in the cohort. We assume that all components of $X^{(1)}$ have finite support, with $X^{(1)}$ having possible values $\{x_1^{(1)}, \ldots, x_I^{(1)}\}$. Let N_{hi} be the number of units in the cohort with $Y = h$ and $X^{(1)} = x_i^{(1)}$ for $h = 0, 1$ and $i = 1, \ldots, I$. In the second phase, we draw a simple random sample of n_{hi} of these N_{hi} units and measure the remaining components of X, resulting in sample values $\{x_{hij} \text{ for } h = 0, 1; i = 1, \ldots, I; \text{ and } j = 1, \ldots, n_{hi}\}$. Components measured at the final stage may be discrete or continuous.

The full likelihood,

$$\prod_{h,i} \left\{ \text{pr}(Y = h, X^{(1)} = x_i^{(1)})^{N_{hi}} \prod_{j=1}^{n_{hi}} \text{pr}(x_{hij} \mid Y = h, X^{(1)} = x_i^{(1)}) \right\},$$

depends not only on β, the parameter of interest, but also on the conditional distribution of X given $X^{(1)} = x_i^{(1)}$ for $i = 1, \ldots, I$. We are not interested in these distributions for their own sake, and we are certainly not interested in modelling them in situations of any complexity, so we want methods for making inferences about β that avoid the need for even thinking about them.

The *conditional maximum likelihood* approach developed by Breslow & Cain (1988) and the semiparametric *profile likelihood* approach developed by Scott & Wild (1997) are both closely related to a simple prospective scheme in which units are examined sequentially, retained in the sample with known probability r_{hi} if $(Y = h, X^{(1)} = x_i^{(1)})$ and otherwise discarded. Under this scheme, the probability that $Y = h$, given that $(X^{(1)} = x_i^{(1)}, X = x)$ and that the unit is selected, would be

$$\text{pr}(Y = h \mid X^{(1)} = x_i^{(1)}, X = x) = \frac{p_h(x; \beta) r_{hi}}{p_1(x; \beta) r_{1i} + p_0(x; \beta) r_{0i}}. \tag{2}$$

Writing $\alpha_i = \log(r_{1i}/r_{0i})$ and using (1), we can express pr$(Y = 1 \mid X^{(1)} = x_i^{(1)}, X = x)$ in the form

$$p_{1i}^*(x; \alpha_i, \beta) = \frac{e^{\alpha_i + x^T\beta}}{1 + e^{\alpha_i + x^T\beta}}. \tag{3}$$

This is the original logistic regression model (1) modified by the inclusion of stratum-specific offsets. If our data had been generated by this slightly modified scheme, the log-likelihood function would be

$$\widetilde{\ell}_2(\beta) = \sum_{h,i,j} \log p_{hi}^*(x_{hij}; \alpha_i, \beta) \tag{4}$$

where $p_{0i}^* = 1 - p_{1i}^*$, and estimation would proceed straightforwardly using an ordinary prospective logistic regression program by fitting model (1) with the $\alpha_i = \log(r_{1i}/r_{0i})$ values included as offsets.

Conditional maximum likelihood:

To adapt these ideas to actual two-phase case-control sampling, suppose that we estimate the retention probability r_{hi} by the sampling fraction n_{hi}/N_{hi}, $(i = 1, 0)$ and thus α_i by

$$\widehat{\alpha}_{i,\text{CML}} = \frac{n_{1i}/N_{1i}}{n_{0i}/N_{0i}}, \tag{5}$$

the log of the relative sampling fractions. The conditional maximum likelihood estimator is obtained if we maximize (4), with α_i replaced by $\widehat{\alpha}_{i,\text{CML}}$, with respect to β. In this context (4) is only a pseudo-likelihood. The parameter estimates can still be obtained from an ordinary logistic regression program by fitting model (1) with the $\widehat{\alpha}_{i,\mathrm{CML}}$ values included as offsets. However, additional computation is usually required to obtain valid standard errors.

Profile likelihood:
If the α_is in the modified model (3) are treated as free parameters, then (3) can be thought of as the original logistic model (1) augmented by a set of unknown stratum-specific intercepts. By adapting the results of Scott & Wild (1997), the profile likelihood estimator, $\widehat{\beta}$, can be obtained as follows. Take the likelihood that would be appropriate for fitting the augmented logistic model (3) prospectively, add forcing terms $c_i(\alpha_i)$ to give

$$\ell_2^*(\phi) = \ell_2^*(\alpha, \beta) = \sum_{h,i,j} \log p_{hi}^*(x_{hij}; \alpha_i, \beta) + \sum_{i=1}^{I} c_i(\alpha_i), \tag{6}$$

and then solve the resulting score equations for $\widehat{\phi} = (\widehat{\alpha}, \widehat{\beta})$. The forcing terms, presented in equation (13) in Section 2.3, have the effect of pushing the unknown α_i towards $\widehat{\alpha}_{i,\mathrm{CML}}$. Inference using (6), including variance estimation and hypothesis testing, is very simple because we can treat $\ell_2^*(\phi)$ almost like an ordinary log-likelihood. We can obtain $\widehat{\beta}$ by solving the pseudo-likelihood equations, $\partial\ell_2^*/\partial\phi = 0$, we can estimate $\mathrm{cov}(\widehat{\beta})$ using the appropriate submatrix of $J_2^*(\widehat{\phi})^{-1}$, where $J_2^*(\phi)$ is the observed pseudo-information matrix, and we can treat the appropriate differences in $-2\ell_2^*$ as chi-squared random variables to test hypotheses about β.

The profile likelihood itself, $\ell_P(\beta)$, is obtained by maximizing the full likelihood over the unknown conditional distributions of X given $X^{(1)} = x_i^{(1)}, i = 1, \ldots, I$ treated nonparametrically. The connection with $\ell_2^*(\alpha, \beta)$ is that $\ell_P(\beta) = \ell_2^*\{\alpha(\beta), \beta\}$ with $\alpha(\beta)$ defined as the solution of $\partial\ell_2^*(\alpha, \beta)/\partial\alpha = 0$. The proviso that $\ell_2^*(\phi)$ can be treated "*almost* like an ordinary log-likelihood" is needed because $\widehat{\phi}$ can correspond to a saddlepoint of ℓ_2^* rather than a maximum so that we cannot obtain $\widehat{\beta}$ by maximizing $\ell_2^*(\phi)$ in general. The profile likelihood estimator can be shown to have full semi-parametric efficiency so that it is more efficient than conditional maximum likelihood; see Breslow, McNeney & Wellner (2003), Lee & Hirose (2009). The difference in efficiency is often small but there are situations when it is appreciable; see Scott & Wild (1991), Lawless et al. (1999), for example.

2.2 Results for three phases

Now suppose that only a subset of the remaining components of X, say $X^{(2)}$, are measured at the second phase of sampling. We assume that $X^{(2)}$ also has finite support, with possible values $\{x_1^{(2)}, \ldots, x_J^{(2)}\}$ say. Let N_{hij} be the number of second-phase sample units taking values $Y = h, X^{(1)} = x_i^{(1)}, X^{(2)} = x_j^{(2)}$ for $h = 0, 1; i = 1, \ldots, I; j = 1, \ldots, J$. Note that $N_{hi+} = n_{hi}$. Then, at the third phase of sampling, we draw a simple random sample of n_{hij} of these N_{hij} units and measure the remaining components of X. This results in sample data $\{x_{hijk}$ for $h = 0, 1,\ i = 1, \ldots, I, j = 1, \ldots, J,$ and $k = 1, \ldots, n_{hij}\}$ collected at phase three and a likelihood of the form

$$\prod_{h,i} \Bigg[\mathrm{pr}(Y = h, X^{(1)} = x_i^{(1)})^{N_{hi}} \prod_{j=1}^{J} \Big\{ \mathrm{pr}(X^{(2)} = x_j^{(2)} \mid Y = h, X^{(1)} = x_i^{(1)})^{N_{hij}} \\ \times \prod_{k=1}^{n_{hij}} \mathrm{pr}(x_{hijk} \mid Y = h, X^{(1)} = x_i^{(1)}, X^{(2)} = x_j^{(2)}) \Big\} \Bigg]. \tag{7}$$

As with two-phase sampling, this contains the joint distribution of X as a nuisance parameter and we want methods that avoid any need to model this joint distribution.

We can extend the profile likelihood method of the previous section directly. Again it is based on a related prospective scheme, now with

$$\text{pr}(Y = 1 \mid X^{(1)} = x_i^{(1)}, X^{(2)} = x_j^{(2)}, x) = p_{1ij}^*(x; \alpha, \beta) = \frac{e^{\alpha_i + \alpha_{ij} + x^T\beta}}{1 + e^{\alpha_i + \alpha_{ij} + x^T\beta}}, \tag{8}$$

as in model (3) but now with an additional intercept term for every cell of the $X^{(1)} \times X^{(2)}$ stratification used to classify the Phase 2 data. The extension of the conditional maximum likelihood method is to estimate β by fitting model (1) by prospective logistic regression with stratum-specific offsets $\widehat{\alpha}_{i,CML} + \widehat{\alpha}_{ij,CML}$ where $\widehat{\alpha}_{ij,CML} = \log\{(n_{1ij}/N_{1ij})/(n_{0ij}/N_{0ij})\}$. We obtain the profile likelihood estimator essentially by expanding (6) to include an additional set of forcing terms stemming from the phase three subsampling to form

$$\ell_3^*(\alpha, \beta) = \sum_{h,i,j,k} \log p_{hij}^*(x_{hijk}; \alpha, \beta) + \sum_i c_i(\alpha_i) + \sum_{i,j} c_{ij}(\alpha_{ij}), \tag{9}$$

with $p_{0ij}^* = 1 - p_{1ij}^*$, and solving the resulting score equations as before. Expressions for the forcing terms $c_t(\alpha_t)$ are given in equation (13). We will show that their derivatives, $dc_t/d\alpha_t$, increase monotonely from $-n_{1t}$ to n_{0t} crossing zero at $\widehat{\alpha}_{t,CML}$. Thus the effect of the added terms is to pull the α_t-component of the solution of the pseudo-score equation towards $\widehat{\alpha}_{t,CML}$. Here, and in much of what follows, we have adopted the notational device of using t to represent an arbitrary cell either in the one-way classification defined by values of $X^{(1)}$ or in the two-way $X^{(1)} \times X^{(2)}$ classification.This notation also allows for extensions to further phases where needed.

Although $\ell_3^*(\phi)$ is not itself a true likelihood, we show in Appendix 1 that the profile likelihood $\ell_P(\beta)$ is equal to $\ell_3^*\{\beta, \alpha(\beta)\}$ where $\alpha(\beta)$ is the solution to $\partial\ell_3^*/\partial\alpha = 0$. A consequence of this equivalence is that, just as in two phase sampling, we can largely act as if the pseudo-loglikelihood $\ell_3^*(\phi)$ is the true log-likelihood for making inferences about β. Specifically, we can obtain $\widehat{\beta}$ by solving the pseudo-score equations obtained by setting the derivatives of (9) to zero, we can estimate $\text{cov}(\widehat{\beta})$ with the appropriate submatrix of $J_3^*(\widehat{\phi})^{-1}$, where $J_3^*(\phi)$ is the observed pseudo-information matrix, and we can treat the appropriate differences in $-2\ell_3^*$ as chi-squared random variables to test hypotheses about β.

To implement this, we need the first and second derivatives of ℓ_3^* with respect to the ϕ. Let us rewrite model (8) in the form

$$\text{logit}\, p_{1ij}^*(x; \phi) = z^T\phi,$$

where the first $T = (I + IJ)$ elements of z indicate the presence or absence of the corresponding components of α, i.e. the first element of z_{hijk}, corresponding to α_1, is equal to 1 if i=1 and 0 otherwise, and so on. The final elements of z are made up of the elements of x. Then we can write

$$U_3^*(\phi) = \frac{\partial\ell_3^*}{\partial\phi} = \sum_{h,i,j,k} z_{hijk}\{y_{hijk} - p_1(z_{hijk}; \phi)\} + \begin{pmatrix} \gamma \\ 0 \end{pmatrix}, \tag{10}$$

and

$$J_3^*(\phi) = -\frac{\partial^2\ell_3^*(\phi)}{\partial\phi\partial\phi^T} = \sum_{h,i,j,k} z_{hijk} z_{hijk}^T p_1(z_{hijk}) p_0(z_{hijk}) - \begin{pmatrix} \text{diag}(A) & 0 \\ 0 & 0 \end{pmatrix}, \tag{11}$$

with γ representing a T-dimensional vector with components γ_t, A representing a T-dimensional vector with components A_t, and $y_{hijk} = h$. The quantities γ_t and A_t are defined below in (12) and (14). Equations (10) and (11) are just the usual score and information expressions for logistic regression except for adjustments to the components corresponding to elements of α made by γ and the A_ts.

It remains to define the elements of γ and A_t. If a cell is fully subsampled, i.e. $n_{1t} = N_{1t}$ and $n_{0t} = N_{0t}$, then $\alpha_t \equiv 0$, $c_t(\alpha_t) \equiv 0$, $\gamma_t = 0$ and $A_t = 0$. For cells that are not fully subsampled, it is convenient to express c_t and related quantities in terms of $\gamma_t(\alpha_t)$, defined as the unique solution of

$$\log\left(\frac{n_{1t}+\gamma_t}{N_{1t}+\gamma_t}\right) - \log\left(\frac{n_{0t}-\gamma_t}{N_{0t}-\gamma_t}\right) = \alpha_t. \tag{12}$$

This corresponds to the γ_0-parameters used in Scott & Wild (1997, pp 60, 65, 68) where γ_0 was shown to equal $N_0 - \widehat{N_0}$, the difference between the number of controls in the population and the number predicted by the fitted model. Working with this parameterization, rather than other possibilities such as those used in Scott & Wild (2001), proved to be critical in obtaining the results of this paper. It is shown in Appendix A1 that

$$c_t(\alpha_t) = N_{1t}\log(N_{1t}+\gamma_t) - n_{1t}\log(n_{1t}+\gamma_t) + N_{0t}\log(N_{0t}-\gamma_t) - n_{0t}\log(n_{0t}-\gamma_t), \tag{13}$$

with $\gamma_t = \gamma_t(\alpha_t)$. Differentiating the expressions in equations (12) and (13) with respect to α_t, it follows that $c_t(\alpha)$ has first derivative $dc_t/d\alpha_t = \gamma_t$ and second derivative $d^2c_t/d\alpha_t^2 = A_t$ where

$$A_t = \left\{\frac{N_{1t}-n_{1t}}{(N_{1t}+\gamma)(n_{1t}+\gamma)} + \frac{N_{0t}-n_{0t}}{(N_{0t}-\gamma)(n_{0t}-\gamma)}\right\}^{-1}. \tag{14}$$

Setting $dc_t/d\alpha_t = 0$ leads to $\alpha_t = \widehat{\alpha}_{t,CML} = \log\{(n_{1t}/N_{1t})/(n_{0t}/N_{0t})\}$. As α_t increases from $-\infty$ to $+\infty$, $\gamma_t(\alpha_t)$ increases monotonically from $-n_{1t}$ to n_{0t}, with derivative $d\gamma_t/d\alpha_t = A_t$.

Computationally, we solve the pseudo-likelihood equations above using the Newton-Rhapson method starting from the conditional maximum likelihood solution and $J_3^*(\widehat{\phi})$ is obtained as a byproduct. The conditional maximum likelihood solution is obtained by performing an ordinary logistic regression on the data from the fully observed units, i.e. those observed at Phase 3, but including the appropriate offsets. To cater for the cases where there are fully subsampled cells, we need to work with a reduced set of parameters $\check{\phi} = (\check{\alpha}, \beta)$. Here, $\check{\alpha}$ contains just the α_t-values for cells that are not fully subsampled. We can write $\check{\phi} = B\phi$. Then, $\partial\ell^*/\partial\check{\phi} = B\partial\ell^*/\partial\phi$ and $\partial^2\ell^*(\phi)/\partial\check{\phi}\partial\check{\phi}^T = B\left(\partial^2\ell^*/\partial\phi\partial\phi\right)B^T$. Additionally, $\phi = B^T\check{\phi}$. Some care is needed in handling the data structures. A software implementation, in the form of an R function, R Development Core Team (2008), is available from Chris Wild.

2.3 Extensions

It is straightforward to extend these results to more phases of sampling and to other binary regression models besides the logistic.

To handle S phases of sampling, we simply have to augment the model in equation (8) with an additional constant term, α_t, and add a corresponding term, $c_t(\alpha_t)$, defined by equation (13), to the pseudo log-likelihood in equation (9) for every cell of the $X^{(1)} \times X^{(1)} \times \ldots \times X^{(s)}$

classification, assuming that $X^{(s)}$ has finite support ($s = 1, \ldots, S-1$). For example, with four phase sampling, our pseudo model would have the form

$$\text{logit}\{p^*_{1ijk}(\alpha, \beta)\} = x^T\beta + \alpha_i + \alpha_{ij} + \alpha_{ijk} \tag{15}$$

and the pseudo log-likelihood would be

$$\ell^*_4(\alpha, \beta) = \sum_{h,i,j,k,l} \log p^*_{hijk}(x_{hijkl}; \alpha, \beta) + \sum_{i=1} c_i(\alpha_i) + \sum_{i,j} c_{ij}(\alpha_{ij}) + \sum_{i,j,k} c_{ijk}(\alpha_{ijk}).$$

We can then use $\ell^*_S(\phi)$ to make inferences about β just as with two or three phases. To adapt (10) and (11), the sums are taken over all individuals observed in the final phase of sampling and γ and A are lengthened in the obvious way.

The extension to non-logistic binary regression models is also simple. Suppose that $\text{pr}(Y = h \mid x) = p_h(x; \beta)$. Then we just replace $x^T\beta$ in the definition of $\text{pr}^*_{ht}(x; \phi)$ by logit$\{p_1(x; \beta)\}$ and proceed exactly as with the logistic. Thus, in four phase sampling, (15) would be replaced by

$$\text{logit}\{p^*_{1ijk}(\alpha, \beta)\} = \text{logit}\{p_1(x; \beta)\} + \alpha_i + \alpha_{ij} + \alpha_{ijk}.$$

All the results for the logistic case, apart from the simplification discussed in the next section that results when the model contains appropriate dummy variables, then apply immediately.

2.4 Special cases

Prentice & Pyke (1979), following earlier work by Anderson (1972), showed that for the two-phase case-control studies ($S = 2$) where the logistic model has a separate intercept β_{0i} for every level of $X^{(1)}$, maximum likelihood inferences about all the coefficients except the intercepts can be obtained by running the sample data through a standard logistic regression program without any modification. Semiparametric efficiency follows from Breslow, McNeney & Wellner (2003) and Lee & Hirose (2009). If terms $\widehat{\alpha}_i = \log\{(n_{1i}/N_{1i})/(n_{0i}/N_{0i})\}$ are added as fixed offsets, then estimates of the β_{0i}s are also valid and the variances can be corrected by subtracting $n_{1i}^{-1} - N_{1i}^{-1} + n_{0i}^{-1} - N_{0i}^{-1}$ from the estimated variance of $\widehat{\beta}_{0i}$ from the logistic program.

The situation where the first phase of a three-phase study is a simple (unstratified) case-control sample and our model (1) includes an intercept, β_0 say, is another special case of interest. Here $I = 1$ and Breslow & Holubkov (1997) noted that for making inferences about all coefficients except β_0, we can act as if we had an two-phase study with a prospective first phase. This was explored further in Lee, Scott & Wild (2007).

We now consider three-phase sampling more generally. Setting $\beta^*_{0i} = \beta_{0i} + \alpha_i$ and $\alpha^* = (\alpha_{11}, \alpha_{12}, \ldots, \alpha_{IJ})^T$, we can write the model in expression (8) in the form

$$\text{logit}\{p^*_{1t}(x; \alpha^*, \beta^*)\} = \alpha_t + x^T\beta^*,$$

where t ranges over all possible values of (i, j) for $i = 1, \ldots, I$ and $j = 1, \ldots, J$. In other words, the three-phase model (8) can be rewritten in the form of the two-phase model (3) with $X^{(1)}$ replaced by $X^{*(1)} = (X^{(1)}, X^{(2)})$ taking $I^* = IJ$ possible values. Then the three-phase pseudo log-likelihood (9) can be written in the form

$$\ell^*_3(\phi) = \ell^*_2(\phi^*) + \sum_i c_i(\alpha_i),$$

with $\phi^* = (\alpha^*, \beta^*)$ and ℓ_2^* as in equation (6). It follows that the pseudo-score equations, $\partial\ell_3^*(\phi)/\partial\phi = 0$, become

$$\frac{\partial\ell_2^*(\phi^*)}{\partial\phi^*} = 0 \text{ and } \frac{dc_i(\alpha_i)}{d\alpha_i} = 0, \text{ for } 1 = 1, \ldots, I,$$

and the pseudo-information matrix becomes

$$J_3^* = -\frac{\partial^2\ell_3^*}{\partial\phi\partial\phi^T} = D_0 \oplus J_2^*,$$

where $J_2^* = -\frac{\partial^2\ell_2^*}{\partial\phi^*\partial\phi^{*T}}$ and D_0 is an $I \times I$ diagonal matrix. Thus, to make inferences about components of β^*, and hence about all the components of β apart from the constant terms, we can treat $\ell_2^*(\phi^*)$ as a two-phase pseudo-loglikelihood. This can be useful as it does not require knowledge of the N_{hi}s. If we want to estimate β_{0i}, and we do know the N_{hi}s, then we need to subtract $\widehat{\alpha}_{i,CML} = \log\{(n_{1i}/N_{1i})/(n_{0i}/N_{0i})\}$, the solution of $dc_i(\alpha_i)/d\alpha_i = 0$, from $\widehat{\beta}_{0i}^*$ and subtract $n_{1i}^{-1} - N_{1i}^{-1} + n_{0i}^{-1} - N_{0i}^{-1}$ from the variance of $\widehat{\beta}_{0i}^*$ estimated from the two-phase program exactly as in the simpler case above.

More generally, it can be shown that where the model has a separate intercept for every level of $X^{(1)} \times \ldots \times X^{(s)}$ for some $s < S$, then the analysis can be reduced to that for $(S - s)$ phases with offsets and variance adjustments needed where estimation of the intercepts is of interest.

3 Example

We use the Wilm's tumor data described in Kulich & Lin (2004) to illustrate the method. The studies from which the data originated were performed by the US National Wilms Tumor Study Group (DAngio et al., 1989; Green et al., 1998). Wilms tumor is a rare kidney cancer occurring in young children. The data relates to 3,915 children who had been treated for Wilms tumor. We take as our binary response variable "relapse within 3 years". The explanatory variables available were histological type of the tumor, classified as favorable versus unfavorable, stage (I $-$ IV), age at diagnosis, and tumor diameter. Breslow & Chatterjee (1999) worked with a slight superset of this data set, but without the age and tumor diameter variables, to construct two-phase data.

Quoting from Kulich & Lin (2004), "Histological type was assessed in two ways. Pathologists at the individual sites analyzed a tumor sample and determined a preliminary local histological type. Each sample was then sent to a central facility, where an experienced pathologist reevaluated it. This reevaluation was an expensive and time-consuming process. The central assessment can be considered the 'true' histological type, and the local assessment can be considered an imprecise surrogate." Although central histology was obtained for all patients in the study, Breslow & Chatterjee (1999) and Kulich & Lin (2004), performed analyses on two-phase data sets obtained post-hoc by subsampling and only using the central histology of subsampled patients in their analysis. They did this to show how well you could do with a cheaper two-phase study that only obtained the expensive measure, central histology, for a subset of patients.

We do the same thing here. We will use tumor diameter to play the role of a variable, e.g. genetic, that is even more expensive than central Histology and will only be obtained at Phase 3.

Unfavourable histologies are coded “1” and favourable are coded as “0”. We set up 24 Phase-1 strata defined by Institutional histology (Inst), Stage and 3 levels of Age (Age ≤ 1, $1 <$ Age ≤ 4 and Age > 4. The numbers of controls and cases falling into these strata are given in the N_{0i} and N_{1i} columns of Table 1. This is the Phase 1 information. There were 3312 controls and 603 cases at Phase 1. At Phase 2, n_{0i} controls ($i = 1, \ldots, 24$) and all cases were taken and the value of central Histology obtained. We sampled $n_{0i} = 100$ units from any cell with $N_{0i} \geq 100$ and retained all units in cells with $N_{0i} \leq 100$. This resulted in 1248 controls and 603 cases being observed at Phase 2. Crossing the new variable, central Histology, with the original set of 24 strata now produces 48 strata. Of those sampled at Phase 2, the numbers falling into each of these 48 strata are given in the N_{0i0}, N_{0i1} columns of Table 1 for controls and the corresponding N_{1i0}, N_{1i1} columns for cases. At Phase 3, we sampled n_{0i1} controls with unfavourable (central) Histology and n_{0i0} controls with favourable Histology. We subsampled 25 units in cells of more than 25 and took all units in smaller cells. Cases with favourable histology were subsampled in the same way. The new variable observed at Phase 3 was Tumor diameter.

Below we present comparative results for a model that fits the full data quite well. The standard errors presented for 3-phase sampling are averages over 1,000 replications of the process of drawing second and third phase samples. We have to expect some loss of information because values of our “most expensive” variable are only collected for a 25% subsample. If we had analysed a random 25% subsample of the full data we would expect, on average, a doubling of the standard errors. We are doing a lot better than that for many of the variables fitted. Even for Tumor diameter, which was only collected at Phase 3, and its interaction with Stage, there is only a 40% increase in standard error. We have almost full efficiency for some of the variables collected at Phase 1 or even central Histology which was only collected at Phase 2. Of course institutional histology, collected at Phase 1, is a very good surrogate for central histology which was collected at Phase 2.

4 Discussion

Multi-phase sampling has a considerable potential for delivering efficiency gains for case-control studies. This paper unifies and extends the methodology for special cases of multi-phase case-control studies given by Prentice and Pyke (1979), Scott and Wild (1997), Breslow and Holubkov (1997), Lawless et al. (1999) and Lee et al (2007) to provide a general semiparametric-efficient solution to the problem of analyzing multi-phase case-control studies that collect discrete or continuous covariate information at the last stage of sampling and discrete covariate information at previous phases. The methods are relatively easy to implement, particularly in the case of logistic regression where we have shown how to make relatively minor modifications to the score vector and information matrix for ordinary prospective logistic regression. The need for covariate data used in all but the last phase to be discrete is an important limitation.

All of our asymptotics are carried out under the assumption that the model to be true. As a referee has pointed out, however, it would be beneficial to investigate whether the asymptotic methods of Newey (1994), which allow for fairly general model misspecifications, can be generalized to multiphase case-control studies. As as shown in Scott and Wild (2002), the parameter estimated by semiparametric maximum likelihood from a case-control study will not generally

Table 1. *Three-phase sampling counts*

					Controls						Cases			
			Ph 1	Phase 2			Phase 3		Ph 1	Phase 2			Phase 3	
	Phase 2				Hist		Hist				Hist		Hist	
	strata				0	1	0	1			0	1	0	1
Inst	Stg	Age	N_{0i}	n_{0i}	N_{0i0}	N_{0i1}	n_{0i0}	n_{0i1}	N_{1i}	n_{1i}	N_{1i0}	N_{1i1}	n_{1i0}	n_{1i1}
0	1	≤ 1	387	100	100	0	25	0	35	35	32	3	25	1
0	1	(1,4]	672	100	97	3	25	3	49	49	48	1	25	1
0	1	> 4	283	100	97	3	25	3	36	36	35	1	25	1
0	2	≤ 1	78	78	76	2	25	2	9	9	8	1	8	1
0	2	(1,4]	432	100	96	4	25	4	60	60	56	4	25	4
0	2	> 4	254	100	96	4	25	4	66	66	58	8	25	8
0	3	≤ 1	40	40	37	3	25	3	4	4	4	0	4	0
0	3	(1,4]	337	100	100	0	25	0	37	37	34	3	25	3
0	3	> 4	296	100	99	1	25	1	63	63	55	8	25	8
0	4	≤ 1	1	1	1	0	1	0	5	5	3	2	3	2
0	4	(1,4]	141	100	97	3	25	3	33	33	31	2	25	2
0	4	> 4	162	100	98	2	25	2	53	53	48	5	25	5
1	1	≤ 1	8	8	1	7	1	7	8	8	0	8	0	8
1	1	(1,4]	36	36	4	32	4	32	7	7	1	6	1	6
1	1	> 4	19	19	1	18	1	18	3	3	0	3	0	3
1	2	≤ 1	4	4	0	4	0	4	7	7	0	7	0	7
1	2	(1,4]	32	32	5	27	5	27	11	11	0	11	0	11
1	2	> 4	25	25	4	21	4	21	15	15	0	15	0	15
1	3	≤ 1	2	2	0	2	0	2	17	17	0	17	0	17
1	3	(1,4]	41	41	8	33	8	33	21	21	0	21	0	21
1	3	> 4	26	26	6	20	6	20	22	22	2	20	2	20
1	4	≤ 1	0	0	0	0	0	0	12	12	0	12	0	12
1	4	(1,4]	22	22	13	9	13	9	15	15	2	13	2	13
1	4	> 4	14	14	9	5	9	5	15	15	2	13	2	13
		Totals	3312	1248	1045	203	327	203	603	603	419	184	247	181

Table 2. *Full-data analysis compared with 3-phase analysis*

	Full data			3-phase			Ratio
	Estimate	se	z	Estimate	se	z	of ses
Intercept	-4·08	0·390	-10·4	4·02	0·538	7·5	1·38
Histology	1·30	0·125	10·4	1·33	0·133	10·0	1·06
Stage	0·81	0·149	5·5	0·86	0·204	4·2	1·37
Age ≤ 1	-0·29	0·180	-1·6	-0·26	0·187	-1·4	1·04
$1 <$ age < 4	-0·47	0·102	-4·6	-0·47	0·105	-4·4	1·03
Hist$\times$ Age ≤ 1	1·77	0·347	5·1	1·61	0·351	4·6	1·01
Tumor diam·	0·14	0·031	4·6	0·13	0·045	3·0	1·45
Stage$\times$Tumor	-0·04	0·012	-3·4	-0·04	0·017	-2·6	1·42

[3-phase results are means from 1,000 subsamples]

be the same as the parameter estimated from from complete-cohort data when there is model misspecification. Using the broader form of asymptotics would, for example, facilitate investigating asymptotic efficiencies in a way that penalises departures from the complete-cohort limit. The issues are subtle, however, and our earlier work casts some doubt on whether this is always the relevant comparison. We hope to explore this in the future.

Appendix 1

Derivation of the estimating equation

All derivations are for an arbitrary regression function of the form $p_h(x, \beta) = \text{pr}(Y = h \mid X = x)$.

Set $X^T = (X^{(1)^T}, X^{(2)^T}, X^{(3)^T})$ and assume first that $X^{(3)}$ has finite support, taking on values $x_k^{(3)}$. Let N_{hijk} denote the number of times the value $x_k^{(3)}$ appears in the hij stratum and let $p_{hijk}(\beta) = \text{pr}(Y = h \mid X^{(1)} = x_i^{(1)}, X^{(2)} = x_j^{(2)}, X^{(3)} = x_k^{(3)})$. We parameterize the log-likelihood in terms of the basic parameters $g_{ijk} = \text{pr}(X^{(3)} = x_k^{(3)} \mid X^{(1)} = x_i^{(1)}, X^{(2)} = x_j^{(2)})$, $\xi_{ij} = \text{pr}(X^{(2)} = x_j^{(2)} \mid X^{(1)} = x_i^{(1)})$ and $\zeta_i = \text{pr}(X^{(1)} = x_i^{(1)})$. It is convenient to introduce the notation

$$\begin{aligned} \psi_{hij} &= \sum_k p_{hijk}(\beta) g_{ijk} = \text{pr}(Y = h \mid X^{(1)} = x_i^{(1)}, X^{(2)} = x_j^{(2)}), \\ \pi_{hi} &= \sum_j \psi_{hij} \xi_{ij} = \text{pr}(Y = h | X^{(1)} = x_i^{(1)}). \end{aligned}$$

Then from equation (7) in Section 2.2 and recalling that $N_{hi+} = n_{hi}$, the log-likelihood is

$$\begin{aligned} \ell(\beta, \zeta, \xi, g) &= \sum_{hijk} n_{hijk} \log p_{hijk}(\beta) + \sum_{hi} (N_{hi} - n_{hi}) \log \pi_{hi} + \sum_{hij} (N_{hij} - n_{hij}) \log \psi_{hij} \\ &\quad + \sum_{ij} N_{+ij} \log \xi_{ij} + \sum_{ijk} n_{+ijk} \log g_{ijk} + \sum_i N_{+i} \log \zeta_i. \qquad (16) \end{aligned}$$

We want to maximize $\ell(\beta, \zeta, \xi, g)$ with respect to ζ, ξ and g to obtain the profile likelihood of β. We see that terms involving β, which is the quantity of interest, cannot be factored out from the nuisance parameters ξ and the g_{ijk}'s, which are of little or no interest in their own right. However, the parameter ζ is orthogonal to the other parameters, and can be ignored in what follows. Accordingly we drop the last term from the log-likelihood and write $\ell(\beta, \xi, g)$.

We introduce Lagrange multipliers η_i and η_{ij} to take care of the constraints $\sum_j \xi_{ij} = 1$ and $\sum_k g_{ijk} = 1$. Differentiating (16) with respect to ξ_{ij} and setting the result equal to η_i leads to

$$\frac{N_{+ij}}{\xi_{ij}} + \sum_h (N_{hi} - n_{hi}) \frac{\psi_{hij}}{\pi_{hi}} = \eta_i. \tag{17}$$

Multiplying (17) through by ξ_{ij} and summing over j gives $\eta_i = N_{+i}$ so that the maximizing values of ξ_{ij} satisfy

$$\xi_{ij} = \frac{N_{+ij}}{\sum_h \frac{\mu_{hi}\psi_{hij}}{\pi_{hi}}} \tag{18}$$

where $\mu_{hi} = N_{+i}\pi_{hi} - (N_{hi} - n_{hi})$. Similarly, differentiating (16) with respect to g_{ijk} and setting the result equal to η_{ij} leads to

$$\frac{n_{+ijk}}{g_{ijk}} + \sum_h \left\{ (N_{hi} - n_{hi}) \frac{\xi_{ij}}{\pi_{hi}} + (N_{hij} - n_{hij})/\psi_{hij} \right\} p_{hijk} = \eta_{ij}. \tag{19}$$

Multiplying (19) through by g_{ijk}, summing over k and applying (17) then gives $\eta_{ij} = N_{+i}\xi_{ij}$. Thus the maximizing values of g_{ijk} satisfy the equations

$$g_{ijk} = \frac{n_{+ijk}}{\sum_h \frac{\mu_{hij}}{\psi_{hij}} p_{hijk}(\beta)}, \tag{20}$$

where

$$\begin{aligned} \mu_{hij} &= N_{+ij}\pi_{hij} - (N_{hij} - n_{hij}) \\ \pi_{hij} &= \frac{\mu_{hi}\psi_{hij}/\pi_{hi}}{\mu_{0i}\psi_{0ij}/\pi_{0i} + \mu_{1i}\psi_{1ij}/\pi_{1i}}. \end{aligned} \tag{21}$$

Substituting the expressions (18) and (20) for ξ_{ij} and g_{ijk} into (16), we obtain the profile likelihood

$$\begin{aligned} \ell_P(\beta) &= \sum_{hijk} n_{hijk} \log p^*_{hijk}(\beta) + \sum_{hi} (N_{hi} \log \pi_{hi} - n_{hi} \log \mu_{hi}) \\ &\quad + \sum_{hij} (N_{hij} \log \pi_{hij} - n_{hij} \log \mu_{hij}) \end{aligned} \tag{22}$$

where

$$p^*_{hijk} = \frac{\frac{\mu_{hij}}{\psi_{hij}} p_{hijk}}{\sum_h \frac{\mu_{hij}}{\psi_{hij}} p_{hijk}}.$$

Note that (21) and the fact that $\psi_{+ij} = 1$ imply that

$$\psi_{hij} = \frac{\pi_{hi}\pi_{hij}/\mu_{hi}}{\sum_h \pi_{hi}\pi_{hij}/\mu_{hi}}$$

so that $\ell_P(\beta)$ is a function of the π's. These parameters are not free, but rather satisfy the equations

$$\sum_k n_{+ijk} p^*_{hijk} = \mu_{hij}, \quad \sum_j N_{+ij} \pi_{hij} = \mu_{hi}, \tag{23}$$

which come from substituting the expressions from (18) and (20) into the definitions of π_{ij} and ψ_{ijk}.

We have reparameterized our profile log-likelihood in terms of the π's, which must satisfy the equations (23). Since $\pi_{+i} = \pi_{+ij} = 1$ there are $I + IJ$ free parameters. Next, we introduce a further parameterization corresponding to that used by Scott & Wild (1997) in the two-phase case. Put $\gamma_i = N_{+i}\pi_{1i} - N_{1i}$ and $\gamma_{ij} = N_{+ij}\pi_{1ij} - N_{1ij}$. Then $\mu_{1i} = n_{1i} + \gamma_i$ and $\mu_{1ij} = n_{1ij} + \gamma_{ij}$, and the profile log-likelihood can be written in terms of these new parameters as

$$\ell_P(\beta) = \sum_{hijk} n_{hijk} \log p^*_{hijk}(\beta) + \sum_i c_i(\alpha_i) + \sum_{hij} c_{ij}(\alpha_{ij})$$

where α_t is defined as a function of γ_t by (12) and c_t is given by (13). Here, as in Section 2.2, we use the subscript t to mean either i or the double subscript ij. In terms of α_t, p^*_{1ijk} can be written as

$$\text{logit}(p^*_{1ijk}) = \text{logit}(p_{1ijk}) + \alpha_i + \alpha_{ij}.$$

The conditions (23) can be written in terms of the γ_t's as

$$\sum_k n_{+ijk} p^*_{1ijk} = n_{1ij} + \gamma_{ij}, \quad \gamma_{i+} = \gamma_i. \tag{24}$$

Now consider the function ℓ^*_3 defined by (9), where the α_t's are regarded as free parameters. We have shown that

$$\ell_P(\beta) = \ell^*_3\{\beta, \alpha(\beta)\},$$

where the elements α_t of $\alpha(\beta)$ are given by (12), and the γ_t's satisfy equations (24).

Finally, we show that the equations (23) are implied by the derivative conditions $\partial\ell^*_3/\partial\alpha_t = 0$. We have $\partial\ell^*_3/\partial\alpha_{ij} = n_{1ij+} - \sum_k n_{+ijk} p^*_{1ijk} + \gamma_{ij}$, so that $\partial\ell^*_3/\partial\alpha_{ij} = 0$ implies the first part of (24). Similarly,

$$\frac{\partial\ell^*_3}{\partial\alpha_i} = n_{1i++} - \sum_{jk} n_{+ijk} p^*_{1ijk} + \gamma_i.$$

Thus, adding the equations $\partial\ell^*_3/\partial\alpha_{ij} = 0$ over j gives $n_{1i++} = \sum_{jk} n_{+ijk} p^*_{1ijk} + \gamma_{i+}$ so that $\partial\ell^*_3/\partial\alpha_i = 0$ and $\partial\ell^*_3/\partial\alpha_{ij} = 0$ imply equations (24). It follows that $\hat{\beta} = \text{argmax}\, \ell_P(\beta) \text{argmax}\, \ell^*_3(\beta, \alpha(\beta))$ is found by solving the estimating equation $\frac{\partial\ell^*_3}{\partial\phi} = 0$, where $\phi^T = (\beta^T, \alpha^T)$.

Appendix 2

Establishing the efficiency bound

In this Appendix we establish the asymptotic efficiency of the semiparametric maximum likelihood estimate obtained by solving the estimating equation considered above. We begin by calculating the asymptotic variance of $\widehat{\beta}$.

Asymptotic variance of the estimate

Let $N_T = N + n_{++} + n_{+++}$ and set $\quad I^* = -\text{plim}_{N_T \to \infty} {N_T}^{-1} \dfrac{\partial^2 \ell_3^*}{\partial\phi\partial\phi^T}$,

where all derivatives are evaluated at the true values $\phi^{(0)}$. The reason for this non-standard normalization will be come clear below. It can be shown (see Scott & Wild 2001, Lee, Scott & Wild 2006, 2007) that

$$\lim_{N_T \to \infty} N \text{cov}(\widehat{\beta}) = \{I^*_{\beta\beta} - I^*_{\beta\alpha}(I^*_{\alpha\alpha})^{-1} I^*_{\alpha\beta}\}^{-1}, \tag{25}$$

where I^* is partitioned in accordance with (α, β). This result follows from the fact that, under suitable regularity conditions, the solution $\widehat{\phi}$ of $\partial\ell^*\partial\phi = 0$ is asymptotically normal with asymptotic variance $I^{*-1}\Sigma I^{*-1}$ where the matrix Σ is of the form

$$\Sigma = I^* - I^* \begin{pmatrix} 0 & 0 \\ 0^T & D \end{pmatrix} I^*$$

for some matrix D. Thus, the asymptotic variance of $\widehat{\phi}$ is

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

and it follows from the partitioned matrix inverse formula that the asymptotic variance matrix of $\widehat{\beta}$ is given by (25).

We now derive an explicit expression for I^* under a different but equivalent three-phase sampling scheme. Suppose that

1. (Phase 1). We take a random sample of N individuals from the population of cases and controls. Then $\{N_{hi}\}$, the number out of N with $Y = h$ and $X^{(1)} = x_i^{(1)}$, have a multinomial distribution with probabilities $\Delta_{hi} = \text{pr}(Y = h, X^{(1)} = x_i^{(1)})$. This is the same as the original sampling scheme.

2. (Phase 2). For $h = 0, 1$ and $i_1 = 1, \ldots, I$, we take n_{hi} individuals sampled, independently of what happens at Phase 1, from the conditional distribution of $X^{(2)} = x_j^{(2)}$, given $Y = h$ and $X^{(1)} = x_i^{(1)}$. Let N_{hij} be the number of these having $X^{(2)} = x_j^{(2)}$. Then the $\{N_{hij}\}$ have a multinomial distribution with probabilities $\Delta_{hij} = \text{pr}(X^{(2)} = x_j^{(2)} \mid Y = h, X^{(1)} = x_i^{(1)})$.

3. (Phase 3). For $h = 0, 1$ and $i = 1, \ldots, I,\ j = 1, \ldots, J$, we take a sample of n_{hij} individuals, independently of what happens at Phase 2, from the conditional distribution of $X^{(3)}$, given $Y = h, X^{(1)} = x_i^{(1)}, X^{(2)} = x_j^{(2)}$, with density

$$p_{hij}(x, \beta) g_{ij}(x^{(3)}) / \psi_{hij},$$

where g_{ij} is the conditional density of $X^{(3)}$ given $X^{(1)} = x_i^{(1)}$, $X^{(2)} = x_j^{(2)}$, and $\psi_{hij} = \mathrm{pr}(Y = h \mid X^{(1)} = x_i^{(1)}, X^{(2)} = x_j^{(2)}) = \int p_{hij}(x, \beta) g_{ij}(x^{(3)})\, dx^{(3)}$. We denote conditional expectation with respect to $X^{(3)}$, given $Y = h, X^{(1)} = x_i^{(1)}, X^{(2)} = x_j^{(2)}$, by E_{hij}.

This sampling scheme has the same likelihood, and hence the same asymptotics, as the one considered previously. Thus, if an estimator is efficient under the new sampling scheme, it will be efficient under the old. For a proof of this for two-phase sampling, see Lee (2007). The general case for arbitrary S is essentially identical.

We work with this new scheme for the remainder of this section. We consider asymptotics where $N/N_T \to w$, $n_{hi}/N_T \to w_{hi}$ and $n_{hij}/N_T \to w_{hij}$. Corresponding to the fact that $N_{hi} \geq n_{hi}$ and $N_{hij} \geq n_{hij}$ in the original scheme, we will assume that $w\Delta_{hi}^{(0)} \geq w_{hi}$ and $w_{hi}\Delta_{hij}^{(0)} \geq w_{hij}$. Here, we are using the additional superscript 0 to denote the true value of the corresponding parameter. We also let $\pi_{hi}^{(0)}$ denote the true value of the conditional probability $\mathrm{pr}(Y = h \mid X^{(1)} = x_i^{(1)})$.

Under the new scheme, and applying the law of large numbers directly to ${N_T}^{-1} \frac{\partial^2 \ell_3^*}{\partial\phi\partial\phi^T}$, where ℓ_3^* is given by (equation10) , we obtain

$$\begin{aligned} I^* &= -\sum_{hij} w_{hij} E_{hij} \left\{ \frac{\partial^2 \log p_{hij}^*(X^{(3)}, \phi)}{\partial\phi\partial\phi^T} \right\} - \sum_i \frac{\partial^2 c_i}{\partial\phi\partial\phi^T} - \sum_{ij} \frac{\partial^2 c_{ij}}{\partial\phi\partial\phi^T} \\ &= -\sum_{hij} w_{hij} E_{hij} \left\{ \frac{\partial^2 \log p_{hij}^*(X^{(3)}, \phi)}{\partial\phi\partial\phi^T} \right\} - \begin{pmatrix} 0 & 0 \\ 0 & A \end{pmatrix} \end{aligned}$$

where A is a $(I + IJ) \times (I + IJ)$ diagonal matrix with entries A_t, since $\partial^2 c_t / \partial\alpha_t^2 \partial\gamma_t / \partial\alpha_t = A_t$. Using the identity

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

and noting that

$$E_{hij} \left\{ \frac{1}{p_{hij}^*} \frac{\partial^2 p_{hij}^*}{\partial\phi\partial\phi^T} \right\} = 0,$$

we finally obtain

$$I^* = \sum_{hij} w_{hij} E_{hij} \left\{ \frac{\partial \log p_{hij}^*(X^{(3)}, \phi)}{\partial\phi} \frac{\partial \log p_{hij}^*(X^{(3)}, \phi)}{\partial\phi^T} \right\} - \begin{pmatrix} 0 & 0 \\ 0 & A \end{pmatrix}. \qquad (26)$$

A general result

We first describe a general result that shows how an efficiency bound can be calculated. Suppose that we have independent observations z of J different types, with respective densities $f_j(z,\beta,\gamma)$ for $j=1,\ldots,J$ where β is a finite dimensional parameter and γ can be infinite-dimensional. Then if $\widehat{\beta}$ is a regular asymptotically linear semi-parametric estimate of β, the covariance matrix of $\widehat{\beta}$ must satisfy $\mathrm{var}(\widehat{\beta}) \geq B$ where B is the semi-parametric efficiency bound. The matrix B may be found as follows. Consider the "expected population log likelihood"

$$\sum_j w_j E_j\{\log f_j(z,\beta,\gamma)\}, \tag{27}$$

where E_j denotes expectation with respect to the true density $f_j(z,\beta^{(0)},\gamma^{(0)})$, and the weights are the limiting proportions of the different types of observations. For fixed β, let $\gamma(\beta)$ be the maximizer of (27). This is called the "least favourable distribution" for the problem. The "efficient scores" S_j^* are given by

$$S_j^* = \left.\frac{\partial \log f_j(z,\beta,\gamma(\beta))}{\partial\beta}\right|_{\beta=\beta^{(0)}}, \quad j=1,\ldots,J$$

and the efficiency bound is given by

$$B^{-1} = \sum_j w_j E_j[S_j^* S_j^{*T}].$$

Thus, to establish the efficiency of our procedure, we need only show that the asymptotic variance of our estimate coincides with B. This approach to semi-parametric efficiency is described in Tsiatis (2006) in the case of a single population, and extended to more than one population in Lee & Hirose (2009). The characterization of the least favourable distribution as the maximizer of an ”expected log-likelihood” was first considered by Newey (1994). Newey's formulation is more general than the one considered here, in that it makes no assumption that the parametric part of the model is correctly specified, as is the case in the Tsiatis formulation.

Application to three-phase sampling

Now we apply the theory sketched above to regression models for data obtained by the modified three-phase sampling scheme described in the previous section. The results obtained also apply to the original three-phase sampling scheme.

First, we parameterize the distributions in phases 1 to 3 of our sampling scheme in terms of the conditional densities $g_{ij}(x_k^{(3)})$ of $X^{(3)}$, given $X^{(1)}=x_i^{(1)}$ and $X^{(2)}=x_j^{(2)}$, the conditional probability $\xi_{ij} = \mathrm{pr}(X^{(2)}=x_j^{(2)}, \mid X^{(1)}=x_i^{(1)})$ and the unconditional probability $\zeta_i = \mathrm{pr}(X^{(1)}=x_i^{(1)})$. The first phase distribution is

$$\Delta_{hi} = \pi_{hi}\zeta_i.$$

For the second phase, the distributions are

$$\Delta_{hij} = \frac{\psi_{hij}\xi_{ij}}{\pi_{hi}},$$

where

$$\begin{aligned} \psi_{hij} &= \int p_{hij}(x,\beta) g_{ij}(x^{(3)})\, dx^{(3)}, \\ \pi_{hi} &= \sum_j \psi_{hij}\xi_{ij}. \end{aligned}$$

Finally, the third phase distributions have densities

$$p_{hij}(x,\beta) g_{ij}(x^{(3)})/\psi_{hij}.$$

The "expected log likelihood" is, up to a constant not involving ζ_i, ξ_{ij} or g_{ij}

$$\mathcal{E} = \sum_{hi} w\Delta^{(0)}_{hi} \log \Delta_{hi} + \sum_{hij} w_{hi}\Delta^{(0)}_{hij} \log \Delta_{hij} + \sum_{hij} w_{hij} E_{hij}[\log g_{ij}(X^{(3)})/\psi_{hij}]. \quad (28)$$

Finding the least favourable distribution

To find the least favourable distribution, we must maximize (28) over the g_{ij}'s and the ξ_{ij}'s for fixed β. Recall that $p_{hij}(x,\beta)$ depends on x only through $x^{(3)}$. As in A1, the parameter ζ_i can be ignored in what follows. From Appendix 3, the maximizing values satisfy the equations

$$\hat{g}_{ij}(x^{(3)},\beta) = \frac{P_{ij}(x^{(3)}) g^{(0)}_{ij}(x^{(3)})}{\sum_h \frac{\mu_{hij}(\beta)}{\psi_{hij}(\beta)} p_{hij}(x,\beta)}, \quad (29)$$

$$\hat{\xi}_{ij}(\beta) = \frac{w^{(ij)}}{\sum_h \mu_{hi}(\beta)\psi_{hij}(\beta)/\pi_{hi}(\beta)}, \quad (30)$$

where

$$P_{ij}(x^{(3)}) = \sum_h \frac{w_{hij}}{\psi^{(0)}_{hij}} p_{hij}(x,\beta^{(0)})$$

and $w^{(ij)} = \sum_h w_{hi}\Delta^{(0)}_{hij}$. The quantities $\mu_{hi}, \mu_{hij}, \pi_{hi}, \pi_{hij}$ are scaled limiting versions of those considered in Appendix 1, and are defined in terms of further quantities γ_t by

$$\mu_{hi} = w_{hi} + \delta_h\gamma_i, \quad \pi_{hi} = \frac{w\Delta^{(0)}_{hi} + \delta_h\gamma_i}{w\Delta^{(0)}_{+i}},$$

$$\mu_{hij} = w_{hij} + \delta_h\gamma_{ij}, \quad \pi_{hi} = \frac{w_{hi}\Delta^{(0)}_{hij} + \delta_h\gamma_{ij}}{w_{hi}\Delta^{(0)}_{+ij}},$$

where we have written $\delta_h = \pm 1$ according as $h = 1$ or 0. Also, as in A1, let

$$\text{logit} p^*_{1ij}(x^{(3)},\phi) = \text{logit} p_{1ij}(x^{(3)},\beta) + \alpha_i + \alpha_{ij},$$

where

$$\alpha_i = \log\left(\frac{w_{1i} + \gamma_i}{w\Delta^{(0)}_{1i} + \gamma_i}\right) - \log\left(\frac{w_{0i} - \gamma_i}{w\Delta^{(0)}_{0i} - \gamma_i}\right),$$

$$\alpha_{ij} = \log\left(\frac{w_{1ij} + \gamma_{ij}}{w_{1i}\Delta^{(0)}_{1ij} + \gamma_{ij}}\right) - \log\left(\frac{w_{0ij} - \gamma_{ij}}{w_{0i}\Delta^{(0)}_{0ij} - \gamma_{ij}}\right).$$

The quantities γ_t satisfy the equations

$$\gamma_{i+} = \gamma_i, \quad \int p^*_{1ij} P_{ij}(x^{(3)}) g^{(0)}_{ij}(x^{(3)})\, dx^{(3)} = w_{1ij} + \gamma_{ij}. \tag{31}$$

The proof of these assertions is almost the same as that given in A1. Note that the γ_t's, and hence the α_t's, are functions of β, although this is suppressed in the notation. When $\beta = \beta^{(0)}$, then $\gamma_t = 0$ is a solution of equations (31).

For the first distribution, the efficient scores are

$$\frac{\partial \log \Delta_{hi}}{\partial \beta}|_{\beta=\beta^{(0)}} = \delta_h \frac{1}{w\Delta^{(0)}_{hi}} \frac{\partial \gamma_i}{\partial \beta}.$$

For the second distribution, the efficient score is

$$\frac{\partial \log \Delta_{hij}}{\partial \beta}\ |_{\beta=\beta^{(0)}} = \delta_h \left(\frac{1}{w_{hi}\Delta^{(0)}_{hij}} \frac{\partial \gamma_{ij}}{\partial \beta} - \frac{1}{w_{hi}} \frac{\partial \gamma_i}{\partial \beta} \right)$$

and for the third

$$\frac{\partial}{\partial \beta} \log p_{hij}(x, \beta) g_{ij}(x^{(3)}) / \psi_{hij} \frac{\partial \log p^*_{hij}(x, \phi)}{\partial \beta} - \delta_h \frac{1}{w_{hij}} \frac{\partial \gamma_{ij}}{\partial \beta}.$$

Note also that

$$E_{hij} \left\{ \frac{\partial \log p^*_{hij}(X, \phi)}{\partial \beta} \right\} \delta_h \frac{1}{w_{hij}} \frac{\partial \gamma_{ij}}{\partial \beta}.$$

Thus, the inverse of the information bound B is, writing $x^{\otimes 2}$ for xx^T,

$$\begin{aligned} B^{-1} &= \sum_{hi} w\Delta^{(0)}_{hi} \left(\frac{1}{w\Delta^{(0)}_{1i}} \frac{\partial \gamma_i}{\partial \beta} \right)^{\otimes 2} + \sum_{hij} w_{hi}\Delta^{(0)}_{hij} \left(\frac{1}{w_{hi}\Delta^{(0)}_{hij}} \frac{\partial \gamma_{ij}}{\partial \beta} - \frac{1}{w_{hi}} \frac{\partial \gamma_i}{\partial \beta} \right)^{\otimes 2} \\ &\quad + \sum_{hij} w_{hij} E_{ij} \left\{ \left(\frac{\partial \log p^*_{hij}(X^{(3)}, \phi)}{\partial \beta} - \delta_h \frac{1}{w_{hij}} \frac{\partial \gamma_{ij}}{\partial \beta} \right)^{\otimes 2} \right\} \\ &= \sum_{hij} w_{hij} E_{ij} \left\{ \left(\frac{\partial \log p^*_{hij}(X^{(3)}, \phi)}{\partial \beta} \right)^{\otimes 2} \right\} - \sum_i A_i^{-1} \left(\frac{\partial \gamma_i}{\partial \beta} \right)^{\otimes 2} - \sum_{ij} A_{ij}^{-1} \left(\frac{\partial \gamma_{ij}}{\partial \beta} \right)^{\otimes 2} \end{aligned}$$

where

$$A_i = \left(\frac{1}{w\Delta^{(0)}_{1i}} - \frac{1}{w_{hi}} \right)^{-1}, \qquad A_{ij} = \left(\frac{1}{w_{hi}\Delta^{(0)}_{1ij}} - \frac{1}{w_{hij}} \right)^{-1}$$

and all derivatives are evaluated at $\beta = \beta^{(0)}$. Since $\partial \gamma_t / \partial \beta = (\partial \gamma_t / \partial \alpha_t) A_t$ we finally obtain

$$B^{-1} = \sum_{hij} w_{hij} E_{ij} \left[\left(\frac{\partial \log p^*_{hij}(X^{(3)}, \phi)}{\partial \beta} \right)^{\otimes 2} \right] - \sum_i A_i \left(\frac{\partial \alpha_i}{\partial \beta} \right)^{\otimes 2} - \sum_{ij} A_i \left(\frac{\partial \alpha_{ij}}{\partial \beta} \right)^{\otimes 2}.$$

Since $\phi^T = (\beta^T, \alpha(\beta)^T)$, the chain rule and (26) imply that

$$B^{-1} = I^*_{\beta\beta} + \left(\frac{\partial \alpha}{\partial \beta} \right)^T I^*_{\alpha\beta} + I^*_{\beta\alpha} \left(\frac{\partial \alpha}{\partial \beta} \right) + \left(\frac{\partial \alpha}{\partial \beta} \right)^T I^*_{\alpha\alpha} \left(\frac{\partial \alpha}{\partial \beta} \right). \tag{32}$$

To calculate $\frac{\partial\alpha}{\partial\beta}$, let $\mathcal{E}$ be the expected log-likelihood (28). Then, by the results in Appendix 3, for each fixed β we have

$$\mathcal{E}(\beta, \hat{\xi}, \hat{g}) = \mathcal{E}^*(\beta, \alpha(\beta))$$

where $\mathcal{E}^*$ is given by (35) in Appendix 3. Using the same arguments as those used in Appendix 1 for the sample case, $\alpha(\beta)$ satisfies

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

for each β. Differentiating again by the chain rule, we get

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

Thus, by (28), we get $\quad I^*_{\beta\alpha} + \left(\dfrac{\partial\alpha}{\partial\beta}\right)^T I^*_{\alpha\alpha} = 0$

so that $\partial\alpha/\partial\beta = -\left(I^*_{\alpha\alpha}\right)^{-1} I^*_{\alpha\beta}$. Substituting this into (32) shows that $B^{-1} = I^*_{\beta\beta} - I^*_{\beta\alpha}(I^*_{\alpha\alpha})^{-1}I^*_{\alpha\beta}$, and so by (25), the estimating equation leads to asymptotically efficient estimates.

Appendix 3

Maximizing the expected log-likelihood

Again, we begin by assuming that the support of the variable $X^{(3)}$ is finite. By following the same argument as in Appendix 1, the Lagrange multiplier argument suggests that the maximizing values of g_{ij} and ξ_{ij} satisfy (29) and (30). In fact, this remains true even when the support of $X^{(3)}$is not finite, at least for values of β in a neighbourhood of $\beta^{(0)}$. We now verify this.

For $\hat{g}_{ij}$ and $\hat{\xi}_{ij}$ as defined in (29) and (30), and for arbitrary densities g_{ij} and probabilities ξ_{ij}, we must show that

$$\begin{aligned} &\sum_{hij} w_{hi}\Delta^{(0)}_{hij} \log\Delta_{hij}(\beta) + \sum_{hij} \frac{w_{hij}}{\psi^{(0)}_{hij}} \int \log\hat{g}_{ij}(x^{(3)}) p_{hij}(x,\beta^{(0)}) g^{(0)}_{ij}(x^{(3)})\,dx^{(3)} - \sum_{hij} w_{hij}\log\psi_{hij}(\beta) \\ \geq\ &\sum_{hij} w_{hi_1}\Delta^{(0)}_{hij}\log\Delta^*_{hij} + \sum_{hij}\frac{w_{hij}}{\psi^{(0)}_{hij}}\int\log g_{ij}(x^{(3)})p_{hij}(x,\beta^{(0)})g^{(0)}_{ij}(x^{(3)})\,dx^{(3)} - \sum_{hij} w_{hij}\log{\psi_{hij}}^* \qquad (33)\end{aligned}$$

where ${\psi_{hij}}^* = \int g_{ij}(x^{(3)})p_{hij}(x,\beta)\,dx^{(3)}$ and $\Delta^*_{hij} = {\psi_{hij}}^*\xi_{ij}/\sum_j {\psi_{hij}}^*\xi_{ij}$. The inequality (33) is equivalent to

$$\sum_{hij} w_{hi}\Delta^{(0)}_{hij}\log\frac{\Delta_{hij}}{\Delta^*_{hij}} - \sum_{hij} w_{hij}\log\frac{\psi_{hij}}{{\psi_{hij}}^*} + \sum_{hij}\frac{w_{hij}}{\psi^{(0)}_{hij}}\int\log\frac{\hat{g_{ij}}(x^{(3)})}{g_{ij}(x^{(3)})}P_{ij}(x^{(3)})g^{(0)}_{ij}(x^{(3)})\,dx^{(3)} \geq 0.(34)$$

When $\beta = \beta^{(0)}$, $\hat{g}_{ij}(x^{(3)},\beta) = g^{(0)}(x^{(3)})$ and $\Delta_{ij} = \Delta^{(0)}_{ij}$, the left hand side of (34) becomes

$$\sum_{hij} w_{hi}\Delta^{(0)}_{hij}\log\frac{\Delta^{(0)}_{hij}}{\Delta^*_{hij}} - \sum_{hij} w_{hij}\log\frac{\psi^{(0)}_{hij}}{{\psi_{hij}}^*} + \sum_{hij}\frac{w_{hij}}{\psi^{(0)}_{hij}}\int\log\frac{g^{(0)}_{ij}(x^{(3)})}{g_{ij}(x^{(3)})}P_{ij}(x^{(3)})g^{(0)}_{ij}(x^{(3)})\,dx^{(3)}.$$

An argument based on the Kullback-Leibler information inequality shows that

$$\frac{w_{hij}}{\psi^{(0)}_{hij}} \int \log \frac{g^{(0)}_{ij}(x^{(3)})}{g_{ij}(x^{(3)})} P_{ij}(x^{(3)}) g^{(0)}_{ij}(x^{(3)})\, dx^{(3)} > w_{hij} \log \frac{\psi^{(0)}_{hij}}{{\psi_{hij}}^*},$$

provided $g_{ij} \neq g^{(0)}_{ij}$. Moreover, the Kullback-Leibler inequality implies that

$$\sum_{hij} w_i \Delta^{(0)}_{hij} \log \frac{\Delta^{(0)}_{hij}}{\Delta_{hij}} \geq 0.$$

Hence, the right-hand side of (34) is strictly positive at $\beta = \beta^{(0)}$, and by a continuity argument is non-negative for all β in some neighbourhood of $\beta^{(0)}$. Thus, (29) and (30) do indeed maximize the expected log-likelihood. Substituting these maximizing values back into (32), we get the analogue of (22) for the expected log-likelihood:

$$\mathcal{E}(\beta, \hat{\xi}, \hat{g}) = \mathcal{E}^*(\beta, \alpha(\beta))$$

where

$$\mathcal{E}^*(\beta, \alpha) = \sum_{hij} w_{hij} E_{hij}[\log p^*_{hij}(X^{(3)}, \phi)] + \sum_i c_i(\alpha_i) + \sum_{ij} c_{ij}(\alpha_{ij}). \qquad (35)$$

References

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

Arbogast, P.G., Lin, D.Y., Siscovick, D.S. and Schwartz, S.M. (2002) Estimating incidence rates from population-based case-control studies in the presence of nonrespondents. *Biometrical Journal*, **44**, 2:227–239.

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

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

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., 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.

Chatterjee, N. and Chen, Y.H. (2007). Maximum likelihood inference on a mixed conditionally and marginally specified regression model for genetic epidemiologic studies with two-phase sampling. *J. R. Statist. Soc.* B, **69**, 123–142.

Cochran, W.G. (1977). *Sampling Techniques* (3rd ed). New York: Wiley.

D’Angio, G. J., Breslow, N., Beckwith, J. B., et al. (1989). Treatment of Wilms tumor: Results of the third National Wilms Tumor Study. *Cancer*, **64**, 349–360.

Green, D. M., Breslow, N. E., Beckwith, J. B., et al. (1998). Comparison between single-dose and divided-dose administration of Dactinomycin and Doxorubicin for patients with Wilms’ tumor: A report from the National Wilms’ Tumor Study Group. *J. Clinical Oncology*, **16**, 237–245.

Kulich, M. and Lin, D.Y. (2004). Improving the efficiency of relative-risk estimation in case-cohort studies. *J. Am. Statistit. Assoc.*, **99**, 832–844.

Lawless, J.F., Kalbfleish, J.D 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. (2007). Semi-parametric efficiency bounds for regression models under choice-based sampling. *J. Appl. Math. & Decision Sci.*, vol. 2007, Article ID 86180, 23 pages. doi:10.1155/2007/86180.

Lee, A.J. and Hirose, Y. (2009). Semi-parametric efficiency bounds for regression models under response-selective sampling: the profile likelihood approach. *Ann. Inst. Stat. Math.*, to appear, DOI: 10.1007/s10463-008-0205-1.

Lee, A.J., Scott, A.J. and Wild, C.J. (2006). Fitting binary regression models with case-augmented samples. *Biometrika* **93**, 385–397.

Lee, A.J., Scott, A.J. and Wild, C.J. (2007). On the Breslow-Holubkov estimator. *Lifetime Data Analysis*, **13**, 545–563.

Lee, A.J., Scott, A.J. and Wild, C.J. (2009). Efficient estimation in multi-phase case-control studies. Unpublished manuscript, available at http://www.stat.auckland.ac.nz/∼lee/threephase.pdf

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

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

R Development Core Team. (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. URL http://www.R-project.org.

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. *J. Statist. Plan. Infer.*, **96**, 3–27.

Scott, A.J. and Wild, C.J. (2002). On the robustness of weighted methods for fitting model to case-control data by maximum likelihood. *J. R. Statist. Soc, B*, **64**, 207-220.

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

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.

Whittemore, A.S. and Halpern, (1997). Multi-stage sampling in genetic epidemiology. *Stat. Med.*, **16**, 153–167.