STATISTICS IN MEDICINE, VOL. 16, 103–116 (1997)

WEIGHTED LIKELIHOOD, PSEUDO-LIKELIHOOD AND MAXIMUM LIKELIHOOD METHODS FOR LOGISTIC REGRESSION ANALYSIS OF TWO-STAGE DATA

NORMAN E. BRESLOW AND RICHARD HOLUBKOV

Department of Biostatistics, University of Washington, Seattle, WA 98195-7232, U.S.A.

SUMMARY

General approaches to the fitting of binary response models to data collected in two-stage and other stratified sampling designs include weighted likelihood, pseudo-likelihood and full maximum likelihood. In previous work the authors developed the large sample theory and methodology for fitting of logistic regression models to two-stage case-control data using full maximum likelihood. The present paper describes computational algorithms that permit efficient estimation of regression coefficients using weighted, pseudo- and full maximum likelihood. It also presents results of a simulation study involving continuous covariables where maximum likelihood clearly outperformed the other two methods and discusses the analysis of data from three *bona fide* case-control studies that illustrate some important relationships among the three methods. A concluding section discusses the application of two-stage methods to case-control studies with validation subsampling for control of measurement error.

1. INTRODUCTION

Walker[1] and White[2] introduced two-stage sampling as a means of improving the efficiency of the case-control design in situations where collection of covariable information is limited by considerations of cost or feasibility. At the first stage of sampling, disease cases and controls are drawn at random from the study population and classified into strata on the basis of the exposure variable. Collection of additional covariables, needed for adjusted relative risk estimates, is restricted to a smaller number of cases and controls that are randomly selected from within each stratum during a second stage of sampling. Breslow and Cain[3] demonstrated that, by balancing the numbers of cases and controls sampled from each stratum at the second stage, substantial improvements in efficiency could be achieved in comparison with a standard case-control design. Methods of data analysis for two-stage studies may be useful in other situations where certain covariable data are missing by design or are simply missing at random[4] within defined strata of cases and controls.

Development of estimation strategies that efficiently combine information from the first and second stages of sampling has proved challenging to statistical research workers. Breslow and Cain[3] proposed a so-called CML estimate that was modelled on 'conditional maximum likelihood' estimation for choice-based studies in econometrics.[5,6] Since it involves unbiased estimating equations derived from a product of conditional probabilities, rather than from a conditional likelihood *per se*, this approach is more accurately described and will be referred to in the sequel as pseudo-likelihood or PL.[7] A refined PL estimate that attempts to make better use of the

CCC 0277–6715/97/010103–14

second stage data was proposed by Schill *et al.*[8] Flanders and Greenland[9] advocated a sample survey technique[10] that we shall term weighted likelihood or WL. In the econometrics literature it is known as weighted exogenous sampling maximum likelihood.[5] Working with a slightly different design that involves independent samples of cases and controls from each stratum at the first stage of sampling, Scott and Wild[11,12] and Wild[13] developed a full maximum likelihood (ML) approach that corresponds to the econometricians' full information concentrated likelihood.[14] Breslow and Holubkov[15] derived the same estimate by solving a constrained maximum likelihood problem for the two-stage case-control design, presented an efficient computational algorithm and developed a simple covariance formula. Robins *et al.*[16] also developed a fully efficient estimate in the course of a general approach to the regression analysis of incomplete data from the viewpoint of semi-parametric inference.

The present paper is intended as a practical introduction to these methodologies that compares their features by means of simulations and a series of illustrative examples. After introducing some notation and stating the basic model, we present each estimate in terms of the algorithm used for its computation. Readers are referred to the companion paper[15] and to the doctoral dissertation of Holubkov[17] for derivation of the algorithms and theoretical details. A small simulation study, part of extensive work[17] that will be reported elsewhere, demonstrates that there are substantial differences in the efficiencies of the three estimates with certain data configurations. The methods are applied to data from case-control studies of perinatal mortality,[18] of occupational lung cancer[19] and of smoking and lung cancer.[20] We conclude with a discussion of their potential applicability to measurement error problems with validation sampling at the second stage.[21]

2. THE ESTIMATION PROBLEM

Suppose that N_1 cases of disease and N_0 controls are sampled at random from the population at risk. Each subject is classified on the basis of one or more explanatory variables into one of J strata, resulting in counts N_{1j} of cases and N_{0j} of controls in stratum j. Any further data beyond those required for assignment of stratum at this first stage of sampling are discarded. Thus it may be desirable to maintain a reasonably fine degree of stratification within the constraints of the available sample size. Let $N = N_+$ denote the total sample size and note that $N_{i+} = N_i$ for $i = 0, 1$.

At the second stage, n_{ij} subjects are selected at random from among the N_{ij} in each of $2J$ cells and additional covariable information is collected for each of them. For notational convenience only, we assume that there is a finite number K of possible combinations of covariable values at this second stage and we denote by $X = k$ the event that the kth such combination is observed. Let n_{ijk} denote the frequency of second-stage subjects having disease status i, stratum j and covariable class k. There is no limitation on the size of K. The methods may be applied to continuous covariables collected at the second stage by having each subject occupy a unique covariable class.

Define D to be a binary disease indicator taking values $i = 1$ for cases and $i = 0$ for controls and let S be a random stratum indicator with values $j = 1, \ldots, J$. As a model for disease risk in the population we assume

$$p^*_{1jk} = 1 - p^*_{0jk} = \Pr(D = 1 | S = j, X = k) = (1 + \mathrm{e}^{-x^{\mathrm{t}}_{jk}\beta})^{-1} \tag{1}$$

where x_{jk} is a p-vector of covariables and x^{t} denotes the transpose of x. This choice of the linear logistic model is predicated on the knowledge that, except for a constant term that we shall assume is always present, the regression coefficients β are interpretable as log odds ratios (log relative risks) and are estimable from case-control studies.[22] The marginal distributions $\Pr(D=i)$ and $\Pr(X=k, S=j)$ are left unspecified.

Following Prentice and Pyke[23] and Schill *et al.*,[8] define stratum and covariable probabilities conditional on being sampled at the first or second stages by

$$Q_j = \sum_{i=0}^{1} \frac{N_i}{N} \Pr(S=j \mid D=i) \tag{2}$$

$$q_{jk} = \sum_{i=0}^{1} \frac{n_{ij}}{n_{+j}} \Pr(X=k \mid D=i, S=j) \tag{3}$$

and similarly conditioned disease probabilities given stratum or covariables by

$$P_{1j} = 1 - P_{0j} = \frac{N_1 \mathrm{e}^{\delta_j}}{N_0 + N_1 \mathrm{e}^{\delta_j}} \tag{4}$$

$$p_{1jk} = 1 - p_{0jk} = \frac{n_{1j}\mathrm{e}^{-\alpha - \delta_j + x_{jk}^{\mathrm{t}}\beta}}{n_{0j} + n_{1j}\mathrm{e}^{-\alpha - \delta_j - x_{jk}^{\mathrm{t}}\beta}} \tag{5}$$

where $\alpha = \log\{\Pr(D=1)/\Pr(D=0)\}$ is the marginal log odds of disease and $\delta_j = \log\{\Pr(S=j \mid D=1)/\Pr(S=j \mid D=0)\}$. The parameter α is aliased with the constant term in (1), which is uninterpretable unless the disease risk in the population is known. Differences in the δ_j represent log odds ratios of disease probabilities.

According to Scott and Wild[12] and Schill *et al.*,[8] the likelihood of the data $\{N_{ij}\}$, $\{n_{ijk}\}$ is proportional to

$$\left\{\prod_{i=0}^{1} \prod_{j=1}^{J} P_{ij}^{N_{ij}} \prod_{k=1}^{K} p_{ijk}^{n_{ijk}}\right\} \left\{\prod_{j=1}^{J} Q_j^{N_{+j}}\right\} \left\{\prod_{j=1}^{J} \prod_{k=1}^{K} q_{jk}^{n_{+jk}}\right\}. \tag{6}$$

The p β's, J δ's, J Q's and JK q's that occur in this expression are, however, constrained by the equations $N_i = N \sum_{j=1}^{J} P_{ij} Q_j$ for $i = 0, 1$ and $n_{ij} = n_{+j} \sum_{k=1}^{K} p_{ijk} q_{jk}$ for $i = 0, 1, j = 1, \ldots, J$, that is by the requirement that the probabilities of being a case or control at stage one, or of being a case or control in stratum j at stage two, are fixed by design. The only parameters of real interest are the regression coefficients β and perhaps also, on occasion, $\delta = (\delta_1, \ldots, \delta_J)^{\mathrm{t}}$.

3. THE THREE ESTIMATES

The WL estimate is obtained by fitting the logistic regression model (1) to the second-stage data using the inverses f_{ij}^{-1} of the sampling fractions $f_{ij} = n_{ij}/N_{ij}$ as *prior weights*[24] for the observations in the (i, j) cell. Since cases and controls are weighted differently, this means the model is fitted with $2JK$ separate data records consisting of binary indicators d_{ijk} of case-control status ($d_{1jk} = 1$, $d_{0jk} = 0$), vectors x_{jk} of covariables (for $i = 0$ or 1) and replicates n_{ijk}. Let X denote the corresponding design matrix of dimension $2JK \times p$ and let V^* be the $2JK \times 2JK$ diagonal matrix with diagonal elements $f_{ij}^{-1} n_{ijk} p_{0jk}^{*} p_{1jk}^{*}$. Then $X^{\mathrm{t}} V^* X$ is the information matrix that arises from the weighted logistic regression problem. The covariance matrix for the WL estimate is $(X^{\mathrm{t}} V^* X)^{-1} G^* (X^{\mathrm{t}} V^* X)^{-1}$ where G^*, given by

$$\sum_{ij} f_{ij}^{-2} \left\{ \sum_k n_{ijk} u_{ijk}^{*\oplus 2} - \frac{(1-f_{ij})}{n_{ij}} \left(\sum_k n_{ijk} u_{ijk}^{*} \right)^{\oplus 2} \right\} - \sum_i \frac{1}{N_i} \left(\sum_j f_{ij}^{-1} \sum_k n_{ijk} u_{ijk}^{*} \right)^{\oplus 2} \tag{7}$$

is calculated from the logistic regression scores $u_{ijk}^{*} = (d_{ijk} - p_{1jk}^{*}) x_{jk}$ and $u^{\oplus 2}$ for any vector u denotes the matrix uu^{t}. This is the 'information sandwich' covariance matrix that arises from the theory of M-estimation;[25] the matrix of negative partial derivatives of the estimating equation (score) vector with respect to the parameters β is $X^{t}V^{*}X$, G^{*} accounts both for the observed within cell variability of the scores and for the variability in the first-stage data used as prior weights.

The PL estimates ignore the constraints on Q and q and utilize only the pseudo-likelihood given by the first factor in (6), namely

$$L_1L_2 = \left\{\prod_{i=0}^{1}\prod_{j=1}^{J} P_{ij}^{N_{ij}}\right\}\left\{\prod_{i=0}^{1}\prod_{j=1}^{J}\prod_{k=1}^{K} p_{ijk}^{n_{ijk}}\right\}. \tag{8}$$

Breslow and Cain[3] maximized L_1 in δ to find $\hat{\delta}_j = \log\{(N_{1j}N_0)/(N_{0j}N_1)\}$, which quantities were inserted into L_2 before maximizing it as a function of β alone. In practice their estimate is obtained by fitting the logistic regression model (1) to the second-stage data with the term $\log\{(n_{1j}N_{0j}N_1)/(n_{0j}N_{1j}N_0)\}$ added as an *offset*[24] to the covariable vector for observations in stratum *j*. The scores obtained by differentiating $\log L_2$ are $u_{ijk} = (d_{ijk} - p_{1jk})\,x_{jk}$ and the second derivative is $-X^{t}VX$ where V is diagonal with diagonal elements $n_{ijk}p_{0jk}p_{1jk}$. With $W_j = \sum_k n_{+jk}p_{0jk}p_{1jk}x_{jk}$ and $W_+ = \sum_j W_j$, the estimated asymptotic covariance matrix may be written $(X^{t}VX)^{-1}\,\{X^{t}VX - C^{*}\}(X^{t}VX)^{-1}$ where $C^{*} = \sum_{i,j}\{(n_{ij}^{-1} - N_{ij}^{-1})\,W_j^{\oplus 2} + N_i^{-1}W_+^{\oplus 2}\}$. An empirical version of the covariance matrix, which uses the observed within-stratum variability of the scores, is obtained by replacing the middle term $X^{t}VX - C^{*}$ by $G_1^{e} + G_2$ where $G_1^{e} = \sum_{i,j}\{\sum_k n_{ijk}u_{ijk}^{\oplus 2} - n_{ij}^{-1}\,(\sum_k n_{ijk}u_{ijk})^{\oplus 2}\}$ and $G_2 = \sum_i(\sum_j N_{ij}^{-1}W_j^{\oplus 2} - N_i^{-1}W_+^{\oplus 2})$.

The PL estimate of Schill *et al.*[8] maximixes L_1L_2 jointly as a function of the $p + J$ parameters $\gamma = (\beta^{t}, \delta^{t})^{t}$. It also may be obtained by fitting a logistic regression model with offsets jointly to the first- and second-stage data, as follows. Define *augmented covariable vectors* $\tilde{x}_{jk} = (x_{jk}^{t}, -e_j^{t})^{t}$ of length $p + J$ where e_j is the J-vector with a 1 in the *j*th location and 0's elsewhere. Similarly set $\tilde{x}_{j0} = (0, e_j^{t})^{t}$. The augmented design matrix $\tilde{X}$ of dimension $(2J + 2JK)\times(p + J)$ contains the vectors $\tilde{x}_{j0}^{t}$ in its first $2J$ rows. The remaining rows contain the vectors $\tilde{x}_{jk}^{t}$ in locations corresponding to stratum *j* and covariable class *k*. Construct an augmented 'outcome' vector $\tilde{d}$ of length $(2J + 2JK)$ that contains ones or zeros in locations corresponding to cases or controls, respectively, and place the data $\{N_{ij}\}$ and $\{n_{ijk}\}$ in a $2J + 2JK$ dimensional vector of replicates. Let the augmented vector of offsets contain $\log(N_1/N_0)$ in the first $2J$ locations and $\log(n_{1j}/n_{0j})$ in those of the remaining $2JK$ locations that correspond to stratum *j*. The solution to this logistic regression problem is the Schill *et al.* version of the PL estimate. See the last example in Section 5 for further detail. Its covariance matrix may be estimated by $(\tilde{X}^{t}\tilde{V}\tilde{X})^{-1}\{\tilde{X}^{t}\tilde{V}\tilde{X} - \tilde{C}\}(\tilde{X}^{t}\tilde{V}\tilde{X})^{-1}$ where now the correction matrix is $\tilde{C} = (N_0^{-1} + N_1^{-1})\tilde{W}_j^{\oplus 2} + \sum_j(n_{0j}^{-1} + n_{1j}^{-1})\tilde{W}_j^{\oplus 2}$ with $\tilde{W}_0 = \sum_j N_{+j}P_{0j}P_{1j}\tilde{x}_{j0}$ and $\tilde{W}_j = \sum_k n_{+jk}p_{0jk}p_{1jk}\tilde{x}_{jk}$. For an empirical version, simply replace the middle term $\tilde{X}^{t}\tilde{V}\tilde{X} - \tilde{C}$ by the sum of the within-stratum covariance matrices of the score contributions from stage one and two, that is, by

$$\sum_{i=0}^{1}\left\{\sum_{j=1}^{J} N_{ij}\tilde{u}_{ij0}^{\oplus 2} - \frac{1}{N_i}\left(\sum_{j=1}^{J} N_{ij}\tilde{u}_{ij0}\right)^{\oplus 2}\right\} + \sum_{i=0}^{1}\sum_{j=1}^{J}\left\{\sum_{k=1}^{K} n_{ijk}\tilde{u}_j^{\oplus 2} - \frac{1}{n_{ij}}\left(\sum_{k=1}^{K} n_{ijk}\tilde{u}_{ijk}\right)^{\oplus 2}\right\}$$

where $\tilde{u}_{ij0} = (\tilde{d}_{ij0} - P_{1j})\tilde{x}_{j0}$ and $\tilde{u}_{ijk} = (\tilde{d}_{ijk} - p_{1jk})\tilde{x}_{jk}$.

By concentrating the Lagrangian that arises from the constrained maximum likelihood problem, Breslow and Holubkov[15] showed that ML estimates of γ may be obtained from the estimating equations

$$U(\gamma) = \sum_{j=1}^{J}\left[T_j\tilde{x}_{j0} + \sum_{k=1}^{K}\left\{n_{1jk} - \frac{(n_{1j} - T_j)n_{0j}n_{+jk}p_{1jk}}{n_{0j}n_{1j} - T_j(n_{0j} - n_{+j}p_{0jk})}\right\}\tilde{x}_{jk}\right] = 0 \tag{9}$$

where $T_j = N_{1j} - N_{+j}P_{1j}$. Solution of these equations by standard algorithms is eased by calculation of the gradient

$$\partial U/\partial\gamma^{\mathrm{t}} = -\sum_{j=1}^{J} N_{+j}P_{0j}P_{1j}\tilde{x}_{j0}^{\oplus 2} - \sum_{j=1}^{J}\sum_{k=1}^{K} \frac{n_{0j}n_{1j}n_{+jk}p_{0jk}p_{1jk}n_{+j}N_{+j}P_{0j}P_{1j}}{\{n_{0j}n_{1j} - T_j(n_{0j} - n_{+j}p_{0jk})\}^2}\,\tilde{x}_{jk}\tilde{x}_{j0}^{\mathrm{t}}$$

$$- \sum_{j=1}^{J}\sum_{k=1}^{K} \frac{n_{0j}n_{1j}n_{+jk}p_{0jk}p_{1jk}(n_{1j} - T_j)(n_{0j} + T_j)}{\{n_{0j}n_{1j} - T_j(n_{0j} - n_{+j}p_{0jk})\}^2}\,\tilde{x}_{jk}^{\oplus 2}. \qquad (10)$$

ML estimates of the nuisance parameters are $\hat{Q}_j = N_{+j}/N$ and

$$\hat{q}_{jk} = \left(\frac{n_{+jk}}{n_{+j}}\right)\frac{n_{0j}n_{1j}}{n_{0j}n_{1j} + T_j(n_{1j} - n_{+j}p_{1jk})}. \qquad (11)$$

Breslow and Holubkov also showed by means of a linearization argument that an estimate of γ that is asymptotically equivalent to ML may be obtained by iterative use of standard programs for logistic regression. Starting with the Schill *et al.* PL estimate $\hat{\gamma}$ one calculates fitted values $\hat{P}_{ij}$ and $\hat{p}_{ijk}$ and then $\hat{q}_{jk}$ from (11). These are substituted into the equation

$$\eta_j = \frac{n_{+j}^2}{n_{0j}n_{1j}}\sum_{k=1}^{K} \hat{q}_{jk}\hat{p}_{0jk}\hat{p}_{1jk}\tilde{x}_{jk} \qquad (12)$$

to yield $p + J$ dimensional vectors η_j of *constructed variables*. In the next iteration, the covariable vector for the *j*th stratum of first-stage data in the augmented design matrix is taken to be $\tilde{x}_{j0} + \eta_j$ while the offsets for the *j*th stratum of stage-one data are changed to $\log(N_1/N_0)$-$\eta_j^{\mathrm{t}}\,\hat{\gamma}$. A new estimate $\hat{\gamma}$ is calculated and the procedure is iterated to convergence.

The covariance matrix for the ML estimate or its asymptotic equivalent is estimated by

$$(\tilde{X}^{\mathrm{t}}\tilde{V}\tilde{X} + C_1)^{-1}(\tilde{X}^{\mathrm{t}}\tilde{V}\tilde{X} + C_2)(\tilde{X}^{\mathrm{t}}\tilde{V}\tilde{X} + C_1^{\mathrm{t}})^{-1} \qquad (13)$$

where

$$C_1 = \sum_{j=1}^{J} N_{+j}P_{0j}P_{1j}\eta_j\tilde{x}_{j0}^{\mathrm{t}}$$

and

$$C_2 = C_1 - \frac{N}{N_0N_1}\left(\sum_{j=1}^{J} N_{+j}P_{0j}P_{1j}\eta_j\right)\left(\sum_{j=1}^{J} N_{+j}P_{0j}P_{1j}\tilde{x}_{j0}\right)^{\mathrm{t}}$$

$$+ C_1^{\mathrm{t}} - \frac{N}{N_0N_1}\left(\sum_{j=1}^{J} N_{+j}P_{0j}P_{1j}\tilde{x}_{j0}\right)\left(\sum_{j=1}^{J} N_{+j}P_{0j}P_{1j}\eta_j\right)^{\mathrm{t}}$$

$$+ \sum_{j=1}^{J} N_{+j}P_{0j}P_{1j}\eta_j^{\oplus 2} - \frac{N}{N_0N_1}\left(\sum_{j=1}^{J} N_{+j}P_{0j}P_{1j}\eta_j\right)^{\oplus 2}.$$

Note that the augmented design matrix $\tilde{X}$ as originally defined, before addition of the η_j terms, is used in this calculation. The linearized version of the estimating equations (9) also may be used to construct an empirical version of the covariance matrix analogous to that already derived for the PL estimate. The middle term in (13) is replaced by the sum of $2(J + 1)$ empirical covariance matrices calculated from the contributions to the final series of estimating equations, one such covariance matrix for each of the sampling strata used at the first- and second-stages of data collection.

Table I. Population frequencies and covariable means for the simulation study

	Controls			Cases		
	$X_1 = -1$	$X_1 = 0$	$X_1 = 1$	$X_1 = -1$	$X_1 = 0$	$X_1 = 1$
Frequency (X_1)	0·3333	0·3333	0·3333	0·1793	0·3796	0·4411
$\mathrm{E}(X_2 \mid X_1)$	0·0	2·0	2·0	0·3	2·3	2·3

4. A SIMULATION STUDY

We consider a simple problem based on the Olkin and Tate[26] mixed discrete and continuous correlation model. This involves three strata, a single modelled stratum variable and a single normally distributed covariable that has a constant variance of one and a mean depending upon stratum and case-control status. Table I shows the frequency distributions of the stratum variable X_1 and the conditional means of the covariable distributions $X_2 | X_1$ for cases and controls. The conditional probabilities of disease given stratum and covariable satisfy a linear logistic model with regression coefficients of 0·15 for X_1 and 0·3 for X_2. The correlation between X_1 and X_2 in the control population is 0·686.

At stage one, 1000 (X_1, X_2) observations on cases and 1000 such observations on controls are drawn from their respective probability distributions as specified in Table I. After classification into stratum, six subsamples of constant size $n_{ij} = 20$ are drawn and the X_2 values are recorded. X_2 values for the remaining subjects are discarded. The asymptotic standard errors of the (X_1, X_2) regression coefficients for this set-up are (0·158, 0·157) for WL, (0·167, 0·154) for Breslow–Cain PL, (0·168, 0·155) for Schill *et al.* PL and (0·112, 0·102) for ML. The corresponding efficiencies, ratios of asymptotic variances relative to ML, are (50 per cent, 42 per cent) for WL and (45 per cent, 43–44 per cent) for the two PL estimates. Ten thousand replications of data from this design were generated in the simulation study, a new set of 2000 (X_1, X_2) observations being generated for each replication.

Table II reports the means and standard deviations of the regression coefficients estimated by each method. Also shown are the means and standard deviations of the estimated standard errors, for comparison with the asymptotic values given above, and the proportion of samples for which the true value of the regression coefficient fell below or above the endpoints of the 90 per cent confidence interval based on the estimated standard error. Some bias was evident in the regression coefficients estimated by all four procedures and some asymmetry was evident in the tails of the ML distributions. The actual standard errors were slightly larger than the means of their estimated values, this being most noticeable for WL. None the less, the finite sample efficiencies were essentially equal to their asymptotic counterparts. In other simulations we found that the biases in the estimated coefficients and standard errors, the asymmetries in the tails of the ML distributions and the improved efficiency of ML over WL and PL were all accentuated with smaller stage two sample sizes. The biases decreased with larger stage-two sample sizes and became negligible when no subsampling was required (Table II). Simulations involving discrete data indicated that the asymptotic theory was a good guide to small sample performance even when the n_{ij} varied in size.[17]

A second study with 10,000 cases and 10,000 controls at stage one was simulated for which the regression coefficients were $-0{\cdot}15$ and $-0{\cdot}30$ and the control means $-2{\cdot}5$, $-2{\cdot}5$ and 1·5 in the three strata, giving a correlation of 0.884 between X_1 and X_2 for controls. Otherwise, the set-up was identical to that just described. This had asymptotic efficiencies of (17 per cent, 9 per cent) for

Table II. Results of the simulation study

Variable	Regression Coefficient		Standard Error		Outside CI	
	Mean	SD	Mean	SD	Low	High
Weighted likelihood						
X_1	0·1402	0·1650	0·1610	0·0230	0·044	0·057
X_2	0·3112	0·1640	0·1583	0·0154	0·053	0·055
Breslow–Cain pseudo-likelihood						
X_1	0·1398	0·1736	0·1694	0·0220	0·040	0·056
X_2	0·3107	0·1603	0·1570	0·0138	0·051	0·047
Schill *et al.* pseudo-likelihood						
X_1	0·1397	0·1744	0·1722	0·0217	0·041	0·056
X_2	0·3108	0·1618	0·1585	0·0136	0·051	0·048
Maximum likelihood						
X_1	0·1406	0·1176	0·1157	0·0203	0·034	0·061
X_2	0·3096	0·1017	0·0990	0·0183	0·061	0·036
No subsampling ($n_{ij} \equiv N_{ij}$)						
X_1	0·1515	0·0691	0·0695	0·0008	0·045	0·059
X_2	0·2994	0·0419	0·0420	0·0008	0·047	0·052

WL and (13 per cent, 11 per cent) for both PL estimates of the (X_1, X_2) coefficients. The biases were more pronounced for WL and PL, and their standard errors had greater variability, than those estimated by ML. These results, although not typical of actual practice, demonstrate that substantial efficiency loss may accompany the failure to use the full ML estimate. Several groups report similar results for another atypical design involving a single binary covariable in place of X_2.[11,15,16]

5. ILLUSTRATIVE ANALYSES

5.1. The Leicestershire Perinatal Mortality Study

Table III shows the distributions of total births, sampled controls and perinatal deaths ascertained in Leicestershire during the years 1978–1987. The study design is described by Clarke and Clayton[27] and an analysis of data from 1976–1981 was given by Clayton.[18] Controls were initially pair matched to cases on the basis of exact time and intended place of delivery. For purposes of the present and earlier analyses, however, the exact pair matching was ignored and only the degree of stratification shown in Table III was retained. Thus the data are here regarded as having arisen from a two-stage study with complete enumeration of cases and controls at the first stage and frequency matched sampling of controls within each of 20 strata at the second stage. Perinatal deaths occurring in one or more twins or other multiple births were excluded from analysis since they are so different biologically. However, the corresponding controls were retained and for this reason the controls slightly outnumber the cases in almost every stratum. Control sampling fractions vary more than tenfold, from 0·21 per cent of births intended for delivery in general practioner units (GPU) during 1980–1981 to 2·27 per cent of births intended for the Royal Infirmary during 1978–1979. Retention of the first-stage data thus is important in order to be able to estimate secular trends in perinatal mortality and to account for the fact that several risk factors were correlated with the stratification factors. The matching had been

Table III. Numbers of births, sampled controls and perinatal deaths in Leicestershire

Intended place	Year of birth				
of delivery*	1978–79	1980–81	1982–83	1984–85	1985–87
Other consultant units (OCU)	2968	3162	2882	2766	2678
	37	29	33	23	30
	36	23	25	22	25
Leicester Royal Infirmary (LRI)	10507	10155	10421	11091	11754
	239	145	134	103	129
	221	131	119	99	120
Leicester General Hospital (LGH)	4381	5699	6018	6387	6444
	67	75	65	73	59
	65	67	55	65	50
General practitioner units (GPU)	4133	4321	3128	2868	2599
	15	9	16	9	8
	14	9	15	10	8

* 1375 births intended for delivery at home were excluded

Table IV. Log odds ratios (β) of perinatal mortality

Maternal	Per cent positive		Log odds ratios ($\hat{\beta} \pm$ SE)	
risk factor	Cases	Controls	Crude	Adjusted
Asian origin	19·8	13·9	0·419 ± 0·109	0·523 ± 0·104
First birth	50·8	45·1	0·230 ± 0·081	0·300 ± 0·081
Social class V	12·1	8·7	0·378 ± 0·133	0·307 ± 0·132
Age 35 + yrs	8·5	6·2	0·344 ± 0·156	0·391 ± 0·155
Smoker	35·5	32·4	0·138 ± 0·085	0·179 ± 0·084
GPU → OTH*	16·7	11·8	0·406 ± 0·116	0·677 ± 0·113
GPU → GPU†	4·6	4·4	0·044 ± 0·194	− 1·378 ± 0·139

* Intended delivery general practitioner, actual delivery elsewhere

† Intended and actual delivery at general practitioner unit

undertaken primarily as a matter of administrative convenience, rather than in an attempt to control for the effects of known risk factors, and an analysis was desired that would reflect the results obtained had the controls been selected completely at random from the source population.

Seven binary covariables were selected for the present analysis, which is illustrative only and not intended as a comprehensive treatment of these data. Their distributions among sampled cases and controls, with corresponding log odds ratios, are shown in Table IV. Also shown are the log relative risks and standard errors estimated by ML from the first- and second-stage data using a model containing only the constant term and a single binary covariable. The most dramatic difference between the crude odds ratios and those adjusted for the variable control sampling fractions are for the final risk factor, which equals one only if the delivery had been intended for a GPU and actually took place there. Since deliveries intended for GPUs were grossly underrepresented in the control sample, the substantially lower risk of perinatal death associated with them was obscured, and the associated standard error was overestimated, when only the second-stage data were considered. Other coefficients that had notable changes in value are those for Asian origin and a change from GPU to another location for actual delivery. The former results

Table V. A two-stage analysis with covariables depending only on stratum

Maternal	Log odds ratio ($\hat{\beta} \pm$ SE)				
risk factor	LR*	WL	PL: B&C	PL: Schill	ML
Constant	− 4·730 ± 0·088	− 0·169 ± 0·084	− 0·146 ± 0·084	− 0·147 ± 0·083	− 0·169 ± 0·083
Period	− 0·161 ± 0·021	− 0·161 ± 0·022	− 0·161 ± 0·020	− 0·161 ± 0·021	− 0·161 ± 0·021
LRI†	0·369 ± 0·096	0·369 ± 0·096	0·361 ± 0·096	0·361 ± 0·096	0·369 ± 0·096
LGH†	0·181 ± 0·105	0·181 ± 0·106	0·164 ± 0·106	0·165 ± 0·106	0·181 ± 0·106
GPU†	− 1·052 ± 0·160	− 1·052 ± 0·160	− 0·999 ± 0·164	− 1·007 ± 0·162	− 1·052 ± 0·160

* LR = ordinary logistic regression analysis of stage one data only

† Intended place of delivery; OCU as baseline; see Table III for key

Table VI. Multiple logistic regression analyses of the Leicestershire data

Maternal	Log odds ratio ($\hat{\beta} \pm$ SE)			
risk factor	WL	PL: B&C	PL: Schill	ML
Constant	− 0·350 ± 0·065	− 0·353 ± 0·066	− 0·351 ± 0·066	− 0·377 ± 0·065
Period	− 0·154 ± 0·024	− 0·165 ± 0·021	− 0·164 ± 0·022	− 0·161 ± 0·022
Asian origin	0·578 ± 0·116	0·597 ± 0·115	0·608 ± 0·116	0·582 ± 0·114
First birth	0·272 ± 0·086	0·273 ± 0·084	0·269 ± 0·084	0·285 ± 0·083
Social class V	0·421 ± 0·139	0·406 ± 0·136	0·398 ± 0·136	0·432 ± 0·135
Age 35 + yr	0·564 ± 0·166	0·576 ± 0·161	0·562 ± 0·161	0·613 ± 0·160
Smoker	0·198 ± 0·080	0·222 ± 0·086	0·221 ± 0·086	0·228 ± 0·085
GPU → OTH	0·471 ± 0·122	0·496 ± 0·120	0·498 ± 0·120	0·502 ± 0·119
GPU → GPU	− 1·177 ± 0·148	− 1·110 ± 0·148	− 1·125 ± 0·148	− 1·158 ± 0·145

from the fact that the General Hospital, which is located in an area with a large Asian population that generally seeks care there, also draws high risk pregnancies from other districts having fewer Asians.[18] The hospital specific odds ratio thus underestimated the population value.

Two multiple logistic regression analyses were conducted. The first involved covariables that depended only on stratum and that were thus known for all subjects sampled at stage one: period of birth, coded -2 for 1978–79, …, 2 for 1986–87; and intended place of delivery, with other consultant units serving as baseline. Results are shown in Table V. Note that the usual two-stage models were fitted with $n_{ij} < N_{ij}$ even though the covariable values depended only on stratum. Apart from the constant term, the ML results agreed with those obtained by fitting a logistic regression model to the stage one data alone. The WL regression coefficients also agreed with those from the simple logistic analysis while the WL (empirical) standard errors agreed with the ML empirical standard errors. The two PL procedures, by contrast, failed to exactly replicate the ‘correct’ logistic analysis.

The second multiple logistic regression analysis included all the binary risk variables shown in Table IV plus calendar period coded as above. Regression coefficients and standard errors estimated by each of the four methods are shown in Table VI. No simple logistic regression analysis results are available for comparison here since several of the covariables were measured only for subjects sampled at stage two. From a practical viewpoint, results obtained with all four methods tell much the same story for these data.

Table VII. Four analyses of occupational lung cancer risk

Risk variable	Log relative risk ($\hat{\beta} \pm$ SE) WL	PL : B&C	PL : Schill	ML
Constant	– 2·352 ± 0·665	– 2·317 ± 0·608	– 2·345 ± 0·610	– 2·301 ± 0·596
Age 60–69 yrs	– 0·130 ± 0·251	– 0·137 ± 0·253	– 0·137 ± 0·253	– 0·131 ± 0·255
Age 70 +	0·014 ± 0·310	– 0·008 ± 0·299	0·007 ± 0·298	0·024 ± 0·296
Mild exposure	0·746 ± 0·302	0·783 ± 0·307	0·771 ± 0·304	0·736 ± 0·305
Heavy exposure	0·849 ± 0·306	0·819 ± 0·304	0·825 ± 0·304	0·843 ± 0·303
1–9 pack-yrs	1·341 ± 0·715	1·398 ± 0·717	1·417 ± 0·719	1·359 ± 0·707
10–19	2·216 ± 0·704	2·187 ± 0·689	2·208 ± 0·690	2·158 ± 0·678
20–39	2·321 ± 0·681	2·310 ± 0·668	2·345 ± 0·669	2·264 ± 0·656
40+	2·760 ± 0·655	2·732 ± 0·646	2·755 ± 0·647	2·698 ± 0·634

5.2. Case-control study of occupational lung cancer

Schill *et al.*[8] analysed data on 146 male lung cancer cases and 315 sex and age matched controls from a pilot study of occupational lung cancer risk carried out in Northern Germany.[19] Data on occupational exposure in three categories, age in three groups and tobacco consumption in five levels of pack-years were collected for all 461 subjects. In order to simulate a two-stage design, they ignored the tobacco data except for subjects in a subsample that comprised about half the controls and three-fourths of the cases. Frequencies of subjects classified by exposure, age and disease status at the first and second stages of sampling are shown in their Table II.

We reanalysed these data using all four methods. Except for the constant and age terms for the Breslow–Cain PL estimate, for which the original values were in error, the results shown in Table VII for the two PL estimates agree with the original analysis to at least three and in most cases four decimal places. We have parameterized the coefficients for the three age groups into a constant plus two age effects, rather than three separate constants as in the original analysis. As noted by Schill *et al.*, the fact that the first-stage sample was stratified on age renders the constant and age coefficients uninterpretable but has no effect on the validity of the other parameters. Once again there is little to choose among the four sets of estimates and standard errors. The ML standard errors for the covariable (tobacco) effects are slightly smaller than the others in this example.

5.3. A case-control study of lung cancer and smoking

Our final analysis considered preliminary data on gender, age, smoking and disease status from a case-control study of radon gas exposure and lung cancer.[20] Table VIII shows the age × gender distributions of cases and controls after the first stage of sampling. Controls were initially identified by random digit dialling (age under 65) or by sampling at random from Health Care Financing Administration rosters (age 65 and older). All the cases and a subset of the controls were selected for telephone interview. Control selection used a randomized recruitment process in which the probabilities of recruitment varied by age and gender.[28] Both cases and subsampled controls were classified into one of four smoking categories coded never smoked ($k = 1$), ex-smoker ($k = 2$), light smoker ($k = 3$) and heavy smoker ($k = 4$). Those for whom smoking status was indeterminate were omitted from the second-stage sample, which consisted of 663/809 = 81·9 per cent of the cases and 3760/11886 = 31·6 per cent of the controls. Thus the analysis effectively assumed that the smoking data were missing at random[4] *within* each of the 20

Table VIII. Stage-one sample sizes N_{ihj} for the smoking and lung cancer study

Age group (j)	Age range (yrs)	Controls ($i = 0$) Female ($h = 1$)	Controls ($i = 0$) Male ($h = 2$)	Cases ($i = 1$) Female ($h = 1$)	Cases ($i = 1$) Male ($h = 1$)
1	40–59	2738	2589	67	100
2	60–64	468	437	45	97
3	65–69	1147	1010	74	123
4	70–74	1119	903	55	131
5	75–79	838	637	40	77
	Totals	6310	5576	281	528

Table IX. Four analyses of smoking and lung cancer

Risk variable	Model term	Log relative risk ($\hat{\beta} \pm$ SE) WL	PL : B&C	PL : Schill	ML
Age 40–59 yrs	α_1	− 5·686 ± 0·154	− 5·711 ± 0·151	− 5·686 ± 0·152	− 5·726 ± 0·153
Age 60–64	α_2	− 4·079 ± 0·162	− 4·061 ± 0·162	− 4·043 ± 0·162	− 4·078 ± 0·163
Age 65–69	α_3	− 4·447 ± 0·150	− 4·458 ± 0·147	− 4·434 ± 0·148	− 4·472 ± 0·149
Age 70–74	α_4	− 4·275 ± 0·146	− 4·251 ± 0·145	− 4·231 ± 0·146	− 4·259 ± 0·147
Age 75–79	α_5	− 3·988 ± 0·153	− 3·993 ± 0·154	− 3·973 ± 0·154	− 4·006 ± 0·155
Ex-smoking	β_2	1·761 ± 0·170	1·746 ± 0·170	1·712 ± 0·171	1·765 ± 0·170
Light smoking	β_3	3·036 ± 0·179	3·059 ± 0·177	3·041 ± 0·177	3·069 ± 0·179
Heavy smoking	β_4	3·863 ± 0·149	3·882 ± 0·147	3·854 ± 0·148	3·898 ± 0·149

cells shown in Table VIII; conditional on gender and age group in addition to disease outcome, the true smoking status would not influence the probability that the smoking status was unknown. This is a somewhat less stringent assumption than that required for missing data in a single-stage case-control study, namely that missingness is independent of covariables conditional on the disease outcome alone. Note that the $J = 10$ strata are here jointly indexed by $h = 1, 2$ and $j = 1, \ldots, 5$.

Wacholder and Weinberg[20] found that a model with age and smoking effects fit the data nearly as well as one that included also gender. Thus the model equation was

$$\mathrm{logit}\,(p^*_{1hjk}) = \mathrm{logit}\,[\Pr(D = 1 | \mathrm{Gender} = h,\ \mathrm{Age} = j,\ \mathrm{Smoking} = k)] = \alpha_j + \beta_k$$

with the constraint $\beta_1 = 0$ imposed so that ‘Never smoked’ was the baseline category. The logistic regression model actually fitted for the Schill PL analysis was

$$\mathrm{logit}\,(P_{1hj}) = \log\,(N_1/N_0) + \delta_{hj}$$

for the stage-one data and

$$\mathrm{logit}(p_{1hjk}) = \log(n_{1hj}/n_{0hj}) - \delta_{hj} + \alpha_j + \beta_k$$

for stage two, where P_{1hj} and p_{1hjk} are the conditional disease probabilities and δ_{hj} the odds ratio parameters defined in Section 2. The δ_{hj} estimates are not shown since these parameters are of limited interest.

Table IX reports the results. The Breslow–Cain PL and the ML estimates agreed with those reported by Wacholder and Weinberg to two decimal places, in spite of the fact that the somewhat cumbersome EM approach they used required replacing two zero n_{ijk} frequencies by 0·1. None of the four methods described here are affected by the presence of zeros in the table of stage-two frequencies. Our standard errors also agreed with those reported by Wacholder and Weinberg provided that the empirical version of the ML standard error was used rather than the one shown in Table IX.

6. THE MEASUREMENT ERROR PROBLEM

The methodology described herein was developed specifically for the two-stage design suggested by Walker[1] and White.[2] A key feature of this design is the fact that the frequencies n_{ij} of subjects selected by the investigator at stage two may depend on both exposure stratum and disease. A critical assumption (1) is that the probability of disease development depends only on the covariables x_{jk} that are explicitly modelled and not otherwise on stratum. Of course, the methods are applicable when the second-stage sample is drawn at random from among cases and controls, without regard to the strata assigned at stage one. One simply considers the data conditional on the observed values of the n_{ij}. Indeed, in the second example above the stage two sampling frequencies depended only on case or control status.

The case-control study with validation sampling[21] corresponds to our set-up provided that one interprets S as an indicator of the values taken by discrete covariables that are measured with error for all subjects sampled at stage one. We assume that the true covariables X are measured only for n_1 cases and n_0 controls selected at random for the validation sample. Suppose the disease probabilities in the population depend on the vector x_k of true covariables through the logistic model $\Pr(D=1|X=k)=\{1+\exp(-\alpha-x_k^t\beta)\}^{-1}$. Our assumption (1) then implies that D and S are conditionally independent given X and is equivalent to the assumption of non-differential measurement error: $\Pr(S=j|X=k, D=0)=\Pr(S=j|X=k, D=1)$. In this case S is known as a surrogate for X. Provided that the second-stage frequencies n_{ijk} arose from random subsampling of cases and controls, the likelihood function is

$$\prod_{i,j}\Pr(S=j|D=i)^{M_{ij}}\prod_{i,j,k}\Pr(S=j,X=k|D=i)^{n_{ijk}}$$

$$=\prod_{i,j}\left\{\sum_k \Pr(X=k|D=i)\Pr(S=j|X=k,D=i)\right\}^{M_{ij}}$$

$$\times\prod_{i,j,k}\{\Pr(X=k|D=i)\Pr(S=j|X=k,D=i)\}^{n_{ijk}} \qquad (14)$$

where $M_{ij}=N_{ij}-n_{ij}$. When X is binary, or takes a small number of discrete values, Carroll and colleagues[21] demonstrated that it is feasible to maximize (14) by numerical means as a function of the odds ratio parameters β specified by the logistic model, the marginal covariable probabilities $\Pr(X=k)$ and the error rates $\Pr(S=j|X=k)$ that are here assumed equal for cases and controls. We replicated the Carroll *et al.* analysis of data shown in their Table 1, ignoring the strong evidence against our assumption of non-differential measurement error, so as to achieve results comparable to those shown in the second part of their Table 2. Here X was an indicator of exposure to herpes simplex virus type 2 measured by an accurate bioassay (1 = positive, 0 = negative), S was the result of a less accurate bioassay and D denoted cases of cervical cancer or controls. Using a numerical routine to calculate the second derivatives of the logarithm of the likelihood (14), we found the ML estimate and standard error $\hat{\beta}=0{\cdot}958\pm0{\cdot}237$.

Since $\Pr(S=j, X=k|D=i)/\Pr(S=j|D=i)=\Pr(X=k|D=i, S=j)$, (14) may also be written

$$\prod_{i,j}\Pr(S=j|D=i)^{N_{ij}}\prod_{i,j,k}\Pr(X=k|S=j, D=i)^{n_{ijk}} \tag{15}$$

which is proportional to the full likelihood of the data from a two-stage study.[10,11] Equation (6) is also proportional to (15).[15,17] Thus our ML estimation method provides a fully efficient analysis for problems with validation subsampling, non-differential measurement error and discrete surrogates. The two-stage ML estimate for the data in Carroll *et al.*'s Table 1 was $\hat{\beta}=0{\cdot}958\pm0{\cdot}256$. While $\hat{\beta}$ is identical to the coefficient obtained from the more cumbersome numerical approach, as it must be, the two standard errors differ. They are based on two asymptotic formulae that need agree only when the data fit the model exactly, and, as already mentioned, these particular data do not fit the model well. In view of its apparent complexity when the true covariables X take on a larger number of values, Carroll *et al.* eschew the full ML estimate in favour of a simpler PL approach to the likelihood (14).

7. DISCUSSION

Several methods for fitting of logistic regression models to data from two-stage case-control studies are now available that may be implemented using standard computer software. Whereas our simulation study in Section 4 demonstrated the possibility of substantial losses in efficiency by failure to use the full ML estimate, there was little difference between the four estimates and their standard errors in the actual examples in Section 5. Extensive simulations[17] confirm that serious efficiency loss is confined to relatively small regions of a multi-dimensional model space. The two PL procedures tend to give very similar estimates and standard errors, even when they differ from the WL and ML procedures. When the second-stage sample sizes are substantially smaller than those at stage one, the ML estimates display greater bias and larger standard errors than do those from simple logistic regression models fitted to complete stage-one data (see Table II and related discussion), with the bias and efficiency loss both reduced when the correlation between strata and covariables is high.

Two-stage designs offer the prospect of substantial efficiency gains over standard case-control designs by allowing the investigator to balance the numbers of cases and controls sampled within each stratum for the second stage.[3] Such gains should be even greater now that procedures for fully efficient ML estimation are available. The same procedures provide a computationally efficient solution to ML estimation in measurement error problems with validation subsampling, provided that the measurement error is non-differential and that the surrogates take on a relatively small number of discrete values.

ACKNOWLEDGEMENTS

We are indebted to Professor Michael Clarke and Dr. David Clayton for sharing the data from the Leicestershire perinatal mortality study and to Dr. Walter Schill for the data from the occupational lung cancer study. This work was supported in part by grants R01 CA40644 and T32 ES07262 from the United States Public Health Service.

REFERENCES

1. Walker, A. M. 'Anamorphic analysis: sampling and estimation for covariate effects when both exposure and disease are known', *Biometrics*, **38**, 1025–1032 (1982).

2. White, J. E. ‘A two stage design for the study of the relationship between a rare exposure and a rare disease’, *American Journal of Epidemiology*, **115**, 119–128 (1982).
3. Breslow, N. E. and Cain, K. C. ‘Logistic regression for two-stage case-control data’, *Biometrika*, **75**, 11–20 (1988).
4. Rubin, D. ‘Inference and missing data’, *Biometrika*, **63**, 581–592 (1976).
5. Manski, C. F. and McFadden D. ‘Alternative estimators and sample designs for discrete choice analysis’, in Manski, C. F. and McFadden, D. (eds), *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, Cambridge, MA, 1981, pp. 2–50.
6. Hsieh, D. A., Manski, C. F. and McFadden, D. ‘Estimation of response probabilities from augmented retrospective observations’, *Journal of the American Statistical Association*, **80**, 651–662 (1985).
7. Besag, J. ‘Efficiency of pseudolikelihood estimation for simple Gaussian random fields’, *Biometrika*, **64**, 616–618 (1977).
8. Schill, W., Jöckel, K-H., Drescher, K. and Timm, J. ‘Logistic analysis in case-control studies under validation sampling’, *Biometrika*, **80**, 339–352 (1993).
9. Flanders, W. D. and Greenland, S. ‘Analytic methods for two-stage case-control studies and other stratified designs’, *Statistics in Medicine*, **10**, 739–747 (1991).
10. Horvitz, D. G. and Thompson, D. J. ‘A generalization of sampling without replacement from a finite universe’, *Journal of the American Statistical Association*, **47**, 663–685 (1952).
11. Scott, A. J. and Wild, C. J. ‘Fitting logistic models in stratified case-control studies’, *Biometrics*, **47**, 497–510 (1991).
12. Scott, A. J. and Wild, C. J. ‘Maximum likelihood estimation of case-control data’, unpublished manuscript, 1993.
13. Wild, C. J. ‘Fitting prospective regression models to case-control data’, *Biometrika*, **78**, 705–717 (1991).
14. Cosslett, S. R. ‘Maximum likelihood estimator for choice-based samples’, *Econometrica*, **49**, 1289–1316 (1981).
15. Breslow, N. E. and Holubkov, R. ‘Fitting logistic regression models to data from two-stage case-control studies’, Technical Report No. 136, Department of Biostatistics, University of Washington, Seattle, 1995.
16. Robins, J. M., Rotnitzky, A. and Zhao, L. P. ‘Estimation of regression coefficients when some regressors are not always observed’, *Journal of the American Statistical Association*, **89**, 846–866 (1994).
17. Holubkov, R. ‘Maximum likelihood estimation in two-stage case-control studies’, Ph.D. Dissertation. University of Washington, Seattle, 1995.
18. Clayton, D. G. ‘The Leicestershire perinatal mortality study: a case study of multi-group discriminant analysis with complex sampling’, *Statistics in Medicine*, **2**, 229–242 (1983).
19. Jöckel, K.-H., Ahrens, W., Wichmann, H.-E., Becher, H., Bolm-Audorff, U., Jahn, I., Molik, B., Greiser, E. and Timm, J. ‘Occupational and environmental hazards associated with lung cancer’, *International Journal of Epidemiolology*, **21**, 202–213 (1993).
20. Wacholder, S. and Weinberg, C. R. ‘Flexible maximum likelihood methods for assessing joint effects in case-control studies with complex sampling’, *Biometrics*, **50**, 350–357 (1994).
21. Carroll, R. J., Gail, M. H. and Lubin, J. H. ‘Case-control studies with errors in covariates’, *Journal of the American Statistical Association*, **88**, 185–199 (1993).
22. Breslow, N. E. and Day, N.E. *Statistical Methods in Cancer Research*: *The Case-Control Study*, International Agency for Research on Cancer, Lyon, 1980.
23. Prentice, R. L. and Pyke, R. ‘Logistic disease incidence models and case-control studies’, *Biometrika*, **66**, 403–411 (1979).
24. McCullagh, P. and Nelder, J. A. *Generalized Linear Models*, 2nd edn, Chapman and Hall, London, 1989.
25. Huber, P. J. ‘The behavior of maximum likelihood estimates under nonstandard conditions’, In Neyman, (ed), *Proceedings of the 5th Berkeley Symposium in Mathematical Statistics and Probability*, University of California Press, Berkeley, 1967, pp. 221–233.
26. Olkin, I. and Tate, R. F. ‘Multivariate correlation models with mixed discrete and continuous variables’, *Annals of Mathematical Statistics*, **32**, 448–465 (1961).
27. Clarke, M. and Clayton, D. ‘The design and interpretation of case-control studies of perinatal mortality’, *American Journal of Epidemiology*, **113**, 636–645 (1981).
28. Weinberg, C. R. and Sandler, D. P. ‘Randomized recruitment in case-control studies’, *American Journal of Epidemiology*, **134**, 421–432 (1991).