Biometrika (1997), **84**, 1, *pp.* 57–71
Printed in Great Britain

Fitting regression models to case-control data by maximum likelihood

BY A. J. SCOTT AND C. J. WILD
Department of Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand
e-mail: scott@stat.auckland.ac.nz wild@stat.auckland.ac.nz

SUMMARY

We consider fitting categorical regression models to data obtained by either stratified or nonstratified case-control, or response selective, sampling from a finite population with known population totals in each response category. With certain models, such as the logistic with appropriate constant terms, a method variously known as conditional maximum likelihood (Breslow & Cain, 1988) or pseudo-conditional likelihood (Wild, 1991), which involves the prospective fitting of a pseudo-model, results in maximum likelihood estimates of case-control data. We extend these results by showing the maximum likelihood estimates for any model can be found by iterating this process with a simple updating of offset parameters. Attention is also paid to estimation of the asymptotic covariance matrix. One benefit of the results of this paper is the ability to obtain maximum likelihood estimates of the parameters of logistic models for stratified case-control studies, compare Breslow & Cain (1988), Scott & Wild (1991), using an ordinary logistic regression program, even when the stratum constants are modelled.

Some key words: Case-control study; Conditional maximum likelihood; Logistic regression; Maximum likelihood; Pseudo-likelihood; Response selective sampling; Stratified case-control study.

1. INTRODUCTION

Consider an arbitrary prospective regression model for a categorical response variable Y with $I \geqslant 2$ categories and a vector of covariates, x. We denote the model by

$$P_i(x; \theta) = \mathrm{pr}(Y = i \mid x; \theta), \tag{1}$$

for $i = 1, 2, \ldots, I$, where θ denotes a vector of unknown parameters. In the case of a binary response variable, $I = 2$, important examples of the model (1) include the logistic, the probit and the complementary log–log regression models. When $I > 2$, examples include the proportional odds model, the multivariate probit and multivariate logistic regression models. We want to obtain maximum likelihood estimates of the parameters θ from unmatched case-control data supplemented by information on population totals as discussed by Scott & Wild (1991) and Wild (1991).

In the unstratified case, what we mean by 'case-control data supplemented by information on population totals' is the following. There is a finite population of N individuals which is, or is regarded as, a random sample from the joint distribution of (Y, x). All that is known about the finite population is that there are N_i individuals with $Y = i$ $(i = 1, \ldots, I)$. A case-control sample has been drawn from the finite population and the

values of the variables in x have been recorded for each sampled individual. More specifically, from each response class $Y=i$, a simple random sample of size n_i has been taken ($n_i \leqslant N_i$), and x has been recorded for each individual thus sampled. This results in data $\{x_{ij}: j=1,\ldots,n_i\}$ for $i=1,\ldots,I$.

The results in this paper can be regarded as generalisations of the classic results of Anderson (1972) and Prentice & Pyke (1979). They show that, for the special case of a binary logistic model with a constant term, maximum likelihood estimates of all the regression parameters except the constant term can be found by simply fitting the original logistic model as if we had a random sample from the whole population, completely ignoring the case-control structure. Moreover the corresponding estimated standard errors are also valid. Special cases were known even before 1972; see Cox (1970, Ch. 6) for example. In this case, supplementary information on the population totals N_i is only needed to make inferences about the constant term, which is often of secondary importance. This is a particularly attractive feature since this supplementary information is often difficult or impossible to obtain. Unfortunately this feature does not carry over to arbitrary categorical regression models: in general, supplementary information such as knowledge of the population totals, N_i, is necessary for all parameters in (1) to be identified and estimated.

In the case-control sampling scheme described above, a sample of fixed size n_i is taken from $\mathrm{pr}(x \mid Y=i)$ for each response category $i=1,\ldots,I$. Consider an alternative sampling scheme in which we choose $Y=i$ with probability n_i/n, and then sample x from $\mathrm{pr}(x \mid Y=i)$ $(i=1,\ldots,I)$. If this is repeated independently n times, samples of random, rather than fixed, size are obtained from each $Y=i$ category. Let $\mathrm{pr}^*(.)$ refer to probabilities under the new sampling scheme. Then, by Bayes theorem,

$$\mathrm{pr}^*(Y=i \mid x) = \frac{\mu_i P_i(x)}{\sum_{l=1}^{I} \mu_l P_l(x)}, \tag{2}$$

where $\mu_i = n_i/\{n\,\mathrm{pr}(Y=i)\}$ is the ratio of the probability that individuals fall into the ith class in the sample to the probability that individuals fall into this class in the population. We will call (2) 'the pseudo-model'. The same model (2) also arises, conditionally upon the data being recorded, if units are sampled from the joint distribution of (Y, x) but are only recorded with probability proportional to μ_j if $Y=j$.

Consider the log likelihood for θ that would have been appropriate for the case-control data if these had been obtained by sampling independently from the pseudo-model, namely

$$L^*(\theta) = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \log \mathrm{pr}^*(Y=i \mid x_{ij}). \tag{3}$$

It can be shown that consistent estimates of θ under the actual case-control sampling scheme can be obtained by maximising the pseudo-likelihood L^* provided each μ_i in (2) is replaced by a consistent estimator. The method, which is called 'conditional maximum likelihood' by Hsieh, Manski & McFadden (1985) and Breslow & Cain (1988) and 'pseudo-conditional likelihood' by Wild (1991), corresponds to maximising L^* with $\mathrm{pr}(Y=i)$ replaced by its natural estimate, N_i/N, or equivalently μ_i in (2) replaced by the sampling fraction n_i/N_i. For simplicity, we shall use the term pseudo-likelihood in the rest of this paper.

Wild (1991) describes both full maximum likelihood estimation and pseudo-likelihood estimation. In most cases, the pseudo-likelihood method is considerably simpler to implement. For a class of models called 'multiplicative intercept models' by Hsieh et al.

(1985), which will be discussed in more detail in § 2, the pseudo-likelihood procedure produces full maximum likelihood estimates. The most important members of this class are, in the binary case, logistic regression models which contain an intercept term, and when $I > 2$, multivariate logistic models with a full set of intercept terms. For the logistic model for a binary response the true prospective model is

$$\text{pr}(Y = i \mid x;\, \beta) = \exp(\beta^{\mathrm{T}} x)/\{1 + \exp(\beta^{\mathrm{T}} x)\},$$

and, in the pseudo-likelihood approach, the corresponding pseudo-model (2), to be fitted to the case-control data as though the sampling was prospective, is given by

$$\text{pr}^*(Y = i \mid x;\, \beta) = \exp(c + \beta^{\mathrm{T}} x)/\{1 + \exp(c + \beta^{\mathrm{T}} x)\},$$

where $c = \log(\mu_1/\mu_2)$. Implementation of the pseudo-likelihood method is particularly simple in this case since the pseudo-model can be fitted using widely available logistic regression programs such as GLIM that allow for the incorporation of a fixed offset in the model. Where the logistic model contains a constant term, the estimates are maximum likelihood and only the constant term is affected by the case-control sampling. Valid standard errors can be obtained by making a very simple modification to the printed standard error given for the intercept term: see Scott & Wild (1991) for details. Slight generalisations apply to the multivariate logistic.

Although pseudo-likelihood estimates are not in general maximum likelihood, we show in this paper how we can produce maximum likelihood estimates for any model by iterating the pseudo-likelihood procedure with a simple updating of the μ_i values between iterations. We also discuss how the 'information' matrix obtained by naively fitting the pseudo-model can be modified to produce a valid asymptotic variance–covariance matrix for the maximum likelihood estimate $\hat{\theta}$. As well as being practically useful, our approach, which involves finding a profile likelihood for θ, gives some insight into the relationship between the two estimation procedures.

Section 3 generalises these results to stratified case-control studies; compare Breslow & Cain (1988), Scott & Wild (1991). Section 3·1 contains a brief description of the nature of such studies and their uses, while § 3·2 discusses maximum likelihood estimation. In the special case of logistic regression models, it is well known that maximum likelihood estimates can be obtained directly from a standard logistic regression program if a complete set of stratum constants is included. We show that maximum likelihood estimates can again be obtained from a standard logistic regression program by iteration, even when stratum effects are modelled as a function of explanatory variables. Derivations are given in Appendix 1.

2. Unstratified case-control studies

2·1. *Maximum likelihood estimates*

Let $F(x)$ denote the marginal distribution function of x in the whole population. From equation (3) of Wild (1991), the likelihood function for the supplemented case-control data is

$$\prod_{i=1}^{I} \left\{ \prod_{j=1}^{n_i} \text{pr}(x_{ij} \mid Y = i) \right\} \text{pr}(Y = i)^{N_i} = \prod_{i=1}^{I} \left\{ \prod_{j=1}^{n_i} P_i(x_{ij};\, \theta)\, \text{pr}(x_{sij}) \right\} \left\{ \int P_i(x;\, \theta)\, dF(x) \right\}^{N_i - n_i}. \tag{4}$$

This likelihood function is a function of θ, which is the quantity of interest, and $F(x)$,

which is a nuisance parameter. If x is discrete and can take only a small number of possible values, then there is only a small number of nuisance parameters and we can obtain the maximum likelihood estimator of θ by maximising the likelihood directly, as in Scott & Wild (1991). This rapidly becomes unworkable as the range of possible values for x increases. The following theorem gives a method for calculating the maximum likelihood estimator, $\hat{\theta}$, of θ that applies to both discrete and continuous covariates. The derivation is given in Appendix 1. Asymptotic normality of $\hat{\theta}$ follows from standard results on likelihood if the support of x is finite and from the results in Wild (1991) in more general cases.

THEOREM 1. *Under supplemented case-control sampling, the maximum likelihood estimate of θ satisfies*

$$\frac{\partial L^*}{\partial \theta} \equiv \frac{\partial}{\partial \theta} \sum \sum \log \mathrm{pr}^*(Y = i | x_{ij}) = 0, \tag{5}$$

where

$$\mathrm{pr}^*(Y = i | x) = \frac{\mu_i P_i(x; \theta)}{\sum_{l=1}^{I} \mu_l P_l(x; \theta)}, \tag{6}$$

$$\mu_i = \frac{n_i - \gamma_i}{N_i - \gamma_i}, \tag{7}$$

$$\gamma_i = n_i - \sum_{l=1}^{I} \sum_{j=1}^{n_l} \mathrm{pr}^*(Y = i | x_{lj}) \tag{8}$$

for $i = 1, \ldots, I$.

Recall that L^*, with the μ_i's treated as constants, is the pseudo log-likelihood so that

$$\frac{\partial L^*}{\partial \theta} = 0$$

is the pseudo-likelihood equation. The theorem suggests the following straightforward algorithm for obtaining the maximum likelihood estimate. Begin with $\gamma_i^{(0)} = 0$, giving $\mu_i^{(0)} = n_i / N_i$ as in pseudo-likelihood, or conditional maximum likelihood, estimation. Then, for $a = 0, 1, \ldots,$

(i) obtain $\theta^{(a)}$ to maximise $L^{*(a)}(\theta)$;

(ii) obtain

$$\gamma_i^{(a+1)} = n_i - \sum_{l=1}^{I} \sum_{j=1}^{n_l} \mathrm{pr}^{*(a)}(Y = i | x_{lj})$$

and hence $\mu_i^{(a+1)}$, where $\mathrm{pr}^{*(a)}(Y = i | x_{lj})$ is obtained by substituting $\theta^{(a)}$ and $\mu^{(a)}$ in (6). Then iterate to convergence.

Thus the algorithm consists of repeatedly running the pseudo-likelihood program, updating γ_i, and thus μ_i, between runs by using (8).

By inspection of (8), it is obvious that γ_i can be interpreted as measuring the fit of the pseudo-model to the sample count in response class $Y = i$ for $i = 1, \ldots, I$. Because the pseudo-model is artificial, this interpretation is not compelling. However, one can show that, for fixed θ, $\gamma_i = N_i - N_+ \hat{q}_i$, where $\hat{q}_i$ is the maximum likelihood estimate of $\mathrm{pr}(Y = i)$. This gives us a much more satisfactory interpretation of γ_i as measuring the fit of the original model to the population count in the ith response class.

2·2. *Asymptotic variance estimates*

We now turn our attention to finding an estimate of the asymptotic variance covariance matrix of the maximum likelihood estimate $\hat{\theta}$. Let

$$\mathscr{I}^* = -\left.\frac{\partial^2 L^*}{\partial\theta\,\partial\theta^{\mathrm{T}}}\right|_{\theta=\hat{\theta}} \tag{9}$$

be the observed information matrix obtained from the pseudo-likelihood program at the final iteration of the process described in § 2·1. With a slight modification of the argument in Wild (1991), it can be shown that $\mathscr{I}^{*-1}$ gives an asymptotically conservative estimator of $\mathrm{var}(\hat{\theta})$.

In many cases, this may well be all that is needed. However, the modifications to $\mathscr{I}^{*-1}$ needed to produce a consistent estimator of $\mathrm{var}(\hat{\theta})$ are relatively straightforward. Again, derivations are given in Appendix 1. Let

$$B = \sum_x \frac{\partial P^*(x)}{\partial\theta},$$

$$W = \sum_x [\mathrm{diag}\{P^*(x)\} - P^*(x)P^*(x)^{\mathrm{T}}],$$

where $P^*(x)$ denotes the $(I-1)$ vector with ith element, $P_i^*(x) = \mathrm{pr}^*(Y=i\,|\,x)$ as defined in (6) and (8) for $i=1,\ldots,I-1$. In the above, differentiation with respect to θ in the expression for B treats γ as a constant, all expressions are evaluated at $\theta=\hat{\theta}$, $\gamma=\hat{\gamma}$, and sums over x denote sums over all sampled individuals. Then the modification takes the form

$$\mathrm{var}_{\mathrm{est}}(\hat{\theta}) = (\mathscr{I}^* + BK^{-1}B^{\mathrm{T}})^{-1}, \tag{10}$$

where $K = A_0^{-1} - W$, with

$$A_0 = \mathrm{diag}(a_1,\ldots,a_{I-1}) + a_I J, \quad a_i = \frac{1}{n_i-\gamma_i} - \frac{1}{N_i-\gamma_i},$$

and J is a matrix of ones. Calculation is straightforward in any program that allows matrix manipulations. Note that the individual terms in the expressions for B and K have all been evaluated within the pseudo-likelihood program so that it is simple to modify the program to produce the corrected variance estimator.

When the response is binary, B and K reduce to

$$b := \sum_x \frac{\partial P_1^*(x)}{\partial\theta},$$

$$k := (a_1+a_2)^{-1} - \sum_x P_1^*(x)\{1-P_1^*(x)\}.$$

2·3. *Multiplicative intercept models*

The results above simplify considerably for multiplicative intercept models. We have already noted that the class of multiplicative intercept models includes the logistic and multivariate logistic models, but it is broader than this. The class includes any model that can be written in the form

$$\mathrm{pr}(Y=i\,|\,x) = \frac{e^{\theta_{0i}}Q_i(x;\,\theta_1)}{\sum_{l=1}^{I} e^{\theta_{0l}}Q_l(x;\,\theta_1)}. \tag{11}$$

Identifiability constraints such as $\theta_{0I} = 0$, $Q_i(x = \theta; \theta_{1I}) = 1$ and $Q_I(x; \theta_{1I}) \equiv 1$ are required in (11). We adopt these particular constraints here for convenience, but the form of the results is independent of the chosen constraints. The essential feature of multiplicative intercept models is that a full set of parameters,

$$e^{\theta_{0i}} = \left.\frac{\text{pr}(Y=i|x)}{\text{pr}(Y=I|x)}\right|_{x=0},$$

is available to describe all the odds at baseline and that each odds ratio,

$$Q_i(x; \theta_1) = \frac{\text{pr}(Y=i|x)}{\text{pr}(Y=I|x)} \Big/ \frac{\text{pr}(Y=i|0)}{\text{pr}(Y=I|0)}, \tag{12}$$

can be modelled independently by an arbitrary function, Q_i, of the covariates. This allows the use of additive models for the effects of the covariates on the odds ratios, as in Storer, Wacholder & Breslow (1983), for example.

For multiplicative intercept models, the pseudo-model (2) is given by

$$\text{pr}(Y=i|x) = \frac{e^{c_i+\theta_{0i}} Q_i(x; \theta_1)}{\sum_{l=1}^{I} e^{c_l+\theta_{0l}} Q_l(x; \theta_1)},$$

where $c_i = \log(\mu_i/\mu_I)$. Thus the first simplification is that, as noted for the logistic, the pseudo-model can be fitted using a program written for fitting the original prospective model, provided that the program allows for offsets. Furthermore, maximum likelihood and pseudo-likelihood estimates coincide and the algorithm stops at the first iteration. This assertion can be substantiated by looking at the equation $\partial L^*/\partial\theta_{0i} = 0$. This equation tells us that at the pseudo-likelihood estimate we have

$$n_i = \sum_{l=1}^{I} \sum_{j=1}^{n_l} \text{pr}^*(Y=i|x_{lj})$$

and thus, from (8), $\gamma_i = 0$. It follows that the pseudo-likelihood estimate satisfies the conditions for the maximum likelihood estimate given in Theorem 1.

Finally, the modification needed to produce a consistent estimator of $\text{var}(\hat{\theta})$ is particularly simple for multiplicative intercept models. In the binary case, it is straightforward to show that the first column of $\mathcal{I}^*$, namely $\partial^2 L^*/(\partial\theta_0\,\partial\theta)$, is equal to b. Thus, $\mathcal{I}^{*-1} b = (1, 0)^{\mathrm{T}}$. Moreover, for such a model: (i) $\gamma_i = 0$, and (ii) the first element of b, namely $\sum \partial P_1^*(x)/\partial\theta_0$, is equal to $\sum P_1^*(x)\{1 - P_1^*(x)\}$. It then follows immediately from (10) that

$$\text{var}_{\text{est}}(\hat{\theta}) = \text{var}^*_{\text{est}}(\hat{\theta}) - \begin{pmatrix} d & 0^{\mathrm{T}} \\ 0 & 0 \end{pmatrix},$$

where $\text{var}^*_{\text{est}}(\hat{\theta}) = \mathcal{I}^{*-1}$ and

$$d = \left(\frac{1}{n_1} + \frac{1}{n_2}\right) - \left(\frac{1}{N_1} + \frac{1}{N_2}\right).$$

This means that the pseudo-likelihood variance estimate for the odds ratio parameter, $\text{var}^*_{\text{est}}(\hat{\theta}_1)$, needs no correction, just as for logistic regression (Weinberg & Wacholder, 1993). The correction to $\text{var}(\hat{\theta}_0)$ is made up of two parts. The first $n_1^{-1} + n_2^{-1}$, represents the reduction in variance from stratifying into cases and controls; compare simple random sampling versus proportionally allocated stratified sampling. The second, $N_1^{-1} + N_2^{-1}$, represents the increase in variance from replacing μ_i by n_i/N_i for $i = 1, 2$.

Similar results can be obtained for multiplicative intercept models with $I > 2$. It turns out (Scott & Wild, 1986) that

$$\mathrm{var}_{\mathrm{est}}(\hat{\theta}) = \mathrm{var}^{*}_{\mathrm{est}}(\hat{\theta}) - \begin{pmatrix} A_0 & 0^{\mathrm{T}} \\ 0 & 0 \end{pmatrix}.$$

Again the variance estimator for the odds ratio parameter needs no correction at all.

2·4. *Examples*

The data given in Table 1 were obtained by sampling from the results of a population cross-sectional study of people under 35 in Northern Malawi, presented as a three-way contingency table in Clayton & Hills (1993, pp. 156, 175). Cases are new cases of leprosy. Controls are those in the population without leprosy. The data were broken down into 5-year age groups. The variable 'Age' gives the age-group midpoints. The variable 'Scar' refers to the presence or absence of a BCG vaccination scar (1 = present, 0 = absent). The data in Table 1 were obtained by taking all of the cases in the original study and sampling an equal number of controls from the $N_2 = 80\,622$ controls in the population. The number of controls in each combination of age and scar was then observed. The last column of the table, which gives the numbers in the control population for each age group, will be needed in § 3·3.

Since the linear logistic model

$$\mathrm{logit}(p) = \beta_0 + \beta_1 \,\mathrm{Scar} + \beta_2 X,$$

where $X = 100(\mathrm{Age} + 7{\cdot}5)^{-2}$, provides an excellent fit to the whole-population data given in Clayton & Hills (1993), we will use binary regression models in Scar and X throughout. For this data set, $n_1 = N_1 = 260$, $n_2 = 260$ and $N_2 = 80\,622$. The results from fitting a linear logistic model and a complementary log–log model to these data are given in Table 2. The true standard errors from maximum likelihood are denoted by 'SE'. The standard errors obtained from the pseudo-model, or equivalently $\mathscr{I}^{*-1}$, are labelled 'SE*'. As expected from the theory, for the logistic model the standard errors from the pseudo-model are correct for coefficients other than the intercept, while $\mathrm{SE}^{*}(\hat{\beta}_0)$ is too large by the expected amount. The reduction in standard error need not be confined to the intercept for a binary regression model with complementary log–log link. Nevertheless, as can be seen from Table 2, the results of a complementary log–log analysis were essentially identical to the logistic analysis. This might also have been expected since $p = e^{z}/(1 + e^{z})$ and

Table 1. *Simple case-control sample*

	Scar = 0		Scar = 1		Total		Popn
Age	Case	Control	Case	Control	Case	Control	Control
2·5	1	24	1	31	2	55	19312
7·5	11	22	14	39	25	61	17327
12·5	28	23	22	27	50	50	13172
17·5	16	5	28	22	44	27	10325
22·5	20	9	19	12	39	21	8026
27·5	36	17	11	5	47	22	5981
32·5	47	21	6	3	53	24	6479
Total					260	260	80622

Age is age-group midpoint.

Table 2. *Analysis of Table* 1 *data*

	Logistic			Comp. log–log		
	Coef.	SE	SE*	Coef.	SE	SE*
Intercept	−4·510	0·160	0·172	−4·514	0·160	0·171
Scar	−0·302	0·197	0·197	−0·301	0·197	0·197
X	−4·310	0·579	0·579	−4·304	0·578	0·578

$X = 100\ (\text{age} + 7{\cdot}5)^{-2}$.
SE* taken from $\mathscr{I}^{*-1}$. Comp. log–log, complementary log–log.

$p = 1 - \exp(-e^z)$ are hard to distinguish for small p, that is large negative z. In rare-disease situations, the probabilities will usually be confined to such a region.

For nonlogistic models, the variance corrections for case-control sampling with supplementary information on the population totals may not be confined to the intercept and may be smeared across all the other coefficients. However, we have not found any simple case-control sampling situation in which this is true to any appreciable extent. We explored binary regressions with a single linear covariate using both the complementary log–log and probits, covering a wide range of values of p and vastly differing sampling rates. We were unable to find any situations in which standard errors from $\text{var}^*_{\text{est}}(\hat{\beta}) = \mathscr{I}^{*-1}$ were off by more than two per cent for all coefficients apart from the intercept. However, the smearing referred to above can have a substantial effect in the stratified case-control studies treated in § 3·1.

3. STRATIFIED CASE-CONTROL STUDIES

3·1. *Character and rationale*

We now extend the ideas of § 2 to the stratified studies considered in Fears & Brown (1986), Breslow & Zhao (1988) and Scott & Wild (1991). Here we have S strata, indexed by $s = 1, \ldots, S$, defined by discrete or categorical covariates that may or may not be included in the model. Strata could be defined by such things as geographical region, or age-group and sex. Data of the following type are available on the finite population. For each stratum, the numbers of individuals falling into each response category, N_{si}, are known. In addition, for each stratum, a sample of size n_{si} ($n_{si} \leqslant N_{si}$) is taken from response category $Y = i$ ($i = 1, \ldots, I$), and the covariate values of the sampled individuals are measured, resulting in case-control data x_{sij} ($j = 1, \ldots, n_{si}$). This sampling scheme includes the situation in which all members of some response categories are taken, $n_{si} = N_{si}$, as a special case.

In the scenario above, population counts of the form N_{si} may be available from official statistics, or be obtainable from such sources as routine hospital records. However, the same type of data can also arise in a two-stage study with a prospective first stage. Suppose that we have a set of discrete covariates upon which information can be obtained cheaply. It may be possible to run a large study in which the responses and measurements on the cheap covariates are obtained for every subject. The combinations of levels of these covariates then define the strata. At the second stage, case-control subsampling of smaller numbers of individuals is undertaken within strata to obtain information about the remaining more expensive covariates. A scheme like this has been considered by Breslow & Cain (1988), and by N. E. Breslow and R. Holubkov in a private communication, although the

schemes differ slightly in that their first stage is also a case-control sample, whereas our first stage is prospective.

The notation of Scott & Wild (1991) distinguished between covariates that modelled differences between strata, and thus took on the same value for every individual sampled from a stratum, and covariates whose values varied within strata. Such a distinction is unnecessary in this context. The model to be fitted can still be described by (1), where x can contain both within and between stratum information. Indeed, it is even possible to stratify on ranges of a variable, e.g. ages by decade, and then include that variable as a continuous covariate for those individuals included in the case-control samples.

3·2. *Maximum likelihood estimation*

We allow the distribution of the covariate x to differ from stratum to stratum. Again these distributions are handled nonparametrically with no attempt being made to model either the distributions or relationships between them. Let $F_s(x)$ denote the marginal distribution function of x within stratum s. From (4) and the independence of data from different strata, the likelihood function is

$$\prod_{s=1}^{S}\left[\prod_{i=1}^{I}\left\{\prod_{j=1}^{n_{si}}\mathrm{pr}(x_{sij}\,|\,Y=i,s)\right\}\mathrm{pr}(Y=i\,|\,s)^{N_{si}}\right]$$

$$=\prod_{s=1}^{S}\left[\prod_{i=1}^{I}\left\{\prod_{j=1}^{n_{si}}\mathrm{P}_i(\mathrm{x};\,\theta)\,\mathrm{pr}(x_{sij}\,|\,s)\right\}\left\{\int P_i(x;\,\theta)\,dF_s(x)\,dx\right\}^{N_{si}-n_{si}}\right].$$

An extension of Theorem 1 applies with only trivial modifications to its proof. Since case-control sampling is undertaken within each stratum and different sampling rates may be used in different strata, each stratum gives rise to a different pseudo-model, namely

$$\mathrm{pr}^*(Y=i\,|\,x,\,\text{stratum } s)=\frac{\mu_{si}P_i(x)}{\sum_{l=1}^{I}\mu_{sl}P_l(x)}, \tag{13}$$

with corresponding pseudo-loglikelihood

$$L^*(\theta)=\sum_{s=1}^{S}\sum_{i=1}^{I}\sum_{j=1}^{n_{si}}\log\mathrm{pr}^*(Y=i\,|\,x_{sij},\,s). \tag{14}$$

The maximum likelihood estimator of θ satisfies

$$\frac{\partial L^*}{\partial\theta}\equiv\frac{\partial}{\partial\theta}\sum_{s=1}^{S}\sum_{i=1}^{I}\sum_{j=1}^{n_{si}}\log\mathrm{pr}^*(Y=i\,|\,x_{sij},\,s), \tag{15}$$

where $\mathrm{pr}^*(.)$ is defined in (13) with

$$\mu_{si}=\frac{n_{si}-\gamma_{si}}{N_{si}-\gamma_{si}},\quad \gamma_{si}=n_{si}-\sum_{l=1}^{I}\sum_{j=1}^{n_{si}}\mathrm{pr}^*(Y=i\,|\,x_{slj},\,s)\quad(i=1,\ldots,I). \tag{16}$$

The algorithm is as set out in § 2·2, except that we now have a set of γ-parameters for each stratum to be updated using (16).

For a logistic model, the pseudo-model is given by

$$\mathrm{pr}^*(Y=i\,|\,x,\,s;\,\beta)=\exp(c_s+\beta^{\mathrm{T}}x)/\{1+\exp(c_s+\beta^{\mathrm{T}}x)\},$$

where $c_s=\log(\mu_{s1}/\mu_{s2})$, so that the pseudo-conditional likelihood step can again be per-

formed using any logistic regression program which permits the inclusion of offsets. The only change is that observations in different strata now have different offsets. The starting values required for the pseudo-models are now $\mu_{si}^{(0)} = n_{si}/N_{si}$. If the logistic model contains a separate constant term for each stratum, the pseudo-likelihood estimates from the first step are maximum likelihood estimates. This applies more widely to multiplicative intercept models with a complete set of constant terms for each stratum. Fears & Brown (1986) suggested that the pseudo-likelihood estimates were maximum likelihood estimates for any logistic model, whether or not it has a full set of stratum constants. This has been shown to be false (Breslow & Cain, 1988; Breslow & Zhao, 1988) but the result above gives a partial justification to the claim in that the pseudo-likelihood estimate is the first step in the iteration to the maximum likelihood estimate.

As in the unstratified case, it can be shown that $\mathscr{I}^{*-1}$, where

$$\mathscr{I}^* = -\frac{\partial^2 L^*}{\partial\theta\,\partial\theta^{\mathrm{T}}}\bigg|_{\theta=\hat{\theta}} \tag{17}$$

calculated at the final pseudo-likelihood iteration, gives an asymptotically conservative estimator of $\mathrm{var}(\hat{\theta})$. The modifications needed to produce a consistent estimator take the form $\mathrm{var}_{\mathrm{est}}(\hat{\theta}) = \mathscr{I}_P^{-1}$, where

$$\mathscr{I}_P = \mathscr{I}^* + \sum_{s=1}^{S} B_s K_s^{-1} B_s^{\mathrm{T}}, \tag{18}$$

with

$$B_s = \sum_{x\in s} \frac{\partial P^*(x)}{\partial\theta}, \quad K_s = A_{0s}^{-1} - W_s,$$

and where

$$W_s = \sum_{x\in s} [\mathrm{diag}\{P^*(x)\} - P^*(x)P^*(x)^{\mathrm{T}}],$$

$$A_{0s} = \mathrm{diag}(a_{s1}, \ldots, a_{sI-1}) + a_{sI}J,$$

with

$$a_{si} = \frac{1}{n_{si}-\gamma_{si}} - \frac{1}{N_{si}-\gamma_{si}}.$$

In the case of logistic models with a full set of stratum constants, the modification to $\mathscr{I}^{*-1}$ only affects the constant terms.

3·3. *Post-stratification*

Even if a study has not been stratified by design, we can still get some of the efficiency gains by using post-stratification, provided that the stratum totals are known. To illustrate this, we carried out a post-stratified analysis of the data in Table 1 incorporating the information given there about the control population totals in each age-group stratum. The results of a logistic analysis linear in Scar and X are given in Table 3. There is a reduction in all standard errors with the incorporation of the additional information. The reduction is particularly noticeable in the terms which model age-stratum effects, namely, the intercept and coefficient of X. The increase in precision of Scar at approximately 10%

Table 3. *Post-stratified analysis of Table 1 data*

	Unstratified analysis			Post-stratified analysis		
	Coef.	SE	SE*	Coef.	SE	SE*
Intercept	−4·510	0·160	0·172	−4·481	0·114	0·170
Scar	−0·302	0·197	0·197	−0·421	0·178	0·199
X	−4·310	0·579	0·579	−4·091	0·449	0·541

$X = 100\,(\text{age} + 7{\cdot}5)^{-2}$.
SE* taken from $\mathcal{I}^{*-1}$.

is modest, 20% reduction in variance, perhaps reflecting the lack of relationship between Age, equivalently X, and Scar.

Appendix 1

Derivations

We start the proof to Theorem 1 by deriving the profile log-likelihood, $L_P(\theta)$, which is obtained by maximising the full likelihood (4) over the nuisance parameter $F(x)$.

Lemma 1. *Under supplemented case-control sampling the profile log-likelihood for* θ *is*

$$L_P(\theta) = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \log \text{pr}^*(Y = i \mid x_{ij}) + \sum N_i \log(N_i - \gamma_i) - \sum n_i \log(n_i - \gamma_i), \tag{A1·1}$$

where $\text{pr}^*(Y = i \mid x)$, μ_i *and* γ_i *are given by equations* (6), (7) *and* (8) *for* $i = 1, \ldots, I$, *respectively.*

We note that the γ_i's satisfy the constraint $\sum \gamma_i = 0$. This follows from (8) and the fact that $\sum_i \text{pr}^*(Y = i \mid x) = 1$.

Proof of Lemma 1. If the support of x is finite, obtaining the profile likelihood $L_P(\theta)$ involves maximisation with respect to a fixed finite number of parameters. In the semiparametric case where the distribution of x is left unspecified, the nonparametric maximum likelihood estimate is discrete with all its probability mass being placed on the observed covariate values; compare Wild (1991), Gill, Vardi & Wellner (1988). Thus, we can prove the result by working with a discrete distribution for x.

It is convenient to make the following notational change which applies for the duration of this proof only. We assume that x is discrete and can take on values $x_j\,(j = 1, \ldots, J)$. When n_i individuals are sampled from the N_i in the subpopulation with $Y = i$, we observe n_{ij} with $x = x_j$. We use a '+' in place of a subscript to denote summation over that subscript. Thus, $n_i = n_{i+}$.

From (4) the likelihood can be written as

$$\left\{ \prod_{i=1}^{I} \prod_{j=1}^{J} (P_{ij}\delta_j)^{n_{ij}} \right\} \left\{ \prod_{i=1}^{I} \left(\sum_l P_{il}\delta_l \right)^{N_i - n_{i+}} \right\},$$

where $P_{ij}(\theta) = \text{pr}(Y = i \mid x_j;\, \theta)$ and $\delta_j = \text{pr}(x_j)$. Thus, the log-likelihood is given by

$$L(\theta, \delta) = \sum_i \sum_j n_{ij} \log P_{ij} + \sum_j n_{+j} \log \delta_j + \sum_i (N_i - n_{i+}) \log \left(\sum_l P_{il}\delta_l \right).$$

To find the profile likelihood of θ, we need to replace δ by $\hat{\delta}(\theta)$, the value obtained by maximising the log likelihood over δ for fixed θ. If we introduce a Lagrange multiplier η to take care of the constraint $\sum \delta_j = 1$ and set the derivative with respect to δ_j equal to zero, we obtain

$$\frac{n_{+j}}{\delta_j} + \sum_i (N_i - n_i) \frac{P_{ij}}{\sum_l P_{il}\delta_l} + \eta = 0. \tag{A1·2}$$

Multiplying through by δ_j and summing over j then gives $\eta = -N$. Substituting this into (A1·2) gives the following system of equations for the $\hat{\delta}_j(\theta)$'s:

$$\delta_j = \frac{n_{+j}}{N - \sum_i \{(N_i - n_i)P_{ij}/(\sum_l P_{il}\delta_l)\}} = \frac{n_{+j}}{N\sum_i \mu_i P_{ij}}, \tag{A1·3}$$

where

$$\mu_i = 1 - \frac{N_i - n_i}{N\sum P_{il}\delta_l}.$$

Substituting the expression for δ_j back into the definition of μ_i, we obtain

$$\frac{\mu_i}{1-\mu_i} = \frac{\sum_j n_{+j}P^*_{ij}}{N_i - n_i}, \tag{A1·4}$$

where

$$P^*_{ij} = \frac{\mu_i P_{ij}}{\sum_l \mu_l P_{lj}}. \tag{A1·5}$$

From (A1·4),

$$\mu_i = \frac{n_i - \gamma_i}{N_i - \gamma_i}, \tag{A1·6}$$

where

$$\gamma_i = n_i - \sum_j n_{+j}P^*_{ij}. \tag{A1·7}$$

We can substitute (A1·3) back into (4), and use (A1·4) and (A1·5) to obtain the profile loglikelihood which we can express, ignoring additive constants, as

$$L_P(\theta) = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log\left(\frac{P_{ij}}{\sum_l \mu_l P_{lj}}\right) + \sum_{i=1}^{I}(N_i - n_{i+})\log\left(\sum_l \frac{n_{+l}P_{il}}{\sum_k \mu_k P_{kl}}\right)$$

$$= \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log\left(\frac{\mu_i P_{ij}}{\sum_l \mu_l P_{lj}}\right) - \sum n_i \log\mu_i - (N_i - n_i)\log(1-\mu_i) \tag{A1·8}$$

$$= \sum\sum n_{ij}\log P^*_{ij} + \sum N_i \log(N_i - \gamma_i) - \sum n_i \log(n_i - \gamma_i). \tag{A1·9}$$

Lemma 1 as stated follows from translating the notation used within this proof, which lists the distinct values of x as x_j $(j = 1, \ldots, J)$ together with multiplicities n_{ij}, back into the notation used elsewhere, namely x_{ij} $(i = 1, \ldots, I, j = 1, \ldots, n_i)$, which keeps track of the value of x for each individual sampled. □

The profile likelihood in Lemma 1 is implicitly defined, being represented in terms of θ and a vector $\gamma = (\gamma_1, \ldots, \gamma_I)^{\mathrm{T}}$ which is itself a function of θ. To write this function down explicitly would require solution of the system of equations (6), (7) and (8). However, Theorem 1 allows us to characterise and find the maximum likelihood estimate without ever having to obtain $\gamma = \gamma(\theta)$ explicitly.

Proof of Theorem 1. Recall that $\sum \gamma_i = 0$. Let $\breve{\gamma}$ contain the first $I - 1$ elements of γ. Consider the profile loglikelihood (A1·1). Now, from (6), $\mathrm{pr}^*(Y = i|x)$ is a function of θ and $\mu = \mu(\breve{\gamma})$ from (7). Thus, we can think of the profile loglikelihood (A1·1) being of the form

$$L_P(\theta) = g(\theta, \breve{\gamma}), \tag{A1·10}$$

where $\breve{\gamma} = \breve{\gamma}(\theta)$. The maximum likelihood estimate of θ therefore satisfies

$$0^{\mathrm{T}} = \frac{\partial \log L_P}{\partial \theta^{\mathrm{T}}} = \frac{\partial g}{\partial \theta^{\mathrm{T}}} + \frac{\partial g}{\partial \breve{\gamma}^{\mathrm{T}}}\frac{\partial \breve{\gamma}}{\partial \theta^{\mathrm{T}}}. \tag{A1·11}$$

The maximiser of L^* satisfies $\partial g/\partial\theta = 0$. The first and second derivatives of g are given in Appendix 2. It is immediately clear that each component of $\partial g/\partial\breve{\gamma}$ is zero when the γ_i's satisfy (8). □

Now consider variance estimation. We consider only the case when the support of x is finite, so that the profile loglikelihood L_P has been obtained from a likelihood with a finite-dimensional parameter vector. Thus, it follows from Richards (1961), Seber & Wild (1989, § 2.2.3) that we can obtain a consistent estimator of the asymptotic covariance of $\hat{\theta}$ as

$$\mathrm{var}_{\mathrm{est}}(\hat{\theta}) = \mathscr{I}_P^{-1}, \tag{A1·12}$$

where

$$\mathscr{I}_P = -\left.\frac{\partial^2 L_P}{\partial\theta\,\partial\theta^{\mathrm{T}}}\right|_{\theta=\hat{\theta}}. \tag{A1·13}$$

Now, thinking again of L_P as being of the form $g(\theta, \breve{\gamma})$ as in the proof of Theorem 1,

$$\mathscr{I}_P = \mathscr{I}^* - \left(\frac{\partial\breve{\gamma}^{\mathrm{T}}}{\partial\theta}\frac{\partial^2 g}{\partial\breve{\gamma}\,\partial\theta^{\mathrm{T}}} + \frac{\partial^2 g}{\partial\theta\,\partial\breve{\gamma}^{\mathrm{T}}}\frac{\partial\breve{\gamma}}{\partial\theta^{\mathrm{T}}} + \frac{\partial\breve{\gamma}^{\mathrm{T}}}{\partial\theta}\frac{\partial^2 g}{\partial\breve{\gamma}\,\partial\breve{\gamma}^{\mathrm{T}}}\frac{\partial\breve{\gamma}}{\partial\theta^{\mathrm{T}}}\right). \tag{A1·14}$$

We note that the omitted term in the expansion (A1·14) (Seber & Wild, 1989, p. 682), namely

$$\sum\frac{\partial g}{\partial\gamma_j}\frac{\partial^2\gamma_j}{\partial\theta\,\partial\theta^{\mathrm{T}}},$$

is zero at $\hat{\theta}$ since $\partial g/\partial\gamma = 0$ at $\hat{\theta}$. Expressions for $\partial g/\partial\breve{\gamma}$, $\partial^2 g/\partial\theta\,\partial\breve{\gamma}^{\mathrm{T}}$ and $\partial\breve{\gamma}/\partial\theta^{\mathrm{T}}$ are derived in Appendix 2. Substitution of these expressions in (A1·14) leads to the expression in (10).

Equation (A1·14) allows the 'observed' information matrix from the pseudo-likelihood fit to be modified to give the required observed information matrix for the profile likelihood. We stress that the above results apply to observed and not expected information. The observed information under the prospective pseudo-model is not in general the same as the expected information, although the two coincide for logistic models and we would expect them to be close in large samples. Note that our variance estimator is asymptotically equivalent to the more complicated alternative suggested in Wild (1991). That estimator is shown to be valid for arbitrary covariate distributions and it seems from limited simulations that the estimator proposed here also works well even when the support of x is not finite. More work is needed on this, however.

The extension to stratified sampling is straightforward. We now have a $\breve{\gamma}$ vector for each stratum and (A1·14) becomes

$$\mathscr{I}_P = \mathscr{I}^* - \sum_{s=1}^{S}\left(\frac{\partial\breve{\gamma}_s^{\mathrm{T}}}{\partial\theta}\frac{\partial^2 g}{\partial\breve{\gamma}_s\,\partial\theta^{\mathrm{T}}} + \frac{\partial^2 g}{\partial\theta\,\partial\breve{\gamma}_s^{\mathrm{T}}}\frac{\partial\breve{\gamma}_s}{\partial\theta^{\mathrm{T}}} + \frac{\partial\breve{\gamma}_s^{\mathrm{T}}}{\partial\theta}\frac{\partial^2 g}{\partial\breve{\gamma}_s\,\partial\breve{\gamma}_s^{\mathrm{T}}}\frac{\partial\breve{\gamma}_s}{\partial\theta^{\mathrm{T}}}\right). \tag{A1·15}$$

The derivative matrices given in Appendix 2 now apply to a single stratum and require obvious changes, for example γ_i to γ_{si} and N_i to N_{si}, and the sums over x refer only to individuals within the sth stratum.

Appendix 2

Derivatives

Let

$$g(\theta, \gamma) = \sum_{i=1}^{I}\sum_{j=1}^{n_i}\log P_i^*(x_{ij}) + \sum N_i\log(N_i - \gamma_i) - \sum n_i\log(n_i - \gamma_i),$$

where

$$P_i^*(x) = \frac{\mu_i P_i(x;\,\theta)}{\sum_{l=1}^{I}\mu_l P_l(x;\,\theta)}, \quad \mu_i = (n_i - \gamma_i)/(N_i - \gamma_i).$$

Using

$$\sum_x h(x) := \sum_{i=1}^{I} \sum_{j=1}^{n_i} h(x_{ij})$$

to denote sums over the covariate values of all sampled individuals, we have

$$\frac{\partial g}{\partial \gamma_i} = a_i \left[\gamma_i - \left\{ n_i - \sum_x P_i^*(x) \right\} \right], \tag{A2·1}$$

where

$$a_i = \frac{(N_i - n_i)}{(n_i - \gamma_i)(N_i - \gamma_i)} = \frac{1}{n_i - \gamma_i} - \frac{1}{N_i - \gamma_i}.$$

Let

$$B = \sum_x \frac{\partial P^*(x)}{\partial \theta}, \quad W = \sum_x [\mathrm{diag}\{P^*(x)\} - P^*(x)P^*(x)^{\mathrm{T}}],$$

where $P^*(x)$ denotes the $(I-1)$ vector with ith element $P_i^*(x)$ for $i = 1, \ldots, I$. Then, differentiating (A2·1) with respect to γ_j and θ_j respectively, we find

$$\frac{\partial^2 g}{\partial \gamma \, \partial \theta^{\mathrm{T}}} = AC^{\mathrm{T}}B^{\mathrm{T}}, \quad \frac{\partial^2 g}{\partial \gamma \, \partial \gamma^{\mathrm{T}}} = A - AC^{\mathrm{T}}WCA,$$

where $A = \mathrm{diag}(a_1, \ldots, a_I)$ and $C = (I_{I-1} - 1)$ with I_{I-1} a $(I-1) \times (I-1)$ identity matrix and 1 a $(I-1)$-vector of ones.

We have made no use of the constraint $\sum \gamma_i = 0$ in the above. The expressions needed in Appendix 1 require constrained derivatives using the parameter vector $\breve{\gamma}$ which contains the first $I-1$ γ_i's. Now $\gamma = C^{\mathrm{T}}\breve{\gamma}$, so that

$$\frac{\partial^2 g}{\partial \breve{\gamma} \, \partial \theta^{\mathrm{T}}} = C \frac{\partial^2 g}{\partial \gamma \, \partial \theta^{\mathrm{T}}} = CAC^{\mathrm{T}}B^{\mathrm{T}} = A_0 B^{\mathrm{T}},$$

$$\frac{\partial^2 g}{\partial \breve{\gamma} \, \partial \breve{\gamma}^{\mathrm{T}}} = C \frac{\partial^2 g}{\partial \gamma \, \partial \gamma^{\mathrm{T}}} C^{\mathrm{T}} = A_0 - A_0 W A_0,$$

where $A_0 = CAC^{\mathrm{T}}$.

We also need an expression for $\partial \breve{\gamma} / \partial \theta^{\mathrm{T}}$. This can be obtained by differentiating (8) with respect to θ to give

$$(I - C^{\mathrm{T}}WCA)C^{\mathrm{T}} \frac{\partial \breve{\gamma}}{\partial \theta^{\mathrm{T}}} = -C^{\mathrm{T}}B^{\mathrm{T}}.$$

Pre-multiplying by CA leads to

$$\frac{\partial \breve{\gamma}}{\partial \theta^{\mathrm{T}}} = -(A_0 - A_0 W A_0)^{-1} A_0 B^{\mathrm{T}}.$$

Substituting the expressions for $\partial \breve{\gamma} / \partial \theta^{\mathrm{T}}$, $\partial^2 g / \partial \breve{\gamma} \, \partial \theta^{\mathrm{T}}$ and $\partial^2 g / \partial \breve{\gamma} \, \partial \breve{\gamma}^{\mathrm{T}}$ in (A2·1) leads to

$$\mathscr{I}_P = \mathscr{I}^* + BK^{-1}B^{\mathrm{T}},$$

where

$$K = A_0^{-1} - W.$$

In the binary case $C = (1, -1)$, $\breve{\gamma} = \gamma_1$ and

$$\frac{\partial^2 g}{\partial \theta \, \partial \breve{\gamma}} = a_+ b, \quad \frac{\partial^2 g}{\partial \breve{\gamma}^2} = a_+(1 - a_+ w), \quad \frac{\partial \breve{\gamma}}{\partial \theta} = -\frac{b}{1 - a_+ w},$$

where

$$a_+ = a_1 + a_2, \quad b = \sum_x \frac{\partial P_1^*(x)}{\partial \theta}, \quad w = \sum_x P_1^*(x)\{1 - P_1^*(x)\}.$$

References

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

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

Breslow, N. E. & Zhao, L. P. (1988). Logistic regression for stratified case-control studies. *Biometrics* **44**, 891–99.

Clayton, D. & Hills, M. (1993). *Statistical Models in Epidemiology*. Oxford: Oxford University Press.

Cox, D. R. (1970). *The Analysis of Binary Data*. London: Chapman and Hall.

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

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

Hsieh, D. A., Manski, C. F. & McFadden, D. (1985). Estimation of response probabilities from augmented retrospective observations. *J. Am. Statist. Assoc.* **80**, 651–62.

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

Richards, F. S. G. (1961). A method of maximum likelihood estimation. *J. R. Statist. Soc.* B **23**, 469–76.

Scott A. J. & Wild, C. J. (1986). Fitting logistic models under case-control or choice based sampling. *J. R. Statist. Soc.* B **48**, 170–82.

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

Seber, G. A. F. & Wild, C. J. (1989). *Nonlinear Regression*. New York: Wiley.

Storer, B. E., Wacholder, S. & Breslow, N. E. (1982). Maximum likelihood fitting of general risk models to stratified data. *Appl. Statist.* **32**, 172–81.

Weinberg, C. R. & Wacholder, S. (1993). Prospective analysis of case-control data under general multiplicative-intercept risk models. *Biometrika* **80**, 461–5.

Wild, C. J. (1991). Fitting prospective regression models to case-control data. *Biometrika* **78**, 705–17.

[*Received November* 1995. *Revised June* 1996]