Biometrika (1979), 66, 3, *pp.* 403–11
Printed in Great Britain

Logistic disease incidence models and case-control studies

BY R. L. PRENTICE
Fred Hutchinson Cancer Research Center, Seattle, Washington
AND R. PYKE
Department of Mathematics, University of Washington

SUMMARY

The probability of disease development in a defined time period is described by a logistic regression model. A model for the regression variable, given disease status, is induced and is applied to case-control data. It is shown that the odds ratio estimators and their asymptotic variance matrices may be obtained by applying the original logistic regression model to the case-control study as if the data had been obtained in a prospective study. This result gives a flexible and convenient method of analysis for a range of case-control studies in which stratum sizes are reasonably large. The work extends Anderson's (1972) results on logistic discrimination and generalizes the findings of Breslow & Powers (1978) on the equivalence of odds ratio estimators when both prospective and retrospective logistic models are applied to case-control data.

Some key words: Asymptotic likelihood; Case-control study; Constrained maximum likelihood; Logistic regression; Odds ratio; Retrospective sampling.

1. INTRODUCTION

The case-control study is a primary tool for the study of factors related to disease incidence. Epidemiological studies may concentrate, for example, on exposures from industrial, environmental or iatrogenic sources or on personal, demographic or genetic characteristics. A prospective or cohort study would identify variously exposed subjects who would then be followed for disease development. A case-control study, in which diseased cases and disease-free control study subjects are identified and 'followed back' to ascertain their exposure levels, often brings about important economies in cost and study duration, particularly with rare diseases.

There are a number of possible case-control study designs. One of the most common, sometimes referred to as a cumulative incidence study (Miettinen, 1976), utilizes all or a random sample of cases that occur in a specified time period, the case accession period, in a defined population. The control sample is then randomly selected from individuals that are disease-free at the end of the case accession period. Various degrees of matching or stratification can be built into the design and case and control sampling fractions can be allowed to vary among such matched sets or strata.

Suppose that k mutually exclusive and exhaustive disease groups are defined and let $D = i$ denote the development of the ith disease during the defined accession period. Let $D = 0$ indicate the disease-free state at the end of the accession period. Suppose that a regression vector $z = (z_1, \ldots, z_p)$ is to be related to disease incidence. The vector z would typically include characteristics or exposures of primary interest along with auxiliary variables thought to be related to disease incidence, as well as product terms that would permit certain interactions between primary and auxiliary variables in relation to disease incidence. The regression vector pertains to the study subject at the beginning of the case accession period.

Let $\mathrm{pr}(D=i\,|\,z)$ denote the probability that an individual with regression vector z develops disease $D=i$ in the defined accession period. A prospective study in which initially disease-free individuals are followed throughout the accession period to observe disease incidence would involve direct sampling from $\mathrm{pr}(D\,|\,z)$ and would permit estimation of $\mathrm{pr}(D\,|\,z)$ or any of its derivatives.

In comparison, the unstratified case-control study described above involves direct sampling from $\mathrm{pr}(z\,|\,D)$; in fact, the sampling structure is that of a $(k+1)$-sample problem with separate samples obtained for each value of $D=0,1,\ldots,k$. In that $\mathrm{pr}(z\,|\,D)$ does not completely determine $\mathrm{pr}(D\,|\,z)$, the full prospective model cannot be estimated from case-control data alone. Odds ratios calculated from $\mathrm{pr}(D\,|\,z)$ can, however, be estimated. The 'odds' for disease $D=i$ for an individual with characteristics z, relative to that for an individual with some standard regression vector z_0, is

$$\{\mathrm{pr}(D=i\,|\,z)/\mathrm{pr}(D=0\,|\,z)\}\{\mathrm{pr}(D=i\,|\,z_0)/\mathrm{pr}(D=0\,|\,z_0)\}^{-1} \quad (i=1,\ldots,k). \tag{1}$$

If the probability of disease is small, (1) will closely approximate $\mathrm{pr}(D=i\,|\,z)/\mathrm{pr}(D=i\,|\,z_0)$, which is sometimes referred to as a relative risk. Let $\mathrm{pr}(D)$ and $\mathrm{pr}(z)$ represent marginal probability functions or probability density functions in the population as a whole. Substitution of

$$\mathrm{pr}(D\,|\,z)=\mathrm{pr}(z\,|\,D)\,\mathrm{pr}(D)/\mathrm{pr}(z) \tag{2}$$

into each of the terms of (1) shows that the odds ratios can be written

$$\{\mathrm{pr}(z\,|\,D=i)/\mathrm{pr}(z_0\,|\,D=i)\}\{\mathrm{pr}(z\,|\,D=0)/\mathrm{pr}(z_0\,|\,D=0)\}^{-1} \quad (i=1,\ldots,k). \tag{3}$$

It follows that the odds ratio (1) can be estimated from case-control data.

The analysis of case-control data may proceed by specifying and fitting a statistical model $\mathrm{pr}(z\,|\,D)$ after which quantities of interest such as odds ratios can be estimated (Prentice, 1976). In fact, the classical procedure of Mantel & Haenszel (1959) can be viewed as a special case of this approach. When several, possibly continuous, exposure variables are simultaneously under study, such retrospective modelling is likely to involve a large number of parameters and to be unduly cumbersome. The alternative approach of specifying a prospective model, $\mathrm{pr}(D\,|\,z)$, which is used to induce a model, $\mathrm{pr}(z\,|\,D)$, for application to the case-control data, is then likely to be preferable. In the next section a prospective logistic regression model is used to induce a model $\mathrm{pr}(z\,|\,D)$, which also turns out to be of logistic form.

2. PROSPECTIVE AND RETROSPECTIVE LOGISTIC MODELS

A logistic regression model for disease incidence during the defined accession period would specify

$$\mathrm{pr}(D=i\,|\,z)=\exp(\alpha_i+z\beta_i)\Big/\sum_{l=0}^{k}\exp(\alpha_l+z\beta_l) \quad (i=0,\ldots,k), \tag{4}$$

where β_i is a $p\times 1$ real parameter, with $\alpha_0=0$, $\beta_0=0$ for uniqueness. The odds ratios (1) are easily calculated to be

$$\exp\{(z-z_0)\beta_i\} \quad (i=1,\ldots,k). \tag{5}$$

Further beginning with (5) one can recover (4) upon defining

$$\alpha_i=\log\{\mathrm{pr}(D=i\,|\,z)/\mathrm{pr}(D=0\,|\,z_0)\}-z_0\beta_i \quad (i=1,\ldots,k).$$

Similarly the odds ratio representations (3) and (5) allow one to calculate

$$\text{pr}(z \mid D = i) = c_i \exp\{\gamma(z) + z\beta_i\} \quad (i = 0, \ldots, k), \tag{6}$$

where $\gamma(z) = \log\{\text{pr}(z \mid D = 0)/\text{pr}(z_0 \mid D = 0)\}$ for all z and $c_i = c_i(\gamma, \beta_i)$ is a normalization constant. The induced model (6) is again of logistic form. The prospective model (4) and the retrospective model (6) are precisely equivalent provided that α_i $(i = 1, \ldots, k)$ in (4) and the function $\gamma(.)$ in (6) are unrestricted. Both (4) and (6) simply capture the notion (5) that regression variables affect the odds ratios in a multiplicative manner.

Assuming a prospective logistic model (4), the case-control study analysis involves fitting (6) to the observed data and thereby estimating the odds ratio parameters β_i $(i = 1, \ldots, k)$. If the sample space for z is finite, (6) is an ordinary logistic model and standard likelihood methods may be applied. More generally, however, nonstandard estimation theory will be required because of the unspecified nuisance function $\gamma(.)$. The next two sections develop the required likelihood equations and asymptotic distribution theory. This is followed by a discussion of the relationship to Anderson's (1972, 1973) results.

3. Maximum likelihood estimation

Suppose now that n_0 controls and n_i cases of disease i $(i = 1, \ldots, k)$ are randomly selected from their respective subpopulations. Note that some of the n_i can be identically zero for all values of $n = n_0 + \ldots + n_k$. If $n_0 = 0$ only differences among the β_i can be estimated. For simplicity of notation assume $n_i > 0$ $(i = 0, \ldots, k)$ and let z_{ij} $(j = 1, \ldots, n_i)$ denote the n_i regressor variables in disease group i $(i = 1, \ldots, k)$. The likelihood function can be written

$$\prod_{i=0}^{k} \prod_{j=1}^{n_i} \text{pr}(z_{ij} \mid D_i),$$

where, according to (6), $\text{pr}(z \mid D = i) = c_i \exp\{\gamma(z) + z\beta_i\}$.

Reparameterization clarifies the estimation problem. Set

$$q(z) = \{\exp \gamma(z)\} \sum_{l=0}^{k} (n_l/n)\, c_l \exp(z\beta_l).$$

Note that the integral of $q(.)$ over the sample has value 1 since (6) is a probability density function for each i. In fact $q(.)$ is the marginal probability density function for z under the case-control sampling scheme in which $\text{pr}(D = i) = n_i/n$. Substitution for $\exp\{\gamma(z)\}$ in (6) gives

$$\text{pr}(z \mid D = i) = \left\{\exp(\delta_i + z\beta_i) \Big/ \sum_{l=0}^{k} \exp(\delta_l + z\beta_l)\right\} q(z) n n_i^{-1}, \tag{7}$$

where $\delta_l = \log(c_l n_l n^{-1})$. The likelihood function is therefore proportional to

$$L = \left\{\prod_{i=0}^{k} \prod_{j=1}^{n_i} p_i(z_{ij})\right\} \left\{\prod_{i=0}^{k} \prod_{j=1}^{n_i} q(z_{ij})\right\} = L_1 L_2, \tag{8}$$

where we have written

$$p_i(z) = \exp(\delta_i + z\beta_i) \Big/ \sum_{l=0}^{k} \exp(\delta_l + z\beta_l) \quad (i = 0, \ldots, k).$$

The parameters $\theta^{\mathrm{T}} = (\theta_1^{\mathrm{T}}, \theta_2^{\mathrm{T}}) = (\delta_1, \ldots, \delta_k, \beta_1^{\mathrm{T}}, \ldots, \beta_k^{\mathrm{T}})$ and $q(.)$ are restricted only by the fact that (7) is a probability distribution for each i; that is,

$$n_i n^{-1} = \int p_i(z)\, q(z)\, dz \quad (i = 0, \ldots, k), \tag{9}$$

where the integral sign denotes integration or summation or both, depending on the sample space for z.

Consider first maximization of (8) without regard for the constraints (9). Because of the likelihood factorization 'unconstrained' maximum likelihood estimators $\hat{\theta}_1^{\mathrm{T}} = (\delta_1, \ldots, \delta_k)$ and $\hat{\theta}_2^{\mathrm{T}} = (\hat{\beta}_1^{\mathrm{T}}, \ldots, \hat{\beta}_k^{\mathrm{T}})$ are solutions to

$$\partial \log L_1/\partial \delta_i = n_i - \sum_{m=0}^{k} \sum_{j=1}^{n_m} p_i(z_{mj}) = 0 \quad (i = 1, \ldots, k),$$

$$\partial \log L_1/\partial \beta_i = \sum_{j=1}^{n_i} z_{ij}^{\mathrm{T}} - \sum_{m=0}^{k} \sum_{j=1}^{n_m} z_{mj}^{\mathrm{T}} p_i(z_{mj}) = 0 \quad (i = 1, \ldots, k). \tag{10}$$

Note that $\partial \log L_1/\partial \beta_i$ in (10) is a $p \times 1$ vector. The corresponding maximum likelihood estimator of the $q(.)$ distribution is the empirical probability function, $\hat{q}(.)$, that assigns mass s/n to any value of z that is observed with multiplicity s, and assigns value zero elsewhere.

The likelihood function constrained by (9) can be at most as large as that evaluated at the unconstrained maximum likelihood estimators $\hat{\theta}^{\mathrm{T}} = (\hat{\theta}_1^{\mathrm{T}}, \hat{\theta}_2^{\mathrm{T}})$, $\hat{q}(.)$. Because of the choice of parameterization it happens that the constraints are satisfied by $\hat{\theta}$ and $\hat{q}(.)$; in fact, substitution of these values into the last k equations of (9) gives precisely the first k equations of (10). This, along with the fact that $\int \hat{q}(z)\,dz = 1$, shows that $\{\hat{\theta}, \hat{q}(.)\}$ are the desired constrained maximum likelihood estimators. In particular the maximum likelihood estimators $\hat{\beta}_i$ $(i = 1, \ldots, k)$ are solutions to (10). Note that if the prospective model (4) were applied to the case-control data, as if the sampling were prospective, the likelihood equations would be exactly (10) with δ's replaced by α's. The maximum likelihood estimators of the odds ratio parameters $\hat{\beta}_i$ $(i = 1, \ldots, k)$ can therefore be obtained from such an application of (4). The next section gives similar results for the asymptotic distribution of $\hat{\beta}_1, \ldots, \hat{\beta}_k$.

4. Asymptotic distribution of odds ratio estimators

Standard constrained maximum likelihood distribution theory (Aitchison & Silvey, 1958) could be applied to (8), subject to the constraints (9), provided that the sample space for z is finite. Such theory could also be used with arbitrary sample space if $q(.)$ were restricted to be known up to a finite dimensional parameter. Distributional results that are applicable for arbitrary sample space for z and without restrictions on $q(.)$ will, however, require a specialized development.

The desired distribution theory for $\hat{\theta}$ can be developed in a rather classical manner by considering the asymptotic behaviour of solutions to (10) under the $k+1$ sample structure of the case-control study. Since much of the development is standard, only an outline is given at some points.

A first-order Taylor expansion of the left-hand side of (10) about the 'true' θ^0 gives

$$0 = \partial \log L_1/\partial \hat{\theta} = \partial \log L_1/\partial \theta^0 + (\partial^2 \log L_1/\partial \theta^* \partial \theta^*)(\hat{\theta} - \theta^0), \tag{11}$$

where θ^* is between $\hat{\theta}$ and θ^0. For arbitrary θ the matrix $I(\theta) = -n^{-1} \partial^2 \log L_1/\partial\theta\,\partial\theta$ will be positive-definite, under slight restrictions on the z_{ij} values, since this matrix is also the observed information matrix from a prospective likelihood with disease probabilities given by $p_i(z)$ $(i = 0, \ldots, k)$. The expectation matrix $G(\theta) = E\{I(\theta)\}$, where E denotes expectation under θ^0, will be positive-definite under even weaker conditions to ensure that the distributions $\mathrm{pr}(z \mid D = i)$ $(i = 0, \ldots, k)$ do not degenerate. Expression (11) can be rewritten

$$n^{\frac{1}{2}}(\hat{\theta} - \theta^0) = I(\theta^*)^{-1} S(\theta^0), \tag{12}$$

where $S(\theta) = n^{-\frac{1}{2}} \partial \log L_1/\partial\theta$. Assuming that $n_i n^{-1} \to \rho_i > 0$ $(i = 0, \ldots, k)$ as $n \to \infty$, we may derive an asymptotic normal distribution for $n^{\frac{1}{2}}(\hat{\theta} - \theta^0)$. The derivation involves an application of the central limit theorem to $S(\theta^0)$ and an argument to show $I(\theta^*)$ to be a consistent estimator of $G(\theta^0)$. The results differ from classical results on the asymptotic distribution of the maximum likelihood estimator only through the fact that the contributions to the score statistic, $S(\theta^0)$, from the individual samples do not in general have mean zero. Correspondingly the variance matrix for $S(\theta^0)$ is not $G(\theta^0)$.

Consider first the asymptotic distribution of $S(\theta^0)$. Straightforward calculations using (7) and (9) give

$$E(\partial \log L_1/\partial\delta_i^0) = n_i - n(a_{i0} + \ldots + a_{ik}) = 0, \quad E(\partial \log L_1/\partial\beta_i^0) = nb_{i.} - n(b_{i0} + \ldots + b_{ik}) = 0,$$

where $a_{ij} = \int p_i(z)\, p_j(z)\, q(z)\, dz$, $b_{ij} = \int z^{\mathrm{T}} p_i(z)\, p_j(z)\, q(z)\, dz$ and $b_{i.} = b_{i0} + \ldots + b_{ik}$. It follows that $E\{S(\theta^0)\} = 0$ even though the contributions to this expectation from specific disease groups are nonzero. One can write

$$S(\theta^0) = \sum_{i=0}^{k} (n/n_i)^{-\frac{1}{2}} \left[n_i^{-\frac{1}{2}} \sum_{j=1}^{n_i} \{\partial \log p_i(z_{ij})/\partial\theta^0 - \mu_i^0\} \right] + n^{-\frac{1}{2}} \sum_{i=0}^{k} n_i \mu_i^0,$$

where μ_i^0 is the expectation, in the ith disease group, of $\partial \log p_i(z)/\partial\theta^0$. The central limit theorem can be applied to the terms in square brackets for $i = 0, \ldots, k$. Also $\Sigma n_i \mu_i^0 = 0$ since $E\{S(\theta^0)\} = 0$. It follows that $S(\theta^0)$ is asymptotically normal with mean 0 and variance matrix

$$\Sigma = E\{(\partial \log L_1/\partial\theta^0)\,(\partial \log L_1/\partial\theta^0)^{\mathrm{T}}\}.$$

The consistency of $\hat{\theta}$ as an estimator of θ^0 is shown in the Appendix.

The consistency of $I(\theta^*)$ as an estimator of $G(\theta)$ remains to be shown. The strong law of large numbers gives the almost sure convergence, under $\theta = \theta^0$, of $I(\theta')$ to its expectation $G(\theta')$ for any θ'. In fact it is possible to show the convergence to be uniform in a neighbourhood of θ^0. This along with the consistency of $\hat{\theta}$ and the fact that θ^*, defined in (11), lies between $\hat{\theta}$ and θ^0 implies the consistency of $I(\theta^*)$ as an estimator of $G(\theta)$, provided that this expectation matrix exists. A sufficient condition for such existence is $E(\|z_i\|^2) < \infty$ for each disease group $(i = 0, \ldots, k)$.

Returning to (12), one now has an asymptotic normal distribution for $n^{\frac{1}{2}}(\hat{\theta} - \theta^0)$ that has mean zero and variance matrix $G^{-1}\Sigma G^{-1}$, where $G = G(\theta^0)$. The remainder of this section is devoted to showing that the asymptotic variance matrix $(G^{-1}\Sigma G^{-1})_{22}$ corresponding to $\hat{\theta}_2^{\mathrm{T}} = (\hat{\beta}_1^{\mathrm{T}}, \ldots, \hat{\beta}_k^{\mathrm{T}})$ can be shown to equal $(G^{-1})_{22}$, where the subscripts (2, 2) indicate the lower right $kp \times kp$ submatrix.

Simple calculations give

$$E(-n^{-1}\, \partial^2 \log L_1/\partial\delta_i^0\, \partial\delta_j^0) = n^{-1}(\Delta_{ij} n_i - n a_{ij}),$$

where $\Delta_{ij} = 1$ or zero according to whether or not $i = j$. It follows that $G_{11} = N - A$ where $N = \mathrm{diag}\,(n_1 n^{-1}, \ldots, n_k n^{-1})$ and $A = ((a_{ij}))$ is the $k \times k$ matrix defined by $i, j = 1, \ldots, k$. Similarly

$$E\{n^{-1}(\partial \log L_1/\partial\delta_i^0)\,(\partial \log L_1/\partial\delta_j^0)\}$$
$$= a_{ij} - n \sum_{m=1}^{k} n_m^{-1} a_{im} a_{mj} - n n_0^{-1} \left(n_i n^{-1} - \sum_{l=1}^{k} a_{il} \right) \left(n_j n^{-1} - \sum_{r=1}^{k} a_{rj} \right).$$

In matrix notation this can be written

$$\begin{aligned} \Sigma_{11} &= A - AXA - (NXN - N) + (AXN - A) + (NXA - A) \\ &= N - A - (N - A)\, X (N - A) \\ &= G_{11} - G_{11} X G_{11}, \end{aligned}$$

where X is the $k \times k$ matrix with elements $x_{ij} = nn_0^{-1} + \Delta_{ij} nn_i^{-1}$. Very similar calculations show $\Sigma_{12} = G_{12} - G_{11} X G_{12}$, $\Sigma_{21} = G_{21} - G_{21} X G_{11}$ and $\Sigma_{22} = G_{22} - G_{21} X G_{12}$. Multiplication gives

$$G^{-1} \Sigma G^{-1} = G^{-1} - \begin{bmatrix} X & 0 \\ 0 & 0 \end{bmatrix},$$

so that $(G^{-1} \Sigma G^{-1})_{22} = (G^{-1})_{22}$ as required.

The asymptotic distribution $n^{\frac{1}{2}}(\hat{\theta}_2 - \theta_2^0)$, where $\hat{\theta}_2^{\mathrm{T}} = (\hat{\beta}_1^{\mathrm{T}}, \ldots, \hat{\beta}_k^{\mathrm{T}})$, then has mean zero and variance matrix $(G^{-1})_{22}$. As noted above this variance matrix will be consistently estimated by $\{I(\hat{\theta})^{-1}\}_{22}$. An asymptotic distribution for $n^{\frac{1}{2}}(\hat{\theta} - \theta^0)$ with mean zero and variance matrix $\{I(\hat{\theta})^{-1}\}_{22}$ is precisely the distributional statement that would arise if the prospective model (4) were directly applied to the case-control data, as if a prospective study had been conducted. This result indicates a very convenient computational procedure for adapting and fitting a logistic disease probability model to the type of case-control study described above.

5. Relationship to other work

The parameterization used in § 3 is basically that used by Anderson (1972). The above development attempts to clarify his work by defining new parameters $\delta_1, \ldots, \delta_k$ and $q(.)$ in terms of the induced model $\mathrm{pr}(z \mid D)$. This approach gives rise naturally to the constraints (9), whereas Anderson appears to have merely assumed that the same relationships would hold among $n_i n^{-1}$ $(i = 1, \ldots, k)$ and the new parameters as holds among the population quantities $\Pi_i = \mathrm{pr}(D = i)$ and the parameters of (4). It also makes clear that the population parameters Π_i need not be known (Farewell, 1979). The development of § 3 also simplifies Anderson's work by avoiding the Lagrange multiplier procedure to find the maximum likelihood estimators.

The most important distinction between this work and the earlier work of Anderson, however, arises from the fact that in §§ 3 and 4 no restriction whatever is required on the sample space for z. The regressor variable may then include continuous or mixed discrete and continuous components. In contrast Anderson suggested some arbitrary grouping on the regressor variable in order to make its sample space finite. Further, the development of § 4, using the $k+1$ sample structure, indicates that in most circumstances the asymptotic distribution of $\hat{\beta}_i$ $(i = 1, \ldots, k)$ will provide an adequate approximation to the actual sampling distribution provided each n_i $(i = 0, \ldots, k)$ is moderately large. Such an indication is not apparent from Anderson's work since his development utilized the joint asymptotic distribution of $\hat{\theta}$ and nuisance parameter estimators.

Another possible approach to model fitting using the likelihood (8) would involve specification of a parametric model $q(z) = q(z; \tau)$ with $\tau^{\mathrm{T}} = (\tau_1, \ldots, \tau_m)$. Since $q(z)$ can be directly estimated, the goodness of fit of such a specified model could be checked and, of course, standard constrained likelihood methods could be applied to the corresponding estimator of $(\theta^{\mathrm{T}}, \tau^{\mathrm{T}})$. Though the details are omitted here, such a procedure shows maximum likelihood estimators of the odds ratio parameters β_i $(i = 1, \ldots, k)$ to have precisely the same asymptotic distribution as that arising in § 4, regardless of the choice of parameterization. This means that assumption or knowledge on $q(.)$ does not aid in the estimation of the odds ratio parameters, at least in large samples.

6. More general designs and models

Suppose that the case-control design is relaxed to that of random sampling of cases and controls among subsets or strata of the defined population. Suppose that strata $s = s(x)$ are

defined in terms of auxiliary variables $x = (x_1, \ldots, x_q)$. The auxiliary variables will describe study subject characteristics at the beginning of the case accession period. The regressor variable $z = (z_1, \ldots, z_p)$ may be defined to include exposure and other variables as well as interaction terms between x and exposure variables.

A prospective logistic disease incidence model of the form (4) may be specified in each of t strata giving for $i = 0, \ldots, k$; $s = 1, \ldots, t$

$$\text{pr}(D = i \mid z, s) = \exp(\alpha_{is} + z\beta_i) \left\{ \sum_{l=0}^{k} \exp(\alpha_{ls} + z\beta_l) \right\}^{-1}. \tag{13}$$

The odds ratio in each stratum is again of the form $\exp(\{z - z_0\}\beta_i)$ $(i = 1, \ldots, k)$ and the induced retrospective logistic model can be written

$$\text{pr}(z \mid D = i, s) = c_{is} \exp\{\gamma_s(z) + z\beta_i\}. \tag{14}$$

Essentially the same arguments as in §§ 3 and 4 may be used to show the equivalence of $\hat{\beta}_i$ $(i = 1, \ldots, k)$ and the corresponding standard errors under application of (14) or under direct application of (13) to the case-control data. Stratum sizes must be allowed to become large as the overall sample size becomes large, with a nonzero limiting ratio, in order that the asymptotic likelihood theory apply.

Some case-control studies involve the assimilation of matched sets each containing a specified number of cases and controls. A model of the form (13) may again be considered with the stratum variable s distinguishing matched sets. A retrospective logistic model (14) will be induced for each matched set. One cannot, however, assume the same distribution theory will hold, at least unless the matched sets have reasonably large sample sizes. With small stratum sizes one may instead eliminate the 'nuisance' function $\gamma_s(z)$ in (14) by conditioning on the set of z values observed in that stratum (Prentice & Breslow, 1978; Breslow *et al.* 1978; Farewell, 1979). Suppose that n_0^s controls and n_i^s cases $(i = 1, \ldots, k)$ are selected in stratum or matched set s. Let z_{ij}^s $(i = 0, \ldots, k; j = 1, \ldots, n_i^s)$ be the corresponding regressor variables. The probability that the z values align themselves with disease groups as observed, given the set of observed z values, is easily shown to be

$$\exp\left(\sum_{i=1}^{k} \sum_{j=1}^{n_i^s} z_{ij}^s \beta_i \right) \Big/ \sum_{l \in R(n_*^s, \ldots, n_k^s)} \exp\left(\sum_{i=1}^{k} \sum_{j=1}^{n_i^s} z_{lij}^s \beta_i \right), \tag{15}$$

where $R(n_0^s, \ldots, n_k^s)$ represents the set of all divisions of $n^s = n_0^s + \ldots + n_k^s$ numbers into $(k+1)$ sets of size $n_0^s, \ldots, n_k^s$ and $l = (l_{01}, \ldots, l_{0n_0^s}, l_{11}, \ldots, l_{kn_k^s})$. Estimation based on likelihood factors of the form (15) is feasible computationally in circumstances in which at most one of the n_i^s $(i = 0, \ldots, k)$ is large. A combination of conditional inference on small strata and 'unconditional' inference as in §§ 3 and 4 for large strata seems reasonable though specific criteria would need to be developed.

Further insight into the relationship between the application of prospective and retrospective logistic models to case-control data can be obtained by modelling an auxiliary variable $x = (x_1, \ldots, x_q)$ rather than forming strata on the basis of x values. One may specify

$$\text{pr}(D = i \mid z, x) = \exp\{\alpha_i(x) + z\beta_i\} \Big/ \sum_{l=0}^{k} \exp\{\alpha_l(x) + z\beta_l\}. \tag{16}$$

Inversion of the model while conditioning on x gives

$$\text{pr}(z \mid D = i, x) = c_i(x) \exp\{\gamma(z, x) + z\beta_i\}. \tag{17}$$

The two models will be equivalent if $\alpha_i(x)$ $(i = 1, \ldots, k)$ in (16) and $\gamma(z, x)$ in (17) are unrestricted. As a special case one may suppose the sample space of x to be finite and $\alpha_i(x)$ to

be saturated with parameters. The above discussion on stratification, with each x value defining a stratum, then implies that the maximum likelihood estimators $\hat{\beta}_i$ $(i = 1, \ldots, k)$ from (17) and their asymptotic variances are the same as would be obtained by direct fitting of (16) to the case-control data. This result generalizes the finding of Breslow & Powers (1978) to situations in which the 'exposure' variable z need not be discrete and to situations in which k may exceed one. If the sample space for x is not finite it will usually be necessary to restrict $\alpha_i(x)$ or $\gamma(z, x)$ to yield useful estimation techniques. As Breslow & Powers suggest, one expects the two analyses to yield progressively more similar results as the modelling of $\alpha_i(x)$ and $\gamma(z, x)$ is permitted to approach saturation.

This work was supported by the National Institute of General Medical Sciences and by the National Cancer Institute. Helpful discussions with Norman Breslow, Vern Farewell and Jay Lubin are also acknowledged.

APPENDIX

Consistency of $\hat{\theta}$ as an estimator of θ^0

Since $L_1(\theta)$ is continuous and differentiable it takes a maximum on the closed sphere $\|\theta - \theta^0\| < \varepsilon$, for any $\varepsilon > 0$. If it can be shown that, for sufficiently large n, this maximum does not occur on the boundary with probability arbitrarily close to one, then one has a local maximum $\hat{\theta} = \hat{\theta}(n)$, within the sphere, for which $S(\hat{\theta}) = 0$. Standard techniques using a sequence of values of ε that converge to zero then permit the production of a consistent sequence of estimators $\hat{\theta}(n)$.

Define

$$H(\theta) = -\sum_{i=0}^{k} \rho_i E\{\log p_0(z_i;\, \theta)\},$$

where z_i is sampled from disease group $D = i$ and the dependence of $p_0(z) = 1/\Sigma \exp(\delta_l + z\beta_l)$ on θ has been made explicit by writing $p_0(z;\, \theta)$. The matrix $H''(\theta) = \partial^2 \log H(\theta)/\partial\theta\, \partial\theta$ will be positive-definite under slight conditions to ensure the nondegeneracy of the distributions (7); in fact $H''(\theta)$ differs from $G(\theta)$ only in the replacement of n_i/n by ρ_i $(i = 0, \ldots, k)$. For specified ε consider a boundary point $\tilde{\theta}$ such that $\|\tilde{\theta} - \theta^0\| = \varepsilon$. A Taylor expansion

$$H(\tilde{\theta}) = H(\theta^0) + (\tilde{\theta} - \theta^0)^{\mathrm{T}} H'(\theta^0) + (\tilde{\theta} - \theta^0)^{\mathrm{T}} H''(\theta^*)\, (\tilde{\theta} - \theta^0)$$

along with the positive-definiteness of $H''(\theta^*)$ implies

$$0 < \{\theta^{0\mathrm{T}} H'(\theta^0) - H(\theta^0)\} - \{\tilde{\theta}^{\mathrm{T}} H'(\theta^0) - H(\tilde{\theta})\}. \tag{A1}$$

The strong law of large numbers applied to each of the $k+1$ samples, along with the convergence of n_i/n to ρ_i, imply the almost sure convergence

$$\left\| n^{-1} \sum_{i=0}^{k} \sum_{j=1}^{n_i} \log p_0(z_{ij};\, \theta) + H(\theta) \right\| \to 0 \tag{A2}$$

for $\theta = \tilde{\theta}$ and for $\theta = \theta^0$. The fact that $E\{S(\theta^0)\} = 0$ implies that

$$n^{-1} E\left\{ \sum_{i=0}^{k} \sum_{j=1}^{n_i} \partial(\delta_i^0 + z_{ij}\beta_i^0)/\partial\theta^0 \right\} = -\sum_{i=0}^{k} (n_i/n)\, E\{\partial \log p_0(z_i;\, \theta^0)/\partial\theta^0\},$$

the right-hand side of which can be made arbitrarily close to $H'(\theta^0)$. The strong law of large numbers then gives the almost sure convergence

$$\left\| n^{-1} \sum_{i=0}^{k} \sum_{j=1}^{n_i} \partial(\delta_i^0 + z_{ij}\beta_i)/\partial\theta^0 - H'(\theta^0) \right\| \to 0. \tag{A3}$$

Substitution from (A2) and (A3) into (A1) then gives, with limiting probability one, $0 < L_1(\theta^0) - L_1(\tilde{\theta})$, so that the maximum does not occur on the boundary with limiting probability one, as required.

References

Aitchison, J. & Silvey, S. D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Ann. Math. Statist.* **29**, 813–28.

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

Anderson, J. A. (1973). Logistic discrimination with medical applications. In *Discriminant Analysis and Applications*, Ed. T. Cacoullos, pp. 1–15. New York: Academic Press.

Breslow, N. E., Day, N. E., Halvorsen, K. T., Prentice, R. L. & Sabai, C. (1978). Estimation of multiple relative risk functions in matched case-control studies. *Am. J. Epid.* **108**, 299–307.

Breslow, N. E. & Powers, W. (1978). Are there two logistic regressions for retrospective studies? *Biometrics* **34**, 100–5.

Farewell, V. T. (1979). Some results on the estimation of logistic models based on retrospective data. *Biometrika* **66**, 27–32.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Nat. Cancer Inst.* **22**, 719–48.

Miettinen, O. S. (1976). Estimability and estimation in case-referrent studies. *Am. J. Epid.* **104**, 226–35.

Prentice, R. L. (1976). Use of the logistic model in retrospective studies. *Biometrics* **32**, 599–606.

Prentice, R. L. & Breslow, N. E. (1978). Retrospective studies and failure time models. *Biometrika* **65**, 153–8.

[*Received May* 1978. *Revised May* 1979]