Biometrika (1995), **82**, 1, *pp.* 57–67
Printed in Great Britain

Logistic regression of family data from case-control studies

By ALICE S. WHITTEMORE

Department of Health Research and Policy, Stanford University School of Medicine, Stanford, California 94305, U.S.A.

Summary

Multivariate regression models are applied to binary disease data in families identified from case-control studies. Attention is restricted to 'marginal' or reproducible models, i.e. those whose parameters have the same interpretations in the marginal distributions for all subsets of a family, with a logistic specification of each individual's marginal disease probability. For such models, it is shown that the case-control family data can be analysed as if they were obtained from a prospective study, with the baseline disease probabilities of case and control probands differing from that of their relatives. This result extends that of Anderson (1972) and Prentice & Pyke (1979) for the probands' data to include disease outcomes and covariates for their families. It contrasts with inconsistent estimates of parameters in nonreproducible models that result when the case-control sampling design is ignored (Tosteson, Rosner & Redline, 1991). The contrast underscores the need to check the plausibility of the reproducibility assumption, which requires that the covariates be independent of any unmeasured factors responsible for the correlation of familial disease occurrence, before analysing case-control data prospectively.

Some key words: Case-control design; Logistic regression; Loglinear model; Marginal model; Ovarian cancer; Prospective design; Reproducibility.

1. Introduction

Family data from case-control studies are frequently used to study relationships between disease and environmental or genetic characteristics. Such a case-control study identifies a sample of diseased cases and an independent sample of disease-free controls, and for each identified individual, hereafter called a proband, determines his environmental covariates, his family structure, and the disease status and covariates of his relatives. A 'prospective' study, by contrast, identifies a sample of probands with specified covariates, and for each proband, determines his disease status, his family structure, and the disease status and covariates of his relatives. The case-control design is used for rare diseases because large numbers of probands with the disease can be ascertained economically.

For the case-control data shown in Table 1, the probands are women with and without a diagnosis of ovarian cancer. Each proband was asked whether her mother had been diagnosed with ovarian cancer, and about her own and her mother's reproductive histories. In this example, the proband's family consists only of her mother. The covariates are categories of parity, defined as a woman's total number of childbirths. Parity has been classified according to the three categories zero births, one–two births and three or more births. The scientific objectives are to determine: (i) the relationship between disease risk and parity, allowing for any correlation in disease between mothers and daughters; and

(ii) the familial association in disease risk due to genetic factors, after adjustment for possible mother-daughter similarities in parity.

Table 1. *Ovarian cancer and parity among 769 probands and their mothers*

		Number of probands			
		Case		Control	
Number of childbirths		Cancer in mother		Cancer in mother	
Mother	Proband	Yes	No	Yes	No
1–2	0	1	29	0	22
	1–2	1	36	0	33
	3+	0	41	0	38
3+	0	3	85	1	50
	1–2	1	105	0	103
	3+	1	84	0	135

Combined data from Casagrande et al. (1979) and Cramer et al. (1983).

The analysis of such data could proceed by fitting a model for the joint distribution of the entire family's disease and covariate data, conditional on the proband's disease status. Then quantities of interest, such as odds-ratios relating disease to covariates and correlation coefficients for disease among family members, could be estimated. However when several covariates are simultaneously under study, such modelling is likely to involve many parameters and to be cumbersome and vulnerable to model misspecification. A more tractable approach is to specify a model for the joint distribution of the family disease responses given all covariates, and use it to induce one for the distribution of the family disease and covariate data conditional on the proband's disease status. We apply this approach to a class of 'marginal' models (Liang, Zeger & Qaqish, 1992), wherein regression parameters measure associations between response and covariates in the marginal distributions of the family members. Specifically, we show that, for this class of models, odds-ratios relating disease to covariates and correlation coefficients relating disease occurrence among family members can be estimated as if the data were obtained from a prospective study. The only adjustment needed to accomodate the case-control design is to allow the logistic intercept parameter for the probands' marginal disease probabilities to differ from that of their relatives. This result extends the work of Anderson (1972) and Prentice & Pyke (1979) for logistic regression of the probands' data to include data for their families. It contrasts with findings, for a different class of loglinear models, that ignoring the case-control design gives inconsistent estimates of odds-ratios relating response in pairs of relatives (Tosteson et al., 1991). This distinction indicates the need to check the appropriateness of marginal models for the particular application before analysing case-control data prospectively.

2. Logistic regression of the probands' data

We begin with a brief review of logistic regression for the case and control probands themselves. Let y be an indicator for an individual's disease response, assuming the value one if he has the disease and zero otherwise, and let z be a column vector of covariates to be related to his disease risk. We assume

$$p_z \equiv \operatorname{pr}(y=1|z) = \frac{e^{\alpha+\beta z}}{1+e^{\alpha+\beta z}}. \tag{1}$$

As noted by Anderson (1972) and in a more general context by Prentice & Pyke (1979), (1) puts constraints on the distribution of z conditional on disease status. Specifically, Bayes' rule and (1) imply that

$$\frac{\operatorname{pr}(z|y=1)}{\operatorname{pr}(z|y=0)} = \frac{\operatorname{pr}(y=0)}{\operatorname{pr}(y=1)} e^{\alpha+\beta z} \equiv \frac{1-\eta}{\eta} e^{\alpha+\beta z}. \tag{2}$$

The marginal disease probability η cannot be estimated from case-control data; indeed, the proportions of diseased and disease-free probands are controlled by the investigator. It is therefore useful to regard the probands as members of a second, hypothetical population of individuals whose disease probability is given by another parameter π, namely the proportion of probands who are cases, but whose covariate distributions $\operatorname{pr}(z|y=1)$ and $\operatorname{pr}(z|y=0)$ still satisfy (2). In this population, from Bayes' rule,

$$\operatorname{pr}(y=1|z) \equiv P_z = \frac{\pi \operatorname{pr}(z|y=1)}{\pi \operatorname{pr}(z|y=1) + (1-\pi)\operatorname{pr}(z|y=0)}.$$

By using (2) in this expression we get

$$P_z = \frac{e^{\delta+\beta z}}{1+e^{\delta+\beta z}}, \tag{3}$$

where

$$\delta = \alpha + \log\left\{\frac{\pi(1-\eta)}{(1-\pi)\eta}\right\}.$$

Also in this population, the marginal density of the covariates is

$$\phi(z) = \pi \operatorname{pr}(z|y=1) + (1-\pi)\operatorname{pr}(z|y=0),$$

and

$$\begin{aligned} \operatorname{pr}(z|y=1) &= P_z\phi(z)/\pi, \\ \operatorname{pr}(z|y=0) &= (1-P_z)\phi(z)/(1-\pi). \end{aligned} \tag{4}$$

We note that each proband's contribution (4) to a likelihood function is proportional to $P_z^y(1-P_z)^{1-y}$, with P_z given by the logistic form (3). This is the contribution from a prospective study of the hypothetical population. Comparison of (1) and (3) shows that P_z and p_z differ only in their intercept parameters δ and α, and not in their regression coefficient β. This similarity suggests that β can be estimated by pretending that the probands were sampled from a prospective study of the hypothetical population. A complication arises because the quantities δ, α, β, and $\phi(.)$ are linked by the constraint

$$\int P_z\phi(z)\,dz = \pi \tag{5}$$

needed to ensure that (4) are probability distributions. Nevertheless Anderson (1972) and Prentice & Pyke (1979) verified that asymptotic inferences for β can proceed as if a prospective study had been conducted.

3. Extension to family data

We want to extend these results to data on both probands and their families. The objective is to infer relationships between covariates and disease in family members and to evaluate disease correlations among family members, conditional on their covariates. Suppose we have sampled case and control probands from independent families, and for each proband we have recorded his covariates, the size of his family and the disease response and covariates of each relative. The issue is to determine the circumstances under which such case-control data can be analysed as if they were collected from a prospective study.

To address this issue, we let $Y=(y_1, \ldots, y_m)$ and $Z=(z_1, \ldots, z_m)$ denote disease responses and covariates, respectively, for a family of size m, and consider models for the joint distribution $\operatorname{pr}(Y|Z)$. Two classes of models have been proposed for this problem. In both classes the natural parameters are modelled as functions of linear combinations of covariates and a regression parameter β. The classes differ in the roles played by the natural parameters, and thus in the interpretation of β. The regression parameters in the first class of 'marginal' models (Liang et al., 1992) measure associations between disease and covariates in the univariate marginals, i.e. for randomly selected individuals. An example of such models is the parametric family proposed by Bahadur (1961), discussed in § 5. In contrast, the regression parameters of the second class of log-linear models (Bishop, Feinberg & Holland, 1975) have interpretations in terms of conditional distributions of a subset of the responses, given the others. Use of these loglinear models for families of varying size m has been criticised on the grounds that the interpretation of β must vary with family size. This is because the components of β derive meaning only from the conditional distribution of an individual's response, given the responses of the other family members, and the number of binary variables conditioned on, from which β derives its meaning, depends on m. See Prentice (1988) for further discussion.

These considerations suggest that a model $\operatorname{pr}(Y|Z)$ for the analysis of family data should be reproducible, in the sense that its regression parameters have the same interpretation in the full model as in the marginal distributions obtained by summing over the responses of one or more individuals. In particular, the marginal distribution for a subset of r individuals $i_1, \ldots, i_r$ should depend on Z only through $z_{i_1}, \ldots, z_{i_r}$. When $r=1$ this condition is

$$\operatorname{pr}(y_i=1|Z)=\operatorname{pr}(y_i=1|z_i)\equiv p_{z_i}. \tag{6}$$

We shall see in the next section that condition (6), with the p_{z_i} given by the logistic model (1), is sufficient to extend the results of Anderson and Prentice & Pyke to family data.

Despite the desirability and mathematical convenience of the reproducibility assumption, it is important to recognise that it implies a certain type of independence between the covariates and the sources of disease correlation in the family. Therefore there is need to examine the plausibility of this independence for a given application. Specifically, suppose the disease correlation is due to a latent variable $U=(u_1, \ldots, u_m)$ representing unmeasured genetic or other factors in the family. That is, the family members' responses are conditionally independent, given U and Z, so that

$$\operatorname{pr}(Y|Z)=\sum_U \operatorname{pr}(U|Z)\prod_{i=1}^m \operatorname{pr}(y_i|u_i, z_i). \tag{7}$$

In general, summing (7) over $(y_2, \ldots, y_m)$ will not yield (6). It will yield (6) if $\operatorname{pr}(U|Z)=\operatorname{pr}(U)$, that is, if the covariates Z are independent of the source U of disease correlation.

For the data in Table 1, for example, consider hypothesis A: unmeasured genetic factors strongly affect both ovarian cancer risk and fertility. Under this hypothesis, parity would not be independent of the latent genetic factors, and indeed, the marginal probability of ovarian cancer for a woman would depend not only on her own parity but also on that of her mother, in violation of (6). In fact, hypothesis A is implausible because parity is determined largely by choice and is therefore a poor surrogate for fertility. In the present application then, the independence assumption (6) seems plausible.

There are many other applications in which assumption (6) is reasonable. It is implicit in all studies of personal environmental characteristics in relation to diseases having a familial component, e.g. personal occupational exposures in relation to lung cancer, or personal reproductive characteristics in relation to breast cancer. Yet in some situations, such as that of hypothesis A, unmeasured disease-causing factors, e.g. genetic factors or unmeasured dietary factors, may be strongly correlated within families and with the measured covariates. For such situations, assumption (6) would be violated, and indeed summary measures of disease-covariate association for individual family members would be inappropriate.

4. Maximum likelihood estimates

We assume that $\mathrm{pr}\,(Y|Z)$ satisfies (6), where the marginal disease probabilities p_{z_i} are given by the logistic function (1). Examples of such models are discussed in the next section. A prospective study involves sampling from

$$\mathrm{pr}\,(Y, Z_{-1}|z_1) = \mathrm{pr}\,(Y|Z)\,\mathrm{pr}\,(Z_{-1}|z_1),$$

where the subscript 1 indexes the proband and the subscript -1 indicates a vector with its first component deleted. Using equation (6) we may write $\mathrm{pr}\,(Y|Z) = \mathrm{pr}\,(y_1|z_1)\,\mathrm{pr}\,(Y_{-1}|y_1, Z)$. Hence

$$\mathrm{pr}\,(Y, Z_{-1}|z_1) = \mathrm{pr}\,(y_1|z_1)\,\mathrm{pr}\,(Y_{-1}|y_1, Z)\,\mathrm{pr}\,(Z_{-1}|z_1).$$

So the prospective likelihood based on a sample of n independent vectors $(Y_j, Z_j) = (y_{j1}, \ldots, y_{jm_j}, z_{j1}, \ldots, z_{jm_j})$ $(j = 1, \ldots, n)$ is

$$\begin{aligned} L_P &= \prod_{j=1}^{n} \mathrm{pr}\,(Y_j, Z_{j,-1}|z_{j1}) \\ &= \left\{\prod_{j=1}^{n} \mathrm{pr}\,(y_{j1}|z_{j1})\right\} \left\{\prod_{j=1}^{n} \mathrm{pr}\,(Y_{j,-1}|Z_j, y_{j1})\right\} \left\{\prod_{j=1}^{n} \mathrm{pr}\,(Z_{j,-1}|z_{j1})\right\} \\ &= L_P^{(1)}(\beta, \alpha) \times L_P^{(2)}(\beta, \alpha, \rho) \times \mathrm{const}, \end{aligned} \tag{8}$$

where ρ denotes additional parameters in $\mathrm{pr}\,(Y|Z)$ and the constant is independent of $\theta = (\beta, \alpha, \rho)$. Under our assumptions,

$$L_P^{(1)}(\beta, \alpha) = \prod_{j-1}^{n} p_{z_{j1}}^{y_{j1}} q_{z_{j1}}^{1-y_{j1}},$$

with $p_z = 1 - q_z$ given by (1).

In comparison, the case-control study involves two separate samples, one from $\mathrm{pr}\,(Y_{-1}, Z|y_1 = 1)$ and one from $\mathrm{pr}\,(Y_{-1}, Z|y_1 = 0)$. Equation (6) implies that y_1 and Z_{-1} are conditionally independent, given z_1:

$$\mathrm{pr}\,(Z|y_1) = \mathrm{pr}\,(z_1|y_1)\,\mathrm{pr}\,(Z_{-1}|z_1).$$

Thus

$$\mathrm{pr}\,(Y_{-1}, Z|y_1) = \mathrm{pr}\,(Y_{-1}|Z, y_1)\,\mathrm{pr}\,(Z|y_1) = \mathrm{pr}\,(z_1|y_1)\,\mathrm{pr}\,(Y_{-1}|Z, y_1)\,\mathrm{pr}\,(Z_{-1}|z_1).$$

So the retrospective likelihood, based on a sample of n_1 case vectors

$$(Y_j, Z_j) = (y_{j1}, \ldots, y_{jm_j}, z_{j1}, \ldots, z_{jm_j}) \quad (j = 1, \ldots, n_1),$$

and an independent sample of n_0 control vectors

$$(Y_j, Z_j) = (y_{j1}, \ldots, y_{jm_j}, z_{j1}, \ldots, z_{jm_j}) \quad (j = n_1 + 1, \ldots, n_1 + n_0 = n)$$

is

$$\begin{aligned} L_R &= \prod_{j=1}^{n} \mathrm{pr}\,(Y_{j,-1}, Z_j|y_{j1}) \\ &= \left\{\prod_{j=1}^{n} \mathrm{pr}\,(z_{j1}|y_{j1})\right\}\left\{\prod_{j=1}^{n} \mathrm{pr}\,(Y_{j,-1}|Z_j, y_{j1})\right\}\left\{\prod_{j=1}^{n} \mathrm{pr}\,(Z_{j,-1}|z_{j1})\right\} \\ &= L_R^{(1)}(\delta, \beta) \times L_P^{(2)}(\beta, \alpha, \rho) \times L_R^{(3)} \times \text{const.} \end{aligned} \tag{9}$$

Here we have used (4) for the factors $\mathrm{pr}\,(z_{j1}|y_{j1})$, so that

$$L_R^{(1)}(\delta, \beta) = \prod_{j=1}^{n} P_{z_{j1}}^{y_{j1}}(1 - P_{z_{rj1}})^{1-y_{j1}},$$

with P_z given by (3),

$$L_R^{(3)} = \prod_{j=1}^{n} \phi(z_{j1}),$$

and the factors $\pi^{y_{j1}}(1-\pi)^{1-y_{j1}}$ have been absorbed in the constant.

We want to maximise $\log L_R$ with respect to $\vartheta = (\delta, \theta) = (\delta, \beta, \alpha, \rho)$ and $\phi(.)$, subject to the constraint (5). Because of the way (9) factors, the maximum likelihood estimate of $\phi(.)$ puts mass $1/n$ at each observed value of z_{j1} $(j = 1, \ldots, n)$. Thus (5) becomes

$$\sum_{j=1}^{n} P_{z_{j1}} = n_1, \tag{10}$$

where we have used $\pi = n_1/n$. Consider now the unconstrained maximum likelihood estimate $\hat{\vartheta}$ of ϑ. The score equations corresponding to (9) with $\phi(.) = \hat{\phi}(.)$ are

$$0 = \partial \log L_R/\partial\delta = n_1 - \sum_{j=1}^{n} P_{z_{j1}}, \quad 0 = \partial \log L_R/\partial\theta. \tag{11}$$

By virtue of the first equation in (11), $\hat{\vartheta}$ automatically satisfies the constraint (10), and therefore it is the desired constrained maximum likelihood estimate. Comparison of (9) with (8) shows that $\hat{\vartheta}$ is precisely the maximum of a prospective likelihood L_P^* that differs from L_P only by its inclusion of an additional parameter δ for the intercept in the marginal disease probabilities of the probands.

The Appendix shows that the asymptotic distribution of $n^{\frac{1}{2}}(\hat{\theta} - \theta)$ is normal with mean zero and with variance estimated consistently by the following procedure. Fit the model for $\mathrm{pr}\,(Y|Z)$, with different baseline marginal probabilities for probands and their relatives, directly to the data as if a prospective study had been conducted. Then delete the first row and column of the inverse of the resulting observed information matrix.

5. Example

To analyse the data in Table 1 we consider the class of models proposed by Bahadur (1961). For $m = 2$, the models are of the form

$$\operatorname{pr}(Y|Z) = \left(\prod_{i=1}^{2} p_i^{y_i} q_i^{1-y_i} \right)(1 + \rho t_1 t_2), \tag{12}$$

where $t_i = (y_i - p_i)(p_i q_i)^{-\frac{1}{2}}$ is the ith standardised response, $i = 1, 2$. The correlation coefficient ρ satisfies the well-known constraints

$$-\min\{(w_1 w_2)^{\frac{1}{2}}, (w_1 w_2)^{-\frac{1}{2}}\} \leqslant \rho \leqslant \min\{(w_1 w_2^{-1})^{\frac{1}{2}}, (w_2 w_1^{-1})^{\frac{1}{2}}\}, \tag{13}$$

with $w_i = p_i/q_i$ $(i = 1, 2)$. Model specification proceeds by indicating how ρ and the marginal response probabilities p_i depend on Z. A model is reproducible if each p_i depends only on z_i.

We fit model (12) to the data in Table 1, with $p_i = p_{z_i}$ given by (1) and with various assumptions for ρ. A woman's covariate vector z consists of two indicators: PAR1, which equals one if she had one or two births and zero otherwise; and PAR2, which equals one if she is nulliparous and zero otherwise. Thus e^{β_1} and e^{β_2} represent, respectively, odds-ratios relating risk in women with one or two and no childbirths to that of women with three or more births.

Table 2 shows parameter estimates and maximised loglikelihoods for four models. Models I and III, with ρ fixed at zero, assume that all women's data are independent. The estimates for α and δ in Model I are simply the logits of overall disease prevalence in mothers and probands, respectively. Model III assumes common β's for mothers and daughters, but allows separate intercept parameters. Models II and IV extend Models I and III respectively, by allowing a common nonzero correlation coefficient for all mother-daughter pairs. Models II and IV assume that familial correlation in disease risk due to genetic factors is independent of the women's reproductive histories.

Table 2. *Parameter estimates for data of Table* 1

Model	α	δ	β_1	β_2	ρ	Likelihood
I	−4·555	0·0130	—	—	—	577·5
II	−4·554	0·0148	—	—	0·0726	575·5
III	−4·658	−0·3079	0·3485	0·7881	—	568·5
IV	−4·656	−0·2989	0·3401	0·7582	0·0371	567·6

Comparison of Models III and IV indicates that adjustment for familial correlation in disease risk has little effect on the coefficient estimates associated with parity. In contrast, comparison of Models II and IV shows that adjustment for parity has reduced $\hat{\rho}$ from 0·0726 to 0·0371, and reduced the likelihood ratio statistic testing $\rho = 0$ from 4·0 ($p < 0{\cdot}05$) to 1·8 ($p = 0{\cdot}15$). For both Models II and IV, $\hat{\rho}$ lies in the interior of the interval defined by the constraints (13). The similarity in estimates of β from Models III and IV supports a suggestion of McDonald (1993) for regression of bivariate binary data when the correlation ρ is a nuisance: estimate β from Model III, replacing the incorrect binomial variance estimate for $\hat{\beta}$ with the empirical variance estimate suggested by Liang & Zeger (1986).

In summary, the data suggest that low parity is positively associated with ovarian cancer risk, and that, after adjustment for parity, there is little evidence for familial correlation in disease risk.

6. Discussion

The preceding result extends to studies with $K \geqslant 1$ different disease responses, with the multinomial marginal disease probabilities now specified by a polytomous logistic regression model. Indeed, the work of Anderson (1972) and Prentice & Pyke (1979) is set in this more general context. However for correlated family data the number of parameters increases rapidly with K; even a minimal model allowing only a single nonzero correlation coefficient for each pair of diseases among pairs of relatives has $2K$ intercept parameters, sK regression coefficients where s is the number of covariates, and $K(K+1)/2$ correlation coefficients. Therefore such extension is likely to be feasible only for data from very large studies.

The result also extends to the more general designs discussed by Prentice & Pyke, whereby stratified samples are obtained from the relevant populations, with stratification based on the values of one or more ancillary variables. For example in matched case-control studies probands are sampled in matched sets, each containing a specified number of cases and controls. Once a proband has been sampled, his family's disease and covariate data are then gathered. With this design, inferences for β could be based on a hybrid likelihood, similar to the product $L_R^{(1)} L_P^{(2)}$ of (9), but with the factor $L_R^{(1)}$ replaced by a conditional logistic likelihood component that depends only on β. As in ordinary logistic regression, the conditional likelihood component arises by considering a conditional probability in each stratum, with the conditioning taken on the set of observed covariates for all probands in the stratum.

Tosteson et al. (1991) considered loglinear models for multivariate binary data whose association parameters are the pairwise odds-ratios

$$\frac{\operatorname{pr}(y_i = y_j = 1)\operatorname{pr}(y_i = y_j = 0)}{\operatorname{pr}(y_i = 1, y_j = 0)\operatorname{pr}(y_i = 0, y_j = 1)},$$

where the probabilities are conditioned on the responses of the other family members. They showed for $\beta = 0$ that these odds-ratios are estimated inconsistently when the data are analysed as if they had been sampled prospectively. Because the natural parameters in the models considered by Tosteson et al. (1991) specify conditional rather than marginal probabilities, their prospective analysis did not allow the probands' marginal response probability to differ from that of their relatives. In applications for which the independence assumption (6) is appropriate, consistent estimates for conditional or marginal pairwise odds-ratios, or other quantities of interest, can be obtained by estimating ϑ using the methods of § 5, and then calculating the quantities using the joint distribution (12) with $\theta = \hat{\theta}$.

For prospective data, quasilikelihood estimation of parameters in the first- and possibly second-order marginals of the joint distribution of family disease responses given family covariates provides an alternative to likelihood-based inference based on a full parametric model. The quasilikelihood approach, e.g. Liang et al. (1991), Prentice (1988), avoids misspecification of the joint distribution of the data, and can entail little loss of efficiency. However its application to case-control data has been unclear. Some investigators have simply ignored the probands' data and analysed instead only that of their relatives, modelling the latter's disease probabilities and pairwise odds-ratios in terms of their own covariates, the proband's disease status, and possibly the proband's covariates, e.g., Liang et al. (1992), Zhao & Le Marchand (1992). This strategy loses efficiency by ignoring the dependence beween the probands' disease status and the probands' covariates, a loss which can be severe for studies in which the most reliable and complete covariate data pertain to the probands. In Table 1 for example, odds-ratios relating ovarian cancer to nulliparity are inestimable from the mothers' data, since these women are all parous by definition. In applications for which independence assumption (6) is plausible, the present result suggests a strategy for quasilikelihood estimation from case-control data: construct quasilikelihood equations based on marginals of a full prospective model, accomodating the case-control design simply by allowing the intercept parameter for the probands' disease probabilities to differ from that of their relatives.

Acknowledgement

This work was supported by an Outstanding Investigator Grant from the U.S. National Institutes of Health. The author thanks Joseph B. Keller for helpful discussion, Jerry Halpern for computing assistance, and an Associate Editor for constructive comments.

Appendix

Asymptotic distribution of $\hat{\vartheta}$

We describe briefly the asymptotic distribution of the maximum likelihood estimate $\hat{\vartheta}$ under the two-sample case-control design, assuming that $n_1 n^{-1} \to \pi > 0$ as $n \to \infty$.

A first-order Taylor expansion of $\hat{\vartheta}$ about the true value $\vartheta_0 = (\delta_0, \theta_0)$ gives

$$n^{\frac{1}{2}}(\hat{\vartheta} - \vartheta_0) = I(\vartheta^*)^{-1} U(\vartheta_0), \tag{A1}$$

where

$$U(\vartheta) = n^{-\frac{1}{2}} \partial \log L_R / \partial \vartheta, \quad I(\vartheta) = -n^{-1} \partial^2 \log L_R / \partial \vartheta \, \partial \vartheta, \quad \| \vartheta^* - \vartheta_0 \| < \| \hat{\vartheta} - \vartheta_0 \|.$$

For arbitrary ϑ the information matrix $I(\vartheta)$ and its expectation $G(\vartheta) = E\{I(\vartheta)\}$ are positive definite under mild restrictions on the z values (Prentice & Pyke, 1979).

The first objective is to show that $E\{U(\vartheta_0)\} = 0$. To do so, we write

$$U(\vartheta) = U_1(\vartheta) + U_2(\vartheta), \tag{A2}$$

where, setting $l_1 = \log L_R^{(1)}$ and $l_2 = \log L_P^{(2)}$,

$$U_1(\vartheta) = n^{-\frac{1}{2}} \partial l_1 / \partial \vartheta = n^{-\frac{1}{2}} \begin{pmatrix} \partial l_1 / \partial \delta \\ \partial l_1 / \partial \beta \\ 0 \\ 0 \end{pmatrix},$$

$$U_2(\vartheta) = n^{-\frac{1}{2}} \partial l_2 / \partial \vartheta = n^{-\frac{1}{2}} \begin{pmatrix} 0 \\ \partial l_2 / \partial \beta \\ \partial l_2 / \partial \alpha \\ \partial l_2 / \partial \rho \end{pmatrix}.$$

A straightforward extension of the arguments of Prentice & Pyke shows that $E(\partial l_1 / \partial \delta) = 0$ and $E(\partial l_1 / \partial \beta) = 0$, where the expectation is evaluated at $\vartheta = \vartheta_0$. Thus, from (A2), $E\{U(\vartheta_0)\} = E\{U_2(\vartheta_0)\}$, and we need only show that $E\{U_2(\vartheta_0)\} = 0$. To do so, we note that, for any scalar function $a(y_1, z_1)$,

$$\begin{aligned} &E_{\mathrm{pr}(Y_{-1},Z|y_1)}\{a(y_1, z_1)\ \partial \log \mathrm{pr}\,(Y_{-1}|Z, y_1)/\partial\vartheta\} \\ &\qquad = E_{\mathrm{pr}(Z|y_1)}[a(y_1, z_1)E_{\mathrm{pr}(Y_{-1}|Z,y_1)}\{\partial \log \mathrm{pr}\,(Y_{-1}|Z, y_1)/\partial\vartheta\}] \\ &\qquad = E_{\mathrm{pr}(Z|y_1)}\{a(y_1, z_1)\cdot 0\} = 0, \end{aligned} \tag{A3}$$

where $E_{\mathrm{pr}(.)}(.)$ denotes expectation with respect to pr (.). In particular, taking expectations of each summand of $U_2(\vartheta_0)$ and using (A3) with $a \equiv 1$ gives $E\{U_2(\vartheta_0)\} = 0$ as required.

The likelihood theory outlined by Prentice & Pyke can now be used to show that $U(\vartheta_0)$ is asymptotically normal with variance $V = E\{U(\vartheta_0)U(\vartheta_0)^{\mathrm{T}}\}$, that $\hat\vartheta$ is consistent for ϑ_0, and that $I(\hat\vartheta)$ is consistent for $G(\vartheta)$. Thus it follows from (A1) that $n^{\frac{1}{2}}(\hat\vartheta - \vartheta_0)$ is asymptotically normal with mean zero and variance $G^{-1}VG^{-1}$.

It remains to show that the asymptotic variance $(G^{-1}VG^{-1})_{22}$ of $\hat\theta$ equals $(G^{-1})_{22}$, where $\hat\theta = (\hat\beta, \hat\alpha, \hat\rho)$, and the subscript 22 denotes the submatrix obtained by deleting the first row and column of the given matrix. In analogy with (A2), we write

$$G = G_1 + G_2,$$

where

$$G_1 = \begin{pmatrix} G_{1;11} & G_{1;12} \\ G_{1;12}^{\mathrm{T}} & G_{1;22} \end{pmatrix},$$

with

$$G_{1;11} = -n^{-1}E(\partial^2 l_1/\partial\delta\, \partial\delta), \quad G_{1;12} = -n^{-1}E\{(\partial^2 l_1/\partial\delta\, \partial\beta, 0, 0)\},$$

$$G_{1;22} = -n^{-1}E\begin{pmatrix} \partial^2 l_1/\partial\beta\, \partial\beta & 0 \\ 0 & 0 \end{pmatrix}, \quad G_2 = E\begin{pmatrix} 0 & 0 \\ 0 & \partial^2 l_2/\partial\theta\, \partial\theta \end{pmatrix}.$$

Also, we use (A2) to write

$$UU^{\mathrm{T}} = U_1U_1^{\mathrm{T}} + U_1U_2^{\mathrm{T}} + (U_1U_2^{\mathrm{T}})^{\mathrm{T}} + U_2U_2^{\mathrm{T}}. \tag{A4}$$

Expanding $U_1U_2^{\mathrm{T}}$, taking expectations of the summands, and using (A3) with $a(y_1, z_1) = \log(P_{z_1}^{y_1}Q_{z_1}^{1-y_1})$, gives $E(U_1U_2^{\mathrm{T}}) = 0$. Thus

$$V = E(UU^{\mathrm{T}}) = E(U_1U_1^{\mathrm{T}}) + E(U_2U_2^{\mathrm{T}}) \equiv V_1 + V_2.$$

Straightforward calculations, outlined by Prentice & Pyke, show that

$$V_1 = G_1 - x(G_{1;11}, G_{1;12})^{\mathrm{T}}(G_{1;11}, G_{1;12}),$$

where $x = (n/n_0) + (n/n_1)$. Moreover the usual likelihood theory applied to the likelihood component $L_P^{(2)}$ for the relatives' data implies that $V_2 = G_2$. Thus $V = G - x(G_{1;11}, G_{1;12})^{\mathrm{T}}(G_{1;11}, G_{1;12})$. Multiplication gives

$$G^{-1}VG^{-1} = G^{-1} - \begin{pmatrix} x & 0 \\ 0 & 0 \end{pmatrix},$$

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

In summary, the asymptotic distribution of $n^{\frac{1}{2}}(\hat\theta - \theta_0)$ is normal with mean zero and variance $(G^{-1})_{22}$, which is estimated consistently by $\{I(\theta)^{-1}\}_{22}$. An asymptotic normal distribution for $n^{\frac{1}{2}}(\hat\theta - \theta_0)$ with mean zero and variance estimated by $\{I(\theta)^{-1}\}_{22}$ is just the result obtained if the model for pr $(Y|Z)$, with different baseline marginal probabilities for probands and their relatives, were applied directly to the data as if a prospective study had been conducted.

References

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

BAHADUR, R. R. (1961). A representation of the joint distribution of responses to *n* dichotomous outcomes. In *Studies in Item Analysis and Prediction*, Stanford Mathematical Studies in the Social Sciences IV, Ed. H. Solomon, pp. 158–68. Stanford University Press.

BISHOP, Y. M. M., FIENBERG, S. E. & HOLLAND, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: Massachusetts Institute of Technology Press.

CASAGRANDE, J. T., LOUIE, E. W., PIKE, M. C., ROY, S., ROSS, R. K. & HENDERSON, B. E. (1979). 'Incessant ovulation' and ovarian cancer. *Lancet* **2**, 170–3.

CRAMER, D. W., HUTCHISON, G. B., WELCH, W. R., SCULLY, R. E. & RYAN, K. J. (1983). Determinants of ovarian cancer risk. I. Reproductive experiences and family history. *J. Nat. Cancer Inst.* **71**, 711–6.

LIANG, K.-Y. & ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

LIANG, K.-Y., ZEGER, S. L. & QAQISH, B. (1992). Multivariate regression analysis for categorical data. *J. R. Statist. Soc.* B **54**, 3–40.

MCDONALD, B. W. (1993). Estimating logistic regression parameters for bivariate binary data. *J. R. Statist. Soc.* B **55**, 391–7.

PRENTICE, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033–48.

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

TOSTESON, T. D., ROSNER, B. & REDLINE, S. (1991). Logistic regression for clustered binary data in proband studies with application to familial aggregation of sleep disorders. *Biometrics* **47**, 1257–65.

ZHAO, L. P. & LE MARCHAND, L. (1992). An analytical method for assessing patterns of familial aggregation in case-control studies. *Genet. Epidem.* **9**, 141–54.

[*Received March* 1993. *Revised July* 1994]