CHAPTER VII

CONDITIONAL LOGISTIC REGRESSION FOR MATCHED SETS

One of the methods for estimating the relative risk parameters $\boldsymbol{\beta}$ in the stratified logistic regression model was conditioning (§ 6.3). We supposed that for a given stratum composed of n_1 cases and n_0 controls we knew the unordered values $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of the exposures for the $n = n_1 + n_0$ subjects, but did not know which values were associated with the cases and which with the controls. The conditional probability of the observed data was calculated (6.15) to be a product of terms of the form

$$\frac{\prod_{j=1}^{n_1} \exp(\sum_{k=1}^{K} \beta_k x_{jk})}{\sum_{\mathbf{l}} \prod_{j=1}^{n_1} \exp(\sum_{k=1}^{K} \beta_k x_{l_j k})}, \tag{7.1}$$

where $\mathbf{l}$ ranged over the $\binom{n}{n_1}$ choices of n_1 integers from among the set $\{1,2, \ldots, n\}$.

With a single binary exposure variable x, coded x = 1 for exposed and x = 0 for unexposed, knowing the unordered x's meant knowing the total number exposed in the stratum, and thus knowing all the marginal totals in the corresponding 2 × 2 table. The complete data were then determined by the number of exposed cases. In these circumstances the conditional probability (7.1) is proportional to the hypergeometric distribution (4.2), used as a starting point for exact statistical inference about the odds ratio in a 2 × 2 table.

The conditional likelihood offers important conceptual advantages as a basis for statistical analysis of the results of a case-control study. First, it depends only on the relative risk parameters of interest and thus allows for construction of exact tests and estimates such as were described in Chapters 4 and 5 for selected problems. Second, precisely the same (conditional) likelihood is obtained whether we regard the data as arising from either (i) a prospective study of n individuals with a given set of exposures $\mathbf{x}_1, \ldots, \mathbf{x}_n$, the conditioning event being the observed number n_1 of cases arising in the sample; or (ii) a case-control study involving n_1 cases and n_0 controls, the conditioning event being the n observed exposure histories. The observation that these two conditional likelihoods agree, which was made in § 4.2 for the 2 × 2 table, confirms the fundamental point that identical methods of analysis are used whether the data have been gathered according to prospective or retrospective sampling plans.

Unfortunately, whenever the strata contain sizeable numbers of both cases and

7. CONDITIONAL LOGISTIC REGRESSION FOR MATCHED SETS

controls, the calculations required for the conditional analysis are extremely costly if not actually impossible even using large computers. Since the analysis based on the unconditional likelihood (6.12) yields essentially equivalent results, it would seem to be the method of choice in such circumstances. The conditional approach is best restricted to matched case-control designs, or to similar situations involving very fine stratification, where its use is in fact essential in order to avoid biased estimates of relative risk. We begin this chapter with an illustration of the magnitude of the bias which arises from analysing matched data with the unconditional model. Next, the conditional model is examined for several of the special problems considered in Chapters 4 and 5; many of the estimates and test statistics discussed earlier for these problems are shown to result from application of the general model. Finally, we explore the full potential of the conditional model for the multivariate analysis of matched data, largely by means of example, and discuss some of the issues which arise in its implementation.

7.1 Bias arising from the unconditional analysis of matched data

Use of the unconditional regression model (6.12) for estimation of relative risks entails explicit estimation of the α stratum parameters in addition to the β coefficients of primary interest. For matched or finely stratified data, the number of α parameters may be of the same order of magnitude as the number of observations and much greater than the number of β's. In such situations, involving a large number of nuisance parameters, it is well known that the usual techniques of likelihood inference can yield seriously biased estimates (Cox & Hinkley, 1974, p. 292). This phenomenon is perhaps best illustrated for the case of 1–1 pair matching with a single binary exposure variable x.

Returning to the general set-up of § 6.2, suppose that each of the I strata consists of a matched case-control pair and that each subject has been classified as exposed $(x = 1)$ or unexposed $(x = 0)$. The outcome for each pair may be represented in the form of a 2×2 table, of which there are four possible configurations, as shown in (5.1). The model to be fitted is of the form

$$\mathrm{pr}_i(y = 1 \mid x) = \frac{\exp(\alpha_i + \beta x)}{1 + \exp(\alpha_i + \beta x)},$$

where $\beta = \log \psi$ is the logarithm of the relative risk, assumed constant across matched sets.

According to a well-known theory developed for logistic or log-linear models (Fienberg, 1977), unconditional maximum likelihood estimates (MLEs) for the parameters α and β are found by fitting frequencies to all cells in the $2 \times 2 \times K$ dimensional configuration such that (i) the fitted frequencies satisfy the model and (ii) their totals agree with the observed totals for each of the two dimensional marginal tables. For the n_{00} concordant pairs in which neither case nor control is exposed, and the n_{11} concordant pairs in which both are exposed, the zeros in the margin require that the fitted frequencies be exactly as observed. Such tables provide no information about the relative risk since, whatever the value of β, the nuisance parameter α_i may be chosen so that fitted and observed frequencies are identical ($\alpha_i = 0$ for tables of the first type and $\alpha_i = -\beta$ for tables of the latter to give probability $^1/_2$ of being a case or control).

The remaining $n_{10} + n_{01}$ discordant pairs have the same marginal configuration, and for these the fitted frequencies are of the form

	Exposure		
	+	−	
Case	μ	$1-\mu$	1
Control	$1-\mu$	μ	1
	1	1	2

where

$$\mu = \mathrm{pr}_i(y = 1 \mid x = 1) = \frac{\exp(\alpha_i + \beta)}{1 + \exp(\alpha_i + \beta)}$$

and

$$1-\mu = \mathrm{pr}_i(y = 1 \mid x = 0) = \frac{\exp(\alpha_i)}{1 + \exp(\alpha_i)},$$

which can be expressed as

$$\psi = \exp(\beta) = \left(\frac{\mu}{1-\mu}\right)^2.$$

The additional constraint satisfied by the fitted frequencies is that the total number of exposed cases, $n_{10} + n_{11}$, must equal the total of the fitted values, namely $(n_{10} + n_{01})\mu + n_{11}$. This implies $\hat{\mu} = n_{10}/(n_{10} + n_{01})$ and thus that the unconditional MLE of the relative risk is

$$\hat{\psi} = \left(\frac{\hat{\mu}}{1-\hat{\mu}}\right)^2 = \left(\frac{n_{10}}{n_{01}}\right)^2,$$

the square of the ratio of discordant pairs (Andersen, 1973, p. 69).

The estimate based on the more appropriate conditional model has already been presented in § 5.2. There we noted that the distribution of n_{10} given the total $n_{10} + n_{01}$ of discordant pairs was binomial with parameter $\pi = \psi/(1+\psi)$. It followed that the conditional MLE was the simple ratio of discordant pairs

$$\hat{\psi} = \frac{n_{10}}{n_{01}}.$$

Thus the *unconditional analysis of matched pair data results in an estimate of the odds ratio which is the square of the correct, conditional one:* a relative risk of 2 will tend to be estimated as 4 by this approach, and that of $^1/_2$ by $^1/_4$.

While the disparity between conditional and unconditional analyses is particularly dramatic for matched pairs, it persists even with other types of fine stratification. Pike, Hill and Smith (1979) have investigated by numerical means the extent of the bias in unconditional estimates obtained from a large number of strata, each having a fixed number of cases and controls. Except for matched pairs, the bias depends slightly on the proportion of the control population which is exposed, as well as on the true odds ratio. Table 7.1 presents an extension of their results. For sets having 2 cases and 2 controls each, a true odds ratio of 2 tends to be estimated in the range from 2.51 to 2.53, depending upon whether the exposure probability for controls is 0.1 or 0.3. Even with 10 cases and 10 controls per set, an asymptotic bias of approximately 4% remains for estimating a true odds ratio of $\psi = 2$, and of about 15% for estimating $\psi = 10$.

These calculations demonstrate the need for considerable caution in fitting unconditional logistic regression equations containing many strata or other nuisance parameters to limited sets of data. There are basically two choices: *one should either use individual case-control matching in the design and the conditional likelihood for analysis; or else the stratum sizes for an unconditional analysis should be kept relatively large, whether the strata are formed at the design stage or* post hoc.

7.2 Multivariate analysis for matched 1:M designs: general methodology

One design which occurs often in practice, and for which the conditional likelihood (7.1) takes a particularly simple form, is where each case is individually matched to one or several controls. The number of controls per case may either be a fixed number, M, say, or else may be allowed to vary from set to set. We considered such designs in § 5.3 and § 5.4 for estimation of the relative risk associated with a single binary exposure variable.

Suppose that the i^{th} of I matched sets contains M_i controls in addition to the case. Denote by $\mathbf{x}_{i0} = (x_{i01}, \ldots, x_{i0K})$ the K-vector of exposures for the case in this set and by $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijK})$ the exposure vector for the j^{th} control ($j = 1, \ldots, M_i$). In other words, x_{ijk} represents the value of the k^{th} exposure variable for the case ($j = 0$) or j^{th} control in the i^{th} matched set. We may then write the conditional likelihood in the form (Liddell, McDonald & Thomas, 1977; Breslow et al., 1978):

$$\prod_{i=1}^{I} \frac{\exp(\sum_{k=1}^{K} \beta_k x_{i0k})}{\sum_{j=0}^{M_i} \exp(\sum_{k=1}^{K} \beta_k x_{ijk})}$$

$$= \prod_{i=1}^{I} \frac{1}{1 + \sum_{j=1}^{M_i} \exp\{\sum_{k=1}^{K} \beta_k (x_{ijk} - x_{i0k})\}} \, . \tag{7.2}$$

It follows from this expression that if any of the x's are matching variables, taking the same value for each member of a matched set, their contribution to the likelihood is zero and the corresponding β cannot be estimated. This is a reminder that matched designs preclude the analysis of relative risk associated with the matching variables. However by defining some x's to be interaction or cross-product terms involving both risk factors and matching variables, we may model how relative risk changes from one matched set to the next.

Table 7.1 Asymptotic mean values of unconditional maximum likelihood estimates of the odds ratio from matched sets consisting of n_1 cases and n_0 controls

True odds ratio	No. of controls per set (n_0)	Proportion of controls positive $p_0 = 0.1$				$p_0 = 0.3$				$p_0 = 0.7$			
		No. of cases per set (n_1)				No. of cases per set (n_1)				No. of cases per set (n_1)			
		1	2	4	10	1	2	4	10	1	2	4	10
1.5	1	2.25	1.81	1.64	1.55	2.25	1.83	1.65	1.56	2.25	1.86	1.67	1.57
	2	1.87	1.72	1.62	1.55	1.85	1.72	1.62	1.55	1.82	1.72	1.63	1.56
	4	1.68	1.63	1.59	1.54	1.67	1.63	1.59	1.55	1.65	1.62	1.59	1.55
	10	1.57	1.56	1.55	1.53	1.57	1.56	1.55	1.53	1.56	1.55	1.55	1.53
2	1	4.00	2.72	2.32	2.12	4.00	2.82	2.37	2.14	4.00	2.94	2.45	2.18
	2	2.97	2.51	2.27	2.11	2.90	2.53	2.29	2.13	2.76	2.52	2.32	2.15
	4	2.47	2.32	2.21	2.10	2.42	2.31	2.21	2.11	2.34	2.28	2.21	2.12
	10	2.19	2.16	2.12	2.07	2.16	2.14	2.12	2.08	2.12	2.12	2.10	2.07
5	1	25.00	10.45	6.98	5.64	25.00	12.68	8.12	6.05	25.00	14.42	9.44	6.67
	2	14.26	8.69	6.66	5.61	12.81	9.11	7.19	5.91	10.08	8.57	7.39	6.24
	4	9.30	7.40	6.31	5.55	8.20	7.22	6.46	5.74	6.83	6.58	6.27	5.84
	10	6.59	6.21	5.84	5.44	6.08	5.93	5.75	5.49	5.60	5.57	5.53	5.43
10	1	100.00	35.66	17.90	12.20	100.00	47.28	24.77	14.60	100.00	53.34	30.55	17.64
	2	50.95	24.85	16.08	12.05	42.71	26.49	18.59	13.61	27.15	21.74	18.07	14.60
	4	28.03	18.80	14.53	11.83	21.54	17.67	15.03	12.67	14.95	14.35	13.67	12.66
	10	16.16	14.28	12.81	11.44	13.34	12.87	12.34	11.60	11.46	11.42	11.34	11.18

If there is but a single matched control per case, the conditional likelihood simplifies even further to

$$\prod_{i=1}^{I} \frac{1}{1+\exp\{\sum_{k=1}^{K}\beta_k(x_{i1k}-x_{i0k})\}} \, . \tag{7.3}$$

This may be recognized as the unconditional likelihood for the logistic regression model where the sampling unit is the pair and the regression variables are the *differences* in exposures for case *versus* control. The constant (α) term is assumed to be equal to 0 and each pair corresponds to a positive outcome ($y = 1$). This correspondence permits GLIM or other widely available computer programmes for unconditional logistic regression to be used to fit the conditional model to matched pair data (Holford, White & Kelsey, 1978).

While not yet incorporated into any of the familiar statistical packages, computer programmes are available to perform the conditional analysis for both matched (Appendix IV) and more generally stratified designs (Appendix V), using the likelihoods (7.2) and (7.1), respectively (Smith et al., 1981). These programmes calculate the following: (i) the (conditional) MLEs of the relative risk parameters; (ii) minus twice the maximized logarithm of the conditional likelihood, used as a measure of goodness of fit; (iii) the (conditional) information matrix, or negative of the matrix of second partial derivatives of the log likelihood, evaluated at the MLE; and (iv) the score statistic for testing the significance of each new set of variables added in a series of hierarchical models. These quantities are used to make inferences about the relative risk just as described in § 6.4 for the unconditional model. For example, the difference between goodness-of-fit (G) measures for a sequence of hierarchical models, in which each succeeding model represents a generalization of the preceding one, may be used to test the significance of the additional estimated parameters. This difference has an asymptotic chi-square distribution, with degrees of freedom equal to the number of additional variables incorporated in the regression equation, provided of course that the β coefficients of these variables are truly zero. Similarly, asymptotic variances and covariances of the parameter estimates in any particular model are obtained from the inverse information matrix printed out by the programme.

Now that the technology exists for conditional logistic modelling, all the types of multivariate analysis of stratified samples which were discussed in Chapter 6 can also be carried out with matched case-control data. In the next few sections we introduce these techniques by re-analysing the data already considered in Chapter 5. This will serve to indicate where the model yields results identical with the "classical" techniques, and where it goes beyond them. Later sections will extend the applications to exploit fully the potential of the model.

7.3 Matched pairs with dichotomous and polytomous exposures: applications

Our first application of the general conditional model is to analyse in this framework the matched pair data already considered at the end of § 5.2. There we used the 63 pairs consisting of the case and the first control in each matched set from the Los Angeles study of endometrial cancer (Mack et al., 1976). The analysis was directed towards obtaining an overall relative risk for oestrogens, detecting a possible interaction with age for the risk associated with gall-bladder disease, and examining the joint effects of gall-bladder disease and hypertension. Further analysis of these same matched pairs was carried out in § 5.5 to investigate the relative risks attached to different dose levels of conjugated oestrogens.

In order to carry out parallel analyses in the context of the logistic model, we defined a number of regression variables as shown in Table 7.2. The first four of these (EST, GALL, HYP, AGEGP) are dichotomous indicators for history of oestrogen use, gall-bladder disease, hypertension, and age, respectively. AGE is a continuous variable, given in years. In cases where the ages of case and control differed, although this was never by more than a year or two, AGE and AGEGP were defined as the age of the case. Hence they represent perfect matching variables which are constant within each matched set. The three binary variables, DOS1, DOS2 and DOS3, represent the four dose levels of conjugated oestrogen and thus should always appear in any equation as a group or not at all. The last variable, DOS, represents the coded dose levels of this same factor, and is used to test specifically for a trend in risk with increasing dose.

Table 7.3 shows the results of a number of regression analyses of the variables defined in Table 7.2. The statistic G for the model with no parameters, i.e., all β's assumed equal to zero, evaluates the goodness of fit to the data of the null hypothesis that none of the regression variables affects risk. Part A of the table considers the relative risk associated with a history (yes or no) of exposure to any oestrogen, as indicated by the binary variable EST. The estimated relative risk is $\hat{\psi} = \exp(\hat{\beta}) = \exp(2.269) = 9.67$, which is precisely the value found in § 5.2 as the ratio 29/3 of discordant pairs. This

Table 7.2 Definition of regression variables used in the matched pairs analysis

Variable	Code		
EST	0 1	No Yes	History of any oestrogen use
GALL	0 1	No Yes	History of gall-bladder disease
HYP	0 1	No Yes	History of hypertension
AGEGP	0 1	Age 55–69 years Age 70–83 years	
AGE	Age in years (55–83)		
DOS 1	1 0	0.1–0.299 mg/day conjugated oestrogens otherwise	
DOS 2	1 0	0.3–0.625 mg/day conjugated oestrogens otherwise	
DOS 3	1 0	0.626+ mg/day conjugated oestrogens otherwise	
DOS	0 1 2 3	None 0.1–0.299 mg/day 0.3–0.625 mg/day 0.626+ mg/day	conjugated oestrogen

Table 7.3 . Results of fitting the conditional logistic regression model to matched pairs consisting of the case and first matched control: Los Angeles study of endometrial cancer

No. of parameters	Goodness of fit (G)	Score test[a]	Regression coefficients ± standard error for each variable in equation		
0	87.34				
			A. Any oestrogens		
			EST		
1	62.89	21.13	2.269 ± 0.606		
			B. Gall-bladder disease and age		
			GALL	GALL × AGEGP	GALL × (AGE-70)
1	83.65	3.56	0.956 ± 0.526		
2	81.87	1.68	1.946 ± 1.069	–1.540 ± 1.249	
2	83.31	0.35[b]	1.052 ± 0.566		–0.066 ± 0.113
			C. Hypertension/Gall-bladder disease		
			GALL	HYP	GALL × HYP
1	86.53	0.81		0.325 ± 0.364	
2	82.79	3.61	0.970 ± 0.531	0.348 ± 0.364	
3	80.84	2.01	1.517 ± 0.699	0.627 ± 0.435	–1.548 ± 1.125
			D. Gall-bladder disease/Hypertension		
			GALL	HYP	GALL × HYP
1	83.65	3.56	0.956 ± 0.526		
2	82.79	0.86	0.970 ± 0.531	0.348 ± 0.377	
3	80.84	2.01	1.517 ± 0.699	0.627 ± 0.435	–1.548 ± 1.125
			E. Dose levels of conjugated oestrogen		
			DOS1	DOS2	DOS3
3	62.98	16.96	1.524 ± 0.618	1.266 ± 0.569	2.120 ± 0.693
			F. Coded dose of conjugated oestrogen		
			DOS	DOS × AGE	
1	65.50	14.71	0.690 ± 0.202		
2	65.50	0.00	0.693 ± 0.282	–0.001 ± 0.403	

[a] Score statistic comparing each model with the preceding model in each set, unless otherwise indicated. The first model in each set is compared with the model in which all β's are 0.

[b] After fitting one parameter model with GALL only

reflects the fact that the conditional likelihood (7.2) is identical (up to a constant of proportionality) to that used earlier as a basis of inference (5.3), so that the two analyses are entirely equivalent. Likewise, the score statistic for the test of the null hypothesis, H_0: $\psi = 1$, is identical with the *uncorrected* (for continuity) value of the χ^2 defined in (5.4), namely

$$\frac{|29-3|^2}{29+3} = 21.13.$$

This illustrates the point that many of the elementary tests are in fact score tests based on the model (Day & Byar, 1979). The corrected chi-square value is of course the more accurate and preferred one, but it has not been incorporated in the computer programme written for the general regression analysis, since it is not applicable in other situations.

Two other statistics are available for testing the null hypothesis. These are the differences in goodness-of-fit measures, 87.34–62.89 = 24.45, and the square of the standardized regression coefficient, $(2.269/0.606)^2 = 13.99$, each of which also has a nominal χ_1^2 distribution under the null hypothesis. Although the three values are somewhat disparate with these data, they all indicate a highly significant effect. The test based on the corrected score statistic is preferred when available, as this comes closest to the corresponding exact test.

Asymptotic 95% confidence limits for ψ are calculated as $\exp(2.269 \pm 1.96 \times 0.606) = (2.9, 31.7)$, the upper limit being noticeably smaller than that based on the exact conditional (binomial) distribution ($\psi_U = 49.6$) or the normal approximation to it ($\psi_U = 39.7$) which were calculated in § 5.2.

Part B of Table 7.3 presents the relative risk estimate for gall-bladder disease and its relationship to age. Just as for EST, the estimate of relative risk associated with GALL, $\exp(0.956) = 2.6 = 13/5$, and the (uncorrected) score statistic, $3.56 = (13-5)^2/18$, must agree with the values found earlier. There is better concordance between the three available tests of the null hypothesis in this (less extreme) case: 87.34–83.65 = 3.69 for the test based on G, and $(0.956/0.526)^2 = 3.30$ for that based on the standardized coefficient, are the other two values besides the score test.

For the second model in Part B the coefficient of GALL represents the log relative risk for those under 70 years of age, $\exp(1.946) = 7.0 = 7/1$, while the sum of the coefficients for GALL and GALL × AGEGP gives the log relative risk for those 70 and over, $\exp(1.946-1.540) = 1.50 = 6/4$. These are the same results as found before. Similarly, the score statistic for the additional parameter GALL × AGEGP, which tests the equality of the relative risk estimates in the two age groups, is identical to the uncorrected chi-square test for equality of the proportions 7/8 and 6/10, namely

$$\chi^2 = \frac{(7 \times 4 - 6 \times 1)^2 \times 18}{8 \times 10 \times 13 \times 5} = 1.68.$$

In § 5.2 we reported the corrected value of this chi-square as $\chi^2 = 0.59$.

The third line of Part B of the table introduces an interaction term with the continuous matching variable AGE. Here the coefficient of GALL gives the estimated relative risk for someone aged 70, $\exp(1.052) = 2.86$, while the relative risk for other ages is determined from $\exp\{1.052-0.066(\mathrm{AGE}-70)\}$. In other words, the RR is estimated to decline by a factor $\exp(-0.066) = 0.936$ for each year of age above 70 and increase by a factor $\exp(0.066) = 1.068$ for each year below. However this tendency has no statistical significance; all three of the available tests for homogeneity give a chi-square of about 0.35 ($p = 0.56$). Such continuous variable modelling is of course not available with the elementary techniques.

Part C of Table 7.3 illustrates the increased analytical power which is available using regression methods. In order to estimate and test the relative risk of gall-bladder disease, while controlling for hypertension, we start with an equation containing the