# 7. CONDITIONAL LOGISTIC REGRESSION FOR MATCHED SETS

CHAPTER VII

# CONDITIONAL LOGISTIC REGRESSION FOR MATCHED SETS

One of the methods for estimating the relative risk parameters **β** 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

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

where **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 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 bε 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 $\alpha$ stratum parameters in addition to the $\beta$ coefficients of primary interest. For matched or finely stratified data, the number of $\alpha$ parameters may be of the same order of magnitude as the number of observations and much greater than the number of $\beta$'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 $\alpha$ and $\beta$ 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 $\beta$, the nuisance parameter $\alpha_i$ 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 | $\mu$ | $1-\mu$ | 1 |
| Control | $1-\mu$ | $\mu$ | 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):

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

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

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 $\beta$ 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<br>$\psi$ | No. of controls<br>per set ($n_0$) | Proportion of controls positive<br>$p_0$ = 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 ($\alpha$) 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 $\beta$ 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 $\beta$'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<br>1 | No<br>Yes | History of any oestrogen use |
| GALL | 0<br>1 | No<br>Yes | History of gall-bladder disease |
| HYP | 0<br>1 | No<br>Yes | History of hypertension |
| AGEGP | 0<br>1 | Age 55–69 years<br>Age 70–83 years | |
| AGE | Age in years (55–83) | | |
| DOS 1 | 1<br>0 | 0.1–0.299 mg/day conjugated oestrogens<br>otherwise | |
| DOS 2 | 1<br>0 | 0.3–0.625 mg/day conjugated oestrogens<br>otherwise | |
| DOS 3 | 1<br>0 | 0.626+ mg/day conjugated oestrogens<br>otherwise | |
| DOS | 0<br>1<br>2<br>3 | None<br>0.1–0.299 mg/day<br>0.3–0.625 mg/day<br>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 $\beta$'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 $\chi^2$ 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 $\chi_1^2$ 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 $\psi$ 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 single variable HYP. When we add to this a second term for gall-bladder disease (line 2, part C), the model then specifies that the relative risks associated with these two variables are multiplicative, and moreover that their joint effect is multiplicative with those of the matching variables. The relative risk for GALL, adjusted for the multiplicative effects of hypertension, is estimated as $\hat{\psi} = \exp(0.970) = 2.65$, scarcely different from the unadjusted value. Likewise the null hypothesis that $\psi = 1$ is tested by $\chi^2 = 3.61$ (uncorrected), which is also rather close to the unadjusted value. By way of contrast, the adjusted estimate of RR for GALL obtained in § 5.2, where we restricted attention to the eight case-control pairs which were homogeneous for HYP and heterogeneous for GALL, gives the relatively unstable value of $\hat{\psi} = 7/1$. The difference is explained by the fact that the model uses all the case-control pairs which are discordant for at least one of GALL and HYP (see Table 7.4) to estimate the main effects of both variables. The five pairs which are discordant for both variables, not used in the elementary analysis, now contribute to the estimate of the coefficient of GALL.

In case the reader is left with the impression that something has been gained for nothing by this procedure, we hasten to point out that the elementary estimate is strictly valid under a weaker set of assumptions than that based on the model. In Chapter 5 we effectively assumed only that the relative risk of GALL was constant with respect to HYP and the matching variables. The modelling procedure supposes in addition that HYP combines multiplicatively with the matching variables; it could lead to biased estimates of the coefficient of GALL if interactions were present. Of course, in some situations, such interactions involving the matching and other confounding variables might also be modelled and added to the equation as a means of further adjustment. For example, if we suspected that not only the main effects of HYP but also the interaction between HYP and AGE were confounding the estimate of the GALL coefficient, we would fit the equation with terms for GALL, HYP and HYP × AGE. Fortunately, the higher order interactions which might necessitate such a procedure are rarely present in epidemiological studies (Miettinen, 1974).

Further insight into the assumptions which underlie the model is given by consideration of line 3 of Part C, Table 7.3. Here the addition of the interaction term GALL × HYP allows us to estimate the relative risk of each possible combination of exposures to these two risk factors, relative to those who are exposed to neither. Thus $\hat{\psi}_{10} = \exp(1.517) = 4.56$ is the estimated RR for those with gall-bladder disease only, $\hat{\psi}_{01} = \exp(0.627) = 1.87$ for those with hypertension only, and $\hat{\psi}_{11} = \exp(1.517 + 0.627 - 1.548) = 1.81$ for those having a positive history of both diseases. In summary, the relative risks are given by this bizarre-looking table:

| | | Gall-bladder disease | |
|---|---|---|---|
| | | − | + |
| Hypertension | − | 1.00 | 4.56 |
| | + | 1.87 | 1.81 |

However the interaction effect is not significant, as indicated by the score statistic comparing lines 2 and 3 of Table 7.3, Part C.

In effect what we have now done is to create out of GALL and HYP a joint risk variable with four exposure categories: (–, –), (–, +), (+, –), and (+, +). The estimation problem is as described in § 5.5 for matched-pair studies with a polytomous risk variable. Table 7.4 presents the distribution of the 63 matched pairs according to the joint response of case and control, following the format of Table 5.5. We readily verify that the maximum likelihood equations (5.30) for data of this type, namely

$$14+1+0 = 20\frac{\psi_{01}}{1+\psi_{01}} + 5\frac{\psi_{01}}{\psi_{01}+\psi_{10}} + 1\frac{\psi_{01}}{\psi_{01}+\psi_{11}}$$

$$6+4+0 = 7\frac{\psi_{10}}{1+\psi_{10}} + 5\frac{\psi_{10}}{\psi_{01}+\psi_{10}} + 1\frac{\psi_{10}}{\psi_{10}+\psi_{11}}$$

$$2+1+1 = 5\frac{\psi_{11}}{1+\psi_{11}} + 1\frac{\psi_{11}}{\psi_{01}+\psi_{11}} + 1\frac{\psi_{11}}{\psi_{10}+\psi_{11}} ,$$

are solved by the estimates just derived using the general computer programme.

The analysis shown in Part D of Table 7.3 is identical with that in Part C except for the order of entry of the variables into the equation. If our interest is in the effects of GALL after adjustment for HYP, we would follow the sequence shown in Part D. In this example, the estimated coefficients and standard errors are not much affected by the presence of the other variable in the equation, which means that they are not confounded to any appreciable degree.

Another example of the analysis of matched-pair data with a polytomous exposure variable was presented at the end of § 5.5. There we estimated the relative risks of endometrial cancer for each of three increasing dose levels of conjugated oestrogens, using the no-dose category as baseline. In order to carry out an essentially identical analysis in the present framework, we first define the three indicator variables DOS1, DOS2 and DOS3, whose $\beta$ coefficients represent the log odds ratios for each of the

Table 7.4 Histories of gall-bladder and hypertensive disease for cases and matched controls: Los Angeles study of endometrial cancer

| Exposures of cases | | Exposures of controls | | | | Total |
|---|---|---|---|---|---|---|
| Gall bladder | Hypertension | – – | – + | + – | + + | |
| – | – | 15 | 6 | 1 | 3 | 25 |
| – | + | 14 | 6 | 1 | 0 | 21 |
| + | – | 6 | 4 | 2 | 0 | 12 |
| + | + | 2 | 1 | 1 | 1 | 5 |
| Total | | 37 | 17 | 5 | 4 | 63 |

dose levels shows in Table 7.2 relative to baseline. The conditional logistic regression model (7.3) in this case is merely a restatement of the model (5.29), in which the odds ratios corresponding to each category of exposure are assumed to be constant over the matching variables. By definition they satisfy the consistency relationship discussed earlier in § 5.5.

Part E of Table 7.3 presents the results. Regression coefficients for the three dose variables do indeed correspond to the odds ratios already estimated: exp(1.524) = 4.59 for the 0.1–0.299 mg/day dose level; exp(1.266) = 3.55 for 0.3–0.625 mg/day; and exp(2.120) = 8.33 for over 0.625 mg/day. Likewise the score statistic for testing the null hypothesis is identical with the statistic (5.32) derived earlier, taking the value 16.96 for these data. The only important additional quantities available from the computer fit of the model are the standard errors of the parameter estimates, which enable us to put approximate confidence limits on the estimated relative risks. For example, $\exp(1.524 \pm 1.96 \times 0.618) = (1.37, 15.4)$ are the 95% limits for the 0.1–0.299 mg/day category.

In order to test for a trend in risk with increasing dose we use the single, coded dose variable DOS. Estimated relative risks for the three dose levels are then exp(0.690) = 1.99, $\exp(2 \times 0.690) = 3.98$ and $\exp(3 \times 0.690) = 7.94$, respectively. Comparing the G statistics for the two dose-response models yields 65.50–62.98 = 2.52, nominally a chi-square with two degrees of freedom, for testing the extent to which the linear trend adequately explains the variation in risk between dose levels. The observed departure from trend is not statistically significant ($p = 0.28$). On the other hand, the trend itself is highly significant ($p<0.0001$) as demonstrated by the value 14.71 for the score statistic. This too is identical to the trend statistic derived earlier (5.33), except that the continuity correction is not used by the computer programme. Note that there is not the slightest hint of interaction between dose and age (line 2, part F, Table 7.3).

In summary, analyses of matched-pair data *via* the conditional logistic model yield results identical to those of the "classical" procedures presented earlier for binary and polytomous risk factors. This is hardly surprising, as the previously discussed methods were themselves based on conditional likelihoods worked out in detail for each separate problem. Nevertheless it is an important fact since it shows that the very general methodology developed here is well integrated with the techniques used in the past. Even more important, of course, are extensions to problems involving multiple and/or continuous risk variables which we next consider in the more general context of 1 : M matching.

### 7.4 1:M matching with single and multiple exposure variables: applications

While the regression variables defined in Table 7.2 have so far in this Chapter been used exclusively with the matched-pair data, their coefficients can in fact be better estimated by taking account of the full complement of controls selected for each case. Table 7.5 presents the results of several analyses, based on the conditional likelihood (7.2), which used all the available data. Since no information was available regarding the dose and/or duration of conjugated oestrogen use by certain of the women, their data records were excluded from the analysis when fitting equations containing these variables. While a missing value for the case leads to exclusion of the entire matched

Table 7.5 Results of fitting several conditional logistic regression models to the matched sets consisting of one case and four controls: Los Angeles study of endometrial cancer

| No. of parameters | Goodness of fit G | Score test | Regression coefficients ± standard error for each variable in the equation | | |
|---|---|---|---|---|---|
| A. Oestrogen use and age level (based on all 63 matched sets, 315 observations) | | | | | |
| | | | EST | EST × AGE1 | EST × AGE2 |
| 0 | 202.79 | – | | | |
| 1 | 167.44 | 31.16 | 2.074 ± 0.421 | | |
| 3 | 166.76 | 0.76 | 1.431 ± 0.826 | 0.847 ± 1.034 | 0.780 ± 1.154 |
| B. Oestrogen use and coded age level (based on all 63 matched sets, 315 observations) | | | | | |
| | | | EST | EST × AGE3 | |
| 1 | 167.44 | 31.16 | 2.074 ± 0.421 | | |
| 2 | 167.05 | 0.39 | 1.664 ± 0.750 | 0.385 ± 0.616 | |
| C. Conjugated oestrogen use and age (based on 59 matched sets, 291 observations) | | | | | |
| | | | CEST | CEST × AGE1 | CEST × AGE2 |
| 0 | 188.13 | | | | |
| 1 | 159.22 | 27.57 | 1.710 ± 0.354 | | |
| 3 | 158.28 | 0.89 | 1.583 ± 0.815 | –0.081 ± 0.930 | 0.764 ± 1.143 |

set, a missing value in a control record might simply mean that the number of controls in that set was reduced by one.

In order to estimate the overall relative risk associated with a history of exposure to any oestrogen, we employed the general purpose computer programme with the single binary variable EST (Part A, Table 7.5). This yields $\hat{\psi} = \exp(2.074) = 7.95$, which is of course the same value as found in § 5.3 by solving the equation (5.17) for conditional maximum likelihood estimation. The standard error $0.421 = \sqrt{0.177}$, given by formula (5.21), has already been used to place an approximate 95% confidence interval of $\exp(2.074 \pm 1.96 \times 0.421) = (3.5,\ 18.1)$ about the point estimate. Likewise the score test statistic is identical to the summary chi-square defined in (5.19), but calculated without the continuity correction so as to give $(110–13)^2/302 = 31.16$ in place of the corrected value 29.57 found earlier.

Continuing the lines of the analysis shown in Table 5.2, we investigated a possible difference in the relative risk for EST in the three age groups 55–64, 65–74 and 75+ by adding to the regression equation interaction terms involving EST and age. In order to account for the breakdown of age into three groups, two binary indicator variables were defined: AGE1 = 1 for 65–74 years, and 0 otherwise; and AGE2 = 1 for 75+ years, 0 otherwise. Thus, from line 2, Part A, Table 7.5, $\exp(1.431) = 4.18$ is the estimated relative risk for women aged 55–64 years, $\exp(1.431 + 0.847) = 9.76$ for those 65–74 years, and $\exp(1.431 + 0.780) = 9.12$ for the 75+ year olds, these results agreeing with those shown in Table 5.2. While there is an apparent increase in the relative risk for the women aged 65 or more years, the score test of 0.76 shows that the differences are not statistically significant ($p = 0.68$). Note that this value agrees with that calculated earlier from the explicit formula (5.23) for the score test of interaction.

A single degree of freedom test for a trend in relative risk with increasing age is obtained by fitting a single interaction term as shown in Part B of Table 7.5. Coding AGE3 to be 0, 1 or 2 according to the subject's age group, the resulting score test for interaction is the uncorrected version of the statistic (5.24), taking the value 0.39. The corrected value calculated earlier was 0.09. Estimated relative risks for the three age categories are in this case $\exp(1.664) = 5.28$, $\exp(1.664 + 0.385) = 7.76$ and $\exp(1.664 + 2 \times 0.385) = 11.40$, respectively. However since there is no evidence that the apparent trend is real, such estimates would not normally be reported.

The flexibility of the regression approach is particularly evident when dealing with matched sets containing a variable number of controls. Part C of Table 7.5 presents

Table 7.6 Matched univariate analysis of Los Angeles study of endometrial cancer: all cases and controls used except as noted

| Variable | Levels | RR | $\chi^2$[a] | DF | p |
|---|---|---|---|---|---|
| Gall-bladder disease | Yes | 3.69 | 13.83 | 1 | 0.0002 |
| | No | 1.00 | | | |
| Hypertension | Yes | 1.51 | 1.85 | 1 | 0.18 |
| | No | 1.00 | | | |
| Obesity | Yes | 1.76 | 5.70 | 2 | 0.06 |
| | No | 1.00 | | | |
| | Unk | 0.63 | | | |
| Obesity | Yes | 2.02 | 5.16 | 1 | 0.02 |
| | No/Unk | 1.00 | | | |
| Other drugs (non-oestrogen) | Yes | 3.90 | 10.38 | 1 | 0.001 |
| | No | 1.00 | | | |
| Any oestrogens | Yes | 7.96 | 31.16 | 1 | <0.00001 |
| | No | 1.00 | | | |
| Conjugated oestrogens[b]: dose in mg/day | None | 1.00 | 33.22 | 3 | <0.00001 |
| | 0.1–0.299 | 4.11 | | | |
| | 0.3–0.625 | 4.86 | | | |
| | 0.625+ | 10.97 | | | |
| | Trend[c] | 5.53 | 27.57 | 1 | <0.00001 |
| Conjugated oestrogens[d]: duration in months | None | 1.00 | 34.93 | 4 | <0.00001 |
| | 1–11 | 2.66 | | | |
| | 12–47 | 4.17 | | | |
| | 48–95 | 8.13 | | | |
| | 96+ | 10.41 | | | |
| | Trend[e] | 1.81 | 34.79 | 1 | <0.00001 |

[a] Uncorrected score test

[b] Based on 59 sets, 291 observations

[c] Regression on coded dose levels: 0 = none; 1 = 0.1–0.299 mg/day; 2 = 0.3–0.625 mg/day; 3 = 0.625+ mg/day

[d] Based on 57 sets, 277 observations

[e] Regression on coded duration: 0 = none; 1 = 1–11 months; ...; 4 = 96+ months

the regression analysis of the data considered in § 5.4 on use of conjugated oestrogens. Of 59 matched sets for whom the case history of conjugated oestrogen use was known, 55 had the full complement of 4 controls while for each of the 4 others, one control was lacking information. Running the computer programme with a single binary variable CEST representing the history of use of conjugated oestrogens, we easily replicate the results already obtained: $\hat{\psi} = \exp(1.710) = 5.53$ for the estimate of relative risk and $\chi^2 = 27.57$ for the uncorrected chi-square test of the null hypothesis. It is also easy to test for constancy of the relative risk over the three age groups by addition of the interaction variables CEST×AGE1 and CEST×AGE2 to the equation. The score test for this addition, which is the generalization of (5.24) discussed in § 5.4, yields the value $\chi^2_2 = 0.89$ ($p = 0.64$). We did not report this result earlier because of the labour involved in the hand calculation.

Thus far in this section we have used the general methods for matched data analysis primarily in order to replicate the results already reported in Chapter 5 for particular elementary problems. The emphasis has been on demonstrating the concordance between the quantities in the computerized regression analysis, and those calculated earlier from grouped data. In the remainder of the section we carry out a full-scale multivariate analysis of the Los Angeles data much as one would do in actual practice.

As an initial step in this process, Table 7.6, which summarizes and extends the results obtained so far, presents relative risk estimates and tests of their statistical significance for each risk variable individually. Comparing the entries there with those in Table 5.1 we see that there is little to choose between the matched and unmatched analyses for this particular example (see § 7.6, however). The rather large number of "unknown" responses for obesity indicated lack of information on this item in the medical record. Grouping these with the negatives led to only a slight decrease in the goodness of fit ($\chi^2_1 = 0.75$, $p = 0.39$) and to a slight increase in the relative risk associated with a positive history. We therefore decided to use the dichotomy positive *versus* negative/unknown in the subsequent multivariate analyses. This meant that the final analyses used the five binary variables GALL-bladder disease, HYPertension, OBesity, NON-oestrogen drugs and any oESTrogen, none of which had missing values. There were also two polytomous variables representing DOSe and DURation of conjugated oestrogen, both of which had missing values.

Table 7.7 presents the results for a series of multivariate analyses involving the five binary risk factors and several of their two-factor interactions. Model 2 contains just the main effects of each variable. Their $\beta$ coefficients have been exponentiated for presentation so as to facilitate their interpretation in terms of relative risk. In fact the estimates of RR for gall-bladder disease and oestrogen use do not change much from the univariate analysis (Table 7.6), while those for the other three variables are all somewhat smaller. The coefficient for hypertension becomes slightly negative, while those for obesity and non-oestrogen drugs are reduced to non-significant levels. The reduction for non-oestrogen drugs is particularly striking, and inspection of the original data indicates this is due to a high degree of confounding with oestrogen use: for the controls, only 16 or 21.1% of 76 who did not take non-oestrogen drugs had a history of oestrogen use, *versus* 111 or 63.1% of 176 who did take non-oestrogen drugs (Table 7.8).

Models 3–5 explore the consequences of dropping from the equation those variables which do not have significant main effects. The confounding between other drugs and

Table 7.7 Matched multivariate analysis of five binary risk factors and their interactions: Los Angeles study of endometrial cancer

| Model | No. of parameters | Goodness of fit G | Score test[a] | Relative risks (exponentiated regression coefficients) for each variable in the equation<br>Standardized regression coefficients in parentheses | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | GALL | HYP | OB | NON | EST | GALL×EST | OB×EST | NON×EST | GALL × OB | GALL×NON |
| 1 | 0 | 202.79 | | | | | | | | | | | |
| 2 | 5 | 153.74 | 42.75 | 3.63<br>(3.12) | 0.82<br>(–0.55) | 1.61<br>(1.31) | 1.95<br>(1.31) | 6.78<br>(4.21) | | | | | |
| 3 | 4 | 154.04 | 42.63[b] | 3.59<br>(3.10) | | 1.58<br>(1.27) | 1.85<br>(1.23) | 6.58<br>(4.17) | | | | | |
| 4 | 3 | 155.64 | 41.74[b] | 3.58<br>(3.10) | | 1.67<br>(1.43) | | 7.69<br>(4.62) | | | | | |
| 5 | 3 | 155.68 | 41.42[b] | 3.63<br>(3.12) | | | 2.00<br>(1.39) | 6.79<br>(4.22) | | | | | |
| 6 | 2 | 157.74 | 39.92[b] | 3.58<br>(3.10) | | | | 8.29<br>(4.81) | | | | | |
| 7 | 3 | 153.46 | 4.66 | 18.07<br>(3.28) | | | | 14.88<br>(4.41) | 0.128<br>(–2.06) | | | | |
| 8 | 4 | 151.58 | 4.39[c] | 17.19<br>(3.23) | | 1.63<br>(1.35) | | 13.74<br>(4.27) | 0.136<br>(–2.01) | | | | |
| 9 | 5 | 151.50 | 0.43 | 17.84<br>(3.26) | | 2.85<br>(1.13) | | 19.92<br>(3.43) | 0.132<br>(–2.02) | 0.532<br>(–0.65) | | | |
| 10 | 6 | 151.00 | 0.17 | 14.78<br>(2.72) | | 2.46<br>(0.93) | | 18.98<br>(3.43) | 0.127<br>(–2.05) | 0.576<br>(–0.56) | | 1.39<br>(0.41) | |
| 11 | 4 | 151.17 | 4.92[d] | 19.45<br>(3.32) | | | 2.14<br>(1.46) | 12.28<br>(4.03) | 0.120<br>(–2.12) | | | | |
| 12 | 5 | 148.75 | 2.23 | 22.51<br>(3.27) | | | 8.63<br>(1.80) | 54.60<br>(3.07) | 0.103<br>(–2.14) | | 0.156<br>(–1.42) | | |
| 13 | 6 | 148.04 | 0.68 | 9.36<br>(1.55) | | | 4.94<br>(1.23) | 40.36<br>(2.94) | 0.089<br>(–2.22) | | 0.225<br>(–1.11) | | 2.98<br>(0.81) |

[a] Score test with respect to preceding model, unless otherwise noted
[b] Score test for all variables in model (with respect to Model 1)
[c] Score test *versus* Model 4
[d] Score test *versus* Model 5

Table 7.8 Joint distribution of cases and controls according to selected risk factors: Los Angeles study of endometrial cancer

A. Gall-bladder disease and oestrogens

| | Gall-bladder disease negative | | Gall-bladder disease positive | | Totals |
|---|---|---|---|---|---|
| | Oestrogen− | Oestrogen+ | Oestrogen− | Oestrogen+ | |
| Cases | 3 | 43 | 4 | 13 | 63 |
| Controls | 117 | 111 | 8 | 16 | 252 |
| Relative risks | | | | | |
| Unmatched | 1.0 | 15.1 | 19.5 | 31.7 | |
| Matched[a] | 1.0 | 14.9 | 18.1 | 34.5 | |

B. Oestrogen and non-oestrogen drug use

| | Other drugs negative | | Other drugs positive | | Totals |
|---|---|---|---|---|---|
| | Oestrogen− | Oestrogen+ | Oestrogen− | Oestrogen+ | |
| Cases | 1 | 6 | 6 | 50 | 63 |
| Controls | 60 | 16 | 65 | 111 | 252 |
| Relative risks | | | | | |
| Unmatched | 1.0 | 22.5 | 5.5 | 27.0 | |
| Matched[b] | 1.0 | 54.6 | 8.6 | 73.5 | |

[a] From Model 7, Table 7.7.

[b] From Model 12, Table 7.7 (hence adjusted for gall-bladder disease)

oestrogen is evident from the fact that the coefficient for the latter depends most noticeably on whether or not the former is present. Subtracting the goodness-of-fit statistics between Models 6 and 2 yields $\chi_3^2 = 4.00$ ($p = 0.26$) for testing the joint contribution of hypertension, obesity and non-oestrogen drug use to the equation.

The contrast between Models 7 and 6 shows that there is a strong and statistically significant ($p = 0.03$) *negative* interaction between the two variables that have substantial main effects on risk, namely gall-bladder disease and oestrogens. The basic data contributing to this negative interaction are shown in Part A of Table 7.8, together with relative risks estimated *via* the model, e.g., $RR = 14.9 \times 18.1 \times 0.128 = 34.5$ for the double exposure category. The interaction effect itself is perhaps best illustrated by contrasting the RR of 14.9 for oestrogens among those who had no history of gall-bladder disease with the RR of $34.5/18.1 = 1.9$ among those with such a history.

Similar negative interactions are evident in Models 10 and 12 for obesity with oestrogens, and other drugs with oestrogens, respectively. From the unmatched data, shown in Part B of Table 7.8, we see that the instability in the regression coefficients for Model 12 stems from the fact that only a single case falls in the joint "non-exposed" category. While they are statistically significant only in the case of gall-bladder disease, the data suggest that there are negative interactions of oestrogen use with the other factors which are possibly linked to endometrial cancer. Given that a woman is already at elevated risk from her history of gall-bladder disease, obesity, or non-oestrogen drug use, the further increase in risk from use of oestrogens is not nearly as important as when she is not exposed to other risk factors. This same observation, that oestrogen use interacts negatively with traditional risk factors for endometrial cancer, such as hypertension and obesity, has been made in other case-control studies (Smith et al., 1975). It suggests that the effects of oestrogen use are more likely to combine additively rather than multiplicatively with those of other factors. Another interesting feature of the relationship, which could not be investigated in the Los Angeles study, is that the excess risk is much smaller among ex-users compared with continuing users of oestrogen (Jick et al., 1979).

So far our analysis has accounted only for the fact of oestrogen use and not of dose or duration. Unfortunately, information about one or both of these items was lacking for nine cancer cases, leading to the exclusion of the corresponding matched sets from the analysis, and for one control in each of seven of the remaining 54 sets. Moreover, the drug tended to be administered at one of a few standard doses, which precluded analysis of this variable as a true continuous variable. Instead both dose and duration were treated as ordered categorical variables, and arbitrary scale values were assigned to the increasing levels for regression analysis of trends (see Tables 7.2 and 7.6).

A series of analyses which investigate the effect of dose and/or duration of conjugated oestrogen exposure on risk is presented in Table 7.9. In part A of the table we first fit the main effect for oestrogen exposure followed by a single variable DOS representing the trend in risk with coded dose level. Since women with EST = 1 but DOS = 0 use oestrogens but not the conjugated variety, the coefficient of EST determines the relative risk for women taking only non-conjugated oestrogens, $\exp(1.451) = 4.3$. Estimated relative risks for the three dose levels of conjugated oestrogen are $\exp(1.451+0.402) = 6.4$, $\exp(1.451+2\times 0.402) = 9.5$ and $\exp(1.451+3\times 0.402) = 14.3$, respectively. The third model is a generalization of the second in that the effects of the individual dose levels are allowed to vary independently rather than being determined by the trend. While the estimated relative risks for dose levels 1 and 2 are rather similar, there is no strong evidence for a deviation from the fitted trend ($\chi^2_2 = 2.41$, p = 0.30). As shown in Model 4, there is a significant trend with duration, even after accounting for the dose effects.

Part B of the table considers in a similar way the effect of duration of exposure. Here there is a smooth progression in risk, and the fit of the linear trend in coded duration level seems quite adequate ($\chi^2_3 = 1.11$, p = 0.78). The trend in dose continues to be significant even after adjustment for duration (Model 3, Part B).

In Part C of the table we simultaneously fit separate effects for both dose and duration. Since the sums of both DOS1 + DOS2 + DOS3 and DUR1 + DUR2 + DUR3 + DUR4 equal the variable CEST defined above, it was necessary to drop one of these indicator variables from the equation in order to avoid linear dependence among the variables and to obtain unique estimates of all coefficients; this explains the absence of DOS1 from the list of variables. Comparing Model 2 with Model 1 shows that the effects of dose and duration are reasonably multiplicative; addition of the linear interaction term results in only a slight improvement in goodness of fit($\chi^2_1 = 0.59$, p = 0.44). In Models 3–6 we consider the effects of some of the other risk factors after

Table 7.9 Multivariate analysis of effects of dose and duration of conjugated oestrogens: Los Angeles study of endometrial cancer

| Model | No. of parameters | Goodness of fit G | Score test[b] | Regression coefficients for each variable in the equation (standardized coefficients in parentheses) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | A. Effect of dose | | | | | | |
| | | | | EST | DOS | DOS1 | DOS2 | DOS3 | DUR | |
| 1 | 1 | 139.86 | 27.22 | 2.088<br>(4.60) | | | | | | |
| 2 | 2 | 135.63 | 4.19 | 1.451<br>(2.59) | 0.402<br>(2.01) | | | | | |
| 3 | 4 | 133.20 | 2.41 | 1.856<br>(2.74) | | 0.029<br>(0.05) | 0.023<br>(0.04) | 1.141<br>(1.80) | | |
| 4 | 5 | 128.32 | 4.82 | 1.987<br>(2.86) | | −1.101<br>(−1.33) | −1.116<br>(−1.32) | −0.013<br>(−0.02) | 0.420<br>(2.15) | |
| | | | | B. Effect of duration | | | | | | |
| | | | | EST | DUR | DUR1 | DUR2 | DUR3 | DUR4 | DOS |
| 1 | 2 | 134.84 | 4.91[c] | 1.431<br>(2.58) | 0.309<br>(2.17) | | | | | |
| 2 | 5 | 133.76 | 1.11 | 1.868<br>(2.73) | | −0.418<br>(−0.58) | 0.122<br>(0.19) | 0.596<br>(0.88) | 0.899<br>(1.43) | |
| 3 | 6 | 129.52 | 4.22 | 1.946<br>(2.83) | | −1.655<br>(−1.70) | −0.876<br>(−1.08) | −0.586<br>(−0.65) | −0.296<br>(−0.34) | 0.578<br>(2.01) |

Table 7.9 (contd)

| | | | | C. Dose, duration and other variables | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | EST | DOS2 | DOS3 | DUR1 | DUR2 | DUR3 | DUR4 | DUR× | GALL | GALL× | NON | OB |
| | | | | | | | | | | | DOS | | EST | | |
| 1 | 7 | 127.63 | 6.25[d] | 2.020 | −0.024 | 1.175 | −0.961 | −0.131 | 0.251 | 0.404 | | | | | |
| | | | 5.35[e] | (2.90) | (−0.05) | (2.07) | (−1.18) | (−0.19) | (0.33) | (0.57) | | | | | |
| 2 | 8 | 127.04 | 0.59 | 2.024 | −0.835 | −0.416 | −0.395 | −0.555 | −0.464 | 0.254 | 0.179 | | | | |
| | | | | (2.91) | (−1.03) | (−0.53) | (−0.35) | (−0.39) | (−0.61) | (0.19) | (0.77) | | | | |
| 3 | 8 | 118.91 | 9.54 | 2.083 | −0.725 | −0.019 | 0.470 | 0.283 | −0.049 | 1.136 | | 1.498 | | | |
| | | | | (2.81) | (−0.86) | (−0.03) | (0.59) | (0.38) | (−0.09) | (1.95) | | (2.93) | | | |
| 4 | 9 | 117.06 | 1.87 | 2.433 | −0.708 | −0.034 | 0.357 | 0.276 | 0.000 | 1.111 | | 2.531 | −1.519 | | |
| | | | | (2.99) | (−0.85) | (−0.05) | (0.45) | (0.37) | (0.00) | (1.92) | | (2.72) | (−1.35) | | |
| 5 | 9 | 116.42 | 2.33 | 1.951 | −0.694 | −0.008 | 0.525 | 0.220 | −0.076 | 1.114 | | 1.521 | | 0.936 | |
| | | | | (2.59) | (−0.82) | (−0.01) | (0.65) | (0.29) | (−0.14) | (1.90) | | (2.96) | | (1.50) | |
| 6 | 9 | 113.58 | 5.22 | 2.195 | −0.908 | −0.140 | 0.356 | 0.231 | −0.228 | 1.242 | | 1.423 | | | 1.059 |
| | | | | (2.86) | (−1.04) | (−0.19) | (0.43) | (0.29) | (−0.41) | (2.06) | | (2.73) | | | (2.24) |

[a] Based on 54 matched sets, 263 observations having known values for both dose and duration of conjugated oestrogen use

[b] Score test relative to preceding model in each Part, unless otherwise indicated

[c] Relative to Model 1, Part A

[d] Relative to Model 2, Part A

[e] Relative to Model 1, Part B

more complete adjustment for oestrogen than was possible using the binary variable EST alone. The coefficients for these variables should be contrasted with those shown in Table 7.7. Gall-bladder disease continues to stand out as an important, independent risk factor with an estimated relative risk of exp(1.498) = 4.5 compared with the 3.6 found earlier (Model 6, Table 7.7). The interaction of gall-bladder disease with oestrogen use is no longer statistically significant when the dose and duration variables are included in the equation. While the coefficient for non-oestrogen drugs is little changed, obesity is now estimated to carry a relative risk of exp(1.059) = 2.9, which is significantly different from 1 at the p = 0.02 level. Part of these differences, of course, may result because slightly different data sets were used.

In conclusion, we can simply reiterate a point which is well illustrated by the preceding example: all the techniques of multivariate analysis which were once restricted to unmatched studies are now available for use with matched data as well.

## 7.5 Combining sets of 2×2 tables

Besides individual case-control matching, another situation in which the calculations based on the exact conditional likelihood may be quite feasible is when information is combined from a set of 2×2 tables. We noted earlier that the conditional likelihood in this case took the form of a product of non-central hypergeometric distributions (see § 4.4 for notation):

$$\prod_{i=1}^{I} \frac{\binom{n_{1i}}{a_i}\binom{n_{0i}}{m_{1i}-a_i}\psi_i^{a_i}}{\sum_u \binom{n_{1i}}{u}\binom{n_{0i}}{m_{1i}-u}\psi_i^{u}}. \qquad (7.4)$$

As usual, the summations in the denominator range over all possible values u which are consistent with the observed marginals in the i$^{th}$ table, namely $\max(0, n_{1i}-m_{0i}) \leq u \leq \min(m_{1i}, n_{1i})$. Calculation of exact tail probabilities (4.6, 4.7) and confidence intervals (4.8, 4.9) based on this distribution requires that all possible sets of tables which are compatible with the given marginals are evaluated. Their number is

$$\prod_{i=1}^{I}\{\min(m_{1i},n_{1i})-\max(0,n_{1i}-m_{0i})\},$$

i.e., the *product* of the number of possible tables at each level, which can rapidly become prohibitively large (Thomas, 1975). On the other hand, evaluation of the log-likelihood function and its first and second derivatives requires calculations which increase only in proportion to the *sum*

$$\sum_{i=1}^{I}\{\min(m_{1i},n_{1i})-\max(0,n_{1i}-m_{0i})\}$$

of the number of possible tables at each level. Hence a conditional likelihood analysis, similar to those already developed in this chapter for matched designs, is often possible for problems involving sets of 2×2 tables, even where the completely exact analysis would be unfeasible. Only if the entries in some of the tables are very large will problems be encountered in the evaluation of the binomial coefficients appearing in (7.4).

Usually cases and controls will have been grouped into strata (tables) on the basis of covariables which are thought either to confound or to modify the effect of exposure on disease. Suppose that a vector $\mathbf{z}_i$ of such covariables is associated with the i<sup>th</sup> table. Then there are several hypotheses about the odds ratios $\psi_i$ which are of interest:

$$H_0\colon\ \psi_i \equiv 1$$

$$H_1\colon\ \psi_i \equiv \psi = \exp(\beta)$$

$$H_2\colon\ \psi_i = \exp(\beta + \Sigma_l \gamma_l z_{il})$$

$$H_3\colon\ \text{No restrictions on } \psi_i.$$

In Chapter 4 we concentrated on the estimation of $\psi$ under $H_1$, tests of the null hypothesis $H_0$, and tests for constancy in the relative risk ($H_1$) against global alternatives ($H_3$). We have remarked on several occasions that these latter may be insensitive to particular patterns of interaction and that a preferred strategy is to model specific variations in the relative risk associated with the covariables using $H_2$. In § 6.12 several such models were fitted to the Oxford Childhood Survey data using unconditional logistic regression in which a separate $\alpha$ parameter was estimated for each stratum. As we saw in § 7.2, however, it is possible seriously to overestimate the relative risk with this procedure if the data are thin. Hence it will often be preferable to use instead the conditional likelihood, which may be written

$$\prod_{i=1}^{I} \frac{\binom{n_{1i}}{a_i}\binom{n_{0i}}{m_{1i}-a_i} \exp\{a_i(\beta + \Sigma_l \gamma_l z_{il})\}}{\sum_u \binom{n_{1i}}{u}\binom{n_{0i}}{m_{1i}-u} \exp\{u(\beta + \Sigma_l \gamma_l z_{il})\}}\,. \tag{7.5}$$

A listing of a computer programme for fitting models of the form $H_2$ to sets of 2×2 tables using the conditional likelihood is given in Appendix VI. This programme may be used as an alternative to that of Thomas (1975) for finding the exact MLE of the relative risk in $H_1$, provided of course that exact tests and confidence intervals are not also desired. Zelen (1971) develops exact tests for the constancy of the odds ratio against alternatives of the form $H_2$ with a single covariable, and also against the global alternative $H_3$. We presented in (4.31) the score statistic based on (7.5) for testing $H_1$ against $H_2$ with a single covariable.

If the data in each table are truly extensive it may be burdensome to evaluate the binomial coefficients in (7.5). In this case an asymptotic procedure is available. Rather than use the exact conditional means and variances of the table entries $a_i$ under hypothesized values for the odds ratios $\psi_i$, which are required by the iterative likelihood fitting procedure, one can use instead the asymptotic means and variances defined by (4.11) and (4.13). This substitution yields likelihood equations and an information matrix which are identical to those obtained by applying a two-stage maximization procedure to the *unconditional* likelihood function whereby one first solves the equations for the $\alpha$ coefficients in terms of $\beta$ and $\gamma$ (Richards, 1961). The estimates $\hat{\beta}$ and $\hat{\gamma}$ so obtained, as well as their standard errors and covariances, are thus identical to those obtained using unconditional logistic regression (Breslow, 1976). The advantage is that the unconditional model is fitted without explicit estimation of all the nuisance parameters. This is a serious consideration if there are many tables, since the required number of parameters may exhaust the capacity of the available computer. Nevertheless, no matter how they are calculated, the unconditional estimates may be subject to bias in such circumstances and the conditional analysis is preferred whenever it is computationally feasible.

To illustrate the use of the conditional likelihood with a set of 2 × 2 tables we found new estimates of the parameters $\beta$ and $\gamma_1$, representing the log relative risk of obstetric radiation and its linear decrease with calendar time, which we estimated earlier from the Oxford Childhood Cancer Survey Data using unconditional logistic regression (6.12). We recall that several estimates for these parameters were made depending on the degree of polynomial adjustment for the stratifying variables age and calendar year. In fact, for the last line in Table 6.17 where the confounding effects of age and year were completely saturated, we avoided explicit estimation of separate $\alpha$ parameters for each of the 120 2 × 2 tables by using the technique just discussed.

The parameter estimates and standard errors calculated directly from the conditional likelihood (7.5) were

$$\hat{\beta} = 0.5165 \pm 0.0564$$

and

$$\hat{\gamma}_1 = -0.0385 \pm 0.0144\ .$$

It is of considerable theoretical interest that these quantities are closer to those obtained from the unconditional fifth degree polynomial model than to those obtained with the saturated model (see last two lines, Table 6.17). This suggests that the confounding effects of age and year are suitably accounted for by the polynomial regression, and that inclusion of additional nuisance parameters in the equation serves only to increase bias of the type considered in § 7.1. However, because of the exceptionally large sample (over 5 000 cases and controls) the inflation of the relative risk estimates due to the excess of nuisance parameters was not terribly serious.

## 7.6 Effect of ignoring the matching

Prior to the advent of methods for the multivariate analysis of case-control studies, in particular those based on the conditional likelihood (7.2), it was common practice to ignore the matching in the analysis. In simple problems one often found that taking explicit account of the matched pairs or sets did not seriously alter the estimate of relative risk. With the Los Angeles study of endometrial cancer, for example, there were only slight differences between the unmatched (Table 7.5) and matched (Table 7.6) estimates for each risk variable considered individually. However, the agreement is not always as good, and there has been considerable confusion regarding the conditions under which incorporation of the matching in the analysis is necessary.

A sufficient and widely-quoted condition for the 'poolability' of data across matched sets or strata is that the *stratification variables are either:* (i) *conditionally independent of disease status given the risk factors; or* (ii) *conditionally independent of the risk factors given disease status.* If either of these conditions is satisfied, both pooled and matched analyses provide (asymptotically) unbiased estimates of the relative risk for a dichotomous exposure (Bishop, Fienberg & Holland, 1975). [Whittemore (1978) has shown that, contrary to popular belief, both types of analyses may sometimes yield equivalent results even if conditions (i) and (ii) are both violated.] In matched studies condition (i) is more relevant since the matching variables are guaranteed to be uncorrelated with disease in the sample as a whole. Of course this does not ensure that they have the same distributions among cases and controls conditionally, within categories defined by the risk factors. Therefore an unmatched analysis may give biased results.

One result of using an unmatched analysis with data collected in a matched design, however, is that the *direction of the bias tends towards conservatism.* Relative risk estimates from the pooled data tend on average to be closer to unity than those calculated from the matched sets. This phenomenon was noted in § 3.4 when pooling data from two 2 × 2 tables, where the ratio of cases to controls in each table was constant. Seigel and Greenhouse (1973) show that the same thing happens if matched pairs are formed at random from among the cases and controls within each of two strata, and the data are then pooled for analysis. Armitage (1975) gives a slightly more general formulation. He supposes that there are I matched sets with exposure probabilities $p_{1i} = 1-q_{1i}$ for the cases and $p_{0i} = 1-q_{0i}$ for the controls, and that the odds ratio $\psi = p_{1i}q_{0i}/(p_{0i}q_{1i})$ is constant across all sets. It follows that the estimate of relative risk calculated as the cross-products ratio from the 2 × 2 table formed by pooling all the data tends towards the value

$$\frac{\Sigma p_{1i}\Sigma q_{0i}}{\Sigma p_{0i}\Sigma q_{1i}}$$

$$= \psi \frac{\Sigma q_{1i}\vartheta_i \Sigma q_{0i}}{\Sigma q_{0i}\vartheta_i \Sigma q_{1i}} \tag{7.6}$$

where $\vartheta_i = p_{0i}/q_{0i}$. For $\psi > 1$ the bias term multiplying $\psi$ in (7.6) is less than one, unless the exposure probabilities $p_{0i}$ are constant across sets (in which case there is no bias). Similarly, for $\psi < 1$, the bias term exceeds unity. Thus, failure to account for the matching in the analysis can (and often does) result in conservatively biased estimates of the relative risk.

A related question is to consider the cost, in terms of a loss of efficiency in the analysis, of using a matched analysis when in fact the matching was unnecessary to avoid bias. Suppose that the exposure probabilities $p_{0i}$ in the above model are all equal to the constant $p_0$, so that both matched and unmatched analyses tend to estimate correctly the true odds ratio $\psi$. According to (4.18), the large sample variance of the pooled estimate of log $\psi$ is

$$\frac{1}{I}\left\{\frac{1}{p_1}+\frac{1}{q_1}+\frac{1}{p_0}+\frac{1}{q_0}\right\}=\frac{p_1q_1+p_0q_0}{Ip_1q_1p_0q_0}.$$

Standard calculations show that the large sample variance of the estimate of log $\psi$ based on the matched pairs in this situation is

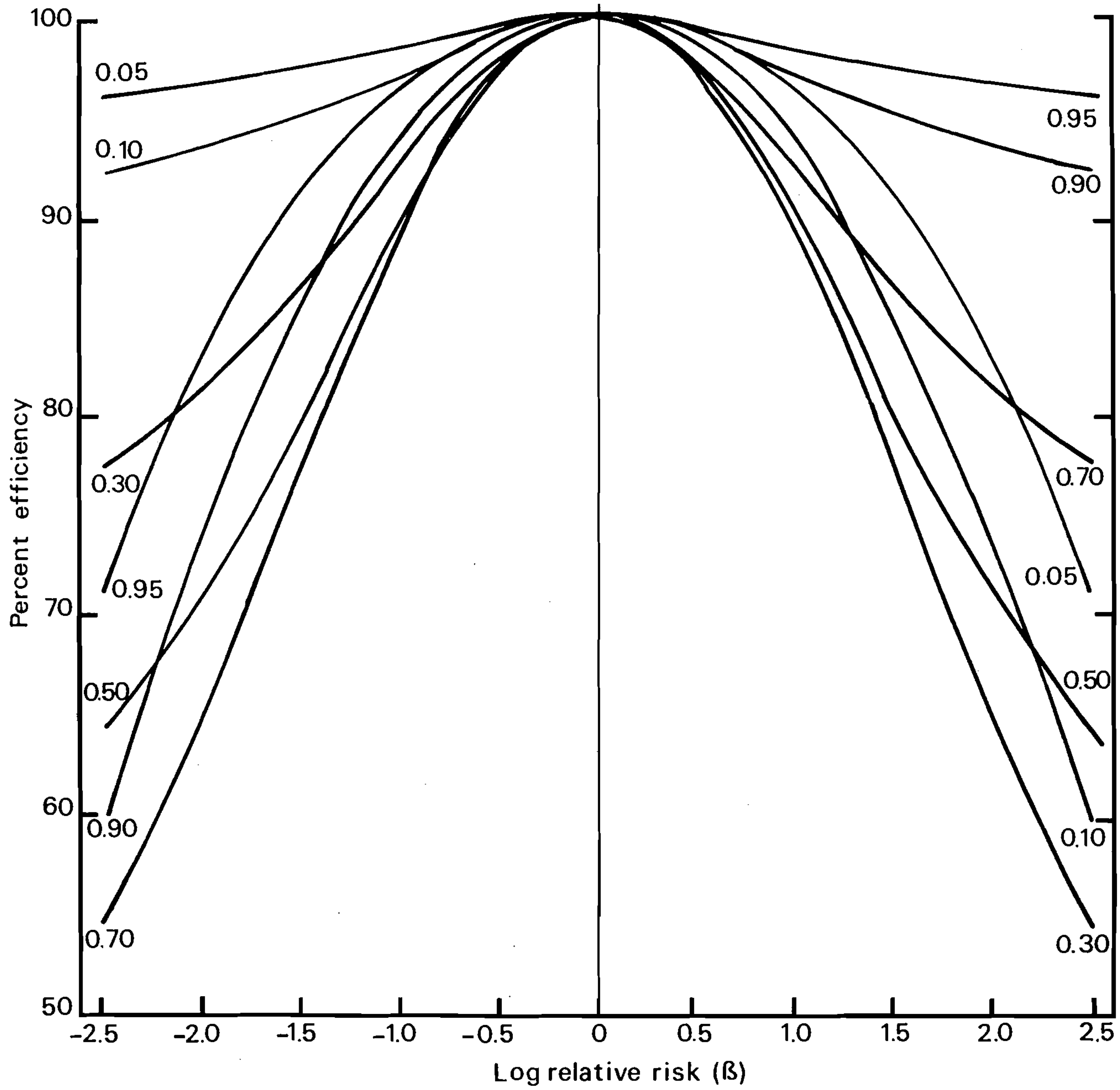
$$\frac{p_1q_0+q_1p_0}{Ip_1q_1p_0q_0}.$$

Consequently, using the ratio of variances to measure the relative precision of the two estimates, the efficiency of the matched pairs analysis when pairing at random is

$$\text{eff} = \frac{p_1q_1 + p_0q_0}{p_1q_0 + p_0q_1}. \qquad (7.7)$$

When $\psi = 1$, i.e., $p_1 = p_0$, the matched pairs estimate is thus seen to be fully efficient. Otherwise eff $< 1$, reflecting the loss in information due to the random pairing. Nevertheless Figure 7.1 shows that the loss, which tends to be worse for intermediate values

Fig. 7.1 Loss in efficiency with a matched-pair design of using a matched statistical analysis, when the matching was unnecessary to avoid bias. Different curves correspond to different proportions exposed in the control population.



of $p_0$, is not terribly important unless the odds ratios being estimated are rather extreme. Pike, Hill and Smith (1979) reach similar conclusions on the basis of studies of the power of the chi-square test of the null hypothesis computed from the matched *versus* unmatched data.

While no additional theoretical studies have yet been made, it is likely that these same general conclusions regarding the bias and efficiency of matched *versus* unmatched *analyses* apply also to the estimation of multiple relative risk functions. Two numerical examples will serve to illustrate the basic points. The first contrasts the fitting of both conditional and unconditional logistic regression analyses to data from an IARC sponsored study of oesophageal cancer occurring among Singapore Chinese (de Jong et al., 1974). The analysis was based on 80 male cases and on 320 matched controls whose ages were within five years of the corresponding case. Two controls for each case were drawn from the same hospital ward as the case, while two others were selected from an orthopaedic unit. However, as there were no important differences in exposure histories between the two control groups, they were not separated in the analysis.

Table 7.10 Coefficients (± standard errors) of variables in the multiple relative risk function, estimated using linear logistic regression analyses appropriate for both matched and unmatched samples. IARC study of oesophageal cancer among Singapore Chinese[a]

| Variables in equation[b] | Matched analysis<br>Coefficient ± S.E. | Unmatched analysis<br>Coefficient ± S.E. |
|---|---|---|
| | A. Interaction term excluded | |
| $x_0$ Constant | | −3.2062 ± 0.3650 |
| $x_1$ Dialect | 1.2570 ± 0.3273 | 1.4145 ± 0.3301 |
| $x_2$ Samsu | 0.5064 ± 0.2936 | 0.5352 ± 0.2766 |
| $x_3$ Cigarettes | 0.0122 ± 0.0099 | 0.0121 ± 0.0095 |
| $x_4$ Beverage temperature | 0.7846 ± 0.1640 | 0.7556 ± 0.1493 |
| Goodness-of-fit statistic (G) | 197.43 | 336.23 |
| | B. Interaction term included | |
| $x_0$ Constant | | −3.2123 ± 0.3661 |
| $x_1$ Dialect | 1.2559 ± 0.3280 | 1.4200 ± 0.3312 |
| $x_2$ Samsu | 0.5072 ± 0.2941 | 0.5303 ± 0.2774 |
| $x_3$ Cigarettes | 0.0123 ± 0.0099 | 0.0124 ± 0.0096 |
| $x_4$ Beverage temperature | 0.7872 ± 0.1726 | 0.7447 ± 0.1563 |
| $x_5 = x_4 \times$ (age-60) | −0.0009 ± 0.0179 | 0.0034 ± 0.0147 |
| Goodness-of-fit statistic (G) | 197.43 | 336.18 |

[a] de Jong et al. (1974)

[b] Coding of risk variables:

$x_1 =$ 1 Hokkien/Teochew
0 Cantonese/other

$x_3 =$ No. of cigarettes/day average

$x_2 =$ 1 Drinkers (Samsu)
0 Abstainers

$x_4 =$ No. of beverages (0–3) drunk "burning hot"

Information was obtained regarding diet, alcohol and tobacco usage, and on various social factors including dialect group, which indicates the patient's ancestral origin within China. Only four variables are considered here: dialect group, cigarettes, samsu (a distilled liquor made from a mixture of grains) and beverage temperature (the number of beverages among tea, coffee and barley wine that the patient reported drinking at "burning hot" temperatures). The coding of these variables has been simplified from that used in the original analysis, and an interaction term between beverage temperature and age (a matching variable) was introduced to see if the log relative risk for beverage temperature changed linearly with age.

Table 7.10 presents the estimated regression coefficients and standard errors obtained by fitting the unconditional logistic model with a single stratum parameter $\alpha$ to the pooled data. Shown for comparison are the same quantities estimated from the conditional likelihood. With the exception of that for dialect group, the standard errors of the matched analysis are slightly larger than those for the unmatched. Small changes are evident in the regression coefficients themselves, so that this is evidently a situation in which the matching variables either have little relationship to the exposures conditional on disease status or else have little relationship to disease status conditional on exposure. As a partial confirmation of the latter interpretation, Table 7.11 shows that cases and controls have roughly equivalent average ages even within the levels of each risk factor. This analysis is incomplete, since it involves only averages and ignores possible higher order interactions of age with risk factor combinations. Nevertheless, it is consistent with the notion that the matching variables are conditionally independent of disease status given the exposures, and thus that the requirements for 'poolability'of matched data are satisfied.

Table 7.11 Average ages ± standard errors for cases and controls within levels of each risk factor: IARC study of oesophageal cancer among Singapore Chinese[a]

| Risk factor | Level | Cases | | Controls | | Totals | |
|---|---|---|---|---|---|---|---|
| | | n | Mean ± S.E. | n | Mean ± S.E. | n | Mean ± S.E. |
| Dialect group | Hokkien/Teochew | 66 | 61.3 ± 1.0 | 160 | 60.6 ± 0.8 | 226 | 60.8 ± 0.6 |
| | Cantonese/other | 14 | 65.4 ± 2.6 | 160 | 63.0 ± 0.7 | 174 | 63.2 ± 0.6 |
| Samsu | Drinkers | 40 | 63.6 ± 1.2 | 109 | 62.4 ± 0.8 | 149 | 62.7 ± 0.7 |
| | Abstainers | 40 | 60.5 ± 1.4 | 211 | 61.5 ± 0.6 | 251 | 61.4 ± 0.6 |
| Cigarettes | None | 8 | 63.6 ± 5.4 | 55 | 62.8 ± 1.3 | 63 | 62.9 ± 1.3 |
| | 1–10 per day | 14 | 65.9 ± 1.9 | 81 | 63.7 ± 1.0 | 95 | 64.0 ± 0.9 |
| | 11–20 per day | 35 | 61.7 ± 1.0 | 115 | 62.2 ± 0.8 | 150 | 62.1 ± 0.7 |
| | 21+ per day | 23 | 59.6 ± 1.8 | 69 | 58.2 ± 1.0 | 92 | 58.5 ± 0.9 |
| Beverage temperature (no. "burning hot") | 0 | 41 | 60.8 ± 1.4 | 261 | 61.5 ± 0.6 | 302 | 61.4 ± 0.5 |
| | 1 | 13 | 62.2 ± 2.1 | 31 | 62.8 ± 1.6 | 44 | 62.6 ± 1.3 |
| | 2 | 18 | 65.3 ± 1.9 | 25 | 63.6 ± 1.9 | 43 | 64.3 ± 1.3 |
| | 3 | 8 | 60.5 ± 2.8 | 3 | 66.3 ± 3.2 | 11 | 62.1 ± 2.3 |
| Totals | All | 80 | 62.0 ± 0.9 | 320 | 61.8 ± 0.5 | 400 | 61.9 ± 0.4 |

[a] de Jong et al. (1974)

Table 7.12 Coefficients (± standard errors) of variables in the multiple relative risk function, using a variety of analyses: Iran/IARC case-control study of oesophageal cancer in the Caspian littoral of Iran[a]

| Variables in equation | Type of analysis | | | | | |
|---|---|---|---|---|---|---|
| | | Stratified into | | | | |
| | Fully matched | 7 Regions, 4 Age groups | 4 Regions, 4 Age groups | 4 Regions | 4 Age groups | Unmatched |
| Social class | –1.125 ± 0.254 | –0.808 ± 0.212 | –0.782 ± 0.206 | –0.745 ± 0.201 | –0.684 ± 0.180 | –0.682 ± 0.179 |
| Ownership of garden | –0.815 ± 0.250 | –0.614 ± 0.222 | –0.602 ± 0.219 | –0.592 ± 0.218 | –0.326 ± 0.191 | –0.307 ± 0.190 |
| Consumption of raw green vegetables | –0.552 ± 0.220 | –0.459 ± 0.203 | –0.439 ± 0.199 | –0.432 ± 0.198 | –0.429 ± 0.188 | –0.440 ± 0.187 |
| Consumption of cucumbers | –0.640 ± 0.217 | –0.539 ± 0.196 | –0.548 ± 0.192 | –0.562 ± 0.192 | –0.466 ± 0.182 | –0.449 ± 0.181 |
| Goodness-of-fit (G) | 375.38[b] | 776.54 | 777.60 | 780.80 | 787.04 | 789.56 |

[a] Cook-Mozaffari et al. (1979)

[b] Based on the conditional model and hence not comparable to the others

In general one must anticipate that the degree to which the matching variables are incorporated in the analysis will affect the estimates of relative risk. An example which better illustrates this phenomenon is provided by the joint Iran/IARC study of oesophageal cancer on the Caspian littoral (Cook-Mozaffari et al., 1979). In that part of the world both cancer incidence and many environmental variables show marked geographical variation. Cases and controls were therefore individually matched according to village of residence, as well as for age. Just as in the preceding example, the data were analysed using both the conditional fully matched analysis based on (7.2) and the unconditional analysis based on (6.10) in which the entire sample was considered as a single stratum. Intermediate between these two extremes, additional analyses were performed which incorporated various levels of stratification by age and by geographical area, the latter grouping the villages into regions with roughly homogeneous incidence.

Table 7.12 presents the results for males for four risk variables which appeared to be the best indicators of socioeconomic and dietary status. Substantial bias of the regression coefficients towards the origin, indicating a lesser effect on risk, is evident with the coarsely stratified and unmatched analyses. This confirms the theoretical results regarding the direction of the bias which were noted above to hold for the univariate situation. While the standard errors of the estimates increase as greater account is taken of the matching, the changes are not great and seem a small price to pay for avoiding bias.

In summary, both theoretical and numerical studies confirm that the pooling of matched or stratified samples for analysis will result in relative risk estimates which are conservatively biased in comparison with those which would be obtained using the appropriate matched analysis. In some situations, where the matching was not essential to avoid bias, the pooled and matched estimates may scarcely differ at all. Even then, however, the additional information gained from the pooled data, as reflected in the variances of the estimates, is not great. Consequently, now that appropriate and flexible methods are available for doing so, the *matching should be accounted for in the analysis whenever it has been incorporated in the design.*

While the availability of methods for multivariate analysis of matched samples certainly makes such designs more attractive, it does not follow that they should always be used. Close pair matching may result in a number of cases being lost from the study for want of an appropriate match. It may also impose severe administrative costs which could be avoided with a less restrictive design. Increasing use is being made of "population controls" obtained as an age-stratified random sample of the population from which the cases were diagnosed. Many epidemiologists believe that this is the best way to avoid the selection bias inherent in other choices of the control population. The confounding effects of other factors which are causally related to disease may be accounted for by post-hoc stratification of the sample, or by modelling them in the analysis. Such designs and analyses accomplish many of the aims intended by the use of matching, and constitute a practical alternative which may be preferred in many situations.

## REFERENCES

Andersen, E.B. (1973) *Conditional Inference and Models for Measuring,* Copenhagen, Mental Hygienisk Forlag., p. 69

Armitage, P. (1975) *The use of the cross-ratio in aetiological surveys.* In: Gani, J., ed., *Perspectives in Probability and Statistics,* London, Academic Press, pp. 349–355

Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975) *Discrete Multivariate Analysis: Theory and Practice,* Cambridge, Mass., MIT Press

Breslow, N.E. (1976) Regression analysis of the log odds ratio: a method for retrospective studies. *Biometrics, 32,* 409–416

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. Epidemiol., 108,* 299–307

Cook-Mozaffari, P.J., Azordegan, F., Day, N.E., Ressicaud, A., Sabai, C. & Aramesh, B. (1979) Oesophageal cancer studies in the Caspian littoral of Iran: results of a case-control study. *Br. J. Cancer, 39,* 293–309

Cox, D.R. & Hinkley, D.V. (1974) *Theoretical Statistics,* London, Chapman & Hall

Day, N.E. & Byar, D. (1979) Testing hypotheses in case-control studies: equivalence of Mantel-Haenszel statistics and logit score tests. *Biometrics, 35,* 623–630

de Jong, U.W., Breslow, N.E., Goh Ewe Hong, J., Sridharan, M. & Shanmugaratnam, K. (1974) Aetiological factors in oesophageal cancer in Singapore Chinese. *Int. J. Cancer, 13,* 291–303

Fienberg, S.E. (1977) *The Analysis of Cross-Classified Categorical Data,* Cambridge, Mass., MIT Press

Holford, T.R., White, C. & Kelsey, J.L. (1978) Multivariate analysis for matched case-control studies. *Am. J. Epidemiol., 107,* 245–256

Jick, H., Watkins, R.N., Hunter, J.R., Dinan, B.J., Madsen, S., Rothman, K.J. & Walker, A.M. (1979) Replacement estrogens and endometrial cancer. *New Engl. J. Med., 300,* 218–222

Liddell, F.D.K., McDonald, J.C. & Thomas, D.C. (1977) Methods of cohort analysis: appraisal by application to asbestos mining. *J. R. stat. Soc. Ser. A, 140,* 469–491

Mack, T.M., Pike, M.C., Henderson, B.E., Pfeffer, R. I., Gerkins, V.R., Arthur, B.S. & Brown, S.E. (1976) Estrogens and endometrial cancer in a retirement community. *New Engl. J. Med., 294,* 1262–1267

Miettinen, O.S. (1974) Confounding and effect modification. *Am. J. Epidemiol., 100,* 350–353

Pike, M.C., Hill, A.P. & Smith, P.G. (1980) Bias and efficiency in logistic analyses of stratified case-control studies. *Int. J. Epidemiol., 9,* 89–95

Richards, F.S.G. (1961) A method of maximum likelihood estimation. *J. R. stat. Soc. B., 23,* 469–475

Seigel, D.G. & Greenhouse, S.W. (1973) Multiple relative risk functions in case-control studies. *Am. J. Epidemiol., 97,* 324–331

Smith, D.C., Prentice, R., Thompson, D.J. & Herrmann, W.L. (1975) Association of exogenous estrogen and endometrial carcinoma. *New Engl. J. Med., 293,* 1164–1167

Smith, P.G., Pike, M.C., Hill, A.P., Breslow, N.E. & Day, N.E. (1981) Multivariate conditional logistic analysis of stratum-matched case-control studies (submitted for publication)

Thomas, D.G. (1975) Exact and asymptotic methods for the combination of 2×2 tables. *Comput. biomed. Res., 8,* 423–446

Whittemore, A.S. (1978) Collapsibility of multidimensional contingency tables. *J. R. stat. Soc. B., 40,* 328–340

Zelen, M. (1971) The analysis of several 2×2 tables. *Biometrika, 58,* 129–137

## LIST OF SYMBOLS – CHAPTER 7 (in order of appearance)

| | |
|---|---|
| $\beta_k$ | log relative risk associated with unit change in kth risk variable |
| $\mathbf{x}_j$ | vector of risk variables for jth study subject; $\mathbf{x}_j = (x_{j1}, \ldots, x_{jk})$ |
| $n_1$ | number of cases |
| $n_0$ | number of controls |
| n | total number of study subjects |
| $\mathbf{l}$ | denotes a partition of the integers from 1 to n into two groups, one of size $n_1$ and the other of size $n_0 = n - n_1$; e.g., if $n_1{}' = 2$ and $n_0 = 3$ a possible partition is $l_1 = 3, l_2 = 4, l_3 = 1, l_4 = 2, l_5 = 5$ or $\mathbf{l} = (3,4,1,2,5)$ |
| $\alpha_i$ | logit of disease probability for an individual with a standard ($\mathbf{x} = \mathbf{0}$) set of risk variables in the ith stratum |
| $pr_i(y = 1/x)$ | disease probability in the ith stratum for an individual with value x for the risk variable |
| $\psi$ | odds ratio |
| $\beta$ | log relative risk (binary exposure) |
| $n_{00}$ | number of matched pairs with neither case nor control exposed |
| $n_{01}$ | number of matched pairs with case unexposed and control exposed |
| $n_{10}$ | number of matched pairs with case exposed and control unexposed |
| $n_{11}$ | number of matched pairs with both case and control exposed |
| $\mu$ | in discordant matched pairs with a binary exposure variable, denotes the fitted number of exposed cases under the unconditional model |
| $\pi$ | conditional probability that in a discordant matched pair it is the case which is exposed |
| M | number of controls per case (fixed) |
| $M_i$ | number of controls per case in the ith matched set |
| I | number of matched sets |
| $x_{ijk}$ | value of kth exposure variable (k = 1, ..., K) for case (j = 0) or jth control (j = 1, ..., $M_i$) in the ith matched set |
| $\mathbf{x}_{ij}$ | $(x_{ij1}, \ldots, x_{ijK})$ exposure vector for jth subject in ith set |
| G | goodness-of-fit statistic based on the (conditional) log likelihood |
| $a_i$ | number of exposed cases in ith of I 2×2 tables |
| $n_{1i}$ | number of cases in ith table |
| $n_{0i}$ | number of controls in ith table |
| $\psi_i$ | (expected) odds ratio associated with ith of I 2×2 tables |
| $z_{il}$ | value of lth covariable for ith 2×2 table |

| | |
|---|---|
| $\mathbf{z}_i$ | vector of covariable values for $i^{th}$ table |
| $\gamma$ | vector of interaction parameters in logistic model for a series of 2×2 tables |
| $p_{1i}$ | exposure probability for cases in the $i^{th}$ stratum |
| $q_{1i}$ | $1-p_{1i}$ |
| $p_{0i}$ | exposure probability for controls in the $i^{th}$ stratum |
| $q_{0i}$ | $1-p_{0i}$ |
| $\vartheta_i$ | $p_{0i}/q_{0i}$ |