EXAMPLE: BCG VACCINATION AND LEPROSY

The data in Table 16.1 are from a rather unusual example of a case-control study in which the controls were obtained from a 100% cross-sectional survey of the study base.* The aim of the study was to investigate whether BCG vaccination in early childhood, whose purpose is to protect against tuberculosis, confers any protection against leprosy, which is caused by a closely related bacillus. New cases of leprosy reported during a given period in a defined geographical area were examined for presence or absence of the characteristic scar left by BCG vaccination. During approximately the same period, a 100% survey of the population of this area had been carried out, and this survey included examination for BCG scar. The tabulated data refer only to subjects under 35, because persons over the age of 35 at the time of the study would have been children at a time when vaccination was not widely available.

---

*From Fine, P.E.M. *et al.* (1986) *The Lancet*, August 30 1986, 499–502.

Table 16.1. BCG scar status in new leprosy cases and in a healthy population survey

BCG scar	Leprosy cases	Population survey
Present	101	46 028
Absent	159	34 594

Table 16.2. A simulated study with 1000 controls

BCG scar	Leprosy cases	Population survey
Present	101	554
Absent	159	446

Exercise 16.1. Estimate the odds of BCG vaccination for leprosy cases and for the controls. Estimate the odds ratio and hence the extent of protection against leprosy afforded by vaccination.

18
Comparison of odds within strata

This chapter deals with methods for analysing stratified case-control studies which closely parallel the methods for cohort studies discussed in Chapter 15.

18.1 The constant odds ratio model

As an example we return to the study of the effect of BCG vaccination upon the incidence of leprosy. Since leprosy incidence increases with age among young people, age is certainly a variable which would have been controlled in an experiment. In Chapter 16 it was shown that BCG-vaccinated individuals had just under one half of the incidence of leprosy as compared with unvaccinated persons, but age was ignored in the analysis. This could have biased the estimated effect of BCG vaccination because BCG vaccination in the area (Northern Malawi) was introduced gradually in infants and young children, so that people who were older during the study period, having been born at earlier dates, were less likely to have been vaccinated. As a result, on average the vaccinated group will be younger than the unvaccinated group. This means that, even if BCG vaccination were totally ineffective, one would expect to observe lower rates in vaccinated members of the base cohort, simply as a result of their relative youth.

Table 18.1 subdivides these data by strata corresponding to 5-year age

Table 18.1. BCG vaccination and leprosy by age

	BCG scar				Odds
	Leprosy cases		Healthy population		ratio
Age	Absent	Present	Absent	Present	estimate
0–4	1	1	7593	11719	0.65
5–9	11	14	7143	10184	0.89
10–14	28	22	5611	7561	0.58
15–19	16	28	2208	8117	0.48
20–24	20	19	2438	5588	0.41
25–29	36	11	4356	1625	0.82
30–34	47	6	5245	1234	0.54

bands. The table also shows age-specific odds ratios. Although there is random variation, there is no systematic trend of the odds ratio with age, and it seems reasonable to make the assumption that the odds ratio parameter is the same in all age bands. In the next section we show how an estimate of this common odds ratio can be calculated.

18.2 An estimate of the common odds ratio

In the prospective approach to the analysis, the assumption of a common odds ratio implies that ω_1^t/ω_0^t is constant, so that the model can be expressed in terms of the odds ratio parameter θ and the ω_0^t parameters. Alternatively, in the retrospective approach the model is expressed in terms of θ and the parameters Ω_0^t. In both approaches, replacing the nuisance parameters by their estimates leads to the profile likelihood for θ. If there are not too many strata, and the data are not too sparse in each stratum, then the profile likelihood for θ can be used to find the most likely value and the supported range. For coarsely stratified data sets such as Table 18.1, these conditions are met. Such an analysis is not feasible by hand, but would usually be carried out on a computer using *logistic regression* (see Chapter 23).

When the data are very finely stratified so that each stratum contains very few cases and controls, the profile likelihood approach can be unreliable, and the hypergeometric likelihood should be used. The total log likelihood is then obtained by adding together the hypergeometric log likelihoods for the different strata. Again, the most likely value M and the standard deviation S cannot usually be computed by hand, but would be carried out using a *conditional* logistic regression program (see Chapter 29). However, the calculations for the score test for $\theta = 1$ are straightforward. For a single stratum the score under the hypergeometric likelihood is

$$U = D_1 - E_1$$

where D_1 is the observed number of exposed cases and $E_1 = DN_1/N$ is the expected number under the null hypothesis. The score variance is

$$V = \frac{DHN_0N_1}{(N)^2(N-1)}.$$

Since every stratum contributes additively to the overall log likelihood, the overall score is a sum of contributions from each stratum of exactly the same form as above. Thus, the score is

$$U = \sum (D_1^t - E_1^t)$$

where

$$E_1^t = D^t \frac{N_1^t}{N^t}.$$

and the overall score variance is

$$V = \sum \frac{D^t H^t N_0^t N_1^t}{(N^t)^2(N^t - 1)}.$$

Exercise 18.1. Show that the first age band in Table 18.1 makes a contribution of -0.21 to U and 0.48 to V.

The overall test statistic is obtained by repeating these calculations for each stratum and yields

$$U = -0.21 - 0.69 - 6.68 - 6.56 - 8.11 - 1.76 - 4.06 = -28.07$$

and

$$V = 0.48 + 6.05 + 12.18 + 7.38 + 8.22 + 9.22 + 8.09 = 51.62.$$

The approximate chi-squared value on one degree of freedom is

$$(U)^2/V = 787.92/51.62 = 15.26.$$

The statistic U has a negative sign because the exposure is protective — the observed number of vaccinated cases is less than would have been expected had vaccination been ineffective.

Exercise 18.2. Verify that, when there is only one case per stratum, the test becomes identical to the log rank test discussed in section 15.5.

This test was proposed by Mantel and Haenszel. They also proposed a way of calculating a nearly most likely value for θ. This is suggested by an algebraic rearrangement of the equation for the score:

$$\begin{aligned} U &= \sum (D_1^t - E_1^t) \\ &= \sum \frac{D_1^t H_0^t - D_0^t H_1^t}{N^t} \\ &= \sum Q^t - \sum R^t, \end{aligned}$$

where $Q^t = D_1^t H_0^t / N^t$ and $R^t = D_0^t H_1^t / N^t$. The usual estimate of the odds ratio in stratum t is Q^t/R^t, and this suggests estimating the common odds ratio, θ, by

$$\frac{Q^1 + Q^2 + \ldots}{R^1 + R^2 + \ldots} = \frac{Q}{R}.$$

When the true value of θ is close to 1, this *Mantel–Haenszel estimate* is almost as precise as the the most likely value of θ according to the hypergeometric likelihood. It can only be improved upon for odds ratios which differ substantially from one.

Exercise 18.3. Show that the Mantel–Haenszel estimate of the odds ratio for the data of Table 18.1 is 0.587.

Note that allowing for confounding by age has weakened the estimated protective effect of vaccination. This is now about 41% rather than 52% — a modest adjustment. This is in accord with the general experience that confounding only causes substantial modification of rate ratios in quite extreme circumstances.

The usefulness of the Mantel–Haenszel estimate in practice was limited by the fact that, rather surprisingly, no expression was available for its standard deviation until relatively recently. Several estimates have now been proposed, most of them rather awkward to calculate. For most prac-

23.2 Logistic regression

In logistic regression the original parameters are odds parameters and these are expressed in terms of new parameters in the same way as for the rate parameter. The most important application of logistic regression is to case-control studies and we shall use the study of BCG and leprosy as an illustration.

For convenience the data from this study are repeated in Table 23.2, which shows the numbers of cases and controls by age and BCG vaccination. Taking a prospective view the response parameter is the odds of being a case rather than a control, so a useful way of summarizing these data is to

Table 23.2. Cases of leprosy and controls by age and BCG scar

Age	Leprosy cases		Healthy controls	
	Scar −	Scar +	Scar −	Scar +
0–4	1	1	7 593	11 719
5–9	11	14	7 143	10 184
10–14	28	22	5 611	7 561
15–19	16	28	2 208	8 117
20–24	20	19	2 438	5 588
25–29	36	11	4 356	1 625
30–34	47	6	5 245	1 234

Table 23.3. Case/control ratio ($\times 10^3$) by age and BCG scar

	BCG scar	
Age	Absent	Present
0–4	0.13	0.08
5–9	1.54	1.37
10–14	4.99	2.91
15–19	7.25	3.45
20–24	8.20	3.40
25–29	8.26	6.77
30–34	8.96	4.86

show the estimated value of this parameter, which is the case/control ratio, for different levels of age and BCG vaccination. This summary is given in Table 23.3 and shows a consistently lower case/control ratio for those with a BCG scar than for those without. It also shows that the case/control ratio increases sharply with age in both groups.

Because there are many subjects in this study the data are entered to the computer program as frequency records. Table 23.4 shows the data as an array of frequency records ready for computer input. Programs often require the data to be entered as the number of cases and the total number of subjects for each record, rather than as the number of cases and the number of controls. The change is easily made by deriving a new variable equal to the variable for the number of cases plus the variable for the number of controls.

The log likelihood contribution for a frequency record in which N subjects split as D cases and H controls takes the Bernoulli form

$$D\log(\omega) - N\log(1+\omega),$$

where ω is the odds, given by the model, that a subject in that frequency

Table 23.4. The BCG data as frequency records

Cases	Total	Scar	Age
1	7594	0	0
1	11720	1	0
11	7154	0	1
14	10198	1	1
28	5639	0	2
22	7583	1	2
16	2224	0	3
28	8145	1	3
20	2458	0	4
19	5607	1	4
36	4392	0	5
11	1636	1	5
47	5292	0	6
6	1240	1	6

record is a case rather than a control. When fitting a regression model the total log likelihood is expressed in terms of new parameters using the regression equations and most likely values of the new parameters are found. For individual records the log likelihood is

$$d\log(\omega) - \log(1+\omega),$$

where $d = 1$ for a case and $d = 0$ for a control. The sum of the log likelihoods for all subjects contributing to a frequency record is equal to

$$D\log(\omega) - N\log(1+\omega),$$

which is the same as the log likelihood for the frequency record.

The regression model

$$\log\,(\text{Odds}) = \text{Corner} + \text{Age} + \text{BCG},$$

expresses the constraint that the odds ratio for BCG vaccination is constant over age groups. Apart from the corner, all the parameters in this model are odds ratios. The BCG parameter compares the odds of being a case for subjects who are BCG positive to the odds of being a case for subjects who are BCG negative. The six age parameters compare the odds of being a case for subjects in the age groups 1–6 to the odds of being a case in age group 0. The most likely values of these parameters (on a log scale) are shown in Table 23.5.

Exercise 23.1. What is the most likely value of the odds ratio for BCG vac-

Table 23.5. Output from a logistic regression program

Parameter	Estimate	SD
Corner	−8.880	0.7093
Age(1)	2.624	0.7340
Age(2)	3.583	0.7203
Age(3)	3.824	0.7228
Age(4)	3.900	0.7244
Age(5)	4.156	0.7224
Age(6)	4.158	0.7213
BCG(1)	−0.547	0.1409

cination? Does this seem about right, from Table 23.3? Compare this estimate with the Mantel–Haenszel estimate given in Chapter 18.

The parameters in the model

$$\log(\text{Odds}) = \text{Corner} + \text{Age} + \text{BCG},$$

apart from the corner, refer to changes in the log odds of being a case.