29
Conditional logistic regression

In an individually matched case-control study, it is necessary to introduce a new parameter for every case-control set, if the matching is to be preserved in the analysis. This means that the number of parameters in the model exceeds the number of cases and in this case the profile likelihood does not lead to sensible estimates. Instead the nuisance parameters must be eliminated using a conditional likelihood. In Chapter 19 we indicated how this is done for a simple binary exposure. In this chapter we show how to use a conditional likelihood with the logistic regression model.

29.1 The logistic model

Suppose we wish to fit a logistic regression model which contains parameters for the case-control sets in addition to parameters for the effects of two explanatory variables A and B. Using a categorical variable to define the set to which each subject belongs, the model would be written

$$\log(\text{Odds}) = \text{Corner} + \text{Set} + \text{A} + \text{B}.$$

The model can also be written in the multiplicative form as

$$\text{Odds} = \text{Corner} \times \text{Set} \times \text{A} \times \text{B}.$$

For the case where A has three levels and B has two levels, the parameters in this model are Corner, A(1), A(2), B(1), together with

$$\text{Set}(1),\ \text{Set}(2),\ \cdots,\ \text{Set}(N-1)$$

where N is the number of case-control sets. These set parameters are those used in standard logistic regression models, but they are no longer the most convenient choice. It is now more convenient to choose a separate corner for each set, namely the odds parameter for each set when A and B are at level 0. The corner for the first case-control set is the corner parameter referred to above, the corner for the second case-control set is

$$\text{Corner} \times \text{Set}(1),$$

and so on. This corresponds to splitting the terms in the model into two groups, as follows:

$$\text{Odds} = \boxed{\text{Corner} \times \text{Set}} \times \boxed{\text{A} \times \text{B}}.$$

The first part of the model contains the separate corners, and these are the nuisance parameters to be eliminated, while the second part contains the effects of interest. When a conditional logistic program is used to fit this model the nuisance parameters are eliminated using conditional likelihood and estimates of the effects of A and B are reported. No estimates of either the corner or the set parameters are obtained in this method, so none can be reported.

To see how the nuisance parameters are eliminated using conditional likelihood it is convenient to return to the algebraic notation for parameters using Greek letters. For any particular case-control set let the corner parameter be ω_{C}. Let the odds for any subject in the set be ω_i, where $i = 1, 2, \ldots$, indexes the subjects within the case-control set, and write

$$\omega_i = \omega_{\text{C}}\theta_i,$$

so that θ_i is the ratio of the odds for subject i to the corner odds. The way θ is related to the effects of A and B is determined by the $\boxed{\text{A}\times\text{B}}$ part of the model. The corner parameter refers to subjects within the set with both A and B at level 0, so that the value of θ for such subjects is 1. For subjects with A at level 1 and B at level 0,

$$\theta = \text{A}(1),$$

for subjects with A at level 1 and B at level 1,

$$\theta = \text{A}(1) \times \text{B}(1),$$

and so on.

To be specific about which case-control set is being referred to, the parameters should be written with superscripts t, as in

$$\omega_i^t = \omega_{\text{C}}^t\theta_i^t.$$

where $t = 0, 1, 2, \ldots$ refers to the levels of the variable defining set membership. The parameters ω_{C}^t correspond to the

$$\boxed{\text{Corner} \times \text{Set}}$$

part of the model, and are the nuisance parameters to be eliminated. In the rest of this chapter we shall derive the contribution to the conditional

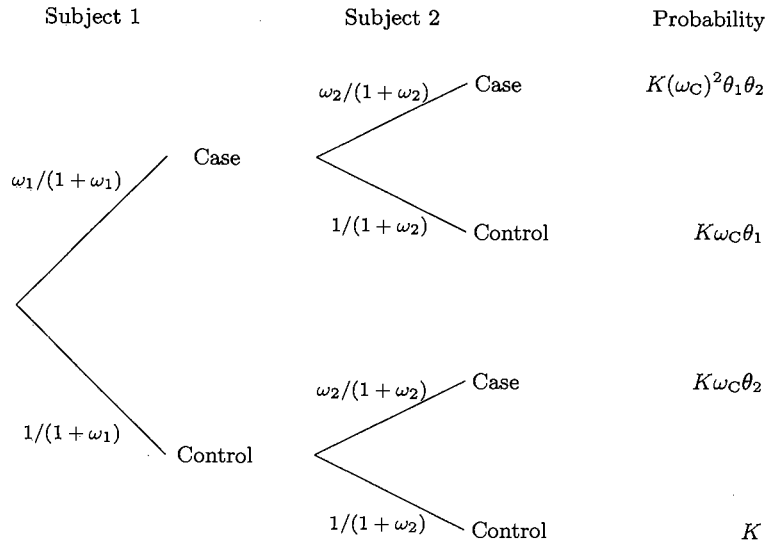

Fig. 29.1. Disease status for two subjects in a case-control study.

log likelihood for a single case-control set, and shall therefore omit the t superscript. The total log likelihood is found by adding the contributions from the single sets.

29.2 The conditional likelihood for 1:1 matched sets

First we derive the contribution for case-control studies with one case and one control in each set. The possible case or control status for any two subjects are represented as a probability tree in Fig. 29.1. Using the relationship between odds and probability, the probabilities that subject 1 is a case or a control are $\omega_1/(1+\omega_1)$ and $1/(1+\omega_1)$ respectively. Similarly, the probabilities for subject 2 are $\omega_2/(1+\omega_2)$ and $1/(1+\omega_2)$. The probabilities of the outcomes for the pair of subjects are obtained by multiplying along branches of the tree in the usual way. The last column of the figure shows such probabilities, after writing

$$\omega_1 = \omega_C\theta_1, \qquad \omega_2 = \omega_C\theta_2,$$

and

$$K = \frac{1}{1+\omega_1} \times \frac{1}{1+\omega_2}.$$

These probabilities refer to any two subjects from the study base. Conditional on the fact that one of the subjects is a case and the other is a control, the probability that subject 1 is the case is

$$\frac{K\omega_C\theta_1}{K\omega_C\theta_1 + K\omega_C\theta_2} = \frac{\theta_1}{\theta_1+\theta_2}.$$

and the probability that subject 2 is the case is

$$\theta_2/(\theta_1+\theta_2).$$

The contribution to the log likelihood of the case-control set is, therefore

$$\log\left(\frac{\theta_{(\text{for case})}}{\theta_{(\text{for case})} + \theta_{(\text{for control})}}\right).$$

This way of writing the log likelihood makes it clear that it does not depend on the arbitrary numbering of the subjects in the pair but only on the expressions for θ in terms of A(1), A(2) and B(1), the parameters to be estimated. The total log likelihood thus depends only on A(1), A(2), and B(1), and the nuisance parameters ω_C^t have been eliminated.

Exercise 29.1. Table 29.1 shows the data for the first two case-control sets in a 1:1 matched study. The set variable indicates which set each subject belongs to, and case or control status is indicated using a variable taking the value 1 for cases and 0 for controls. Illustrative parameter values for the multiplicative effects of the explanatory variables age and exposure, where age has three levels ($< 55, 55-64, 65-74$) and exposure has two levels, are shown below.

Parameter	Value
Age (1)	×1.5
Age (2)	×3.0
Exposure (1)	×5.0

The corner is defined as unexposed and age < 55. Calculate the values of θ predicted by the model for these four subjects. Calculate the log likelihood contributions for the two sets.

Before leaving the 1:1 case we shall verify that the method of obtaining the log likelihood described above gives the same answer as the method described in Chapter 19, for a binary exposure. The model is now

$$\text{Odds} = \boxed{\text{Corner} \times \text{Set}} \times \boxed{\text{Exposure}}$$

which has only one parameter, Exposure(1), apart from the nuisance parameters. This parameter is the multiplicative effect of exposure and we shall refer to it as ϕ. The values of θ for the case and control are determined

Table 29.1. Data file for a 1:1 matched case-control study

Subject	Set	Case/control	Age	Exposure
1	1	1	48	1
2	1	0	64	0
3	2	1	52	1
4	2	0	70	1
...				

Table 29.2. Likelihood contributions for the 1:1 matched study

Exposure	θ for case	θ for control	Likelihood
Neither	1	1	$1/(1+1) = 1/2$
Both	ϕ	ϕ	$\phi/(\phi+\phi) = 1/2$
Case only	ϕ	1	$\phi/(\phi+1)$
Control only	1	ϕ	$1/(1+\phi)$

by whether or not they were exposed. For example, if the case was not exposed then $\theta = 1$, while if the case was exposed then $\theta = \phi$. Similarly for the control. Table 29.2 sets out the four possible outcomes for each case-control set and the corresponding contributions to the log likelihood. The first two outcomes, in which the exposure status of case and control is the same, lead to log likelihood contributions which do not depend upon the parameter, and can be ignored. If N_1 and N_2 are the frequency of occurrence of the remaining outcomes, the total log likelihood is

$$N_1 \log\left(\frac{\phi}{1+\phi}\right) + N_2 \log\left(\frac{1}{1+\phi}\right)$$

which is the same as we obtained in Chapter 19, except that here we have called the effect ϕ rather than θ to avoid confusion.

29.3 The conditional likelihood for 1: m matched sets

We now extend the above argument to sets with one case and m controls. If the sampling had not been carried out deliberately so as to obtain a single case and m controls in the set, the probability that subject 1 is a case and the remaining m subjects are controls would be

$$\frac{\omega_1}{1+\omega_1} \times \frac{1}{1+\omega_2} \times \frac{1}{1+\omega_3} \times \cdots,$$

and making the substitutions

$$\begin{aligned} \omega_i &= \omega_C \theta_i \\ K &= \frac{1}{1+\omega_1} \times \frac{1}{1+\omega_2} \times \frac{1}{1+\omega_3} \times \cdots \end{aligned}$$

this may be written as $K\omega_C\theta_1$. Similarly, the probability that subject 2 is a case and all other subjects controls is $K\omega_C\theta_2$, and so on. The sum of probabilities for all the outcomes in which one member of the set is a case and all other members are controls is

$$K\omega_C(\theta_1 + \theta_2 + \theta_3 + \cdots)$$

so that the conditional probability that subject 1 is the case is:

$$\frac{K\omega_C\theta_1}{K\omega_C(\theta_1 + \theta_2 + \theta_3 + \cdots)} = \frac{\theta_1}{\theta_1 + \theta_2 + \theta_3 + \cdots}.$$

The contribution of one set to the log likelihood is, therefore,

$$\log\left(\theta_{(\text{for case})} \Big/ \sum_{\text{Case-control set}} \theta\right).$$

The total log likelihood is obtained by adding the contributions for all case-control sets.

From the form of this log likelihood it is clear that the conditional approach does not allow estimation of multiplicative effects of variables used in matching. Since all subjects in the set share the same value for such a variable its multiplicative effect will cancel out in the ratio of θ for the case to the sum of all θ's in the case-control set. However, interaction terms involving matching variables *can* be fitted. For example, for a case-control study in which sex was one of the matching variables, the sex effect cannot be estimated but the parameters for interaction between sex and exposure can be, because they will not occur in all of the θ's from the same case-control set.

29.4 Sets containing more than one case ★

The conditional argument can be generalized quite easily to allow for case-control sets containing more than one case, although the computation of the log likelihood may become rather lengthy. The idea is illustrated for a set containing two cases and one control. Fig. 29.2 shows the probability tree for case/control status of a set of three subjects. In three of the eight possible outcomes there are two cases and one control. The probabilities for these branches are written to the right of the figure, again using the

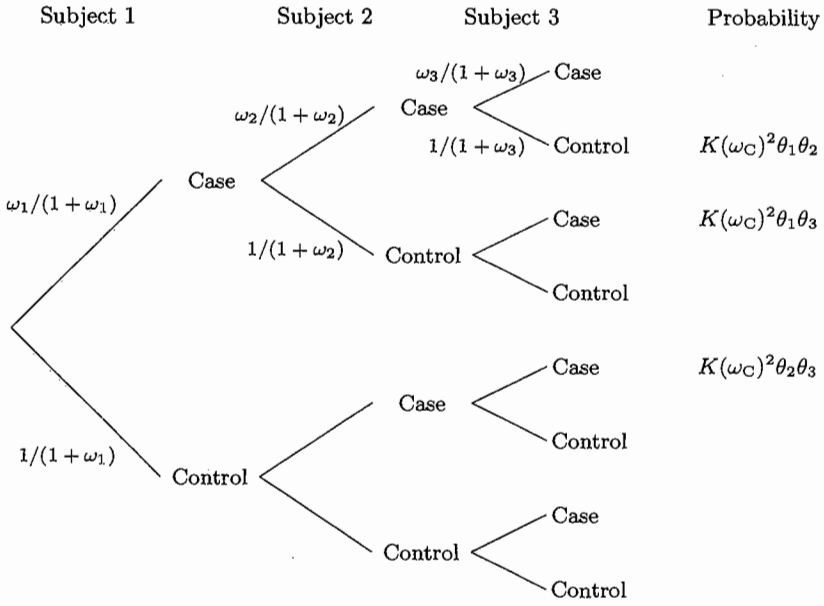

Fig. 29.2. Sets with two cases and one control.

abbreviation

$$K = \frac{1}{1+\omega_1} \times \frac{1}{1+\omega_2} \times \frac{1}{1+\omega_3}.$$

Conditional on the observed outcome being one of the three with two cases and one control the probability that the cases are subjects 1 and 2 is

$$\frac{K(\omega_C)^2\theta_1\theta_2}{K(\omega_C)^2\theta_1\theta_2 + K(\omega_C)^2\theta_1\theta_3 + K(\omega_C)^2\theta_2\theta_3} = \frac{\theta_1\theta_2}{\theta_1\theta_2 + \theta_1\theta_3 + \theta_2\theta_3}.$$

The log of this conditional probability is the contribution of the set to the log likelihood.

It is easy to see how this argument can be extended to deal with any number of cases and controls in a set. For example, for sets of size 6 containing 3 cases, the conditional probability that subjects 1, 2, and 3 are the cases is

$$\frac{\theta_1\theta_2\theta_3}{\theta_1\theta_2\theta_3 + \theta_1\theta_2\theta_4 + \theta_1\theta_2\theta_5 + \cdots}.$$

The denominator contains a term for each of the 20 ways of selecting three subjects from 6, and does not depend on the way the subjects have been numbered.

Solutions to the exercises

29.1 The values of θ for the four subjects are:

		Multiplicative effects		
Subject	Corner	Age	Exposure	θ
1	1.0		×5.0	5.0
2	1.0	×1.5		1.5
3	1.0		×5.0	5.0
4	1.0	×3.0	×5.0	15.0

Subject 1 is the case in the first set and subject 3 is the case in the second set. The log likelihood contributions are, therefore

$$\log\left(\frac{5.0}{5.0+1.5}\right) + \log\left(\frac{5.0}{5.0+15.0}\right) = -0.262 - 1.386.$$