23
Poisson and logistic regression

In principle the way a computer program goes about fitting a regression model is simple. First the likelihood is specified in terms of the original set of parameters. Then it is expressed in terms of the new parameters using the regression equations, and finally most likely values of these new parameters are found. In studies of event data the two most important likelihoods are Poisson and Bernouilli, and the combinations of these with regression models are called *Poisson* and *logistic* regression respectively. Gaussian regression is the combination of the Gaussian likelihood with regression models and will be discussed in Chapter 34.

23.1 Poisson regression

When a time scale, such as age, is divided into bands and included in a regression model, the observation time for each subject must be split between the bands as described in Chapter 6. This is illustrated in Fig. 23.1, where a single observation time ending in failure (the top line) has been split into three parts, the last of which ends in failure. These parts can then be used to make up frequency records containing the number of failures and the observation time, as was done for the ischaemic heart disease data in Table 23.1, or they can be analysed as though they were individual records.

If they are to be analysed as though they were individual records then each of these new records must contain variables which describe which time band is being referred to, how much observation time is spent in the time band, and whether or not a failure occurs in the time band. Values of

Table 23.1. The IHD data as frequency records

Cases	Person-years	Age	Exposure
4	607.9	0	0
2	311.9	0	1
5	1272.1	1	0
12	878.1	1	1
8	888.9	2	0
14	667.5	2	1

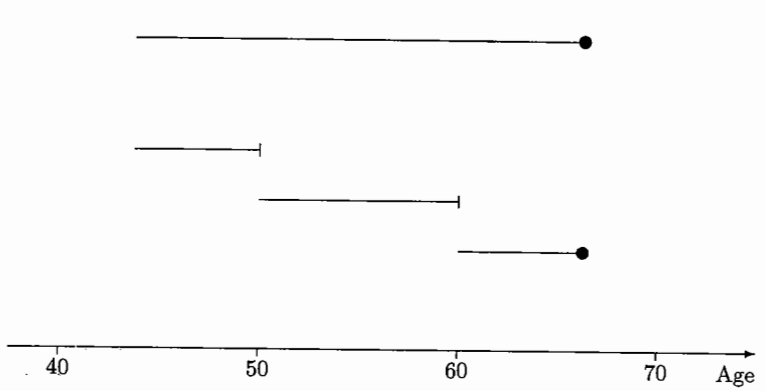

Fig. 23.1. Splitting the follow-up record.

other explanatory variables, such as exposure, must also be included. The idea extends to more than one time scale — each record then refers to an observation of a subject through one cell of a Lexis diagram — but the number of new records can then be many times the number of subjects and analysis becomes cumbersome.

To instruct a computer program to fit a Poisson regression model to the frequency records in Table 23.1 it is first necessary to enter the names of the variables which contain the observation time for the record, the number of failures, the exposure level and the age band. When the Poisson regression option is selected the program automatically assumes that the regression model is of the form

$$\log(\text{Rate}) = \text{Corner} + \text{A} + \text{B} + \ldots,$$

where A, B, etc., are explanatory variables. It is therefore only necessary to instruct the program that the rate for each record is to be calculated from the person-years variable and the number of failures variable, and that exposure and age are to be included in the model as explanatory variables.

The log likelihood for each combination of age band and exposure takes the standard Poisson form. For example when age is at level 2 and exposure is at level 1 the rate parameter is λ_1^2. There are 14 failures and 667.5 person-years so the log likelihood for λ_1^2 is

$$14\log(\lambda_1^2) - 667.5\lambda_1^2.$$

The total log likelihood (in terms of the original parameters) is equal to the sum of the separate log likelihoods for the six cells of the table. This total is expressed (by the computer program) in terms of the four new parameters Corner, Age(1), Age(2), and Exposure(1), using the information provided by the regression model. As usual the most likely values of the log parameters are found on the log scale and some programs leave the user to convert these back to the original scale.

The same log likelihood is obtained from individual records as from frequency records, provided the explanatory variables in the individual records take discrete values in the same way as for the frequency records. For example, the contribution to the log likelihood from a subject with exposure at level 1, age band at level 2, and observation time y, is

$$d\log(\lambda_1^2) - y\lambda_1^2,$$

where d takes the value 1 if the subject fails in this age band and 0 otherwise. Adding this log likelihood over all subjects contributing to the frequency record with exposure at level 1 and age at level 1 gives

$$14\log(\lambda_1^2) - 667.5\lambda_1^2,$$

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

A computer program for Poisson regression can also be used after the confounding effect of age has been allowed for by indirect standardization, that is by calculating the expected number of failures using standard reference rates. This is because the log likelihood for the parameter representing the (common) ratio of age-specific rates in a study group to the age-specific reference rates has the same algebraic form as the log likelihood for a rate parameter; one is obtained from the other by exchanging the person-years and the expected number of failures. With this exchange, the original parameters are now rate ratios expressing age-controlled comparisons of different sections of the study group to the reference rates. The regression model relates these to a smaller number of parameters in the same way as with rates. Note that the parameter estimates in such models are, in effect, ratios of SMRs. For the reasons discussed in Chapter 15, they can be misleading if an inappropriate set of reference rates is used.

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	7593	11719
5–9	11	14	7143	10184
10–14	28	22	5611	7561
15–19	16	28	2208	8117
20–24	20	19	2438	5588
25–29	36	11	4356	1625
30–34	47	6	5245	1234

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. From Chapter 16 we know that the odds of being a case is proportional to the odds of being a failure in the study base, provided the selection of cases and controls is independent of both age and BCG status. More precisely,

$$\text{Odds of being a case} = K\frac{\pi}{1-\pi}$$

where

$$K = \frac{\text{Probability that a failure is sampled as a case}}{\text{Probability that a survivor is sampled as a control}}.$$

On a log scale

$$\log(\text{Odds}) = \log(K) + \log\left(\frac{\pi}{1-\pi}\right),$$

so a change in the log odds of being a case is equal to the corresponding change in the log odds of failure in the study base. It follows that estimates of the effects of age and BCG on the log odds of being a case also estimate the effects of age and BCG on the log odds of failure in the study base. This argument does not apply to the corner (which is not a change in log odds) so unless K is known the corner parameter in the study base cannot be estimated.

Table 23.6. A simulated group-matched study

	BCG scar			
	Cases		Controls	
Age	Absent	Present	Absent	Present
0–4	1	1	3	5
5–9	11	14	48	52
10–14	28	22	67	133
15–19	16	28	46	130
20–24	20	19	50	106
25–29	36	11	126	62
30–34	47	6	174	38

When the disease is rare the probability of failure in the study base is small and the odds of failure are related to the rate λ by

$$\frac{\pi}{1-\pi} \approx \lambda T,$$

where T is the duration of the study. Thus

$$\begin{aligned}\log(\text{Odds}) &= \log(K) + \log\left(\frac{\pi}{1-\pi}\right),\\ &\approx \log(K) + \log(T) + \log(\lambda),\end{aligned}$$

and the same argument shows that effects estimated from a logistic regression model are also estimates of effects on the log rate in the study base.

23.3 Matched case-control studies

In Chapter 18 we presented a simulated group-matched case-control study, based on the BCG study, in which the age distribution of controls is made equal to that of the cases by taking four times as many controls as cases in each age stratum. The results from this study are shown again in Table 23.6.

When estimating the effect of BCG the matching variable, age, cannot be ignored, so the appropriate model to fit is

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

even though the effects of age in this model may be close to zero. The results of fitting this model are shown in Table 23.7. As expected the estimate of the BCG effect is virtually unchanged, although it has a slightly larger standard deviation because it is based on a smaller number of controls.

Table 23.7. Regression output for the group-matched study

Parameter	Estimate	SD
Corner	−1.0670	0.800
Age(1)	−0.0421	0.827
Age(2)	0.0119	0.812
Age(3)	0.0713	0.814
Age(4)	0.0244	0.816
Age(5)	−0.1628	0.814
Age(6)	−0.2380	0.813
BCG(1)	−0.5721	0.155

However, the age effects are very different from the previous output for the whole data set in Table 23.5. They are now all close to zero but this does not mean that age can be omitted from the model. To do so would produce a biased estimate of the BCG effect. Variables which have been used in the matching must be included in the model used to estimate the effects of interest. The same point was made in Chapter 18 where matched case-control studies were analysed by stratifying on the matching variable and using the Mantel–Haenszel method to combine the separate estimates of the effect of interest over strata.

Exercise 23.2. Explain the large differences in the age effects between the two outputs. You may find it helps to make a summary table of case/control ratios based on the data in Table 23.6.

Using a computer program for logistic regression is a convenient way of analyzing group-matched case-control studies and gives correct estimates of odds ratios, at least for variables not used in the matching, provided there are not too many matching strata. However, in individually matched case-control studies each new case introduces its own stratum and, therefore, a new nuisance parameter. This turns out to be one of the situations in which replacing the nuisance parameters by their most likely values and using profile likelihood to estimate the parameters of interest gives the wrong answer. For individually matched studies the likelihood argument of Chapter 19 can be extended to cover regression models. This new method is called *conditional* logistic regression analysis, and will be discussed in Chapter 29.

★ 23.4 Modelling risk and prevalence

The prospective approach to the regression analysis of case-control studies regards the case/control status as the outcome variable. In Chapter 1 we discussed other epidemiological studies in which the outcome of interest

is binary. Most important are studies of risk(sometimes called *cumulative incidence* studies) in which each subject is studied for a fixed period, the outcome being failure or survival, and cross sectional *prevalence studies* in which each subject's present state is recorded as diseased or healthy.

In both these types of study the original parameters are probabilities. For case-control studies, we choose to model odds rather than probabilities because odds ratios are independent of the sampling fractions used and have a ready interpretation as risk or rate ratios in the study base. For risk and prevalence studies there is no such compelling reason to use the odds, although it often proves useful to do so because the log odds is unconstrained and models for the log odds are likely to describe the data better than models for π or $\log(\pi)$.

An alternative to the log odds may be derived from the relationship between π, the probability of failure in a time interval of length T, and λ, the failure rate for this interval. This relationship is given by

$$\text{Cumulative survival probability} = \exp(-\text{ Cumulative failure rate})$$

that is,

$$1 - \pi = \exp(-\lambda T),$$

so

$$\log(1 - \pi) = -\lambda T$$

and

$$\log(-\log(1 - \pi)) = \log(T) + \log(\lambda).$$

Thus models for $\log(-\log(1-\pi))$ may be interpreted as models for $\log(\lambda)$, apart from the corner parameter, and parameters which are estimated from such models may be interpreted as the logarithms of rate ratios. The function $\log(-\log(1-\pi))$ is called the *complementary log-log* transformation of π and some programs allow regression models to be fitted on this scale. Provided π is less than about 0.2 the complementary log-log function does not differ appreciably from the log odds, so in this case regression models for the log odds can also be interpreted as regression models for $\log(\lambda)$.

For diseases in which mortality (and migration) of subjects is unaffected by their contracting the disease, there is a similar relationship between age-specific prevalence and the age-specific incidence rate. In this case, parameters of complementary log-log models for prevalence are identical to parameters of an underlying model for log incidence rates. However in general such an assumption cannot be made and the relationship between effects on prevalence and effects on incidence is complicated.

Solutions to the exercises

23.1 The most likely value of the log of the BCG parameter is -0.547. This corresponds to an odds ratio of $\exp(-0.547) = 0.579$. We therefore estimate that vaccination with BCG reduces the incidence rate of leprosy in the base study to about 58% of what it would be without vaccination. From Chapter 18 the Mantel–Haenszel estimate of the BCG parameter is 0.587.

23.2 The discrepancies between the two outputs is due to the age matching of controls to cases in the second analysis. In the first analysis there is no such matching, and the age parameters refer to the underlying relationship between age and leprosy incidence (incidence increases with age). Matching controls to cases with respect to age has the effect that the sampling probabilities for controls differ between age strata so that K, the constant of proportionality between the odds of being a case and the odds of failure in the study base, now varies between age bands. It follows that the age parameters of the model now include the effect of variation in sampling probabilities, and are not interpretable.