Part II

Regression models

22
Introduction to regression models

One of the main problems discussed in Part I was how to compare two rate parameters, λ_0 and λ_1, using their ratio λ_1/λ_0. To do this the log likelihood for the parameters λ_0 and λ_1 was re-expressed in terms of λ_0 and θ, where $\theta = \lambda_1/\lambda_0$. This technique was then extended to deal with comparisons stratified by a confounding variable by making the assumption that the parameter θ was constant over strata. In this second part of the book, the technique will be further extended to deal with the joint effects of several exposures and to take account of several confounding variables.

A common theme in all these situations is a change from the original parameters to new parameters which are more relevant to the comparisons of interest. This change can be described by the equations which express the old parameters in terms of the new parameters. These equations are referred to as *regression* equations, and the statistical model is called a *regression model.* To introduce regression models we shall first express some of the comparisons discussed in Part I in these terms. We use models for the rate parameter for illustration, but everything applies equally to models for the odds parameter.

22.1 The comparison of two or more exposure groups

When comparing two rate parameters, λ_0 and λ_1, the regression equations which relate the original parameters to the new ones are

$$\lambda_0 = \lambda_0, \qquad \lambda_1 = \lambda_0\theta,$$

where the first of these simply states that the parameter λ_0 is unchanged.

When there are three groups defined by an exposure variable with three levels, corresponding (for example) to no exposure, moderate exposure, and heavy exposure, the original parameters are λ_0, λ_1, and λ_2, and there are now more ways of choosing new parameters. The most common choice is to change to

$$\lambda_0, \qquad \theta_1 = \lambda_1/\lambda_0, \qquad \theta_2 = \lambda_2/\lambda_0.$$

With this choice of parameters the moderate and heavy exposure groups

Table 22.1. A regression model to compare rates by exposure levels

	Exposure	
Age	0	1
0	λ_0^0	$\lambda_0^0\theta$
1	λ_0^1	$\lambda_0^1\theta$
2	λ_0^2	$\lambda_0^2\theta$

are compared to the unexposed group. The regression equations are now

$$\lambda_0 = \lambda_0, \qquad \lambda_1 = \lambda_0\theta_1, \qquad \lambda_2 = \lambda_0\theta_2.$$

22.2 Stratified comparisons

When the comparison between exposure groups is stratified by a confounding variable such as age the change to new parameters is first made separately for each age band; for two exposure groups the regression equations for age band t are

$$\lambda_0^t = \lambda_0^t \qquad \lambda_1^t = \lambda_0^t\theta^t.$$

The parameter θ^t is age-specific and to impose the constraint that it is constant over age bands it is set equal to the constant value θ, in each age band. The regression equations are now

$$\lambda_0^t = \lambda_0^t \qquad \lambda_1^t = \lambda_0^t\theta.$$

This choice of parameters is the same as for the proportional hazards model, introduced in Chapter 15. The model is written out in full in Table 22.1 for the case of three age bands.

Although our main interest is whether the rate parameter varies with exposure, within age bands, we might also be interested in investigating whether it varies with age, within exposure groups. The parameter θ does not help with this second comparison because it has been chosen to compare the exposure groups. When making the comparison the other way round the age bands are the groups to be compared and the exposure groups are the strata. To combine the comparison across these strata requires the assumption that the rate ratios which compare levels 1 and 2 of age with level 0 are the same in both exposure groups. This way of choosing parameters is shown in Table 22.2, where the parameters ϕ^1 and ϕ^2 are the rate ratios for age, assumed constant within each exposure group. Note that there are two parameters for age because there are three age bands being compared.

Putting these two ways of choosing parameters together gives the regression model shown in Table 22.3. The parameter λ_0^0 has now been written as λ_C, for simplicity and to emphasize that it refers to the (top left-hand) corner of the table. Both sorts of comparison can now be made in the same analysis. It is no longer necessary to regard one variable as the exposure, and the other as a confounder used to define strata; the model treats both types of variable symmetrically. To emphasize this symmetry the term *explanatory* variable is often used to describe both exposures and confounders in regression models. Although this is useful in complex situations where there are many variables, there are also dangers. Although it makes no difference to a computer program whether an explanatory variable is an exposure or confounder it makes a great deal of difference to the person trying to interpret the results. Perhaps the single most important reason for misinterpreting the results of regression analyses is that regression models can be used without the user thinking carefully about the status of different explanatory variables. This will be discussed at greater length in Chapter 27.

Table 22.2. A regression model to compare rates by age bands

	Exposure	
Age	0	1
0	λ_0^0	λ_1^0
1	$\lambda_0^0\phi^1$	$\lambda_1^0\phi^1$
2	$\lambda_0^0\phi^2$	$\lambda_1^0\phi^2$

Table 22.3. A regression model for exposure and age

	Exposure	
Age	0	1
0	λ_C	$\lambda_C\theta$
1	$\lambda_C\phi^1$	$\lambda_C\theta\phi^1$
2	$\lambda_C\phi^2$	$\lambda_C\theta\phi^2$

Exercise 22.1. Table 22.4 shows a set of values for the rate parameters (per 1000 person-years) which satisfy exactly the model shown in Table 22.3. What are the corresponding values of $\lambda_C, \theta, \phi^1, \phi^2$?

Exercise 22.2. When the model in Table 22.3 is fitted to data it imposes the constraint that the rate ratio for exposure is the same in all age bands, and equally, that each of the two rate ratios for age is constant over both levels of exposure. Is the constraint on the rate ratios for age a new constraint, or does it automatically follow whenever the rate ratio for exposure is the same in all age bands?

Table 22.4. Parameter values (per 1000) which obey the constraints

	Exposure	
Age	0	1
0	5.0	15.0
1	12.0	36.0
2	30.0	90.0

Table 22.5. A regression model using names for parameters

	Exposure	
Age	0	1
0	Corner	Corner × Exposure(1)
1	Corner × Age(1)	Corner × Age(1) × Exposure(1)
2	Corner × Age(2)	Corner × Age(2) × Exposure(1)

22.3 Naming conventions

Using Greek letters for parameters is convenient when developing the theory but less so when applying the methods in practice. With many explanatory variables there will be many parameters and it is easy to forget which letter refers to which parameter. For this reason we shall now move to using names for parameters instead of Greek letters.

The first of the parameters in Table 22.3, λ_C, is called the Corner. The θ parameter, which is the effect of exposure controlled for age, is referred to as Exposure(1); when the exposure variable has three levels there are two effects and these are referred to as Exposure(1) and Exposure(2), and so on. When the exposure variable is given a more specific name such as Alcohol then the effects are referred to as Alcohol(1) and Alcohol(2). The ϕ parameters, which are the effects of age controlled for exposure, are referred to as Age(1) and Age(2). The model in Table 22.3 is written using names in Table 22.5.

Because writing out models in full is rather cumbersome, particularly when using names for parameters, we shall use a simple abbreviated form instead. The entries in Tables 22.3 and 22.5 refer to the right-hand sides of the regression equations; the left-hand sides are the original rate parameters which are omitted. Such a set of regression equations is abbreviated to

$$\text{Rate} = \text{Corner} \times \text{Exposure} \times \text{Age}.$$

It is important to remember that this abbreviation is not itself an equation (even though it looks like one!); it represents a set of equations and is shorthand for tables like Table 22.5. The regression model is sometimes abbreviated even further and referred to simply as a *multiplicative model* for exposure and age.

Table 22.6. Energy intake and IHD incidence rates per 1000 person-years

	Unexposed			Exposed			
	(≥ 2750 kcals)			(< 2750 kcals)			Rate
Age	Cases	P-yrs	Rate	Cases	P-yrs	Rate	ratio
40–49	4	607.9	6.58	2	311.9	6.41	0.97
50–59	5	1272.1	3.93	12	878.1	13.67	3.48
60–69	8	888.9	9.00	14	667.5	20.97	2.33

Table 22.7. Estimated values of the parameters for the IHD data

Parameter	Estimate
Corner	0.00444
Exposure(1)	×2.39
Age(1)	×1.14
Age(2)	×2.00

22.4 Estimating the parameters in a regression model

Table 22.6 shows the data from the study of ischaemic heart disease and energy intake. There are two explanatory variables, age with three levels and exposure with two. The two levels of exposure refer to energy intakes above and below 2750 kcals per day.

Although the rate ratio for exposure is rather lower in the first age band than in the other two age bands, it is based on only 6 cases, and a summary based on the assumption of a common rate ratio seems reasonable. In the new terminology this means fitting the regression model

$$\text{Rate} = \text{Corner} \times \text{Exposure} \times \text{Age}.$$

The most likely values of the parameters in this model, obtained from a computer program, are shown in Table 22.7. Note that the most likely value of the Exposure(1) parameter is the same, to two decimal places, as the Mantel–Haenszel estimate of the common rate ratio, given in Chapter 15.

Exercise 22.3. Use the most likely values of the parameters in the regression model, shown in Table 22.7, to predict the rates for the six cells in Table 22.6.

Computer programs differ in the precise details of how the output is

Table 22.8. Estimated parameters and SDs on a log scale

Parameter	Estimate (M)	SD (S)
Corner	−5.4180	0.4420
Exposure(1)	0.8697	0.3080
Age(1)	0.1290	0.4753
Age(2)	0.6920	0.4614

labelled. In particular you may see the word *variable* where we have used *parameter*, and the word *coefficient* where we have used *estimate*. We have used the term *corner* for the parameter which measures the level of response in the first age band of the unexposed group but several other terms are in widespread use, for example *constant*, *intercept*, *grand mean*, and (most cryptically of all) the number 1. We have numbered strata and exposure categories starting from zero, but some programs start numbering from one.

22.5 Gaussian approximations on the log scale

Gaussian approximations to the likelihood are used to obtain approximate confidence intervals for the parameter values. For the simple multiplicative models discussed so far the approximation is always made on the log scale, and in many programs the output is also in terms of logarithms. Table 22.8 shows the output on a log scale for the ischaemic heart data; the second column shows the most likely values (M) of the logarithms of the parameters and exponentials of these give the values on the original scale. For example,

$$\exp(0.8697) = 2.39,$$

which is the rate ratio for exposure. The third column shows the standard deviations (S) of the estimates, obtained from Gaussian approximations to the profile log likelihoods for each parameter. The standard deviation of the effect of exposure, on the log scale, is 0.3080, so the error factor for a 90% confidence interval for this parameter is $\exp(1.645 \times 0.3080) = 1.66$, and the limits are from $2.39/1.66 = 1.44$ to $2.39 \times 1.66 = 3.96$.

Exercise 22.4. Use Table 22.8 to calculate the 90% confidence limits for the first effect of age.

When the regression model is fitted on a log scale it is written in the form

$$\log(\text{Rate}) = \text{Corner} + \text{Exposure} + \text{Age}.$$

Table 22.9. A more complete description of the age effects

Parameter	Estimate	SD
Age(1)	0.1290	0.4753
Age(2)	0.6920	0.4614
Age(2) − Age(1)	0.5630	0.3229

Table 22.10. An abbreviated table for the age effects

Parameter	Estimate	SD	
Age(1)	0.1290	0.4753	0.3229
Age(2)	0.6920	0.4614	

Strictly speaking, the parameters on the right-hand side of this expression should be written as log(Corner) etc., but in practice the log on the left-hand side is enough to signal the fact that the parameter estimates will be on a log scale.

For variables with more than two categories, comparisons other than those with the first category are sometimes of interest. Taking the variable age in the ischaemic heart disease data as an example, the effect of changing from level 1 to level 2 of age is the difference between the two age effects, namely $0.6920 - 0.1290 = 0.5630$. Because the two age effects are based on some common data the standard deviation of their difference cannot be obtained from the simple formula

$$\sqrt{0.4753^2 + 0.4614^2} = 0.6624,$$

which was used in Chapter 13. To obtain the correct standard deviation we usually need to resort to a trick, such as recoding age so that the corner parameter refers to the *second* age band rather than the first. Table 22.9 shows how a fuller analysis of age effects could be reported; an option to obtain output in this form would be a useful feature not currently available in most computer programs.

An abbreviated way of conveying the same information is shown in Table 22.10. This provides the standard deviations for all three comparisons but leaves the user to do the subtraction to find the effect of changing from level 1 to level 2. The method extends naturally for factors with more than three levels; for example, a four-level factor would need a triangular array of 6 standard deviations for the six possible pairwise comparisons.

22.6 Additive models

When comparing two groups, in the first section of this chapter, the two parameters λ_0 and λ_1 were replaced by λ_0 and $\theta = \lambda_1/\lambda_0$. This change of parameters made it possible to estimate the rate ratio θ along with its standard deviation. The parameters could equally well have been changed to λ_0 and $\theta = \lambda_1 - \lambda_0$, thus making it possible to estimate the rate difference instead of the rate ratio.

The choice between the rate ratio and the rate difference is usually an empirical one, depending on which of the two is more closely constant over strata. In the early years of epidemiology, when age was often the only explanatory variable apart from exposure, methods of analysis were all based (implicitly) on multiplicative models. This is because most rates vary so much with age that the rate ratio is almost always more closely constant over age bands than the rate difference. More recently, particularly when investigating the joint effects of several exposures, epidemiologists have shown a greater interest in rate differences.

To impose the constraint that the rate difference is constant over age strata, the regression model

$$\text{Rate} = \text{Corner} + \text{Exposure} + \text{Age}$$

is fitted. This is called an *additive model* for exposure and age. Note that it is the rate and not the log rate which now appears on the left-hand side. The same likelihood techniques are used as with the additive model as with the multiplicative model, but because the estimated values of the parameters in the additive model must be restricted so that they predict positive rates, it is much harder to write foolproof programs to fit these models. We shall return to additive models in Chapter 28.

22.7 Using computer programs

There is a certain amount of specialized terminology connected with computer programs which we shall introduce briefly in this section.

VARIABLES AND RECORDS

The information collected in a study is best viewed as a rectangular table in which the columns refer to the different kinds of information collected for each subject, and the rows to the different subjects. In computer language the columns are called *variables* and the rows are called *records.* Variables such as age and observation time are called *quantitative* because they measure some quantity. Variables such as exposure group are called *categorical* because they record the category into which a subject falls. The different categories are called the *levels* of the variable. Another name for a categorical variable is *factor.* Categorical variables with only two categories (or levels) are also known as *binary* variables.

DERIVED VARIABLES

The raw data which is collected in a study may not be in exactly the right form for analysis. For example, in a follow-up study the observation time will usually be recorded as date of entry to the study and date of exit. The computer can be instructed to derive the observation time from these two dates by subtraction. Another example is where the grouped values of a quantitative variable are required in an analysis; it is then convenient to derive a new categorical variable which records the group into which each subject falls.

VARIABLE NAMES

In order to give instructions to a computer program each of the variables needs a name. These can usually be at least eight characters long and it is a good idea to make full use of this and to choose names which will mean something to you (and someone else) in a year's time.

SUMMARY TABLES

It is always important when using computer programs to keep in close touch with the data you are analyzing. The simplest way of doing this is to start by looking at tables which show the estimated rate or odds parameters for different combinations of the values of the explanatory variables. When there are two explanatory variables the table is called two-way, and so on. Three-way tables are presented as a series of two-way tables. When an explanatory variable is quantitative it will usually be necessary to group the values of the variable before using it to define a table. Only after inspecting various summary tables to get some feel for the main results should you use regression models to explore the data more fully.

FREQUENCY OR INDIVIDUAL RECORDS

Computer programs are generally able to accept either *individual records* or *frequency records* based on groups of subjects. For example, in the ischaemic heart disease study, we could use the data records for each subject, or frequency records showing the number of subjects in each combination of age band and exposure group. Entering a frequency record for 25 subjects has exactly the same effect as entering 25 identical individual records.

When an explanatory variable is quantitative its values must be grouped before frequency records can be formed, while the actual values can be used with individual records. Frequency records can be stored more compactly than individual records, and log likelihood calculations are correspondingly faster, but using frequency records requires two computer programs — one to compute the frequency records and one to carry out the regression analysis — and communication between these programs may be inconvenient. For case-control studies the number of subjects is usually relatively small and the data are usually entered as individual records. For cohort studies there may be tens of thousands of individual records, possibly further subdivided between time-bands, so the data are usually entered as frequency records.

MISSING VALUES

Most studies contain records which have some missing values, and it is essential to have some way of indicating this to the computer program. The most convenient code for a missing value is the character *, but when a program insists on a numeric code it is best to choose some large number like 9999. When there are many variables in a study the analyses are usually on some subset of the variables, and the program will automatically include those records with complete data on the subset being used.

Solutions to the exercises

22.1 $\lambda_C = 5.0$ per 1000, $\theta = 3.0$, $\phi^1 = 2.4$, $\phi^2 = 6.0$.

22.2 It is not a new constraint. Table 22.1 shows that when the rate ratio for exposure is constant over age bands then the rate ratios for age will automatically be constant over exposure groups.

22.3 The predicted rates for the six combinations of age and exposure are

Age	Unexposed	Exposed
40 – 49	4.44	10.61
50 – 59	5.06	12.10
60 – 69	8.88	21.22

22.4 The effect of age level 1 is exp(0.1290) = 1.14. The 90% confidence interval for this effect is

$$1.14 \overset{\times}{\div} \exp(1.645 \times 0.4753)$$

which is from 0.52 to 2.49.

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

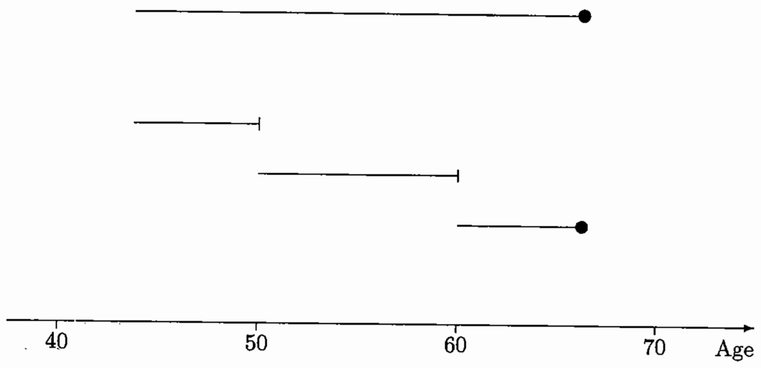


Fig. 23.1. Splitting the follow-up record.

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	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

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

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

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

24
Testing hypotheses

The scientific imagination knows no bounds in the creation of theories and interesting models, but when should such elaboration end? The principle which is invoked to deal with this problem is *Occam's razor*. This principle holds that we should always adopt the simplest explanation consistent with the known facts. Only when the explanation becomes inconsistent are we justified in greater elaboration. Occam's razor has much in common with statistical tests of null hypotheses. Statisticians erect null hypotheses and seek positive evidence against them before accepting alternative explanations. This philosophical position should not be taken to imply that the absence of evidence against a null hypothesis establishes the null hypothesis as being true.

24.1 Tests involving a single parameter

An explanatory variable with two levels requires only one parameter to make a comparison between them. When the comparison is made using a rate ratio (or an odds ratio) the null value is 1.0, or zero on the log scale. The simplest way of testing for a zero null value is to use the Wald test, based on the profile log likelihood for the parameter being tested. This involves referring

$$\left(\frac{M-0}{S}\right)^2$$

to tables of the chi-squared distribution on one degree of freedom, where M is the most likely value of the log of the parameter and S is its standard deviation. These quantities are the ones listed in the computer output under estimate and standard deviation.

Exercise 24.1. Table 24.1 repeats the results of the regression analysis of the ischaemic heart disease data. Carry out the Wald test of the hypothesis of no effect of exposure on IHD incidence.

A log likelihood ratio test based on the profile likelihood for the exposure parameter can also be used to test the hypothesis in Exercise 24.1. The profile log likelihood ratio for a zero exposure effect is the difference between two log likelihoods: (a) the log likelihood when the exposure parameter is

Table 24.1. Program output for the ischaemic heart disease data

Parameter	Estimate	SD
Corner	−5.4180	0.4420
Exposure(1)	0.8697	0.3080
Age(1)	0.1290	0.4753
Age(2)	0.6920	0.4614

zero and the age parameters take their most likely values given that there is no exposure effect, and (b) the log likelihood evaluated when all parameters take their most likely values. The former is obtained by fitting a model which includes age but not exposure, and the latter is obtained by fitting a model which includes both age and exposure. The difference between these two log likelihoods gives the profile log likelihood ratio, and the test is carried out by referring minus twice this value to the chi-squared distribution with one degree of freedom. Some programs report the *deviance*, a quantity closely related to the log likelihood which we shall discuss in a later section of this chapter.

Exercise 24.2. The log likelihoods for the models

$$\begin{aligned}\log(\text{Rate}) &= \text{Corner} + \text{Age} + \text{Exposure}\\ \log(\text{Rate}) &= \text{Corner} + \text{Age}\end{aligned}$$

for the ischaemic heart disease data, are −247.027 and −251.176. How can you tell which likelihood was obtained for which model? Carry out the likelihood ratio test for a zero exposure effect and compare it with the Wald test calculated in the previous exercise.

The score test for a zero exposure effect is found from a quadratic approximation which has the same gradient and curvature as the profile log likelihood at the null value. Since the log likelihood ratio test is easy to obtain using a computer program the score test is rarely carried out, although some programs do offer this option.

24.2 Tests involving several parameters

When a variable has three levels two parameters are required to make comparisons between the levels. A test that just one of these parameters takes its null value is rarely of interest. The hypothesis that both take their null values is usually more relevant, because this corresponds to the variable having no effect on the response. We shall now consider the extension of the likelihood ratio test to cover this situation. A convenient example is provided by the problem of testing the effect of age in the analysis shown in Table 24.1, although this is a hypothesis of no scientific interest!

The same general principle as for one parameter is used: the log likelihood for the model

$$\text{Corner} + \text{Age} + \text{Exposure}$$

which includes the two age parameters, is subtracted from the log likelihood for the model

$$\text{Corner} + \text{Exposure},$$

in which the two age parameters are zero. This gives the log likelihood ratio for testing the hypothesis that both age parameters take their null values. Minus twice the log likelihood ratio is referred to the chi-squared distribution with *two* degrees of freedom, because two parameters have been set to their null values. In this case minus twice the log likelihood ratio is equal to 4.016, and the p-value is 0.134, showing that there is no significant effect of age on ischaemic heart disease in this study.

Exercise 24.3. Does the fact that there is no significant effect of age on incidence in this study mean that there is no need to control for age when comparing exposure groups?

There is some temptation to scan the output for the model which includes both age and exposure and to try to interpret the separate tests of the two parameters for age, rather than making a joint test. Using the Wald test with the results in Table 24.1 shows that the data support both null values for age when tested separately, but it would be unwise to deduce from this that there is no effect of age. This is because both age effects are rather imprecisely estimated, due to the fact that only 6 heart attacks were observed in the first age band. When the corner is located where there is very little data it is common to see effects for both levels 1 and 2 which are small compared to their standard deviations, yet a highly significant effect from level 1 to level 2. The only safe way of testing the effect of age is to make a test of the joint hypothesis that both age effects take their null value. The Wald test can be generalized to do this (as can the score test), but the easiest test to use is the log likelihood ratio test.

24.3 Testing for interaction

The regression model used in the test for an exposure effect imposes the constraint that the effect of exposure is constant over age bands. Similarly for the test for age effects. An important question to ask is whether it is reasonable to impose these constraints, or whether the data better support different exposure effects in each age band, and different age effects in each exposure group. When the effects of exposure vary with age there is said to be *interaction* between exposure and age. Interaction between exposure and age automatically implies interaction between age and exposure and vice versa.

Table 24.2. Definition of interactions in terms of exposure

		Exposure	
		0	1
	0	5.0	15.0
Age	1	12.0	42.0
	2	30.0	135.0
	0	5.0	5.0 × 3.0
Age	1	12.0	12.0 × 3.5
	2	30.0	30.0 × 4.5
	0	5.0	5.0 × 3.0
Age	1	12.0	12.0 × 3.0 × 1.167
	2	30.0	30.0 × 3.0 × 1.5

To test for interaction it is necessary to choose new parameters in a way that allows for separate effects of exposure in the different age bands. This is done by choosing one parameter to measure the effect of exposure in the first age band and two to measure the extent to which the effects of exposure in the other two age bands differ from the effect in the first age band. The way this is done is best illustrated using numerical values for the parameters.

A set of illustrative values for the 6 rate parameters are shown at the top of Table 24.2. The rate ratios for exposure by levels of age are 3.0, 3.5, and 4.5, shown in the middle part of the table, so these rate parameters do not obey a multiplicative model. The extent of the departure from the multiplicative model can be measured by expressing 3.5 and 4.5 as ratios relative to 3.0, as shown in the third part of the table. These ratios, which take the values 1.167 and 1.5 in this case, are called *interaction* parameters.

Table 24.3 shows the same thing in terms of the rate ratios for age by levels of exposure. These rate ratios are 2.4 and 6.0 when exposure is at level 0 but 2.8 and 9.0 when exposure is at level 1. The extent to which these differ, measured as ratios relative to the rate ratios at level 0 of exposure, are again equal to 1.167 and 1.5. Thus the interaction parameters are symmetric in exposure and age.

Table 24.3. Definition of interactions in terms of age

		Exposure	
		0	1
	0	5.0	15.0
Age	1	12.0	42.0
	2	30.0	135.0
	0	5.0	15.0
Age	1	5.0 × 2.4	15.0 × 2.8
	2	5.0 × 6.0	15.0 × 9.0
	0	5.0	15.0
Age	1	5.0 × 2.4	15.0 × 2.4 × 1.167
	2	5.0 × 6.0	15.0 × 6.0 × 1.5

Table 24.4. Definition of interactions in terms of exposure and age

	Exposure	
Age	0	1
0	5.0	5.0 × 3.0
1	5.0 × 2.4	5.0 × 3.0 × 2.4 × 1.167
2	5.0 × 6.0	5.0 × 3.0 × 6.0 × 1.5

Tables 24.2 and 24.3 are combined in Table 24.4. Using the terminology of regression models, the 6 original rate parameters are re-expressed in terms of the corner, the rate ratio for exposure when age is at level 0, the rate ratio for age when exposure is at level 0, and the two interaction parameters. This way of re- expressing the original rate parameters has not resulted in any reduction in the number of parameters; its sole purpose is to assess the extent of the departures from the multiplicative model. We shall write the model with interaction in one or other of the forms

$$\begin{aligned} \text{Rate} &= \text{Corner} \times \text{Exposure} \times \text{Age} \times \text{Exposure}{\cdot}\text{Age} \\ \log(\text{Rate}) &= \text{Corner} + \text{Exposure} + \text{Age} + \text{Exposure}{\cdot}\text{Age}. \end{aligned}$$

To test for interaction it is necessary to fit the model with and without interaction parameters and to measure the log likelihood ratio for these two models. Minus twice this log likelihood ratio is then referred to tables of chi-squared on two degrees of freedom. The chi-squared has two degrees of freedom because the hypothesis being tested is that two interaction parameters take their null values. The instruction to include interaction parameters is done by including the term Age·Exposure in the model description. When this is done the output will include estimated values for the interaction parameters, but these are rarely of much use because they are chosen specifically to make the test for no interaction. If there is interaction then it will usually be best to report the effects of exposure separately for each age band. If there is no interaction then the effects of exposure and age should be obtained from the model without interaction parameters. Further details on how to report interactions are given in Chapter 26.

Table 24.5. Estimates of parameters in the model with interaction

Parameter	Estimate	SD
Corner	−5.0237	0.500
Exposure(1)	−0.0258	0.866
Age(1)	−0.5153	0.671
Age(2)	0.3132	0.612
Age(1)·Exposure(1)	1.2720	1.020
Age(2)·Exposure(1)	0.8719	0.973

Table 24.5 shows the output for the ischaemic heart disease data when fitting the model which includes the interaction between exposure and age. The interaction parameters are given names like Age(1)·Exposure(1) and Age(2)·Exposure(1). In general the number of interaction parameters between a variable on a levels and one on b levels is $(a-1)(b-1)$.

Exercise 24.4. Verify from Table 24.5 that the estimated corner parameter in the model with interaction is now the log of the observed rate for unexposed subjects in age band 0, and the estimated Exposure(1) parameter is now the observed rate ratio (exposed/unexposed) in age band 0. (The observed rates are in Table 22.6.)

24.4 Deviance

The log likelihood for a regression model, evaluated at the most likely values of the parameters, is a measure of *goodness-of-fit* of the model — the greater the log likelihood, the better the fit. Since the absolute value of the log likelihood is not itself of interest there is some advantage in always reporting a log likelihood ratio, compared to some other model. A convenient choice is the *saturated* which includes the maximum possible number of parameters. The output would then include the log likelihood ratio between the model being fitted and the saturated model. For use with tables of chi-squared it is slightly more convenient to report minus twice the log likelihood ratio, a quantity which is called the *deviance* for the model being fitted. Each deviance has degrees of freedom equal to the difference between the number of parameters in the model and the number in the saturated model.

The deviance is a measure of badness of fit; the larger the deviance the worse the fit. Two models are compared by comparing their deviances. The change in deviance is minus twice the log likelihood ratio for the two models because the log likelihood for the saturated model occurs in both deviances and cancels (see Fig. 24.1.) The degrees of freedom for this test are found by subtracting the degrees of freedom for the two deviances. For example, when fitting the models

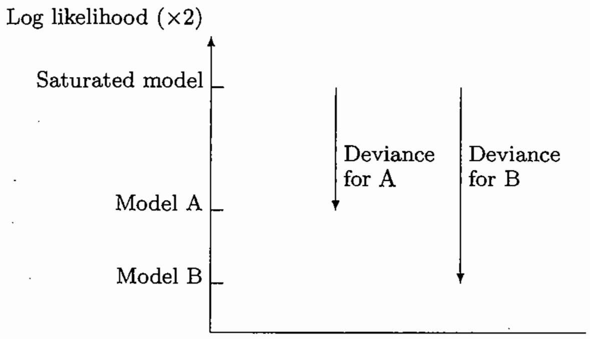


Fig. 24.1. Relationship between deviance and log likelihood

$$
\begin{aligned}
\log(\text{Rate}) &= \text{Corner} + \text{Age} + \text{Exposure}\\
\log(\text{Rate}) &= \text{Corner} + \text{Exposure},
\end{aligned}
$$

to the ischaemic heart disease data the corresponding values for the two deviances were 1.673 and 5.689. The difference between these is 4.016 which is the same as the result obtained earlier in the chapter for minus twice the log likelihood ratio.

Exercise 24.5. How do you know which deviance was obtained for which model? How many degrees of freedom do the two deviances have?

When the data are entered as frequency records the saturated model has the same number of parameters as there are frequency records. In the case of the ischaemic heart disease data there are six records so the saturated model has 6 parameters. All models with six parameters are saturated and have the same log likelihood. The model which includes the interaction parameters between age and exposure has six parameters, and is saturated, so it follows that the deviance for the model

$$\log(\text{Rate}) = \text{Corner} + \text{Age} + \text{Exposure}$$

provides a test of no interaction between age and exposure. It may be referred directly to a chi-squared distribution with two degrees of freedom.

When the data are entered as individual records the saturated model has the same number of parameters as the number of individual records and the deviance measures minus twice the difference between the log likelihood for the fitted model and this saturated model. This is not a test of anything useful. There is no short cut for making a test of no interaction using individual records: it is necessary to obtain the deviances for the models

Table 24.6. Cases (controls) for oral cancer study

	Alcohol			
Tobacco	0	1	2	3
0	10 (38)	7 (27)	4 (12)	5 (8)
1	11 (26)	16 (35)	18 (16)	21 (20)
2	13 (36)	50 (60)	60 (49)	125 (52)
3	9 (8)	16 (19)	27 (14)	91 (27)

Table 24.7. Case/control ratios for the oral cancer data

	Alcohol			
Tobacco	0	1	2	3
0	0.26	0.26	0.33	0.63
1	0.42	0.46	1.13	1.05
2	0.36	0.83	1.22	2.40
3	1.12	0.84	1.93	3.37

with and without the interaction parameters.

24.5 Models with two exposures

Because regression models treat all explanatory variables in the same way, models for studies with two exposures look very similar to models for studies with one exposure and one confounder. However, there are some differences in the way different hypotheses are interpreted.

Table 24.6 repeats the study of oral cancer introduced in Chapter 16, in which the numbers of cases and controls are tabulated by two exposures, alcohol consumption (on four levels) and tobacco consumption (also on four levels). For alcohol the levels are 0, 0.1–0.3, 0.4–1.5, and 1.6+ ounces per day (coded as 0, 1, 2, and 3). For tobacco the levels are 0, 1–19, 20–39, and 40+ cigarettes per day (also coded as 0, 1, 2, and 3). A summary table of case/control ratios by alcohol and tobacco is shown in Table 24.7. Because the frequencies in the table are small, there is a lot of random variation, but there is an overall tendency for the ratios to increase both from left to right along rows, and from top to bottom down columns. This indicates that *both* variables have an effect on cancer incidence; there is an effect of tobacco when alcohol intake is held constant, and vice versa.

An important question is whether the two exposures act independently of one another. In other words, are the effects of tobacco the same at all levels of alcohol, and are the effects of alcohol the same at all levels of tobacco? This question is answered by testing for no interaction between alcohol and tobacco, but it must be emphasized that the test depends on how the effect parameters are defined. When they are defined as ratios the interaction parameters are also ratios and measure departures from a model in which the two exposures combine multiplicatively. By choosing to measure effects as ratios we have therefore chosen to interpret independent action as meaning that the two exposures act multiplicatively. In Chapter 28 we show how the effects can be defined as differences, in which case the interaction parameters are also differences and measure departures from a model in which the two exposures combine additively. In this case we have chosen to interpret independent action as meaning the two exposures act additively.

If there is a significant interaction then it will be necessary to report the effects of alcohol separately as odds ratios for each level of tobacco consumption, and the effects of tobacco separately as odds ratios for each level of alcohol. On the other hand, if there is no significant interaction then the two exposures may be assumed to act independently and we can estimate the effects of alcohol controlled for tobacco and the effects of tobacco controlled for alcohol. Note that even when the two exposures act independently it is still necessary to control each for the other. This is because people's drinking and smoking habits are not independent so ignoring one when studying the other could lead to biased estimates.

The test for no interaction is carried out by comparing the fit of the multiplicative model

$$\log(\text{Odds}) = \text{Corner} + \text{Alcohol} + \text{Tobacco},$$

with that of the model which includes the interaction parameters,

$$\log(\text{Odds}) = \text{Corner} + \text{Alcohol} + \text{Tobacco} + \text{Alcohol} \cdot \text{Tobacco}.$$

Since the second of these models is saturated the test can be based directly on the deviance for the multiplicative model. Provided the data support the hypothesis of no interaction it is then possible to test for an effect of alcohol, controlled for tobacco, by comparing the models

$$\begin{aligned} \log(\text{Odds}) &= \text{Corner} + \text{Alcohol} + \text{Tobacco} \\ \log(\text{Odds}) &= \text{Corner} + \text{Tobacco}. \end{aligned}$$

Similarly the test for an effect of tobacco is made by comparing the models

$$\begin{aligned} \log(\text{Odds}) &= \text{Corner} + \text{Alcohol} + \text{Tobacco} \\ \log(\text{Odds}) &= \text{Corner} + \text{Alcohol}. \end{aligned}$$

In each of these tests the smaller of the two models being compared is obtained from the larger by setting some parameters to zero. The smaller

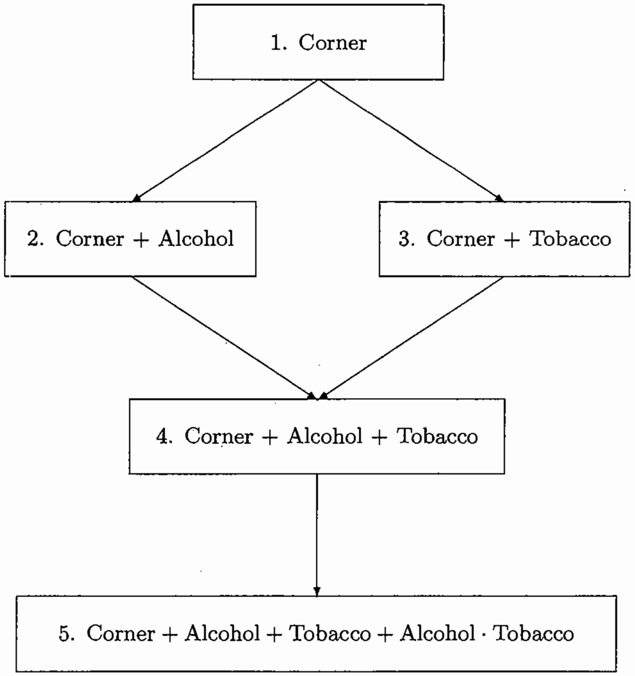

1. Corner
2. Corner + Alcohol
3. Corner + Tobacco
4. Corner + Alcohol + Tobacco
5. Corner + Alcohol + Tobacco + Alcohol · Tobacco

Fig. 24.2. Nesting of models.

model is then said to be *nested* in the larger model. Comparisons between models where neither is nested in the other are not allowed since they do not correspond to a hypothesis in which some parameter values are set equal to zero. Fig. 24.2 shows the five possible models which could be fitted to the alcohol and tobacco data. The arrows indicate nesting so any two models joined by an arrow correspond to a hypothesis which can be tested. For example, a comparison of models 4 and 5 is a test of no interaction, and a comparison of models 4 and 2 is a test of no effect of tobacco (controlling for alcohol). In model 1 both alcohol and tobacco parameters are set to zero so it is nested in all of the other models.

Exercise 24.6. For the models set out in Fig. 24.2, the deviances are (1) 132.561, (2) 37.951, (3) 61.880, and (4) 6.689. What are the degrees of freedom associated with each of these deviances? Carry out the four tests corresponding to the arrows in the figure. What is the interpretation of these tests?

24.6 Goodness-of-fit tests

A question which is often asked is whether a model provides an adequate fit to the data. Because the absolute value of the log likelihood has no meaning this question can only be answered by comparing the model with other more complicated models and asking whether the extra complication is justified. The saturated model represents the most complicated model which could be used and the deviance automatically provides a comparison of the model currently being fitted with the saturated model. For this reason the deviance for a model is often put forward as a test of goodness of fit (really badness-of-fit) of the model. There are several cautions which need to be borne in mind when interpreting the deviance in this way.

1. Comparisons with the saturated model are meaningless when the data are entered as individual records.

2. Comparisons with the saturated model which are on many degrees of freedom will lack power to discriminate; in this case it will be better to make comparisons with models which are less complicated than the saturated model.

3. The deviance is only approximately distributed as chi-squared and this approximation gets worse as the degrees of freedom increase.

24.7 Collinearity

In a study in which tobacco and alcohol consumption were very highly associated it would be very difficult to make an estimate of the effects of alcohol controlled for tobacco (or of the effects of tobacco controlled for alcohol). This is because controlling for tobacco involves fixing the level of tobacco consumption and then estimating the effects of alcohol from subjects whose tobacco consumption is at this level. If alcohol and tobacco are highly associated then nearly all subjects at a fixed tobacco level will have the same level of alcohol consumption and it will therefore be difficult to estimate the effects of alcohol. In extreme cases fixing the level of tobacco might fix the level of alcohol completely, in which case it would be impossible to estimate the effects of alcohol. In such a case the two variables are said to be *collinear*. This situation is not uncommon, particularly when working with derived variables.

Solutions to the exercises

24.1 In the Wald test $(0.8697/0.3080)^2 = 7.97$ is referred to the chi-squared distribution with one degree of freedom, giving a p-value of 0.005.

24.2 The larger likelihood, −247.027, corresponds to the first model because this has more parameters than the second. The log likelihood ratio for the two models is $-251.176 - (-247.027) = -4.149$. Minus twice this is 8.298 which is quite close to the Wald chi-squared value obtained in the previous exercise. Referring 8.30 to the chi-squared distribution with one degree of freedom gives $p = 0.004$.

24.3 No. When taking account of confounding variables it is best to play safe and to control for them regardless of whether their effects are significant or not. Very little is lost by doing this.

24.4 The Corner, Exposure(1), Age(1) and Age(2) parameters are

$$\begin{aligned}
\log(6.580/1000) &= -5.0237 \\
\log(6.412/6.580) &= -0.0258 \\
\log(3.931/6.580) &= -0.5153 \\
\log(9.00/6.58) &= 0.3132.
\end{aligned}$$

24.5 The smaller deviance corresponds to the larger model since this will be a better fit. The degrees of freedom are 2 and 4 respectively.

24.6 The number of parameters in models 1 to 5 are 1, 4, 4, 7, and 16, respectively. The number of parameters in the saturated model is 16, so the degrees of freedom for the deviances are $16 - 1 = 15$, $16 - 4 = 12$, $16 - 4 = 12$, $16 - 7 = 9$, and $16 - 16 = 0$ respectively. Note that model 5 has 16 parameters so it is saturated. The table below shows the comparisons of models in terms of the change in deviance.

Comparison	Change in deviance	Change in df
(1) vs (2)	$132.56 - 37.95 = 94.61$	$15 - 12 = 3$
(1) vs (3)	$132.56 - 61.88 = 70.68$	$15 - 12 = 3$
(2) vs (4)	$37.95 - 6.69 = 31.26$	$12 - 9 = 3$
(3) vs (4)	$61.88 - 6.69 = 55.19$	$12 - 9 = 3$
(4) vs (5)	$6.69 - 0 = 6.69$	$9 - 0 = 9$

The last of these comparisons shows that there is no significant interaction. This means that the next two comparisons (working up from the bottom) make sense. The change in deviance from model 3 to model 4 shows that there is a significant effect of alcohol after controlling for tobacco; similarly the change in deviance from model 2 to model 4 shows that there is a significant effect of tobacco after controlling for alcohol. All of the models can be compared with model 1, but these comparisons have little interest. For example, a comparison of model 1 with model 2 is a test of the alcohol effects (ignoring tobacco) while a comparison of model 1 with model 4 is a joint test of the alcohol effects (controlling for tobacco) *and* the tobacco effects (controlling for alcohol).

25
Models for dose-response

When the subjects in a study receive different levels of exposure, measured on a quantitative or ordered scale, it is likely that any effect of exposure will increase (or decrease) systematically with the level of exposure. This is known as a dose-response relationship, or trend. The existence of such a relationship provides more convincing evidence of a causal effect of exposure than a simple comparison of exposed with unexposed subjects. Some simple procedures for testing for trend were introduced in Chapter 20. These tests are based on a log-linear dose-response relationship, that is, a linear relationship between the log rate parameter (or log odds parameter) and the level of exposure. We now return to this topic and show how such dose-response relationships are easily described as regression models.

25.1 Estimating the dose-response relationship

To illustrate the use of regression models when exposure is measured on a quantitative scale we shall use the case-control study of alcohol and tobacco in oral cancer in which there are two exposure variables, both with four levels. The model

$$\log(\text{Odds}) = \text{Corner} + \text{Alcohol} + \text{Tobacco},$$

in which alcohol and tobacco are categorical variables each with four levels, makes no assumption about dose-response; there are three alcohol parameters and three tobacco parameters. The estimated values of these parameters are shown in Table 25.1. If we were able to assume simple dose-response relationships for these two exposures, we could concentrate the available information into fewer parameters and, as a result, gain power.

To study the dose-response for tobacco consumption it helps to change from the parameters Tobacco(1), Tobacco(2), and Tobacco(3), which are chosen to compare each level of exposure with level 0, to

$$\text{Tobacco(1)}\ , \quad \text{Tobacco(2)}-\text{Tobacco(1)}\ , \quad \text{Tobacco(3)}-\text{Tobacco(2)}\ ,$$

which are chosen to compare each level with the one before.

Exercise 25.1. Use the results of Table 25.1 to write down the estimated values of these new parameters. Repeat the exercise for alcohol.

Table 25.1. Alcohol and tobacco treated as categorical variables

Parameter	Estimate	SD
Corner	−1.6090	0.2654
Alcohol(1)	0.2897	0.2327
Alcohol(2)	0.8437	0.2383
Alcohol(3)	1.3780	0.2256
Tobacco(1)	0.5887	0.2844
Tobacco(2)	1.0260	0.2544
Tobacco(3)	1.4090	0.2823

Table 25.2. The linear effect of tobacco consumption

Alcohol	Tobacco	log(Odds) = Corner + ···
0	0	–
0	1	1×[Tobacco]
0	2	2×[Tobacco]
0	3	3×[Tobacco]
1	0	Alcohol(1)
1	1	Alcohol(1) + 1×[Tobacco]
1	2	Alcohol(1) + 2×[Tobacco]
1	3	Alcohol(1) + 3×[Tobacco]
2	0	Alcohol(2)
2	1	Alcohol(2) + 1×[Tobacco]
2	2	Alcohol(2) + 2×[Tobacco]
2	3	Alcohol(2) + 3×[Tobacco]
3	0	Alcohol(3)
3	1	Alcohol(3) + 1×[Tobacco]
3	2	Alcohol(3) + 2×[Tobacco]
3	3	Alcohol(3) + 3×[Tobacco]

The simplest possible dose-response model would assume that each step in tobacco consumption, from one level to the next, produces the same change in the log odds. This model requires only one parameter for tobacco, namely the common change in log odds per change in level. This parameter is called the *linear effect* of tobacco and we shall write it as [Tobacco], where the brackets are used to distinguish the linear effect parameter from the separate effect parameters for each level. The model is written in full in Table 25.2.

The data from this study are in the form of frequency records containing the number of cases, the total number of cases and controls, alcohol consumption coded as 0, 1, 2, 3, and tobacco consumption coded as 0, 1, 2, 3. We shall write the model of Table 25.2 in the abbreviated form:

Table 25.3. Linear effect of tobacco per level

Parameter	Estimate	SD
Corner	−1.5250	0.219
Alcohol(1)	0.3020	0.232
Alcohol(2)	0.8579	0.237
Alcohol(3)	1.3880	0.225
[Tobacco]	0.4541	0.083

$$\log(\text{Odds}) = \text{Corner} + \text{Alcohol} + [\text{Tobacco}].$$

The regression program output for this model is illustrated in Table 25.3.

Exercise 25.2. How would you report the meaning of the number 0.4541 in Table 25.3?

A more accurate scale for tobacco consumption would be to use the midpoints of the ranges of tobacco use at each level, namely 0, 10, 30, and (say) 50 cigarettes per day. If the tobacco variable were coded in this way then the parameter [Tobacco] would refer to the linear effect per extra cigarette rather than per change of level. If the data were entered as individual records then the individual values for consumption could be used. In view of the uncertainties in measuring tobacco use there is something to be said for sticking to the scale 0, 1, 2, 3.

The reparametrization of the alcohol effects carried out in Exercise 25.1 also suggests a constant effect with increasing level of alcohol consumption. This allows the model to be further simplified to

$$\log(\text{Odds}) = \text{Corner} + [\text{Alcohol}] + [\text{Tobacco}],$$

where the parameter [Alcohol] is the common effect of an increase of one level in alcohol consumption. The regression output for this model is shown in Table 25.4.

Exercise 25.3. Use the output in Table 25.4 to work out what the model predicts for the combined effect of level 3 for tobacco and level 3 for alcohol compared to level 0 for both. Use the output in Table 25.1 to work out the same prediction when tobacco and alcohol are both treated as categorical.

For comparison we also show, in Table 25.5, the regression output for the model where alcohol consumption is measured in approximate mean ounces of alcohol per day for each category (0.0, 0.2, 1.0 and 2.0), and

Table 25.4. Linear effects of alcohol and tobacco per level

Parameter	Estimate	SD
Corner	−1.6290	0.1860
[Alcohol]	0.4901	0.0676
[Tobacco]	0.4517	0.0833

Table 25.5. Alcohol in ounces/day and tobacco in cigarettes/day

Parameter	Estimate	SD
Corner	−1.2657	0.1539
[Alcohol]	0.6484	0.0881
[Tobacco]	0.0253	0.0046

tobacco consumption is measured in approximate cigarettes per day for each category (0, 10, 30, or 50). The [Alcohol] and [Tobacco] parameters now look quite different from those in Table 25.4, but this is because they are measured per ounce of alcohol and per cigarette respectively.

TESTING FOR TREND

Comparison of log likelihoods for the models

$$\log(\text{Odds}) = \text{Corner} + \text{Alcohol} + [\text{Tobacco}]$$

and

$$\log(\text{Odds}) = \text{Corner} + \text{Alcohol}$$

yields a one degree of freedom test for the effect of tobacco controlled for the effect of alcohol. The Mantel extension test described in Chapter 20 is the corresponding score test, which tests the hypothesis that the [Tobacco] parameter takes the value zero.

TESTING FOR DEPARTURE FROM LINEARITY

To test for departures from linearity in the dose-response for tobacco, the models

$$\begin{aligned}\log(\text{Odds}) &= \text{Corner} + \text{Alcohol} + \text{Tobacco}\\ \log(\text{Odds}) &= \text{Corner} + \text{Alcohol} + [\text{Tobacco}],\end{aligned}$$

can be compared. In the first model Tobacco refers to the three effects of a categorical variable with 4 levels, while in the second [Tobacco] refers to the effect of a change of one level in tobacco consumption. The second model is a special case of the first, so they can be compared using a log likelihood ratio test.

Table 25.6. A quadratic dose-response relationship for tobacco

z	$(z)^2$	log(Odds) = Corner + ···
0	0	–
1	1	1×[Tobacco] + 1×[Tobsq]
2	4	2×[Tobacco] + 4×[Tobsq]
3	9	3×[Tobacco] + 9×[Tobsq]

Table 25.7. Predictions from a quadratic relationship

Effect	Predicted from model
Tobacco(1)	[Tobacco] + 1×[Tobsq]
Tobacco(2) − Tobacco(1)	[Tobacco] + 3×[Tobsq]
Tobacco(3) − Tobacco(2)	[Tobacco] + 5×[Tobsq]

Exercise 25.4. (a) How many parameters are there in the two models? (b) Reparametrize the models so that the second model is a special case of the first, with two parameters set to zero. (c) How would you interpret a significant difference between the fit of these two models?

25.2 Quadratic dose-response relationships

The simplest departure from a log-linear dose relationship is a log-quadratic relationship. To fit this model it is necessary to create a new dose variable which takes the values 0, 1, 4, 9, that is the squares of the values used to code tobacco consumption. We shall call this new variable 'tobsq'. The model is then fitted by including both tobacco and tobsq and declaring them as quantitative variables. The regression equations for this model are given in Table 25.6 and these show that when [Tobsq] is zero the dose-response is log-linear. Table 25.7 shows the tobacco effects for each level relative to the previous one, predicted from the quadratic model, and these show that the parameter [Tobsq] measures the degree to which the dose-response relationship departs from linearity.

The log-quadratic model also provides another way of testing for departures from a log-linear dose-response relationship, by comparing the models

$$\begin{aligned}\log(\text{Odds}) &= \text{Corner} + \text{Alcohol} + [\text{Tobacco}]\\ \log(\text{Odds}) &= \text{Corner} + \text{Alcohol} + [\text{Tobacco}] + [\text{Tobsq}].\end{aligned}$$

The comparison of these two models provides a test (on one degree of freedom) which will be sensitive to a departure from linearity in which the effect of tobacco increases with level ([Tobsq]>0), or decreases with level ([Tobsq]<0).

25.3 How many categories?

When collecting data, exposure is often measured as accurately as possible for individuals and only later are the observed values grouped into a relatively small number of categories. For example, the number of previous births would be recorded exactly, but might then be grouped as

$$0,\quad 1-3,\quad 4-6,\quad 7-9,\quad 10+\ .$$

When the variable is to be treated as categorical it is best to keep the number of categories small; three may be enough, and five is usually a maximum number. For exploratory analyses the use of just two categories has the advantage that there is only one effect to interpret, and it can often be easier to see what is going on.

The number of subjects in each category should be roughly the same, and to achieve this tertiles, quartiles or quintiles of the distribution of exposure are often used. Tertiles define three equal-sized groups, quartiles define four equal-sized groups, and quintiles define five such groups. This is quite a sensible way of choosing the grouping intervals provided the actual intervals are reported. A serious disadvantage is that such grouping intervals will vary from study to study, thus making it harder to compare findings.

When the variable is to be treated as quantitative there is no penalty in taking a larger number of categories. In the extreme case the original values are used. However, it is best to avoid the situation where one or two of the subjects have much higher values than all the rest. This can occur with an exposure like the number of previous sexual partners, which might lie between 0 and 10 for most subjects but reach numbers in excess of 100 for a few. In such a case the few subjects with high values can dominate the fit of a model, and it will be best to group the values so that all the high ones fall into a group such as 15 or more.

★ 25.4 Indicator variables

In order to fit a model to data the computer program must use the abbreviated description of the model to form the regression equations. These express the log rate (or log odds) parameter for each record as a linear combination of new parameters. For example, when the variable alcohol is entered in a model as categorical with levels coded 0, 1, 2, and 3, the regression equations include the parameter Alcohol(1) for records in which alcohol is at level 1, the parameter Alcohol(2) for records in which alcohol is at level 2, and the parameter Alcohol(3) for records in which alcohol is at level 3. The way the program does this is to create an *indicator* variable for each parameter. These variables are coded 1 for records which include the parameter and 0 otherwise. The indicator variables A_1, A_2, A_3 for the three alcohol parameters are shown in Table 25.8 alongside the levels of alcohol. Note that A_1, which indicates when Alcohol(1) should be included, takes the value 1 when alcohol is at level 1, and so on.

Table 25.8. Indicator variables for the three alcohol parameters

A_1	A_2	A_3	Level	log(Odds) = Corner + $\cdots$
0	0	0	0	–
1	0	0	1	Alcohol(1)
0	1	0	2	Alcohol(2)
0	0	1	3	Alcohol(3)

Exercise 25.5. Repeat Table 25.8 to show indicator variables for the case where both alcohol and tobacco have four levels.

A variable which is treated as quantitative acts as its own indicator since the way the variable is coded indicates what multiple of the linear effect parameter is to be included in the regression equations. For example, when tobacco is included as a quantitative variable, coded 0, 1, 2, and 3, the equations include the parameter [Tobacco] when tobacco is at level 1, twice the parameter [Tobacco] when tobacco is at level 2, and three times the parameter [Tobacco] when tobacco is at level 3. The coding of the tobacco variable thus indicates which multiple of the parameter is to be included in the model.

INTERACTION PARAMETERS

When interaction terms are included in the model, indicator variables are again used to form the regression equations. For simplicity we shall consider the situation where tobacco has only two levels, 0 for non-smokers and 1 for smokers. The model in which both alcohol and tobacco are categorical, and which contains interaction terms, is shown in full in Table 25.9. Indicator variables A_1, A_2, A_3 have been used for alcohol, and the indicator variable T has been used for tobacco. Note that when tobacco has only two levels, coded 0 and 1, it serves as its own indicator variable.

The indicator variable for Alcohol(1)·Tobacco(1) takes the value 1 when both alcohol and tobacco are at level 1, and 0 otherwise. The indicator variable for Alcohol(2)·Tobacco(1) takes the value 1 when alcohol is at level 2 and exposure is at level 1, and 0 otherwise, and so on. The most convenient way of generating these interaction indicator variables is by multiplying together pairs of the original indicator variables for alcohol and tobacco. This is shown in Table 25.10: the indicator for Alcohol(1)·Tobacco(1) is found from the product of A_1 and T; the indicator for Alcohol(2)·Tobacco(1) is

Table 25.9. The model with interaction between alcohol and tobacco

Alc.	Tob.	log(Odds) = Corner + ···
0	0	–
0	1	Tobacco(1)
1	0	Alcohol(1)
1	1	Alcohol(1) + Tobacco(1) + Alcohol(1)·Tobacco(1)
2	0	Alcohol(2)
2	1	Alcohol(2) + Tobacco(1) + Alcohol(2)·Tobacco(1)
3	0	Alcohol(3)
3	1	Alcohol(3) + Tobacco(1) + Alcohol(3)·Tobacco(1)

Table 25.10. Indicator variables for interaction parameters

A_1	A_2	A_3	T	$A_1 \cdot T$	$A_2 \cdot T$	$A_3 \cdot T$
0	0	0	0	0	0	0
0	0	0	1	0	0	0
1	0	0	0	0	0	0
1	0	0	1	1	0	0
0	1	0	0	0	0	0
0	1	0	1	0	1	0
0	0	1	0	0	0	0
0	0	1	1	0	0	1

made up from product of A_2 and T, and so on. When the categorical variables are on a and b levels respectively there are $(a-1)(b-1)$ new indicators for the interaction parameters.

In the first regression programs it was left to the user to create indicator variables for all parameters other than those referring to quantitative variables. Although it is rarely necessary to do this today, indicator variables are still important when we wish to use a non-standard parametrization of a regression model.

★ 25.5 The zero level of exposure

The level of exposure which is coded zero is often qualitatively different from the other levels. For example, zero previous births represents a very different biological experience from any other point on this scale. In such cases it may be better to omit the zero level when estimating the dose-response relationship, by allowing the response of at zero dose to differ from the general relationship (see Fig. 25.1). A parameter for each of these comparisons can be included in a model by using the indicator variable for non-smokers to fit the model

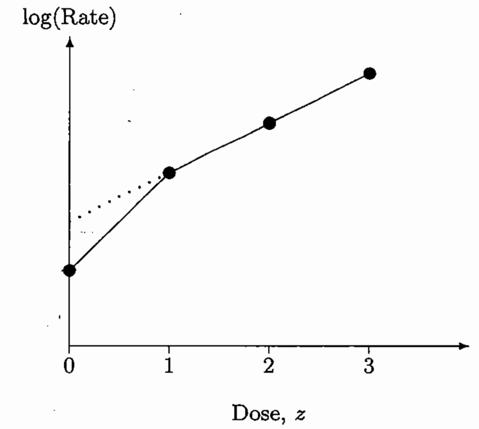


Fig. 25.1. Separating zero exposure from the dose-response.

Table 25.11. Separating zero exposure from the dose-response

Tobacco	Non-smoker	log(Odds) = Corner + ···
0	1	[Non-smoker]
1	0	1×[Tobacco]
2	0	2×[Tobacco]
3	0	3×[Tobacco]

$$\log(\text{Odds}) = \text{Corner} + [\text{Non-smoker}] + [\text{Tobacco}].$$

The regression equations for all four dose levels are shown in Table 25.11. The parameter [Non-smoker] measures the discrepancy between the log odds for non-smokers and that predicted by extrapolation of the dose-response line to zero dose.

25.6 Using indicators to reparametrize the model ★

Indicator variables provide a convenient way of changing from one set of parameters to another. We shall give one example, namely changing from parameters which compare each level with level 0, to parameters which compare each level with the one before. Using tobacco as an example, the first set of parameters are Tobacco(1), Tobacco(2), and Tobacco(3). We shall call the new parameters Tobdiff(1), Tobdiff(2), and Tobdiff(3). The

Table 25.12. Indicators to compare each level with the one before

Tobacco	D_1	D_2	D_2
0	0	0	0
1	1	0	0
2	1	1	0
3	1	1	1

relationship between the new parameters and the old is

$$\begin{aligned}
\text{Tobdiff}(1) &= \text{Tobacco}(1)\\
\text{Tobdiff}(2) &= \text{Tobacco}(2) - \text{Tobacco}(1)\\
\text{Tobdiff}(3) &= \text{Tobacco}(3) - \text{Tobacco}(2).
\end{aligned}$$

This relationship may be inverted to give the old in terms of the new as

$$\begin{aligned}
\text{Tobacco}(1) &= \text{Tobdiff}(1)\\
\text{Tobacco}(2) &= \text{Tobdiff}(1) + \text{Tobdiff}(2)\\
\text{Tobacco}(3) &= \text{Tobdiff}(1) + \text{Tobdiff}(2) + \text{Tobdiff}(3)
\end{aligned}$$

Let the indicator variables for Tobdiff(1), Tobdiff(2), Tobdiff(3), be denoted by D_1, D_2, D_3. The first of these should indicate Tobdiff(1) when tobacco is at level 1, 2, or 3; the second should indicate Tobdiff(2) when tobacco is at level 2 or 3; and the third should indicate Tobdiff(3) when tobacco is at level 3. Their values are shown in Table 25.12.

Solutions to the exercises

25.1 The estimates of the new parameters will be

Tobacco(1)	0.5887
Tobacco(2)−Tobacco(1)	0.4373
Tobacco(3)−Tobacco(2)	0.3830

and

Alcohol(1)	0.2897
Alcohol(2)−Alcohol(1)	0.5540
Alcohol(3)−Alcohol(2)	0.5343

25.2 The parameter represents the change in log odds for each increase in level of tobacco consumption.

25.3 The combined effect on the log odds is

$$+(3 \times 0.4901) + (3 \times 0.4517) = 2.8254.$$

This corresponds to a multiplicative effect of ×16.87 on the odds. When alcohol and tobacco are both treated as categorical the combined effect on the log odds is

$$+1.3780 + 1.4090 = 2.7870$$

which corresponds to a multiplicative effect of ×16.23 on the odds.

25.4 (a) The first model has 7 parameters, the second has 5. (b) Starting with Tobacco(1), Tobacco(2), and Tobacco(3), change to the parameters New(1), New(2), and New(3), where

$$\begin{aligned}
\text{New}(1) &= \text{Tobacco}(1)\\
\text{New}(2) &= \{\text{Tobacco}(2) - \text{Tobacco}(1)\} - \text{Tobacco}(1)\\
\text{New}(3) &= \{\text{Tobacco}(3) - \text{Tobacco}(2)\} - \text{Tobacco}(1).
\end{aligned}$$

Then New(1) measures the effect of changing level from 0 to 1; New(2) measures the difference between this and the effect of changing level from 1 to 2; New(3) measures the difference between this and changing level from 2 to 3. The model with all three parameters allows separate effects of changing level while the model with New(2) and New(3) equal to zero imposes the constraint that there is a common effect of changing level.
(c) When the first model is a significantly better fit than the second model it means that there is a significant departure from linearity in the dose-response.

25.5 Let $A_1, A_2, A_3, T_1, T_2, T_3$ be the indicator variables for alcohol and tobacco. The table below shows how these variables are coded and the regression model which is fitted when all the indicators are included.

A_1	A_2	A_3	T_1	T_2	T_3	log(Odds) = Corner + ⋯
0	0	0	0	0	0	–
0	0	0	1	0	0	Tobacco(1)
0	0	0	0	1	0	Tobacco(2)
0	0	0	0	0	1	Tobacco(3)
1	0	0	0	0	0	Alcohol(1)
1	0	0	1	0	0	Alcohol(1) + Tobacco(1)
1	0	0	0	1	0	Alcohol(1) + Tobacco(2)
1	0	0	0	0	1	Alcohol(1) + Tobacco(3)
0	1	0	0	0	0	Alcohol(2)
0	1	0	1	0	0	Alcohol(2) + Tobacco(1)
0	1	0	0	1	0	Alcohol(2) + Tobacco(2)
0	1	0	0	0	1	Alcohol(2) + Tobacco(3)
0	0	1	0	0	0	Alcohol(3)
0	0	1	1	0	0	Alcohol(3) + Tobacco(1)
0	0	1	0	1	0	Alcohol(3) + Tobacco(2)
0	0	1	0	0	1	Alcohol(3) + Tobacco(3)

26
More about interaction

In this chapter we draw together some of the ideas of the previous chapters, particularly those relating to interaction, and consider studies with several explanatory variables. The first stage in the analysis of such studies is to classify the explanatory variables into those whose effects are of interest (the exposures), and those whose effects are of no interest, but which must be included in the model (the confounders). In order to illustrate the problems which arise with several confounders we introduce a new example in Table 26.1* This shows the proportion of subjects with monoclonal gammapathy by age, sex, and work. Work can be agricultural or non-agricultural and is the exposure of interest. Age and sex are confounders.

26.1 Interaction between confounders

To control for the confounding effect of both age and sex using stratification it would be necessary to form $5 \times 2 = 10$ age– sex strata. The separate estimates of the effect of work for each stratum would then be pooled over strata using the Mantel–Haenszel method. The same thing can be done by fitting the model

$$\log(\text{Odds}) = \text{Corner} + \text{Age} + \text{Sex} + \text{Age} \cdot \text{Sex} + \text{Work},$$

which includes age–sex interaction parameters. The total number of parameters for the corner, age, sex, and the age–sex interaction is $1+4+1+4 = 10$, which is the same as the number of the age–sex strata. Fitting the model with interaction does the same job as age–sex stratification, which has one parameter for each of the 10 strata.†

It is also possible to control for age and sex by omitting the interaction term and fitting the model

$$\log(\text{Odds}) = \text{Corner} + \text{Age} + \text{Sex} + \text{Work}.$$

*From Healy, M. (1988) *GLIM. An Introduction*, Oxford Science Publications.

†The abbreviation Age⋆Sex is sometimes used for the group of terms

$$\text{Age} + \text{Sex} + \text{Age} \cdot \text{Sex}$$

Table 26.1. Prevalence of monoclonal gammapathy

	Agricultural (0)		Non-agricultural (1)	
Age	Male (0)	Female (1)	Male (0)	Female (1)
< 40 (0)	1/1590	1/1926	2/1527	0/712
40–49 (1)	12/2345	7/2677	3/854	0/401
50–59 (2)	24/2787	15/2902	5/675	4/312
60–69 (3)	53/2489	38/3145	3/184	1/80
70+ (4)	95/2381	63/2918	2/75	0/20

The estimated effect of work is −0.134 with standard deviation 0.244 in the model with interaction and −0.136 with standard deviation 0.243 in the model without. In this case, therefore, omitting the interaction term makes almost no difference.

Exercise 26.1. How should the effect of work be interpreted in terms of disease prevalence?

When using stratification or logistic regression to control for confounders it is best to keep the number of parameters in the model as low as possible. This is because both techniques are based on profile likelihood which can be unreliable when there are too many parameters to eliminate. Including interactions can require a lot of extra parameters, possibly too many to deal with by using profile likelihood. For example, if one confounder has 45 levels and another has 6 levels, then the model with interaction requires $5 \times 44 = 220$ extra parameters. Even when none of the confounders has a large number of levels it will still take many extra parameters to include interactions when there are a lot of them. For example, 10 confounders each with 3 levels require 180 extra parameters to include interactions between all possible pairs. In the monoclonal gammapathy example the model with interaction has 11 parameters while the model without interaction has only 7. By fitting a model without interaction we have reduced the number of parameters from 11 to 7. This is not a great saving and little is lost in this case by playing safe and fitting a model with the interaction.

It is possible, of course, to test for interaction between any pair of confounders. For the monoclonal example the deviance for the model with age–sex interaction is 6.771 on 9 degrees of freedom, and the deviance for the model without interaction is 7.649 on 13 degrees of freedom. The difference between these two deviances is only $7.649 - 6.771 = 0.878$, on 4 degrees of freedom, so the interaction is not significant. Unfortunately such a test has only sufficient power to be useful when based on a few degrees of freedom, and these are just the situations where nothing much is gained by omitting interactions. Thus the decision about whether or not to include interactions must usually be taken on other grounds. As a general rule, interactions between a confounder with many levels, and any other confounder, are omitted. For confounders with fewer levels it is only necessary to consider interaction between those pairs in which both are known to be very strongly related to the outcome. It is then probably best to include the interaction term for such pairs as a matter of course. Age and sex often form such a pair, and are usually controlled for by using a model which includes the age–sex interaction.

It can happen that a confounding variable has too many levels to be included into a logistic regression model, even before considering interactions. This occurs with matched case-control studies in which controls are individually matched to each case. Each case-control set then corresponds to a level of the categorical variable which defines the sets. The effects of this variable are of no interest but they must be included in the model when estimating the effects of other more interesting variables. The way out of this dilemma is to use conditional logistic regression (see Chapter 29) which uses a conditional likelihood in place of the profile likelihood.

26.2 Interaction between exposure and confounders

When controlling the effect of an exposure for the confounding effects of other variables there is a basic assumption that there is no interaction between exposure and the confounding variables. This assumption can be tested by comparing the model without interaction with a model containing the appropriate interaction term.

For example, when using the model

$$\log(\text{Odds}) = \text{Corner} + \text{Age} + \text{Sex} + \text{Work}$$

to control the effect of work for age and sex, there is an assumption of no interaction between work and age and no interaction between work and sex. To test the work and age interaction we compare the model without interactions with the model

$$\log(\text{Odds}) = \text{Corner} + \text{Age} + \text{Sex} + \text{Work} + \text{Work} \cdot \text{Age}.$$

To test the work and sex interaction we compare the model without interactions with

$$\log(\text{Odds}) = \text{Corner} + \text{Age} + \text{Sex} + \text{Work} + \text{Work} \cdot \text{Sex}.$$

Exercise 26.2. Use the deviances in Table 26.2 to test for interaction between work and the other two variables.

Table 26.2. Testing for interaction

Model	Deviance
Corner + Age + Sex + Work	7.65
Corner + Age + Sex + Work + Work·Age	5.81
Corner + Age + Sex + Work + Work·Sex	7.24

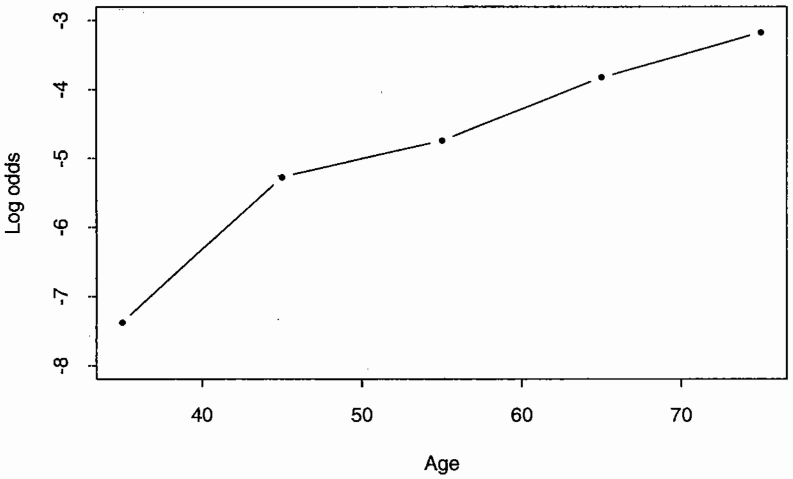


Fig. 26.1. Log prevalence odds by age

26.3 Confounders measured on a quantitative scale

The variable age in Table 26.1 is measured on a quantitative scale (years) which has been divided into five groups. When controlling for age we have the choice between treating it as categorical with five levels, treating it as quantitative with values equal to the mid- points of the five age groups, or treating it as quantitative with values on the original scale. The last of these alternatives is only possible when the data are in the form of individual records.

Fig. 26.1 shows a plot of the log of the prevalence odds against the mid-points of the age bands (35, 45, 55, 65, and 75 years) for male agricultural workers. The plot shows that the log odds increases approximately linearly with age. Plots for the other three groups in the study also show a roughly log-linear relationship with age.

Exercise 26.3. From Fig. 26.1 make a rough estimate by eye of the gradient of the line relating log odds to age. Express your answer per 10 years of age.

The model which assumes a log-linear relationship between odds and age for each work–sex combination has fewer parameters than the model which ignores the quantitative nature of the age scale, and this suggests that there may be some advantage in treating age as quantitative with values equal to mid-points of the five age groups. Making this modification to the model with age, sex, and work, we obtain

Table 26.3. A quadratic relationship with age

Parameter	Estimate	SD
Corner	−6.682	0.344
Work(1)	−0.148	0.243
[Age]	1.204	0.264
[Agesq]	−0.084	0.049
Sex(1)	−0.583	0.115

$$\log(\text{Odds}) = \text{Corner} + [\text{Age}] + \text{Sex} + \text{Work},$$

where [Age] refers to the effect for a change in age of one year. There are now only 4 parameters in this model and the work effect is −0.186 compared to −0.134 using the model in which age was treated as a categorical variable. This difference is large in comparison with the size of the effect, even though in neither analysis does the effect achieve statistical significance. The reason for the difference is that the relationship with age is not entirely linear.

We can test for linearity using a log-quadratic model for the relationship between log odds and age. The parameters in this model are estimated by fitting the model

$$\log(\text{Odds}) = \text{Corner} + [\text{Age}] + [\text{Agesq}] + \text{Sex} + \text{Work},$$

where the variable agesq takes as values the squares of the values of age. The results are shown in Table 26.3. When both [Age] and [Agesq] are included the deviance is 8.93 on 15 degrees of freedom — 3.13 less than when only [Age] is included. Referring this difference to the chi-squared distribution on 1 degree of freedom shows it to be significant at the 0.10 level. This would not normally be considered very convincing evidence of departure from linearity, but note that the estimate of the work effect is now in rather better agreement with earlier values.

The important lesson to be learned from this example is that the effect of a strong confounder such as age must be properly modelled, and that the yardstick of statistical significance may not be adequate for deciding upon the appropriate level of complexity. When the data are grouped in frequency records it is best to treat the variable as categorical; when using

Table 26.4. Interaction between age (quantitative) and work

Parameter	Estimate	SD
Corner	−6.211	0.201
Work(1)	−0.299	0.471
[Age]	0.763	0.058
Sex(1)	−0.584	0.115
[Age]·Work(1)	0.053	0.188

individual records it is best to err on the side of over-detailed modelling and to fit quadratic or even cubic dose-response relationships.

26.4 Interaction between categorical and quantitative variables

One situation where it can be valuable to treat a variable as quantitative is when testing for interaction; the resulting reduction in the number of parameters needed to measure interaction means that the test will be more powerful.

We have seen how to test for interaction between age and work when both are categorical variables, but what if age is a quantitative variable? The model without interaction, in which age is quantitative, is

$$\log(\text{Odds}) = \text{Corner} + [\text{Age}] + \text{Sex} + \text{Work}.$$

To test for interaction between work and quantitative age this is compared with

$$\log(\text{Odds}) = \text{Corner} + [\text{Age}] + \text{Sex} + \text{Work} + [\text{Age}] \cdot \text{Work}.$$

The model without interaction assumes that the gradient of the log-linear relationship of log odds with age is the same in both work groups, while the model which contains the interaction term allows for different gradients in the two work groups. The [Age].Work parameter measures the extent to which the gradient in the second work group differs from the gradient in the first, and its null value, corresponding to no interaction, is zero. Output for the model which includes the interaction between the linear effect of age and work is shown in Table 26.4.

Exercise 26.4. Use the output in Table 26.4 to test for interaction between age as a quantitative variable and work.

Exercise 26.5. How many parameters would there be for the interaction term [Age]·Work if there were three categories of work?

For a variable which is very strongly related to the response, such as age in this example, it may be necessary to model the relationship with age more closely than by using a linear relationship. Even so, the linear part of any new relationship will be the main part and it is worth testing for interaction just with this linear part. For example, if a quadratic relationship with age is used, as in the model

Table 26.5. Interaction between [Age] and Work

Parameter	Estimate	SD
Corner	−7.064	0.553
Age(1)	1.666	0.567
Age(2)	2.394	0.562
Age(3)	3.239	0.562
Age(4)	3.860	0.559
Sex(1)	−0.585	0.115
Work(1)	0.046	0.544
[Age]·Work(1)	−0.083	0.220

$$\log(\text{Odds}) = \text{Corner} + [\text{Age}] + [\text{Agesq}] + \text{Sex} + \text{Work},$$

then the interaction of work with the linear effect of age is tested by including the term [Age]·Work in the model. It is also possible to test for the interaction of work with the linear effect of age when the effect of age is modelled by a categorical variable. This is done by comparing

$$\log(\text{Odds}) = \text{Corner} + \text{Age} + \text{Sex} + \text{Work}.$$

with

$$\log(\text{Odds}) = \text{Corner} + \text{Age} + \text{Sex} + \text{Work} + [\text{Age}] \cdot \text{Work}.$$

This is a more powerful way of testing for interaction than including the term Age·Work (which has four parameters), provided the relationship with age is predominantly linear. Table 26.5 shows the results of this analysis, with quantitative age coded 0 to 4. The deviance for this model is 7.51, which is only a little smaller than the deviance for the model without interaction. Thus there is no evidence that the work effect varies with age. The same conclusion is reached by comparing the estimate of the interaction parameter with its standard deviation. Since the estimate of the work effect in the model without interaction is also not significant, it seems clear that these data provide no evidence for a relationship between agricultural work and the prevalence of monoclonal gammapathy.

Table 26.6. Model in terms of separate work parameters

Age	Work	log(Odds) = Corner + ⋯
0	0	–
1	0	Age(1)
2	0	Age(2)
3	0	Age(3)
4	0	Age(4)
0	1	Wbyage(1)
1	1	Wbyage(2) + Age(1)
2	1	Wbyage(3) + Age(2)
3	1	Wbyage(4) + Age(3)
4	1	Wbyage(5) + Age(4)

⋆ 26.5 What to do when there is interaction

Interaction parameters are chosen specifically to test for interaction; their estimated values are of no use in themselves. When there is interaction it is necessary to reparametrize so that the new parameters provide a satisfactory summary of the data in this situation. Indicator variables are a useful way of doing this.

Suppose, for example, that in a study of work and age there was an interaction between them. The most sensible way of reporting the results would be to estimate the effect of work separately for each level of age, but few packages allow this as a standard option. One way of doing it is by separating the data into age groups and analyzing these separately. Another is to reparametrize so that instead of one work parameter and four work·age parameters, we use five work parameters, one for each age group. Writing these separate work parameters as Wbyage, short for work by age, the model is shown in Table 26.6.

The values taken by the indicator variables for the age parameters are the same as before. The indicator variable for Wbyage(1) takes the value 1 when work is at level 1 and age is at level 0, and 0 otherwise; the indicator for Wbyage(2) takes the value 1 when work is at level 1 and age is at level 1, and 0 otherwise; and so on. One advantage of using indicator variables is that it is then possible to include another variable in the model with the indicators. This model imposes the constraint that the indicator effects are the same within the levels of this extra variable and provides estimates of their common values. It would not be possible to do this if the data were subdivided on age because subdividing on age is equivalent to fitting interaction terms of all variables with age.

When there is interaction between two exposures it is commonly reported by creating a new categorical variable with a level for each combination of the levels of the two exposures. For two exposures, each on four levels, the new variable would have 16 levels, with level 0 corresponding to level zero on both exposures and level 16 corresponding to level 3 on both exposures. There are 15 parameters for this new variable, measuring the ratio of the rate (or odds) for each one of the levels relative to the zero level. These are entered in the model in place of the 6 parameters for the two exposures and the 9 parameters for their interaction. The estimated parameters would be displayed in a four by four table, with the levels of one exposure determining the rows and the levels of the other determining the columns.

Table 26.7. Rate parameters per 100 000 person-years

	A	
B	0	1
0	5.0	15.0
1	20.0	λ

26.6 Interaction is scale-dependent ⋆

Interaction parameters are chosen to measure departures from a model. When the effects of variables are measured as ratios interaction parameters are ratios, chosen to measure departures from a multiplicative model. When the effects of variables are measured as differences (see Chapter 28) interaction parameters are differences chosen to measure departures from an additive model. Thus interaction depends on how the effects are measured. For example, consider two explanatory variables, A and B, each with two levels. Values for three of the parameters involved are shown in Table 26.7. For the moment the fourth parameter, λ, is left unspecified. When effects are measured as ratios the effect of A when B is at level 0 is $15/5 = 3$, and the effect of A when B is at level 1 is $\lambda/20$. The interaction parameter is the ratio of these two effects which is $\lambda/60$. When effects are measured as differences the effect of A when B is at level 0 is $15 - 5 = 10$, and the effect of A when B is at level 1 is $\lambda - 20$. The interaction parameter is now the difference between these two effects, which is $\lambda - 30$. It follows that if $\lambda = 60$ there is no departure from the multiplicative model but there is a departure from the additive model. Similarly if $\lambda = 30$ there is no departure from the additive model but there is a departure from the multiplicative model.

The choice between measuring effects as ratios or differences is usually an empirical one, with the investigator preferring to measure effects in such a way as to minimize the interaction, but there are sometimes biological grounds for preferring one method to the other.

Solutions to the exercises

26.1 The multiplicative effect of work is the ratio of the prevalence odds for non-agricultural workers to the prevalence odds for agricultural workers.

26.2 The degrees of freedom for the deviances are

$$\begin{aligned} 20-(1+4+1+1) &= 13 \\ 20-(1+4+1+1+4) &= 9 \\ 20-(1+4+1+1+1) &= 12 \end{aligned}$$

The change of deviance with inclusion of the Work.Age interaction is 1.84 with 4 degrees of freedom, and for the Work.Sex interaction it is 0.41 with 1 degree of freedom. Neither is significant.

26.3 The change in log odds over the age range of 35 to 75 is approximately +4. The gradient is therefore approximately +1 per 10 year age band.

26.4 The Wald test for interaction between the linear effect of age and work is

$$\left(\frac{0.053}{0.188}\right)^2 = 0.079,$$

which is not significant.

26.5 There would be two parameters for this interaction term.

27
Choice and interpretation of models

Previous chapters have illustrated the use of regression models using simple bodies of data containing relatively few variables. More commonly, we are faced with large data files containing many variables. Sometimes derived variables such as Quetelet's weight-for-height index are included in the model in addition to or in place of the original variables. In such situations it can be difficult to know where to begin, and all too easy to lose one's way. This chapter offers some guidance towards the sensible use of regression methods.

27.1 Variable selection strategies

A lot has been written about the process of finding the 'best' regression model in problems involving many variables. Much of this activity has been concerned with the search for an optimal strategy, and the relative merits of different approaches have been hotly debated. Many computer programs implement one or more of these strategies in an automatic model selection option called *stepwise regression.* These programs usually work by a combination of the *step-up* strategy (examining the effect of inclusion of variables not yet in the model) and the *step-down* strategy (examining the effect of of removing variables currently in the model). With the recent increased speed and reduced cost of computers, some programs now offer an exhaustive search of *all subsets* from a list of possible explanatory variables.

In assessing the value of such procedures it is important to note that regression models have two very different uses in epidemiology. Historically they were first used to derive *risk scores* designed to classify subjects into graded categories with respect to risk of developing disease. Later, when attention turned to interpretation of the parameter estimates and the close relationship between regression and stratification methods became apparent, regression models became important tools for analyses whose aim was the advancement of scientific knowledge. For convenience we refer to these two uses as *prediction* and *explanation*, respectively.

When the aim is prediction, the best model is the one which best predicts the fate of a future subject. This is a well defined task and automatic strategies to find the model which is best in this sense are potentially useful. However, when used for explanation the best model will depend on the scientific questions being asked, and automatic selection strategies have no place.

An important tool for assessing how well a model predicts the fate of a future subject is *cross-validation* — a technique in which each subject in turn is removed from the dataset and the actual outcome for that subject is compared with the predicted outcome using the model based on the remaining observations. The deviance for a model will always decrease with the introduction of more parameters, but prediction of future observations is not always improved. There comes a point at which increasing the complexity of the model to gain a slightly better fit to the observed data will reduce the accuracy of its predictions. Cross-validation measures the predictive properties of the model directly and therefore reflects the adverse consequences of fitting too many parameters.

Cross-validation is potentially expensive in computer time, but simple approximate criteria have been developed which allow the assessment of whether any step up or down in an automatic model selection procedure would be expected to improve prediction. The best known is *Akaike's information criterion*, namely

$$(\text{Reduction in deviance}) - 2 \times (\text{Increase in number of parameters}).$$

If this is positive the increased complexity would be expected to improve prediction and if negative, to degrade prediction.

27.2 Explanatory variables and natural experiments

This book has been entirely concerned with the use of models whose aim is explanation. In such analyses there is a clear distinction between the roles of exposures and confounders but this distinction is lost when using regression models — both become explanatory variables. Ignoring the distinctions between different types of explanatory variable is appropriate when using regression models for prediction, since all variables have the same role, but in a scientific analysis of data different explanatory variables may play quite different roles.

The distinction between exposure and confounder, as described in this book, relies heavily on the idea of experiments of nature. An exposure is something which we can intervene to change while a confounder is a variable which we would have held constant had we designed the experiment rather than leaving it to nature. It is helpful to think of regression analysis as simulating an experiment, in the same way. For example, the effects of A in the model

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

are the effects of changing the level of A in a simulated experiment in

which B and C are held constant. Similarly, the effects of B are the effects of changing the level of B in a simulated experiment in which the levels of A and C are held constant. Thus regression analysis does not simulate a single experiment but many. This flexibility of the regression approach is undoubtedly useful, but in practice it can also become its most serious weakness. To extend our analogy, the data analyst is in a position like that of an experimental scientist who has the capability to plan and carry out many experiments within a single day. Not surprisingly a cool head is required! Before embarking on a regression analysis it is essential to spend an hour or so, preferably away from the computer, to list the main scientific questions and to think how these can be answered by fitting a series of models. Analyses which follow such thought are always simpler and more incisive than those which are born of uncritical use of the computer or worse, of a stepwise regression program.

It will rarely be necessary to include a large number of variables in the analysis, because only a few exposures are of genuine scientific interest in any one study, and there are usually very few variables of sufficient *a priori* importance for their potential confounding effect to be controlled for. Most scientists are aware of the dangers of analyses which search a long list of potentially relevant exposures. These are known as *data dredging* or *blind fishing* and carry a considerable danger of false positive findings. Such analyses are as likely to impede scientific progress as to advance it. There are similar dangers if a long list of potential confounders is searched, either with a view to explaining the observed relationship between disease and exposure or to enhancing it — findings will inevitably be biased. Confounders should be chosen *a priori* and not on the basis of statistical significance. In particular, variables which have been used in the design, such as matching variables, must be included in the analysis.

Recently there has been some dispute between 'modellers', who support the use of regression models, and 'stratifiers' who argue for a return to the methods described in Part I of this book. Logically this dispute is based on a false distinction — there is no real difference between the methods. In practice the difference lies in the inflexibility of the older methods which thereby imposes a certain discipline on the analyst. Firstly, since stratification methods treat exposures and confounders differently, any change in the role of a variable requires a new set of computations. This forces us to keep in touch with the underlying scientific questions. Secondly, since strata must be defined by cross classification, relatively few confounders can be dealt with and we are forced to control only for confounders of *a priori* importance. These restraints can be helpful in keeping a data analysis on the right tracks but once the need for such discipline is recognized, there are significant advantages to the regression modelling approach.

EXAMPLE: DIETARY FAT AND TOTAL ENERGY INTAKE

The analogy between regression models and imaginary experiments is very useful in making decisions about whether to include a variable in a regression model or not. An interesting illustration arises in nutritional epidemiology when considering the relationship between total energy intake and the incidence of coronary heart disease. This relationship was first detected because relationships were observed between intake and disease risk for a large number of nutrients — the more that was eaten, the lower the risk. A relationship with total energy intake, possibly reflecting energy expenditure, was considered the most likely explanation.

However, once this relationship is recognized, how should the relationship between risk and other aspects of the diet, notably fat intake, be analysed? One way is to measure *nutrient density*, which is the ratio of daily intake of fat to the total energy intake. This approach is open to the criticism that such nutrient densities are not usually independent of total energy intake — subjects with high energy intakes typically have a different pattern of nutrient densities from subjects with low energy intakes.

If energy intake is to be regarded as a confounder, then it should be controlled for, either by stratification or with a regression model. In the latter case we fit a model such as

$$\log(\text{Rate}) = \text{Corner} + \text{Fat} + \text{Energy}$$

and interpret the parameters representing the effect of fat in terms of an experiment in which fat intake is varied but the total energy content of the diet is held constant. Of course, such an experiment would require other constituents of the diet such as carbohydrate to vary in order to maintain the total energy intake and this must be born in mind when interpreting parameters.

Exercise 27.1. How would you interpret the effect of fat in the model

$$\log(\text{Rate}) = \text{Corner} + \text{Fat} + \text{Carbohydrate} + \text{Energy}?$$

Other authors have approached the problem of allowing for total energy expenditure by dividing total calories between calories from fat and calories from other sources, and fitting the model

$$\log(\text{Rate}) = \text{Corner} + \text{Fat-calories} + \text{Other-calories}.$$

The parameters representing the effect of fat intake must now be interpreted in terms of an experiment in which fat intake is varied while intake of other calories is held constant. In this experiment a reduction of fat intake would result in a reduction of total energy intake. Such an experiment would be difficult to interpret, even if it could be carried out.

Finally we should point out that a real public health intervention to reduce dietary fat intakes would be unlikely to mimic either of the above imaginary experiments. When dietary fat intake is reduced in free–living subjects, some of the energy intake is made up from other sources, but typically there is a net reduction in energy intake. This demonstrates that the use of models to predict the effect of intervention usually requires considerable extra knowledge. In particular, we need to have some understanding of the mechanism by which change will be effected.

27.3 Endogenous and exogenous explanatory variables

The 'effects' of an explanatory variable are defined in terms of differences in log rate (or log odds) between groups of subjects with different levels of the variable. Thus the effect of cigarette smoking is defined by contrasting rates in smokers and non-smokers, and the effect of serum cholesterol concentration (classified as high or low) is defined as the difference in log rate between subjects with high cholestrol concentration and subjects with low cholesterol concentration. This language encourages people to interpret 'effects' as the change in rates to be expected as a result of intervention to change the level, but this is a big step. How are the subjects to alter their level? For a variable like serum cholesterol there is no direct way to alter its level and any intervention would have to be indirect, for example by change of diet or by cholesterol lowering drugs. However, there is no guarantee that such mechanisms will bear any relationship to the mechanism which led the the study subjects to have different levels in the first place. The effect of indirectly changing the levels of serum cholesterol in a group of subjects may be completely different from that estimated by comparing groups of subjects who just happen to have different levels of cholesterol.

The same problem arises in an even more acute form when studying the effects of two or more interrelated variables, such as blood pressure and obesity in relation to the incidence of coronary disease. The effect of blood pressure controlled for obesity might now be interpreted as the expected effect of changing blood pressure while keeping obesity constant. However, is it be possible to intervene to change blood pressure while keeping obesity constant? While this could be achieved, for example by using drug treatment, this method of intervention would bear little relation to the mechanism that led subjects to their current levels in the first place, and it might have different effects. Intervention aimed at life style changes are more likely to duplicate these conditions, but might be expected to change both blood pressure and obesity simultaneously. In this case the estimated effects of blood pressure controlled for obesity, or obesity controlled for blood pressure could be poor predictors of the effect of the intervention.

The position is much clearer when considering environmental exposures, such as radiation dose, occupational exposure to toxic chemicals, and even cigarette smoking. In such cases, it is entirely reasonable to imagine an experiment in which exposure of groups is directly varied without any consequent change in other variables, and the parameters of regression models are easier to interpret.

Variables such as cholesterol concentration, blood pressure, and obesity are called *endogenous*. The word endogenous means 'growing from within'. Variables such as smoking, diet and occupation are called *exogenous*. The distinction between endogenous and exogenous variables is borrowed from the behavioural sciences and, although the distinction is not hard and fast, is useful in drawing attention to the different assumptions which it is necessary to make for the two kinds of variable when interpreting the parameters of regression models as expected effects following intervention.

27.4 Interpretation of interaction

An underlying theme of this chapter is that while distinctions between different types of explanatory variable are not relevant to the mechanical process of estimating the parameters of a regression model, they are essential to the strategy adopted in the analysis and to the interpretation of results. This is particularly true when dealing with interaction. The word describes a purely mathematical concept in regression models. Its relationship to the scientific language of epidemiology requires further consideration of the nature of the variables involved.

We shall first consider interaction between two confounders. There seems to be no word to describe this in epidemiology, almost certainly because the phenomenon is of no scientific interest. Whether we include such terms in a model or not is a purely technical matter of trading the number of parameters against freedom from assumptions. Usually if there are two strong confounders such as age and sex, the gain in efficiency from assuming no interaction between them is extremely modest and it will usually be safer to include an interaction term regardless of its significance. However, if we are worried about the aggregate effect of five or six weak confounders, then omission of interaction terms is unlikely to have a major effect on estimates of parameters of interest.

Interaction between a confounder and an exposure of interest is known in epidemiology as *effect modification* and is clearly of considerable scientific importance, since the *consistency* of an effect in diverse study groups would usually be considered relevant to labelling a relationship as 'causal', in the sense of predicting the effect of future interventions. The ease with which we can test for such interaction in the framework of regression models represents a clear advance over earlier stratification methods in which the absence of such interaction is a hidden assumption.

Finally, the question of interaction between two exposures of interest is usually of considerable importance, both for the scientific interpretation of an analysis and for its implications for preventive intervention. We shall deal with this in more detail in Chapter 28.

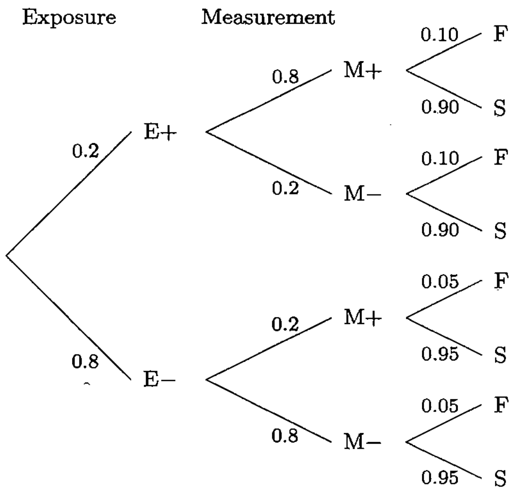


Fig. 27.1. Misclassification of exposure.

27.5 Errors of measurement of explanatory variables

In the models discussed in this book it is assumed that explanatory variables are correctly measured. This assumption is often unjustified in practice, but epidemiologists have generally been prepared to ignore measurement errors. Some have believed that to do so is justifiable providing there is no relationship between errors of measurement of exposure and disease outcome, that is if there is no *differential misclassification*. This is now known to be false.

To illustrate the effect of ignoring measurement error we consider the hypothetical situation illustrated in Fig. 27.1, in which exposure E is measured imperfectly by measurement M. As a result of this misclassification there is a probability of 0.2 that an exposed subject is misclassified as unexposed, and a probability of 0.2 that an unexposed subject is misclassified as exposed. The probability of failure depends only on true exposure, taking the value 0.1 for exposed subjects and 0.05 for unexposed subjects. An epidemiological study observes only the marginal relationship between measured exposure and failure.

Exercise 27.2. Calculate probabilities for each of the eight tips of the tree in Fig. 27.1. By collapsing over exposure categories, calculate the probabilities for each of the four possible combination of measured exposure and disease (failure) status. Hence derive the probability tree expressing the probability of failure

Table 27.1. Diastolic blood pressure (DBP) and rate ratios for stroke

Baseline	Rate	Mean DBP	
DBP	ratio	at baseline	after 2 years
≤ 69	0.276	63.6	72.7
70–79	0.395	73.8	77.0
80–89	0.595	83.6	83.0
90–99	1.000	93.5	91.2
100–109	1.904	103.4	99.2
≥ 110	3.875	116.4	107.3

conditional upon measured exposure.

It is clear from this exercise that the effect of exposure is decreased by the measurement error: whereas the risk ratio for true exposure is 2, the risk ratio for measured exposure is only 1.42. It is worth noting that 20% misclassification would be regarded as acceptable in many branches of epidemiology.

Similar considerations apply when exposure takes on more than two levels. The observed dose-response relationship between measured exposure and disease outcome is less steep than the underlying relationship with true exposure, under any realistic assumptions about the dose-response relationship. This is illustrated by the data of Table 27.1 which concern the relationship between diastolic blood pressure and subsequent incidence of stroke.* These data are taken from a re-analysis of seven cohort studies, and the first two columns of the table summarize the relationship between diastolic blood pressure at a single initial visit (the 'baseline' measurement) and subsequent incidence. Note that in the rate ratios the fourth category is taken as reference. These were obtained by fitting the model

$$\log(\text{Rate}) = \text{Corner} + \text{Study} + \text{DBP}$$

where study is a categorical variable with one level for each study, so that confounding of the relationship due to differences between the study cohorts is eliminated. The third column shows the mean of the baseline diastolic pressures for each of the five categories. The log rate ratios are plotted against the mean baseline values in Fig. 27.2 (solid line). This line represents the apparent dose-response relationship between a single measurement of diastolic blood pressure and the incidence of stroke. It is approximately log-linear, so that essentially the same relationship would have been obtained by fitting the model

$$\log(\text{Rate}) = \text{Corner} + \text{Study} + [\text{DBP}],$$

*From Macmahon, S. *et al.* (1990), *The Lancet*, **335**, 765–774.

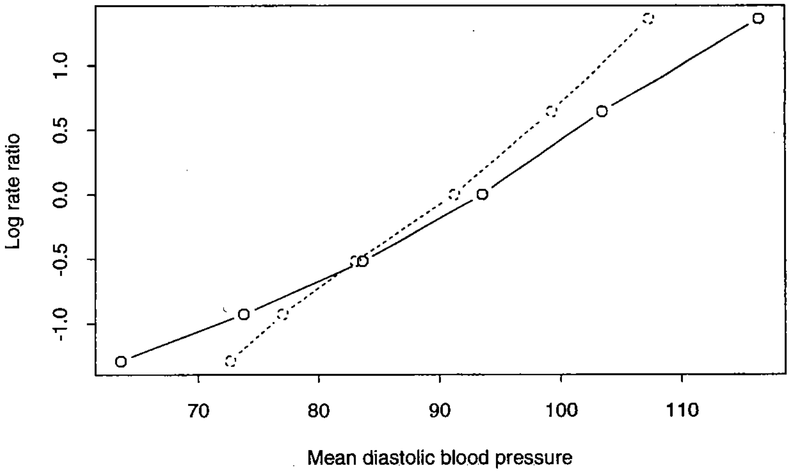


Fig. 27.2. Apparent and true dose-response relationships.

where [DBP] is measured per mm Hg. However, this line is a poor representation of the true relationship between blood pressure and the incidence of stroke. Blood pressure is subject to both short-term fluctuations and to measurement errors, neither of which will be reflected in the risk of stroke which is determined by the longer-term average level of blood pressure. The final column of Table 27.1 shows the mean blood pressure taken two years later in representative samples taken from each of the five groups. These figures provide a better estimate of long-term average blood pressure in the six groups as the short-term fluctuations and measurement errors are washed out. Plotting the rate ratios for stroke against these new values for mean diastolic blood pressure provides a truer estimate of the relationship between stroke incidence and the long- term average level of diastolic pressure. This plot is shown in Fig. 27.2 as a broken line and clearly represents a stronger relationship than the apparent relationship based on a single baseline measurement. This finding is true in general. When an explanatory variable suffers from measurement error or within subject variability the linear effects of this variable will be closer to zero than when there is no error or variability. This is known as *regression dilution*

This second example demonstrates both the attenuation of relationships owing to exposure measurement error and one of the methods which has been suggested for correcting for it. An alternative approach is to formally adopt probability models such as that illustrated in Fig. 27.1 and to estimate the conditional probabilities for every branch of the tree. Validation substudies are required in order to estimate the misclassification probabilities. A difficulty with this approach is that when there are several levels of exposure, the number of parameters in the model can become very large.

In summary, when exposures are subject to measurement error, the apparent exposure effects will be less pronounced than the true underlying relationships. When confounders are measured inaccurately, the consequences are even more serious. Since the relationship between disease and confounder is not correctly estimated in these circumstances, it follows that the analysis will not properly control for confounding. If both exposure and confounder are measured inaccurately, there exists the possibility that the two sets of errors may be interrelated, so that the apparent relationship between exposure and confounder may be quite different from that between the underlying variables. In these circumstances models for relationships between measured exposure, measured confounder, and response have no interpretation in terms of an imaginary experimental intervention and may be scientifically meaningless. Such might well be the position in our example involving dietary fat and total energy intake. Measured intakes of total energy and of each specific nutrient are usually derived from the same dietary records, taken over a period of several days. Not only are such measurements very imperfect measures of long-term intake, but it is reasonable to believe that errors in the measured fat intake will be closely related to errors in measured energy intake, since the former is an important contributor to the latter. Regression models which include total energy as well specific nutrients may, therefore, not be interpretable in practice.

Solutions to the exercises

27.1 The parameter(s) measure the effect of changes in fat intake while holding both total energy intake and carbohydrate intake constant. To reduce fat intake while holding both total energy and carbohydrate intake constant would be very difficult for an individual to do and would require large changes in other components of the total energy intake, such as protein.

27.2 From top to bottom the probabilities are 0.016, 0.144, 0.004, 0.036, 0.008, 0.152, 0.032, and 0.608. The remaining calculations are shown in Fig. 27.3. The probability of failure conditional upon having been measured as exposed is 0.075, while the failure probability conditional upon having been measured as unexposed is 0.053.

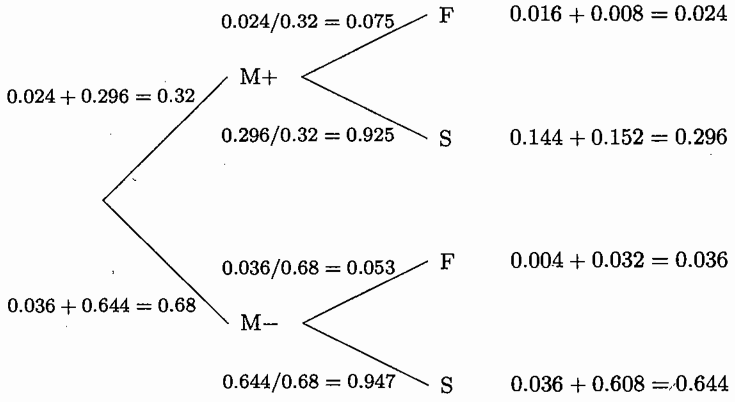


Fig. 27.3. Failure probabilities conditional upon measured exposure.

28

★ Additivity and synergism

When discussing the way two exposures combine to influence the risk of disease the word interaction is used to refer to departures from either multiplicative or additive models. In general these models have no biological basis and interaction is therefore a purely statistical concept. The interaction parameters are chosen solely to test hypotheses and are not useful for describing the data when there is interaction. The word *synergism* is often used, in a similar sense, to refer to departures from a biological model for the independent action of two exposures. When the joint effect of two exposures is greater than would be expected from the separate effects, according to such a model, the exposures are said to display positive synergism. Synergism is therefore a particular kind of interaction but precisely what kind depends on the biological model for independent action.

Epidemiologists often use the word synergism without specifying precisely what they mean by independent action. In other words they use it in a statistical sense. When used in this way synergism is generally measured as a departure from an additive model. This suggests an ill-defined biological model which predicts that the rate for the joint effect of two exposures is the sum of the rates for the separate effects. An example of such a model is shown in Fig. 28.1 which refers to a situation where disease is caused by one or other of two *precipitating* events. Exposure A influences the chance of the first event occurring, while exposure B influences the chance of the second event occurring. When A and B act independently their effects on the rate will be additive because

$$\text{Rate(Event 1 } \textit{or } 2) = \text{Rate(Event 1)} + \text{Rate(Event 2)}.$$

In cases like this it makes sense to fit an additive model so that departures from this model can be measured and used to test whether the two exposures act independently. In this chapter we consider some of the special problems which arise when using additive regression models.

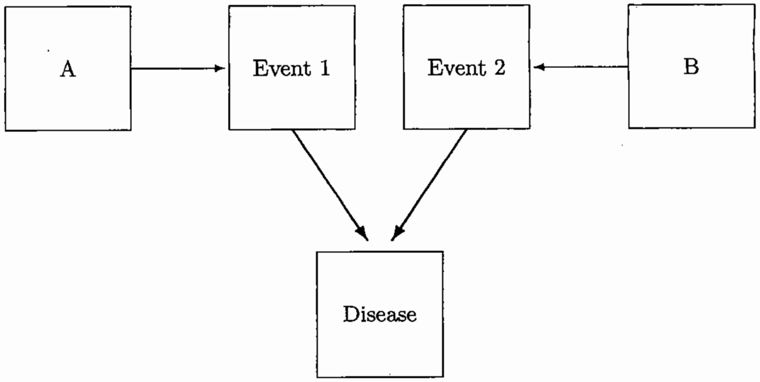


Fig. 28.1. Two precipitating events for disease.

28.1 Fitting additive models

With additive models effects are measured as differences between rates (or odds) parameters rather than as ratios. The use of stratification to control the additive effects of an exposure for confounding would be based on the assumption that the difference between the rate parameters for the different levels of exposure is constant over the strata. Formulating the same problem in terms of regression models the effects of an exposure controlled for a confounder are found by fitting the additive model for the rate,

$$\text{Rate} = \text{Corner} + \text{Exposure} + \text{Confounder}.$$

The assumption that the additive effect of the exposure is the same for all strata formed by the confounder is expressed by the fact that the model is additive, with no interaction terms.

Additive models are fitted to data by choosing parameters to maximize the log likelihood in the same way as for multiplicative models, but the calculations are different and require different computer programs. Similarly log likelihood ratios are used to test hypotheses in the same way as for multiplicative models. In practice additive models can be more troublesome to fit than multiplicative models because the most likely parameter values do not *necessarily* predict rates which are greater than zero. It is then rather difficult to know what to do. Should one treat this as evidence that the additive model is a poor fit, or should one find most likely values subject to the constraint that they predict positive rates? Generally the latter policy is followed, but it can be difficult to implement.*

*This problem does not arise with multiplicative models because these are fitted as additive models for the log rate and the log rate is not constrained to be positive.

28.2 Discriminating between additive and multiplicative models

When there are rival biological grounds for choosing an additive model and a multiplicative model the investigator will wish to discriminate between the two models by seeing which fits the data best. The deviances for the two models provide an informal way of looking at this but they cannot be compared in a formal test because the additive and multiplicative models are not nested. The solution to this technical problem is to find an *extended model* which contains both additive and multiplicative models as special cases. One such model is

$$\frac{(\text{Rate})^\rho - 1}{\rho} = \text{Corner} + \text{A} + \text{B},$$

where ρ is a parameter yet to be determined. In this model A and B refer to parameters which measure differences in the value of

$$\frac{(\text{Rate})^\rho - 1}{\rho}.$$

As ρ approaches 1 the model reduces to

$$\text{Rate} - 1.0 = \text{Corner} + \text{A} + \text{B}$$

in which the A and B parameters measure differences in the rate. As ρ approaches zero, the left-hand side of the model approaches the log of the rate [†] , so the model reduces to

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

in which the A and B parameters measure differences in the log rate. The two extremes of the extended model therefore correspond to an additive model ($\rho = 0$) and a multiplicative model ($\rho = 1$). When this extended model is fitted for a range of values for ρ, including $\rho = 1$ and $\rho = 0$; a comparison of the log likelihoods for the different values of ρ will indicate which is the most likely value for ρ and whether the additive or multiplicative model is preferred. It may turn out, of course, that both models provide an adequate fit, or that neither model is acceptable. We do not advocate the use of the model with values of ρ other than zero or one, because effect parameters measured as differences in the value of

$$\frac{(\text{Rate})^\rho - 1}{\rho}$$

[†] This follows because, for small ρ,

$$R^\rho = [\exp(\log(R))]^\rho = \exp[\rho \log(R)] \approx 1 + \rho \log(R).$$

would be hard to interpret. The sole purpose of the extended model is to provide a framework in which to choose between additive and multiplicative models.

Using the extended model to discriminate between multiplicative and additive models involves fitting a non-standard regression model for each of a range of values of ρ. Even with software which allows non-standard models this can be quite a lot of work.

28.3 Additive models with case-control studies

There are some special problems which arise when trying to fit additive models to data from case-control studies. To illustrate these we shall consider a case-control study of the joint effect of two exposures A and B in which the ratio of sampling probabilities is

$$K = \frac{\text{Probability of selecting a failure as a case}}{\text{Probability of selecting a survivor as a control}}.$$

We showed in Chapter 23 that parameters which are defined as ratios of the odds of being a case are also ratios of the corresponding odds of failure in the study base. Unfortunately this does not apply to additive models. Parameters which are defined as differences in the odds of being a case are K times the corresponding differences in the odds of being a failure in the study base. The factor K, which relates the odds of being a case to the odds of faliure, cancels in ratios but not in differences. It follows that fitting an additive model to case-control data tells us nothing about the additive effects on the odds of failure in the study base except in those rare cases where the value of K is known. It is still possible, of course, to test hypotheses about zero parameter values since a zero additive effect on the odds of being a case corresponds to a zero additive effect on the odds of being a failure in the study base.

Although it is not possible to estimate the additive effects of A and B on the odds of failure in the study base it is still possible to estimate the ratio of these effects to the corner. This is less satisfactory than estimating differences in the odds themselves, but better than nothing. These new parameters are estimated by fitting the model

$$\text{Odds} = \text{Corner} \times (1.0 + \text{A} + \text{B}).$$

When the model is written in this way the corner parameter is still the odds of being a case when A and B are at level zero, but the A and B parameters are now differences in the ratio

$$\frac{\text{Odds}}{\text{Corner}}.$$

Table 28.1. Estrogen replacement, weight, and endometrial cancer

Weight	Estrogen replacement			
	No		Yes	
(kg)	Cases	Controls	Cases	Controls
< 57	12	183	20	61
57–75	45	378	37	113
> 75	42	140	9	23

This model can be fitted to data using likelihood in the same sort of way as for conventional models but special software is required.

Exercise 28.1. Table 28.1 shows results of a case-control study relating endometrial cancer incidence to use of estrogen therapy and body weight. Calculate odds ratios for each category of weight and estrogen use relative to the corner (top left corner cell). Obtain differences in these odds ratios for estrogen replacement yes compared to estrogen replacement no, at each level of weight. Do the data appear consistent with an additive model?

When a case-control study is stratified by age at time of diagnosis, and controls are sampled separately in each age stratum, there will be a different value of K for each stratum. To make sure the A and B parameters do not depend on these K's the parameters must now be defined as differences in the value of

$$\frac{\text{Odds}}{\text{Age specific corner}},$$

where the age specific corners are the odds in each age stratum when A and B are both at level 0. The A and B parameters will then equal the corresponding differences in the ratio of the odds of failure to the age specific corners in the study base.

Assuming that the new A and B parameters are constant over age strata, their common value can be estimated by fitting the model

$$\text{Odds} = \text{Corner} \times \text{Age} \times (1.0 + \text{A} + \text{B})\,.$$

where age is a categorical variable with one level for each age stratum. The $\boxed{\text{Corner} \times \text{Age}}$ part of the model corresponds to fitting separate corner parameters for each age stratum. This model again requires special software.

28.4 Discriminating between models using case-control studies

The extended model containing the extra parameter ρ can also be used to compare the fit of a multiplicative model with an additive model using data from a case-control study. The two models we wish to compare are

$$\text{Odds} = \text{Corner} \times \text{A} \times \text{B},$$

in which A and B parameters are ratios of odds, and

$$\text{Odds} = \text{Corner} \times (1.0 + \text{A} + \text{B}),$$

in which the A and B parameters are differences in the ratios of odds to the corner. The multiplicative model can also be written in the form

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

in which the A and B parameters are defined as differences in log odds. The extended model is now

$$\frac{(\text{Odds}/\text{Corner})^{\rho} - 1.0}{\rho} = \text{A} + \text{B}.$$

As ρ approaches 0 this model approaches

$$\log(\text{Odds}/\text{Corner}) = \text{A} + \text{B},$$

which simplifies to

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

This is the multiplicative model written in log form, apart from the fact that because the corner parameter is on the original scale in the extended model it appears as log(Corner). As ρ approaches 1, the extended model approaches

$$\text{Odds} = \text{Corner} \times (1.0 + \text{A} + \text{B}),$$

which is the additive model.

The procedure for comparing the fit of a multiplicative and an additive model is illustrated by fitting the extended model to the data in Table 28.1 for a range of values of ρ. To actually do this involved fitting a non-standard model for each of these values. The resulting log likelihood ratios are shown in Fig. 28.2. At $\rho = 0$ the log likelihood ratio is -2.774 and at $\rho = 1$ it is -0.408. To test for the adequacy of the multiplicative model we take $\rho = 0$ as the null value. Minus twice the log likelihood ratio for $\rho = 0$ is 5.548 ($p \approx 0.02$), so the data do not support this model. To test for the adequacy of the additive model we take $\rho = 1$ for the null value. Minus twice the log likelihood ratio for $\rho = 1$ is 0.816 ($p > 0.10$) so the data are consistent with the additive model.

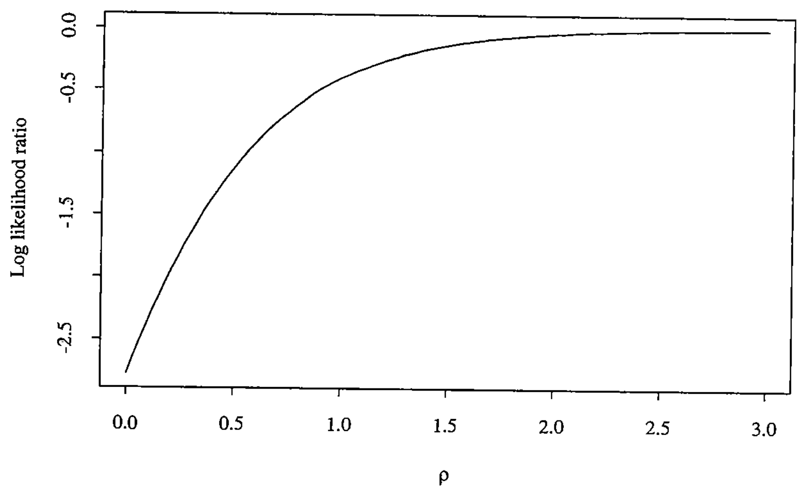


Fig. 28.2. The log likelihood ratio for ρ.

The most frequent outcome when comparing the fit of multiplicative and additive models is that both provide an acceptable description of the data. This has been taken by some epidemiologists as a serious flaw in the modern modelling approach to statistical analysis, since additive and multiplicative models have radically different public health implications (notably in relation to the targeting of interventions). This difficulty is indeed serious, but it is attributable more to an attempt to extrapolate beyond the data than to any shortcomings in statistical methodology.

A good example of this arises in attempts to study the implication of different dose-response relationships for the carcinogenic effect of ionizing radiation. The public health problem (if there is one) is one of relatively large populations exposed to low doses, but the available epidemiological studies have concentrated upon high exposure groups — A-bomb survivors, irradiated patient groups and so on. Additive and multiplicative dose-response models make similar predictions at high doses so these studies are poorly discriminated. However, they make very different predictions for subjects receiving low dose exposure. If data were available for subjects receiving low dose exposure the two models would be easily discriminated; the problem lies in trying to discriminate between them using data from a range of dose levels for which the two models make the same predictions.

Exercise 28.2. We plan to reduce the total burden of disease in a community by attempting to eliminate exposure A but another explanatory variable, B, is also known to be important. Should the intervention be targeted on individuals whose exposure to B is greatest? Consider how the answer to this question depends on whether the effects of A and B on the rate are additive or multiplicative.

Solutions to the exercises

28.1 The odds ratios are shown below.

Weight	Estrogen replacement		
(kg)	No	Yes	Difference
< 57	1.00	5.00	4.00
57–75	1.82	4.99	3.17
> 75	4.58	5.97	1.39

The additive model does not appear to fit particularly well as the differences between the odds ratios for the two estrogen groups seems to fall with increasing weight. Further examination of the table suggests the possibility that there is only a relationship with weight when there is no estrogen replacement.

28.2 Consider a population classified according to the two factors A and B. When these act additively or multiplicatively, the rates follow one of the following patterns:

	Additive model			Multiplicative model		
	A		Potential	A		Potential
B	No	Yes	reduction	No	Yes	reduction
No	1	3	2	1	3	2
Yes	3	5	2	3	9	6

When the multiplicative model holds the reduction in rates by eliminating exposure A is greater in the B-Yes group than in the B-No group. It would therefore be cost effective to target intervention at the high- risk section of the population. When the additive model holds this is no longer the case — there is an equal potential reduction in both sections of the population, and targeted intervention makes little sense.

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 $i = 1, 2, \dots$, 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 $t = 0, 1, 2, \dots$ 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

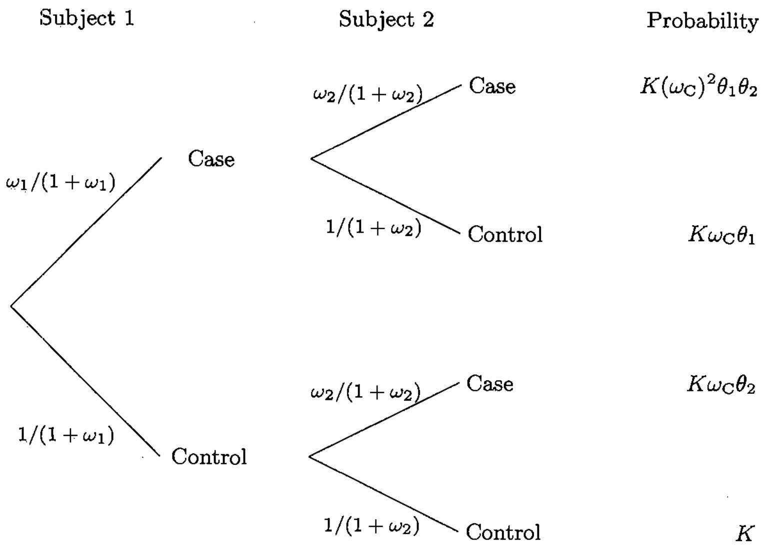


Fig. 29.1. Disease status for two subjects in a case-control study.

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 $\boxed{\star}$

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

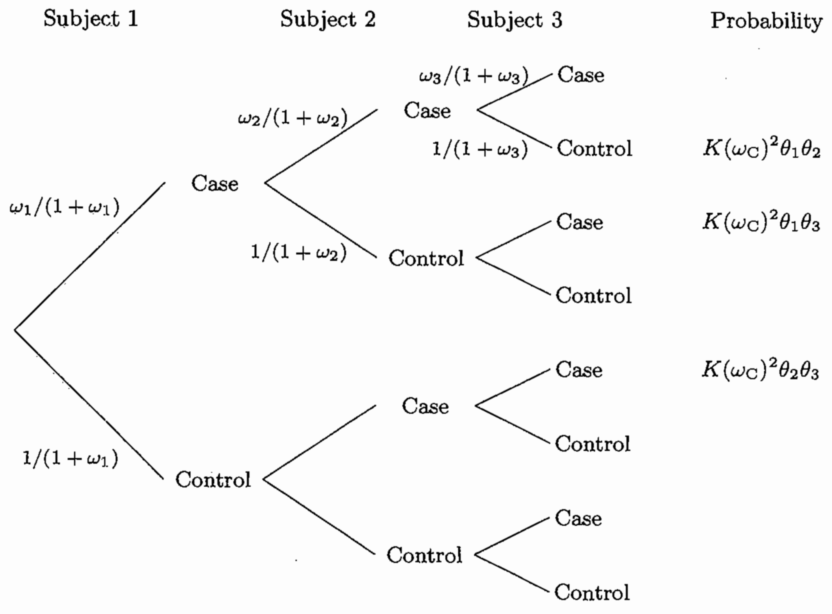


Fig. 29.2. Sets with two cases and one control.

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

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

30
Cox's method for follow-up studies

When using Poisson regression models to analyse data from follow-up studies, time is divided into fairly broad bands such as 5 or 10 years of age. Age is the most common time scale but in some applications other time scales may be more relevant. This point is discussed in more detail in the next chapter, but for the moment we refer to the time scale simply as time. Cox's method is very similar to Poisson regression but is based on a much finer subdivision of time.

30.1 Choosing parameters

When there are two explanatory variables, A and B, and the rate is allowed to vary with time, the multiplicative model for the rate takes the form

$$\text{Rate} = \text{Corner} \times \text{Time} \times \text{A} \times \text{B}.$$

Here time is a categorical variable with one level for each time band. Again we split the model into two parts, as in

$$\text{Rate} = \boxed{\text{Corner} \times \text{Time}} \times \boxed{\text{A} \times \text{B}}\,.$$

Algebraically this corresponds to a reparametrization of the model as

$$\lambda_i^t = \lambda_{\text{C}}^t \theta_i,$$

where λ_{C}^t is a corner parameter measuring the rate for time band t when A and B are both at level 0, and θ_i is the rate ratio which compares the rate for subject i, in time band t, to the corner rate for that time band. The parameters λ_{C}^t correspond to the

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

part of the model and the parameters θ_i to the

$$\boxed{\text{A} \times \text{B}}$$

part of the model.

30.2 The profile likelihood

The parameters λ_{C}^t are also called the *baseline* rates, and are generally nuisance parameters. The main interest is in the parameters of the second part of the model. The profile likelihood for the parameters in the second part of the model is obtained by deriving formulae for the most likely values of the nuisance parameters, λ_{C}^t, and substituting these into the expression for the log likelihood. The number of nuisance parameters depends upon the number of time bands into which the total study period has been partitioned. For the present we shall consider a finite number of bands, but in the next section the argument is generalized to the case where time is divided into clicks.

The contribution of subject i to the log likelihood is the sum of contributions for each time band. These have the Poisson form:

$$d_i^t \log(\lambda_i^t) - y_i^t \lambda_i^t$$

where y_i^t is the observation time in time-band t and d_i^t indicates whether the event occurred ($d = 1$) or not ($d = 0$). The total log likelihood is the sum of such terms over all subjects (i) and all time bands (t). Rewriting λ_i^t as $\lambda_{\text{C}}^t \theta_i$, this becomes

$$\sum_{i,t} \left[d_i^t \log(\lambda_{\text{C}}^t \theta_i) - y_i^t \lambda_{\text{C}}^t \theta_i \right].$$

The rules of calculus show that, given the θ_i, the most likely values of the baseline rates λ_{C}^t are

$$\frac{d^t}{\sum_i y_i^t \theta_i},$$

where d^t represents the total number of events occurring in time band t. Substituting these values into the expression for the log likelihood yields a profile log likelihood which depends only on the parameters in the second part of the model. This is

$$\sum_{j,t} d_j^t \log \left(\frac{\theta_j}{\sum_i y_i^t \theta_i} \right).$$

30.3 Time divided into clicks

The profile log likelihood derived by stratifying the follow-up interval into bands provides a satisfactory method for regression analysis of cohort studies; but although this is the approach used with frequency records it is rarely used with individual records. The reason for this is that a further generalization offers increased flexibility without seriously compromising either statistical or computational efficiency. In this generalization the time scale is subdivided into clicks which can contain no more than one event, thus allowing rates to vary continuously over time.

The consequence of this generalization for the profile log likelihood are quite minor. First consider the effect upon the observation times, y_i^t. If the duration of the time bands is h and we allow h to become very small, almost every y_i^t will become either zero (if subject i was not observed at click t) or h (if subject i was observed). In these circumstances, it is convenient to redefine y_i^t to be *at risk indicators* taking on the values 0 or 1 respectively. The observation times then become hy_i^t and the profile log likelihood for the rate ratio model becomes

$$\sum_{j,t} d_j^t \log\left(\frac{\theta_j}{\sum_i hy_i^t\theta_i}\right),$$

which may be further simplified to

$$\sum_{j,t} d_j^t \log\left(\frac{\theta_j}{\sum_i y_i^t\theta_i}\right) - D\log(h).$$

Since the term $D\log(h)$ does not depend upon any parameters, it may be omitted.

Examination of the profile likelihood equation shows it to be constructed of a sum of terms, in which d_j^t is a multiplier which takes on the value 1 for clicks in which an event occurs, and 0 everywhere else. Thus the profile log likelihood receives an additive contribution for every failure event. Each of these is the log of a ratio whose numerator is the rate ratio, θ_j, predicted by the model for subject j in whom the event occurred (the *case*), and whose denominator,

$$\sum_i y_i^t\theta_i$$

is the sum of rate ratios, θ_i, for those subjects under observation at t, the time of occurrence of the failure.

The collection of subjects contributing to the denominator is known as the *risk set* for the observed failure. Using this terminology the profile likelihood can be written

$$\sum_{\text{Failures}} \log\left(\theta_{(\text{for case})} \Big/ \sum_{\text{Risk set}} \theta\right).$$

The ratio in brackets is the conditional probability that, given a failure occurred in this set of subjects, it occurred in the case rather than in some other member of the risk set. The profile log likelihood therefore

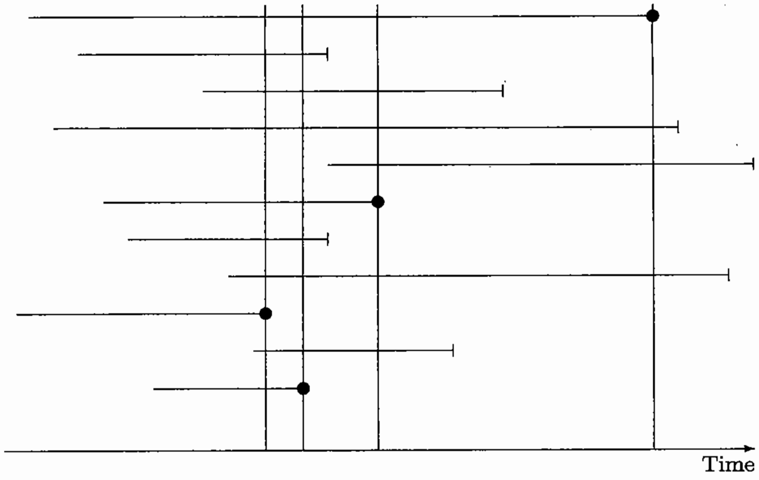


Fig. 30.1. Composition of risk sets.

corresponds exactly with the conditional log likelihood obtained for individually matched case-control studies, and analysis of a cohort study using the above profile likelihood is equivalent to its analysis as a matched case-control study in which each case is matched on time with all other members of the corresponding risk set. The composition of risk sets is illustrated by Fig. 30.1. The risk set for each failure contains all subjects whose observation lines cross the appropriate vertical, including the subject in whom the defining event occurred.

The recognition that this likelihood is a profile likelihood came some years after Cox's original proposal of the method, in which he called it the *partial likelihood.** This name has stuck, and is in general use, so we shall continue to use it, but we emphasize that partial likelihood is the profile likelihood for the parameters in the second part of the regression model when Cox's method has been used to eliminate the parameters in the first half. Because a very large number of nuisance parameters have been eliminated — infinitely many, in fact, we have no right to expect that the partial likelihood will maintain the properties of likelihood. In the present application, however, it has been proved to behave the same way as a true

*Cox originally used an argument identical to that we used in Chapter 29 for individually matched case-control studies and referred to it as a *conditional* likelihood. There are, however, difficulties with this argument when applied in the present context. While each term which contributes to the log likelihood is indeed the logarithm of a conditional probability, the total is not. A later paper correcting this error introduced the term partial likelihood.

Table 30.1. A cohort of 10 subjects

		Entry to Study		End of Study	
Subject	Sex	Date	Age	Date	Age
A	F	13/ 6/65	29.3	31/12/89	53.8
B	M	23/10/72	25.2	31/12/89	42.4
C	M	3/ 3/59	22.1	31/12/89	52.8
D	F	10/10/67	32.2	31/12/89	54.4
E	M	2/ 1/60	33.1	4/ 7/79	52.6
F	M	9/ 1/75	42.1	31/12/89	57.1
G	F	5/ 8/53	35.2	3/10/68	50.4
H	M	10/10/69	27.0	31/12/89	47.2
I	M	2/ 3/72	44.8	31/12/89	62.7
J	F	1/11/70	51.5	31/12/89	70.6

likelihood as the amount of data increases.

The composition of risk sets (and hence the results of the analysis) depend upon the choice of time scale for the analysis, as is demonstrated by the following exercise.

Exercise 30.1. The data set out in Table 30.1 refer to 10 subjects from a cohort study. Subjects *E* and *G* died at the second date while the remaining eight subjects survived until the date of analysis (31/12/89). List the members of the risk sets for both deaths when the appropriate time scale is (a) calendar date (b) age (c) time since entry into the study.

The difference between these analyses is that they represent three different models. In each case the λ_C^t parameters represent variation of baseline rates along different time scales.

30.4 Choice of time scale

Our derivation of Cox's method allows for time to be interpreted in the most appropriate manner for a particular analysis. Usually this will mean the time scale with the strongest relationship to failure rate. Regrettably it is still the case that some major software packages do not allow such flexibility. This reflects the fact that the method was motivated by problems of survival following medical treatment. In such studies the appropriate time scale is time since start of follow-up so that all observation of all subjects starts at time zero. In such studies, risk sets always become smaller (as a result of failure and censoring) as time advances.

On other time scales there will be *late entry* of subjects (observation starting at time > 0) and risk sets may be supplemented by new entrants as time advances. In order to be able to select the most appropriate time scale for an analysis, the software must be capable of allowing for late entry.

30.5 Confounders other than time

The confounding effect of time is allowed for by including time in the first part of the model. For example, taking age as the time variable, the multiplicative model

$$\text{Rate} = \boxed{\text{Corner} \times \text{Age}} \times \boxed{\text{A} \times \text{B}},$$

includes the effect of age in the baseline rate parameters. The most obvious way to deal with another confounder, such as sex, is to include it in the second part of the model, as in

$$\text{Rate} = \boxed{\text{Corner} \times \text{Age}} \times \boxed{\text{Sex} \times \text{A} \times \text{B}}.$$

This model assumes that the effect of sex is constant with age so that the baseline rates for males are a constant multiple of those for females. To extend the model to allow for different patterns of baseline rates for each sex, the interaction between age and sex must be included in the model. When the age scale is divided into clicks this interaction term involves a very large number of parameters, so it is best to absorb these parameters in the baseline rate part of the model, giving

$$\text{Rate} = \boxed{\text{Corner} \times \text{Age} \times \text{Sex} \times \text{Age·Sex}} \times \boxed{\text{A} \times \text{B}}.$$

This model has the effect of allowing different sets of baseline rate parameters for males and females. If we estimate these algebraically as before, we find that the profile likelihood for the rate ratio part of the model still has the form of a partial likelihood:

$$\sum_{\text{Failures}} \log\left(\theta_{(\text{for case})} \Big/ \sum_{\text{Risk set}} \theta\right)$$

but the risk set is now restricted to contain only those subjects who (a) were under study at the time of failure of the case, and (b) belonged to the same sex as the case. Thus the analysis simulates a matched case-control study in which controls are matched to cases with respect to sex.

This extension of Cox's method is usually referred to as a stratified analysis, although more properly it should be referred to as *doubly* stratified — Cox's method stratifies by time alone, while the extended method stratifies by both time and a further variable. In our example stratification is by age and sex.

Exercise 30.2. Repeat Exercise 30.1 for an analysis which is to be stratified by sex.

It can be seen from the last exercise that when an analysis is doubly stratified the risk sets contain fewer subjects than when it is stratified on time alone. Rather unexpectedly, therefore, the effect of adopting a more complicated model is to *reduce* the amount of computation required to estimate the parameters of interest. Further stratification can be introduced but there is a limit. If a study is overstratified, some risk sets will contain only the case, there being no other subjects matching the case in respect of all stratifying variables. Such sets make no contribution to the profile likelihood, so the information from these events is lost.

⋆ 30.6 Estimating the baseline rates

In some circumstances the dependence of rates upon time is of some interest, and we would wish to estimate the baseline rates, λ^t_C. In this section we shall show that the plot of the most likely estimate of the baseline rate against time turns out to be very similar in form to the Aalen– Nelson estimator introduced in Chapter 5.

Given the values of the parameters in the second part of the model the most likely values of the baseline rates, λ^t_C, were shown in Section 30.2 to be

$$\frac{d^t}{\sum_i y^t_i \theta_i}.$$

where θ_i is given by the second part of the model. When we divide time into clicks of duration h and redefine y^t_i to be 0 or 1 at-risk indicators, this expression becomes

$$\frac{d^t}{\sum_i h y^t_i \theta_i}.$$

In most clicks no failure occurs, $d^t = 0$, and the estimate of the rate is zero. In a click in which a failure occurs, $d^t = 1$, the estimated rate is

$$\frac{1}{h \sum_i y^t_i \theta_i},$$

which becomes very large as h becomes very small. However, the *cumulative* baseline rate increases at each click by the amounts $h\lambda^t_C$, and the estimated values of these are either zero or

$$\frac{1}{\sum_i y^t_i \theta_i}$$

when a failure occurs. Thus the cumulative baseline rate is estimated by stepped curve with jumps at the observed failure times. This is called the Aalen–Breslow estimate and is illustrated in Fig. 30.2. The height of the jump at each failure time is now given by

$$1 \Big/ \sum_{\text{Risk set}} \theta$$

rather than by

$$1\,/(\text{Number of subjects at risk})$$

as in the simpler case discussed in Chapter 5. As noted there, examination of the cumulative rate plot allows us to assess the dependence of failure rate on time.

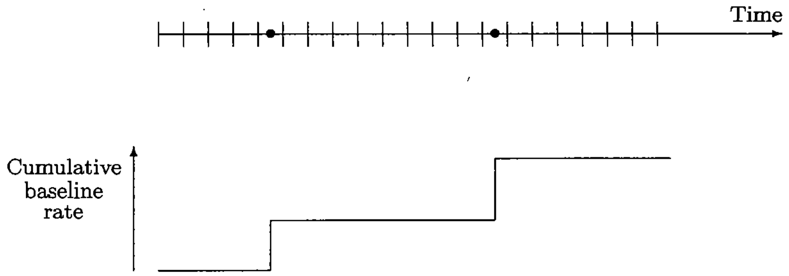


Fig. 30.2. The Aalen–Breslow estimate of the cumulative baseline rate.

Solutions to the exercises

30.1 When date is the time scale, membership of risk sets is determined by whether or not the subject was observed at the date of occurrence of the death. The risk sets corresponding to the two deaths are as follows:

Date of death	Subjects in risk set
3/10/68	*A, C, D, E, G (case)*
4/ 7/79	*A, B, C, D, E (case), F, H, I, J*

The risk set corresponding to the death of subject *G* contains fewer individuals since it occurred at a date earlier than some subjects had joined the cohort.

When age is the time scale, risk set membership is determined by whether the subject was observed at the age at which the death occurred. The risk sets are now as follows:

Age at death	Subjects in risk set
50.4	*A, C, D, E, F, G (case),I*
52.6	*A, C, D, E (case), F, I, J*

When time in study is the scale, the risk sets are as follows:

Time in study at death	Subjects in risk set
15.2 yrs	*A, B, C, D, E, G (case), H, I, J*
19.5 yrs	*A, C, D, E (case), H*

30.2 Since subject G is female and subject E is male, the risk set for the failure of G contains only female subjects and risk sets for the failure of E contains only males. When date is the time scale, the risk sets corresponding to the two deaths are as follows:

Date of death	Subjects in risk set
3/10/68	*A, D, G (case)*
4/ 7/79	*B, C, E (case), H, I*

When age is the time scale, the risk sets are

Age at death	Subjects in risk set
50.4	*A, D, G (case)*
52.6	*C, E (case), F, I*

When time in study is the scale, the risk sets are:

Time in study at death	Subjects in risk set
15.2 yrs	*A, D, G (case), J*
19.5 yrs	*C, E (case), H*

31
Time-varying explanatory variables

Cox's method provides a convenient way of controlling for time in the analysis of follow-up studies. In its simple form the method assumes that other explanatory variables do not change with time. In this chapter we show how the method can be extended to allow for this. We also discuss the closely related problem of analysis strategies when rates vary in relation to more than one time scale, and draw attention to some dangers and difficulties.

31.1 The model and the likelihood

We have seen that Cox's method amounts to dividing the multiplicative model for rates into two parts:

$$\text{Rate} = \boxed{\text{Corner} \times \text{Time}} \times \boxed{\text{A} \times \text{B} \times \cdots}\,.$$

The first part refers to the baseline rates while the second part specifies how the rate ratio

$$\theta_i = \frac{\text{Rate for subject } i \text{ at time } t}{\text{Baseline rate at time } t}$$

is related to the explanatory variables A, B, etc.. On a log scale

$$\log(\text{Rate}) = \boxed{\text{Corner} + \text{Time}} + \boxed{\text{A} + \text{B} + \cdots}\,.$$

In the simple form of the method θ_i is assumed to be independent of time.

The extension of Cox's method with which we are now concerned allows the relationship between θ_i and the explanatory variables to vary with time. This would be necessary, for example, when studying levels of hazardous industrial exposures in occupational studies and when studying changing treatments in long term follow-up studies of chronic disease aetiology. Indeed *most* explanatory variables of interest to epidemiologists vary with time if follow-up is over a sufficiently long period.

Allowing the rate ratio part of the model to change over time involves only a simple change to the contribution

$$\log\left(\theta_{(\text{for case})} \Big/ \sum_{\text{Risk set}} \theta\right),$$

from each risk set to the partial log likelihood. Since the model now predicts different values of θ at different times the contribution of each risk set must now be calculated using the values of θ current at the time of occurrence of the failure.

COMPUTATION

When it comes to computing the likelihood and finding the values of parameters which maximize it this simple change turns out to have major consequences, and computation times can increase by several orders of magnitude. To understand why the computation is so heavy it helps to look at the simpler version of Cox's method to see why this does *not* involve heavy computations. There are two reasons. First, for any particular set of values for the parameters, the value of θ only needs to be worked out once for each subject. Second, the value of $\sum\theta$ does not have to be calculated from scratch for each risk set because the equivalent term from the previous risk set can be updated by subtracting the values of θ for all subjects lost to follow-up in the intervening period and adding the contributions of those newly joining the cohort. Other terms needed in the computation of gradient and curvature of the log likelihood can be updated in a similar way.

When the model allows the rate ratios θ to change over time a subject who appears in several risk sets can have different values of θ in each. This means that not only must the values of θ be re-calculated for each risk set but $\sum\theta$ and other gradient and curvature terms must be calculated from scratch. The result is that the computing time rises dramatically.

Some reduction in computing time can be achieved by sampling the risk sets. The algebraic equivalence of the partial likelihood in Cox's method and the conditional likelihood for matched case-control studies means that analyzing a cohort study using Cox's method is the same as analyzing it as a case-control study in which each incident case is individually matched with a control set in which the controls are all other subjects under study at the moment of incidence. Since a case-control study which draws many controls for each case provides very little more information than one which draws only a few, we shall lose little by taking a random sample of controls drawn from each risk set rather than using the entire risk set. Sampling risk sets in this way creates what is called a *nested case-control study.* Such studies offer a number of practical advantages in addition to considerable computational savings and will be discussed further in Chapter 33.

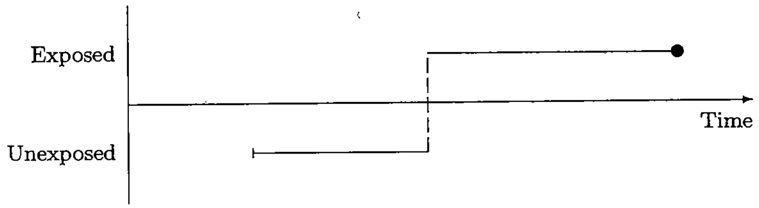


Fig. 31.1. Changing exposure group.

31.2 Changing exposure group

One simple but important way in which an explanatory variable can change with time arises when a subject can change from being unexposed to being exposed group (or vice versa) during the course of follow-up (see Fig. 31.1). This is most easily dealt with by splitting the follow-up for such subjects into an unexposed part and an exposed part, and treating the parts as distinct subjects. The data can then be analysed using the simple form of Cox's method in which the explanatory variables do not change with time. The validity of the analysis depends on a relatively strong assumption concerning the *reasons* for the change of exposure group, namely that transfer is unrelated to the subsequent probability of failure. If the transfer mechanism operates in a way that selects particularly high or low risk subjects then subsequent comparisons will be distorted. This is another example of selection bias. More formally, it is required that transfer must be independent of subsequent failure conditional upon the values of all other variables in the model. If transfer and failure are both strongly related to age (say) there will be an overall association between transfer time and outcome, but this will not bias estimates of other effects providing there is no relationship between transfer time and outcome *for subjects of the same age*, and providing the model takes proper account of the relationship between age and failure rate. Similar considerations apply when there are more than two categories of exposure or when the level of exposure varies continuously.

Exercise 31.1. Subjects enter a heart transplant programme as unexposed on joining a waiting list for a transplant, and switch to the exposed group on receiving the transplant. Do you think the assumptions discussed above are likely to be met in this case?

31.3 Time scales as explanatory variables

Another very common form of time-dependent explanatory variable is an additional time scale. For example, in a clinical study in which survival

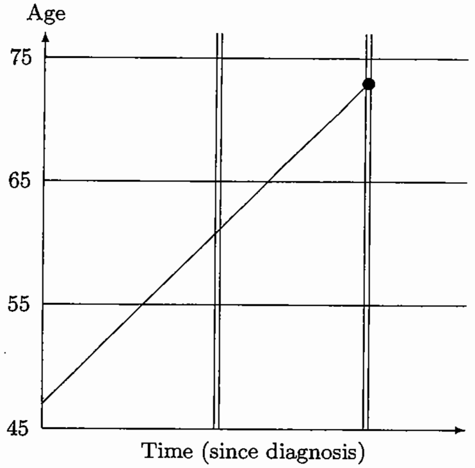


Fig. 31.2. Follow-up by age and time.

is analysed largely in relation to time since diagnosis, it will usually be necessary to control the comparison of different treatments for the age of the subjects receiving them. For short studies this can be achieved by including age at diagnosis, which is fixed for every subject. When follow-up is over many years it is better to include age itself, which varies with time. Fig. 31.2 illustrates follow-up of a subject in which observation time is classified by time since diagnosis and age. The risk sets are determined by the times of occurrence of failures. Two such times are illustrated in the figure by narrow vertical bands. One corresponds to the risk set for the failure of the subject shown while the other is an earlier failure. The subject shown contributes to both risk sets, but is of a different age on the two occasions.

One possible analysis would be to include time since diagnosis in the first part of the model, so that this is the time scale which is used to determine the risk sets, and to include age as a time varying explanatory variable in the second part of the model. This could be done either by dividing the age scale into 5- or 10-year bands and treating it as a categorical variable, as in

$$\log(\text{Rate}) = \boxed{\text{Corner} + \text{Time}} + \boxed{\text{Age} + \text{A} + \text{B} + \cdots},$$

or by treating age as a quantitative and fitting linear effects, and possibly quadratic effects too, as in

$$\log(\text{Rate}) = \boxed{\text{Corner} + \text{Time}} + \boxed{[\text{Age}] + [\text{Age-sq}] + \text{A} + \text{B} + \cdots}.$$

When the partial log likelihood is formed for either of these analyses each risk set contributes a term of the form $\log(\theta/\sum\theta)$ where the values of θ for the subjects in the risk set are determined by the relationship between $\log(\theta)$ and the parameters in the second part of the model. As an example of this computational process consider the model

$$\log(\text{Rate}) = \boxed{\text{Corner} + \text{Time}} + \boxed{\text{Age} + \text{A} + \text{B}}$$

where age has five levels, A has two levels and B has three levels. The parameters in the second part of the model are then Age(1), $\cdots$, Age(4), A(1), B(1) and B(2). Now consider a subject, at level 1 for A and level 2 for B, who appears as a survivor in the risk sets at two failure times, and suppose that this subject is in age band 3 at the time of the first failure, and in age band 4 at the time of the second failure.

Exercise 31.2. Write down an expression, in terms of the parameters, for the values of $\log(\theta)$ for this subject, in the two risk sets.

When there are two time scales a natural question to be considered is which should be included in the baseline rates part of the model and which should be included in the rate ratio part. The choice depends on the way that rates vary along each time scale. If this variation is to be modelled in the rate ratio part of the model then we must either divide the scale into broad bands or fit simple mathematical functions of time, such as linear or quadratic. The former strategy is adequate if the variation of rates is not too rapid, while the latter is only possible if the variation is regular enough to describe by simple mathematical functions. If variation is both rapid and irregular neither approach works very well and the variation should be modelled in the baseline rates. Thus if it is suspected that variation along one scale will be rapid and irregular this should be the scale whose effects are modelled by the baseline rates, and other scales should be treated as time varying explanatory variables. If variation is smooth along all scales it is better to use the scale with the strongest effects for the baseline rates.

Exercise 31.3. Discuss appropriate strategies for modelling the effects of age and calendar time on incidence of (a) a chronic degenerative disease, and (b) an infectious disease.

31.4 Dependencies between time scales $\boxed{\star}$

Different time scales are not truly different variables but the same variable measured from different origins. It is therefore impossible for a subject to advance one year on one scale without simultaneously advancing one year on other time scales. For example, we cannot pass through a year of calendar time without advancing a year in age — would that we could! This dependency between time scales can lead to difficulties when trying to interpret the estimated effects of changes on these time scales.

As an illustration we shall return to the example of age and time since diagnosis in a clinical follow-up study. Let us first consider the model

$$\log(\text{Rate}) = \boxed{\text{Corner} + \text{Time}} + \boxed{[\text{Age-at-diagnosis}] + \cdots},$$

in which the effect of time since diagnosis is the main time scale and is included in the first part of the model, while age at diagnosis is included as a linear effect in the second. The parameter [Age-at diagnosis] measures the change in the log rate per one year change in age, holding time since diagnosis constant at any arbitrary value. Fig. 31.3 shows two subjects who are diagnosed at ages 47 and 61 respectively; if we assume these subjects have the same values for any other explanatory variables the difference in log rate predicted by the model, at diagnosis, or at any value of time since diagnosis, is

$$(61 - 47) \times [\text{Age-at-diagnosis}] = 14 \times [\text{Age-at-diagnosis}].$$

Now consider the model

$$\log(\text{Rate}) = \boxed{\text{Corner} + \text{Time}} + \boxed{[\text{Age}] + \cdots}$$

in which age varies with time. The two subjects in Fig. 31.3 have a 14 year age difference at diagnosis, so this model predicts a difference in log rates between the two subjects of 14 × [Age] at diagnosis. Because these two subjects have a 14 year age difference not only at diagnosis but at any time after diagnosis, the model also predicts a difference of 14 × [Age] at any value of time since diagnosis. Thus both models predict a constant difference in log rate at any value of time since diagnosis. In the one case the prediction is 14 × [Age-at-diagnosis], in the other the prediction is 14 × [Age]. This is true for any pair of subjects; the models make identical predictions and cannot be differentiated, the [Age-at-diagnosis] parameter in the first model is making the same comparison as the [Age] parameter in the second.

There may well be scientific interest in discriminating between models in which the age at diagnosis determines prognosis, and models in which age itself is the determinant, but if we were to fit the model

$$\log(\text{Rate}) = \boxed{\text{Corner} + \text{Time}} + \boxed{[\text{Age}] + [\text{Age-at-diagnosis}] + \cdots},$$

in order to try and separate the linear effect of age controlled for time since diagnosis from the linear effect of age at diagnosis controlled for time since diagnosis, we would run into difficulties. When time since diagnosis and age are held constant, there can be no further variation in age at diagnosis so that the [Age-at-diagnosis] parameter cannot be estimated. Likewise, time since diagnosis and age at diagnosis uniquely determine age so that the [Age] parameter cannot be estimated. Again the two subjects shown in Fig. 31.3 demonstrate the problem. The new model also predicts that the difference in log rates remains constant at any value of time since diagnosis but this difference is now equal to

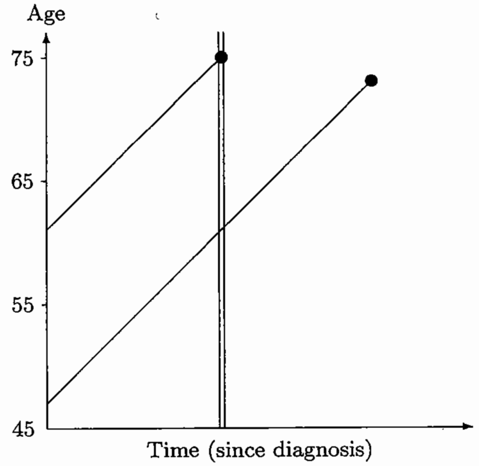


Fig. 31.3. Observation of two subjects.

$$14 \times [\text{Age}] + 14 \times [\text{Age-at-diagnosis}] = 14 \times ([\text{Age}] + [\text{Age-at-diagnosis}]),$$

where the parameters [Age] and [Age-at-diagnosis] now refer to the new model which contains both linear effects. Because any values for the two parameters which have the same sum, make the same predictions, the parameters cannot be estimated individually. They are said to be *non-identifiable* or *aliased.*

A computer program will usually warn the user when two parameters are non-identifiable and then omit one of them from the model. This is quite useful when the object is to control for age and age at diagnosis, but if the object is to disentangle their effects, what the computer program is saying is that we are attempting the impossible.

The non-identifiability of parameters for different time scales refers to their linear effects. When we come to fit models with non-linear terms, things are not so bad. Consider for example the predictions of the model

$$\log(\text{Rate}) = \boxed{\text{Corner} + \text{Time}} + \boxed{[\text{Age}] + [\text{Age-sq}] + \cdots}$$

for the two subjects shown in Fig. 31.3. At the time of diagnosis the model predicts a difference in log rates of

$$(61 - 47) \times [\text{Age}] + (61^2 - 47^2) \times [\text{Age-sq}] = 14 \times [\text{Age}] + 1512 \times [\text{Age-sq}].$$

However, 5 years after diagnosis, their ages are 66 and 52 and the model predicts a difference in log rates of

$$(66 - 52) \times [\text{Age}] + (66^2 - 52^2) \times [\text{Age-sq}] = 14 \times [\text{Age}] + 1652 \times [\text{Age-sq}].$$

In the model with non-linear effects, therefore, the difference between log rates for the two subjects does vary with time since diagnosis. The model

$$\log(\text{Rate}) = \boxed{\text{Corner} + \text{Time}} + \boxed{[\text{Age-at-diagnosis}] + [\text{Age-at-diagnosis-sq}] + \cdots}$$

predicts a difference in log rates of

$$(61 - 47) \times [\text{Age-at-diagnosis}] + (61^2 - 47^2) \times [\text{Age-at-diagnosis-sq}]$$

throughout the follow-up, and this is a different prediction than the one obtained from the model with age and age-squared. The linear parts of the two predictions are still the same and cannot be separately estimated, but the non-linear parts are different and can be.

Similarly, if we were to fit the model

$$\log(\text{Rate}) = \boxed{\text{Corner} + \text{Time}} + \boxed{\begin{array}{l}[\text{Age}] + [\text{Age-sq}] + [\text{Age-at-diagnosis}] + \\ [\text{Age-at-diagnosis-sq}] + \cdots\end{array}},$$

the parameters [Age] and [Age-at-diagnosis] are not identifiable while the parameters [Age-sq] and [Age-at-diagnosis-sq] can be estimated. The same is true for any other non-linear component of the relationships.

★ 31.5 Discrete time bands

In the above discussion the time variables are measured exactly; when the time scales are divided into discrete bands the position is slightly more complicated. To illustrate this we shall return to the two subjects of Fig. 31.3 and imagine a model in which age has been grouped into 5-year bands but time since diagnosis is still measured exactly. At the beginning of follow-up one subject is in the 45–49 band and the other is in the 60–64 band. However, after three years the former subject has moved into the 50–54 band while the latter remains in the 60–64 band. It will appear to a computer program that the age difference between the subjects has narrowed! As a result the program will not spot the underlying non-identifiability of models such as

$$\log(\text{Rate}) = \boxed{\text{Corner} + \text{Time}} + \boxed{\text{Age} + \text{Age-diag} + \cdots}$$

and fit them without complaint. However, the linear components of the relationships with age and age at diagnosis have only become estimable because of the inaccuracy introduced by banding and the resulting parameter estimates are uninterpretable.

31.6 Modelling vital rates ★

A familiar example of these problems arises in 'age-period-cohort' modelling of mortality and other vital rates, where the aim is to disentangle the dependence of rates upon age, calendar time (period), and date of birth (birth cohort). This comparison raises exactly the same problem as above and has provoked a lot of discussion in the epidemiological literature. Much of this has been based on the misconception that the problem is a shortcoming of current statistical methods and that its solution awaits only methodological advances. This is not the case. The difficulty is inescapable and arises from the fact that subjects cannot move in one time scale without an identical move in others.

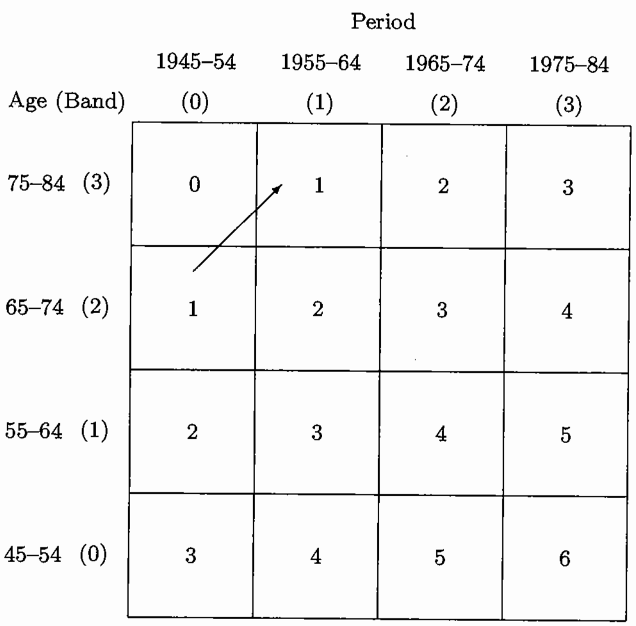
Fig. 31.4 shows a table in which both both age and calendar period have been divided into 10-year bands. Tables of rates, classified in this way, are frequently available from official published sources, and allow effects of year of birth (*birth cohort* effects) to be estimated approximately. If we remember that observation of individual subjects is represented by diagonal lines in the age and calendar time Lexis diagram (illustrated by the arrow), it is clear that diagonal groupings of cells in the table correspond *approximately* to birth cohorts. The cell labelled 0 refers to subjects born around 1870, those labelled 1 to subjects born around 1880, and so on. Although this correspondence is only approximate, the new discrete codings for age period and cohort behave very much like the underlying continuous scales. In particular, they are linearly dependent. In our example,

$$\text{Cohort} = 3 + \text{Period} - \text{Age}.$$

This means that when two are fixed the third is also fixed and in models such as

$$\log(\text{Rate}) = \text{Corner} + [\text{Age}] + [\text{Period}] + [\text{Cohort}]$$

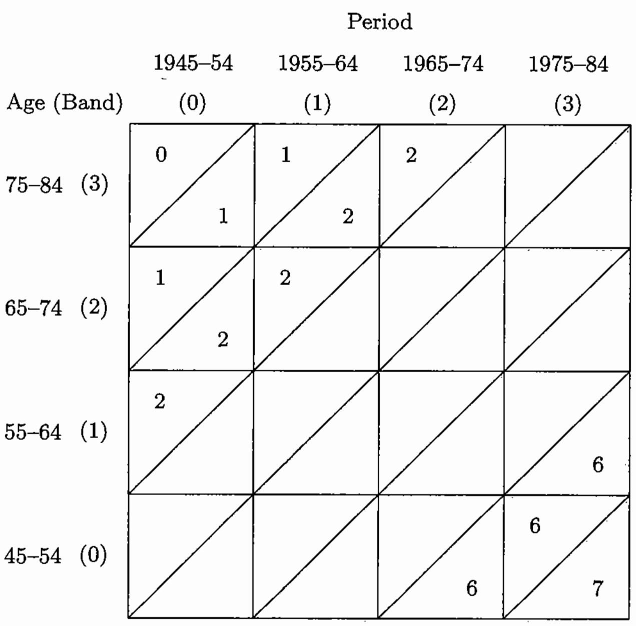
the parameters are unidentifiable, and it is impossible to disentangle the linear effects of all three variables.



	Period			
	1945–54	1955–64	1965–74	1975–84
Age (Band)	(0)	(1)	(2)	(3)
75–84 (3)	0	1	2	3
65–74 (2)	1	2	3	4
55–64 (1)	2	3	4	5
45–54 (0)	3	4	5	6

Fig. 31.4. Approximate birth cohorts.

Some investigators have returned to the raw data in order to allocate subjects to their true birth cohort. This avoids the approximation in Fig. 31.4 but leads to a serious fallacy. Fig. 31.5 shows how the exact birth cohorts move across the Lexis diagram. The cell labelled 0 refers to the 1860–69 birth cohort, those labelled 1 to the 1870–79 cohort, and so on. The discrete codings no longer behave like the underlying scales. For example, birth cohort 1 is observed in 3 cells; the transition from the first to the second involves a change of age band (from 65–74 to 75–84) without change in calendar period, while the transition from second to third corresponds to a move through calendar time without change in age! Looked at naively it would appear that, by grouping, we have created a natural experiment in which subjects can age instantaneously and travel in time without ageing. The fallacy lies in the fact that the regions are triangular and that regions shaped ◸ disproportionately represent ages towards the upper end of the 10-year band and dates towards the lower end of the period, while regions shaped ◿ disproportionately represent ages at the lower end of the band and periods at the upper end. Unfortunately, computer programs have no way of knowing this. They will believe that a miraculous natural experiment has been observed, and estimate separate linear effects for all three scales without complaint. This uncritical behaviour of computer programs (which can't know better) has been hailed by some epidemiologists and statisticians (who should) as a 'solution' to the identifiability 'problem'. The reverse is the case; the computer solution is fallacious, being based entirely on grouping inaccuracies, and the resultant estimates are uninterpretable. It is worth pointing out that this pitfall is not confined to the age-period-cohort problem, but can be encountered whenever more than one time scale is involved in an analysis.



	Period			
	1945–54	1955–64	1965–74	1975–84
Age (Band)	(0)	(1)	(2)	(3)
75–84 (3)	0 1	1 2	2	
65–74 (2)	1 2	2		
55–64 (1)	2			6
45–54 (0)			6	6 7

Fig. 31.5. Exact birth cohorts.

Solutions to the exercises

31.1 When a heart becomes available for transplantation and there is more than one patient eligible to receive it, there is potential selection bias. A controlled study would *randomize* such choices to exclude selection bias, but in an observational study it will always be difficult to know whether the recipient was selected because the clinician felt that this patient was most likely to benefit. Such selection would cause serious bias in a simple analysis. In theory this can be offset by including in the analysis any prognostic factors likely to have been used by the clinician making the decision, but in practice one can rarely be sure that all relevant factors have been taken into account. We shall discuss this example in more detail in Chapter 32.

31.2 For the first risk set

$$\log(\theta) = \text{Age}(3) + \text{A}(1) + \text{B}(2).$$

For the second risk set

$$\log(\theta) = \text{Age}(4) + \text{A}(1) + \text{B}(2).$$

31.3 Incidence rates of chronic degenerative diseases such as ischaemic heart disease and most cancers rise steeply with age. In such diseases age may usually be thought of as a surrogate for the cumulative damage inflicted by a large number of influences throughout life. Such cumulative damage will be reflected in a *smooth* increase of rates with age so that simple linear or quadratic models for the age effect are usually satisfactory. Grouping age by 5 or 10 year bands will also work quite well. Age relationships for incidence of infectious diseases are usually more complicated. Increasing immunity with age will produce a smoothly decreasing curve, but where transmission of the infectious agent depends upon various social influences such as schooling, employment, sexual activity etc., these may give rise to rather irregular age curves. Simple mathematical functions for age-incidence curves are therefore less likely to be useful. Grouping may also be difficult because of abrupt changes in incidence due to age related changes in social behaviour.

32
Three examples ★

This chapter describes three studies where the explanatory variables change with time and where the analysis has been helped by the statistical methods discussed in immediately preceding chapters. The first is a clinical follow-up study of heart transplant patients and has already been introduced in Exercise 31.1. The second is an epidemiological study into the effects of bereavement in old people. The third is concerned with the important problem of estimating the parameters of cancer screening programmes to help public health administrators in planning such services.

32.1 Mortality following heart transplantation

The first example concerns the survival of patients in the Stanford heart transplant program.[*] The basic nature of the data is illustrated in Fig. 32.1. The follow-up of patients starts as soon as they are enrolled in the program to await a suitable heart. In this phase of the follow-up, patients are in the *pre-transplant* state. When a heart becomes available, and if selected, transplantation takes place and the patient transfers into the *post-transplant* state. The diagram shows two patients, one of whom dies some time after transplantation while the other dies while awaiting a suitable heart.

The diagram also indicates (by the two vertical lines) a stratification by time in programme. In this time band there is some person-time pre-transplant and some post-transplant. This allows comparison of mortality in post-transplant patients with that in controls who are still awaiting transplantation. The possible biases in this comparison were the subject of Exercise 31.1. Here we are more concerned with the mechanics of the analysis. In this comparison it would be necessary to control for such variables as age (either itself, or at enrollment into the programme), date when enrolled, date when transplanted, and prognostic factors such as record of previous surgery. Multiplicative models fitted using Cox's method can be used to do this.

[*] Crowley, J. and Hu, M., *Journal of the American Statistical Association*, **72**, 27–36.

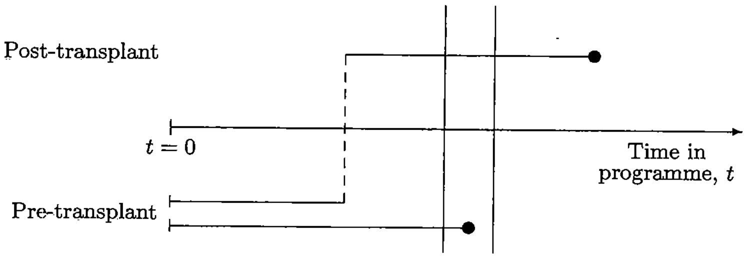


Fig. 32.1. Mortality following heart transplant.

These models are based on the assumption that

$$\frac{\text{Mortality rate for transplanted patient}}{\text{Mortality rate for untransplanted patient}} = \text{Constant},$$

that is, the rate ratio does not vary either with time since entry into the program or with time since transplantation. The latter seems very unlikely. We might even expect an initial adverse effect of transplantation (rate ratio greater than 1) which would later be replaced by a beneficial effect (rate ratio less than 1). The assumption can be relaxed by allowing the transplantation effect to vary with time since transplantation — a variable whose evolution over time can be demonstrated by adding a further axis to the follow-up diagram, as in Fig. 32.2.

Exercise 32.1. Time since transplant can be included in the model for the rate ratio in a number of ways. Perhaps the simplest is to include time since transplant as a quantitative variable as in

$$\log(\text{Rate}) = \text{Corner} + \text{Time} + \text{Transplant} + \text{Transplant} \cdot [\text{Time-since-transplant}]\,,$$

where time is time in program. What signs would you expect for the two parameters of this model? Sketch the graph showing how the rate ratio would vary with time since transplant in this model. (You should assume that Time-since-transplant is coded zero until transplantation occurs.)

Other potential effect modifiers are age at transplantation, time spent awaiting transplantation, and closeness of matching of tissue type with the donor.

32.2 Bereavement in the elderly

The second example is drawn from a study of the effect of bereavement (death of spouse) in an elderly population.[†] There is some empirical evidence that, for a period following the death of a spouse, the mortality rate of the surviving partner is elevated. Fig. 32.3 shows a plausible relationship between mortality rate, expressed relative to mortality in persons with surviving partners, and time since death of spouse. Such a relationship can be modelled by a simple function such as

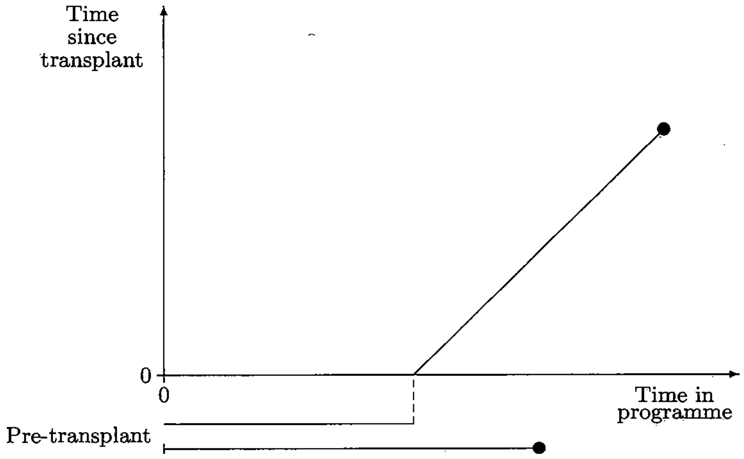


Fig. 32.2. Incorporating time since transplantation.

$$\text{Rate ratio} = \alpha + \beta \exp(-\gamma t),$$

where α, β, and γ are parameters. At $t = 0$ the rate ratio is $\alpha + \beta$ and, with the passage of time since bereavement, it falls away to α. The parameter γ controls how soon the rate ratio dies away.

Fig. 32.4 shows follow-up of four subjects in a cohort study by calendar time and by time since loss of spouse. Before bereavement, subjects are followed through time, thus allowing measurement of baseline mortality rates. Following death of a spouse, observation may be represented by diagonals in the Lexis diagram formed by plotting calendar time against time since bereavement. Our diagram shows the pattern of observation of two couples. For the sake of clarity, the diagram has been simplified by omitting age, although this must be included in the analysis. In a fuller representation, observation of subjects with living spouses would be represented by lines in an age by calendar time Lexis diagram, while bereaved subjects would be represented by lines in a three-dimensional diagram formed by age, calendar time and time since bereavement.

The analysis of this study must relate mortality rates to all three time

[†] Jagger, C. and Sutton, C.J., *Statistics in Medicine*, **10**, 395–404.

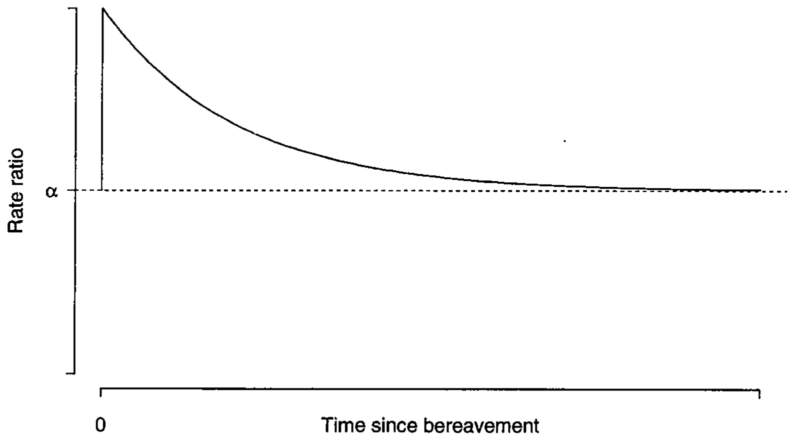


Fig. 32.3. Mortality following bereavement.

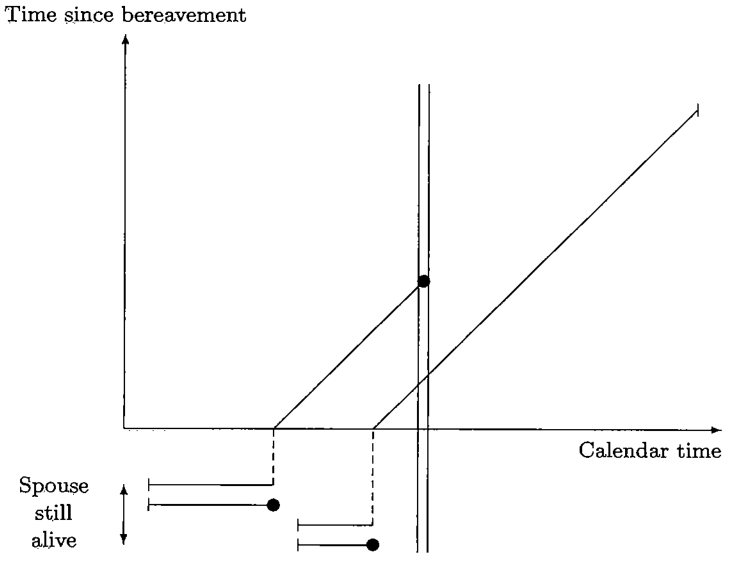


Fig. 32.4. A study of mortality following bereavement.

scales. The effect of time since bereavement is modelled by

$$\text{Rate ratio} = \alpha + \beta \exp(-\gamma t),$$

which describes the relationship using three parameters. For modelling the effects of age and calendar time, all three possibilities discussed in Chapter 31 are open to us. A frequent recommendation is that the scale used in the construction of risk sets should be that with the strongest relationship with event occurrence, and this would argue for age being dealt with in this way. However, mortality in the elderly also varies quite markedly with calendar time, owing to climatic fluctuations, influenza epidemics, and so on. While the age relationship is a smoothly increasing function and may easily be modelled by a linear or quadratic function, the relationship with calendar time is very irregular. It follows that a better strategy is to take calendar time as the scale for definition of risk sets, and to include age in the model as a time-dependent continuous quantitative variable.

Fig. 32.4 illustrates the construction of the risk set in calendar time. The risk set corresponding to each death consists of all those subjects under study in the time slice containing it — illustrated by the vertical band in the diagram. Two of our four subjects belong to the indicated risk set — one as the case. At the relevant date, both have been bereaved and the model would assign them different values of θ (> 1.0) according to the time since their bereavement.

The analysis could also be carried out by creating a nested case- control study by sampling risk sets. This possibility also suggests the design of a *true* case-control study.

Exercise 32.2. Describe a case-control study into mortality following bereavement which mirrors the analysis described above. What sources of bias can you foresee?

32.3 Estimating the parameters of a screening test

Our final example concerns the estimation of the parameters of a cancer screening programme.[‡] The aim of such programmes is to detect cancer during the *preclinical detectable phase* (PCDP) — the period, prior to the time at which the disease would have been detected symptomatically, during which there is some possibility of detecting the disease by screening. Two parameters which it is important to know are the *sojourn time* (the name given to the duration of the PCDP) and the *sensitivity*, defined as the probability of detecting disease by screening during the PCDP. We shall denote these parameters by τ and π respectively, so that π is the probability that screening would detect the disease if applied within a period of

[‡] Day, N.E. and Walter, S.D., *Biometrics*, **40**, 1–14.

duration τ before the time at which the disease would have been discovered anyway.

Interpretation of these parameters and comparisons between different population groups and screening tests requires some care. In general, a better test will lead to increases in both π and τ. More rapid development of tumours will be reflected in decreased values for τ, since the disease will move through the PCDP more quickly. Finally, τ will also be affected by factors which determine rapidity of diagnosis in the absence of screening, so that populations with better access to medical services will usually have smaller values for τ.

We shall now show how these parameters may be estimated from studies of *interval tumours* — incident cases detected by normal clinical means in the intervals between screening appointments. Let us consider the expected variation of incidence following a negative screening test under our simple model, assuming first that the test is 100% sensitive (i.e. $\pi = 1.0$). In this case, there would be zero incidence of interval tumours for a period of length τ following the negative screen, since all the tumours which would have arisen in this period will have been detected at screening. Conversely, after a time τ has elapsed since screening, the rate of diagnosis of interval tumours will return to the normal incidence rate in an unscreened population, since no tumour detected in this period could possibly have been found at the screening appointment. Thus, the rate ratio

$$\frac{\text{Incidence rate of interval tumours following negative screening test}}{\text{Incidence rate in the unscreened population}}$$

will be 0 until time τ following screening, and then jump to 1. Making allowance for less than 100% sensitivity leads to the relationship shown in Fig. 32.5; the proportion of the normal incidence seen in the period after screening is contributed by those cases missed by the screening test.

This model is clearly oversimplified, and we would not expect to observe anything so clearly defined in practice. A more realistic model may be obtained either by allowing for sojourn times to vary or, alternatively, allowing the sensitivity of the test to vary smoothly throughout the PCDP from zero up to π. These models are indistinguishable and lead to a predicted incidence pattern such as is shown in Fig. 32.6. The curve shown is a simple exponential function of time elapsed since negative screen,

$$\text{Rate ratio} = 1 - \pi \exp\left(-\frac{\text{Time since screen}}{\tau}\right).$$

The parameters of this curve, π and τ, may be thought of as the sensitivity and mean sojourn time respectively.

Fig. 32.7 illustrates observation of four subjects in a follow-up study. Three of these enter the study prior to having been screened but are screened during follow-up, while the fourth enters the study some time after a negative screening test. Two of the subjects subsequently develop interval tumours. In an analysis with calendar time as the major time scale,

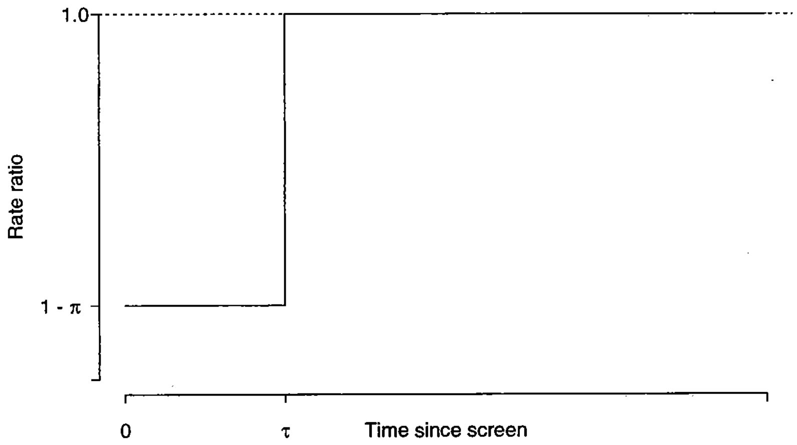


Fig. 32.5. Incidence following a negative screen.

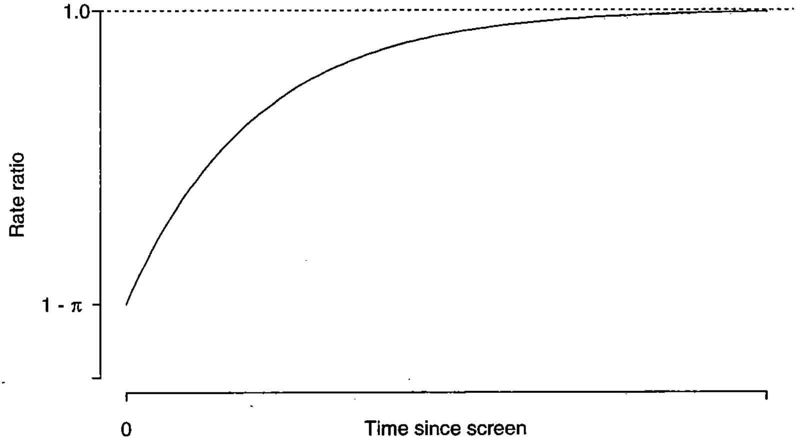


Fig. 32.6. A more realistic evolution of incidence.

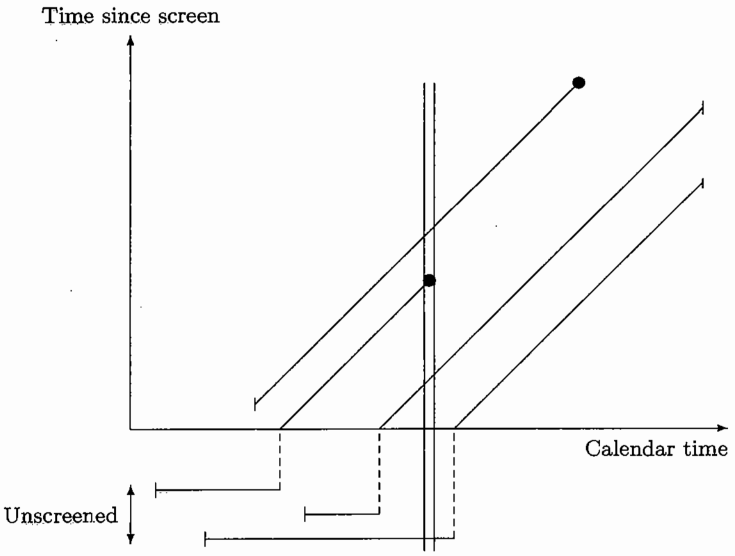


Fig. 32.7. A follow-up study of incidence following a negative screen.

these cases will be compared with risk sets comprising all individuals under study at the date of diagnosis. In the diagram this is illustrated for the first case by the vertical band. It can be seen that all four of the indicated subjects fall into this risk set; one is still unscreened and is assigned $\theta = 1$ by the model, while the other three have different times since their negative screening test and, for any values of τ and π, a model such as that illustrated by Fig. 32.6 assigns three different values of θ to the others. Each interval tumour contributes similarly to the log likelihood, and computer programs may be used to maximize this with respect to τ and π to obtain best estimates of these quantities. Approximate confidence intervals may be found in the usual way from the curvature of the profile log- likelihoods.

Exercise 32.3. What assumption concerning selection of subjects for screening must hold for this analysis to yield unbiased results?

The above discussion slightly over-simplifies the analysis. In particular, it will be necessary to allow for age in the model. As in our previous example, sampling risk sets to create a nested case-control study will avoid some computation, and also suggests a true case-control design.

Exercise 32.4. Describe a case-control study to investigate sensitivity and sojourn time of a screening test for breast cancer. Would you expect to obtain approximately the same results as in a cohort study?

Solutions to the exercises

32.1 The Transplant main effect measures the log rate ratio immediately following transplantation. We might expect this to be positive immediately after surgery, corresponding to an elevated mortality rate, but then to decrease with time, giving way eventually to a beneficial effect. In this case the interaction parameter would be negative.
The predictions of the model in terms of the log rate ratio are shown in Fig. 32.8. The parameter α is the Transplant initial effect and is shown here as positive, indicating an adverse effect. The slope of the line is the Transplant·Time interaction parameter and is shown as negative. This model predicts that transplantation will have an increasingly beneficial effect with increased time from transplantation. The horizontal dotted line represents the level of mortality in untransplanted controls. On the original scale, the rate ratio initially jumps to $\exp(\alpha)$ immediately after transplant but then falls exponentially towards zero.

32.2 The events of interest are deaths in elderly people, let us say those over 70 years of age. A geographically based case-control study would include as cases all such deaths amongst residents of a town or county. Each time such a death occurs, a set of controls would be drawn from the study base. Matching of controls to cases for age and sex would improve the efficiency of the study. Information concerning vital status of spouse and, where appropriate, date of death of spouse, would be obtained retrospectively for all cases and controls. This study would run little risk of information bias, since the relevant data are on public record. However, selection bias could be a problem. These are some of the problems:

- A suitable, accurate, sampling frame may not be available.
- Refusal to participate by potential controls could lead to 'volunteer' bias in the control group finally obtained.
- Migration away from the sampling frame as a result of bereavement is a very real possibility. A bereaved old person may not be able to care for him or herself and might be forced to go into residential care or to live with relatives.

These problems do not exist when a cohort of identified subjects is followed prospectively.

32.3 It must be assumed that individuals selected for screening would have the same subsequent incidence rates as those not selected. This assumption would not be violated by a screening policy which varies with age, providing confounding by age is dealt with in the analysis. However, if patients are referred to screening as a result of early non-specific symptoms, there would be some bias.

32.4 A population based screening programme requires a computer register to generate screening invitations, so this register can form the study base. The study would be of newly diagnosed cases who were not diagnosed as a result of routine screening and whose names could be found on the computer register. Controls for each case would then be drawn from this register. If carried out carefully, it is difficult to see any reason why such a study should give different answers from a cohort study. Indeed, the existence of the computer register means that the study is really nested within a cohort study (see Chapter 33).

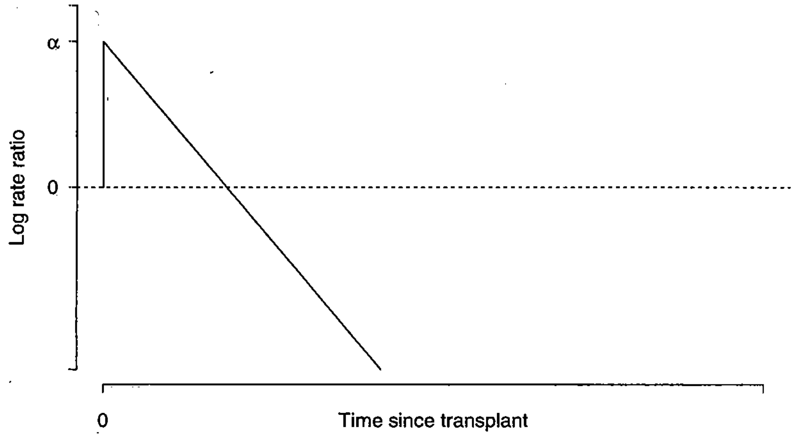


Fig. 32.8. Log rate ratio against time since transplant.

33
Nested case-control studies

Any cohort study can be used to generate a case-control study by sampling the cohort for controls to use in place of the full cohort. The case-control study is then said to be *nested* in the cohort study. For each case the controls are chosen from those members of the cohort who are at risk at that moment, in other words from the risk set defined by the case. Although the idea of nested case-control studies predates Cox's method for the analysis of cohort studies, the design and analysis of such studies has been greatly clarified by the ideas of partial likelihood and risk sets.

33.1 Reasons for using a nested case-control study

The main reason for using a nested study is to reduce the labour and cost of data collection by collecting complete data only for those subjects who are chosen for the nested study. For example, in cardiovascular epidemiology the habitual energy expenditure of subjects has been measured using detailed diary records in which subjects record their physical activities in 15-minute blocks. Coding these diary records into energy expenditure is time consuming and expensive, but with a nested case-control design this conversion is only needed for the cases and their controls. Similar considerations apply to coding diary records in cohort studies in nutritional epidemiology, and to expensive laboratory analyses on biological specimens — these can be collected for all subjects in the cohort but "banked" and analyzed only for cases and their controls.

Another use of nested case-control studies is when an on-going cohort study is to be used to address a question about an exposure or confounder not measured in the original design. Data collection can be restricted to those subjects in a nested study. For example, suppose that routine health service monitoring data shows differences in mortality between groups of patients but, because information is not available on important confounders, it is not possible to exclude confounding as an explanation. A more detailed abstraction of medical records in a nested case-control study could make it possible to measure the confounders in the nested study and hence to control for them.

The final reason for using a nested case-control study is to avoid the computational burden associated with time-dependent explanatory vari-

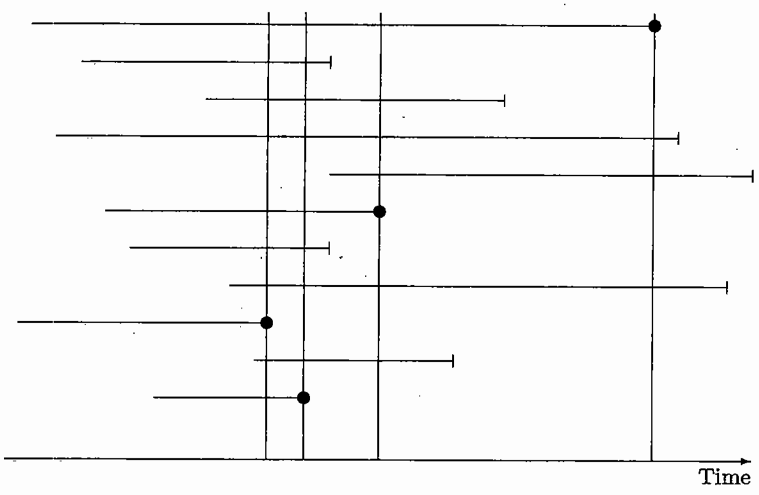


Fig. 33.1. Definition of risk sets.

ables. This problem was discussed briefly in Chapter 31, where we indicated that a natural design for such studies is to randomly sample the *risk sets* on which a full analysis by Cox's method would be based. In this chapter we discuss this suggestion in more detail.

33.2 Sampling risk sets

In nested case-control studies, controls are drawn for each case from the corresponding risk set. Fig. 33.1 shows the risk sets for a follow-up study of eleven subjects, four of whom fail. Corresponding to each of these four events is a risk set containing all those subjects under study at the moment of event occurrence — that is, all subjects whose observation lines cross the relevant vertical. To select controls we ignore the case and choose a random sample of the remaining subjects in the risk set. Sampling of a risk set must be carried out independently both of the sampling of other risk sets and of any later failure or censoring of its members.

Exercise 33.1. What are the sizes of the four risk sets? Indicate how you would select a single control for each case.

In the analysis of the full cohort study using Cox's method, each of the events contributes a term of the form

$$\log\left(\theta_{(\text{for case})} \Big/ \sum_{\text{Risk set}} \theta\right)$$

to the log partial likelihood. When the risk sets are sampled this becomes

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

which is identical to the log likelihood contribution of a matched case-control set in a conditional logistic regression analysis.

CAN THE SAME SUBJECT BE INCLUDED MORE THAN ONCE?

In the procedure for sampling risk sets described above the same subject can be selected as a control more than once and may eventually become a case. This will not happen very often for rare events but when it does it should be permitted. Any intervention in the sampling procedure to prevent its happening violates the requirement for independent sampling of risk sets.

A second aspect of this question is illustrated by the fourth subject shown in Fig. 33.1 who belongs to all four risk sets. If this subject is drawn as a control in one of these risk sets it is tempting to use him or her as an extra control in the other sets. Including a subject in all samples for which he/she is eligible represents an extremely *interdependent* method of sampling risk sets. The result is that the successive terms which contribute to the partial likelihood are no longer independent — each term does not contribute quite as much *new* information as it appears. When this dependence is taken into account the expected gain in precision as a result of multiple use of controls largely evaporates. However, there may be other advantages. One is that, because controls are no longer tied to a particular risk set, they can be randomly selected at the time of recruitment into the cohort study. This design has been called a *case-cohort* study, and some logistic advantages have been claimed. One situation in which it might be considered is for studies in which several different types of event are of interest — for example, occurrence of several different cancers. Independent sampling of risk sets leads to a different set of controls for each type of event while the case-cohort design allows a single control sample to be used for all outcomes. Against this must be weighed the fact that a more complex analysis is required to take account of the interdependency in the sampling of controls.

HOW MANY CONTROLS?

If there are m times as many controls as cases, the precision of the case-control study compared to the cohort study is given by

$$\frac{\text{SD of estimate from case-control data}}{\text{SD of estimate from entire study base}} = \sqrt{1 + \frac{1}{m}}.$$

This formula applies to the simple situation where the exposure effect is small and there is no control for confounding, but it can also be used as a rough guide more generally. Since $\sqrt{1+1/m}$ is only slightly greater than 1 for $m > 5$ little accuracy is lost by taking five or at most ten controls for each case, rather than the whole risk set.

33.3 Matching

In an occupational study of lung cancer, smoking will be a strong confounder, and the comparisón of occupational groups should therefore be controlled for smoking. An overall sample of (say) five controls per case could lead to a very different ratio within smokers and non- smokers. Since there will be many more cases among the smokers than among the non-smokers it is likely that there will fewer than five controls per case among smokers and many more than five per case among non-smokers. In such cases it would be better to match controls to cases with respect to smoking habits. Of course, this requires that smoking data are available for the entire cohort. The contribution to the log likelihood now becomes

$$\log\left(\theta_{(\text{for case})} \Big/ \sum \theta\right)$$

where the $\sum \theta$ denominator refers to summation over the case and the matched controls. Matching the controls to the cases on smoking does not allow estimation of the smoking effect, but when smoking is a confounder this need not concern us.

⋆ 33.4 Counter-matching

In the previous section we discussed the situation where the values of the confounding variables are known for all subjects in the cohort and a nested case-control study is used to reduce the cost of measuring the exposure. Matching controls to cases on the confounding variables can improve the precision of the comparison of exposure groups although, as a side-effect, the effects of the confounding variables cannot be estimated. What about the opposite situation in which the exposure variable is measured for all subjects in the cohort and a nested case-control study is used to reduce the cost of measuring the confounding variables? In this case it would be disastrous to match the controls to the cases on exposure since we would then be unable to estimate the effect of exposure. However, the information available for the full cohort can still be used to sample controls more efficiently.

To illustrate this we consider first the case in which all subjects are classified as exposed or unexposed. For any particular risk set let the numbers of exposed and unexposed subjects be N_1 and N_0 respectively, and suppose we are to draw m controls. The nested case-control set will contain $n = m + 1$ subjects (the case plus m controls). Let the split of these n subjects between exposed and unexposed be n_1 and n_0. When controls are drawn by simple random sampling of the risk sets this can produce a very uneven split of exposed and unexposed subjects and lead to inefficiency. The efficiency of the study can be improved by fixing the split in advance — usually to be 50:50.

For example, suppose that there are 10 exposed and 100 unexposed subjects in the risk set and we wish to select a sample of 5 exposed and 5 unexposed, including the case which defines the risk set. If the case is exposed this means we need 4 exposed controls and 5 unexposed controls. If the case is unexposed we need 5 exposed controls and 4 unexposed controls. For a sample of one exposed and one unexposed an exposed case will always be paired with an unexposed control and an unexposed case with an exposed control. It is from this that the term *counter-matching* is derived.

When sampling in this way the contribution of each risk set to the partial log likelihood must be adjusted to reflect the fact that the exposure distribution in the sample is different from the exposure distribution in the risk set. The modified log partial likelihood contributions take the form

$$\log\left((W\theta)_{(\text{for case})} \Bigg/ \sum_{\text{Case-control set}} (W\theta)\right),$$

where W are *risk weights* for each subject which compensate for the sampling. These weights take the values

$$W = \begin{cases} N_1/n_1 & \text{for an exposed subject} \\ N_0/n_0 & \text{for an unexposed subject.} \end{cases}$$

Note that the choice of weight depends only on exposure status and not upon whether the subject is a case or a control.

Exercise 33.2. What are the weights for exposed and unexposed subjects in a risk set with $N_1 = 10$ exposed subjects and $N_0 = 100$ unexposed subjects, in a 1:1 counter-matched study?

Exercise 33.3. For the special case where there are no confounders θ takes the value 1 for an unexposed subject and the value ϕ for an exposed subject, where ϕ is the (multiplicative) exposure effect. Show that, using the correct weights, the partial log likelihood contribution for the 1:1 sampled set is identical to the contribution of this risk set to the full cohort analysis.

The design and analysis extends readily to the case where there are more than two exposure categories. If the risk set contains N_i subjects in exposure category i and the case-control set is to contain n_i, then we draw either $n_i - 1$ or n_i controls at random according to whether or not the case falls into this category. The risk weight for subjects in this category is N_i/n_i.

The same design and analysis may be used when exposure data is difficult or expensive to collect, but in which we have a surrogate measure available for all subjects. If exposure is rare, it makes sense to use the surrogate exposure measurements to construct a more efficient nested study in which there is a more even split between exposed and unexposed subjects. In a 1:1 study, for example, a case classified as exposed by the surrogate measure would be paired with a control classified as unexposed, and a case classified as unexposed paired with a control classified as exposed. Remembering that in the 1:1 study only exposure discordant pairs are informative for the estimation of the exposure effect, this design is more efficient since it should increase the number of such pairs.

An area in which counter-matching by surrogate exposure measurement could prove particularly useful is pharmacoepidemiology. Exposure to any one drug is rare and can usually only be ascertained after detailed checking of medical records. However, a simple questionnaire might be very successful at identifying a subgroup particularly likely to have taken the drug of interest. The nested case-control study should contain all subjects in the group likely to have taken the drug, and a random sample of the remainder. With this design, the introduction of the correct risk weights into the partial likelihood analysis provides a valid estimate of the drug effect.

☆ 33.5 Two-stage sampling of controls

Both matching and counter-matching require that some information is available for all subjects in the cohort. The general rule is that, when this concerns a confounder we should consider using it for matching controls to cases while, if it concerns an exposure of interest, we should consider counter-matching.

Similar ideas may be useful even when we have no such data for the full cohort or, indeed, in a conventional case-control study. The information to be used in the final matching or counter-matching is collected in an initial study but complete data collection is only followed through in a subsample. This is known as a *two-stage* case-control study.

Solutions to the exercises

33.1 The risk set for the first event contains 10 subjects, the others contain 9, 7, and 4 subjects respectively. A control for the first case is selected at random from the remaining 9 subjects in the risk set. Similarly the remaining controls are sampled at random from the 8, 6, and 3 eligible subjects in the remaining risk sets.

33.2 In the 1:1 counter-matched study each set contains $n = 2$ subjects, 1 exposed and 1 unexposed so that $n_1 = n_0 = 1$. The risk weights used in the analysis are therefore,

$$W = \begin{cases} 10 & \text{for an exposed subject} \\ 100 & \text{for an unexposed subject.} \end{cases}$$

33.3 Suppose the case is exposed. Using the whole risk set the contribution to the log partial likelihood is

$$\log\left(\frac{\phi}{10\times\phi+100\times 1}\right).$$

Using the 1:1 counter-matched design, the contribution to the partial log likelihood is

$$\log\left(\frac{(10\phi)}{(10\phi)+(100)}\right) = \log(10) + \log\left(\frac{\phi}{10\times\phi+100\times 1}\right).$$

These two expressions are the same except for a constant term, log(10), which does not depend on ϕ and can be ignored. The same is true when the case is unexposed.

34
Gaussian regression models

Most of this book has been about events such as the incidence of disease or mortality. Although events are particularly important in epidemiology, in some studies the response of interest is a quantitative measurement such as blood pressure. The most widely used probability model for such responses is the Gaussian model, described in Chapter 8. In this chapter we show how regression models are used in conjunction with the Gaussian probability model. We shall call this combination *Gaussian regression* although it is more usual for it to be called simply regression or *multiple regression* because it was developed before other regression methods.

34.1 Models for the mean

The Gaussian probability model differs from the binary model in having two parameters instead of one. These are μ, the mean, and σ, the standard deviation. In the simplest situation changing the level of an explanatory variable changes the value of μ but leaves σ unchanged. The distributions of response for a comparison of exposed and unexposed subjects predicted by such a model is illustrated in Fig. 34.1. The effect of exposure is measured by the difference between the means, $\mu_1 - \mu_0$.

To control for confounding by age, using stratification, we would stratify by age and make the assumption that $\mu_1 - \mu_0$ is constant across age groups. This is equivalent to fitting the regression model

$$\text{Mean} = \text{Corner} + \text{Age} + \text{Exposure}.$$

The effect of exposure in this model is simply the (common) difference between mean responses for exposed and unexposed subjects within age groups.

To illustrate such models we shall use some additional data from the study of diet and coronary heart disease. These concern daily intake of fibre which is the response variable. Age and occupation are the explanatory variables, both with three levels.* Table 34.1 shows a simple summary of these data in which a separate estimate of mean and standard deviation has been calculated for each of the nine age–occupation groups. The main interest is in differences between occupations and inspection of the estimated means suggests that there is a systematic tendency for bank clerks to eat more fibre than the drivers and conductors. There is no obvious systematic variation in the standard deviation parameters, so the assumption that changing the levels of age and occupation does not affect σ is reasonable.

*Unpublished data

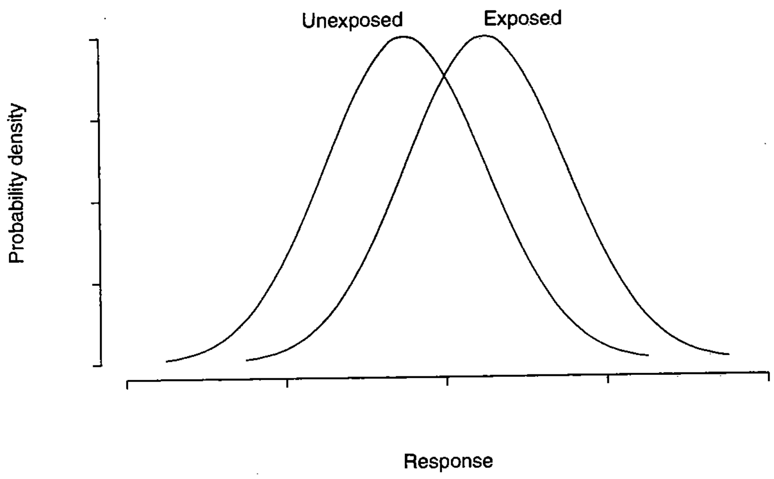


Fig. 34.1. Effect of exposure on the mean response.

The additive regression model relating the mean daily intake of fibre to the effects of age and occupation is

$$\text{Mean} = \text{Corner} + \text{Age} + \text{Work}.$$

When both age and work are treated as categorical this has five parameters in all, namely the Corner, Age(1), Age(2), Work(1), and Work(2) parameters. These are called the *regression parameters* to distinguish them from σ, the common standard deviation, which is called the *residual standard deviation*. The square of σ is called the *residual variance*.

34.2 Likelihood, sums of squares, and deviance

From Chapter 8, the log likelihood for a study of size N is

$$-N\log(\sigma) - \frac{1}{2}\sum_{\text{Subjects}}\left(\frac{x-\mu}{\sigma}\right)^2.$$

Table 34.1. Dietary fibre intake (gm/day) by age and occupation

		Occupation		
Age		Bus driver	Bus conductor	Bank clerk
< 45	N	23	16	38
	Mean	16.1	17.2	19.1
	SD	3.91	5.00	5.53
45 – 49	N	30	29	57
	Mean	16.3	17.0	18.5
	SD	4.22	5.42	6.88
50+	N	45	39	56
	Mean	16.6	14.8	17.6
	SD	6.28	4.48	5.43
All	N	98	84	151
	Mean	16.4	16.0	18.34
	SD	5.17	5.00	6.04

However, in contrast with Chapter 8, the mean parameter μ is not a single constant but can vary from subject to subject according to the regression model. In our example μ can take nine different values according to the combination of age and occupation. For estimating the regression parameters the $N\log(\sigma)$ term in the log likelihood can be ignored, and because σ is assumed to be the same for all subjects the parameter values which minimize the sum of squared differences,

$$\sum(x-\mu)^2,$$

will also maximize the log likelihood, regardless of the value of σ. Thus the most likely values of the regression parameters do not depend on σ. Because they minimize a sum of squared differences they are also called *least squares estimates.* The minimum value which this sum of squared differences takes is known as the *residual sum of squares.*

For example, Table 34.2 shows the parameter estimates for the model

$$\text{Mean} = \text{Corner} + \text{Work}$$

for the dietary fibre data. The table shows most likely values for the three parameters in this model, together with their standard deviations. The standard deviation of each regression parameter has been calculated from the profile log likelihood obtained by maximizing the log likelihood with respect to all the other regression parameters. Although the estimated values of these parameters do not depend on σ their standard deviations do, and in constructing the table σ has been taken equal to 5.5401 (we shall see where this value comes from later in the chapter).

Table 34.2. Effects of work on fibre intake (gm/day)

Parameter	Estimate	SD
Corner	16.425	0.560
Work(1)	−0.402	0.824
Work(2)	1.911	0.719

Exercise 34.1. Use the results in Table 34.2 to find the 90% confidence interval for the Work(1) parameter.

34.3 Analysis of deviance

The deviance for any fitted model is defined as minus twice the log likelihood ratio, when this compares the fitted model with a *saturated* model which has a parameter for each record. When the records refer to individual subjects the saturated model has $\mu = x$ so the deviance is

$$\sum\left(\frac{x-\mu}{\sigma}\right)^2.$$

This is proportional to the residual sum of squares for that model.[†] As before, the degrees of freedom for the deviance are equal to the the number of parameters in the saturated regression model, which is equal to the number of subjects N, less the number of parameters in the regression model which has been fitted. These are also the degrees of freedom for the residual sums of squares.

The deviance can be used to compare models in the same way as in Chapter 24, but all calculations are first done in terms of residual sums of squares and later converted to deviances by dividing by a suitable estimate of the square of σ. The residual sums of squares are obtained from the *analysis of variance* table which is usually in the output when a Gaussian regression model is fitted. For example, the analysis of variance table produced when fitting the model

$$\text{Mean} = \text{Corner} + \text{Age} + \text{Work}$$

to the data in Table 34.1 would look something like Table 34.3. The most important line in this table is the middle one labelled 'Error' which gives

[†] In the original definition of the idea of deviance, this was called the *scaled* deviance because of its dependence on the unknown scale parameter σ and the word deviance was reserved for its value when σ is taken as 1. However, this usage has not received widespread acceptance.

Table 34.3. Analysis of variance for the variable work

Source	DF	SSq
Model	2	369.891
Error	330	10128.636
Total	332	10498.527

the residual sum of squares for the model which has been fitted and its degrees of freedom. Since the number of subjects is $N = 333$ and the regression model has three parameters, the degrees of freedom here are $333 - 3 = 330$. The last line of the table, headed 'Total' gives the same information for the degenerate model

$$\text{Mean} = \text{Corner}$$

in which the mean response is the same for all subjects. This regression model has only one parameter so the degrees of freedom for its residual sum of squares and deviance are 332. The line labelled 'Model' is obtained by subtracting the degrees of freedom and the residual sum of squares for the error and total lines. When this difference in residual sum of squares is converted to a difference in deviance by division by the square of a suitable estimate of σ, it provides us with a test of the null hypothesis that all parameters in the model, other than the corner parameter, are zero. In this case this would be a test of the difference between occupations.

With more than one explanatory variable, testing the hypothesis that all the parameters in the model are zero is rarely of any interest. The only use of analysis of variance tables for such models is to obtain the residual sum of squared deviations from the second line. By fitting a series of models a more useful table can be constructed, as follows. Table 34.4 shows the residual sums of squares extracted from the analysis of variance tables for five models fitted to the fibre data. Changes in residual sums of squares from one model to another can be converted to deviances and used to test a variety of hypotheses. For example, the effects of work controlled for age can be tested by using the change in residual sum of squares between models 3 and 4.

ESTIMATING σ

Using the joint likelihood for the regression parameters and σ it can be shown, using calculus, that the most likely value of σ is

$$\sqrt{\frac{\text{Residual sum of squares}}{N}}.$$

Table 34.4. Analysis of deviance ($\sigma = 5.5445$)

Mean = Corner + ···	DF	SSq	Deviance
1. –	332	10498.527	341.510
2. Work	330	10128.636	329.478
3. Age	330	10384.702	337.807
4. Age + Work	328	10048.456	326.870
5. Age + Work + Age·Work	324	9960.268	324.000

This is the value of σ which maximizes the total likelihood and it therefore also maximizes the profile likelihood for σ. When the number of regression parameters is large compared with the number of subjects, it is preferable to use a conditional likelihood which depends only on σ, rather than the profile likelihood. The most likely value of σ is then equal to the residual sum of squares divided by its degrees of freedom. For example, the value of σ used throughout Table 34.4 was

$$\sigma = \sqrt{9960.268/324} = 5.5445$$

which is the conditional estimate obtained from model 5, although the overall most likely value is

$$\sigma = \sqrt{9960.268/333} = 5.4691$$

It can be seen that the use of the degrees of freedom in place of N has a negligible effect for a study of this size. The reason why σ is generally estimated from the conditional likelihood can be illustrated by a simple argument. If we imagine a study of 10 subjects and fit a regression model with 10 parameters it will fit the observations exactly. The overall most likely value of σ would be zero but the reality is that we have no data for estimating σ. Only when we add an eleventh subject to our study do we start collecting information about σ. It follows that the *effective* size of the study for the purposes of estimating σ is given by the N minus the number of regression parameters — the degrees of freedom — and the estimated value of σ should be

$$\sqrt{\frac{\text{Residual sum of squares}}{\text{Degrees of freedom}}}.$$

One consequence of using this estimate is that the deviance for the model used to estimate σ is equal to its degrees of freedom.

A test for interaction between work and age may be obtained by comparing the deviances for models 4 and 5. The difference in deviance is $326.870 - 324.000 = 2.870$ with $326 - 324 = 2$ degrees of freedom. Referring this to the chi-squared distribution shows this to be clearly non-

Table 34.5. Effects of age and work on fibre intake (gm/day)

Parameter	Estimate	SD
Corner	16.430	0.560
Age(1)	−0.223	0.814
Age(2)	−1.118	0.788
Work(1)	−0.387	0.824
Work(2)	1.828	0.720

significant so that we are reassured concerning our assumption of constant occupational effects over age groups.

The parameter estimates for model 4 are shown in Table 34.5. Note, however, that the value of σ used to calculate the standard deviations of the parameters is slightly different from that used in Table 34.4. This is because, whereas the estimate of σ used in Table 34.4 was obtained from model 5, Table 34.5 refers to model 4 and it is therefore logical to estimate σ using this model, that is by

$$\sigma = \sqrt{10048.456/328} = 5.5349.$$

The significance of the occupational effect, controlled for age, can be tested by comparing the deviances for models 4 and 3. However, since this test only makes sense when there is no interaction, deviances should properly be calculated using the model 4 estimate of σ rather than that used in Table 34.4.

Exercise 34.2. Carry out the test for the effect of occupation controlled for age.

Similarly, the value of σ used to calculate standard deviations of parameter estimates in Table 34.2 is obtained from model 2,

$$\sigma = \sqrt{10128.636/330} = 5.5401$$

and this is the value which would be used if we wished to compare models 1 and 2. In practice the difference between the possible estimates of σ are usually inconsequential except in very small studies.

F RATIO TESTS

The tests discussed above refer changes in deviance to the appropriate chi-squared distribution. If the value of σ were a known constant, these would be *exact tests*. However, when σ is estimated they are only approximate. Exact tests which take account of the fact that σ is estimated may be carried out using *F distributions*, tables of which are readily available. Instead of referring the change in deviance to the chi-square distribution, we divide it by the corresponding degrees of freedom to obtain the *F ratio*. For example, the change in deviance for the test for interaction was 2.870, with two degrees of freedom, so the corresponding F ratio is 1.435. To obtain the exact p-value, the F ratio is referred to the correct F distribution. However, to select the correct F distribution, we must specify two different numbers of degrees of freedom. The first, called the *numerator* degrees of freedom, is the same as the degrees of freedom for the approximate chi-squared test while the second, called the *denominator* degrees of freedom, is the number of degrees of freedom used to estimate σ. In our example these are 2 and 334 respectively.

In practice there is only a noticeable difference between F ratio tests and the approximate chi-squared test in small studies. In our example, the p-value obtained from the chi-squared distribution is 0.2381 while that obtained from the F distribution is 0.2396. Since the F ratio test is only exact if the assumptions of Gaussian distribution shape and constancy of σ are true, they are not usually worth the (admittedly slight) extra trouble.

34.4 Multiplicative models ⋆

A basic assumption in the Gaussian regression model is that changes in the explanatory variables affect the mean level of response but not the variability. However, it is commonly the case that as the level of response goes up, so does its variability. A simple multiplicative model acting at the individual level would explain this, for if the effect of changing the level of work is to double the values of the individual responses, then the standard deviation of these individual values will also get doubled. On a log scale, however, the effect of doubling the response will be to add log(2) to the log response, leaving the standard deviation of the log responses unchanged. This suggests that when the effects appear to act multiplicatively at an individual level, the log response should be analysed in place of the response.

There is some suggestion in Table 34.1 that standard deviation of fibre intake goes up with the mean, so that a multiplicative model may be more appropriate. This suggests analysing log fibre intakes rather than fibre intakes themselves. Inspection of the data suggests that the distribution of log fibre intake is closer to the Gaussian shape than the distribution of fibre intake, and this is another point in favour of analysing log fibre intakes. When the Gaussian regression model

$$\text{Mean} = \text{Corner} + \text{Age} + \text{Work}.$$

is fitted to the logs of the fibre intakes we obtain the parameter estimates shown in Table 34.6.

The effect parameters shown in this table are additive effects upon log fibre intake and these should be exponentiated to express them as multi-

Table 34.6. Effects of age and work on log fibre intake

Parameter	Estimate	SD
Corner	2.8039	0.0430
Age(1)	−0.0253	0.0445
Age(2)	−0.0800	0.0431
Work(1)	−0.0345	0.0451
Work(2)	0.0962	0.0394

plicative effects on fibre intake. The error factor method can be used to calculate confidence intervals for the multiplicative effects.

Exercise 34.3. Express the estimates of the Work parameters as multiplicative effects, and calculate 90% confidence intervals.

Apart from this change in the way the parameter estimates are interpreted the use of the log response in place of the response does not affect matters. Models are compared using residual sums of squares in the same way as before.

If the effect of the explanatory variables is multiplicative at a group level, but not at an individual level, so that σ is constant, a multiplicative model such as

$$\text{Mean} = \text{Corner} \times \text{Age} \times \text{Work},$$

can be fitted to the data on the original scale. Computer programs are available for fitting such models but the need for them rarely arises because the idea of an explanatory variable acting multiplicatively at a group level but not at an individual level is rather implausible.

Solutions to the exercises

34.1 The 90% confidence interval is from $-0.402 - 1.645 \times 0.824 = -1.757$ to $-0.402 + 1.645 \times 0.824 = 0.953$. The lower limit is a reduction of 1.757 gm, the upper limit is an increase of 0.953 gm.

34.2 The appropriate value for σ is 5.5349, taken from the model which includes both age and work. The deviance for this model is then 328.000, and the deviance for the model which includes age alone is

$$10384.702/5.5349^2 = 338.982.$$

The change in deviances is $338.982 - 328.000 = 10.982$ on 2 degrees of freedom, for which $p = 0.004$ (from the chi-squared distribution on two degrees of freedom.

34.3 The Work(1) parameter is estimated as −0.0345, and since

$$\exp(-0.0345) = 0.966,$$

the fibre intakes of conductors are 0.966 times those of drivers. The 90% confidence interval for this ratio is found from the error factor

$$\exp(1.645 \times 0.0451) = 1.077,$$

to be from $0.966/1.077 = 0.897$ to $0.966 \times 1.077 = 1.04$. Similarly, the multiplicative effect of Work(2) is 1.101 with 90% confidence interval from 1.032 to 1.175.

35 Postscript

No scientific methodology stands still and statistical modelling is no exception. In this book we have deliberately restricted our attention to well-established methods which have become a routine part of modern epidemiology, and omitted newer developments, even though some of these will undoubtedly make important contributions to epidemiology in the future. Two areas in particular are worth mentioning. The first is the extension of the models discussed in this book to deal with errors of measurement of explanatory variables (see Chapter 27). The second concerns the extension of these models to *longitudinal studies* in which the response is measured on several different occasions for each subject.

The methods we have described concentrate on the analysis of response at the level of the individual subject. Even when these analyses have been carried out using frequency records this has been purely for computational convenience and parameters still refer to the effects upon the response for an individual subject. However, some epidemiological research is based upon the behaviour of aggregated groups of individuals, for example the inhabitants of a country, region, or town. Statistical analysis then concentrates on description and 'explanation' of differences in the aggregate responses of such groups in time and space. By analogy with the discipline of economics, such activity could be termed *macro-epidemiology.* We have not dealt with it in this book, firstly because this field is currently undergoing active development, and secondly because new likelihoods and fitting procedures become necessary as a result of the more complicated probability models which are a necessary response to lack of data at the subject level.

Some further reading

A good elementary introduction to statistical modelling using the computer program GLIM is:

Healy, M. (1988) *GLIM. An Introduction.* Oxford Science Publications, Oxford University Press, Oxford.

The reader who requires more mathematical details can find them in a

number of statistical texts. General treatments of regression model, including Poisson and logistic regression, are given by the following authors.

Aitkin, M., Anderson, D., Francis, B., and Hinde, J. (1989) *Statistical modelling in GLIM.* Oxford Science Publications, Oxford University Press, Oxford.

McCullagh, M. and Nelder, J.A. (1989) *Generalized linear models* (2nd edn). Chapman and Hall, London.

Descriptions of modern statistical approaches to the analysis of life tables and survival data are given by the following authors.

Cox, D.R. and Oakes, D. (1984) *The analysis of survival data.* Chapman and Hall, London.

Kalbfleisch, J.D. and Prentice, R.L. (1980) *The statistical analysis of failure time data.* Wiley, New York.

A detailed exposition of a more general mathematical approach to modelling event occurrence in time is to be found in:

Andersen, P.K., Borgan, Ø., Gill, R.D., and Keiding, N. (1993) *Statistical models based on counting processes.* Springer, New York.

Intermediate in technical level between these purely statistical texts and this book are:

Breslow, N.E. and Day, N. (1980) *Statistical methods in cancer epidemiology. Vol. I – The analysis of case-control studies.* IARC Scientific Publications No. 32. International Agency for Research on Cancer, Lyon.

Breslow, N.E. and Day, N. (1987) *Statistical methods in cancer epidemiology. Vol. II – The design and analysis of cohort studies.* IARC Scientific Publications No. 82. International Agency for Research on Cancer, Lyon.

A collection of papers dealing with very recent research in epidemiological modelling is:

Moolgavkar, S.H. and Prentice, R.L. (ed.) (1986) *Modern statistical methods in chronic disease epidemiology.* Wiley, New York.

An extensive review of the more recent statistical literature is:

Gail, M.H. (1991) A bibliography and comments on the use of statistical models in epidemiology in the 1980s. *Statistics in Medicine*, **10**, 1819–95.

Part III

Appendices

Appendix A
Exponentials and logarithms

Raising 10 to different powers is a familiar operation. For example,

$$10^1 = 10,\ 10^2 = 100,\ 10^3 = 1000,\ \cdots$$

Mathematically this is regarded as a rule for getting from the power (1, 2, 3, etc.) to the value of 10 raised to that power (10, 100, 1000, etc.). The power is often referred to as the *exponent* and 10 raised to a power is called an *exponential* with base 10.

Raising 10 to a power can be extended to cover fractional powers using the convention that $10^{\frac{1}{2}}$ stands for the square root of 10, $10^{\frac{1}{3}}$ stands for the cube root of 10, and so on. The rule can also be extended to cover negative powers using the convention that 10^{-1} stands for $1/10 = 0.1$. Table A.1 shows the rule for obtaining 10^x from x for a variety of values of x.

Now suppose that we wish to go the other way and, starting with a value of 10^x, find the value of x. For example, starting with 1000 gives $x = 3$, while starting with 0.1 gives $x = -1$. Starting with any positive number y, the value of x which makes $10^x = y$ is called the *logarithm* of y with the base 10 and is written $\log_{10}(y)$. Taking logarithms with base 10 is the inverse operation to exponentiation with base 10. Thus $10^3 = 1000$ and $\log_{10}(1000) = 3$.

Table A.1. Rules for finding 10^x from x

x	$y = 10^x$
0	1
1	10
2	100
3	1000
-1	0.1
-2	0.01
-3	0.001
$\frac{1}{2}$	$\sqrt{10}$
$\frac{1}{3}$	$\sqrt[3]{10}$

Table A.2. Multiplication using logarithms

Number		Logarithm
7.2	⟶	0.8573
16.9	⟶	1.2279
121.7	⟵	2.0852

Logarithms were introduced as a computational device in the seventeenth century to avoid multiplication and division. Tables were prepared so that the logarithm of any number could be looked up. Similarly, tables of exponentials were prepared so that logarithms could be converted back to the original numbers. These tables of exponentials were called *antilogarithms*. The use of logarithms to multiply 7.2 by 16.9 is shown in Table A.2. Arrows from left to right refer to looking up logarithms while arrows from right to left refer to looking up antilogarithms (exponentiation). The result line follows from addition on the logarithmic (right-hand) side or multiplication on the exponential (left-hand) side. The widespread availability of cheap electronic calculators means that nobody now uses logarithms for multiplication or division. However, their mathematical property of converting multiplication to addition, embodied in

$$\log(7.2 \times 16.9) = \log(7.2) + \log(16.9)$$

is still very useful. Another useful property which follows from this is that

$$\log(7.2^2) = 2 \times \log(7.2)$$

$$\log(7.2^3) = 3 \times \log(7.2)$$

and so on.

Raising 2 to a power is called exponentiation with base 2. The inverse process produces logarithms to the base 2 and these are written $\log_2(y)$. Both exponentials and logarithms can be defined with respect to any base. Fig. A.1 shows plots of the exponential functions 10^x, 3^x, e^x, and 2^x, where the symbol e represents the number 2.71828183. The number e is chosen so that the tangent to the plot of e^x versus x drawn at $x = 0$ has a slope of exactly 1 (shown by the broken line). It follows that *when x is very small,*

$$e^x \approx 1 + x.$$

and, therefore,

$$\log_e(1 + x) \approx x.$$

Logarithms to the base e are referred to as *natural* logarithms, and it is the above property that makes them 'natural'. The natural logarithm function is sometimes written as $\ln(y)$, but in this book we shall *always* use logarithms to the base e, and write them simply as $\log(y)$. We also write the exponential function with base e as $\exp(x)$. Note, however, that many electronic calculators assign an entirely different meaning to a key marked *exp*.

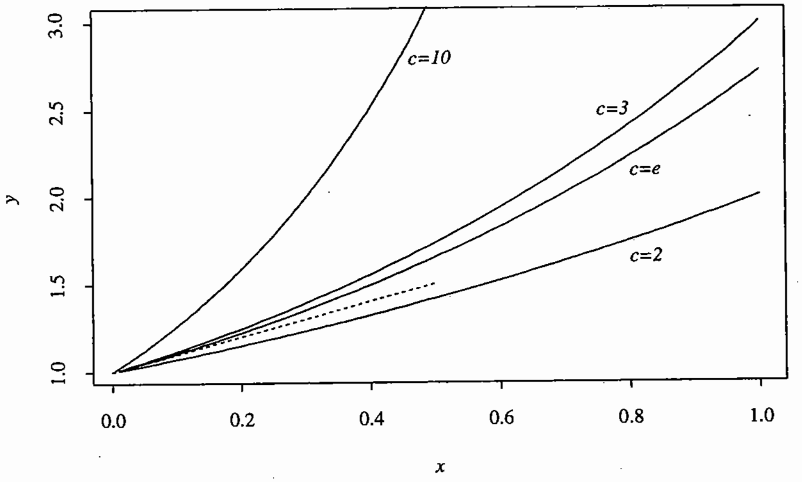


Fig. A.1. Plots of the function $y = c^x$

The logarithms of the same number, using different bases, are related by a simple constant multiplier. For example

$$\log_e(y) = \log_{10}(y) \times 2.3026$$

where $2.3026 = \log_e(10)$. Similarly

$$\log_2(y) = \log_{10}(y) \times 3.3219$$

where $3.3219 = \log_2(10)$.

Appendix B

★ Some basic calculus

The *gradient* of the graph of y versus x measures the rate at which y is increasing (or decreasing) at any point on the graph. It is most easily defined for a straight line graph, such as the one in Fig. B.1. In this case the rate of increase or decrease is the same at any point on the graph, and is measured by the ratio of the *rise* to the *run*. For a straight line relationship in which y *decreases* with x the gradient is negative. Gradients have units equal to those of y/x. The central idea of calculus is that over a small run any curve is approximately a straight line and the gradient of the curve at any point in the run is approximately equal to the gradient of this line.

Differential calculus consists of a number of simple rules which are used to evaluate gradients of curves for which the y co-ordinate of any point on the curve is given by some function of the x co-ordinate. The most useful of these are shown in Table B.1. A further very important rule is that the gradient of a function constructed as the *sum* of two simpler functions is the sum of the gradients of the constituent functions so that, for example, the gradient of $x + \log(x)$ is $1 + 1/x$.

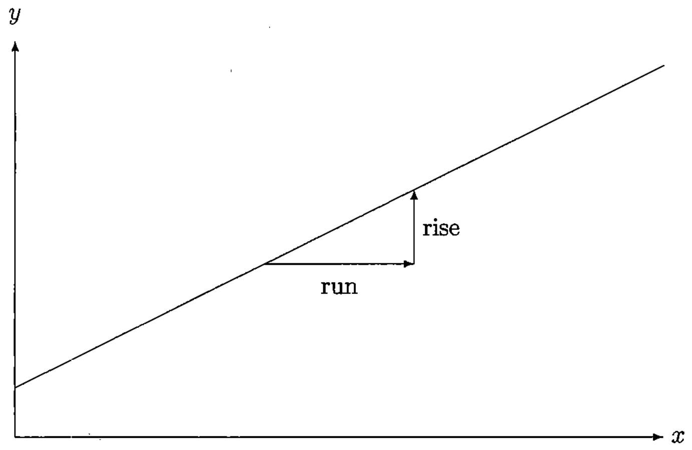


Fig. B.1. Gradient of a straight line graph

Table B.1. Gradients of some simple functions of x

Function	Gradient
c (constant)	0
x	1
$-x$	-1
cx	c
$(x)^2$	$2x$
$(x)^m$	$m(x)^{m-1}$
$\frac{1}{x} = (x)^{-1}$	$-(x)^{-2} = -\frac{1}{(x)^2}$
$\exp(x)$	$\exp(x)$
$\log(x)$	$\frac{1}{x}$
$(c+x)^2$	$2(c+x)$
$(c-x)^2$	$-2(c-x)$
$\log(c+x)$	$\frac{1}{c+x}$
$\log(c-x)$	$-\frac{1}{c-x}$

The use of these rules is now illustrated by finding the gradient of the log likelihood for a rate λ, based on D cases and Y person years. The log likelihood for λ is

$$D\log(\lambda) - \lambda Y.$$

From Table B.1 the gradient of $\log(\lambda)$ is $1/\lambda$ and the gradient of λ is 1. Hence the gradient of the log likelihood is

$$\frac{D}{\lambda} - Y.$$

The maximum value of the log likelihood occurs when the gradient is zero, that is, when $\lambda = D/Y$, so the most likely value of λ is D/Y.

The curvature of the log likelihood curve at the peak is important in determining the range of supported values. A highly curved peak corresponds to a narrow range. The curvature at a point on a curve is a measure of how fast the gradient is changing from one value of x to the next; if the gradient is changing quickly then the curvature is high, while if the gradient is changing slowly the curvature is low. For log likelihood curves the gradient changes from a positive quantity (on the left) to a negative quantity (on the right) so the gradient decreases as x increases and the curvature is negative.

The curvature of a curve, at a point, is defined to be the rate of change of the gradient of the curve at that point. The way that Table B.1 can be used to find curvature is now illustrated using the log likelihood for λ

again. The gradient of the log likelihood at any value of λ has been shown to be

$$\frac{D}{\lambda} - Y.$$

From Table B.1 the gradient of a constant is zero and the gradient of $1/\lambda$ is $-1/(\lambda)^2$, so the curvature of the log likelihood at any value of λ is

$$-\frac{D}{(\lambda)^2}.$$

Appendix C
Approximate profile likelihoods ★

This appendix describes the mathematics underlying Gaussian approximation of profile log likelihoods.

C.1 The difference between two parameters

We shall start with an important special case. Consider a model with two parameters, β_1 and β_0, and suppose that our main interest is in the *difference*

$$\gamma = \beta_1 - \beta_0.$$

We shall further assume that the log likelihoods for β_1 and β_0 are based on two independent sets of data so that the total log likelihood is the sum of the two separate log likelihoods.

Fig. C.1 illustrates the construction of the profile likelihood for γ. The upper panel of the figure shows the total log likelihood obtained by adding the log likelihoods for β_1 and β_0. Contours are shown for log likelihood ratios of $-5, -4, \ldots, -1$. The four diagonal lines correspond to different values of γ. For example, the top leftmost line represents values of β_1, β_0 satisfying

$$\beta_1 - \beta_0 = 0$$

so that this line corresponds to $\gamma = 0$. Similarly, the remaining lines correspond to values of γ of 0.5, 1.0, and 1.5 respectively. To find the profile likelihood for γ, we find the maximum value of the log likelihood along each of these lines. This maximum is plotted against γ in the lower panel of the figure.

The Gaussian approximation of the profile log likelihood can be obtained from making use of the relationship between gradients and curvatures of the total log likelihood (upper panel), and the gradient and curvature of the profile log likelihood (lower panel). These relationships can be derived using the laws of calculus but are only quoted here.

If, at the maximum of the log likelihood along the line $\beta_1 - \beta_0 = \gamma$, the gradient is G_1 with respect to β_1 and G_0 with respect to β_0 the gradient

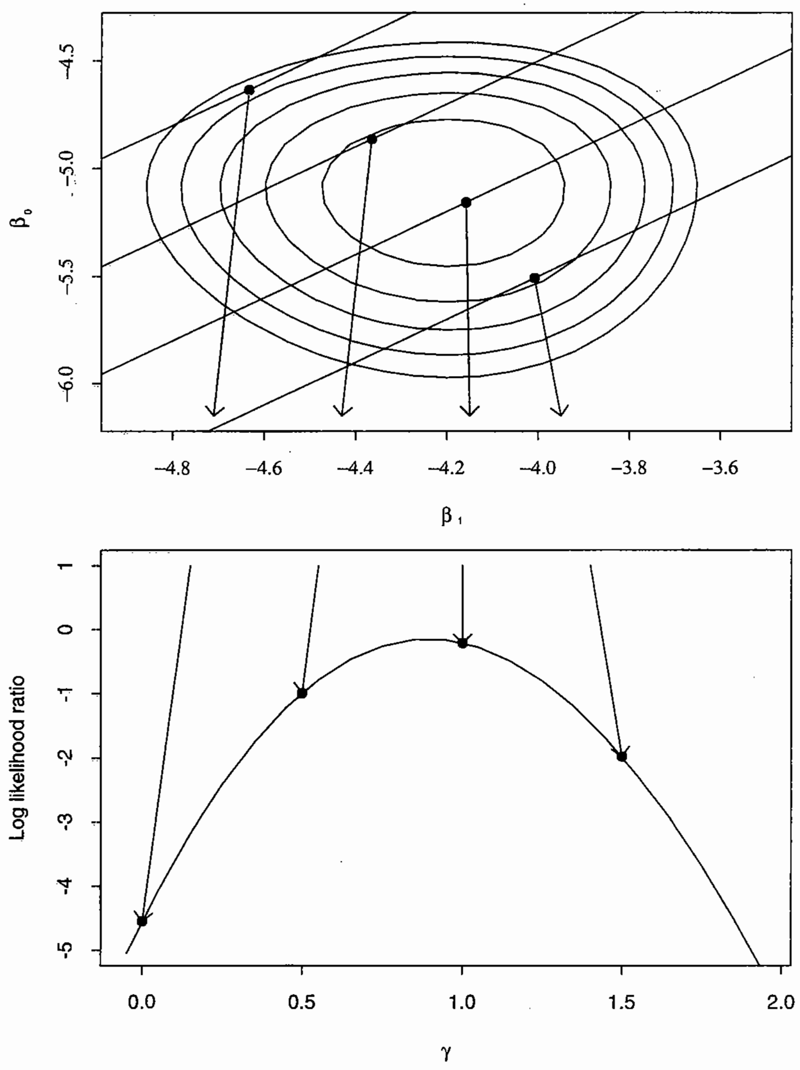


Fig. C.1. The profile log likelihood

of the profile log likelihood at γ is G, where

$$G = G_1 = -G_0.$$

If C_1, C_0 are the corresponding curvatures with respect to β_1 and β_0, then the curvature of the profile log likelihood at γ is C, where

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_0}.$$

From these results it follows directly that, if the most likely values of β_1 and β_0 are M_1 and M_0 respectively, and the corresponding standard deviations of the estimates are S_1 and S_0, then the most likely value of γ is

$$M = M_1 - M_0,$$

and the standard deviation of the estimate is

$$S = \sqrt{(S_1)^2 + (S_0)^2}.$$

THE RATE RATIO REVISITED

As an example, we shall apply use these general rules to the problem of estimating and testing the logarithm of the rate ratio. Let λ_0 and λ_1 be the two rate parameters and define

$$\beta_1 = \log(\lambda_1), \qquad \beta_0 = \log(\lambda_0)$$

then

$$\begin{aligned}\gamma &= \beta_1 - \beta_0\\ &= \log\left(\frac{\lambda_1}{\lambda_0}\right)\\ &= \log(\theta),\end{aligned}$$

the log of the rate ratio.

If, in the exposed group, D_1 cases are observed in Y_1 person-years, and in the unexposed group D_0 cases are observed in Y_0 person-years, the total log likelihood is

$$D_1 \log(\lambda_1) - \lambda_1 Y_1 \quad + \quad D_0 \log(\lambda_0) - \lambda_0 Y_1.$$

The gradients of this with respect to β_1 and β_0 are

$$G_1 = D_1 - \lambda_1 Y_1 \qquad G_0 = D_0 - \lambda_0 Y_0,$$

and the curvatures are

$$C_1 = -\lambda_1 Y_1 \qquad C_0 = -\lambda_0 Y_0.$$

The most likely values for β_1 and β_0 are

$$M_1 = \log(D_1/Y_1), \qquad M_0 = \log(D_0/Y_0)$$

and the corresponding standard deviations are

$$S_1 = \sqrt{1/D_1}, \qquad S_0 = \sqrt{1/D_0}.$$

Using the rules given at the end of the last section, the Gaussian approximation for the profile log likelihood for $\gamma = \log(\theta)$ has

$$\begin{aligned} M &= \log(D_1/Y_1) - \log(D_0/Y_0) \\ &= \log\left(\frac{D_1/Y_1}{D_0/Y_0}\right), \end{aligned}$$

and

$$S = \sqrt{\frac{1}{D_1} + \frac{1}{D_0}}.$$

These expressions are identical to those obtained in Chapter 13.

The Wald test is also based on the Gaussian approximation shown above. The score test is obtained from the gradient and curvature of the profile log likelihood at the null value of the parameter, $\gamma = 0$. Here λ_1 and λ_0 are equal and their most likely common value is D/Y so that the gradients and curvatures are

$$\begin{aligned} G_1 &= D_1 - E_1 & \qquad G_0 &= D_0 - E_0 \\ C_1 &= -E_1 & \qquad C_0 &= -E_0 \end{aligned}$$

where $E_1 = (D/Y)Y_1$ and $E_0 = (D/Y)Y_0$ represent 'expected' numbers of failures in the two groups under the null hypothesis. The score, U, is given by either G_1 or $-G_0$ (it can easily be verified that these are identical). The score variance is minus the curvature of the profile log likelihood and, using the relationship

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_0}.$$

this is

$$\begin{aligned} V &= \left(\frac{1}{E_1} + \frac{1}{E_0}\right)^{-1} \end{aligned}$$

$$= \frac{E_1 E_0}{E}$$

Since $D = E$, this can also be written

$$\begin{aligned} V &= D\frac{E_1}{E}\frac{E_0}{E} \\ &= D\frac{E_1}{E}\left(1 - \frac{E_1}{E}\right) \end{aligned}$$

and this agrees with the expression given in Chapter 13.

THE DIFFERENCE BETWEEN TWO MEANS

A second example is the difference between two mean parameters in a Gaussian model for responses measured on a continuous metric scale. For example, we might wish to compare blood pressure in two groups of subjects. We shall let μ_1 and μ_0 represent the mean parameters for the two groups and assume that the standard deviation of responses about the mean is the same in both groups, σ let us say. As in Chapter 8 we shall assume σ to be a known constant although, in practice, it would also have to be estimated from the data.

Exercise C.1. Derive expressions for the most likely value and for the standard deviation of the estimate of the parameter

$$\gamma = \mu_1 - \mu_0.$$

C.2 Weighted sums

Similar results hold for more general problems. For example, the parameter of interest may be defined as

$$\gamma = W_1\beta_1 + W_0\beta_0$$

where W_1 and W_0 are known constants. In this case the same argument illustrated in Fig. C.1 may be applied, but the parallel lines corresponding to fixed values of γ now have different slopes. The relationship between gradients in the total log likelihood and the gradient of the profile likelihood is now

$$G = \frac{G_1}{W_1} = \frac{G_0}{W_0}$$

and for the curvatures we have

$$\frac{1}{C} = \frac{(W_1)^2}{C_1} + \frac{(W_0)^2}{C_0}.$$

These results generalize in an obvious way to a function of more than two parameters, of the form

$$\gamma = W_1\beta_1 + W_2\beta_2 + W_3\beta_3 + \cdots \quad ,$$

the gradient of the profile log likelihood now being

$$G = \frac{G_1}{W_1} = \frac{G_2}{W_2} = \frac{G_3}{W_3} = \cdots$$

and its curvature

$$\frac{1}{C} = \frac{(W_1)^2}{C_1} + \frac{(W_2)^2}{C_2} + \frac{(W_3)^2}{C_3} + \cdots \quad .$$

If the most likely values of $\beta_1, \beta_2, \ldots$ are $M_1, M_2, \ldots$ with standard deviations $S_1, S_2, \ldots$, then the most likely value of γ is

$$M = W_1M_1 + W_2M_2 + W_3M_3 + \cdots$$

with standard deviation

$$S = \sqrt{(W_1S_1)^2 + (W_2S_2)^2 + (W_3S_3)^2 + \cdots} \quad .$$

Solutions to the exercises

C.1 The log likelihoods for μ_1 and μ_0 are Gaussian with most likely values M_1 and M_0 — the arithmetic means of the N_1 observations in the first group and the N_0 observations in the second. The corresponding standard deviations are

$$S_1 = \frac{\sigma}{\sqrt{N_1}}, \qquad S_0 = \frac{\sigma}{\sqrt{N_0}}.$$

It follows from the results of this section that the profile log likelihood for $\mu_1 - \mu_0$ has most likely value $M_1 - M_0$ and standard deviation

$$\sqrt{\frac{(\sigma)^2}{N_1} + \frac{(\sigma)^2}{N_0}} = \sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_0}}.$$

Appendix D
Table of the chi-squared distribution

Probability	Degrees of freedom, ν				
p	1	2	3	4	5
0.50	0.455	1.386	2.366	3.357	4.351
0.25	1.323	2.773	4.108	5.385	6.626
0.10	2.706	4.605	6.251	7.779	9.2367
0.075	3.170	5.181	6.905	8.496	10.008
0.050	3.841	5.991	7.815	9.488	11.070
0.025	5.024	7.378	9.348	11.143	12.833
0.0100	6.635	9.210	11.345	13.277	15.086
0.0075	7.149	9.786	11.966	13.937	15.780
0.0050	7.879	10.597	12.838	14.860	16.750
0.0025	9.141	11.983	14.320	16.424	18.386
0.0010	10.828	13.816	16.266	18.467	20.515

Probability	Degrees of freedom, ν				
p	6	7	8	9	10
0.50	5.348	6.346	7.344	8.343	9.342
0.25	7.841	9.037	10.219	11.389	12.549
0.10	10.645	12.017	13.362	14.684	15.987
0.075	11.466	12.883	14.270	15.631	16.971
0.050	12.592	14.067	15.507	16.919	18.307
0.025	14.449	16.013	17.535	19.023	20.483
0.0100	16.812	18.475	20.090	21.666	23.209
0.0075	17.537	19.229	20.870	22.471	24.038
0.0050	18.548	20.278	21.955	23.589	25.188
0.0025	20.249	22.040	23.774	25.462	27.112
0.0010	22.458	24.322	26.124	27.877	29.588

The above tables give the value that a variable, distributed according to the chi-squared distribution with ν degrees of freedom, will exceed with probability p. For example, a variable distributed according to the chi-squared distribution with one degree of freedom has a probability of $p = 0.1$ of exceeding the value 2.706.

Index