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)

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	$\lambda_C\phi^2$	$\lambda_C\theta\phi^2$

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.

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 (≥ 2750 kcals)			Exposed (< 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.