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:

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

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

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

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 + ···
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

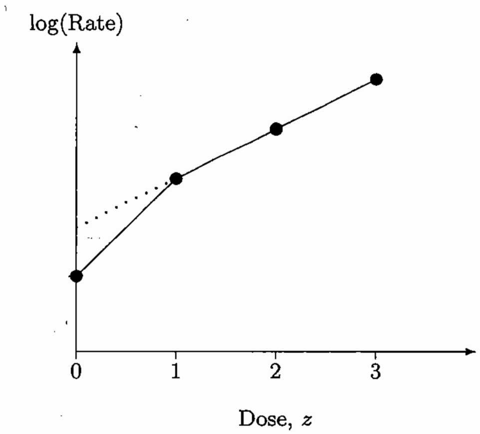

Fig. 25.1. Separating zero exposure from the dose-response.

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]

non-smokers to fit the model

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

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 $\times 16.87$ 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 $\times 16.23$ 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)