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

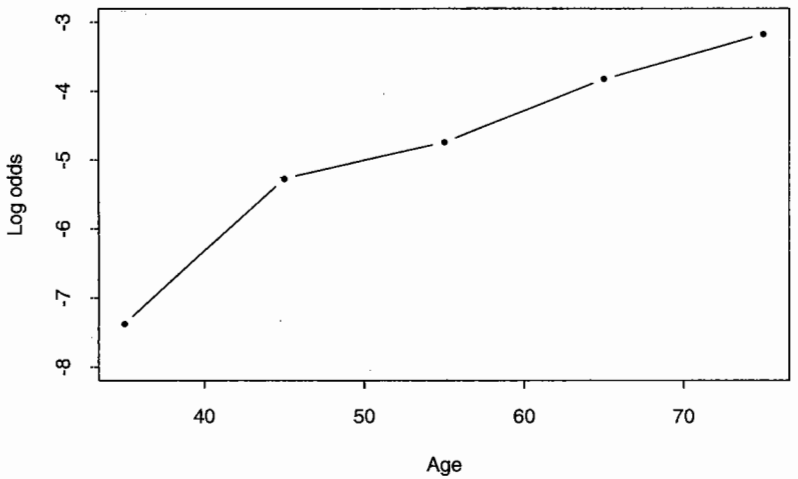

Fig. 26.1. Log prevalence odds by age

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

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

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.

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

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

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

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

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

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

Table 26.7. Rate parameters per 100 000 person-years

	A	
B	0	1
0	5.0	15.0
1	20.0	λ

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.

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.