explained in Chapter 24, the effects in this model should be interpreted as t effect of smoking controlled for social class and the effect of social class controll for smoking. To test the null hypothesis that there is no effect of social class aft controlling for smoking, we compare:

1 the log likelihood of model 2, which includes only smoking, with
2 the log likelihood of model 4 which also includes social class, with the additi corresponding to the effect of social class controlled for smoking.

The likelihood ratio test statistic is:

$$\begin{aligned} \text{LRS} &= -2 \times (L_{exc} - L_{inc}) = -2 \times (L_{\text{model 2}} - L_{\text{model 4}}) \\ &= -2 \times (-1195.513 + 1191.119) = 8.788 \end{aligned}$$

d.f. = number of additional parameters in the inclusive model = 7 − 2 = 5

$P = 0.118$

There is therefore no good evidence for an association between social cla and rates of myocardial infarction, other than that which acts through smokin However, we should be aware that for an ordered categorical variab such as social class a more powerful approach may be to derive a *test for tre* by including social class as a linear effect in the model, rather than as categorical variable. Modelling linear effects is discussed in detail in Secti 29.6.

29.5 INVESTIGATING INTERACTION (EFFECT MODIFICATION) IN REGRESSION MODELS

Interaction was introduced in Section 18.5, where we explained that there is *interaction* between the effects of two exposures if the effect of one exposure vari according to the level of the other exposure. For example, the protective effect breastfeeding against infectious diseases in early infancy is more pronounc among infants living in poor environmental conditions than among those livi in areas with adequate water supply and sanitation facilities. We also explain that an alternative term for interaction is **effect modification**. In this example, can think of this as the quality of environmental conditions *modifying* the effect breastfeeding. Finally, we noted that the most flexible approach to examin interaction is to use regression models, but that when we are using Mantel Haenszel methods to control for confounding an alternative is to use a χ^2 te for effect modification, commonly called a **χ^2 test of heterogeneity**. Interactio effect modification and heterogeneity are three different ways of describing exactl the same thing.

We have also seen that regression models including the effect of two or mo exposures make the assumption that there is *no interaction* between the exposure We now describe how to test this assumption by introducing **interaction terms** int the regression model.

Example 29.2

We will explain this in the context of the onchocerciasis dataset used throughout Chapters 19 and 20, where logistic regression was used to examine the effects of area of residence (forest or savannah) and of age group on the odds of microfilarial (*mf*) infection. We found strong associations of both area of residence and of age group with the odds of *mf* infection. We will do three things:

1 Remind ourselves of the results of the standard logistic regression model including both area and age group, which assumes that is there is *no interaction* between the two. In other words, it assumes that the effect of area is the same in each of the four age groups, and (correspondingly) that the effect of age is the same in the each of the two areas, and that any observed differences are due to sampling variation. Unless you are already familiar with how such models work, we strongly suggest that you read Section 20.2 where this is explained in detail, before continuing with this section.
2 We will then describe how to specify a regression model incorporating an interaction between the effects of area and age group, and how to interpret the regression output from such a model.
3 We will then calculate a likelihood ratio statistic using the log likelihoods of these two models to test the null hypothesis that there is no interaction between the effects of area and age group.

Model with two exposures and no interaction

Table 29.3 summarizes the results from the logistic regression model for *mf* infection including both area and age group, described in Section 20.2. Part (a) of the table shows the set of equations for the eight subgroups of the data that define the model in terms of its parameters. Note that the exposure effects represent odds ratios, and that they are *multiplicative*, since logistic regression models the *log odds*. The eight subgroups can be divided into four different types:

1 The *baseline* subgroup, consisting of those in the baseline groups of both area and age, namely those aged 5–9 years living in a savannah area. This is represented by the Baseline parameter in the model.
2 One subgroup consisting of those in the baseline group for age, but *not* for area, namely those aged 5–9 years living in a rainforest area. This subgroup is '*exposed to area but not to age*'. Its relative odds of *mf* infection compared to the baseline is modelled by the Area parameter.
3 Three subgroups corresponding to those in each of the three non-baseline age groups, but who are in the baseline group for area, namely those living in savannah areas aged 10–19 years, 20–39 years, or 40 years or more. These subgroups are '*exposed to age but not area*'. Their relative odds of *mf* infection compared to the baseline are modelled by the three age group parameters, Agegrp(1), Agegrp(2) and Agegrp(3), respectively.

4 Three subgroups corresponding to those in each of the three non-baseline age groups who are also in the non-baseline group for area, namely those living in rainforest areas aged 10–19 years, 20–39 years, or 40 years or more. These subgroups are '*exposed to both area and age*'. If we assume that there is *no interaction* between the two exposures, the relative odds of *mf* infection in these three subgroups compared to the baseline are modelled by multiplying together the Area parameter and the relevant age group parameter. This gives Area × Agegrp(1), Area × Agegrp(2) and Area × Agegrp(3), respectively.

The model for the odds of *mf* infection in the eight subgroups therefore contains just five parameters. This is made possible by the assumption of *no interaction*. The parameter estimates are shown in part (b) of Table 29.3. Part (c) shows the values obtained when these estimates are inserted into the equations in part (a) to give estimated values of the odds of *mf* infection according to area and age group. The observed odds of *mf* infection in each group are also shown.

Model incorporating an interaction between the two exposures

We now describe how to specify an alternative regression model incorporating an interaction between the effects of the two exposures. We no longer assume that the

Table 29.3 Results from the logistic regression model for *mf* infection, including both area of residence and age group, assuming *no interaction* between the effects of area and age group.

(a) Odds of *mf* infection by area and age group, expressed in terms of the parameters of the logistic regression model: Odds = Baseline × Area × Age group.

	Odds of *mf* infection	
Age group	Savannah areas (Unexposed)	Rainforest areas (Exposed)
0 (5–9 years)	Baseline	Baseline × Area
1 (10–19 years)	Baseline × Agegrp(1)	Baseline × Area × Agegrp(1)
2 (20–39 years)	Baseline × Agegrp(2)	Baseline × Area × Agegrp(2)
3 (≥ 40 years)	Baseline × Agegrp(3)	Baseline × Area × Agegrp(3)

(b) Parameter estimates obtained by fitting the model.

	Baseline	Area	Agegrp(1)	Agegrp(2)	Agegrp(3)
Odds ratio	0.147	3.083	2.599	9.765	17.64

(c) Odds of *mf* infection by area and age group, as estimated from the logistic regression model, and as observed.

	Savannah areas: odds of *mf* infection		Rainforest areas: odds of *mf* infection	
Age group	Estimated	Observed	Estimated	Observed
0 (5–9 years)	0.147	0.208	0.147 × 3.083 = 0.453	0.380
1 (10–19 years)	0.147 × 2.599 = 0.382	0.440	0.147 × 3.083 × 2.599 = 1.178	1.116
2 (20–39 years)	0.147 × 9.765 = 1.435	1.447	0.147 × 3.083 × 9.765 = 4.426	4.400
3 (≥ 40 years)	0.147 × 17.64 = 2.593	2.182	0.147 × 3.083 × 17.64 = 7.993	10.32

relative odds of *mf* infection in the subgroups '*exposed to both age and area*' can be modelled by multiplying the area and age effects together. Instead we introduce extra parameters, called **interaction parameters**, as shown in Table 29.4(a). These allow the effect of area to be different in the four age groups and, correspondingly, the effects of age to be different in the two areas. An interaction parameter is denoted by the exposure parameters for the subgroup written with a full stop between them. The three interaction parameters in this example are denoted Area.Agegrp(1), Area.Agegrp(2) and Area.Agegrp(3).

This new model is fitted using seven indicator variables as shown in Box 29.1. The parameter estimates for this model are shown in Table 29.4(b). Table 29.4(c) shows the values obtained when these are inserted into the equations in part (a). Note that:

1 Since this model has *eight* parameters, the same as the number of area × age subgroups, there is an exact agreement between the estimated odds of *mf* infection in each subgroup and the observed odds, as shown in Tables 29.3(c) and 20.3.
2 Including interaction terms leads to different estimates of the baseline, area and age group parameters than those obtained in the model assuming no interaction. It is important to realize that the interpretation of the area and age group parameters is also different.
   - The Area parameter estimate (1.8275) is the odds ratio for area *in the baseline age group*. In the model assuming no interaction, the Area parameter estimate (3.083) is a weighted average of the odds ratios for area in the four age groups and is interpreted as the odds ratio for area after controlling for age group.
   - Similarly, the age group parameter estimates represent the effect of age in the *baseline area group*, in other words the effect among those living in savannah areas.
3 The estimates for the interaction parameters are all greater than one. This corresponds to a synergistic effect between area and each of the age groups, with the combined effect more than simply the combination of the separate effects. A value of one for an interaction term is equivalent to no interaction effect. A value less than one would mean that the combined effect of both exposures is less than the combination of their separate effects.
4 The interaction parameters allow the area effect to be different in the four age groups. They can be used to calculate age-specific area odds ratios as follows:
   - The Area parameter estimate equals 1.8275, and is the area odds ratio (comparing those living in rainforest areas with those living in savannah areas) in the *baseline* age group (5–9 years).
   - Multiplying the Area parameter estimate by the interaction parameter estimate Area.Agegrp(1) gives the odds ratio for area in age group 1 (10–19 years):

$$\text{OR for area in age group 1} = \text{Area} \times \text{Area.Agegrp(1)}$$
$$= 1.8275 \times 1.3878 = 2.5362$$

Table 29.4 Logistic regression model for *mf* infection, including both area of residence and age group, and *incorporating an interaction* between their effects.

(a) Odds of *mf* infection by area and age group, expressed in terms of the parameters of the logistic regres model, with the interaction parameters shown in bold: Odds = Baseline × Area × Agegroup × Area.Agegrou

	Odds of *mf* infection	
Age group	Savannah areas (Unexposed)	Rainforest areas (Exposed)
0 (5–9 years)	Baseline	Baseline × Area
1 (10–19 years)	Baseline × Agegrp(1)	Baseline × Area × Agegrp(1) × **Area.Agegrp(1)**
2 (20–39 years)	Baseline × Agegrp(2)	Baseline × Area × Agegrp(2) × **Area.Agegrp(2)**
3 (≥40 years)	Baseline × Agegrp(3)	Baseline × Area × Agegrp(3) × **Area.Agegrp(3)**

(b) Computer output showing the results from fitting the model (interaction parameters shown in bold).

	Odds ratio	z	$P > \|z\|$	95 % CI
Area.Agegrp(1)	**1.3878**	**0.708**	**0.479**	**0.560 to 3.43**
Area.Agegrp(2)	**1.6638**	**1.227**	**0.220**	**0.738 to 3.75**
Area.Agegrp(3)	**2.5881**	**2.171**	**0.030**	**1.097 to 6.10**
Area	1.8275	1.730	0.084	0.923 to 3.619
Agegrp(1)	2.1175	1.998	0.046	1.015 to 4.420
Agegrp(2)	6.9639	6.284	0.000	3.802 to 12.76
Agegrp(3)	10.500	7.362	0.000	5.614 to 19.64
Constant (Baseline)	0.2078	−5.72	0.000	0.121 to 0.356

(c) Odds of *mf* infection by area and age group, as estimated from the logistic regression model, with interaction parameters shown in bold.

	Odds of *mf* infection	
Age group	Savannah areas	Rainforest areas
0 (5–9 years)	0.2078	$0.2078 \times 1.8275 = 0.380$
1 (10–19 years)	$0.2078 \times 2.1175 = 0.440$	$0.2078 \times 1.8275 \times 2.1175 \times \mathbf{1.3878} = 1.116$
2 (20–39 years)	$0.2078 \times 6.9639 = 1.447$	$0.2078 \times 1.8275 \times 6.9639 \times \mathbf{1.6638} = 4.400$
3 (≥40 years)	$0.2078 \times 10.500 = 2.182$	$0.2078 \times 1.8275 \times 10.500 \times \mathbf{2.5881} = 10.32$

Similarly,

$$\text{OR for area in age group 2} = \text{Area} \times \text{Area.Agegrp(2)}$$
$$= 1.8275 \times 1.6638 = 3.0406$$

and

$$\text{OR for area in age group 3} = \text{Area} \times \text{Area.Agegrp(3)}$$
$$= 1.8275 \times 2.5881 = 4.7300$$

These four age-group-specific area odds ratios are the same as those show Tables 20.3 and 20.4.

5 In exactly the same way, the interaction parameters can be used to calculate area-specific age group odds ratios. For example:

$$\text{OR for age group 1 in rainforest areas} = \text{Agegrp(1)} \times \text{Area.Agegrp(1)} = 2.1175 \times 1.3878 = 2.9386$$

6 An alternative expression of these same relationships is to note that the interaction parameter Area.Agegrp(1) is equal to the *ratio* of the odds ratios for area in age group 1 and age group 0, presented in Tables 20.3 and 20.4. For example:

$$\text{Area.Agegrp(1)} = \frac{\text{OR for area in age group 1}}{\text{OR for area in age group 0}} = \frac{2.5362}{1.8275} = 1.3878$$

If there is no interaction then the area odds ratios are the same in each age group and the interaction parameter equals 1.

7 Alternatively, we can express the interaction parameter Area.Agegrp(1) as the ratio of the odds ratios for age group 1 (compared to age group 0), in area 1 and area 0:

$$\text{Area.Agegrp(1)} = \frac{\text{OR for age group 1 in area 1}}{\text{OR for age group 1 in area 0}} = \frac{2.9386}{2.1175} = 1.3878$$

(The odds ratios for age group 1 were calculated using the raw data presented in Table 20.3).

8 The other interaction parameter estimates all have similar interpretations: for example the estimate for Area.Agegrp(2) equals the ratio of the area odds ratios in age group 2 and age group 0, and equivalently it equals the ratio of the odds ratios for age group 2 (compared to age group 0) in area 1 and area 0.

9 For a model allowing for interaction between two binary exposure variables, the *P*-value corresponding to the interaction parameter estimate corresponds to a Wald test of the null hypothesis that there is no interaction. When, as in this example, there is more than one interaction parameter, the individual *P*-values corresponding to the interaction parameters are not useful in assessing the evidence for interaction: we describe how to derive the appropriate likelihood ratio test later in this section.

Table 29.5 summarizes the interpretation of the interaction parameters for different types of regression models.

Table 29.5 Interpretation of interaction parameters.

Type of regression model	Interpretation of interaction parameters
Linear	Difference between mean differences
Logistic	Ratio of odds ratios
Poisson	Ratio of rate ratios

BOX 29.1 USING INDICATOR VARIABLES TO INVESTIGATE INTERACTION IN REGRESSION MODELS

Values of the seven indicator variables used in a model to examine the interaction between area (binary variable) and age group (4 groups):

Age group	Area	Area	Age(1)	Age(2)	Age(3)	Area.Age(1)	Area.Age(2)	Area.Age(3)
5–9 years (0)	Savannah	0	0	0	0	0	0	0
	Forest	1	0	0	0	0	0	0
10–19 (1)	Savannah	0	1	0	0	0	0	0
	Forest	1	1	0	0	1	0	0
20–39 years (2)	Savannah	0	0	1	0	0	0	0
	Forest	1	0	1	0	0	1	0
≥40 years (3)	Savannah	0	0	0	1	0	0	0
	Forest	1	0	0	1	0	0	1

Likelihood ratio test for interaction

To test the null hypothesis that there is no interaction between area and age group, we need to compare the log likelihoods obtained in the two models excluding and including the interaction parameters. These are shown in Table 29.6. The likelihood ratio test statistic is:

$$\text{LRS} = -2 \times (L_{exc} - L_{inc}) = -2 \times (-692.407 + 689.773) = 5.268$$

$$\text{d.f.} = \text{number of additional parameters in the inclusive model} = 8 - 5 = 3$$

$$P = 0.153$$

Therefore this analysis provides little evidence of interaction between the effects of area and age on the odds of microfilarial infection

Table 29.6 Log likelihood values obtained from the logistic regression models for *mf* infection by area of residence and age group, (a) assuming *no* interaction, and (b) incorporating an interaction between the effects of area and age group.

Model	Exposure(s) in model	No. of parameters	Log likelihood
(a) exc	Area and Agegrp	5	−692.407
(b) inc	Area, Agegrp and Area.Agegrp	8	−689.773

Interactions with continuous variables

It is straightforward to incorporate an interaction between the effects of a continuous exposure variable (x) and a binary exposure variable (b, coded as 0 for

unexposed and 1 for exposed individuals) in a regression model, by multiplying the values of the two exposures together to create a new variable ($x.b$) representing the interaction, as shown in Table 29.7. This new variable equals 0 for those unexposed to exposure b, and the value of exposure x for those exposed to b. The regression coefficient for $x.b$ then corresponds to the difference between the slope in individuals exposed to b and the slope in individuals not exposed to b, and the evidence for an interaction may be assessed either using the Wald P-value for $x.b$, or by omitting $x.b$ from the model and performing a likelihood ratio test.

To examine interactions between two continuous exposure variables w and x, it is usual to create a new variable $w.x$ by multiplying w by x. If the regression coefficient for $w.x$ is 0 (1 for models with exposure effects reported as ratios) then there is no evidence of interaction.

Table 29.7 Creating a variable to represent an interaction between a continuous and a binary exposure variable.

Continuous exposure (x)	Binary exposure (b)	Interaction variable ($x.b$)
x	0 (unexposed)	0
x	1 (exposed)	x

Confounding and interaction

Note that confounding and interaction may coexist. If there is clear evidence of an interaction between the exposure and the confounder, it is no longer adequate to report the effect of the exposure controlled for the confounder, since this assumes the effect of the exposure to be the same at each level of the confounder. This is not the case when interaction is present. Instead, we should report *separate* exposure effects for each *stratum* of the confounder. We can derive these by performing a separate regression to examine the association between the exposure and outcome variables, for each level of the confounding variable.

It is possible to derive stratum-specific effects in regression models by including appropriate indicator variables, or combining regression coefficients as was done in Table 29.4(c). This has the advantage of allowing estimation of such effects, controlled for the effects of other exposure variables. Confidence intervals for such combinations of regression coefficients need to take into account the covariance (a measure of the association) between the individual regression coefficients: some statistical packages provide commands to combine regression coefficients and derive corresponding confidence intervals.

An advantage of Mantel–Haenszel methods is that because the stratum-specific exposure effects tend to be presented in computer output, we are encouraged to look for evidence of interaction. In regression models we have to fit interaction terms explicitly to do this.