CHAPTER 19

Logistic regression: comparing two or more exposure groups

19.1 INTRODUCTION

In this chapter we introduce **logistic regression**, the method most commonly used for the analysis of *binary* outcome variables. We show how it can be used to examine the effect of a *single* exposure variable, and in particular, how it can be used to:

- Compare a binary outcome variable between two exposure (or treatment) groups.
- Compare more than two exposure groups.
- Examine the effect of an ordered or continuous exposure variable.

We will see that it gives *very similar results* to the methods for analysing *odds ratios* described in Chapters 16, 17 and 18, and is an alternative to them. We will also see how logistic regression provides a flexible means of analysing the association between a binary outcome and a *number* of exposure variables. In the next chapter, we will explain how it is used to control for confounding. We will also briefly describe the regression analysis of risk ratios, and methods for the analysis of categorical outcomes with more than two levels.

We will explain the principles of logistic regression modelling in detail in the next section, in the simple context of comparing two exposure groups. In particular, we will show how it is based on modelling odds ratios, and explain how to interpret the computer output from a logistic regression analysis. We will then introduce the general form of the logistic regression equation, and explain where the name 'logistic' comes from. Finally we will explain how to fit logistic regression models for categorical, ordered or continuous exposure variables.

Links between multiple regression models for the analysis of numerical outcomes, the logistic regression models introduced here, and other types of regression model introduced later in the book, are discussed in detail in Chapter 29.

19.2 LOGISTIC REGRESSION FOR COMPARING TWO EXPOSURE GROUPS

Introducing the logistic regression model

We will start by showing, in the simple case of two exposure groups, how logistic regression models the association between binary outcomes and exposure variables in terms of odds ratios. Recall from Chapter 16 that the *exposure odds ratio* (OR) is defined as:

$$\text{Exposure odds ratio} = \frac{\text{Odds in exposed group}}{\text{Odds in unexposed group}}$$

If we re-express this as:

$$\text{Odds in exposed} = \text{Odds in unexposed} \times \text{Exposure odds ratio}$$

then we have the basis for a simple model for the odds of the outcome, which expresses the odds in each group in terms of two **model parameters**. These are:

1 The **baseline** odds. We use the term **baseline** to refer to the exposure group against which all the other groups will be compared. When there are just two exposure groups as here, then the baseline odds are the odds in the unexposed group. We will use the parameter name 'Baseline' to refer to the odds in the baseline group.
2 The **exposure odds ratio**. This expresses the effect of the exposure on the odds of disease. We will use the parameter name 'Exposure' to refer to the exposure odds ratio.

Table 19.1 shows the odds in each of the two exposure groups, in terms of the parameters of the logistic regression model.

Table 19.1 Odds of the outcome in terms of the parameters of a logistic regression model comparing two exposure groups.

Exposure group	Odds of outcome	Odds of outcome, in terms of the parameter names
Exposed (group 1)	Baseline odds × exposure odds ratio	Baseline × Exposure
Unexposed (group 0)	Baseline odds	Baseline

The logistic regression model defined by the two equations for the odds of the outcome shown in Table 19.1 can be abbreviated to:

$$\text{Odds} = \text{Baseline} \times \text{Exposure}$$

Since the two parameters in this model *multiply* together, the model is said to be **multiplicative**. This is in contrast to the multiple regression models described in Chapter 11, in which the effects of different exposures were *additive*. If there were two exposures (A and B), the model would be:

$$\text{Odds} = \text{Baseline} \times \text{Exposure(A)} \times \text{Exposure(B)}$$

Thus if, for example, exposure A doubled the odds of disease and exposure B trebled it, a person exposed to both would have a six times greater odds of disease than a person in the baseline group exposed to neither. We describe such models in detail in the next chapter.

Example 19.1

All our examples of logistic regression models are based on data from a study of onchocerciasis ('river blindness') in Sierra Leone (McMahon *et al*. 1988, *Trans Roy Soc Trop Med Hyg* **82**; 595–600), in which subjects were classified according to whether they lived in villages in savannah (grassland) or rainforest areas. In addition, subjects were classified as infected if microfilariae (*mf*) of *Onchocerciasis volvulus* were found in skin snips taken from the iliac crest. The study included persons aged 5 years and above. Table 19.2 shows that the prevalence of microfilarial infection appears to be greater for individuals living in rainforest areas compared to those living in the savannah; the associated odds ratio is $2.540/1.052 = 2.413$.

We will now show how to use logistic regression to examine the association between area of residence and microfilarial infection in these data. To use a **computer package to fit a logistic regression model**, it is necessary to specify just two items:

1 The *name of the outcome* variable, which in this case is *mf*. The **required convention for coding** is to code the outcome event (D) as 1, and the absence of the outcome event (H) as 0. The variable *mf* was therefore coded as 0 for uninfected subjects and 1 for infected subjects.
2 The *name of the exposure* variable(s). In this example, we have just one exposure variable, which is called *area*. The required *convention for coding* is that used throughout this book; thus *area* was coded as 0 for subjects living in savannah areas (the *baseline* or '*unexposed*' group) and 1 for subjects living in rainforest areas (the '*exposed*' group).

Table 19.2 Numbers and percentages of individuals infected with onchocerciasis according to their area of residence, in a study of 1302 individuals in Sierra Leone.

	Microfilarial infection			
Area of residence	Yes	No	Total	Odds of infection
Rainforest	$d_1 = 541$ (71.7%)	$h_1 = 213$ (28.3%)	754	$541/213 = 2.540$
Savannah (baseline group)	$d_0 = 281$ (51.3%)	$h_0 = 267$ (48.7%)	548	$281/267 = 1.052$
Total	822	480	1302	

Table 19.3 First ten lines of the computer dataset from the study of onchocerciasis.

id	*mf*	Area
1	1	0
2	1	1
3	1	0
4	0	1
5	0	0
6	0	1
7	1	0
8	1	1
9	1	1
10	1	1

The first ten lines of the dataset, when entered on the computer, are shown in Table 19.3. For example, subject number 1 lived in a savannah area and was infected, number 2 lived in a rainforest area and was also infected, whereas subject number 4 lived in a rainforest area but was not infected.

The **logistic regression model** that will be fitted is:

$$\text{Odds of } mf \text{ infection} = \text{Baseline} \times \text{Area}$$

Its two parameters are:

1 baseline: the odds of infection in the baseline group (subjects living in savannah areas); and

2 area: the odds ratio comparing odds of infection among subjects living in rainforest areas with that among those living in savannah areas.

Table 19.4 shows the computer output obtained from fitting this model. The two *rows* in the output correspond to the two *parameters* of the logistic regression model; area is our exposure of interest, and the **constant** term refers to the baseline group. The same format is used for both parameters, and is based on what makes sense for interpretation of the effect of exposure. This means that some of the information presented for the constant (baseline) parameter is not of interest.

Table 19.4 Logistic regression output for the model relating odds of infection to area of residence, in 1302 subjects participating in a study of onchocerciasis in Sierra Leone.

	Odds ratio	*z*	$P > \|z\|$	95% CI
Area	2.413	7.487	0.000	1.916 to 3.039
Constant	1.052	0.598	0.550	0.890 to 1.244

The column labelled 'Odds ratio' contains the **parameter estimates**:

1 For the first row, labelled 'area', this is the *odds ratio* (2.413) comparing rainforest (area 1) with savannah (area 0). This is identical to the odds ratio which was calculated directly from the raw data (see Table 19.3).
2 For the second row, labelled 'constant', this is the *odds of infection in the baseline group* (1.052 = odds of infection in the savannah area, see Table 19.3). As we will see, this apparently inconsistent labelling is because output from regression models is labelled in a uniform way.

The remaining columns present z statistics, P-values and 95% confidence intervals corresponding to the model parameters. The values for *area* are exactly the same as those that would be obtained by following the procedures described in Section 16.7 for the calculation of a 95% confidence interval for an odds ratio, and the associated Wald test. They will be explained in more detail in the explanation of Table 19.5 below.

The logistic regression model on a log scale

As described in Chapter 16, confidence intervals for odds ratios are derived by using the standard error of the *log* odds ratio to calculate a confidence interval for the *log* odds ratio. The results are then *antilogged* to express them in terms of the original scale. The same is true for logistic regression models; they are *fitted on a log scale*. Table 19.5 shows the two equations that define the logistic regression model for the comparison of two exposure groups. The middle column shows the model for the odds of the outcome, as described above. Using the rules of logarithms (see p. 156, Section 16.5), it follows that corresponding equations on the log scale for the log of the odds of the outcome are as shown in the right-hand column. Note that as in the rest of the book all logs are to the base e (natural logarithms) unless they are explicitly denoted as logs to the base 10 by $\log_{10}$ (see Section 13.2).

Table 19.5 Equations defining the logistic regression model for the comparison of two exposure groups.

Exposure group	Odds of outcome	Log odds of outcome
Exposed (group 1)	Baseline odds × exposure OR	Log(baseline odds) + log(exposure OR)
Unexposed (group 0)	Baseline odds	Log(baseline odds)

Using the parameter names introduced earlier in this section, the logistic regression model on the log scale can be written:

$$\log(\text{Odds}) = \log(\text{Baseline}) + \log(\text{Exposure odds ratio})$$

In practice, we abbreviate it to:

$$\log(\text{Odds}) = \text{Baseline} + \text{Exposure}$$

since it is clear from the context that *output on the log scale refers to log odds and log odds ratios*. Note that whereas the exposure effect on the odds ratio scale is *multiplicative*, the exposure effect on the log scale is *additive*.

Example 19.1 (continued)

In this example, the model on the log scale is:

$$\log(\text{Odds of } mf \text{ infection}) = \text{Baseline} + \text{Area}$$

where

1 *baseline* is the *log odds* of infection in the savannah areas; and
2 *area* is the *log odds ratio* comparing the odds of infection in rainforest areas with that in savannah areas.

Table 19.6 shows the results obtained on the log scale, for this model. We will explain each item in the table, and then discuss how the results relate to those on the odds ratio scale, shown in Table 19.4.

Table 19.6 Logistic regression output (log scale) for the association between microfilarial infection and area of residence.

	Coefficient	s.e.	z	$P > \|z\|$	95% CI
Area	0.881	0.118	7.487	0.000	0.650 to 1.112
Constant	0.0511	0.0854	0.598	0.550	−0.116 to 0.219

1 The two *rows* in the output correspond to the terms in the model; area is our exposure of interest, and as before the **constant term** corresponds to the baseline group.
2 The *first* column gives the results for the **regression coefficients** (corresponding to the parameter estimates on a log scale):
   (a) For the row labelled 'area', this is the **log odds ratio** comparing rainforest with savannah. It agrees with what would be obtained if it were calculated directly from Table 19.3, and with the value in Table 19.4:

$$\log \text{OR} = \log(2.540/1.052) = \log(2.413) = 0.881$$

   (b) For the row labelled 'constant', this is the **log odds in the baseline group** (the group with exposure level 0), i.e. the log odds of microfilarial infection in the savannah:

$$\log \text{odds} = \log(281/267) = \log(1.052) = 0.0511.$$

3 The *second* column gives the standard error(s) of the regression coefficient(s). In the simple example of a binary exposure variable, as we have here, the standard errors of the regression coefficients are exactly the same as those derived using the formulae given in Chapter 16. Thus:

(a) s.e.(log OR comparing rainforest with savannah) is:

$$\begin{aligned}\sqrt{(1/d_1 + 1/h_1 + 1/d_0 + 1/h_0)} &= \sqrt{(1/541 + 1/213 + 1/281 + 1/267)} \\ &= 0.118\end{aligned}$$

(b) s.e.(log odds in savannah) is:

$$\sqrt{(1/d_0 + 1/h_0)} = \sqrt{(1/281 + 1/267)} = 0.0854$$

4 The 95% confidence intervals for the regression coefficients in the *last* column are derived in the usual way.

(a) For the log OR comparing rainforest with savannah, the 95% CI is:

$$0.881 - (1.96 \times 0.118) \text{ to } 0.881 + (1.96 \times 0.118) = 0.650 \text{ to } 1.112$$

(b) For the log odds in the savannah, the 95% CI is:

$$0.0511 - (1.96 \times 0.0854) \text{ to } 0.0511 + (1.96 \times 0.0854) = -0.116 \text{ to } 0.219$$

5 The z statistic in the *area* row of the third column is used to derive a **Wald test** (see Chapter 28) of the null hypothesis that the *area* coefficient $= 0$, i.e. that the exposure has no effect (since if log OR $= 0$, then OR must be equal to 1). This z statistic is simply the regression coefficient divided by its standard error:

$$z = 0.881/0.118 = 7.487$$

6 The P-value in the *fourth* column is derived from the z statistic in the usual manner (see Table A1 and Chapter 8), and can be used to assess the strength of the evidence against the null hypothesis that the true (population) exposure effect is zero. Thus, the P-value of 0.000 (which should be interpreted as < 0.001) for the log OR comparing rainforest with savannah indicates that there is strong evidence against the null hypothesis that the odds of microfilarial infection are the same in the two areas.

7 We are usually not interested in in the third and fourth columns (the z statistic and its P-value) for the *constant* row. However, for completeness, we will explain their meanings:

(a) The z statistic is the result of testing the null hypothesis that the log odds of infection in the savannah areas are 0 (or, equivalently, that the odds of infection are 1). This would happen if the risk of infection in the savannah areas was 0.5; in other words if people living in the savannah areas were equally likely to be infected as they were to be not infected.
(b) The P-value of 0.550 for the log odds in savannah areas indicates that there is no evidence against this null hypothesis.

Relation between outputs on the ratio and log scales

We will now explain the relationship between the two sets of outputs, since the results in Table 19.4 (output on the original, or ratio, scale) are derived from the results in Table 19.6 (output on the log scale). Once this is understood, it is rarely necessary to refer to the output displayed on the log scale: the most useful results are the odds ratios, confidence intervals and P-values displayed on the original scale, as in Table 19.4.

1 In Table 19.4, the column labelled 'Odds Ratio' contains the *exponentials* (antilogs) of the logistic regression coefficients shown in Table 19.6. Thus the OR comparing rainforest with savannah $= \exp(0.881) = 2.413$.
2 The z statistics and P-values are derived from the log odds ratio and its standard error, and so are identical in the two tables.
3 The 95% confidence intervals in Table 19.4 are derived by antilogging (exponentiating) the confidence intervals on the log scale presented in Table 19.6. Thus the 95% CI for the OR comparing rainforest with savannah is:

$$95\%\ \text{CI} = \exp(0.650)\ \text{to}\ \exp(1.112) = 1.916\ \text{to}\ 3.039$$

This is identical to the 95% CI calculated using the methods described in Section 16.7:

$$95\%\ \text{CI (OR)} = \text{OR}/\text{EF}\ \text{to}\ \text{OR} \times \text{EF},\ \text{where EF} = \exp[1.96 \times \text{s.e.}(\log\ \text{OR})]$$

Note that since the calculations are multiplicative:

$$\frac{\text{Odds ratio}}{\text{Lower confidence limit}} = \frac{\text{Upper confidence limit}}{\text{Odds ratio}}$$

This can be a useful check on confidence limits presented in tables in published papers.

19.3 GENERAL FORM OF THE LOGISTIC REGRESSION EQUATION

We will now introduce the general form of the logistic regression model with several exposure variables, and explain how it corresponds to what we used above in the simple case when we are comparing two exposure groups, and therefore have a single exposure variable in our model. The general form of the logistic regression model is similar to that for multiple regression (see Chapter 11):

$$\text{log odds of outcome} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

The difference is that we are modelling a transformation of the outcome variable, namely the *log of the odds of the outcome*. The quantity on the right-hand side of the equation is known as the **linear predictor** of the log odds of the outcome, given the particular value of the p exposure variables x_1 to x_p. The β's are the **regression coefficients** associated with the p exposure variables.

The transformation of the probability, or risk, π of the outcome into the log odds is known as the **logit function**:

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$

and the name **logistic** is derived from this. Recall from Section 14.6 (Table 14.2) that while probabilities must lie between 0 and 1, odds can take any value between 0 and infinity (∞). The log odds are not constrained at all; they can take any value between $-\infty$ and ∞.

We will now show how the general form of the logistic regression model corresponds to the logistic regression model we used in Section 19.2 for comparing two exposure groups. The general form for comparing two exposure groups is:

$$\text{log odds of outcome} = \beta_0 + \beta_1 x_1$$

where x_1 (the exposure variable) equals 1 for those in the *exposed* group and 0 for those in the *unexposed* group. Table 19.7 shows the value of the log odds predicted

Table 19.7 Log odds of the outcome according to exposure group, as calculated from the linear predictor in the logistic regression equation.

Exposure group	Log odds of outcome, predicted from model	Log odds of outcome, in terms of the parameter names
Exposed ($x_1 = 1$)	$\beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1$	log(Baseline odds) + log(Exposure odds ratio)
Unexposed ($x_1 = 0$)	$\beta_0 + \beta_1 \times 0 = \beta_0$	log(Baseline odds)

from this model in each of the two exposure groups, together with the log odds expressed in terms of the parameter names, as in Section 19.2.

We can see that the first regression coefficient, β_0, corresponds to the log odds in the unexposed (baseline) group. We will now show how the other regression coefficient, β_1, corresponds to the log of the exposure odds ratio. Since:

$$\text{Exposure OR} = \frac{\text{odds in exposed group}}{\text{odds in unexposed group}}$$

it follows from the rules of logarithms (see p. 156) that:

$$\log \text{OR} = \log(\text{odds in exposed group}) - \log(\text{odds in unexposed group})$$

Putting the values predicted from the logistic regression equation (shown in Table 19.7) into this equation gives:

$$\log \text{OR} = \beta_0 + \beta_1 - \beta_0 = \beta_1$$

The equivalent model on the ratio scale is:

$$\text{Odds of disease} = \exp(\beta_0 + \beta_1 x_1) = \exp(\beta_0) \times \exp(\beta_1 x_1)$$

In this *multiplicative model* $\exp(\beta_0)$ corresponds to the odds of disease in the baseline group, and $\exp(\beta_1)$ to the exposure odds ratio. Table 19.8 shows how this model corresponds to the model shown in Table 19.1.

Table 19.8 Odds of outcome according to exposure group, as calculated from the linear predictor in the logistic regression equation.

Exposure group	Odds of outcome, predicted from model	Odds of outcome, in terms of the parameter names
Exposed ($x_1 = 1$)	$\exp(\beta_0) \times \exp(\beta_1)$	Baseline odds $\times$ Exposure odds ratio
Unexposed ($x_1 = 0$)	$\exp(\beta_0)$	Baseline odds

19.4 LOGISTIC REGRESSION FOR COMPARING MORE THAN TWO EXPOSURE GROUPS

We now consider logistic regression models for **categorical exposure variables** with more than two levels. To examine the effect of categorical variables in logistic and other regression models, we look at the effect of each level compared to a **baseline** group. When the exposure is an *ordered* categorical variable, it may also be useful to examine the average change in the log odds per exposure group, as described in Section 19.5.

Table 19.9 Association between age group and microfilarial infection in the onchocerciasis study.

Age group (years)	Coded value in dataset	Microfilarial infection		Odds of infection	Odds ratio compared to the baseline group
		Yes	No		
5–9	0	46	156	$46/156 = 0.295$	1
10–19	1	99	119	$99/119 = 0.832$	$0.832/0.295 = 2.821$
20–39	2	299	125	$299/125 = 2.392$	$2.392/0.295 = 8.112$
≥ 40	3	378	80	$378/80 = 4.725$	$4.725/0.295 = 16.02$
Total		822	480		

Example 19.2

In the onchocerciasis study, introduced in Example 19.1, subjects were classified into four age groups: 5–9, 10–19, 20–39 and ≥ 40 years. Table 19.9 shows the association between age group and microfilarial infection. The odds of infection increased markedly with increasing age. A chi-squared test for association in this table gives $P < 0.001$, so there is clear evidence of an association between age group and infection. We chose the 5–9 year age group as the **baseline** exposure group, because its coded value in the dataset is zero, and calculated odds ratios for each non-baseline group relative to the baseline group.

The corresponding logistic regression model uses this same approach; the effect of each non-baseline age group is expressed in terms of the odds ratio comparing it with the baseline. The parameters of the model, on both the odds and log odds scales, are shown in Table 19.10.

Table 19.10 Odds and log odds of the outcome in terms of the parameters of a logistic regression model comparing four age groups.

Age group	Odds of infection	Log odds of infection
0 (5–9 years)	Baseline	Log(Baseline)
1 (10–19 years)	Baseline × Agegrp(1)	Log(Baseline) + Log(Agegrp(1))
2 (20–39 years)	Baseline × Agegrp(2)	Log(Baseline) + Log(Agegrp(2))
3 ($\geq$ 40 years)	Baseline × Agegrp(3)	Log(Baseline) + Log(Agegrp(3))

Here, Agegrp(1) is the odds ratio (or, on the log scale, the log odds ratio) comparing group 1 (10–19 years) with group 0 (5–9 years, the baseline group), and so on. This regression model has four parameters:

1 the odds of infection in the 5–9 year group (the baseline group); and

2 the *three* odds ratios comparing the non-baseline groups with the baseline.

Using the notation introduced in Section 19.2, the four equations for the odds that define the model in Table 19.10 can be written in abbreviated form as:

$$\text{Odds} = \text{Baseline} \times \text{Agegrp}$$

or on a log scale, as:

$$\log(\text{Odds}) = \text{Baseline} + \text{Agegrp}$$

The effect of categorical variables is modelled in logistic and other regression models by using **indicator variables**, which are created automatically by most statistical packages when an exposure variable is defined as categorical. This is explained further in Box 19.1. Output from this model (expressed on the odds ratio scale, with the constant term omitted) is shown in Table 19.11.

Table 19.11 Logistic regression output (odds ratio scale) for the association between microfilarial infection and age group.

	Odds ratio	z	$P > \|z\|$	95% CI
agegrp(1)	2.821	4.802	0.000	1.848 to 4.308
agegrp(2)	8.112	10.534	0.000	5.495 to 11.98
agegrp(3)	16.024	13.332	0.000	10.658 to 24.09

BOX 19.1 USE OF INDICATOR VARIABLES IN REGRESSION MODELS

To model the effect of an exposure with more than two categories, we estimate the odds ratio for each non-baseline group compared to the baseline. In the logistic regression equation, we represent the exposure by a set of **indicator variables** (variables which take only the values 0 and 1) representing each non-baseline value of the exposure variable. The regression coefficients for these indicator variables are the corresponding (log) odds ratios. For example, to estimate the odds ratios comparing the 10–19, 20–39 and ≥ 40 year groups with the 5–9 year group, we create three indicator variables which we will call ageind_1, ageind_2 and ageind_3 (the name is not important). The table below shows the value of these indicator variables according to age group.

Value of indicator variables for use in logistic regression of the association between microfilarial infection and age group.

Age group	ageind_1	ageind_2	ageind_3
0 (5–9 years)	0	0	0
1 (10–19 years)	1	0	0
2 (20–29 years)	0	1	0
3 (≥ 40 years)	0	0	1

All three of these indicator variables (but not the original variable) are then included in a logistic regression model. Most statistical packages create the indicator variables automatically when the original variable is declared as categorical.

The *P*-values for the three indicator variables (corresponding to the non-baseline age groups) can be used to test the null hypotheses that there is no difference in odds of the outcome between the individual non-baseline exposure groups and the baseline group. However, these are not usually of interest: we need a test, analogous to the χ^2 test for a table with four rows and two columns, of the general null hypothesis that there is no association between age group and infection. We will see how to test such null hypotheses in regression models in Chapter 29, and in the next section we address the special case when the categorical variable is ordered, as is the case here. It is usually a mistake to conclude that there is a difference between one exposure group and the rest based on a particular (small) *P*-value corresponding to one of a set of indicator variables.

19.5 LOGISTIC REGRESSION FOR ORDERED AND CONTINUOUS EXPOSURE VARIABLES

Until now, we have considered logistic regression models for binary or categorical exposure variables. For binary variables, logistic regression estimates the odds ratio comparing the two exposure groups, while for categorical variables we have seen how to estimate odds ratios for each non-baseline group compared to the baseline. This approach does not take account of ordering of the exposure variable. For example, we did not use the fact that subjects aged ≥ 40 years are older than those aged 20–39 years, who in turn are older than those aged 10–19 years and so on.

Example 19.3

The odds of microfilarial infection in each age group in the onchocerciasis dataset are shown in Table 19.9 in Section 19.4, and are displayed in Figure 19.1. We do not have a straight line; the slope of the line increases with increasing age group. In other words, this increase in the odds of infection with increasing age does not appear to be constant.

However, Figure 19.2 shows that there *is* an approximately linear increase in the **log odds** of infection with increasing age group. This log-linear increase means that we are able to express the association between age and the log odds of microfilarial infection by a single linear term (as described below) rather than by a series of indicator variables representing the different groups.

Relation with linear regression models

Logistic regression models can be used to estimate the *most likely* value of the increase in log odds per age group, assuming that the increase is the same in each age group. (We will define the meaning of 'most likely' more precisely in Chapter

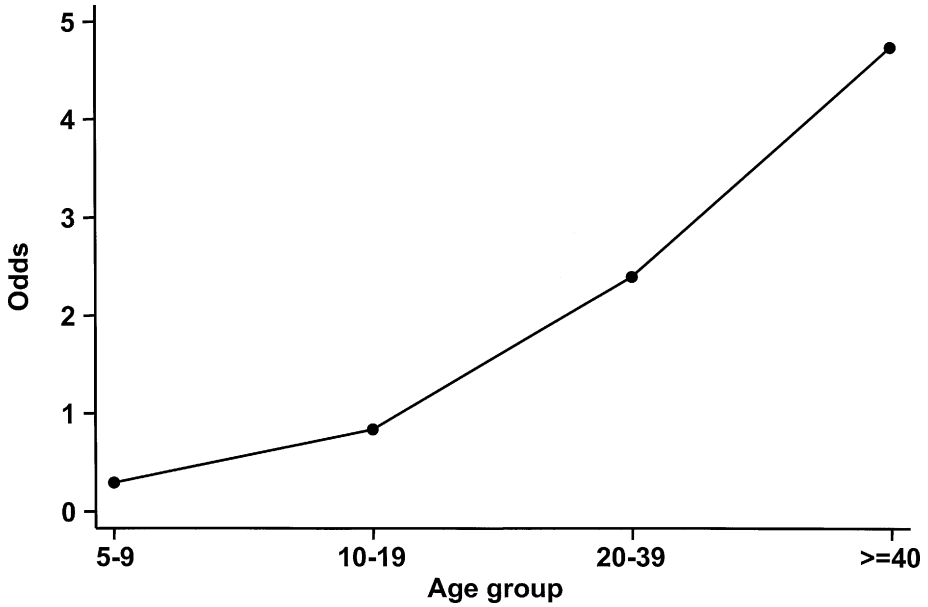

Fig. 19.1 Odds of microfilarial infection according to age group for the onchocerciasis data.

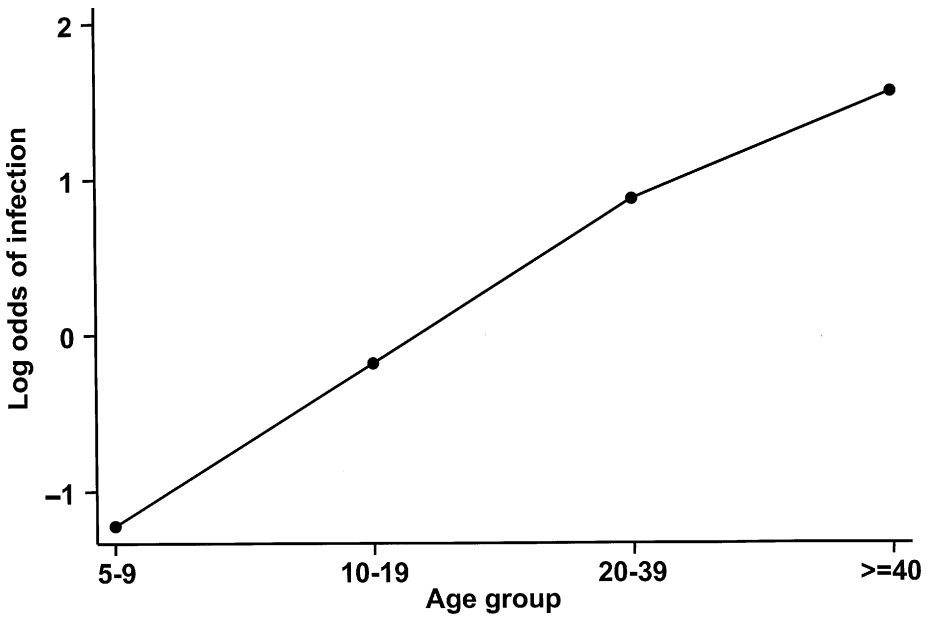

Fig. 19.2 Log odds of microfilarial infection according to age group for the onchocerciasis data.

28.) The model is analogous to the simple linear regression model described in Chapter 11. If we assume that:

$$y = \beta_0 + \beta_1 x$$

then the intercept β_0 is the value of y when $x = 0$, and the slope β_1 represents the increase in y when x increases by 1. Logistic regression models assume that:

$$\log \text{ odds} = \beta_0 + \beta_1 x$$

so that the intercept β_0 is the value of the log odds when $x = 0$, and the slope β_1 represents the increase in log odds when x increases by 1. We will use the notation

$$\log \text{ odds} = \text{Baseline} + [\text{X}]$$

where the square brackets indicate our assumption that variable X has a linear effect on the log odds of the outcome. For the onchocerciasis data, our model is

$$\log \text{ odds} = \text{Baseline} + [\text{Agegrp}]$$

Example 19.3 (continued)

Table 19.12(a) shows logistic regression output for the model assuming a linear effect of logistic regression on the log odds of microfilarial infection. The estimated increase in log odds for every unit increase in age group is 0.930 (95% CI = 0.805 to 1.055). This corresponds to an odds ratio per group of 2.534 (95% CI = 2.236 to 2.871; see output in Table 19.12b). The constant term corresponds to the estimated log odds of microfilarial infection in age group 0 (5–9 years, log odds = −1.115), *assuming a linear relation* between age group and the log odds of infection. It does not therefore numerically equal the baseline term in the

Table 19.12 Logistic regression output for the linear association between the log odds of microfilarial infection and age group (data in Table 19.9).

(a) Output on log scale.

	Coefficient	s.e.	z	$P > \|z\|$	95% CI
Age group	0.930	0.0638	14.587	0.000	0.805 to 1.055
Constant	−1.115	0.127	−8.782	0.000	−1.364 to −0.866

(b) Output on ratio scale.

	Odds ratio	z	$P > \|z\|$	95% CI
Age group	2.534	14.587	0.000	2.236 to 2.871

Table 19.13 Predicted log odds in each age group, derived from a logistic regression model assuming a linear relationship between the log odds of microfilarial infection and age group.

Age group	Logistic regression equation	Predicted log odds
0	log odds = constant + 0 × age group	−1.115 + 0.930 × 0 = −1.115
1	log odds = constant + 1 × age group	−1.115 + 0.930 × 1 = −0.185
2	log odds = constant + 2 × age group	−1.115 + 0.930 × 2 = 0.745
3	log odds = constant + 3 × age group	−1.115 + 0.930 × 3 = 1.674

regression equation when age is included as a categorical variable, as described in Section 19.4.

Substitution of the estimated regression coefficients into the logistic regression equation gives the **predicted log odds** in each age group. These are shown in Table 19.13. Figure 19.3 compares these predicted log odds from logistic regression with the observed log odds in each group. This shows that the linear assumption gives a good approximation to the observed log odds in each group. Section 29.6 describes how to test such linear assumptions.

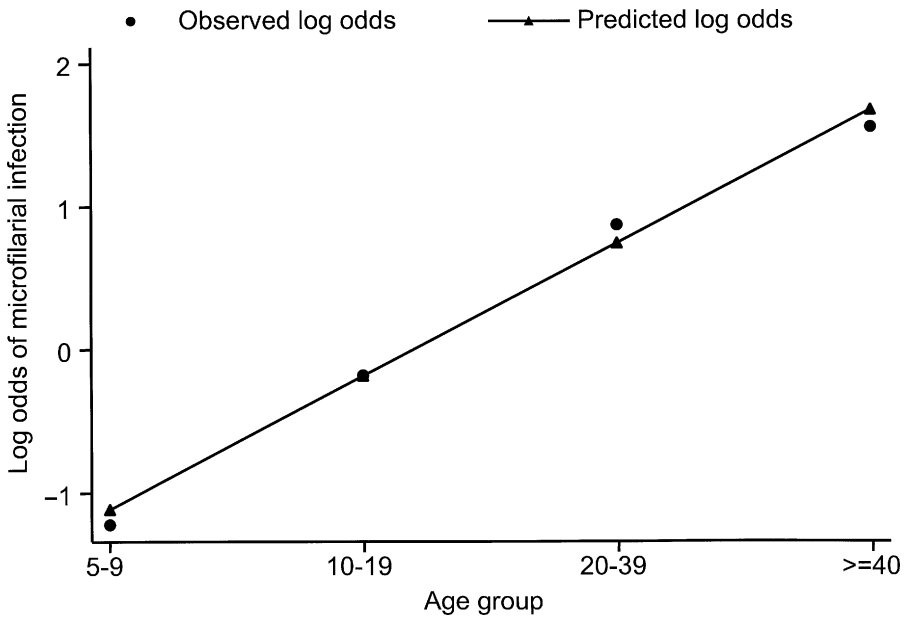

Fig. 19.3 Observed log odds in each age group (circles) and predicted log odds from logistic regression (triangles, connected by line).