16

LOGISTIC REGRESSION

OBJECTIVES

After reading this chapter, you should be able to:

1. Understand logistic regression.
   a. Understand log odds as a measure of disease, and how it relates to a linear combination of predictors.

2. Build and interpret logistic regression models.
   a. Compute and interpret odds ratios derived from a logistic regression model.
   b. Evaluate the effects of predictors on the outcome of interest on a probability scale.
   c. Statistically compare logistic models using both Wald tests and likelihood ratio tests.

3. Understand how logistic regression fits in the family of generalised linear models.

4. Evaluate logistic regression models.
   a. Understand covariate patterns, and how they impact the computation of residuals for logistic regression models.
   b. Understand overdispersion, and how it relates to goodness-of-fit tests.
   c. Compute residuals on the basis of one per covariate pattern, and one per observation.
   d. Select and use the appropriate test(s) to evaluate the goodness of fit of a logistic model.
   e. Determine the effect of changing the threshold (‘cutpoint’) on the sensitivity and specificity of the model.
   f. Generate ROC curves as a method of evaluating the goodness of fit.
   g. Identify and determine the impact of influential observations on a logistic model.

5. Fit a model to a small dataset using exact logistic regression.

6. Fit conditional logistic regression models for matched data.

16.1 INTRODUCTION

As epidemiologists, we often find ourselves in a situation in which the outcome in our study is dichotomous (*ie* Y=0 or 1). Most commonly, this variable represents either the absence or presence of disease or mortality. We can't use linear regression techniques to analyse these data as a function of a set of linear predictors $X=(X_j)$ for the following reasons.

(a) The error terms (ε) are not normally (Gaussian) distributed. In fact, they can only take on 2 values.

$$\text{if } Y=1 \text{ then } \varepsilon = 1-(\beta_0+\sum \beta_j X_j)$$
$$\text{if } Y=0 \text{ then } \varepsilon = -(\beta_0+\sum \beta_j X_j) \qquad \textit{Eq 16.1}$$

(b) The probability of the outcome occurring (*ie* p(Y=1)) depends on the values of the predictor variables (*ie* X). Since the variance of a binomial distribution is a function of the probability (p), the error variance will also vary with the level of X, and consequently the assumption of homoscedasticity will be violated.

(c) The mean responses should be constrained as:

$$0 \le \mathrm{E}(Y) = p \le 1$$

However, with a linear regression model, the predicted values might fall outside of these constraints.

In this chapter, we will explore the use of logistic regression to avoid the problems identified above. The birth weight data used extensively in the previous two chapters will be the primary dataset used in this chapter, but the outcome will be 'low birth weight'. Babies born weighing less than 2,500 gm will be classified as having low birth weight, and factors which influence the probability of this occurring will be investigated. Of the 5,000 observations in this dataset, 371 were classified as low birth weight. Details of the dataset can be found in Chapter 31.

Table 16.1 Selected variables from the low birth weight dataset used in this chapter

Variable	Description
low_bw	birth weight (1=birth weight <2500 gm, 0=weight ≥2500 gm)
smk	smoking (1=smoked during 2nd trimester, 0=non-smoker)
white	mother's race (1=white, 0=all other races)
frace_c3	father's race (1=hispanic, 2=white, 3=black)
previs	number of prenatal visits

16.2 THE LOGISTIC MODEL

One way of getting around the problems described in Section 16.1 is to use a logit transform of the probability of the outcome and model this as a linear function of a set of predictor variables,

$$\ln\left[\frac{p}{1-p}\right] = \beta_0+\sum \beta_j X_j \qquad \textit{Eq 16.2}$$

where ln(p/(1-p)) is the logit transform. This value is the log of the odds of the outcome (because odds=p/(1-p)), so a logistic regression model is sometimes referred to as a log odds model.

Fig. 16.1 shows that, while the logit of p might become very large or very small, p does not go beyond the bounds of 0 and 1. In fact, logit values tend to remain between -7 and +7 as these are associated with very small (<0.001) and very large (>0.999) probabilities, respectively.

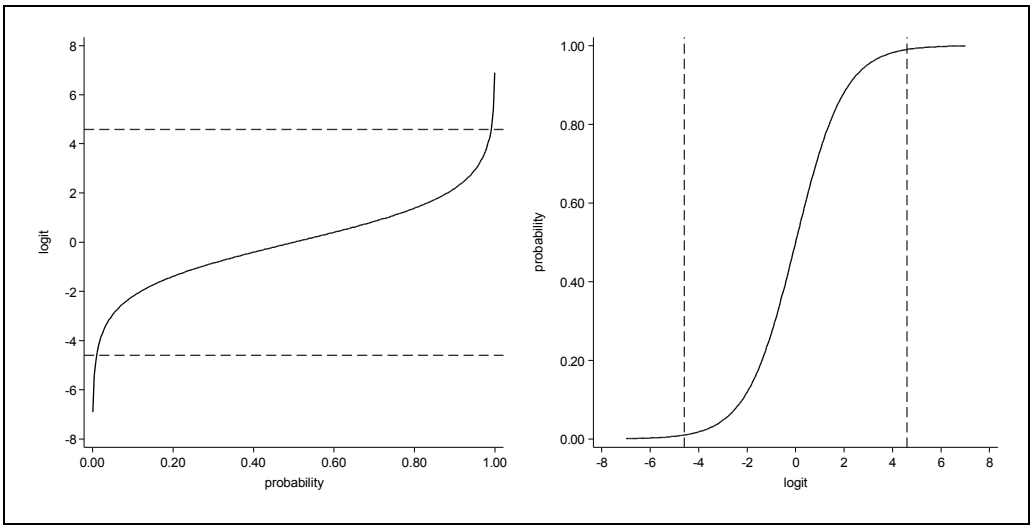

Fig. 16.1 Logit and inverse logit functions

Note Dashed lines are at ± 4.595 which is the logit of 1% and 99%

This transformation leads to the logistic model in which the probability of the outcome can be expressed in 1 of the 2 following ways (they are equivalent).

$$p = \frac{1}{1+e^{-(\beta_0+\sum \beta_j X_j)}} = \frac{e^{(\beta_0+\sum \beta_j X_j)}}{1+e^{(\beta_0+\sum \beta_j X_j)}} \qquad \textit{Eq 16.3}$$

16.3 Odds and odds ratios

Let's look at the simple situation in which the occurrence of disease is the event of interest (Y=0 or 1) and we have a single dichotomous predictor variable (*ie* X=0 or 1). The logistic model is:

$$\ln\left[\frac{p}{1-p}\right] = \beta_0+\beta_1 X_1 \qquad \textit{Eq 16.4}$$

so the odds of disease is:

$$\text{odds} = \frac{p}{1-p} = e^{\beta_0+\beta_1 X} \qquad \textit{Eq 16.5}$$

From this it is a relatively simple process to determine the odds ratio (OR) for disease that is associated with the presence of factor 'X'.

$$\text{if} \quad X=1 \quad \text{odds} = e^{\beta_0+\beta_1}$$
$$\text{if} \quad X=0 \quad \text{odds} = e^{\beta_0}$$

The odds ratio is then:

$$OR = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = \frac{e^{\beta_0}e^{\beta_1}}{e^{\beta_0}} = e^{\beta_1} \qquad \textit{Eq 16.6}$$

This can be extended to the situation in which there are multiple predictors and the *OR* for the k^{th} variable will be e^{β_k}.

16.4 Fitting a logistic regression model

In linear regression, we used least squares techniques to estimate the regression coefficients (or at least the computer did this for us). Because the error term has a Gaussian distribution, this approach produces maximum likelihood estimates of the coefficients. In a logistic model, we use a different maximum likelihood estimation procedure to estimate the coefficients.

The key feature of maximum likelihood estimation is that it estimates values for parameters (the βs) which are most likely to have produced the data that have been observed. Rather than starting with the observed data and computing parameter estimates (as is done with least squares estimates), one determines the likelihood (probability) of the observed data for various combinations of parameter values. The set of parameter values that was most likely to have produced the observed data is that of the maximum likelihood (ML) estimates.

The following simple example demonstrates the maximum likelihood estimation process. Assume that you have a set of serologic results from a sample of 10 students in a high school class and the parameter you want to estimate is the prevalence of the disease. Three of the 10 samples are positive (these are the observed data).

The likelihood (L) of getting 3 positive results from 10 students if the true prevalence is *P* is:

$$\mathrm{L}(P) = \binom{10}{3} P^3(1-P)^7$$

The log likelihood (lnL) is:

$$\ln\mathrm{L}(P) = \ln\left\{\binom{10}{3}\right\} + 3\ln(P) + 7\ln(1-P)$$

In this situation, the maximum value of the lnL can be determined directly, but in many cases an iterative approach is required. If such a procedure was being followed, the steps would be:

(a) Pick a value for the prevalence (perhaps your first guess is 0.2). The likelihood of observing 3 positive students out of 10, if the true prevalence (*P*) is 0.2, is:

$$\mathrm{L}(0.2) = \binom{n}{x} P^x(1-P)^{n-x} = \binom{10}{3} 0.2^3(1-0.2)^{10-3} = 0.201 \qquad \textit{Eq 16.7}$$

The lnL is -1.60.

(b) Pick another prevalence (perhaps your next guess is 0.35) and recompute the likelihood. This turns out to be 0.252 (lnL=-1.38).

(c) Keep repeating this process until you have the estimate of the parameter that gives you the highest likelihood (*ie* maximum likelihood). This would occur at P=0.3 (but you already knew that, didn't you?).

A graph of the relationship between lnL and prevalence (Fig. 16.2) shows the maximum value at P=0.3.

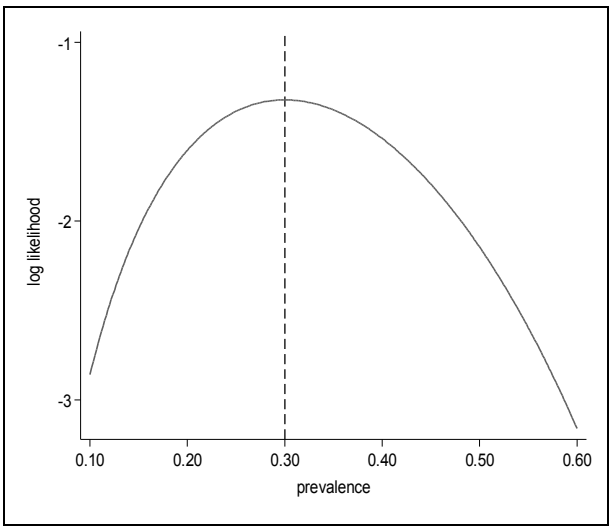

Fig. 16.2 Log likelihood versus prevalence

Of course, the computer doesn't just pick values of parameters at random; there are ways of estimating what the parameter is likely to be and then refining that estimate. Since it is possible to keep refining the estimates to more and more decimal places, you have to specify the **convergence criterion**. Once the estimates change by less than the convergence criterion, the process of refining the estimates is stopped (*ie* convergence has been achieved).

16.5 Assumptions in logistic regression

As with linear regression, there are a number of assumptions inherent in fitting a logistic model. In a logistic model, the outcome Y is dichotomous:

$$Y_i \begin{cases} 1 \\ 0 \end{cases} \qquad p(Y_i=1)=p_i=1-p(Y_i=0) \qquad \textit{Eq 16.8}$$

and 2 important assumptions are independence and linearity.

Independence It is assumed that the observations are independent from each other (the same assumption was made in linear regression). If the data come from people who are in some way clustered, or if multiple measurements are being made on the same individual, this assumption has probably been violated. For example, if data come from patients in multiple clinics, variation between patients in the study population results from the usual variation between patients plus the variation that is due to differences between clinics. This often results in 'over-dispersion' or 'extra-binomial variation' in the data. Some methods of checking this assumption will be presented in Section 16.12.4, and methods of dealing with the problem are discussed in Chapters 20–23.

Linearity As with linear regression, any predictor that is measured on a continuous scale is assumed to have a linear (straight-line) relationship with the outcome. Techniques for evaluating this assumption are presented in Section 15.6.

Note Because the logistic model models the expected probability of disease on the logit scale but the original data are binary (0/1 or no/yes), the logistic model does not have an error term. Consequently, there is no assumption about the distribution of errors. It also means that coefficients in a logistic model represent the effect of a predictor on the logit of the outcome. Presenting effects on the original probability scale is discussed in Section 16.8.5.

16.6 Likelihood ratio statistics

Although the maximum likelihood estimation process produces the largest possible (*ie* maximum) likelihood value, these values are always very, very small, because they are describing the probability of an exact set of observations given the parameter estimates selected. Because of this (and the fact that the estimation process is simpler), computer programs usually work with the log likelihood which will be a moderately sized negative number. Most computer programs print out the log likelihood of the model that has been fit to the data. It is a key component in testing logistic regression models.

16.6.1 Significance of the full model

The test used to determine the overall significance of a logistic model is called the **likelihood ratio test** (*LRT*), as it compares the likelihood of the 'full' model (*ie* with all the predictors included) with the likelihood of the 'null' model (*ie* a model which contains only the intercept). Consequently, it is analogous to the overall *F*-test of the model in linear regressions. The formula for the likelihood ratio test statistic (G^2_0) is:

$$G_0^2 = 2\ln\frac{L}{L_0} = 2(\ln L - \ln L_0) \qquad \textit{Eq 16.9}$$

where L is the likelihood of the full model and L_0 is the likelihood of the null model. The statistic (G^2_0) has an approximate χ^2 distribution with k degrees of freedom (df) (k=number of predictors in the full model). If significant, it suggests that, taken together, the predictors contribute significantly to the prediction of the outcome.

Note When computing an *LRT* statistic, 2 conditions must be met.

1. Both models must be fit using exactly the same observations. If a dataset contains missing values for some predictors in the full model, then these would be omitted from the full model but included when the null model is computed. This must be avoided.
2. The models must be **nested**. This means that the predictors in the simpler model must be a subset of those in the full model. This will not be a problem when the smaller model is the null model, but might be a problem in other situations.

In Example 16.1, a logistic regression model for low birth weight has been fit with 3 predictor variables (-smk-, -white-, -frace_c3-). The likelihood ratio test evaluating the 3 predictors as a group is highly statistically significant ($G^2_0 = 26.47, \text{df} = 4, P < 0.001$).

16.6.2 Comparing full and reduced models

In the preceding section, the *LRT* was used to compare the full and null models, but an *LRT* can also be used to test the contribution of any subset of parameters in much the same way as a

Example 16.1 Comparing logistic regression models
data = bw5k

The log likelihoods from 4 different models were:

Model	Predictors	# of predictors	Log likelihood
null	intercept β_0	1	-1321.85
full	intercept, smk, white, frace_c3 $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$	5	-1308.61
reduced	intercept, smk β_0, β_1	2	-1318.03
saturated	5000 'hypothetical' predictors $\beta_0, \beta_1 \ldots \beta_{n-1}$	5000	0

Overall likelihood ratio test of the full model:
$G_0^2 = 2(-1308.61 - (-1321.85)) = 26.47$ with 4 df (P <0.001)
Taken together, the 3 predictors are highly significant predictors of low birth weight.

Likelihood ratio test comparing the full and reduced models:
$G_0^2 = 2(-1308.61 - (-1318.03)) = 18.83$ with 3 df (P <0.001)
The 2 race predictors (-white- and -frace_c3-) are highly significant predictors.

Likelihood ratio test comparing the saturated and full models:
$G_0^2 = 2(0 - (-1308.61)) = 2617.22$ with 5000 df.
Note This does not have a χ^2 distribution.

multiple partial *F*-test is used in linear regression. The formula is:

$$G_0^2 = 2\ln\frac{L_{full}}{L_{red}} = 2(\ln L_{full} - \ln L_{red}) \qquad \textit{Eq 16.10}$$

where L_{full} and L_{red} refer to the likelihood of the full and reduced models, respectively. As can be seen in Example 16.1, the 2 race predictors (-white-, -frace_c3-) are highly significant predictors of low birth weight. This test is sometimes referred to as the 'improvement χ^2'.

16.6.3 Comparing full and saturated models (deviance)

A special case of the likelihood ratio test is the comparison of the likelihood of the model under investigation to the likelihood of a fully saturated model (1 in which there would be 1 parameter fit for each data point). Since a fully saturated model should perfectly predict the data, the likelihood of the observed data, given this model, should be 1 (or $\ln L_{sat}=0$). This comparison yields a statistic called the **deviance** which is analogous to the **error sum of squares** (SSE) in linear regression. The deviance is a measure of the unexplained variation in the data.

$$D = 2\ \ln\frac{L_{sat}}{L_{full}} = 2(\ln L_{sat} - \ln L_{full}) = -2(\ln L_{full}) \qquad \textit{Eq 16.11}$$

Note The deviance computed in this manner does not have a χ^2 distribution. (See Section 16.12.2 for more discussion of deviance.)

16.7 Wald tests

An alternative approach to evaluating the significance of a single coefficient is to use a test that relates the coefficient to its SE. A Wald test is the ratio of the coefficient to its SE and it follows (asymptotically) a standard normal (Z) distribution. This tests whether the coefficient is significantly different from zero. It is routinely computed by most computer programs and is the most widely used test of the significance of coefficients. However, the estimates of the coefficient and its SE are only estimates, and consequently the normal approximation of its distribution might not be reliable particularly if the sample size is small. To evaluate the significance of variables with a P-value close to the rejection region, it is best to use a likelihood ratio test.

Just as with multiple partial F-tests in linear regression, multiple parameters in a logistic model can be tested with a multiple Wald test. For example, comparing the full and reduced models in Example 16.1 would be equivalent to testing the null hypothesis:

$$H_0 : \beta_2 = \beta_3 = 0$$

In this case, the test statistic is compared with a χ^2 distribution, with the df equal to the number of predictors being tested. In Example 16.1, the Wald χ^2 for comparing the full and reduced models has a value of 20.1 and 3 df. This is a slightly larger test statistic (although this is not always the case) than the likelihood ratio test (χ^2 =18.83), but it is still highly significant.

16.8 Interpretation of coefficients

The coefficients in a logistic regression model represent the amount the logit of the probability of the outcome changes with a unit increase in the predictor. Unfortunately, this is hard to interpret so we usually convert the coefficients into odds ratios. The following sections are based on the model shown in Example 16.2.

$$\ln\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1(\text{smk}) + \beta_2(\text{white}) + \beta_3(\text{frace=hisp}) + \beta_4(\text{frace=black}) + \beta_5(\text{previs})$$

16.8.1 Dichotomous predictor

Coefficients for a dichotomous predictor represent the amount that the log odds of disease increase (or decrease) when the factor is present. These can be easily converted into *OR* by exponentiating the coefficient. For example, the *OR* for -smk- in Example 16.2 is:

$$OR = e^{\beta_1} = e^{0.527} = 1.695$$

If the outcome of interest is relatively rare, the *OR* provides a good approximation of the risk ratio (*RR*). If the data come from a case-control study that used incidence density sampling, the *OR* is a good estimate of the incidence rate ratio (*IR*) in the original population (see Chapter 6).

Example 16.2 Interpreting logistic regression coefficients
data = bw5k

The tables below present results from a logistic regression of -low_bw- on -smk-, -white-, -previs-, and 2 levels of -frace_c3-. The first table presents the effects of the predictors on the logit of the outcome (low birth weight), while the second shows the same results expressed as odds ratios.

Number of obs = 5000
LR chi2 (5) = 42.65
Prob > chi2 = 0.000
Log likelihood = -1300.5253

Predictor	Coef	SE	Z	P	95% CI	
smk	0.527	0.183	2.88	0.004	0.169	0.886
white	-0.321	0.180	-1.78	0.075	-0.673	0.032
frace = hisp	-0.433	0.201	-2.15	0.031	-0.827	-0.039
frace = black	0.196	0.193	1.02	0.308	-0.181	0.574
previs	-0.059	0.015	-4.01	0.000	-0.088	-0.030
constant	-1.673	0.235	-7.11	0.000	-2.134	-1.212

Predictor	OR	SE	95% CI	
smk	1.695	0.310	1.184	2.426
white	0.726	0.131	0.510	1.032
frace = hisp	0.648	0.130	0.437	0.962
frace = black	1.217	0.235	0.834	1.776
previs	0.943	0.014	0.916	0.970

Effect of -smk- Smoking increased the log odds of low birth weight by 0.527, or alternatively, it increased the odds of having a low birth weight baby by 1.7 times. Since low birth weight is a relatively rare condition, it would be reasonable to interpret the odds ratio as a risk ratio and state that smoking increased the risk of having a low birth weight baby by approximately 1.7 times (equivalent to a 70% increase).

Effect of -previs- Increasing the number of prenatal visits from 5 to 15 reduced the log odds of disease by: (15-5)*0.059=0.59 units. Alternatively, it reduces the odds of disease by the factor: $(0.943)^{(15-5)}$=0.556. An increase of 10 in the number of prenatal visits reduces the risk of low birth weight by a factor of approximately 0.56 (equivalent to a 44% reduction).

Effect of -frace_c3- Compared with whites (baseline group), babies from Hispanic fathers had decreased odds of being low birth weight (*OR*=0.65) while those with black fathers had increased odds (*OR*=1.22), although this latter difference was not statistically significant (P=0.308). Collectively, there were significant differences among the father's race groups (P=0.0006, data not shown) and individually, Hispanics had significantly lower odds than whites (P=0.031) and blacks (P<0.001, data not shown).

16.8.2 Continuous predictor

For a continuous predictor, the coefficient (*eg* β_5) represents the change in the log odds of disease for a 1-unit change in the predictor. Similarly, the computed *OR* represents the factor by

which the odds of disease are multiplied for each 1-unit change in the predictor. However, we are often interested in changes of multiple units of the exposure variable(s), such as from x_1 to x_2. For example, for a change from 5 to 15 in the number of prenatal visits, the log odds of disease changes by:

$$\log \text{odds}(x_1, x_2) = (x_2 - x_1) * \beta_5 = (15 - 5) * -0.059 = -0.59 \qquad \textit{Eq 16.12}$$

For this 10-unit change in -previs-, the odds of disease change by:

$$e^{-0.59} = 0.554, \quad \text{or} \quad OR(x_1, x_2) = OR^{(x_2 - x_1)} = 0.943^{(15-5)} = 0.556 \qquad \textit{Eq 16.13}$$

Note This effect of -previs- is based on the assumption that the relationship between -previs- and the log odds of low birth weight is linear. Methods for evaluating this assumption were discussed in Section 15.6.

16.8.3 Categorical predictor

As in linear regression, predictors with multiple categories (*eg* '*j*' categories) must be converted to a series of indicator variables (also called 'dummy' variables) with *j*-1 variables put into the model. The coefficient for each indicator variable represents the effect of that level compared with the category (*ie* the 'baseline') not included in the model. The coefficients are interpreted in the same manner as for any other dichotomous predictor.

Note There are other ways of coding categorical variables, such as hierarchical indicator variables, and these are used in the same way as described in Chapter 14.

When creating indicator variables, the choice of the baseline might be important. In general, we choose one that makes biological sense (*ie* makes some sense as a reference level) and one that has a reasonable number of observations so we are not comparing everything with a category for which the effect can only be estimated very imprecisely. When evaluating the statistical significance of coefficients for categorical variables, it is important NOT to pay much attention to the P-values of individual coefficients. This P-value indicates whether or not the chosen level is statistically different from the baseline level. However, because the choice of the baseline is arbitrary, any category has a range of possible P-values that could be computed. Instead, you should evaluate the statistical significance of all of the categories together with a multiple Wald test or a likelihood ratio test.

In Example 16.2, the variable -frace_c3- was converted to a series of 3 dummy variables and 2 of these (-frace_c3_1-, -frace_c3_3-) were included in the model. These represented Hispanics and blacks, respectively. Consequently, the coefficients represent the effects of these races on the log odds of low birth weight compared with whites (the category that was omitted).

16.8.4 Interpretation of the intercept

Interpretation of the intercept (constant) in the regression model depends on how the data were collected. The intercept represents the logit of the probability of disease if all of the 'risk factors' are absent (*ie* equal to zero). This can be expressed as:

$$\ln\left(\frac{p_0}{1-p_0}\right) = \beta_0 \qquad \textit{Eq 16.14}$$