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 (GLMs).

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

In veterinary epidemiology, we are often in the situation where 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.

$$\begin{aligned}&\text{if } Y=1 \text{ then } \varepsilon=1-(\beta_0+\sum\beta_jX_j)\\&\text{if } Y=0 \text{ then } \varepsilon=-(\beta_0+\sum\beta_jX_j)\end{aligned}$$

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:

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

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 primary dataset used in the examples in this chapter is one derived from a case-control study of *Nocardia* spp mastitis that was carried out during an outbreak of this disease in dairy herds in Nova Scotia, Canada. The data consist of observations from 54 case herds and 54 control herds. The predictors of interest were primarily related to the management of the cows during the dry period and, in particular, the use of specific types of dry-cow mastitis treatment. The variables used in this chapter are presented in Table 16.1. Details of the dataset can be found in Chapter 31.

Table 16.1 Selected variables from the Nocardia dataset

Variable	Description
casecont	Case or control status of the herd (the outcome)
dcpct	Percentage of cows treated with dry-cow treatments
dneo	Use of neomycin-based dry-cow products in the last year (yes/no)
dclox	Use of cloxacillin-based dry-cow products in the last year (yes/no)
dbarn	Categorical variable for barn type (1=freestall; 2=tiestall, 3=other)

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.

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

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.

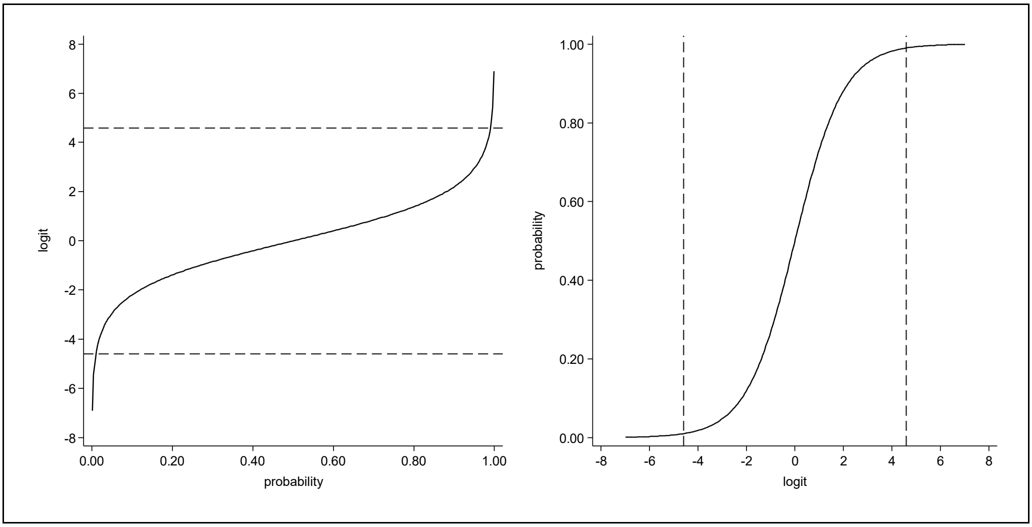

Fig. 16.1 Logit and inverse logit functions

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

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.

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

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

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:

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

Eq 16.4

so the odds of disease is:

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

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:

$$OR = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = \frac{e^{\beta_0}e^{\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

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 *β*s) 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 are the maximum likelihood (ML) estimates.

The following is a very simple example which demonstrates the maximum likelihood estimation process. Assume that you have a set of serologic results from a sample of 10 cows from a dairy herd 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 cows if the true prevalence is *P* is:

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

The log likelihood (lnL) is:

$$1nL(P) = 1n\left\{\binom{10}{3}\right\} + 3\,1n(P) + 7\,1n(1-P)$$

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 probability of observing 3 positive cows out of 10, if the true prevalence (*P*) is 0.2, is:

$$L(0.2) = \binom{n}{x} P^x (1-P)^{n-x} = \binom{10}{3} 0.2^3 (1-0.2)^{10-3} = 0.201$$

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.

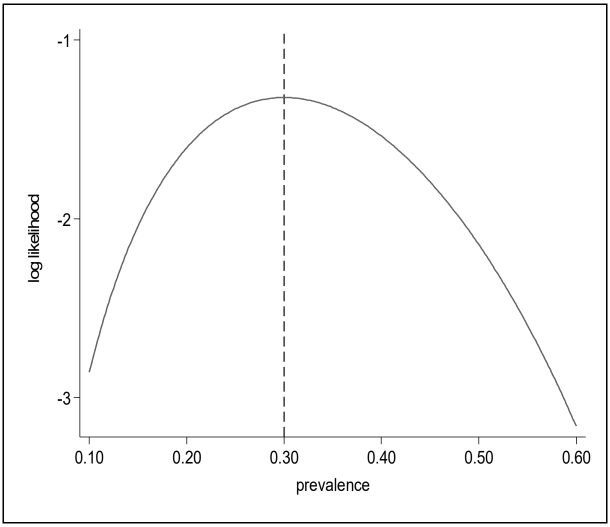

Fig. 16.2 Log likelihood versus prevalence

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:

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

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 animals are maintained in groups or, if multiple measurements are being made on the same individual, this assumption has probably been violated. For example, if animals are kept in herds, variation between animals in the study population results from the usual variation between animals plus the variation that is due to differences between herds. 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 15.6.1.

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 and 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:

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

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 from the case-control study of *Nocardia* spp mastitis has been fit with 3 predictor variables (-dneo-, -dclox-, -dcpct-). The likelihood ratio test evaluating the 3 predictors as a group is highly statistically significant $(G_0^2 = 41.72, \text{df} = 3, P < 0.001)$.

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

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

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 antibiotic specific predictors (-dneo-, -dclox-) are highly significant predictors of case-control status. 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 (one in which there would be one

Example 16.1 Comparing logistic regression models
data = Nocardia

The log likelihoods from 4 different models were:

Model	Predictors	# of predictors	Log likelihood
null	intercept β_0	1	-74.86
full	intercept, dcpct, dneo, dclox $\beta_0, \beta_1, \beta_2, \beta_3$	4	-54.00
reduced	intercept, dcpct β_0, β_1	2	-69.07
saturated	108 'hypothetical' predictors $\beta_0, \beta_1 \ldots \beta_{n-1}$	108	0

Overall likelihood ratio test of the full model:

G_0^2 = 2(-54.00 - (-74.86)) = 41.73 with 3 df (P <0.001)

Taken together, the 3 predictors are highly significant predictors of case-control status.

Likelihood ratio test comparing the full and reduced models:

G_0^2 = 2(-54.00 - (-69.07)) = 30.16 with 2 df (P <0.001)

The 2 antibiotic specific predictors (-dneo- and -dclox-) are highly significant predictors.

Likelihood ratio test comparing the saturated and full models:

G_0^2 = 2(0 - (-54.00)) = 108.00 with 104 df.

Note This does not have a χ^2 distribution.

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 $1\text{nL}_{sat}=0$). 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.

$$D=2\ \ln\frac{L_{sat}}{L_{full}}=2(\ln L_{sat}-\ln L_{full})=-2(\ln L_{full})$$

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. Consequently, 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 21.4 and 2 df. This is a more conservative test statistic (although this is not generally the case) than the likelihood ratio test (χ^2 =30.16), 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.

$$1\text{n}\left[\frac{p}{1-p}\right] = \beta_0+\beta_1(\text{dcpct})+\beta_2(\text{dneo})+\beta_3(\text{dclox})+\beta_4(\text{dbarn\_2})+\beta_5(\text{dbarn\_3})$$

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 -dneo- in Example 16.2 is:

$$OR = e^{\beta_2} = e^{2.685} = 14.7$$

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 in which incidence density sampling was employed, the *OR* is a good estimate of the incidence rate ratio (*IR*) in the original population (see Chapter 6).

16.8.2 Continuous predictor

For a continuous predictor, the coefficient (β1) represents the change in the log odds of disease for a one-unit change in the predictor. Similarly, the computed *OR* represents the factor by which the odds of disease are increased (or decreased) for each one-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 50% to 75% of cows dry-treated, the log odds of disease changes by:

$$\log \text{odds}(x_{1,}x_2)=(x_2-x_1)*\beta_1=(75-50)*0.022=0.55 \qquad \textbf{\textit{Eq 16.12}}$$

For this 25% change in -dcpct-, the odds of disease change by:

$$e^{0.55}=1.73, \quad \text{or} \quad OR(x_{1,}x_2)=OR^{(x_2-x_1)}=1.022^{(75-50)}=1.72 \qquad \textbf{\textit{Eq 16.13}}$$

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 -dbarn- was converted to a series of 3 dummy variables and 2 of these (-dbarn_2-, -dbarn_3-) were included in the model. These represented tiestall and 'other' types of housing, respectively and, consequently, the coefficients represent the effects of these types of housing on the risk of *Nocardia* mastitis compared with freestall barns (the category that was omitted).

Example 16.2 Interpreting logistic regression coefficients
data = Nocardia

The tables below present results from a logistic regression of -casecont- on -dcpct- -dneo- -dclox- and 2 levels of -dbarn-. The first table presents the effects of the predictors on the logit of the outcome (case-control status), while the second shows the same results expressed as odds ratios.

Number of obs = 108
LR chi2 (5) = 47.40
Prob > chi2 = 0.000
Log likelihood = -51.158

Predictor	Coef	SE	Z	P	95% CI	
dcpct	0.022	0.008	2.82	0.005	0.006	0.037
dneo	2.685	0.677	3.96	0.000	1.358	4.013
dclox	-1.235	0.581	-2.13	0.033	-2.374	-0.096
dbarn_2	-1.334	0.632	-2.11	0.035	-2.572	-0.095
dbarn_3	-0.218	1.154	-0.19	0.850	-2.481	2.044
constant	-2.446	0.854	-2.86	0.004	-4.120	-0.771

Predictor	OR	SE	95% CI	
dcpct	1.022	0.008	1.007	1.037
dneo	14.662	9.931	3.888	55.296
dclox	0.291	0.169	0.093	0.908
dbarn_2	0.263	0.166	0.076	0.909
dbarn_3	0.804	0.928	0.084	7.722

Effect of -dneo- Use of neomycin-based products in the herd increased the log odds of *Nocardia* mastitis by 2.685 units. Alternatively, one can say that using neomycin-based products increased the odds 14.7 times. Since *Nocardia* mastitis is a relatively rare condition, it would be reasonable to interpret the odds ratio as a risk ratio and state that use of neomycin-based products increased the risk of *Nocardia* mastitis by approximately 15 times.

Effect of -dcpct- Changing the percentage of dry cows treated from 50% to 75% increases the log odds of disease by: (75-50)*0.022=0.55 units. Alternatively, it increases the odds of disease by: $(1.022)^{(75-50)}$=1.72. An increase of 25% in the percentage of cows dry-treated increases the risk of disease by about 72% (*ie* 1.72 times).

Effect of -dbarn- Tiestall barns (-dbarn_2-) and other barn types (-dbarn_3-) both had lower risks of *Nocardia* mastitis (*ie OR* <1) than did freestall barns (-dbarn_1- was the omitted baseline). However, the multiple Wald test and the likelihood ratio test of the 2 included categories were 0.08 and 0.06, respectively, suggesting that barn type was only borderline significant (0.1 >P >0.05).

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:

$$\ln\left(\frac{p_0}{1-p_0}\right) = \beta_0$$

Eq 16.14

where p_0 equals the probability of disease in this 'non-exposed group'. In a cross-sectional or cohort study, p_0 has real meaning because it represents the frequency of disease in the non-exposed group. However, in a case-control study, p_0 will vary depending on how many cases and controls are selected for inclusion in the study. We don't really know what the frequency of disease is in the non-exposed group because we didn't take a sample from that group. Consequently, the value of the intercept cannot be meaningfully interpreted if the data came from a case-control study.

16.8.5 Presenting effects of factors on the probability scale

As has been presented above, the coefficients from a logistic model represent the change in the log odds of disease that is associated with a unit change in the factor of interest. These can be relatively easily converted to an odds ratio (by exponentiating the coefficient) but there is a limitation to the usefulness of this parameter.

We normally think about the probability of disease (rather than the odds) and the probability of disease is not linearly related to the factor of interest. Consequently, the effect of a unit increase in the factor usually does not increase the probability of disease by a fixed amount. The amount that a unit increase in the factor changes the probability of disease depends on the level of the factor and the levels of other factors in the model.

In Example 16.3, you can see that the effect of a 10% increase in the percentage of cows dry-treated depends heavily on whether it occurs in a herd that uses neomycin or one that uses cloxacillin. It also depends on whether the change is from 10-20% or 80-90%. It is very helpful to generate some graphs of predicted probabilities to get a full understanding of the effects of key variables in your model.

As can be seen, a 10% increase in the level of -dcpct- has a greater effect on the probability of *Nocardia* mastitis in herds using neomycin; furthermore, in the cloxacillin herds, there is a bigger increase in the predicted probability of mastitis going from 80–90% than from 0–10%.

16.9 ASSESSING INTERACTION AND CONFOUNDING

Assessment of interaction and confounding in logistic regression models is similar to the process used in linear regression. **Confounding** is assessed by adding the potential confounding variable to the model and making a subjective decision as to whether or not the coefficient of the variable of interest has changed 'substantially'. In Example 16.4, it appears there is some degree of confounding between -dcpct- and -dclox-.

Interaction is assessed by adding the cross-product term ($X_1 * X_2$) and determining if the coefficient for the term is statistically significant. Estimation of *OR*s in the presence of interaction deserves some attention though. If interaction is present, the *OR* for the variable of interest has to be determined at a predefined level of the interacting variable because it will vary with the level of the interacting variable.

If the interaction is between 2 dichotomous predictors, the coefficients for the main effects and

Example 16.3 Effects of factors on the probability scale

dataset = Nocardia

In this example, a model containing -dcpct-, -dneo- and -dclox- was fit and the predicted probability of *Nocardia* mastitis computed as -dcpct- rose from 0 to 100%. Predicted probabilities were computed separately for neomycin-using herds and cloxacillin-using herds.

Predictor	Coef	SE	Z	P	95% CI	
dcpct	.023	.007	3.15	0.002	.008	.037
dneo	2.212	.578	3.83	0.000	1.080	3.345
dclox	-1.412	.557	-2.53	0.011	-2.505	-.320
constant	-2.984	.772	-3.86	0.000	-4.498	-1.471

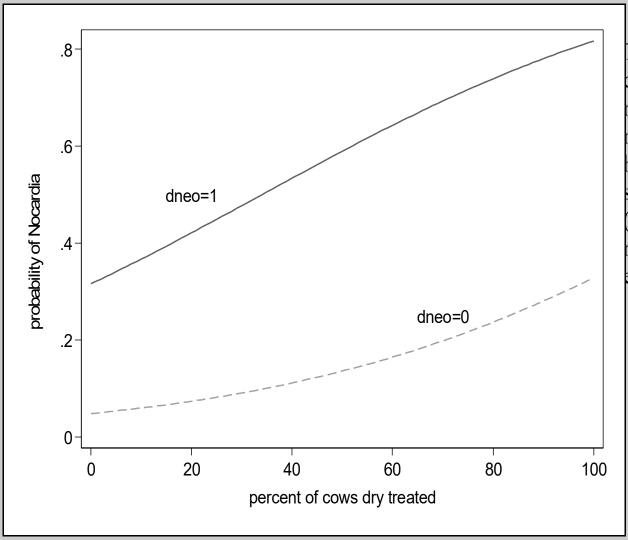

Fig. 16.3 Effect of dry-cow treatment

The effect of a 10% increase in -dcpct- depends on whether the herd is a neomycin-using or not (*ie* the effect is much greater in neomycin-using herds). It also depends on where on the scale of -dcpct- the increase occurs (going from 10-20% in a non-neomycin using herd has a smaller effect than going from 80-90%).

the interaction term have straightforward interpretations. The coefficient for each main effect represents the effect of that variable in observations in which the other variable is absent. In Example 16.5, the coefficient for -dneo- (3.184) is a measure of the effect of neomycin used in herds that don’t use cloxacillin. The interaction term represents the additional effect of having both factors present, over the sum of the 2 individual effects. The results shown in Example 16.5 are summarised in Table 16.2.

Table 16.2 Effect of neomycin and cloxacillin use on the log odds of Nocardia mastitis compared with using neither (from Example 16.5)

		cloxacillin	
		0	1
neomycin	0	0	0.446
	1	3.184	1.078

Note 1.078=3.184+0.446–2.552

Example 16.4 Assessment of confounding
data = Nocardia

First a 'full' model containing -dcpct-, -dneo- and -dclox- was fit, and then -dcpct- was dropped from the model.

	Full model		Reduced model	
Predictor	Coef	SE	Coef	SE
dcpct	0.023	0.007		
dneo	2.213	0.578	2.377	0.550
dclox	-1.413	0.557	-1.010	0.532
constant	-2.984	0.772	-1.480	0.501

When -dcpct- was removed from the model, the coefficient for -dneo- changes very little (-7%), but the coefficient for -dclox- changes by almost 40% suggesting that -dcpct- and -dclox- might be related (acting as confounders for each other).

Higher-order interactions (*eg* 3-way interactions) might also be evaluated (see Section 15.7). **Note:** Interactions in logistic regression are assessed on a multiplicative scale (*ie* in the absence of interaction, the effect of one factor multiplies the log odds of disease by a constant amount, regardless of the level of a second factor). Methods for evaluating interaction on an additive scale have recently been published (Knol *et al*, 2007). (See Section 13.6.2 for a discussion of additive and multiplicative interaction.)

Example 16.5 Assessment of interaction
dataset = Nocardia

Interaction between -dneo- and -dclox- was evaluated by adding their cross-product term:

Predictor	Coef	SE	Z	P	95% CI	
dcpct	0.023	0.008	2.93	0.003	0.007	0.038
dneo	3.184	0.837	3.80	0.000	1.543	4.825
dclox	0.446	1.026	0.43	0.664	-1.565	2.457
neoclox	-2.552	1.205	-2.12	0.034	-4.914	-0.190
constant	-3.777	0.993	-3.80	0.000	-5.724	-1.830

The effect of neomycin and cloxacillin use can be summarised as follows:

neomycin only	log odds goes up by: 3.18 units
cloxacillin only	log odds goes up by: 0.45 units
using both	log odds goes up by: 3.18 + 0.45 -2.55 = 1.08 units

Consequently, using neomycin-based products is much more harmful (increase of 3.18 units in log odds of *Nocardia* mastitis) in herds using neomycin exclusively. If the herd uses cloxacillin as well, the effect of neomycin is only an increase of 0.63 units (1.08-0.45). Alternatively, cloxacillin seems to have a small (insignificant) detrimental effect when used in herds that don't use neomycin (increase of 0.45 units), but in herds that use neomycin, it is highly beneficial (reduces log odds by 2.1 units (3.18-1.08).

16.10 MODEL-BUILDING

In general, the process of building a logistic model is very similar to that of building a linear regression model (see Chapter 15 for details). It might involve any of the following steps.

- laying out a tentative causal diagram to guide your thinking
- unconditional analyses of relationships between predictors and the outcome of interest using a 'liberal' P-value (*eg* unconditional logistic models)
- evaluating linearity of effects of continuous predictors
- evaluation of relationships (correlations) among predictor variables
- automated model-building processes (used with caution)
    - forward selection
    - backward elimination
    - stepwise selection
    - best subset regression
- manual model-building guided by a causal diagram (preferred method) including:
    - evaluation of confounding
    - evaluation of interaction.

One particular feature that must be kept in mind when fitting a logistic model is that data used to build logistic regression models may be binary (0/1) data (also called Bernoulli data) with one observation per study unit or binomial (also called grouped data) with each observation containing the number of positive responses and the number of trials for study units with a certain set of characteristics. A **covariate pattern** is a unique combination of values of predictor variables. For example, if the model contains only 2 dichotomous predictors, there will be 4 covariate patterns: (1,1) (1,0) (0,1) (0,0). The original binary data (n=number of study units) can be converted to binomial data (n=4) with each of the 4 observations having 4 variables: 2 that define the covariate patterns along with variables for the number of positive outcomes and the number of study units within each covariate pattern. On the other hand, if the model contains many continuous variables, there might very well be as many covariate patterns as there are data points (*ie* each covariate pattern will have only one observation in it) and these data are referred to as binary data. This distinction becomes crucial when computing residuals and evaluating the fit of logistic regression models (see Section 16.12.1).

A second fundamental difference relates to the process of evaluating the shape of the relationship between a continuous predictor variable and the outcome of interest. The assumption is that the relationship between the continuous predictor and the log odds of the outcome (not the outcome itself) is linear. Methods of evaluating the linearity of relationships are presented in Chapter 15 and issues related specifically to binomial data are discussed in Section 15.6.1.

Finally, logistic models can be fit directly using maximum likelihood procedures specific for logistic regression models, or they can be fit within the framework of generalised linear models, which are described in the following section.

16.11 GENERALISED LINEAR MODELS

Generalised linear models (GLMs) were developed in the 1970s (Nelder & Wedderburn, 1972) to provide a common framework for a wide range of statistical models, including both continuous and discrete distributions, with model-building and analysis similar to linear models based on the normal distribution (McCullagh & Nelder, 1989). There are 2 key components which need to be specified to fit a general linear model: the **link function** and the **distribution** of the observations.

Link function The cornerstone of GLMs is the link function: the idea that linear modelling of predictors should be allowed to take place on a different scale from the scale of the observations. The link function makes that transition between the observation's mean and the linear modelling. This idea may have been triggered by realising the problems of linear modelling of the observation's mean for bounded distributions. As noted in Section 16.1, modelling disease probabilities as a linear function of predictors may easily lead to predicted values outside the allowed range of probabilities (*ie* between 0 and 1). Consequently, in logistic regression, we model the logit(p)=ln[p/(1-p)] as a linear function of predictors. In GLM terminology, the logit function is the link function. The logit function, which maps the unit interval (0,1) onto the entire real axis (-∞,+∞) is shown on the left of Fig. 16.1. Intuitively, this is like 'stretching' the interval. The graph on the right shows its inverse function, $\text{logit}^{-1}(s)=e^{s}/(1+e^{s})$.

Distribution Data with a wide range of distributions can be modelled in the GLM framework, but the most commonly used distributions are: binomial (including binary), Poisson, negative binomial, Gaussian (normal), inverse Gaussian and gamma.

In theory, the link function used with any specific distribution can be arbitrary, but in practice, it is restricted to a few common choices for each distribution of Y. Each distribution has a 'natural' link function associated with it that is called the **canonical link**. For Gaussian (normal) data, the canonical link is the identity link because the outcome (*Y*) is linked directly to the predictors. For binary/binomial data, the canonical link is the logit link but 2 occasionally used non-canonical links are the probit function (inverse cumulative probability for the standard normal) and the complementary log-log function. The statistical inference using logit and probit links is usually similar, but parameter estimates are scaled roughly by the factor $\pi/\sqrt{3}$ (*ie* logistic regression estimates are numerically larger than those from a probit regression). Table 16.3 lists the canonical links and some commonly used non-canonical links for several distributions. Also, for ordinal data (and a multinomial distribution), the logit is the most common link.

Table 16.3 Selected distributions of outcomes and links used in fitting models in the GLM framework

Distribution of Y	Canonical link	Selected non-canonical links
Gaussian (normal)	identity	log
binary/binomial	logit	probit, complementary log log
Poisson	log	identity
negative binomial	negative binomial	log, identity

Poisson and negative binomial models are discussed in more detail in Chapter 18. The logit (or probit) links are also used for modelling ordinal and multinomial data (see Chapter 17). When choosing among link functions, we would usually use the most common one for the data type at hand, but if the model shows lack of fit, try some of the alternatives and choose the one that gives the best fit to the data. For the sake of completeness, we summarise the discussion by listing all the components of a generalised linear model:

(a) a link function,

(b) a distribution of the outcome Y,

(c) a set of explanatory variables (in a design matrix X), linked to the mean of the i^{th} observation, μ_i=E(Y_i), by the equation:

$$\text{link}(\mu_i)=\beta_0+\beta_1 X_{1i}+\ldots+\beta_k X_{ki} \qquad \textbf{\textit{Eq 16.15}}$$

(d) an assumption of independence between the outcomes.

One important feature of all GLMs with a non-identity link, is that all of the parameters are obtained on a transformed scale and, in order to give meaningful interpretations, we need to do 2 things. First, predicted values need to be back-transformed to the original scale, using the inverse link function. Second, coefficients need to be converted to a more meaningful quantity. This is model specific and for the logistic model, exponentiating the coefficients produces odds ratios.

16.11.1 Estimation methods for GLMs

The standard estimation procedure for GLMs is maximum likelihood (ML) estimation. Due to similarities between different GLMs, generic algorithms for ML estimation applicable for a range of different GLMs were developed early on (the scoring method of Newton-Raphson estimation (McCullagh & Nelder, 1989)). These algorithms were noted to depend only on assumptions about the distribution related to its mean (through the link function) and variance. This triggered an extension of GLMs to include partially specified models involving only the mean and variance but not the full distribution (and likelihood function). When a real likelihood function no longer exists, the estimation is based on a so-called quasi-likelihood function (McCulloch *et al*, 2008). Examples of GLMs with different variance specifications than those derived from the distributions in Table 16.3 are discussed in Chapter 18 on negative binomial models and in Chapter 20 on overdispersion models.

16.11.2 GLM model evaluation

One advantage of the GLM framework is that it has a wide range of statistics and techniques useful for assessing the fit of the model. These include GLM goodness-of-fit statistics (Pearson and deviance χ^2) and the large number of GLM-defined residuals (including Pearson, deviance, Anscombe, partial and score residuals) and other diagnostic parameters (*eg* Cook's distance). Some of these are covered in the following sections.

16.12 EVALUATING LOGISTIC REGRESSION MODELS

There are 2 steps in assessing the fit of the model. The first is to determine if the model fits, in general, using summary measures of goodness of fit or by assessing the predictive ability of the model. The second is to determine whether there are any specific observations (or groups of observations) that do not fit the model or that are having an undue influence on the model. However, before proceeding with either of these 2 areas, it is important to understand the distinction between residuals computed on the basis of 'covariate patterns' (see Section 16.10 and those computed on the basis of 'observations'.

16.12.1 Residuals and covariate patterns

The concept of **covariate patterns** was introduced in Section 16.10. Residuals from logistic models can be computed on the basis of one residual per observation or one residual per covariate pattern. To get a feeling for the difference between these 2 approaches, imagine a covariate pattern 'A' with 2 observations, one disease '+' and one disease '-'. Further assume that the predicted value for the probability of disease in animals with this covariate pattern is 0.5 (Table 16.4).

Table 16.4 Residuals computed on the basis of one per observation and one per covariate pattern

				Residuals	
Observation	Covariate pattern	Disease	Predicted value	One per observation	One per covariate pattern
1	A	1	0.5	positive	0
2	A	0	0.5	negative	

With one residual per observation, we have 2 residuals, of which one will be positive and one will be negative. With residuals computed on the basis of covariate patterns, the predicted value (0.5) exactly equals the observed value (0.5) so the residual is zero. For logistic models, residuals are normally computed on the basis of one per covariate pattern and some of the desirable properties of the residuals only apply if there is a reasonable number of observations in each covariate pattern.

In the following discussion, we will use j to represent the number of covariate patterns, m_j to represent the number of data points in the j^{th} covariate pattern, k to represent the number of predictors in the model (not including the constant) and n is the number of data points in the dataset.

All of the examples in this section are based on the model shown in Example 16.5 (with -dcpct-, -dneo-, -dclox- and the -dneo*dclox- interaction term as predictors). The values of the predictors in this model make up 30 distinct covariate patterns.

16.12.2 Pearson and deviance residuals

Computing residuals for a logistic model is not as straightforward as it is following a linear regression model (*ie* observed value-expected value). A number of different types of residual have been proposed, but the 2 most commonly used are Pearson residuals and deviance residuals.

Pearson residuals are roughly analogous to standardised residuals in linear regression. They are based on the difference between the observed and expected values for a given covariate pattern, but are adjusted based on the precision of the estimate of the observed value (*ie* covariate patterns with a large number of observations will have a more precise estimate than those in which there are few observations). **Pearson residuals** are computed as:

$$r_j = \frac{y_j - m_j p_j}{\sqrt{m_j p_j (1 - p_j)}}$$

where y_j=the number of positive outcomes in the j^{th} covariate pattern and p_j=the predicted probability for the j^{th} covariate pattern. Pearson residuals computed on the basis of one per covariate pattern and one per observation are presented in Example 16.6.

Deviance residuals represent the contribution of each observation to the overall deviance. The sum of deviance residuals computed on the basis of individual observations (rather than covariate patterns) is the deviance (-2*log likelihood) that was observed when comparing the full and saturated models (Section 16.6.3).

Both Pearson and deviance residuals may be standardised to have a mean of zero and unit variance. Standardised residuals are generally used for graphical examination of residual patterns and in particular, standardised deviance residuals are most likely to follow a Normal distribution (Hilbe, 2009). Other residuals (Anscombe, score and partial residuals) are available within the GLM framework, but are beyond the scope of this text.

16.12.3 Goodness-of-fit tests

A variety of tests are available to provide an overall assessment of how well the model fits the observed data. All of these tests are based on the premise that the data will be divided into subsets and within each subset, the predicted number of outcome events will be computed and this will be compared with the observed number of outcome events. Two tests (the Pearson χ^2 and the deviance χ^2) are based on dividing the data up into the natural covariate patterns. A third test (Hosmer-Lemeshow test) is based on a more arbitrary division of the data. Other measures of fit are also described.

Example 16.6 Residuals and covariate patterns
data = Nocardia

A logistic regression model of -casecont- on -dcpct-, -dneo-, -dclox- and the -dneo*dclox- interaction term was fit (see Example 16.5).

It turns out that there were 30 distinct covariate patterns represented in this model. The data for covariate pattern #9 (herds that dry-treated 20% of their cows, and used neomycin-based products but not cloxacillin-based products) are shown below.

cov. pattern	id	case-control	dcpct	dneo	dclox	pred. value	Pearson residual (covariate)	Pearson residual (observ.)
9	22	no	20	yes	no	0.465	0.099	-0.932
9	86	yes	20	yes	no	0.465	0.099	1.073

There were 2 observations in covariate pattern #9 and an observed probability of a positive outcome of 0.5 (1 of the 2 herds was positive). The predicted probability was 0.465 and the Pearson residual computed on the basis of one residual per covariate pattern was a small positive value (0.099). However, when residuals were computed for each observation individually, there was one moderately large positive residual value (1.073 for the case herd) and a negative residual value of a similar magnitude (-0.932) for the control herd.

Pearson and deviance χ^2 tests

The sum of Pearson residuals squared is known as the Pearson χ^2 statistic. When computed on the basis of one per covariate pattern, this statistic has a χ^2 distribution with (j-k-1) df provided that j is much smaller than n (*ie* on average, the m_j are large). j being much smaller than n ensures that the observed probability of the outcome in each covariate pattern is based on a reasonable sample size. If j=n (*ie* binary data), or almost so, the statistic does not follow a χ^2 distribution, so this goodness-of-fit statistic cannot be used.

The Pearson χ^2 indicates whether or not there is sufficient evidence that the observed data do not fit the model (*ie* H_0 is that the model fits the data). If it is not significant, it suggests that there is no reason to assume that the model is not correct (*ie* we accept that the model generally fits the data). **Note** In general, goodness-of-fit tests do not have a lot of power to detect inadequacies in the model.

The sum of the squared deviance residuals computed on the basis of 1 per covariate pattern (*ie* only applicable to binomial data) is called the deviance χ^2. **Note** The term deviance χ^2 is used to differentiate this deviance from that computed on the basis of 1 per observation (discussed in Section 16.6.3). As with the Pearson χ^2, it has a χ^2 distribution with (j-k-1) df. If either the Pearson χ^2 or the deviance χ^2 are significant, you should be suspicious that the model does not fit the data. Example 16.7 shows the Pearson χ^2 and deviance χ^2 for the model presented in Example 16.5.

Hosmer-Lemeshow goodness-of-fit test

If you have binary data (or any situation where j is not much less than n), you can't rely on covariate patterns to divide your data into subsets of sufficient size for a valid goodness-of-fit test. One way to get around this problem is to group the data using some method other than covariate patterns and compare the observed and predicted probabilities of disease (if that is the outcome of interest) in each group. This is the basis of the Hosmer-Lemeshow test (Hosmer & Lemeshow, 2000).

There are 2 ways to group the data. The first is on the basis of percentiles of estimated probability and the second is on fixed values of estimated probability. For example, if you want 10 groups, the first method would take the 10% of the data points with the lowest predicted probabilities of disease and put them in group 1, the next 10% in group 2 *etc*. The second approach would take all data points for which the predicted probability of disease was less than 0.1 and put them in a group (regardless of how many data points fell into that group). In general, the first approach is preferable because it avoids the problem of some groups having very small sample sizes.

Once the data are grouped, a 2*g table is set up (g is the number of groups and should not be <6) with the observed and expected number of cases included in each cell. The expected number of cases in the g=1 row of the table is simply the sum of the estimated probabilities for all subjects in the group. The observed number of cases is simply the number of observations with Y=1. The observed and expected values are compared using a χ^2 statistic with g-2 df. A visual comparison of the observed and expected values will also identify areas where the model might not fit well. Example 16.7 shows the Hosmer-Lemeshow χ^2 along with the observed and expected values.

Example 16.7 Goodness-of-fit tests
data = Nocardia

Goodness-of-fit tests were carried out on the model from Example 16.5. The Hosmer-Lemeshow test was computed with just 7 groups because of the small sample size.

Test	P^2	df	P
Pearson P^2	53.49	25	0.001
Deviance P^2	29.41	25	0.247
Hosmer-Lemeshow	3.85	5	0.572

As can be seen from the P values, there is quite a range of estimates. Since goodness-of-fit tests generally have low power for detecting inabilities of models to adequately fit the data, the general guideline is that if any goodness-of-fit test is statistically significant, you should assume there is a problem with the model and try to correct it. It is also worth noting that with 108 observations and 30 covariate patterns, the average number of observations per covariate pattern is quite low, so the Hosmer-Lemeshow test provides the most reliable evaluation.

A table of the observed and expected values from the Hosmer-Lemeshow test provides some insight into where the model does not fit the data very well.

		Cases	Cases	
Group	p(D+)	observed	Expected	# of herds
1	0.04	1	0.3	11
2	0.18	2	2.2	14
3	0.26	3	3.0	12
4	0.38	1	2.5	7
5	0.41	4	3.9	10
6	0.75	8	8.5	14
7	0.84	35	33.6	40

Proportionally, the largest differences between the observed and expected number of cases is in the first group (lowest predicted probabilities). One possible explanation of this is that some cases might have arisen from mechanisms not included in the model.

16.12.4 Overdispersion and Goodness-of-fit tests

Overdispersion occurs when there is more variation in a set of binomial proportions than would be expected based on the variance of the binomial function. One of the common causes of overdispersion is clustering of data which is discussed in much more detail in Chapters 20 and 22. However, consider the following simple example. Hypothetical data were computed for 10 herds, each with 20 cattle. Each animal was then given a 40% chance of being disease positive (regardless of the herd they were in). The distribution of herd prevalences is shown in the top row of Table 16.5 (labelled 'not clustered') and it has a mean of 0.40 and a standard deviation of 0.098. However, if the disease was highly infectious and affected all of the cows in 4 herds, but was not present in the other 6, the mean prevalence would still be 0.4 but the distribution would look like that shown in the second row (labelled ‘clustered’) and the standard deviation of these values is 0.516. There is clearly much more variability in the herd prevalences as a result of the clustering.

Table 16.5 Hypothetical data showing overdispersion as a result of clustering

Herd	1	2	3	4	5	6	7	8	9	10	mean	sd
Not clustered	0.5	0.3	0.4	0.6	0.45	0.3	0.35	0.4	0.35	0.3	0.395	0.098
Clustered	1	1	1	1	0	0	0	0	0	0	0.400	0.516

The example above shows overdispersion in terms of binomial data (herd proportions). Indeed, the concept of overdispersion really only applies to grouped (binomial) data. Individual level (binary) data cannot be overdispersed. (If the SDs of the above data are computed on the basis of the individual cow observations, both sets of data have a SD of 0.49). Nevertheless, clustered binary data may be ‘implicitly’ overdispersed because reformatting the data into a grouped data format makes the overdispersion obvious (Hilbe, 2009).

Overdispersion can arise in a variety of ways and can be classified as apparent or real overdispersion. **Apparent overdispersion** can be caused by any errors in the logistic model. This can include omission of important explanatory predictors, outlying observations (potentially errors in the data?), failure to account for important interactions in the model, or failure to satisfy the assumption of linearity for continuous predictors. The solution to apparent overdispersion is to fix the model.

Real overdispersion occurs when the true variance in the observed proportions is greater than what would be expected from binomially distributed data. As noted, a common cause of real overdispersion is clustering. Overdispersion may be detected by evaluating the Pearson χ^2 dispersion parameter (and its affiliated Pearson χ^2 statistic) or the Hosmer-Lemeshow goodness-of-fit test. However, both have limitations when you are dealing with binary (or ungrouped binomial) data as can be seen in Example 16.8. (**Note** some references suggest that the deviance χ^2 can also be used to evaluate overdispersion, but recent work (Hilbe, 2009) suggests that the Pearson χ^2 is preferred). Methods of dealing with overdispersion arising from clustering are presented in Chapters 20 and 22.

R^2 (pseudo-R^2)

A number of pseudo R^2-type measures for estimating the amount of variation explained by a

Example 16.8 Detecting overdispersion
data = hypothetical

Some hypothetical data consisting of 100 observations (cows) in each of 10 groups (herds) were constructed so that 5 of the groups had a high proportion of positive outcomes and the other 5 had a low proportion. A group-level predictor (x) which increased the logit of the outcome (y) by 1.0 was then incorporated and the data generated.

If these data are analysed as binary data, the Pearson χ^2 is 0.0 (P-value=1.0) and the overdispersion parameter is 1.002. Both would suggest no problem with overdispersion. However, if the data are collapsed to binomial data and analysed as such, the dispersion parameter is 49.5 which clearly shows the serious problem of overdispersion. This highlights the limitations of goodness-of-fit statistics to detect problems with clustering when binary data are used to build the model.

logistic regression model have been proposed and reviewed (Long & Freese, 2006; Mittlböck & Schemper, 1996; Mittlböck & Schemper, 1999). Details of the various methods are beyond the scope of this text. Unfortunately, the various methods often give widely varying results so interpretation of a value requires specific knowledge of how the measure was computed and what it represents (Hoetker, 2007). For example, for the *Nocardia* mastitis model, estimates from a variety of pseudo-R^2 measures range from 24% to 80% (data not shown). In general, Hosmer and Lemeshow (2000) argue that the pseudo-R^2 is equivalent to the likelihood ratio test for all of the parameters in the model (*ie* comparing the likelihood of the full model to one with only the intercept). It does not compare the fit of the model with the observed values and consequently is better suited for comparing models than for assessing the goodness of fit of a selected model.

16.12.5 Predictive ability of model

A second general approach to assessing the overall usefulness of the model is to assess its predictive ability (*ie* how good a job does it do in predicting the outcome?). This can involve computing the sensitivity and specificity of the model at various probability thresholds and/or generating a receiver operating characteristic (ROC) curve.

Sensitivity and specificity

The ability of the model to correctly classify individuals (or in this example, herds) can be assessed by computing the classification statistics after fitting a model. By default, these are computed by classifying every observation that has a predicted probability $0.5 as positive and those with values <0.5 as negative. However, this cutpoint can be lowered (to increase the sensitivity of the model) or raised (to increase the specificity) similar to the cutpoints for tests (Section 5.6.3). A graph of the sensitivity and specificity vs the potential cutpoint values (2-graph ROC curve—Section 5.5.1) is helpful in selecting an appropriate cutpoint (Example 16.9).

Receiver operating characteristic curves

An ROC curve for the model can also be generated to evaluate the performance of the model at all possible cutpoints. The closer the curve comes to the upper left corner of the graph, the better the predictive ability of the model. If the ROC curve is close to the diagonal line, it indicates that the model has very little predictive ability. The maximum area under an ROC curve is 1.0 (*ie* sensitivity=100% and specificity=100%) while the area will be 0.5 if the curve falls on the diagonal line (*ie* has no predictive ability at all). (See Section 5.5.2 for a more complete discussion of ROC curves.) The predictive ability of the model for *Nocardia* mastitis is shown in Example 16.10.

16.12.6 Identifying important observations

Detecting observations which either do not fit the model well, or which might have an undue influence on the model is an important component of evaluating a logistic regression model, particularly if any of the goodness-of-fit statistics indicate problems with the model.

Outliers

Pearson residuals and deviance residuals represent the square root of the contribution of the covariate pattern to the Pearson and deviance χ^2 statistics, respectively. As with standardised residuals from linear regression, large positive or negative standardised residuals identify points

Example 16.9 Predictive ability of a model—2-graph ROC curve
data = Nocardia

For the model presented in Example 16.5, the classification statistics are:

	Classified (predicted) status		
	T+	T-	
True status	p(D+)≥0.5	p(D+)<0.5	Total
D+	41	13	54
D-	9	45	54
Total	50	58	108

Sensitivity	pr (T+|D+)	75.93%
Specificity	pr(T-|D-)	83.33%
Positive predictive value	pr(D+|T+)	82.00%
Negative predictive value	pr(D-|T-)	77.59%

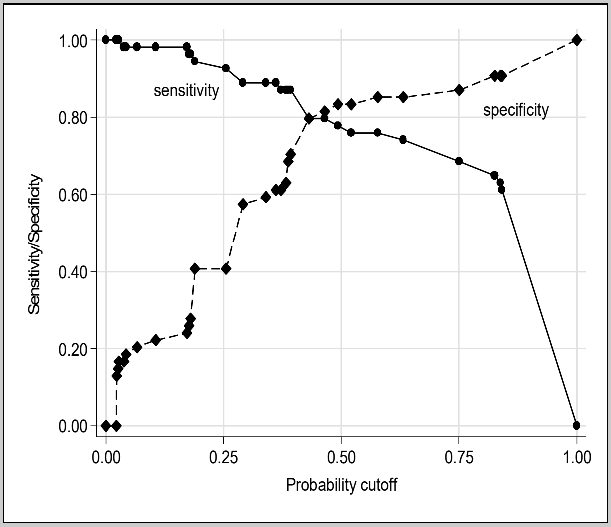

Fig. 16.4 Use of a 2-graph ROC to show the effect of changing cutpoint

At a cutpoint of 0.5, the sensitivity and specificity of the model are roughly balanced. The effect of changing the cutpoint can be evaluated visually in the graph.

In this situation, reducing the cutpoint would reduce specificity quite dramatically and raising it beyond about 0.75 would seriously affect sensitivity.

which are not well fit by the model. If outliers are observed, it is important to try to determine:

(a) Why they are outliers (what are the characteristics of the observations that make them outliers?).
(b) If the data are found to be erroneous, they should be corrected, or failing that, deleted.
(c) If the data are correct, determine if they are having an undue effect on the model.

This last point can be evaluated by looking at other diagnostic parameters (leverage, delta-betas, *etc* (see below) or by refitting the model with the outliers omitted. (Deleting the outliers should only be done for the purpose of evaluating their impact on the model and they must be

Example 16.10 Predictive ability of a model—ROC curve

data = Nocardia

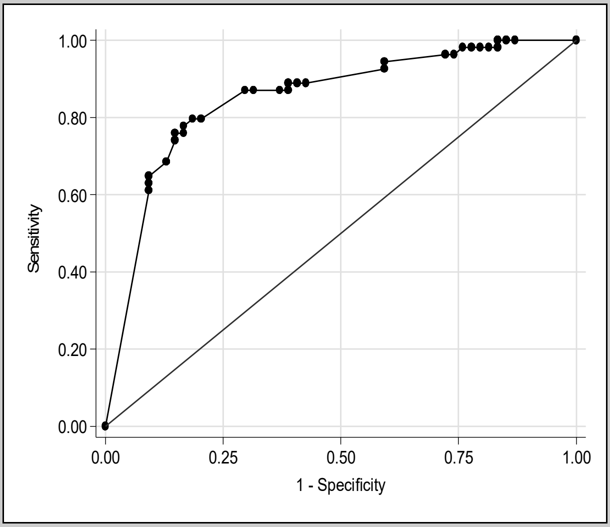

Fig. 16.5 ROC curve

An ROC curve for the *Nocardia* mastitis model is presented in Fig. 16.5.

The ROC curve extends reasonably well into the upper left-hand corner of the graph and the area under the curve is 0.85. Both of these indicate that the model has a moderate predictive ability.

put back in the dataset.) In general, outliers contribute to the lack of fit of a model but often do not have an undue influence on it. An index plot of standardised Pearson residuals (one per covariate pattern) is shown in Example 16.11 and the effect of removing the single observation with a very large standardised residual is shown in the continuation of that example.

Hat matrix and leverage

Another quantity central to the discussion of logistic regression diagnostics is the hat matrix. It is used to calculate leverage values and other diagnostic parameters. The hat matrix is a square matrix of dimension $j * j$ (j=number of covariate patterns) or $n * n$ (n=number of data points) depending on whether the data are binomial or binary. The diagonal elements of the hat matrix are the logistic regression leverage values (h_j) (see Hosmer and Lemeshow, 2000 for details).

As in linear regression, leverage measures the potential impact of an observation (or covariate pattern) on the model. Points with high leverage certainly deserve evaluation given their potential impact.

Unlike leverage values in linear regression models, the leverage of a data point in a logistic model is not exclusively a function of the values of the predictors. Data points that have extreme values of predictor variables (which would have high leverage in linear regression) might, in fact, have low leverage in logistic regression if the predicted value is very large or very small. Observations with extreme values of the predictor(s) will have leverage values that are: highest if the predicted probability lies between 0.1 and 0.3 or 0.7 and 0.9, moderate between 0.3 and 0.7, and low if the predicted probability is <0.1 or >0.9. The covariate patterns with the highest leverage are shown in Example 16.12.

Delta-betas

Values of delta-beta provide an estimate of the effect of the j^{th} covariate pattern on the logistic regression coefficients. These values are analogous to Cook's distance in linear regression

Example 16.11 Identifying important observations

data = Nocardia

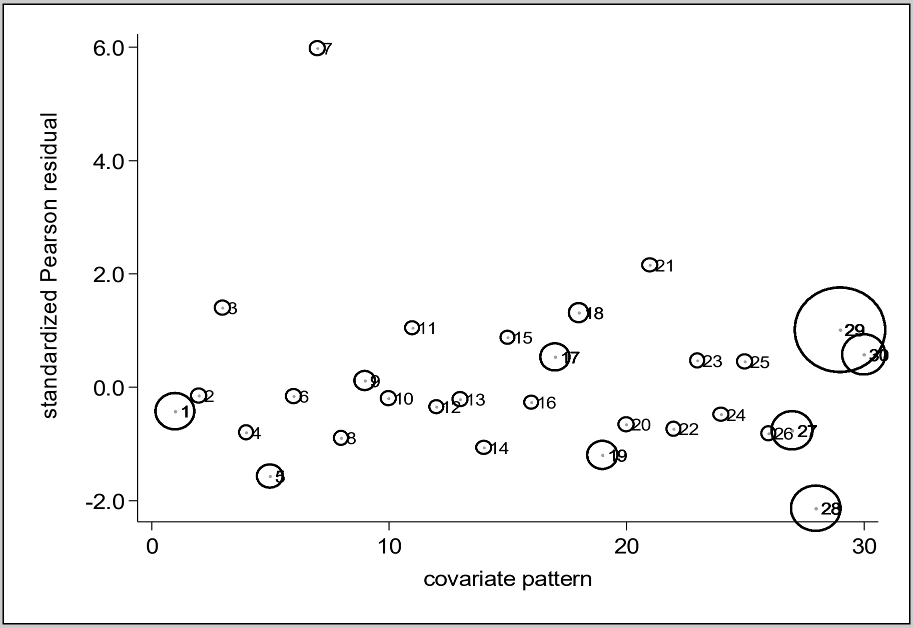

Fig. 16.6 Index plot of standardised residuals

From the model fit in Example 16.3, an index plot of standardised residuals with the covariate pattern identification number used as the marker label and the size of the circles proportional to the number of observations in the covariate pattern identifies one large positive residual. There were no covariate patterns with particularly large negative residuals.

Covariate pattern 7 consisted of a single-case herd which had a very low predicted probability of being a case herd (2.8%). This suggests that *Nocardia* mastitis might have arisen in this herd from some mechanism other than those covered by the predictors in the model, although the possibility of misclassification bias (*ie* false positive cases) cannot be ruled out.

If the model is refit with covariate pattern 7 omitted, the results are as follows:

	full dataset (n=108)		cov. pattern #7 omitted (n=107)	
Predictor	**β**	**SE**	**β**	**SE**
dcpct	0.023	0.008	0.027	0.008
dneo	3.184	0.837	4.035	1.111
dclox	0.446	1.026	1.155	1.244
dneo*dclox	-2.552	1.205	-3.369	1.408
constant	-3.777	0.993	-4.964	1.289

The effect of removing this outlier is that the coefficients for all 4 predictors have moved away from the null (*ie* either larger positive or negative values). This suggests that the model based on the full dataset might provide slightly conservative estimates of the effects of these predictors, but it should be noted that deleting this observation also increases the SEs. However, there is no justification for removing this observation, so the full model should be used).

Example 16.12 Identifying influential observations
data = Nocardia

Based on the model fit in Example 16.5, the covariate patterns with the largest leverage values are:

covariate pattern	# of herds	p(D+)	dcpct	dneo	dclox	predicted value	leverage
30	9	0.444	100	yes	yes	0.392	0.682
27	8	0.125	100	no	no	0.180	0.721
29	38	0.868	100	yes	no	0.841	0.796
28	11	0.182	100	no	yes	0.256	0.932

None of the covariate patterns have a particularly large leverage value, and the outlying observation (covariate pattern 7) did not have high leverage. The covariate patterns with the largest overall delta-betas were determined:

covariate pattern	# of herds	p(D+)	dcpct	dneo	dclox	predicted value	delta- beta
27	8	0.125	100	no	no	0.180	1.525
29	38	0.868	100	yes	no	0.841	3.953
28	11	0.182	100	no	yes	0.256	62.935

The covariate pattern with the largest delta-beta is pattern #28. This covariate pattern is very influential for the coefficients for -dclox- and the interaction term. In fact, if these 11 herds are omitted, it is not possible to get sensible estimates for those terms (data not shown). It is not surprising that covariate pattern 29 also has a large delta beta since this covariate pattern contained 38 observations (approximately 1/3 of the data). Neither the evaluation of the leverage values nor the delta-betas cause particular concern for this model.

The observation that was previously identified as an outlier (covariate pattern 7) is also the covariate pattern with the largest delta-chi-square and delta-deviance values (data not shown).

models.

A single set of values of delta-beta can be calculated—one value for each covariate pattern—and this represents the overall effect of the covariate pattern on the regression model. It is a measure of the distance between the observed set of regression coefficients and a similar set that would be obtained if the observations in the covariate pattern of interest were omitted when building the model. Alternatively, separate sets of delta-betas could be determined for each predictor variable to measure the effect of the covariate pattern on each coefficient in the model.

Values of delta-beta will depend on the leverage that the covariate pattern has, the predicted value, whether or not the model fits the data point well (*ie* is it an outlier?) and also on the number of observations in the covariate pattern. Covariate patterns with a large number of observations will naturally tend to have a large influence on the model, so we want to identify covariate patterns with a large influence but a small m_j, for further investigation.

If a particular pattern has a large delta-beta, it is important to determine why that is. As noted in our example (16.12), when m_j is large, that covariate pattern will likely have a big impact on the model. This is as it should be and need not concern us. However, if it is a covariate pattern with relatively few observations, then it is important to verify that the data are correct and determine if there is a logical explanation for the influence it is exerting.

Other parameters

Two other parameters which measure the overall influence of a covariate pattern on the model are the delta-χ^2 and the delta-deviance. The delta-χ^2 provides an overall estimate of the effect of the j^{th} covariate pattern on the Pearson χ^2 statistic. The delta-deviance provides an overall estimate of the effect of the j^{th} covariate pattern on the deviance χ^2. These 2 measures are overall evaluations of the fit of the model (*ie* they are based on the unexplained variation) so points that are outliers will tend to have large values for the delta-χ^2 and delta-deviance. However, as noted, these observations can only be deleted if you are certain that the data are erroneous.

16.13 Sample size considerations

There are 2 important issues related to sample size in logistic regression analyses. The first relates to the power of the study to detect effects of interest. For a simple logistic regression model with a single dichotomous predictor, the formula for comparing 2 proportions in Eq 2.6 will provide a reasonable estimate of the sample size. For multivariable models, the sample size adjustment shown in Eq 2.10 or 2.11 can be used. The simulation approach described in Section 2.11.8 provides a very flexible method of addressing all sample size issues.

The second issue relates to the adequacy of the obtained sample to support the fitting of a logistic model. In addition to considering the total sample size, the number of positive and negative outcomes in the observed data influence the precision of the estimates of the coefficients in the model. If positive outcomes are rare, then variances might be over- or underestimated and hence parameter estimates and test statistics might be affected. It has been suggested that the dataset should contain a minimum of $10(k+1)$ positive outcomes where k is the number of predictors in the model (not counting the intercept) in order to adequately fit the model (Hosmer & Lemeshow, 2000). The same rationale applies if negative outcomes are rare: there should be $10(k+1)$ negative outcomes in the dataset. It has recently been shown that in some situations this 'rule of 10' is conservative (Vittinghoff & McCulloch, 2007), but it is probably still useful as a general principle.

16.14 Exact logistic regression

In situations in which your dataset is very small or severely unbalanced, ML (or IRLS) estimates of coefficients (and their P-values) may be biased because the estimation procedures rely on asymptotic properties. An alternative approach in these situations is to use exact logistic regression. Exact logistic regression constructs a statistical distribution which can be determined completely and estimates the coefficient and P-value for each independent variable separately, while conditioning out the other predictors in the model. Consequently, the estimates are referred to as conditional maximum likelihood (CML) estimates. The procedure is very computationally intensive and in practical terms can only be used on small datasets with relatively simple models. It is possible to simplify the estimation procedure somewhat by identifying some predictors which need to be conditioned on, but for which you are not interested in the coefficients (*eg* confounders you want to control for). The details of exact logistic regression are beyond the scope of this text, so the reader is referred to (Mehta & Patel, 1995) for more information. Hilbe (2009) suggests that P-values comparable to those obtained

Example 16.13 Exact logistic regression
data = Nocardia

An exact logistic regression model was fit to the *Nocardia* data. The predictor -dcpct- was included in the model but conditioned out so its effects on the terms reflecting the use of neomycin, cloxacillin and their interaction would be controlled for, but the coefficient for -dcpct- was not estimated. For comparison purposes, the model (with -dcpct-) was fit using ordinary logistic regression with robust SEs.

Number of obs = 108

	Exact logistic regression				Robust SE	
Predictor	**Coef**	**P-value**	**95% CI**		**Coef**	**P-value**
dcpct	not estimated				0.023	0.006
dneo	3.079	0.000	1.412	5.425	3.184	0.000
dclox	0.428	1.000	-1.969	3.114	0.446	0.671
dneo*dclox	-2.424	0.102	-5.470	0.365	-2.552	0.042
constant	-3.580	0.000	-6.233	-1.662	-3.777	0.001

The interaction term is no longer significant, which calls into question the validity of the interaction effect observed in the ordinary logistic regression model. The P-value for the interaction term did remain significant (P=0.042) when the model was fit using ordinary logistic regression with robust SE (and this estimate was not much different from the one based on ordinary SEs (P=0.034)). However, the P-value from the exact model would likely be the best estimate of the significance of this term. An exact model without the interaction term was refit using exact logistic regression and the coefficients (and P-values) for -dneo- and -dclox- were 2.13 (P=0.0001) and -1.37 (P=0.020), respectively.

from exact logistic regression can often be obtained in small datasets by using ordinary logistic regression with robust SE (see Sections 14.9.5 and 20.5.4). An example showing the application of exact methods to the *Nocardia* dataset is shown in Example 16.13.

In small datasets, it is also often the situation in which a predictor may predict the outcome perfectly (*eg* all animals in one age group are positive). In this situation, ML and CML estimates are unbounded and cannot be estimated. In these situations, some software implementations for exact logistic regression switch automatically to computing an estimate of the coefficient using a procedure called median unbiased estimation. This at least provides a reasonable estimate of the parameter of interest.

16.15 CONDITIONAL LOGISTIC REGRESSION FOR MATCHED STUDIES

In our discussions of procedures to control confounding, we discussed the technique of matching. The most common application of this technique is in matched case-control studies in which a case is matched with one or more controls on the basis of some factor such as age, breed, herd of origin *etc*. Because there might be one case and a variable number of controls, this is often referred to as 1-M matching, of which 1-1 matching is a special case.

We could analyse the data using regular logistic regression procedures by simply including dummy variables to represent the *j* strata, where a case and its control(s) make up a stratum. Unfortunately, the generally desirable properties of maximum likelihood estimation of a logistic regression model only hold if the sample size is large relative to the number of parameters estimated and this wouldn't be true in a matched study with *j*-1 dummy variables to indicate the strata in addition to the predictors of interest. With matched-pair data (*ie* one case and one control in each matched set), an unconditional logistic regression model including *j*-1 dummy variables produces estimates of the odds ratios of interest that are the square of their true value (*eg* 9 vs 3) (Hosmer & Lemeshow, 2000). This is clearly undesirable.

As we don't really care about the coefficients for the *j* strata variables, we can use a technique known as conditional logistic regression (also called conditional fixed effects logistic regression or McFadden's Choice Model) to analyse matched data. (The conditional likelihood for the *j*th stratum is simply the probability of the observed data conditional on the number of observations in the stratum and the total number of cases in the study). Instead of estimating a parameter for each matched set (stratum) in the data (as an unconditional fixed effects model with indicator variables for strata would do), a conditional model conditions the fixed effects out of the estimation. A conditional logistic model has the following structure.

$$logit(p_i) = \beta_1 X_{1i} + ... + \beta_k X_{ki} \qquad \textbf{\textit{Eq 16.16}}$$

There are 3 limitations associated with the use of conditional models in terms of what can be estimated and which data contribute to the estimation. First coefficients cannot be estimated for predictors that are constant within all matched sets, even if they vary between sets. Consequently, there can be no analysis of the factors used for matching as they will be constant within a set. However, it is possible to include interaction terms between the matching variable and a predictor which varies within sets. Second, conditional models do not estimate an intercept (it is conditioned out). Finally, only sets in which a predictor varies within the set, contribute information to the estimation of the coefficient for that predictor. Example 16.14

Example 16.14 Lack of information from groups with no within-group variation
data = sal_outbreak

An outbreak of *Salmonella* in Funen County of Denmark in 1996 was investigated (see Chapter 31 for description of dataset). The data consisted of 39 cases of *Salmonella typhimurium* phage type 12 and 73 controls matched for age, sex and municipality of residence. Data on numerous food exposures were recorded and a small subset of those data are included in the dataset -sal_outbrk-.

The following table shows the cross-tabulations of case-control status with 4 predictor variables: recently eating pork (-eatpork-), recently eating beef (-eatbeef-), buying meat produced at slaughterhouse A (-slt_a-) and buying meat that came through dealer A (-dlr_a-) for a single matched set.

		eatpork		eatbeef		slt_a		dlr_a	
set		+	-	+	-	+	-	+	-
23	case	1	0	0	1	1	0	0	1
23	control	2	0	1	1	1	1	0	2
	OR	no info.		0		∞		no info.	

As can be seen within set 23, the case and both controls had recently eaten pork, but not products from dealer A, so this matched set provides no useful information with regard to these predictors. While the *OR* for -eatbeef- and -slt_a- are extreme (0 and ∞ respectively), they do provide evidence of negative and positive associations with being a case respectively.

shows why this is true, using data from a matched case-control study of a *Salmonella* outbreak. One consequence of this is that if either the case or all the controls within a set have a missing observation, the entire set is excluded from the analysis because there is no within-set variation. It has been shown that the use of GEE methods (Chapter 23) may be a suitable alternative to conditional logistic regression in situations in which there are many sets with missing data or no within-set variation (Lin *et al*, 2007).

Hypothesis testing in conditional models can be done using Wald tests or (preferably) LRTs in much the same way as for ordinary logistic regression models. Example 16.15 shows simple and multiple conditional logistic regression models for the *Salmonella* data along with a simple ordinary logistic regression.

If data that were collected in a matched-design study are analysed using an unconditional logistic regression model, one of 2 effects can occur. If the matching was done on variables that are confounders (*ie* matching was required to prevent bias) then the estimates from the unconditional analysis will be biased toward the null (*ie* a conservative estimate). If the matching was not necessary to avoid bias, then the coefficients from the unconditional analysis will not be biased, but will be less efficient (*ie* will have wider confidence intervals).

Example 16.15 Simple and multiple conditional logistic regression

data = sal_outbreak

Simple (-slt_a- as the sole predictor) and a multivariable (-slt_a- and its interaction with -gender-) were fit using conditional logistic regression and the results shown below.

Conditional (fixed-effects) logistic regression

```
                                                   Number of obs = 112
                                                    LR chi2(1) = 10.00
                                                 Prob > chi2 = 0.0016
                                        Log likelihood = -35.820042
                                                  Pseudo R2 = 0.1225
```

casecontrol	OR	SE	Z	P>z	95% CI	
slt_a	4.416	2.288	2.870	0.004	1.600	12.191

The odds ratio for -slt_a is 4.42 which is close to the estimate provided by a Mantel-Haenszel stratified (by matched set) analysis (*OR*=3.87).

Conditional (fixed-effects) logistic regression

```
                                                   Number of obs = 112
                                                    LR chi2(2) = 11.24
                                                 Prob > chi2 = 0.0036
                                        Log likelihood = -35.197693
                                                  Pseudo R2 = 0.1377
```

casecontrol	OR	SE	Z	P>z	95% CI	
slt_a	2.895	1.784	1.730	0.084	0.866	9.683
slt_a * gender	3.609	4.456	1.040	0.299	0.321	40.587

As can be seen, the main effect of -gender- was dropped from the model because there is no within-group variation in gender (it was one of the matching variables). The Wald test for the significance of the interaction term yields a P-value of 0.299 which is comparable to a LRT P-value of 0.265 (results not shown).

Consequently, matching should be accounted for in the analysis if it was incorporated into the design of the study (Breslow & Day, 1980). If an ordinary logistic model with -slt_a- as the sole predictor is fit to the *Salmonella* data, the resulting *OR* is 3.21, reflecting the expected bias toward the null.

The evaluation of these models (*ie* regression diagnostics) is not as straightforward as it is for ordinary logistic models (*eg* the Hosmer-Lemeshow goodness-of-fit test is inappropriate). However, some diagnostic parameters are available. Leverage can be computed from the hat matrix and delta χ^2 and delta β statistics can be computed on either an individual basis (reflecting the influence of that individual) or a matched group basis (reflecting the influence of the matched group. Example 16.16 shows some diagnostics for the *Salmonella* outbreak data.

Example 16.16 Conditional logistic regression diagnostics

data = sal_outbreak

Leverage, $\Delta\chi^2$ and $\Delta\beta$ statistics were computed from the model with -slt_a- as the sole predictor. The 3 sets with the largest $\Delta\beta$ values are shown below.

match						group	group
group	casecontrol	slt_a	leverage	$\Delta\chi^2$	$\Delta\beta$	$\Delta\chi^2$	$\Delta\beta$
55	contr	yes	0.007	0.821	0.006	4.545	0.133
55	case	no	0.033	3.723	0.127	4.545	0.133
2	contr	yes	0.001	0.450	0.001	9.012	0.184
2	contr	yes	0.001	0.450	0.001	9.012	0.184
2	case	no	0.022	8.112	0.183	9.012	0.184
9	contr	yes	0.001	0.450	0.001	9.012	0.184
9	contr	yes	0.001	0.450	0.001	9.012	0.184
9	case	no	0.022	8.112	0.183	9.012	0.184

Two sets (2 and 9) had large $\Delta\beta$s. These were sets in which the case did not consume products from slaughterhouse A but both controls did. If these 2 sets are left out of the analysis, the *OR* for -slt_a- increases to 8.01.

References

Breslow N, Day N. Statistical methods in cancer research. Vol 1: the analysis of case-control studies.. Lyon: Intl. Agency for Research on Cancer; 1980.

Hilbe J. Logistic Regression Models. CRC Press: Boca Raton; 2009.

Hoetker G. The use of logit and probit models in strategic management research: critical issues Strategic Management Journal. 2007; 28: 331-43.

Hosmer D, Lemeshow S. Applied Logistic Regression. 2nd Ed. John Wiley and Sons: New York; 2000.

Knol MJ, van der Tweel I, Grobbee DE, Numans ME, Geerlings MI. Estimating interaction on an additive scale between continuous determinants in a logistic regression model Int J Epidemiol. 2007; 36: 1111-8.

Lin I, Lai M, Chuang P. Analysis of matched case-control data with incomplete strata: applying longitudinal approaches Epidemiology. 2007; 18: 446-52.

Long J, Freese J. Regression Models for Categorical Dependent Variables Using Stata. Stata Press: College Station; 2006.

McCullagh P, Nelder J. Generalized Linear Models. 2nd Ed. Chapman & Hall: London; 1989.

McCulloch C, Searle S, Neuhaus J. Generalized, Linear and Mixed Models. Wiley-Blackwell: New York; 2008.

Mehta CR, Patel NR. Exact logistic regression: theory and examples Stat Med. 1995; 14: 2143-60.

Mittlböck M, Schemper M. Explained variation for logistic regression Stat Med. 1996; 15: 1987-97.

Mittlböck M, Schemper M. Computing measures of explained variation for logistic regression models Comput Methods Programs Biomed. 1999; 58: 17-24.

Nelder J, Wedderburn R. Generalized linear models J Royal Stat Soc, Series A. 1972; 135: 370-84.

Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression Am J Epidemiol. 2007; 165: 710-8.