

Chapter 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 different logistic models using both Wald tests and likelihood ratio tests.
 - d. Determine if the relationship between a continuous predictor variable and the log odds of disease is linear.
3. Evaluate logistic regression models
 - a. Understand covariate patterns and how they impact the computation of residuals for logistic regression models.
 - b. Compute residuals on the basis of one per covariate pattern and one per observation.
 - c. Select and use the appropriate test(s) to evaluate the goodness of fit of a logistic model.
 - d. Determine the effect of changing the threshold ('cutpoint') on the sensitivity and specificity of the model.
 - e. Generate ROC curves as a method of evaluating the goodness of fit.
 - f. Identify and determine the impact of influential observations on a logistic model.

copyrighted material

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 presence or absence 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.

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

$$\text{if } Y = 1 \text{ then } \varepsilon = 1 - (\beta_0 + \sum \beta_j X_j)$$

$$\text{if } Y = 0 \text{ then } \varepsilon = -(\beta_0 + \sum \beta_j X_j)$$

Eq 16.1

2. 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.
3. The mean responses should be constrained as:

$$0 \leq E(Y) = p \leq 1$$

However, with a linear regression model, the predicted values may 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 Nova Scotia, Canada, dairy herds. 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.

Table 16.1 Selected variables from the Nocardia dataset

Variable	Description
casecont	case or control status of the herd (the outcome)
dcpcct	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)

Details of the dataset can be found in Chapter 27.

16.2 THE LOGIT TRANSFORM

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.

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

where $\ln \left[\frac{p}{1-p} \right]$ is the logit transform. This value is the log of the odds of the outcome (since $\text{odds} = p/(1-p)$), so a logistic regression model is sometimes referred to as a log odds model.

Fig. 16.1 p vs logit of p

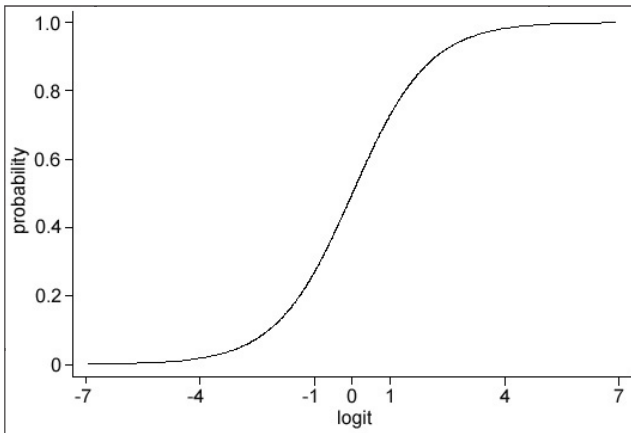


Fig. 16.1 shows that while the logit of p may become very large or very small, p does not go beyond the bounds of 0 and 1. In fact, logit values tend to remain between -7 and +7 as these are associated with very small (<0.001) and very large (>0.999) probabilities, respectively.

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

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

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 probability of disease becomes:

$$p = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \tag{Eq 16.4}$$

copyrighted material

From this, we can compute the odds of disease (ie $p/1-p$). To simplify calculating the odds of disease:

$$\text{let } \alpha = e^{\beta_0 + \beta_1 X} \text{ so } p = \frac{\alpha}{1 + \alpha}$$

Then it follows that:

$$\begin{aligned} \text{odds} &= \frac{p}{1-p} = \frac{\alpha}{1+\alpha} \bigg/ \left(1 - \frac{\alpha}{1+\alpha}\right) \\ &= \frac{\alpha}{1+\alpha} \bigg/ \frac{1+\alpha-\alpha}{1+\alpha} \\ &= \alpha = e^{\beta_0 + \beta_1 X} \end{aligned}$$

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'.

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

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

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).

copyrighted material

The likelihood (L) of getting three positive results from 10 cows if the true prevalence is P:

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

The log likelihood (ln(L)) is:

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

In this situation, the maximum value of the ln(L) 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. You pick a value for the prevalence (perhaps your first guess is 0.2). The probability of observing three positive cows out of 10, if the true prevalence (P) is 0.2, is:

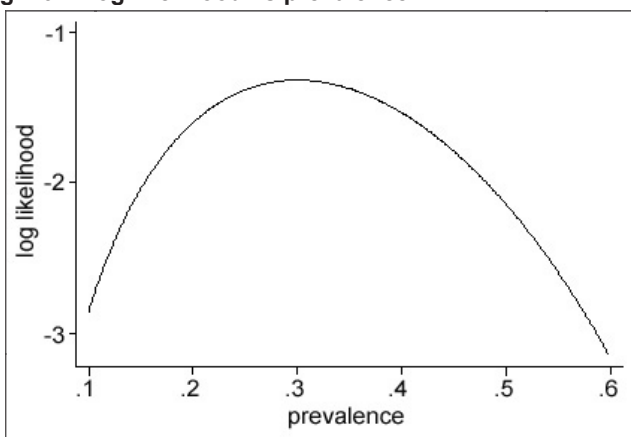
$$L(0.2) = \binom{n}{x} P^x (1-P)^{n-x} = \binom{10}{3} 0.2^3 (1-0.2)^{10-3} = 0.201 \tag{Eq 16.7}$$

The ln(L) is -1.60.

- b. You pick another prevalence (perhaps your next guess is 0.35) and recompute the likelihood. This turns out to be 0.252 (ln(L)=-1.38).
- c. You keep repeating this process until you have the estimate of the parameter that gives you the highest likelihood (ie the maximum likelihood). This would occur at P=0.3 (but you already knew that, didn't you?).

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

Fig. 16.2 Log likelihood vs 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:

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

and two 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 16.11.3 and methods of dealing with the problem are discussed in Chapter 20.

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 16.10.

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

copyrighted material

‘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_0^2) is:

$$G_0^2 = 2 \ln \left(\frac{L}{L_0} \right) = 2(\ln(L) - \ln(L_0)) \quad \text{Eq 16.9}$$

where L is the likelihood of the full model and L_0 is the likelihood of the null model. The statistic (G_0^2) 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, two 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 may 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 three predictor variables (-dneo- -dclox- -dpcpt-). The likelihood ratio test evaluating the three predictors as a group is highly statistically significant ($G_0^2=41.72$, $df=3$, $P < 0.001$).

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:

$$G_0^2 = 2 \ln \left(\frac{L_{\text{full}}}{L_{\text{red}}} \right) = 2 (\ln(L_{\text{full}}) - \ln(L_{\text{red}})) \quad \text{Eq 16.10}$$

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 two 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 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

Example 16.1 Comparing logistic regression models

data=Nocardia

The log likelihoods from four 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 $\beta_0, \beta_1 \dots \beta_{n-1}$	108	0

Overall likelihood ratio test of the full model:

$$G_0^2 = 2(-54.00 - (-74.86)) = 41.72 \text{ with 3 df (P} < 0.001)$$

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

Likelihood ratio test comparing the full and reduced models:

$$G_0^2 = 2(-54.00 - (-69.07)) = 30.16 \text{ with 2 df (P} < 0.001)$$

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

Likelihood ratio test comparing the saturated and full models:

$$G_0^2 = 2(0 - (-54.00)) = 108.00 \text{ with 104 df.}$$

Note This does not have a χ^2 distribution.

copyrighted material

be 1 (or $\ln(L_{\text{sat}}) = 0$). 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.

$$D = 2 \ln \left(\frac{L_{\text{sat}}}{L_{\text{full}}} \right) = 2(\ln(L_{\text{sat}}) - \ln(L_{\text{full}})) = -2(\ln(L_{\text{full}})) \quad \text{Eq 16.11}$$

Note The deviance computed in this manner does not have a χ^2 distribution. (See section 16.11.2 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 may 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 to 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 ($G_0^2=30.16$), 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.

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

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).

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 two levels of -dbarn-. The first table presents the effects of the predictors on the logit of the outcome (case-control status), while the second presents the same results expressed as odds ratios.

Number of obs = 108
LR chi2 (5) = 47.40
Prob > chi2 = 0.000
Log likelihood = -51.168

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: $(1.022)^{(75-50)}=1.73$. An increase of 25% in the percentage of cows dry-treated increases the risk of disease by about 73% (*ie* 1.73 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 tests of the two included categories were 0.08 and 0.06, respectively, suggesting that barn type was only borderline significant ($0.1 > P > 0.05$).