

# Logistic Regression

The goal of a logistic regression analysis is to find the best fitting and most parsimonious, yet biologically reasonable, model to describe the relationship between an outcome (dependent or response variable) and a set of independent (predictor or explanatory) variables. What distinguishes the logistic regression model from the **linear regression** model is that the outcome variable in logistic regression is categorical and most usually *binary* or *dichotomous* (see **Binary Data**).

In any regression problem the key quantity is the mean value of the outcome variable, given the value of the independent variable. This quantity is called the *conditional mean* and will be expressed as  $E(Y|x)$ , where  $Y$  denotes the outcome variable and  $x$  denotes a value of the independent variable. In linear regression we assume that this mean may be expressed as an equation linear in  $x$  (or some transformation of  $x$  or  $Y$ ), such as

$$E(Y|x) = \beta_0 + \beta_1 x.$$

This expression implies that it is possible for  $E(Y|x)$  to take on any value as  $x$  ranges between  $-\infty$  and  $+\infty$ .

Many distribution functions have been proposed for use in the analysis of a dichotomous outcome variable. Cox & Snell [2] discuss some of these. There are two primary reasons for choosing the logistic distribution. These are: (i) from a mathematical point of view it is an extremely flexible and easily used function, and (ii) it lends itself to a biologically meaningful interpretation.

To simplify notation, let  $\pi(x) = E(Y|x)$  represent the conditional mean of  $Y$  given  $x$ . The logistic regression model can be expressed as

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}. \quad (1)$$

The *logit transformation*, defined in terms of  $\pi(x)$ , is as follows:

$$g(x) = \ln \left[ \frac{\pi(x)}{1 - \pi(x)} \right] = \beta_0 + \beta_1 x. \quad (2)$$

The importance of this transformation is that  $g(x)$  has many of the desirable properties of a linear regression model. The logit,  $g(x)$ , is linear in its parameters,

may be continuous, and may range from  $-\infty$  to  $+\infty$  depending on the range of  $x$ .

The second important difference between the linear and logistic regression models concerns the conditional distribution of the outcome variable. In the linear regression model we assume that an observation of the outcome variable may be expressed as  $y = E(Y|x) + \varepsilon$ . The quantity  $\varepsilon$  is called the *error* and expresses an observation's deviation from the conditional mean. The most common assumption is that  $\varepsilon$  follows a normal distribution with mean zero and some variance that is constant across levels of the independent variable. It follows that the conditional distribution of the outcome variable given  $x$  is normal with mean  $E(Y|x)$ , and a variance that is constant. This is not the case with a dichotomous outcome variable. In this situation we may express the value of the outcome variable given  $x$  as  $y = \pi(x) + \varepsilon$ . Here the quantity  $\varepsilon$  may assume one of two possible values. If  $y = 1$ , then  $\varepsilon = 1 - \pi(x)$  with probability  $\pi(x)$ , and if  $y = 0$ , then  $\varepsilon = -\pi(x)$  with probability  $1 - \pi(x)$ . Thus,  $\varepsilon$  has a distribution with mean zero and variance equal to  $\pi(x)[1 - \pi(x)]$ . That is, the conditional distribution of the outcome variable follows a binomial distribution with probability given by the conditional mean,  $\pi(x)$ .

## Fitting the Logistic Regression Model

Suppose we have a sample of  $n$  independent observations of the pair  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$ , where  $y_i$  denotes the value of a dichotomous outcome variable and  $x_i$  is the value of the independent variable for the  $i$ th subject. Furthermore, assume that the outcome variable has been coded as 0 or 1 representing the absence or presence of the characteristic, respectively. To fit the logistic regression model (1) to a set of data requires that we estimate the values of  $\beta_0$  and  $\beta_1$ , the unknown parameters.

In linear regression the method used most often to estimate unknown parameters is **least squares**. In that method we choose those values of  $\beta_0$  and  $\beta_1$  that minimize the sum of squared deviations of the observed values of  $Y$  from the predicted values based upon the model. Under the usual assumptions for linear regression the least squares method yields estimators with a number of desirable statistical properties. Unfortunately, when the least squares method is applied to a model with a dichotomous outcome the estimators

## 2 Logistic Regression

no longer have these same properties.

The general method of estimation that leads to the least squares function under the linear regression model (when the error terms are normally distributed) is **maximum likelihood**. This is the method used to estimate the logistic regression parameters. In a very general sense the maximum likelihood method yields values for the unknown parameters that maximize the probability of obtaining the observed set of data. To apply this method we must first construct a function called the *likelihood function* (see **Likelihood**). This function expresses the probability of the observed data as a function of the unknown parameters. The *maximum likelihood estimators* of these parameters are chosen to be those values that maximize this function. Thus, the resulting estimators are those that agree most closely with the observed data.

If  $Y$  is coded as 0 or 1, then the expression for  $\pi(x)$  given in (1) provides (for an arbitrary value of  $\beta' = (\beta_0, \beta_1)$ , the vector of parameters) the conditional probability that  $Y$  is equal to 1 given  $x$ . This will be denoted  $\Pr(Y = 1|x)$ . It follows that the quantity  $1 - \pi(x)$  gives the conditional probability that  $Y$  is equal to zero given  $x$ ,  $\Pr(Y = 0|x)$ . Thus, for those pairs  $(x_i, y_i)$ , where  $y_i = 1$ , the contribution to the likelihood function is  $\pi(x_i)$ , and for those pairs where  $y_i = 0$ , the contribution to the likelihood function is  $1 - \pi(x_i)$ , where the quantity  $\pi(x_i)$  denotes the value of  $\pi(x)$  computed at  $x_i$ . A convenient way to express the contribution to the likelihood function for the pair  $(x_i, y_i)$  is through the term

$$\xi(x_i) = \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i}. \quad (3)$$

Since the observations are assumed to be independent, the likelihood function is obtained as the product of the terms given in (3) as follows:

$$l(\beta) = \prod_{i=1}^n \xi(x_i). \quad (4)$$

The principle of maximum likelihood states that we use as our estimate of  $\beta$  the value that maximizes the expression in (4). However, it is easier mathematically to work with the log of (4). This expression, the *log likelihood*, is defined as

$$L(\beta) = \ln[l(\beta)] \\ = \sum \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\}. \quad (5)$$

To find the value of  $\beta$  that maximizes  $L(\beta)$  we differentiate  $L(\beta)$  with respect to  $\beta_0$  and  $\beta_1$  and set the resulting expressions equal to zero. These equations are as follows:

$$\sum_{i=1}^n [y_i - \pi(x_i)] = 0 \quad (6)$$

and

$$\sum_{i=1}^n x_i [y_i - \pi(x_i)] = 0, \quad (7)$$

and are called the *likelihood equations*.

In linear regression, the likelihood equations, obtained by differentiating the sum of squared deviations function with respect to  $\beta$ , are linear in the unknown parameters, and thus are easily solved. For logistic regression the expressions in (6) and (7) are nonlinear in  $\beta_0$  and  $\beta_1$ , and thus require special methods for their solution. These methods are iterative in nature and have been programmed into available logistic regression software. McCullagh & Nelder [6] discuss the iterative methods used by most programs. In particular, they show that the solution to (6) and (7) may be obtained using a generalized weighted least squares procedure.

The value of  $\beta$  given by the solution to (6) and (7) is called the maximum likelihood estimate, denoted as  $\hat{\beta}$ . Similarly,  $\hat{\pi}(x_i)$  is the maximum likelihood estimate of  $\pi(x_i)$ . This quantity provides an estimate of the conditional probability that  $Y$  is equal to 1, given that  $x$  is equal to  $x_i$ . As such, it represents the fitted or predicted value for the logistic regression model. An interesting consequence of (6) is that

$$\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{\pi}(x_i).$$

That is, the sum of the observed values of  $y$  is equal to the sum of the predicted (expected) values.

After estimating the coefficients, it is standard practice to assess the significance of the variables in the model. This usually involves testing a statistical hypothesis to determine whether the independent variables in the model are “significantly” related to the outcome variable. One approach to testing for the significance of the coefficient of a variable in any model relates to the following question. *Does the model that includes the variable in question tell us more about the outcome (or response) variable than*

does a model that does not include that variable? This question is answered by comparing the observed values of the response variable with those predicted by each of two models; the first with and the second without the variable in question. The mathematical function used to compare the observed and predicted values depends on the particular problem. If the predicted values with the variable in the model are better, or more accurate in some sense, than when the variable is not in the model, then we feel that the variable in question is “significant”. It is important to note that we are not considering the question of whether the predicted values are an accurate representation of the observed values in an absolute sense (this would be called *goodness of fit*). Instead, our question is posed in a relative sense.

For the purposes of assessing the significance of an independent variable we compute the value of the following statistic:

$$G = -2 \ln \left( \frac{\text{likelihood without the variable}}{\text{likelihood with the variable}} \right). \quad (8)$$

Under the hypothesis that  $\beta_1$  is equal to zero, the statistic  $G$  will follow a chi-square distribution with one degree of freedom. The calculation of the log likelihood and this generalized **likelihood ratio test** are standard features of any good logistic regression package. This makes it possible to check for the significance of the addition of new terms to the model as a matter of routine. In the simple case of a single independent variable, we can first fit a model containing only the constant term. We can then fit a model containing the independent variable along with the constant. This gives rise to a new log likelihood. The likelihood ratio test is obtained by multiplying the difference between the log likelihoods of the two models by  $-2$ .

Another test that is often carried out is the Wald test, which is obtained by comparing the maximum likelihood estimate of the slope parameter,  $\hat{\beta}_1$ , with an estimate of its standard error (see **Likelihood**). The resulting ratio

$$W = \frac{\hat{\beta}_1}{\widehat{\text{se}}(\hat{\beta}_1)},$$

under the hypothesis that  $\beta_1 = 0$ , follows a standard normal distribution. Standard errors of the estimated parameters are routinely printed out by computer

software. Hauck & Donner [3] examined the performance of the Wald test and found that it behaved in an aberrant manner, often failing to reject when the coefficient was significant. They recommended that the likelihood ratio test be used. Jennings [5] has also looked at the adequacy of inferences in logistic regression based on Wald statistics. His conclusions are similar to those of Hauck & Donner.

Both the likelihood ratio test,  $G$ , and the Wald test,  $W$ , require the computation of the maximum likelihood estimate for  $\beta_1$ . For a single variable this is not a difficult or costly computational task. However, for large data sets with many variables, the iterative computation needed to obtain the maximum likelihood estimates can be considerable.

The logistic regression model may be used with matched study designs. Fitting **conditional logistic regression** models requires modifications, which are not discussed here. The reader interested in the conditional logistic regression model may find details in [4, Chapter 7].

### The Multiple Logistic Regression Model

Consider a collection of  $p$  independent variables which will be denoted by the vector  $\mathbf{x}' = (x_1, x_2, \dots, x_p)$ . Assume for the moment that each of these variables is at least interval scaled. Let the conditional probability that the outcome is present be denoted by  $\Pr(Y = 1 | \mathbf{x}) = \pi(\mathbf{x})$ . Then the logit of the multiple logistic regression model is given by

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p, \quad (9)$$

in which case

$$\pi(x) = \frac{\exp[g(\mathbf{x})]}{1 + \exp[g(\mathbf{x})]}. \quad (10)$$

If some of the independent variables are discrete, nominal scaled variables (see **Nominal Data**) such as race, sex, treatment group, and so forth, then it is inappropriate to include them in the model as if they were interval scaled. In this situation a collection of *design variables* (or **dummy variables**) should be used. Most logistic regression software will generate the design variables, and some programs have a choice of several different methods.

In general, if a nominal scaled variable has  $k$  possible values, then  $k - 1$  design variables will be

## 4 Logistic Regression

**Table 1** Code sheet for the variables in the low birth weight data set

Variable	Abbreviation
Identification code	ID
Low birth weight (0 = birth weight $\geq$ 2500 g, 1 = birth weight $<$ 2500 g)	LOW
Age of the mother in years	AGE
Weight in pounds at the last menstrual period	LWT
Race (1 = white, 2 = black, 3 = other)	RACE
Smoking status during pregnancy (1 = yes, 0 = no)	SMOKE
History of premature labor (0 = none, 1 = one, etc.)	PTL
History of hypertension (1 = yes, 0 = no)	HT
Presence of uterine irritability (1 = yes, 0 = no)	UI
Number of physician visits during the first trimester (0 = none, 1 = one, 2 = two, etc.)	FTV
Birth weight (g)	BWT

needed. Suppose, for example, that the  $j$ th independent variable,  $x_j$  has  $k_j$  levels. The  $k_j - 1$  design variables will be denoted as  $D_{ju}$  and the coefficients for these design variables will be denoted as  $\beta_{ju}$ ,  $u = 1, 2, \dots, k_j - 1$ . Thus, the logit for a model with  $p$  variables and the  $j$ th variable being discrete is

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \dots + \sum_{u=1}^{k_j-1} \beta_{ju} D_{ju} + \beta_p x_p.$$

and

$$\sum_{i=1}^n x_{ij} [y_i - \pi(\mathbf{x}_i)] = 0,$$

for  $j = 1, 2, \dots, p$ .

### Fitting the Multiple Logistic Regression Model

Assume that we have a sample of  $n$  independent observations of the pair  $(\mathbf{x}_i, y_i)$ ,  $i = 1, 2, \dots, n$ . As in the univariate case, fitting the model requires that we obtain estimates of the vector  $\beta' = (\beta_0, \beta_1, \dots, \beta_p)$ . The method of estimation used in the multivariate case is the same as in the univariate situation, i.e. maximum likelihood. The likelihood function is nearly identical to that given in (4), with the only change being that  $\pi(\mathbf{x})$  is now defined as in (10). There are  $p + 1$  likelihood equations which are obtained by differentiating the log likelihood function with respect to the  $p + 1$  coefficients. The likelihood equations that result may be expressed as follows:

$$\sum_{i=1}^n [y_i - \pi(\mathbf{x}_i)] = 0$$

**Table 3** Estimated coefficients for a multiple logistic regression model using all variables from the low birth weight data set

Variable	Coeff.	Std. error	$z$	Pr > $ z $	[95% conf. interval]	
AGE	-0.035	0.039	-0.920	0.357	-0.111	0.040
LWT	-0.015	0.007	-2.114	0.035	-0.029	-0.001
SMOKE	0.815	0.420	1.939	0.053	-0.009	1.639
HT	1.824	0.705	2.586	0.010	0.441	3.206
UI	0.702	0.465	1.511	0.131	-0.208	1.613
RACE 1	1.202	0.534	2.253	0.024	0.156	2.248
RACE 2	0.773	0.460	1.681	0.093	-0.128	1.674
FTV01	0.121	0.376	0.323	0.746	-0.615	0.858
PTL01	1.237	0.466	2.654	0.008	0.323	2.148
cons	0.545	1.266	0.430	0.667	-1.937	3.027

**Table 2** Coding of design variables for RACE

RACE	Design variable	
	RACE 1	RACE 2
White	0	0
Black	1	0
Other	0	1

variables are called FTV01 and PTL01.

The results of fitting the logistic regression model to these data are given in Table 3.

In Table 3 the estimated coefficients for the two design variables for race are indicated in the lines denoted by “RACE 1” and “RACE 2”. The estimated logit is given by

$$\begin{aligned} \hat{g}(\mathbf{x}) = & 0.545 - 0.035 \times \text{AGE} - 0.015 \times \text{LWT} \\ & + 0.815 \times \text{SMOKE} + 1.824 \times \text{HT} + 0.702 \\ & \times \text{UI} + 1.202 \times \text{RACE 1} + 0.773 \times \text{RACE 2} \\ & + 0.121 \times \text{FTV01} + 1.237 \times \text{PTL01}. \end{aligned}$$

The fitted values are obtained using the estimated logit,  $\hat{g}(\mathbf{x})$ , as in (10).

### Testing for the Significance of the Model

Once we have fit a particular multiple (multivariate) logistic regression model, we begin the process of

assessment of the model. The first step in this process is usually to assess the significance of the variables in the model. The likelihood ratio test for overall significance of the  $p$  coefficients for the independent variables in the model is performed based on the statistic  $G$  given in (8). The only difference is that the fitted values,  $\hat{\pi}$ , under the model are based on the vector containing  $p + 1$  parameters,  $\hat{\beta}$ . Under the null hypothesis that the  $p$  “slope” coefficients for the covariates in the model are equal to zero, the distribution of  $G$  is **chi-square** with  $p$  **degrees of freedom**.

As an example, consider the fitted model whose estimated coefficients are given in Table 3. For that model the value of the log likelihood is  $L = -98.36$ . A second model, fit with the constant term only, yields  $L = -117.336$ . Hence  $G = -2[(-117.34) - (-98.36)] = 37.94$  and the **P value** for the test is  $\Pr[\chi^2(9) > 37.94] < 0.0001$  (see Table 3). Rejection of the **null hypothesis** (that all of the coefficients are simultaneously equal to zero) has an interpretation analogous to that in multiple linear regression; we may conclude that at least one, and perhaps all  $p$  coefficients are different from zero.

Before concluding that any or all of the coefficients are nonzero, we may wish to look at the univariate Wald test statistics,  $W_j = \hat{\beta}_j / \widehat{\text{se}}(\hat{\beta}_j)$ . These are given in the fourth column (labeled  $z$ ) in Table 3. Under the hypothesis that an individual coefficient is zero, these statistics will follow the **standard normal** distribution. Thus, the value of these statistics may give us an indication of which of the variables

## 6 Logistic Regression

**Table 4** Estimated coefficients for a multiple logistic regression model using the variables LWT, SMOKE, HT, PTL01 and RACE from the low birth weight data set

Variable	Coeff.	Std. err.	$z$	Pr > $ z $	[95% conf. interval]	
LWT	-0.017	0.007	-2.407	0.016	-0.030	-0.003
SMOKE	0.876	0.401	2.186	0.029	0.091	1.661
HT	1.767	0.708	2.495	0.013	0.379	3.156
RACE 1	1.264	0.529	2.387	0.017	0.226	2.301
RACE 2	0.864	0.435	1.986	0.047	0.011	1.717
PTL01	1.231	0.446	2.759	0.006	0.357	2.106
cons	0.095	0.957	0.099	0.921	-1.781	1.970

in the model may or may not be significant. If we use a critical value of 2, which leads to an approximate level of significance (two-tailed) of 0.05, then we would conclude that the variables LWT, SMOKE, HT, PTL01 and possibly RACE are significant, while AGE, UI, and FTV01 are not significant.

Considering that the overall goal is to obtain the best fitting model while minimizing the number of parameters, the next logical step is to fit a reduced model, containing only those variables thought to be significant, and compare it with the full model containing all the variables. The results of fitting the reduced model are given in Table 4.

The difference between the two models is the exclusion of the variables AGE, UI, and FTV01 from the full model. The likelihood ratio test comparing these two models is obtained using the definition of  $G$  given in (8). It has a distribution that is chi-square with three degrees of freedom under the hypothesis that the coefficients for the variables excluded are equal to zero. The value of the test statistic comparing the models in Tables 3 and 4 is  $G = -2[(-100.24) - (-98.36)] = 3.76$  which, with three degrees of freedom, has a  $P$  value of  $P[\chi^2(3) > 3.76] = 0.2886$ . Since the  $P$  value is large, exceeding 0.05, we conclude that the reduced model is as good as the full model. Thus there is no advantage to including AGE, UI, and FTV01 in the model. However, we must not base our models entirely on tests of statistical significance. Numerous other considerations should influence our decision to include or exclude variables from a model.

### Interpretation of the Coefficients of the Logistic Regression Model

After fitting a model the emphasis shifts from the computation and assessment of significance of estimated coefficients to interpretation of their values. The interpretation of any fitted model requires that we can draw practical inferences from the estimated coefficients in the model. The question addressed is: *What do the estimated coefficients in the model tell us about the research questions that motivated the study?* For most models this involves the estimated coefficients for the independent variables in the model. The estimated coefficients for the independent variables represent the slope or rate of change of a function of the dependent variable per unit of change in the independent variable. Thus, interpretation involves two issues: (i) determining the functional relationship between the dependent variable and the independent variable, and (ii) appropriately defining the unit of change for the independent variable.

For a linear regression model we recall that the slope coefficient,  $\beta_1$ , is equal to the difference between the value of the dependent variable at  $x + 1$  and the value of the dependent variable at  $x$ , for any value of  $x$ . In the logistic regression model  $\beta_1 = g(x + 1) - g(x)$ . That is, the slope coefficient represents the change in the logit for a change of one unit in the independent variable  $x$ . Proper interpretation of the coefficient in a logistic regression model depends on being able to place meaning on the difference between two logits. Consider the interpretation of the coefficients for a univariate logistic regression model for each of the possible measurement scales

**Table 5** Values of the logistic regression model when the independent variable is dichotomous

		Independent variable	
		X	
		x = 1	x = 0
Outcome variable	Y		
	y = 1	$\pi(1) = \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}$	$\pi(0) = \frac{\exp \beta_0}{1 + \exp \beta_0}$
	y = 0	$1 - \pi(1) = \frac{1}{1 + \exp(\beta_0 + \beta_1)}$	$1 - \pi(0) = \frac{1}{1 + \exp \beta_0}$
Total		1.0	1.0

of the independent variable.

### Dichotomous Independent Variable

Assume that  $x$  is coded as either 0 or 1. Under this model there are two values of  $\pi(x)$  and equivalently two values of  $1 - \pi(x)$ . These values may be conveniently displayed in a  $2 \times 2$  table, as shown in Table 5.

The **odds** of the outcome being present among individuals with  $x = 1$  is defined as  $\pi(1)/[1 - \pi(1)]$ . Similarly, the odds of the outcome being present among individuals with  $x = 0$  is defined as  $\pi(0)/[1 - \pi(0)]$ . The **odds ratio**, denoted by  $\psi$ , is defined as the ratio of the odds for  $x = 1$  to the odds for  $x = 0$ , and is given by

$$\psi = \frac{\pi(1)/[1 - \pi(1)]}{\pi(0)/[1 - \pi(0)]} \quad (11)$$

The log of the odds ratio, termed log odds ratio, or *log odds*, is

$$\ln(\psi) = \ln \left\{ \frac{\pi(1)/[1 - \pi(1)]}{\pi(0)/[1 - \pi(0)]} \right\} = g(1) - g(0),$$

which is the *logit difference*, where the log of the odds is called the logit and, in this example, these are

$$g(1) = \ln\{\pi(1)/[1 - \pi(1)]\}$$

and

$$g(0) = \ln\{\pi(0)/[1 - \pi(0)]\}.$$

Using the expressions for the logistic regression

model shown in Table 5 the odds ratio is

$$\begin{aligned} \psi &= \frac{\left( \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)} \right) \left( \frac{1}{1 + \exp(\beta_0)} \right)}{\left( \frac{\exp(\beta_0)}{1 + \exp(\beta_0)} \right) \left( \frac{1}{1 + \exp(\beta_0 + \beta_1)} \right)} \\ &= \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1). \end{aligned}$$

Hence, for logistic regression with a dichotomous independent variable

$$\psi = \exp(\beta_1), \quad (12)$$

and the logit difference, or log odds, is

$$\ln(\psi) = \ln[\exp(\beta_1)] = \beta_1.$$

This fact concerning the interpretability of the coefficients is the fundamental reason why logistic regression has proven such a powerful analytic tool for epidemiologic research. A confidence interval (CI) estimate for the odds ratio is obtained by first calculating the endpoints of a **confidence interval** for the coefficient  $\beta_1$ , and then exponentiating these values. In general, the endpoints are given by

$$\exp \left[ \hat{\beta}_1 \pm z_{1-\alpha/2} \times \widehat{\text{se}}(\hat{\beta}_1) \right].$$

Because of the importance of the odds ratio as a measure of association, point and interval estimates are often found in additional columns in tables presenting the results of a logistic regression analysis.

In the previous discussion we noted that the estimate of the odds ratio was  $\hat{\psi} = \exp(\hat{\beta}_1)$ . This is correct when the independent variable has been coded as 0 or 1. This type of coding is called “reference

## 8 Logistic Regression

**Table 6** Cross-classification of hypothetical data on RACE and CHD status for 100 subjects

CHD status	White	Black	Hispanic	Other	Total
Present	5	20	15	10	50
Absent	20	10	10	10	50
Total	25	30	25	20	100
Odds ratio ( $\hat{\psi}$ )	1.0	8.0	6.0	4.0	
95% CI		(2.3, 27.6)	(1.7, 21.3)	(1.1, 14.9)	
$\ln(\hat{\psi})$	0.0	2.08	1.79	1.39	

cell” coding. Other coding could be used. For example, the variable may be coded as  $-1$  or  $+1$ . This type of coding is termed “deviation from means” coding. Evaluation of the logit difference shows that the odds ratio is calculated as  $\hat{\psi} = \exp(2\hat{\beta}_1)$  and if an investigator were simply to exponentiate the coefficient from the computer output of a logistic regression analysis, the wrong estimate of the odds ratio would be obtained. Close attention should be paid to the method used to code design variables.

The method of coding also influences the calculation of the endpoints of the confidence interval. With deviation from means coding, the estimated standard error needed for confidence interval estimation is  $\widehat{\text{se}}(2\hat{\beta}_1)$ , which is  $2 \times \widehat{\text{se}}(\hat{\beta}_1)$ . Thus the endpoints of the confidence interval are

$$\exp \left[ 2\hat{\beta}_1 + z_{1-\alpha/2} \times 2 \times \widehat{\text{se}}(\hat{\beta}_1) \right].$$

In summary, for a dichotomous variable the parameter of interest is the odds ratio. An estimate of this parameter may be obtained from the estimated logistic regression coefficient, regardless of how the variable is coded or scaled. This relationship between the logistic regression coefficient and the odds ratio provides the foundation for our interpretation of all logistic regression results.

### Polytomous Independent Variable

Suppose that instead of two categories the independent variable has  $k > 2$  distinct values (*see Polytomous Data*). For example, we may have variables that denote the county of residence within a state, the clinic used for primary health care within a city, or race. Each of these variables has a fixed number of discrete outcomes and the scale of measurement is nominal.

Suppose that in a study of coronary heart disease (CHD) the variable RACE is coded at four levels, and that the cross-classification of RACE by CHD status yields the data presented in Table 6. These data are hypothetical and have been formulated for ease of computation. The extension to a situation where the variable has more than four levels is not conceptually different, so all the examples in this section use  $k = 4$ .

At the bottom of Table 6 the odds ratio is given for each race, using white as the reference group. For example, for hispanic the estimated odds ratio is  $(15 \times 20)/(5 \times 10) = 6.0$ . The log of the odds ratios are given in the last row of Table 6. This display is typical of what is found in the literature when there is a perceived referent group to which the other groups are to be compared. These same estimates of the odds ratio may be obtained from a logistic regression program with an appropriate choice of design variables. The method for specifying the design variables involves setting all of them equal to zero for the reference group, and then setting a single design variable equal to one for each of the other groups. This is illustrated in Table 7.

Use of any logistic regression program with design variables coded as shown in Table 7 yields the estimated logistic regression coefficients given in Table 8.

**Table 7** Specification of the design variables for RACE using white as the reference group

RACE (code)	Design variables		
	$D_1$	$D_2$	$D_3$
White (1)	0	0	0
Black (2)	1	0	0
Hispanic (3)	0	1	0
Other (4)	0	0	1

**Table 8** Results of fitting the logistic regression model to the data in Table 6 using the design variables in Table 7

Variable	Coeff.	Std. err.	$z$	$P >  z $	[95% conf. interval]	
RACE 1	2.079	0.632	3.288	0.001	0.840	3.319
RACE 2	1.792	0.645	2.776	0.006	0.527	3.057
RACE 3	1.386	0.671	2.067	0.039	0.072	2.701
cons	-1.386	0.500	-2.773	0.006	-2.367	-0.406

  

Variable	Odds ratio	[95% conf. interval]	
RACE 1	8	2.32	27.63
RACE 2	6	1.69	21.26
RACE 3	4	1.07	14.90

A comparison of the estimated coefficients in Table 8 with the log odds in Table 6 shows that  $\ln[\hat{\psi}(\text{black, white})] = \hat{\beta}_{11} = 2.079$ ,  $\ln[\hat{\psi}(\text{hispanic, white})] = \hat{\beta}_{12} = 1.792$ , and  $\ln[\hat{\psi}(\text{other, white})] = \hat{\beta}_{13} = 1.386$ .

In the univariate case the estimates of the standard errors found in the logistic regression output are identical to the estimates obtained using the cell frequencies from the contingency table. For example, the estimated standard error of the estimated coefficient for design variable (1),  $\hat{\beta}_{11}$ , is  $0.6325 = (1/5 + 1/20 + 1/20 + 1/10)^{1/2}$ . A derivation of this result appears in Bishop et al. [1].

Confidence limits for odds ratios may be obtained as follows:

$$\hat{\beta}_{ij} \pm z_{1-\alpha/2} \times \widehat{\text{se}}(\hat{\beta}_{ij}).$$

The corresponding limits for the odds ratio are obtained by exponentiating these limits as follows:

$$\exp[\hat{\beta}_{ij} \pm z_{1-\alpha/2} \times \widehat{\text{se}}(\hat{\beta}_{ij})].$$

### Continuous Independent Variable

When a logistic regression model contains a continuous independent variable, interpretation of the estimated coefficient depends on how it is entered into the model and the particular units of the variable. For purposes of developing the method to interpret the coefficient for a continuous variable, we assume that the logit is linear in the variable.

Under the assumption that the logit is linear in the continuous covariate,  $x$ , the equation for the logit is  $g(x) = \beta_0 + \beta_1 x$ . It follows that the slope coefficient,

$\beta_1$ , gives the change in the log odds for an increase of “1” unit in  $x$ , i.e.  $\beta_1 = g(x+1) - g(x)$  for any value of  $x$ . Most often the value of “1” will not be biologically very interesting. For example, an increase of 1 year in age or of 1 mmHg in systolic blood pressure may be too small to be considered important. A change of 10 years or 10 mmHg might be considered more useful. However, if the range of  $x$  is from zero to one, as might be the case for some created index, then a change of 1 is too large and a change of 0.01 may be more realistic. Hence, to provide a useful interpretation for continuous scaled covariates we need to develop a method for point and interval estimation for an arbitrary change of  $c$  units in the covariate.

The log odds for a change of  $c$  units in  $x$  is obtained from the logit difference  $g(x+c) - g(x) = c\beta_1$  and the associated odds ratio is obtained by exponentiating this logit difference,  $\psi(c) = \psi(x+c, x) = \exp(c\beta_1)$ . An estimate may be obtained by replacing  $\beta_1$  with its maximum likelihood estimate,  $\hat{\beta}_1$ . An estimate of the standard error needed for confidence interval estimation is obtained by multiplying the estimated standard error of  $\hat{\beta}_1$  by  $c$ . Hence the endpoints of the  $100(1-\alpha)\%$  CI estimate of  $\psi(c)$  are

$$\exp[c\hat{\beta}_1 \pm z_{1-\alpha/2} c \widehat{\text{se}}(\hat{\beta}_1)].$$

Since both the point estimate and endpoints of the confidence interval depend on the choice of  $c$ , the particular value of  $c$  should be clearly specified in all tables and calculations.

## 10 Logistic Regression

### Multivariate Case

Often logistic regression analysis is used to *adjust statistically* the estimated effects of each variable in the model for differences in the distributions of and associations among the other independent variables. Applying this concept to a multiple logistic regression model, we may surmise that each estimated coefficient provides an estimate of the log odds adjusting for all other variables included in the model. The term confounder is used by epidemiologists to describe a covariate that is associated with both the outcome variable of interest and a primary independent variable or risk factor. When both associations are present the relationship between the risk factor and the outcome variable is said to be *confounded* (see **Confounder**). The procedure for adjusting for confounding is appropriate when there is no interaction.

If the association between the covariate and an outcome variable is the same within each level of the risk factor, then there is no interaction between the covariate and the risk factor. When interaction is present, the association between the risk factor and the outcome variable differs, or depends in some way on the level of the covariate. That is, the covariate modifies the effect of the risk factor (see **Effect Modification**). Epidemiologists use the term effect modifier to describe a variable that interacts with a risk factor.

The simplest and most commonly used model for including interaction is one in which the logit is also linear in the confounder for the second group, but with a different slope. Alternative models can be formulated which would allow for other than a linear

relationship between the logit and the variables in the model within each group. In any model, interaction is incorporated by the inclusion of appropriate higher order terms.

An important step in the process of modeling a set of data is to determine whether or not there is evidence of interaction in the data. Tables 9 and 10 present the results of fitting a series of logistic regression models to two different sets of hypothetical data. The variables in each of the data sets are the same: SEX, AGE, and CHD. In addition to the estimated coefficients, the log likelihood for each model and minus twice the change (deviance) is given. Recall that minus twice the change in the log likelihood may be used to test for the significance of coefficients for variables added to the model. An interaction is added to the model by creating a variable that is equal to the product of the value of the sex and the value of age.

Examining the results in Table 9 we see that the estimated coefficient for the variable SEX changed from 1.535 in model 1 to 0.979 when AGE was added in model 2. Hence, there is clear evidence of a confounding effect owing to age. When the interaction term “SEX  $\times$  AGE” is added in model 3 we see that the change in the deviance is only 0.52 which, when compared with the chi-square distribution with one degree of freedom, yields a *P* value of 0.47, which clearly is not significant. Note that the coefficient for sex changed from 0.979 to 0.481. This is not surprising since the inclusion of an interaction term, especially when it involves a continuous variable, will usually produce fairly marked changes in the estimated coefficients of dichotomous variables involved in the interaction. Thus, when an interaction

**Table 9** Estimated logistic regression coefficients, log likelihood, and the likelihood ratio test statistic (*G*) for an example showing evidence of confounding but no interaction

Model	Constant	SEX	AGE	SEX $\times$ AGE	Log likelihood	<i>G</i>
1	-1.046	1.535			-61.86	
2	-7.142	0.979	0.167		-49.59	24.54
3	-6.103	0.481	0.139	0.059	-49.33	0.52

**Table 10** Estimated logistic regression coefficients, log likelihood, and the likelihood ratio test statistic (*G*) for an example showing evidence of confounding and interaction

Model	Constant	SEX	AGE	SEX $\times$ AGE	Log likelihood	<i>G</i>
1	-0.847	2.505			-52.52	
2	-6.194	1.734	0.147		-46.79	11.46
3	-3.105	0.047	0.629	0.206	-44.76	4.06

term is present in the model we cannot assess confounding via the change in a coefficient. For these data we would prefer to use model 2 which suggests that age is a confounder but not an effect modifier.

The results in Table 10 show evidence of both confounding and interaction due to age. Comparing model 1 with model 2 we see that the coefficient for sex changes from 2.505 to 1.734. When the age by sex interaction is added to the model we see that the deviance is 4.06, which yields a  $P$  value of 0.04. Since the deviance is significant, we prefer model 3 over model 2, and should regard age as both a confounder and an effect modifier. The net result is that any estimate of the odds ratio for sex should be made with respect to a specific age.

Hence, we see that determining if a covariate,  $X$ , is an effect modifier and/or a confounder involves several issues. Determining effect modification status involves the parametric structure of the logit, while determination of confounder status involves two things. First, the covariate must be associated with the outcome variable. This implies that the logit must have a nonzero slope in the covariate. Secondly, the covariate must be associated with the risk factor. In our example this might be characterized by having a difference in the mean age for males and females. However, the association may be more complex than a simple difference in means. The essence is that we have incomparability in our risk factor groups. This incomparability must be accounted for in the model if we are to obtain a correct, unconfounded estimate of effect for the risk factor.

In practice, the confounder status of a covariate is ascertained by comparing the estimated coefficient for the risk factor variable from models containing and not containing the covariate. Any “biologically important” change in the estimated coefficient for the risk factor would dictate that the covariate is a confounder and should be included in the model, regardless of the statistical significance of the estimated coefficient for the covariate. On the other hand, a covariate is an effect modifier only when the

interaction term added to the model is both biologically meaningful and statistically significant. When a covariate is an effect modifier, its status as a confounder is of secondary importance since the estimate of the effect of the risk factor depends on the specific value of the covariate.

The concepts of adjustment, confounding, interaction, and effect modification may be extended to cover the situations involving any number of variables on any measurement scale(s). The principles for identification and inclusion of confounder and interaction variables into the model are the same regardless of the number of variables and their measurement scales.

Much of this article has been abstracted from [4]. Readers wanting more detail on any topic should consult this reference.

### References

- [1] Bishop, Y.M.M., Fienberg, S.E. & Holland, P. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Boston.
- [2] Cox, D.R. & Snell, E.J. (1989). *The Analysis of Binary Data*, 2nd Ed. Chapman & Hall, London.
- [3] Hauck, W.W. & Donner, A. (1977). Wald’s Test as applied to hypotheses in logit analysis, *Journal of the American Statistical Association* **72**, 851–853.
- [4] Hosmer, D. & Lemeshow, S. (1989). *Applied Logistic Regression*. Wiley, New York.
- [5] Jennings, D.E. (1986). Judging inference adequacy in logistic regression, *Journal of the American Statistical Association* **81**, 471–476.
- [6] McCullagh, P. & Nelder, J.A. (1983). *Generalized Linear Models*. Chapman & Hall, London.

(See also **Categorical Data Analysis; Loglinear Model; Proportional-odds Model; Quantal Response Models**)

STANLEY LEMESHOW & DAVID  
W. HOSMER, JR