Applied Logistic Regression

Second Edition

DAVID W. HOSMER
University of Massachusetts
Amherst, Massachusetts

STANLEY LEMESHOW
The Ohio State University
Columbus, Ohio

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.
New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

CHAPTER 1

Introduction to the Logistic Regression Model

1.1 INTRODUCTION

Regression methods have become an integral component of any data analysis concerned with describing the relationship between a response variable and one or more explanatory variables. It is often the case that the outcome variable is discrete, taking on two or more possible values. Over the last decade the logistic regression model has become, in many fields, the standard method of analysis in this situation.

Before beginning a study of logistic regression it is important to understand that the goal of an analysis using this method is the same as that of any model-building technique used in statistics: 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. These independent variables are often called *covariates*. The most common example of modeling, and one assumed to be familiar to the readers of this text, is the usual linear regression model where the outcome variable is assumed to be continuous.

What distinguishes a logistic regression model from the linear regression model is that the outcome variable in logistic regression is *binary* or *dichotomous*. This difference between logistic and linear regression is reflected both in the choice of a parametric model and in the assumptions. Once this difference is accounted for, the methods employed in an analysis using logistic regression follow the same general principles used in linear regression. Thus, the techniques used in linear regression analysis will motivate our approach to logistic regression. We illustrate both the similarities and differences between logistic regression and linear regression with an example.

Example

Table 1.1 lists age in years (AGE), and presence or absence of evidence of significant coronary heart disease (CHD) for 100 subjects selected to participate in a study. The table also contains an identifier variable (ID) and an age group variable (AGRP). The outcome variable is CHD, which is coded with a value of zero to indicate CHD is absent, or 1 to indicate that it is present in the individual.

It is of interest to explore the relationship between age and the presence or absence of CHD in this study population. Had our outcome variable been continuous rather than binary, we probably would begin by forming a scatterplot of the outcome versus the independent variable. We would use this scatterplot to provide an impression of the nature and strength of any relationship between the outcome and the independent variable. A scatterplot of the data in Table 1.1 is given in Figure 1.1.

In this scatterplot all points fall on one of two parallel lines representing the absence of CHD ($y=0$) and the presence of CHD ($y=1$). There is some tendency for the individuals with no evidence of CHD to be younger than those with evidence of CHD. While this plot does depict the dichotomous nature of the outcome variable quite clearly, it does not provide a clear picture of the nature of the relationship between CHD and age.

A problem with Figure 1.1 is that the variability in CHD at all ages is large. This makes it difficult to describe the functional relationship between age and CHD. One common method of removing some variation while still maintaining the structure of the relationship between the outcome and the independent variable is to create intervals for the independent variable and compute the mean of the outcome variable within each group. In Table 1.2 this strategy is carried out by using the age group variable, AGRP, which categorizes the age data of Table 1.1. Table 1.2 contains, for each age group, the frequency of occurrence of each outcome as well as the mean (or proportion with CHD present) for each group.

By examining this table, a clearer picture of the relationship begins to emerge. It appears that as age increases, the proportion of individuals with evidence of CHD increases. Figure 1.2 presents a plot of the proportion of individuals with CHD versus the midpoint of each age interval. While this provides considerable insight into the relationship between CHD and age in this study, a functional form for this relationship needs to be described. The plot in this figure is similar to what one

Table 1.1 Age and Coronary Heart Disease (CHD) Status of 100 Subjects

ID	AGE	AGRP	CHD	ID	AGE	AGRP	CHD
1	20	1	0	51	44	4	1
2	23	1	0	52	44	4	1
3	24	1	0	53	45	5	0
4	25	1	0	54	45	5	1
5	25	1	1	55	46	5	0
6	26	1	0	56	46	5	1
7	26	1	0	57	47	5	0
8	28	1	0	58	47	5	0
9	28	1	0	59	47	5	1
10	29	1	0	60	48	5	0
11	30	2	0	61	48	5	1
12	30	2	0	62	48	5	1
13	30	2	0	63	49	5	0
14	30	2	0	64	49	5	0
15	30	2	0	65	49	5	1
16	30	2	1	66	50	6	0
17	32	2	0	67	50	6	1
18	32	2	0	68	51	6	0
19	33	2	0	69	52	6	0
20	33	2	0	70	52	6	1
21	34	2	0	71	53	6	1
22	34	2	0	72	53	6	1
23	34	2	1	73	54	6	1
24	34	2	0	74	55	7	0
25	34	2	0	75	55	7	1
26	35	3	0	76	55	7	1
27	35	3	0	77	56	7	1
28	36	3	0	78	56	7	1
29	36	3	1	79	56	7	1
30	36	3	0	80	57	7	0
31	37	3	0	81	57	7	0
32	37	3	1	82	57	7	1
33	37	3	0	83	57	7	1
34	38	3	0	84	57	7	1
35	38	3	0	85	57	7	1
36	39	3	0	86	58	7	0
37	39	3	1	87	58	7	1
38	40	4	0	88	58	7	1
39	40	4	1	89	59	7	1
40	41	4	0	90	59	7	1
41	41	4	0	91	60	8	0
42	42	4	0	92	60	8	1
43	42	4	0	93	61	8	1
44	42	4	0	94	62	8	1
45	42	4	1	95	62	8	1
46	43	4	0	96	63	8	1
47	43	4	0	97	64	8	0
48	43	4	1	98	64	8	1
49	44	4	0	99	65	8	1
50	44	4	0	100	69	8	1

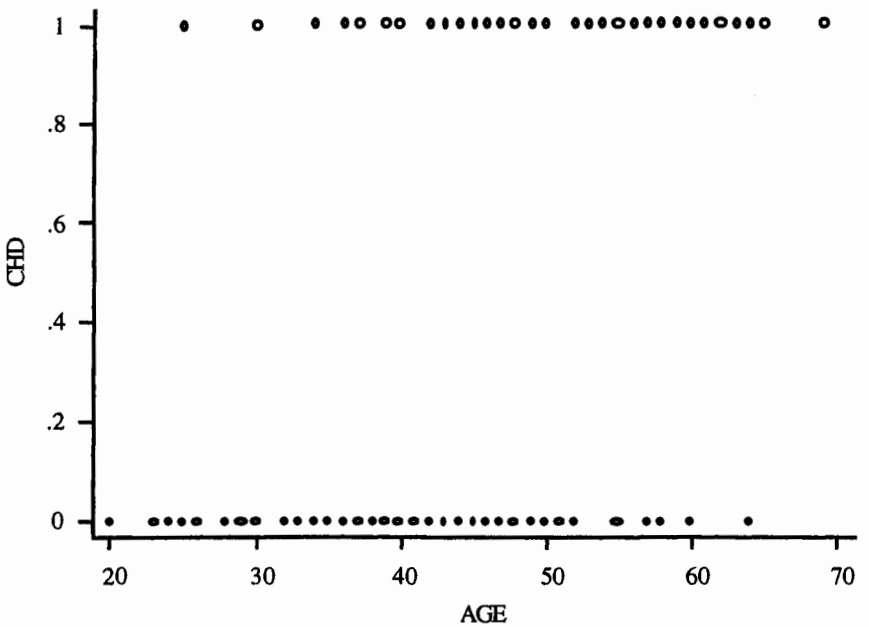

Figure 1.1 Scatterplot of CHD by AGE for 100 subjects.

might obtain if this same process of grouping and averaging were performed in a linear regression. We will note two important differences.

The first difference concerns the nature of the relationship between the outcome and independent variables. 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 "$E(Y \mid x)$" where Y denotes the outcome

Table 1.2 Frequency Table of Age Group by CHD

		CHD		
Age Group	n	Absent	Present	Mean (Proportion)
20 – 29	10	9	1	0.10
30 – 34	15	13	2	0.13
35 – 39	12	9	3	0.25
40 – 44	15	10	5	0.33
45 – 49	13	7	6	0.46
50 – 54	8	3	5	0.63
55 – 59	17	4	13	0.76
60 – 69	10	2	8	0.80
Total	100	57	43	0.43

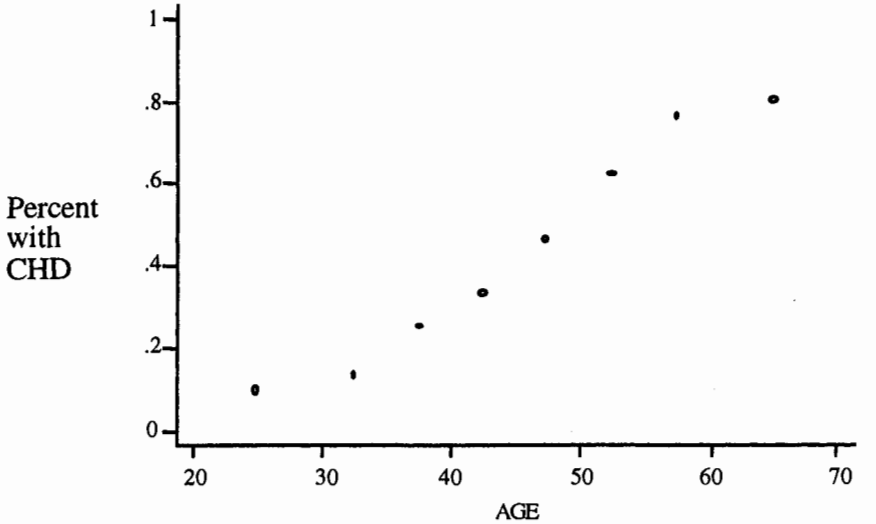

Figure 1.2 Plot of the percentage of subjects with CHD in each age group.

variable and x denotes a value of the independent variable. The quantity $E(Y \mid x)$ is read "the expected value of Y, given the value x." 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

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

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

The column labeled "Mean" in Table 1.2 provides an estimate of $E(Y \mid x)$. We will assume, for purposes of exposition, that the estimated values plotted in Figure 1.2 are close enough to the true values of $E(Y \mid x)$ to provide a reasonable assessment of the relationship between CHD and age. With dichotomous data, the conditional mean must be greater than or equal to zero and less than or equal to 1 [i.e., $0 \leq E(Y \mid x) \leq 1$]. This can be seen in Figure 1.2. In addition, the plot shows that this mean approaches zero and 1 "gradually." The change in the $E(Y \mid x)$ per unit change in x becomes progressively smaller as the conditional mean gets closer to zero or 1. The curve is said to be *S-shaped.* It resembles a plot of a cumulative distribution of a random variable. It

should not seem surprising that some well-known cumulative distributions have been used to provide a model for $E(Y \mid x)$ in the case when Y is dichotomous. The model we will use is that of the logistic distribution.

Many distribution functions have been proposed for use in the analysis of a dichotomous outcome variable. Cox and Snell (1989) discuss some of these. There are two primary reasons for choosing the logistic distribution. First, from a mathematical point of view, it is an extremely flexible and easily used function, and second, it lends itself to a clinically meaningful interpretation. A detailed discussion of the interpretation of the model parameters is given in Chapter 3.

In order to simplify notation, we use the quantity $\pi(x) = E(Y \mid x)$ to represent the conditional mean of Y given x when the logistic distribution is used. The specific form of the logistic regression model we use is:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} . \tag{1.1}$$

A transformation of $\pi(x)$ that is central to our study of logistic regression is the *logit transformation*. This transformation is defined, in terms of $\pi(x)$, as:

$$\begin{aligned} g(x) &= \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] \\ &= \beta_0 + \beta_1 x . \end{aligned}$$

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 $y = E(Y \mid x) + \varepsilon$. The quantity ε is called the *error* and expresses an observation's deviation from the conditional mean. The most common assumption is that ε 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 will be normal with mean $E(Y \mid x)$, 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 ε 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, ε 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, $\pi(x)$.

In summary, we have seen that in a regression analysis when the outcome variable is dichotomous:

(1) The conditional mean of the regression equation must be formulated to be bounded between zero and 1. We have stated that the logistic regression model, $\pi(x)$ given in equation (1.1), satisfies this constraint.

(2) The binomial, not the normal, distribution describes the distribution of the errors and will be the statistical distribution upon which the analysis is based.

(3) The principles that guide an analysis using linear regression will also guide us in logistic regression.

1.2 FITTING THE LOGISTIC REGRESSION MODEL

Suppose we have a sample of n independent observations of the pair (x_i, y_i), $i = 1, 2, \ldots, n$, 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 the presence of the characteristic, respectively. This coding for a dichotomous outcome is used throughout the text. To fit the logistic regression model in equation (1.1) to a set of data requires that we estimate the values of β_0 and β_1, the unknown parameters.

In linear regression, the method used most often for estimating unknown parameters is *least squares*. In that method we choose those values of β_0 and β_1 which 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 method of least squares yields estimators with a number of desirable statistical proper-

CHAPTER 3

Interpretation of the Fitted Logistic Regression Model

3.1 INTRODUCTION

In Chapters 1 and 2 we discussed the methods for fitting and testing for the significance of the logistic regression model. After fitting a model the emphasis shifts from the computation and assessment of significance of the estimated coefficients to the interpretation of their values. Strictly speaking, an assessment of the adequacy of the fitted model should precede any attempt at interpreting it. In the case of logistic regression the methods for assessment of fit are rather technical in nature and thus are deferred until Chapter 5, at which time the reader should have a good working knowledge of the logistic regression model. Thus, we begin this chapter assuming that a logistic regression model has been fit, that the variables in the model are significant in either a clinical or statistical sense, and that the model fits according to some statistical measure of fit.

The interpretation of any fitted model requires that we be able to draw practical inferences from the estimated coefficients in the model. The question being 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. On occasion, the intercept coefficient is of interest; but this is the exception, not the rule. The estimated coefficients for the independent variables represent the slope (i.e., rate of change) of a function of the dependent variable per unit of change in the independent variable. Thus, interpretation involves two issues: determining the functional relationship between the dependent variable and the independent variable, and appropriately defining the unit of change for the independent variable.

The first step is to determine what function of the dependent variable yields a linear function of the independent variables. This is called the *link function* [see McCullagh and Nelder (1983) or Dobson (1990)]. In the case of a linear regression model, it is the identity function since the dependent variable, by definition, is linear in the parameters. (For those unfamiliar with the term "identity function," it is the function $y = y$.) In the logistic regression model the link function is the logit transformation $g(x) = \ln\{\pi(x)/[1-\pi(x)]\} = \beta_0 + \beta_1 x$.

For a linear regression model recall that the slope coefficient, β_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. For example, if $y(x) = \beta_0 + \beta_1 x$, it follows that $\beta_1 = y(x+1) - y(x)$. In this case, the interpretation of the coefficient is relatively straightforward as it expresses the resulting change in the measurement scale of the dependent variable for a unit change in the independent variable. For example, if in a regression of weight on height of male adolescents the slope is 5, then we would conclude that an increase of 1 inch in height is associated with an increase of 5 pounds in weight.

In the logistic regression model, the slope coefficient represents the change in the logit corresponding to a change of one unit in the independent variable (i.e., $\beta_1 = g(x+1) - g(x)$). Proper interpretation of the coefficient in a logistic regression model depends on being able to place meaning on the difference between two logits. Interpretation of this difference is discussed in detail on a case-by-case basis as it relates directly to the definition and meaning of a one-unit change in the independent variable. In the following sections of this chapter we consider the interpretation of the coefficients for a univariate logistic regression model for each of the possible measurement scales of the independent variable. In addition we discuss interpretation of the coefficients in multivariable models.

3.2 DICHOTOMOUS INDEPENDENT VARIABLE

We begin our consideration of the interpretation of logistic regression coefficients with the situation where the independent variable is nominal scale and dichotomous (i.e., measured at two levels). This case provides the conceptual foundation for all the other situations.

We assume that the independent variable, x, is coded as either zero or one. The difference in the logit for a subject with $x = 1$ and $x = 0$ is

$$g(1)-g(0)=[\beta_0+\beta_1]-[\beta_0]=\beta_1.$$

The algebra shown in this equation is rather straightforward. We present it in this level of detail to emphasize that the first step in interpreting the effect of a covariate in a model is to express the desired logit difference in terms of the model. In this case the logit difference is equal to β_1. In order to interpret this result we need to introduce and discuss a measure of association termed the *odds ratio*.

The possible values of the logistic probabilities may be conveniently displayed in a 2×2 table as shown in Table 3.1. 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 OR, is defined as the ratio of the odds for $x=1$ to the odds for $x=0$, and is given by the equation

$$\text{OR}=\frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]}. \tag{3.1}$$

Substituting the expressions for the logistic regression model shown in Table 3.1 into (3.1) we obtain

Table 3.1 Values of the Logistic Regression Model When the Independent Variable Is Dichotomous

Outcome Variable (Y)	Independent Variable (X) $x=1$	$x=0$
$y=1$	$\pi(1)=\dfrac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$\pi(0)=\dfrac{e^{\beta_0}}{1+e^{\beta_0}}$
$y=0$	$1-\pi(1)=\dfrac{1}{1+e^{\beta_0+\beta_1}}$	$1-\pi(0)=\dfrac{1}{1+e^{\beta_0}}$
Total	1.0	1.0

$$\begin{aligned}
\text{OR} &= \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)\Big/\left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)\Big/\left(\frac{1}{1+e^{\beta_0}}\right)} \\
&= \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} \\
&= e^{(\beta_0+\beta_1)-\beta_0} \\
&= e^{\beta_1}.
\end{aligned}$$

Hence, for logistic regression with a dichotomous independent variable coded 1 and 0, the relationship between the odds ratio and the regression coefficient is

$$\text{OR} = e^{\beta_1} . \tag{3.2}$$

This simple relationship between the coefficient and the odds ratio is the fundamental reason why logistic regression has proven to be such a powerful analytic research tool.

The odds ratio is a measure of association which has found wide use, especially in epidemiology, as it approximates how much more likely (or unlikely) it is for the outcome to be present among those with $x = 1$ than among those with $x = 0$. For example, if y denotes the presence or absence of lung cancer and if x denotes whether the person is a smoker, then $\hat{\text{OR}} = 2$ estimates that lung cancer is twice as likely to occur among smokers than among nonsmokers in the study population. As another example, suppose y denotes the presence or absence of heart disease and x denotes whether or not the person engages in regular strenuous physical exercise. If the estimated odds ratio is $\hat{\text{OR}} = 0.5$, then occurrence of heart disease is one half as likely to occur among those who exercise than among those who do not in the study population.

The interpretation given for the odds ratio is based on the fact that in many instances it approximates a quantity called the relative risk. This parameter is equal to the ratio $\pi(1)/\pi(0)$. It follows from (3.1) that the odds ratio approximates the relative risk if $[1-\pi(0)]/[1-\pi(1)] \approx 1$. This holds when $\pi(x)$ is small for both $x = 1$ and 0.

Readers who have not had experience with the odds ratio as a measure of association would be advised to spend some time reading

about this measure in one of the following texts: Breslow and Day (1980), Kelsey, Thompson, and Evans (1986), Rothman and Greenland (1998) and Schlesselman (1982).

An example may help to clarify what the odds ratio is and how it is computed from the results of a logistic regression program or from a 2×2 table. In many examples of logistic regression encountered in the literature we find that a continuous variable has been dichotomized at some biologically meaningful cutpoint. A more detailed discussion of the rationale and implications for the modeling of such a decision is presented in Chapter 4. With this in mind we use the data displayed in Table 1.1 and create a new variable, AGED, which takes on the value 1 if the age of the subject is greater than or equal to 55 and zero otherwise. The result of cross classifying the dichotomized age variable with the outcome variable CHD is presented in Table 3.2.

The data in Table 3.2 tell us that there were 21 subjects with values $(x = 1, y = 1)$, 22 with $(x = 0, y = 1)$, 6 with $(x = 1, y = 0)$, and 51 with $(x = 0, y = 0)$. Hence, for these data, the likelihood function shown in (1.3) simplifies to

$$l(\boldsymbol{\beta}) = \pi(1)^{21} \times [1-\pi(1)]^{6} \times \pi(0)^{22} \times [1-\pi(0)]^{51}.$$

Use of a logistic regression program to obtain the estimates of β_0 and β_1 yields the results shown in Table 3.3.

The estimate of the odds ratio from (3.2) is $\widehat{\text{OR}} = e^{2.094} = 8.1$. Readers who have had some previous experience with the odds ratio undoubtedly wonder why a logistic regression package was used to obtain the maximum likelihood estimate of the odds ratio, when it could have been obtained directly from the cross-product ratio from Table 3.2, namely,

Table 3.2 Cross-Classification of AGE Dichotomized at 55 Years and CHD for 100 Subjects

	AGED(x)		
CHD(y)	≥ 55 (1)	< 55 (0)	Total
Present (1)	21	22	43
Absent (0)	6	51	57
Total	27	73	100

$$\hat{\text{OR}} = \frac{21/6}{22/51} = 8.11.$$

Thus $\hat{\beta}_1 = \ln[(21/6)/(22/51)] = 2.094$. We emphasize here that logistic regression is, in fact, regression even in the simplest case possible. The fact that the data may be formulated in terms of a contingency table provides the basis for interpretation of estimated coefficients as the log of odds ratios.

Along with the point estimate of a parameter, it is a good idea to use a confidence interval estimate to provide additional information about the parameter value. In the case of the odds ratio, OR, for a 2×2 table there is an extensive literature dealing with this problem, much of which is focused on methods when the sample size is small. The reader who wishes to learn more about the available exact and approximate methods should see the papers by Fleiss (1979) and Gart and Thomas (1972). A good summary may be found in the texts by Breslow and Day (1980), Kleinbaum, Kupper, and Morgenstern (1982), and Rothman and Greenland (1998).

The odds ratio, OR, is usually the parameter of interest in a logistic regression due to its ease of interpretation. However, its estimate, $\hat{\text{OR}}$, tends to have a distribution that is skewed. The skewness of the sampling distribution of $\hat{\text{OR}}$ is due to the fact that possible values range between 0 and ∞, with the null value equaling 1. In theory, for large enough sample sizes, the distribution of $\hat{\text{OR}}$ is normal. Unfortunately, this sample size requirement typically exceeds that of most studies. Hence, inferences are usually based on the sampling distribution of $\ln(\hat{\text{OR}}) = \hat{\beta}_1$, which tends to follow a normal distribution for much smaller sample sizes. A $100\times(1-\alpha)\%$ confidence interval (CI) estimate for the odds ratio is obtained by first calculating the endpoints of a con-

Table 3.3 Results of Fitting the Logistic Regression Model to the Data in Table 3.2

Variable	Coeff.	Std. Err.	z	P>|z|
AGED	2.094	0.5285	3.96	<0.001
Constant	−0.841	0.2551	−3.30	0.001

Log likelihood = −58.9795

fidence interval for the coefficient, β_1, and then exponentiating these values. In general, the endpoints are given by the expression

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

As an example, consider the estimation of the odds ratio for the dichotomized variable AGED. The point estimate is $\hat{\mathrm{OR}} = 8.1$ and the endpoints of a 95% CI are

$$\exp(2.094 \pm 1.96 \times 0.529) = (2.9, 22.9).$$

This interval is typical of the confidence intervals seen for odds ratios when the point estimate exceeds 1. The confidence interval is skewed to the right. This confidence interval suggests that CHD among those 55 and older in the study population could be as little as 2.9 times or much as 22.9 times more likely than those under 55, at the 95 percent level of confidence.

Because of the importance of the odds ratio as a measure of association, many software packages automatically provide point and confidence interval estimates based on the exponentiation of each coefficient in a fitted logistic regression model. These quantities provide estimates of odds ratios of interest in only a few special cases (e.g., a dichotomous variable coded zero or one that is not involved in any interactions with other variables). The major goal of this chapter is to provide the methods for using the results of fitted models to provide point and confidence interval estimates of odds ratios that are of interest, regardless of how complex the fitted model may be.

Before concluding the dichotomous variable case, it is important to consider the effect that the coding of the variable has on the computation of the estimated odds ratio. In the previous discussion we noted that the estimate of the odds ratio was $\hat{\mathrm{OR}} = \exp\left(\hat{\beta}_1\right)$. This is correct when the independent variable is coded as 0 or 1. Other coding may require that we calculate the value of the logit difference for the specific coding used, and then exponentiate this difference to estimate the odds ratio.

We illustrate these computations in detail, as they demonstrate the general method for computing estimates of odds ratios in logistic regression. The estimate of the log of the odds ratio for any independent

variable at two different levels, say $x = a$ versus $x = b$, is the difference between the estimated logits computed at these two values,

$$\begin{aligned}\ln\left[\hat{\mathrm{OR}}(a,b)\right] &= \hat{g}(x=a) - \hat{g}(x=b) \\ &= \left(\hat{\beta}_0 + \hat{\beta}_1 \times a\right) - \left(\hat{\beta}_0 + \hat{\beta}_1 \times b\right) \\ &= \hat{\beta}_1 \times (a-b). \end{aligned} \tag{3.3}$$

The estimate of the odds ratio is obtained by exponentiating the logit difference,

$$\hat{\mathrm{OR}}(a,b) = \exp\left[\hat{\beta}_1 \times (a-b)\right]. \tag{3.4}$$

Note that this expression is equal to $\exp\left(\hat{\beta}_1\right)$ only when $(a - b) = 1$. In (3.3) and (3.4) the notation $\hat{\mathrm{OR}}(a,b)$ is used to represent the odds ratio

$$\hat{\mathrm{OR}}(a,b) = \frac{\hat{\pi}(x=a)/\left[1-\hat{\pi}(x=a)\right]}{\hat{\pi}(x=b)/\left[1-\hat{\pi}(x=b)\right]} \tag{3.5}$$

and when $a = 1$ and $b = 0$ we let $\hat{\mathrm{OR}} = \hat{\mathrm{OR}}(1,0)$.

Some software packages offer a choice of methods for coding design variables. The "zero-one" coding used so far in this section is frequently referred to as *reference cell* coding. The reference cell method typically assigns the value of zero to the lower code for x and one to the higher code. For example, if SEX was coded as 1 = male and 2 = female, then the resulting design variable under this method, D, would be coded 0 = male and 1 = female. Exponentiation of the estimated coefficient for D would estimate the odds ratio of female relative to male. This same result would have been obtained had sex been coded originally as 0 = male and 1 = female, and then treating the variable SEX as if it were interval scaled.

Another coding method is frequently referred to as *deviation from means* coding. This method assigns the value of −1 to the lower code, and a value of 1 to the higher code. The coding for the variable SEX discussed above is shown in Table 3.4.

Table 3.4 Illustration of the Coding of the Design Variable Using the Deviation from Means Method

SEX (Code)	Design Variable *D*
Male (1)	-1
Female (2)	1

Suppose we wish to estimate the odds ratio of female versus male when deviation from means coding is used. We do this by using the general method shown in (3.3) and (3.4),

$$\begin{aligned}\ln\left[\hat{\mathrm{OR}}(\text{female},\text{male})\right] &= \hat{g}(\text{female}) - \hat{g}(\text{male}) \\ &= g(D=1) - g(D=-1) \\ &= \left[\hat{\beta}_0 + \hat{\beta}_1 \times (D=1)\right] - \left[\hat{\beta}_0 + \hat{\beta}_1 \times (D=-1)\right] \\ &= 2\hat{\beta}_1\end{aligned}$$

and the estimated odds ratio is $\hat{\mathrm{OR}}(\text{female},\text{male}) = \exp\left(2\hat{\beta}_1\right)$. Thus, if we had exponentiated the coefficient from the computer output we would have obtained the wrong estimate of the odds ratio. This points out quite clearly that we must pay close attention to the method used to code the design variables.

The method of coding also influences the calculation of the endpoints of the confidence interval. For the above example, using the deviation from means coding, the estimated standard error needed for confidence interval estimation is $\hat{\mathrm{SE}}\left(2\hat{\beta}_1\right)$ which is $2 \times \hat{\mathrm{SE}}\left(\hat{\beta}_1\right)$. Thus the endpoints of the confidence interval are

$$\exp\left[2\hat{\beta}_1 \pm z_{1-\alpha/2} 2\hat{\mathrm{SE}}\left(\hat{\beta}_1\right)\right].$$

In general, the endpoints of the confidence interval for the odds ratio given in (3.5) are

$$\exp\left[\hat{\beta}_1(a-b) \pm z_{1-\alpha/2}|a-b| \times \hat{\mathrm{SE}}\left(\hat{\beta}_1\right)\right],$$

where $|a-b|$ is the absolute value of $(a-b)$. Since we can control how we code our dichotomous variables, we recommend that, in most situations, they be coded as 0 or 1 for analysis purposes. Each dichotomous variable is then treated as an interval scale variable.

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. This relationship between the logistic regression coefficient and the odds ratio provides the foundation for our interpretation of all logistic regression results.

3.3 POLYCHOTOMOUS INDEPENDENT VARIABLE

Suppose that instead of two categories the independent variable has $k>2$ distinct values. 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 values and the scale of measurement is nominal. We saw in Chapter 2 that it is inappropriate to model a nominal scale variable as if it were an interval scale variable. Therefore, we must form a set of design variables to represent the categories of the variable. In this section we present methods for creating design variables for polychotomous independent variables. The choice of a particular method depends to some extent on the goals of the analysis and the stage of model development.

We begin by extending the method presented in Table 2.1 for a dichotomous variable. For example, suppose that in a study of CHD the variable RACE is coded at four levels, and that the cross-classification of

Table 3.5 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	1	8	6	4	
95 % CI		(2.3, 27.6)	(1.7, 21.3)	(1.1, 14.9)	
$\ln(\hat{OR})$	0.0	2.08	1.79	1.39	