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	
39	40	4	1	89	59	7	
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	
49	44	4	0	99	65		
50	44	4	0	100	69	8	

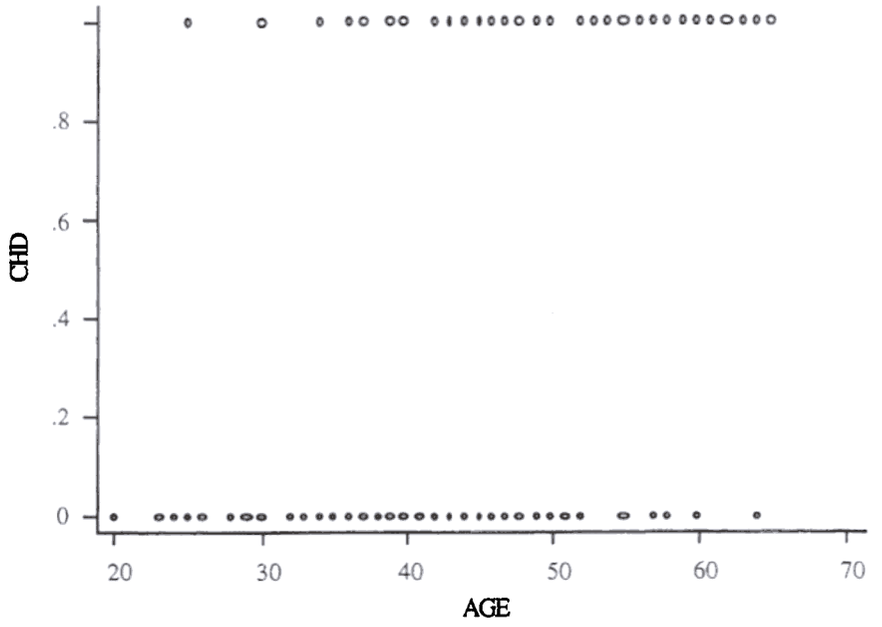

Figure 1.1 Scatterplot of CHD by AGE for 100 subjects.

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

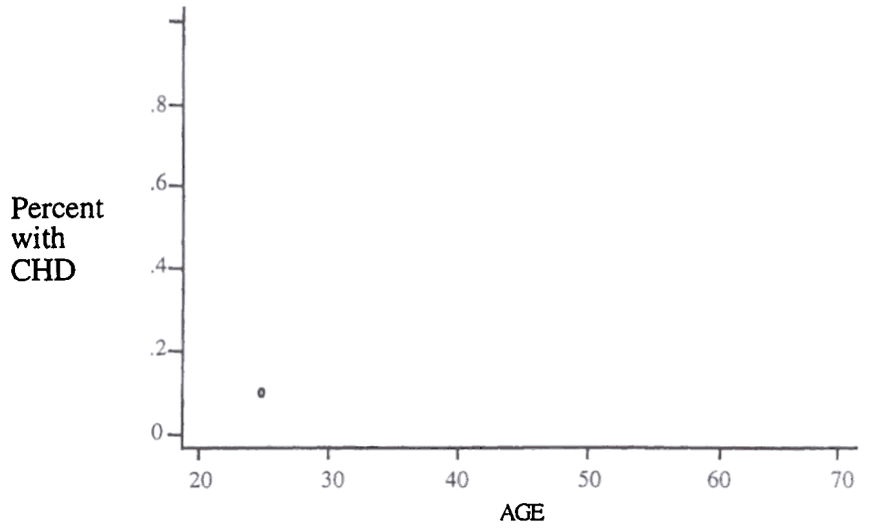

Figure 1.2 Plot of the percentage of subjects with CHD in each age group.

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-

ties. Unfortunately, when the method of least squares is applied to a model with a dichotomous outcome the estimators 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 called *maximum likelihood.* This method will provide the foundation for our approach to estimation with the logistic regression model. In a very general sense the method of maximum likelihood yields values for the unknown parameters which maximize the probability of obtaining the observed set of data. In order to apply this method we must first construct a function, called the *likelihood function.* 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 which agree most closely with the observed data. We now describe how to find these values from the logistic regression model.

If Y is coded as 0 or 1 then the expression for $\pi(x)$ given in equation (1.1) provides (for an arbitrary value of $\boldsymbol{\beta} = (\beta_0, \beta_1)$, the vector of parameters) the conditional probability that Y is equal to 1 given x. This will be denoted as $P(Y = 1 \mid x)$. It follows that the quantity $1 - \pi(x)$ gives the conditional probability that Y is equal to zero given x, $P(Y = 0 \mid x)$. 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 expression

$$\pi(x_i)^{y_i}\left[1 - \pi(x_i)\right]^{1-y_i}$$

Since the observations are assumed to be independent, the likelihood function is obtained as the product of the terms given in expression (1.2) as follows:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}\left[1 - \pi(x_i)\right]^{1-y_i}$$

The principle of maximum likelihood states that we use as our estimate of $\boldsymbol{\beta}$ the value which maximizes the expression in equation (1.3). However, it is easier mathematically to work with the log of equation (1.3). This expression, the *log likelihood*, is defined as

$$L(\boldsymbol{\beta}) = \ln[l(\boldsymbol{\beta})] = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1-y_i)\ln[1-\pi(x_i)]\} \tag{1.4}$$

To find the value of $\boldsymbol{\beta}$ that maximizes $L(\boldsymbol{\beta})$ we differentiate $L(\boldsymbol{\beta})$ with respect to β_0 and β_1 and set the resulting expressions equal to zero. These equations, known as the *likelihood equations*, are:

$$\sum [y_i - \pi(x_i)] = 0 \tag{1.5}$$

and

$$\sum x_i [y_i - \pi(x_i)] = 0. \tag{1.6}$$

In equations (1.5) and (1.6) it is understood that the summation is over i varying from 1 to n. (The practice of suppressing the index and range of summation, when these are clear, is followed throughout the text.)

In linear regression, the likelihood equations, obtained by differentiating the sum of squared deviations function with respect to $\boldsymbol{\beta}$ are linear in the unknown parameters and thus are easily solved. For logistic regression the expressions in equations (1.5) and (1.6) are nonlinear in β_0 and β_1, and thus require special methods for their solution. These methods are iterative in nature and have been programmed into available logistic regression software. For the moment we need not be concerned about these iterative methods and will view them as a computational detail taken care of for us. The interested reader may see the text by McCullagh and Nelder (1989) for a general discussion of the methods used by most programs. In particular, they show that the solution to equations (1.5) and (1.6) may be obtained using an iterative weighted least squares procedure.

The value of $\boldsymbol{\beta}$ given by the solution to equations (1.5) and (1.6) is called the maximum likelihood estimate and will be denoted as $\hat{\boldsymbol{\beta}}$. In general, the use of the symbol "^" denotes the maximum likelihood estimate of the respective quantity. For example, $\hat{\pi}(x_i)$ 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.

Table 1.3 Results of Fitting the Logistic Regression Model to the Data in Table 1.1

Variable	Coeff.	Std. Err.	z	P>|z|

Log likelihood = −53.67656

As such, it represents the fitted or predicted value for the logistic regression model. An interesting consequence of equation (1.5) 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. This property will be especially useful in later chapters when we discuss assessing the fit of the model.

As an example, consider the data given in Table 1.1. Use of a logistic regression software package, with continuous variable AGE as the independent variable, produces the output in Table 1.3. The maximum likelihood estimates of β_0 and β_1 are thus seen to be $\hat{\beta}_0 = -5.309$ and $\hat{\beta}_1 = 0.111$. The fitted values are given by the equation

$$\hat{\pi}(x) = \frac{e^{-5.309+0.111\times\text{AGE}}}{1+e^{-5.309+0.111\times\text{AGE}}}$$

and the estimated logit, $\hat{g}(x)$, is given by the equation

$$\hat{g}(x) = -5.309 + 0. \quad \times \text{AGE}$$

The log likelihood given in Table 1.3 is the value of equation (1.4) computed using $\hat{\beta}_0$ and $\hat{\beta}_1$.

Three additional columns are present in Table 1.3. One contains estimates of the standard errors of the estimated coefficients, the next column displays the ratios of the estimated coefficients to their estimated standard errors and the last column displays a p-value. These quantities are discussed in the next section.

Following the fitting of the model we begin to evaluate its adequacy.