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. Quite often the outcome variable is discrete, taking on two or more possible values. The logistic regression model is the most frequently used regression model for the analysis of these data.

Before beginning a thorough study of the logistic regression model it is important to understand that the goal of an analysis using this model is the same as that of any other regression model used in statistics, that is, to find the best fitting and most parsimonious, clinically interpretable model to describe the relationship between an outcome (dependent or response) variable and a set of independent (predictor or explanatory) variables. The 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 form of the model and its assumptions. Once this difference is accounted for, the methods employed in an analysis using logistic regression follow, more or less, the same general principles used in linear regression. Thus, the techniques used in linear regression analysis motivate our approach to logistic regression. We illustrate both the similarities and differences between logistic regression and linear regression with an example.

Applied Logistic Regression, Third Edition.
David W. Hosmer, Jr., Stanley Lemeshow, and Rodney X. Sturdivant.

Example 1: Table 1.1 lists the age in years (AGE), and presence or absence of evidence of significant coronary heart disease (CHD) for 100 subjects in a hypothetical study of risk factors for heart disease. The table also contains an identifier variable (ID) and an age group variable (AGEGRP). The outcome variable is CHD, which is coded with a value of "0" to indicate that CHD is absent, or "1" to indicate that it is present in the individual. In general, any two values could be used, but we have found it most convenient to use zero and one. We refer to this data set as the CHDAGE data.

It is of interest to explore the relationship between AGE and the presence or absence of CHD in this group. 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 ($y = 0$) or 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.

The main problem with Figure 1.1 is that the variability in CHD at all ages is large. This makes it difficult to see any 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. We use this strategy by grouping age into the categories (AGEGRP) defined in Table 1.1. Table 1.2 contains, for each age group, the frequency of occurrence of each outcome, as well as the percent with CHD present.

By examining this table, a clearer picture of the relationship begins to emerge. It shows that as age increases, the proportion (mean) of individuals with evidence of CHD increases. Figure 1.2 presents a plot of the percent of individuals with CHD versus the midpoint of each age interval. This plot provides considerable insight into the relationship between CHD and AGE in this study, but the functional form for this relationship needs to be described. The plot in this figure is similar to what one might obtain if this same process of grouping and averaging were performed in a linear regression. We 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 is expressed as "$E(Y|x)$" where Y denotes the outcome variable and x denotes a specific value of the independent variable. The quantity $E(Y|x)$ 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

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

ID	AGE	AGEGRP	CHD
1	20	1	0
2	23	1	0
3	24	1	0
4	25	1	0
5	25	1	1
6	26	1	0
7	26	1	0
8	28	1	0
9	28	1	0
10	29	1	0
11	30	2	0
12	30	2	0
13	30	2	0
14	30	2	0
15	30	2	0
16	30	2	1
17	32	2	0
18	32	2	0
19	33	2	0
20	33	2	0
21	34	2	0
22	34	2	0
23	34	2	1
24	34	2	0
25	34	2	0
26	35	3	0
27	35	3	0
28	36	3	0
29	36	3	1
30	36	3	0
31	37	3	0
32	37	3	1
33	37	3	0
34	38	3	0
35	38	3	0
36	39	3	0
37	39	3	1
38	40	4	0
39	40	4	1
40	41	4	0
41	41	4	0
42	42	4	0
43	42	4	0
44	42	4	0

(*continued*)

Table 1.1 (*Continued*)

ID	AGE	AGEGRP	CHD
45	42	4	1
46	43	4	0
47	43	4	0
48	43	4	1
49	44	4	0
50	44	4	0
51	44	4	1
52	44	4	1
53	45	5	0
54	45	5	1
55	46	5	0
56	46	5	1
57	47	5	0
58	47	5	0
59	47	5	1
60	48	5	0
61	48	5	1
62	48	5	1
63	49	5	0
64	49	5	0
65	49	5	1
66	50	6	0
67	50	6	1
68	51	6	0
69	52	6	0
70	52	6	1
71	53	6	1
72	53	6	1
73	54	6	1
74	55	7	0
75	55	7	1
76	55	7	1
77	56	7	1
78	56	7	1
79	56	7	1
80	57	7	0
81	57	7	0
82	57	7	1
83	57	7	1
84	57	7	1
85	57	7	1
86	58	7	0
87	58	7	1
88	58	7	1
89	59	7	1
90	59	7	1

Table 1.1 (*Continued*)

ID	AGE	AGEGRP	CHD
91	60	8	0
92	60	8	1
93	61	8	1
94	62	8	1
95	62	8	1
96	63	8	1
97	64	8	0
98	64	8	1
99	65	8	1
100	69	8	1

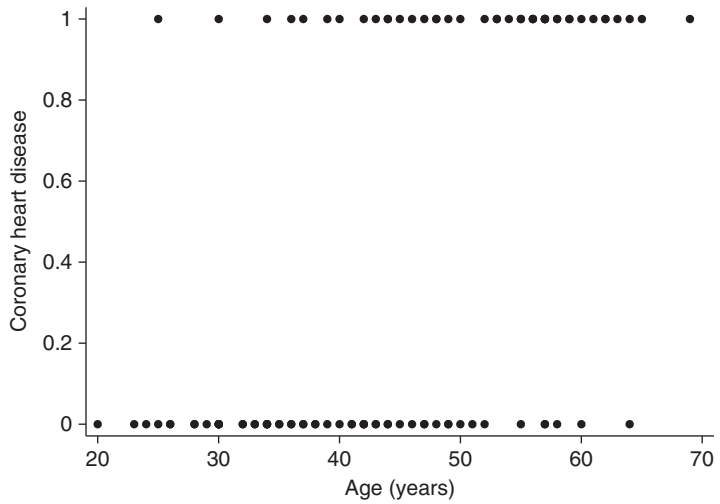

Figure 1.1 Scatterplot of presence or absence of coronary heart disease (CHD) by AGE for 100 subjects.

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 $+\infty$.

The column labeled "Mean" in Table 1.2 provides an estimate of $E(Y|x)$. We assume, for purposes of exposition, that the estimated values plotted in Figure 1.2 are close enough to the true values of $E(Y|x)$ to provide a reasonable assessment of the functional relationship between CHD and AGE. With a dichotomous outcome variable, the conditional mean must be greater than or equal to zero and less than

Table 1.2 Frequency Table of Age Group by CHD

		Coronary Heart Disease		
Age Group	*n*	Absent	Present	Mean
20–29	10	9	1	0.100
30–34	15	13	2	0.133
35–39	12	9	3	0.250
40–44	15	10	5	0.333
45–49	13	7	6	0.462
50–54	8	3	5	0.625
55–59	17	4	13	0.765
60–69	10	2	8	0.800
Total	100	57	43	0.430

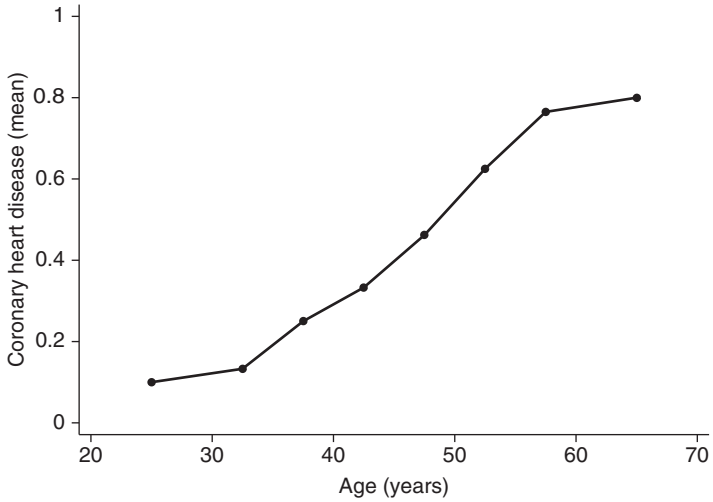

Figure 1.2 Plot of the percentage of subjects with CHD in each AGE group.

or equal to one (i.e., $0 \leq E(Y|x) \leq 1$). This can be seen in Figure 1.2. In addition, the plot shows that this mean approaches zero and one "gradually". The change in the $E(Y|x)$ per unit change in x becomes progressively smaller as the conditional mean gets closer to zero or one. The curve is said to be *S-shaped* and resembles a plot of the cumulative distribution of a continuous random variable. Thus, it should not seem surprising that some well-known cumulative distributions have been used to provide a model for $E(Y|x)$ in the case when Y is dichotomous. The model we use is based on 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. Second, its model parameters provide the basis for clinically meaningful estimates of effect. A detailed discussion of the interpretation of the model parameters is given in Chapter 3.

In order to simplify notation, we use the quantity $\pi(x) = E(Y|x)$ 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 $y = E(Y|x) + \varepsilon$. 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 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 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 shown that in a regression analysis when the outcome variable is dichotomous:

1. The model for the conditional mean of the regression equation must be bounded between zero and one. 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 is the statistical distribution on which the analysis is based.

3. The principles that guide an analysis using linear regression 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 ith 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. Fitting 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 that minimize the sum-of-squared deviations of the observed values of Y from the predicted values based on the model. Under the usual assumptions for linear regression the method of least squares yields estimators with a number of desirable statistical properties. 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 provides the foundation for our approach to estimation with the logistic regression model throughout this text. In a general sense, the method of maximum likelihood yields values for the unknown parameters that 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 the parameters are the values that maximize this function. Thus, the resulting estimators are those that agree most closely with the observed data. We now describe how to find these values for 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 is denoted as $\pi(x)$. 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 expression

$$\pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}. \tag{1.2}$$

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

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}. \tag{1.3}$$

The principle of maximum likelihood states that we use as our estimate of $\boldsymbol{\beta}$ the value that 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 this 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 logistic regression software. For the moment, we need not be concerned about these iterative methods and view them as a computational detail that is taken care of for us. The interested reader may consult 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 is denoted as $\hat{\boldsymbol{\beta}}$. In general, the use of the symbol "$\hat{}$" 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. As such, it represents the fitted or predicted value for the logistic regression model. An interesting consequence of equation (1.5) is that

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

Table 1.3 Results of Fitting the Logistic Regression Model to the CHDAGE Data, $n = 100$

Variable	Coeff.	Std. Err.	z	p
Age	0.111	0.0241	4.61	<0.001
Constant	−5.309	1.1337	−4.68	<0.001

Log-likelihood = −53.676546.

That is, the sum of the observed values of y is equal to the sum of the predicted (expected) values. We use this property 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 $\hat{\beta}_0 = -5.309$ and $\hat{\beta}_1 = 0.111$. The fitted values are given by the equation

$$\hat{\pi}(x) = \frac{e^{-5.309+0.111\times \text{AGE}}}{1 + e^{-5.309+0.111\times \text{AGE}}} \tag{1.7}$$

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

$$\hat{g}(x) = -5.309 + 0.111 \times \text{AGE}. \tag{1.8}$$

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.

1.3 TESTING FOR THE SIGNIFICANCE OF THE COEFFICIENTS

In practice, the modeling of a set of data, as we show in Chapters 4, 7, and 8, is a much more complex process than one of simply fitting and testing. The methods we present in this section, while simplistic, do provide essential building blocks for the more complex process.

After estimating the coefficients, our first look at the fitted model commonly concerns an assessment of the significance of the variables in the model. This usually involves formulation and testing of a statistical hypothesis to determine whether the independent variables in the model are "significantly" related to the outcome variable. The method for performing this test is quite general, and differs from one type of model to the next only in the specific details. We begin by

discussing the general approach for a single independent variable. The multivariable case is considered in Chapter 2.

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 a model that does not include that variable?* This question is answered by comparing the observed values of the response variable to 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 is called *goodness of fit*). Instead, our question is posed in a relative sense. The assessment of goodness of fit is a more complex question that is discussed in detail in Chapter 5.

The general method for assessing significance of variables is easily illustrated in the linear regression model, and its use there motivates the approach used for logistic regression. A comparison of the two approaches highlights the differences between modeling continuous and dichotomous response variables.

In linear regression, one assesses the significance of the slope coefficient by forming what is referred to as an *analysis of variance table*. This table partitions the total sum-of-squared deviations of observations about their mean into two parts: (1) the sum-of-squared deviations of observations about the regression line SSE (or *residual sum-of-squares*) and (2) the sum-of-squares of predicted values, based on the regression model, about the mean of the dependent variable SSR (or *due regression sum-of-squares*). This is just a convenient way of displaying the comparison of observed to predicted values under two models. In linear regression, the comparison of observed and predicted values is based on the square of the distance between the two. If y_i denotes the observed value and $\hat{y}_i$ denotes the predicted value for the ith individual under the model, then the statistic used to evaluate this comparison is

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Under the model not containing the independent variable in question the only parameter is β_0, and $\hat{\beta}_0 = \overline{y}$, the mean of the response variable. In this case, $\hat{y}_i = \overline{y}$ and SSE is equal to the total sum-of-squares. When we include the independent variable in the model, any decrease in SSE is due to the fact that the slope coefficient for the independent variable is not zero. The change in the value of SSE is due to the regression source of variability, denoted SSR. That is,

$$\text{SSR} = \left[\sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2\right] - \left[\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2\right].$$

In linear regression, interest focuses on the size of SSR. A large value suggests that the independent variable is important, whereas a small value suggests that the independent variable is not helpful in predicting the response.

The guiding principle with logistic regression is the same: compare observed values of the response variable to predicted values obtained from models, with and without the variable in question. In logistic regression, comparison of observed to predicted values is based on the log-likelihood function defined in equation (1.4). To better understand this comparison, it is helpful conceptually to think of an observed value of the response variable as also being a predicted value resulting from a saturated model. A saturated model is one that contains as many parameters as there are data points. (A simple example of a saturated model is fitting a linear regression model when there are only two data points, $n = 2$.)

The comparison of observed to predicted values using the likelihood function is based on the following expression:

$$D = -2\ln\left[\frac{\text{(likelihood of the fitted model)}}{\text{(likelihood of the saturated model)}}\right]. \tag{1.9}$$

The quantity inside the large brackets in the expression above is called the *likelihood ratio*. Using minus twice its log is necessary to obtain a quantity whose distribution is known and can therefore be used for hypothesis testing purposes. Such a test is called the *likelihood ratio test*. Using equation (1.4), equation (1.9) becomes

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1 - y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right], \tag{1.10}$$

where $\hat{\pi}_i = \hat{\pi}(x_i)$.

The statistic, D, in equation (1.10) is called the *deviance*, and for logistic regression, it plays the same role that the residual sum-of-squares plays in linear regression. In fact, the deviance as shown in equation (1.10), when computed for linear regression, is identically equal to the SSE.

Furthermore, in a setting as shown in Table 1.1, where the values of the outcome variable are either 0 or 1, the likelihood of the saturated model is identically equal to 1.0. Specifically, it follows from the definition of a saturated model that $\hat{\pi}_i = y_i$ and the likelihood is

$$l(\text{saturated model}) = \prod_{i=1}^{n} y_i^{y_i} \times (1 - y_i)^{(1-y_i)} = 1.0.$$

Thus it follows from equation (1.9) that the deviance is

$$D = -2\ln(\text{likelihood of the fitted model}). \tag{1.11}$$

Some software packages report the value of the deviance in equation (1.11) rather than the log-likelihood for the fitted model. In the context of testing for the significance of a fitted model, we want to emphasize that we think of the deviance in the same way that we think of the residual sum-of-squares in linear regression.

In particular, to assess the significance of an independent variable we compare the value of D with and without the independent variable in the equation. The change in D due to the inclusion of the independent variable in the model is:

$$G = D(\text{model without the variable}) - D(\text{model with the variable}).$$

This statistic, G, plays the same role in logistic regression that the numerator of the partial F-test does in linear regression. Because the likelihood of the saturated model is always common to both values of D being differenced, G can be expressed as

$$G = -2\ln\left[\frac{(\text{likelihood without the variable})}{(\text{likelihood with the variable})}\right]. \tag{1.12}$$

For the specific case of a single independent variable, it is easy to show that when the variable is not in the model, the maximum likelihood estimate of β_0 is $\ln(n_1/n_0)$ where $n_1 = \sum y_i$ and $n_0 = \sum(1 - y_i)$ and the predicted probability for all subjects is constant, and equal to n_1/n. In this setting, the value of G is:

$$G = -2\ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\hat{\pi}_i^{y_i}(1-\hat{\pi}_i)^{(1-y_i)}}\right], \tag{1.13}$$

or

$$G = 2\left\{\sum_{i=1}^{n}\left[y_i\ln\left(\hat{\pi}_i\right) + (1-y_i)\ln(1-\hat{\pi}_i)\right] - \left[n_1\ln\left(n_1\right) + n_0\ln(n_0) - n\ln(n)\right]\right\}. \tag{1.14}$$

Under the hypothesis that β_1 is equal to zero, the statistic G follows a chi-square distribution with 1 degree of freedom. Additional mathematical assumptions are needed; however, for the above case they are rather nonrestrictive, and involve having a sufficiently large sample size, n, and enough subjects with both $y = 0$ and $y = 1$. We discuss in later chapters that, as far as sample size is concerned, the key determinant is $\min(n_0, n_1)$.

As an example, we consider the model fit to the data in Table 1.1, whose estimated coefficients and log-likelihood are given in Table 1.3. For these data the sample size is sufficiently large as $n_1 = 43$ and $n_0 = 57$. Evaluating G as shown in equation (1.14) yields

$$\begin{aligned} G &= 2\{-53.677 - [43\ln(43) + 57\ln(57) - 100\ln(100)]\} \\ &= 2[-53.677 - (-68.331)] = 29.31. \end{aligned}$$

The first term in this expression is the *log-likelihood* from the model containing age (see Table 1.3), and the remainder of the expression simply substitutes n_1 and n_0 into the second part of equation (1.14). We use the symbol $\chi^2(\nu)$ to denote a chi-square random variable with ν degrees of freedom. Using this notation, the p-value associated with this test is $P[\chi^2(1) > 29.31] < 0.001$; thus, we have convincing evidence that AGE is a significant variable in predicting CHD. This is merely a statement of the statistical evidence for this variable. Other important factors to consider before concluding that the variable is clinically important would include the appropriateness of the fitted model, as well as inclusion of other potentially important variables.

As all logistic regression software report either the value of the log-likelihood or the value of D, it is easy to check for the significance of the addition of new terms to the model or to verify a reported value of G. In the simple case of a single independent variable, we first fit a model containing only the constant term. Next, we fit a model containing the independent variable along with the constant. This gives rise to another log-likelihood. The likelihood ratio test is obtained by multiplying the difference between these two values by -2.

In the current example, the log-likelihood for the model containing only a constant term is -68.331. Fitting a model containing the independent variable (AGE) along with the constant term results in the log-likelihood shown in Table 1.3 of -53.677. Multiplying the difference in these log-likelihoods by -2 gives

$$-2 \times [-68.331 - (-53.677)] = -2 \times (-14.655) = 29.31.$$

This result, along with the associated p-value for the chi-square distribution, is commonly reported in logistic regression software packages.

There are two other statistically equivalent tests: the Wald test and the Score test. The assumptions needed for each of these is the same as those of the likelihood ratio test in equation (1.14). A more complete discussion of these three tests and their assumptions may be found in Rao (1973).

The Wald test is equal to the ratio of the maximum likelihood estimate of the slope parameter, $\hat{\beta}_1$, to an estimate of its standard error. Under the null hypothesis and the sample size assumptions, this ratio follows a standard normal distribution. While we have not yet formally discussed how the estimates of the standard errors of the estimated parameters are obtained, they are routinely printed out by computer software. For example, the Wald test for the coefficient for AGE in Table 1.3 is provided in the column headed z and is

$$W = \frac{\hat{\beta}_1}{\widehat{\mathrm{SE}}(\hat{\beta}_1)} = \frac{0.111}{0.024} = 4.61.$$

The two-tailed p-value, provided in the last column of Table 1.3, is $P(|z| > 4.61) < 0.001$, where z denotes a random variable following the standard normal distribution. Some software packages display the statistic $W^2 = z^2$, which is distributed as chi-square with 1 degree of freedom. Hauck and Donner (1977) examined the performance of the Wald test and found that it behaved in an aberrant manner, often failing to reject the null hypothesis when the coefficient was significant using the

likelihood ratio test. Thus, they recommended (and we agree) that the likelihood ratio test is preferred. We note that while the assertions of Hauk and Donner are true, we have never seen huge differences in the values of G and W^2. In practice, the more troubling situation is when the values are close, and one test has $p < 0.05$ and the other has $p > 0.05$. When this occurs, we use the p-value from the likelihood ratio test.

A test for the significance of a variable that does not require computing the estimate of the coefficient is the score test. Proponents of the score test cite this reduced computational effort as its major advantage. Use of the test is limited by the fact that it is not available in many software packages. The score test is based on the distribution theory of the derivatives of the log-likelihood. In general, this is a multivariate test requiring matrix calculations that are discussed in Chapter 2.

In the univariate case, this test is based on the conditional distribution of the derivative in equation (1.6), given the derivative in equation (1.5). In this case, we can write down an expression for the Score test. The test uses the value of equation (1.6) computed using $\beta_0 = \ln(n_1/n_0)$ and $\beta_1 = 0$. As noted earlier, under these parameter values, $\hat{\pi} = n_1/n = \overline{y}$ and the left-hand side of equation (1.6) becomes $\sum x_i(y_i - \overline{y})$. It may be shown that the estimated variance is $\overline{y}(1-\overline{y})\sum (x_i - \overline{x})^2$. The test statistic for the score test (ST) is

$$\mathrm{ST} = \frac{\sum_{i=1}^{n} x_i(y_i - \overline{y})}{\sqrt{\overline{y}(1-\overline{y})\sum_{i=1}^{n}(x_i - \overline{x})^2}}.$$

As an example of the score test, consider the model fit to the data in Table 1.1. The value of the test statistic for this example is

$$\mathrm{ST} = \frac{296.66}{\sqrt{3333.742}} = 5.14$$

and the two tailed p-value is $P(|z| > 5.14) < 0.001$. We note that, for this example, the values of the three test statistics are nearly the same (*note*: $\sqrt{G} = 5.41$).

In summary, the method for testing the significance of the coefficient of a variable in logistic regression is similar to the approach used in linear regression; however, it is based on the likelihood function for a dichotomous outcome variable under the logistic regression model.

1.4 CONFIDENCE INTERVAL ESTIMATION

An important adjunct to testing for significance of the model, discussed in Section 1.3, is calculation and interpretation of confidence intervals for parameters of interest. As is the case in linear regression we can obtain these for the slope, intercept and the "line" (i.e., the logit). In some settings it may be of interest to provide interval estimates for the fitted values (i.e., the predicted probabilities).

The basis for construction of the interval estimators is the same statistical theory we used to formulate the tests for significance of the model. In particular, the confidence interval estimators for the slope and intercept are, most often, based on their respective Wald tests and are sometimes referred to as *Wald-based confidence intervals*. The endpoints of a $100(1-\alpha)\%$ confidence interval for the slope coefficient are

$$\hat{\beta}_1 \pm z_{1-\alpha/2}\widehat{\mathrm{SE}}(\hat{\beta}_1) \tag{1.15}$$

and for the intercept they are

$$\hat{\beta}_0 \pm z_{1-\alpha/2}\widehat{\mathrm{SE}}(\hat{\beta}_0) \tag{1.16}$$

where $z_{1-\alpha/2}$ is the upper $100(1-\alpha/2)\%$ point from the standard normal distribution and $\widehat{\mathrm{SE}}(\cdot)$ denotes a model-based estimator of the standard error of the respective parameter estimator. We defer discussion of the actual formula used for calculating the estimators of the standard errors to Chapter 2. For the moment, we use the fact that estimated values are provided in the output following the fit of a model and, in addition, many packages also provide the endpoints of the interval estimates.

As an example, consider the model fit to the data in Table 1.1 regressing AGE on the presence or absence of CHD. The results are presented in Table 1.3. The endpoints of a 95 percent confidence interval for the slope coefficient from equation (1.15) are $0.111 \pm 1.96 \times 0.0241$, yielding the interval (0.064, 0.158). We defer a detailed discussion of the interpretation of these results to Chapter 3. Briefly, the results suggest that the change in the log-odds of CHD per one year increase in age is 0.111 and the change could be as little as 0.064 or as much as 0.158 with 95 percent confidence.

As is the case with any regression model, the constant term provides an estimate of the response at $x = 0$ unless the independent variable has been centered at some clinically meaningful value. In our example, the constant provides an estimate of the log-odds ratio of CHD at zero years of age. As a result, the constant term, by itself, has no useful clinical interpretation. In any event, from equation (1.16), the endpoints of a 95 percent confidence interval for the constant are $-5.309 \pm 1.96 \times 1.1337$, yielding the interval $(-7.531, -3.087)$.

The logit is the linear part of the logistic regression model and, as such, is most similar to the fitted line in a linear regression model. The estimator of the logit is

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x. \tag{1.17}$$

The estimator of the variance of the estimator of the logit requires obtaining the variance of a sum. In this case it is

$$\widehat{\mathrm{Var}}[\hat{g}(x)] = \widehat{\mathrm{Var}}(\hat{\beta}_0) + x^2\widehat{\mathrm{Var}}(\hat{\beta}_1) + 2x\widehat{\mathrm{Cov}}(\hat{\beta}_0, \hat{\beta}_1). \tag{1.18}$$

In general, the variance of a sum is equal to the sum of the variance of each term and twice the covariance of each possible pair of terms formed from the

Table 1.4 Estimated Covariance Matrix of the Estimated Coefficients in Table 1.3

	Age	Constant
Age	0.000579	
Constant	−0.026677	1.28517

components of the sum. The endpoints of a $100(1-\alpha)\%$ *Wald-based confidence interval* for the logit are

$$\hat{g}(x) \pm z_{1-\alpha/2}\widehat{\mathrm{SE}}[\hat{g}(x)], \tag{1.19}$$

where $\widehat{\mathrm{SE}}[\hat{g}(x)]$ is the positive square root of the variance estimator in equation (1.18).

The estimated logit for the fitted model in Table 1.3 is shown in equation (1.8). In order to evaluate equation (1.18) for a specific age we need the estimated covariance matrix. This matrix can be obtained from the output from all logistic regression software packages. How it is displayed varies from package to package, but the triangular form shown in Table 1.4 is a common one.

The estimated logit from equation (1.8) for a subject of age 50 is

$$\hat{g}(50) = -5.31 + 0.111 \times 50 = 0.240,$$

the estimated variance, using equation (1.18) and the results in Table 1.4, is

$$\widehat{\mathrm{Var}}[\hat{g}(50)] = 1.28517 + (50)^2 \times 0.000579 + 2 \times 50 \times (-0.026677) = 0.0650$$

and the estimated standard error is $\widehat{\mathrm{SE}}[\hat{g}(50)] = 0.2549$. Thus the end points of a 95 percent confidence interval for the logit at age 50 are

$$0.240 \pm 1.96 \times 0.2550 = (-0.260, 0.740).$$

We discuss the interpretation and use of the estimated logit in providing estimates of odds ratios in Chapter 3.

The estimator of the logit and its confidence interval provide the basis for the estimator of the fitted value, in this case the logistic probability, and its associated confidence interval. In particular, using equation (1.7) at age 50 the estimated logistic probability is

$$\hat{\pi}(50) = \frac{e^{\hat{g}(50)}}{1 + e^{\hat{g}(50)}} = \frac{e^{-5.31+0.111\times 50}}{1+e^{-5.31+0.111\times 50}} = 0.560 \tag{1.20}$$

and the endpoints of a 95 percent confidence interval are obtained from the respective endpoints of the confidence interval for the logit. The endpoints of the $100(1-\alpha)\%$ *Wald-based confidence interval* for the fitted value are

$$\frac{e^{\hat{g}(x)\pm z_{1-\alpha/2}\widehat{\mathrm{SE}}[\hat{g}(x)]}}{1 + e^{\hat{g}(x)\pm z_{1-\alpha/2}\widehat{\mathrm{SE}}[\hat{g}(x)]}}. \tag{1.21}$$

Using the example at age 50 to demonstrate the calculations, the lower limit is

$$\frac{e^{-0.260}}{1+e^{-0.260}} = 0.435,$$

and the upper limit is

$$\frac{e^{0.740}}{1+e^{0.740}} = 0.677.$$

We have found that a major mistake often made by data analysts new to logistic regression modeling is to try and apply estimates on the probability scale to individual subjects. The fitted value computed in equation (1.20) is analogous to a particular point on the line obtained from a linear regression. In linear regression each point on the fitted line provides an estimate of the mean of the dependent variable in a population of subjects with covariate value "x". Thus the value of 0.56 in equation (1.20) is an estimate of the mean (i.e., proportion) of 50-year-old subjects in the population sampled that have evidence of CHD. An individual 50-year-old subject either does or does not have evidence of CHD. The confidence interval suggests that this mean could be between 0.435 and 0.677 with 95 percent confidence. We discuss the use and interpretation of fitted values in greater detail in Chapter 3.

One application of fitted logistic regression models that has received a lot of attention in the subject matter literature is using model-based fitted values similar to the one in equation (1.20) to predict the value of a binary dependent value in individual subjects. This process is called *classification* and has a long history in statistics where it is referred to as *discriminant analysis*. We discuss the classification problem in detail in Chapter 4. We also discuss discriminant analysis within the context of a method for obtaining estimators of the coefficients in the next section.

The *coverage*[*†] of the Wald-based confidence interval estimators in equations (1.15) and (1.16) depends on the assumption that the distribution of the maximum likelihood estimators is normal. Potential sensitivity to this assumption is the main reason that the likelihood ratio test is recommended over the Wald test for assessing the significance of individual coefficients, as well as for the overall model. In settings where the number of events ($y = 1$) and/or the sample size is small the normality assumption is suspect and a log-likelihood function-based confidence interval can have better coverage. Until recently routines to compute these intervals were not available in most software packages. Cox and Snell (1989, p. 179–183) discuss the theory behind likelihood intervals, and Venzon and Moolgavkar (1988) describe an efficient way to calculate the end points.

*The remainder of this section is more advanced material that can be skipped on first reading of the text.

†The term coverage of an interval estimator refers to the percent of time confidence intervals computed in a similar manner contain the true parameter value. Research has shown that when the normality assumption does not hold, Wald-based confidence intervals can be too narrow and thus contain the true parameter with a smaller percentage than the stated confidence coefficient.

Royston (2007) describes a STATA [StataCorp (2011)] routine that implements the Venzon and Moolgavkar method that we use for the examples in this text. The SAS package's logistic regression procedure [SAS Institute Inc. (2009)] has the option to obtain likelihood confidence intervals.

The *likelihood-based confidence interval* estimator for a coefficient can be concisely described as the interval of values, β^*, for which the likelihood ratio test would fail to reject the hypothesis, $H_o : \beta = \beta^*$, at the stated $1 - \alpha$ percent significance level. The two end points, β_{lower} and β_{upper}, of this interval for a coefficient are defined as follows:

$$2[l(\hat{\beta}) - l_p(\beta_{\text{upper}})] = 2[l(\hat{\beta}) - l_p(\beta_{\text{lower}})] = \chi^2_{1-\alpha}(1), \tag{1.22}$$

where $l(\hat{\beta})$ is the value of the log-likelihood of the fitted model and $l_p(\beta)$ is the value of the *profile log-likelihood*. A value of the profile log-likelihood is computed by first specifying/fixing a value for the coefficient of interest, for example the slope coefficient for age, and then finding the value of the intercept coefficient, using the Venzon and Moolgavkar method, that maximizes the log-likelihood. This process is repeated over a grid of values of the specified coefficient, for example, values of β^*, until the solutions to equation (1.22) are found. The results can be presented graphically or in standard interval form. We illustrate both in the example below.

As an example, we show in Figure 1.3 a plot of the profile log-likelihood for the coefficient for AGE using the CHDAGE data in Table 1.1. The end points of the 95 percent likelihood interval are $\beta_{\text{lower}} = 0.067$ and $\beta_{\text{upper}} = 0.162$ and are shown in the figure where the two vertical lines intersect the "x" axis. The horizontal line in the figure is drawn at the value

$$-55.5964 = -53.6756 - \left(\frac{3.8416}{2}\right),$$

where -53.6756 is the value of the log-likelihood of the fitted model from Table 1.3 and 3.8416 is the 95th percentile of the chi-square distribution with 1 degree of freedom.

The quantity "Asymmetry" in Figure 1.3 is a measure of asymmetry of the profile log-likelihood that is the difference between the lengths of the upper part of the interval, $\beta_{\text{upper}} - \hat{\beta}$, to the lower part, $\hat{\beta} - \beta_{\text{lower}}$, as a percent of the total length, $\beta_{\text{upper}} - \beta_{\text{lower}}$. In the example the value is

$$A = 100 \times \frac{(0.162 - 0.111) - (0.111 - 0.067)}{(0.162 - 0.067)} \cong 7.5\%.$$

As the upper and lower endpoints of the Wald-based confidence interval in equation (1.15) are equidistant from the maximum likelihood estimator, it has asymmetry $A = 0$.

In this example, the Wald-based confidence interval for the coefficient for age is $(0.064, 0.158)$. The likelihood interval is $(0.067, 0.162)$, which is only 1.1% wider than the Wald-based interval. So there is not a great deal of pure numeric difference in the two intervals and the asymmetry is small. In settings where there

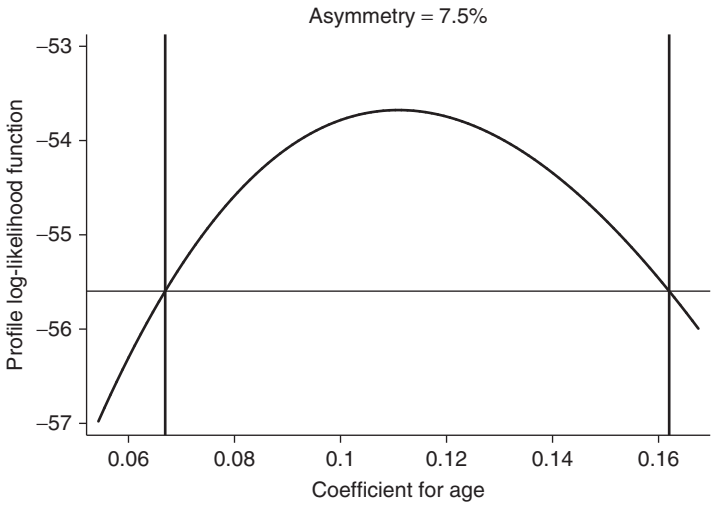

Figure 1.3 Plot of the profile log-likelihood for the coefficient for AGE in the CHDAGE data.

is greater asymmetry in the likelihood-based interval there can be more substantial differences between the two intervals. We return to this point in Chapter 3 where we discuss the interpretation of estimated coefficients. In addition, we include an exercise at the end of this chapter where there is a pronounced difference between the Wald and likelihood confidence interval estimators.

Methods to extend the likelihood intervals to functions of more than one coefficient such as the estimated logit function and probability are not available in current software packages.

1.5 OTHER ESTIMATION METHODS

The method of maximum likelihood described in Section 1.2 is the estimation method used in the logistic regression routines of the major software packages. However, two other methods have been and may still be used for estimating the coefficients. These methods are: (1) noniterative weighted least squares, and (2) discriminant function analysis.

A linear models approach to the analysis of categorical data proposed by Grizzle et al. (1969) [Grizzle, Starmer, and Koch (GSK) method] uses estimators based on noniterative weighted least squares. They demonstrate that the logistic regression model is an example of a general class of models that can be handled by their methods. We should add that the maximum likelihood estimators are usually calculated using an iterative reweighted least squares algorithm, and are also technically "least squares" estimators. The GSK method requires one iteration and is used in SAS's GENMOD procedure to fit a logistic regression model containing only categorical covariates.

A major limitation of the GSK method is that we must have an estimate of $\pi(x)$ that is not zero or 1 for most values of x. An example where we could use both maximum likelihood and GSK's noniterative weighted least squares is the data in Table 1.2. In cases such as this, the two methods are *asymptotically equivalent*, meaning that as n gets large, the distributional properties of the two estimators become identical. The GSK method could not be used with the data in Table 1.1.

The discriminant function approach to estimation of the coefficients is of historical importance as it was popularized by Cornfield (1962) in some of the earliest work on logistic regression. These estimators take their name from the fact that the posterior probability in the usual discriminant function model is the logistic regression function given in equation (1.1). More precisely, if the independent variable, X, follows a normal distribution within each of two groups (subpopulations) defined by the two values of Y and has different means and the same variance, then the conditional distribution of Y given $X = x$ is the logistic regression model. That is, if

$$X|Y \sim \mathrm{N}(\mu_j, \sigma^2),\ j = 0, 1$$

then $P(Y = 1|x) = \pi(x)$. The symbol "$\sim$" is read "is distributed" and the "$N(\mu, \sigma^2)$" denotes the normal distribution with mean equal to μ and variance equal to σ^2. Under these assumptions it is easy to show [Lachenbruch (1975)] that the logistic coefficients are

$$\beta_0 = \ln\left(\frac{\theta_1}{\theta_0}\right) - 0.5(\mu_1^2 - \mu_0^2)/\sigma^2 \tag{1.23}$$

and

$$\beta_1 = (\mu_1 - \mu_0)/\sigma^2, \tag{1.24}$$

where $\theta_j = P(Y = j),\ j = 0,\ 1$. The discriminant function estimators of β_0 and β_1 are obtained by substituting estimators for $\mu_j, \theta_j,\ j = 0,\ 1$ and σ^2 into the above equations. The estimators usually used are $\hat{\mu}_j = \overline{x}_j$, the mean of x in the subgroup defined by $y = j,\ j = 0, 1, \theta_1 = n_1/n$ the mean of y with $\hat{\theta}_0 = 1 - \hat{\theta}_1$ and

$$\hat{\sigma}^2 = [(n_0 - 1)s_0^2 + (n_1 - 1)s_1^2]/(n_0 + n_1 - 2),$$

where s_j^2 is the unbiased estimator of σ^2 computed within the subgroup of the data defined by $y = j,\ j = 0, 1$. The above expressions are for a single variable x and multivariable expressions are presented in Chapter 2.

It is natural to ask why, if the discriminant function estimators are so easy to compute, they are not used in place of the maximum likelihood estimators? Halpern et al. (1971) and Hosmer et al. (1983) compared the two methods when the model contains a mixture of continuous and discrete variables, with the general conclusion that the discriminant function estimators are sensitive to the assumption of normality. In particular, the estimators of the coefficients for non-normally distributed variables are biased away from zero when the coefficient is, in fact, different from zero. The practical implication of this is that for dichotomous independent variables (that

occur in many situations), the discriminant function estimators overestimate the magnitude of the coefficient. Lyles et al. (2009) describe a clever linear regression-based approach to compute the discriminant function estimator of the coefficient for a single continuous variable that, when their assumptions of normality hold, has better statistical properties than the maximum likelihood estimator. We discuss their multivariable extension and some of its practical limitations in Chapter 2.

At this point it may be helpful to delineate more carefully the various uses of the term *maximum likelihood*, as it applies to the estimation of the logistic regression coefficients. Under the assumptions of the discriminant function model stated above, the estimators obtained from equations (1.23) and (1.24) are maximum likelihood estimators. The estimators obtained from equations (1.5) and (1.6) are based on the conditional distribution of Y given X and, as such, are technically "conditional maximum likelihood estimators". It is common practice to drop the word "conditional" when describing the estimators given in equations (1.5) and (1.6). In this text, we use the word *conditional* to describe estimators in logistic regression with matched data as discussed in Chapter 7.

In summary there are alternative methods of estimation for some data configurations that are computationally quicker; however, we use the maximum likelihood method described in Section 1.2 throughout the rest of this text.

1.6 DATA SETS USED IN EXAMPLES AND EXERCISES

A number of different data sets are used in the examples as well as the exercises for the purpose of demonstrating various aspects of logistic regression modeling. Six of the data sets used throughout the text are described below. Other data sets are introduced as needed in later chapters. Some of the data sets were used in the previous editions of this text, for example the ICU and Low Birth Weight data, while others are new to this edition. All data sets used in this text may be obtained from links to web sites at John Wiley & Sons Inc. and the University of Massachusetts given in the Preface.

1.6.1 The ICU Study

The ICU study data set consists of a sample of 200 subjects who were part of a much larger study on survival of patients following admission to an adult intensive care unit (ICU). The major goal of this study was to develop a logistic regression model to predict the probability of survival to hospital discharge of these patients. A number of publications have appeared that have focused on various facets of this problem. The reader wishing to learn more about the clinical aspects of this study should start with Lemeshow et al. (1988). For a more up-to-date discussion of modeling the outcome of ICU patients the reader is referred to Lemeshow and Le Gall (1994) and to Lemeshow et al. (1993). The actual observed variable values have been modified to protect subject confidentiality. A code sheet for the variables to be considered in this text is given in Table 1.5. We refer to this data set as the ICU data.

Table 1.5 Code Sheet for the Variables in the ICU Data

Variable	Description	Codes/Values	Name
1	Identification code	ID number	ID
2	Vital status at hospital discharge	0 = Lived 1 = Died	STA
3	Age	Years	AGE
4	Gender	0 = Male 1 = Female	GENDER
5	Race	1 = White 2 = Black 3 = Other	RACE
6	Service at ICU admission	0 = Medical 1 = Surgical	SER
7	Cancer part of present problem	0 = No 1 = Yes	CAN
8	History of chronic renal failure	0 = No 1 = Yes	CRN
9	Infection probable at ICU admission	0 = No 1 = Yes	INF
10	CPR prior to ICU admission	0 = No 1 = Yes	CPR
11	Systolic blood pressure at ICU admission	mm Hg	SYS
12	Heart rate at ICU admission	Beats/min	HRA
13	Previous admission to an ICU within 6 months	0 = No 1 = Yes	PRE
14	Type of admission	0 = Elective 1 = Emergency	TYPE
15	Long bone, multiple, neck, single area, or hip fracture	0 = No 1 = Yes	FRA
16	PO_2 from initial blood gases	0 = >60 1 = ≤60	PO2
17	PH from initial blood gases	0 = ≥7.25 1 = <7.25	PH
18	PCO_2 from initial blood gases	0 = ≤45 1 = >45	PCO
19	Bicarbonate from initial blood gases	0 = ≥18 1 = <18	BIC
20	Creatinine from initial blood gases	0 = ≤2.0 1 = >2.0	CRE
21	Level of consciousness at ICU admission	0 = No coma or deep stupor 1 = Deep stupor 2 = Coma	LOC

Table 1.6 Code Sheet for the Variables in the Low Birth Weight Data

Variable	Description	Codes/Values	Name
1	Identification code	1–189	ID
2	Low birth weight	0 = ≥2500 g 1 = <2500 g	LOW
3	Age of mother	Years	AGE
4	Weight of mother at last menstrual period	Pounds	LWT
5	Race	1 = White 2 = Black 3 = Other	RACE
6	Smoking status during pregnancy	0 = No 1 = Yes	SMOKE
7	History of premature labor	0 = None 1 = One 2 = Two, etc.	PTL
8	History of hypertension	0 = No 1 = Yes	HT
9	Presence of uterine irritability	0 = No 1 = Yes	UI
10	Number of physician visits during the first trimester	0 = None 1 = One 2 = Two, etc.	FTV
11	Recorded birth weight	Grams	BWT

1.6.2 The Low Birth Weight Study

Low birth weight, defined as birth weight less than 2500 grams, is an outcome that has been of concern to physicians for years. This is because of the fact that infant mortality rates and birth defect rates are higher for low birth weight babies. A woman's behavior during pregnancy (including diet, smoking habits, and receiving prenatal care) can greatly alter the chances of carrying the baby to term, and, consequently, of delivering a baby of normal birth weight.

Data were collected as part of a larger study at Baystate Medical Center in Springfield, Massachusetts. This data set contains information on 189 births to women seen in the obstetrics clinic. Fifty-nine of these births were low birth weight. The variables identified in the code sheet given in Table 1.6 have been shown to be associated with low birth weight in the obstetrical literature. The goal of the current study was to determine whether these variables were risk factors in the clinic population being served by Baystate Medical Center. Actual observed variable values have been modified to protect subject confidentiality. We refer to this data set as the LOWBWT data.

1.6.3 The Global Longitudinal Study of Osteoporosis in Women

The Global Longitudinal Study of Osteoporosis in Women (GLOW) is an international study of osteoporosis in women over 55 years of age being coordinated at the

Table 1.7 Code Sheet for Variables in the GLOW Study

Variable	Description	Codes/Values	Name
1	Identification code	1–n	SUB_ID
2	Study site	1–6	SITE_ID
3	Physician ID code	128 unique codes	PHY_ID
4	History of prior fracture	1 = Yes 0 = No	PRIORFRAC
5	Age at enrollment	Years	AGE
6	Weight at enrollment	Kilograms	WEIGHT
7	Height at enrollment	Centimeters	HEIGHT
8	Body mass index	kg/m^2	BMI
9	Menopause before age 45	1 = Yes 0 = No	PREMENO
10	Mother had hip fracture	1 = Yes 0 = No	MOMFRAC
11	Arms are needed to stand from a chair	1 = Yes 0 = No	ARMASSIST
12	Former or current smoker	1 = Yes 0 = No	SMOKE
13	Self-reported risk of fracture	1 = Less than others of the same age 2 = Same as others of the same age 3 = Greater than others of the same age	RATERISK
14	Fracture risk score	Composite risk score[a]	FRACSCORE
15	Any fracture in first year	1 = Yes 0 = No	FRACTURE

[a]FRACSCORE = 0 × (AGE ≤ 60) + 1 × (60 < AGE ≤ 65) + 2 × (65 < AGE ≤ 70) + 3 × (70 < AGE ≤ 75) + 4 × (75 < AGE ≤ 80) + 5 × (80 < AGE ≤ 85) + 6 × (AGE > 85) + (PRIORFRAC = 1) + (MOMFRAC = 1) + (WEIGHT < 56.8) + 2 × (ARMASSIST = 1) + (SMOKE = 1).

Center for Outcomes Research (COR) at the University of Massachusetts/Worcester by its Director, Dr. Frederick Anderson, Jr. The study has enrolled over 60,000 women aged 55 and older in ten countries. The major goals of the study are to use the data to provide insights into the management of fracture risk, patient experience with prevention and treatment of fractures and distribution of risk factors among older women on an international scale over the follow up period. Complete details on the study as well as a list of GLOW publications may be found at the Center for Outcomes Research web site, www.outcomes-umassmed.org/glow.

Data used here come from six sites in the United States and include a few selected potential risk factors for fracture from the baseline questionnaire. The outcome variable is any fracture in the first year of follow up. The incident first-year fracture rate among the 21,000 subjects enrolled in these six sites is about 4 percent. In order to have a data set of a manageable size, $n = 500$, for this text we have over sampled the fractures and under sampled the non-fractures. As a

result associations and conclusions from modeling these data do not apply to the study cohort as a whole. Data have been modified to protect subject confidentiality. We thank Dr. Gordon Fitzgerald of COR for his help in obtaining these data sets. A code sheet for the variables is shown in Table 1.7. This data set is named the GLOW500 data.

1.6.4 The Adolescent Placement Study

Fontanella et al. (2008) present results from a study of determinants of aftercare placement for psychiatrically hospitalized adolescents and have made the data, suitably modified to protect confidentiality, available to us. It is not our intent to repeat

Table 1.8 Code Sheet for Variables in the Adolescent Placement Study

Variable	Description	Codes/Values	Name
1	Identification code	1–508	ID
2	Placement	0 = Outpatient 1 = Day treatment 2 = Intermediate residential 3 = Residential	PLACE
3	Placement combined	0 = Outpatient or day treatment 1 = Intermediate residential 2 = Residential	PLACE3
3	Age at admission	Years	AGE
4	Race	0 = White 1 = Nonwhite	RACE
5	Gender	0 = Female 1 = Male	GENDER
6	Neuropsychiatric disturbance	0 = None 1 = Mild 2 = Moderate 3 = Severe	NEURO
7	Emotional disturbance	0 = Not severe 1 = Severe	EMOT
8	Danger to others	0 = Unlikely 1 = Possible 2 = Probable 3 = Likely	DANGER
9	Elopement risk	0 = No risk 1 = At risk	ELOPE
10	Length of hospitalization	Days	LOS
11	Behavioral symptoms score[a]	0–9	BEHAV
12	State custody	0 = No 1 = Yes	CUSTD
13	History of violence	0 = No 1 = Yes	VIOL

[a]Behavioral symptom score is based on the sum of three symptom subscales (oppositional behavior, impulsivity, and conduct disorder) from the CSPI.

the detailed analyses reported in their paper, but rather to use the data to motivate and describe methods for modeling a multinomial or ordinal scaled outcome using logistic regression models. As such, we selected a subset of variables, which are described in Table 1.8. This data set is referred to as the *APS data*.

1.6.5 The Burn Injury Study

The April 2008 release (Version 4.0) of the National Burn Repository research dataset (National Burn Repository 2007 Report, Dataset Version 4.0 accessed on 12/05/2008 at: http://www.ameriburn.org/2007NBRAnnualReport.pdf) includes information on a total of 306,304 burn related hospitalizations that occurred between 1973 and 2007. Available information included patient demographics, total burn surface area, presence of inhalation injury, and blinded trauma center identifiers. The outcome of interest is survival to hospital discharge. Osler et al. (2010) selected a subset of approximately 40,000 subjects treated between 2000 and 2007 at 40 different burn facilities to develop a new predictive logistic regression model (see the paper for the details on how this subset was selected). To obtain a much smaller data set for use in this text we over sampled subjects who died in hospital and under sampled subjects who lived to obtain a data set with $n = 1000$ and achieve a sample with 15 percent in hospital mortality. As such, all analyses and inferences contained in this text do not apply to the sample of 40,000, the original data from the registry or the population of burn injury patients as a whole. These data are used here to illustrate methods when prediction is the final goal as well as to demonstrate various model building techniques. The variables are described in Table 1.9 and the data are referred to as the *BURN1000 data*.

Table 1.9 Code Sheet for Variables in the Burn Study

Variable	Description	Codes/Values	Name
1	Identification code	1–1000	ID
2	Burn facility	1–40	FACILITY
3	Hospital discharge status	0 = Alive 1 = Dead	DEATH
4	Age at admission	Years	AGE
5	Gender	0 = Female 1 = Male	GENDER
6	Race	0 = Non-White 1 = White	RACE
7	Total burn surface area	0–100%	TBSA
8	Burn involved inhalation injury	0 = No 1 = Yes	INH_INJ
9	Flame involved in burn injury	0 = No 1 = Yes	FLAME

Table 1.10 Code Sheet for Variables in the Myopia Study

Variable	Variable Description	Values/Labels	Variable Name
1	Subject identifier	Integer (range 1–1503)	ID
2	Year subject entered the study	Year	STUDYYEAR
3	Myopia within the first 5 yr of follow up[a]	0 = No 1 = Yes	MYOPIC
4	Age at first visit	Years	AGE
5	Gender	0 = Male 1 = Female	GENDER
6	Spherical equivalent refraction[b]	Diopter	SPHEQ
7	Axial length[c]	mm	AL
8	Anterior chamber depth[d]	mm	ACD
9	Lens thickness[e]	mm	LT
10	Vitreous chamber depth[f]	mm	VCD
11	How many hours per week outside of school the child spent engaging in sports/outdoor activities	Hours per week	SPORTHR
12	How many hours per week outside of school the child spent reading for pleasure	Hours per week	READHR
13	How many hours per week outside of school the child spent playing video/computer games or working on the computer	Hours per week	COMPHR
14	How many hours per week outside of school the child spent reading or studying for school assignments	Hours per week	STUDYHR
15	How many hours per week outside of school the child spent watching television	Hours per week	TVHR
16	Composite of near-work activities	Hours per week	DIOPTERHR
17	Was the subject's mother myopic?[g]	0 = No 1 = Yes	MOMMY
18	Was the subject's father myopic?	0 = No 1 = Yes	DADMY

[a] MYOPIC is defined as SPHEQ <= −0.75D.

[b] A measure of the eye's effective focusing power. Eyes that are "normal" (don't require glasses or contact lenses) have spherical equivalents between −0.25 diopters (D) and +1.00 D. The more negative the spherical equivalent, the more myopic the subject.

[c] The length of eye from front to back.

[d] The length from front to back of the aqueous-containing space of the eye between the cornea and the iris.

[e] The length from front to back of the crystalline lens.

[f] The length from front to back of the aqueous-containing space of the eye in front of the retina.

[g] DIOPTERHR = 3 × (READHR + STUDYHR) + 2 × COMPHR + TVHR.

Table 1.11 Variables in the Modified NHANES Data Set

Variable	Description	Code/values	Name
1	Identification code	1–6482	ID
2	Gender	0 = Male, 1 = Female	GENDER
3	Age at screening	Years	AGE
4	Marital status	1 = Married 2 = Widowed 3 = Divorced 4 = Separated 5 = Never married 6 = Living together	MARSTAT
5	Statistical weight	4084.478–153810.3	SAMPLEWT
6	Pseudo-PSU	1, 2	PSU
7	Pseudo-stratum	1–15	STRATA
8	Total cholesterol	mg/dl	TCHOL
9	HDL-cholesterol	mg/dl	HDL
10	Systolic blood pressure	mm Hg	SYSBP
11	Diastolic blood pressure	mm Hg	DBP
12	Weight	kg	WT
13	Standing height	cm	HT
14	Body mass index	kg/m^2	BMI
15	Vigorous work activity	0 = Yes, 1 = No	VIGWRK
16	Moderate work activity	0 = Yes, 1 = No	MODWRK
17	Walk or bicycle	0 = Yes, 1 = No	WLKBIK
18	Vigorous recreational activities	0 = Yes, 1 = No	VIGRECEXR
19	Moderate recreational activities	0 = Yes, 1 = No	MODRECEXR
20	Minutes of sedentary activity per week	Minutes	SEDMIN
21	BMI > 35	0 = No, 1 = Yes	OBESE

1.6.6 The Myopia Study

Myopia, more commonly referred to as *nearsightedness*, is an eye condition where an individual has difficulty seeing things at a distance. This condition is primarily because the eyeball is too long. In an eye that sees normally, the image of what is being viewed is transmitted to the back portion of the eye, or retina, and hits the retina to form a clear picture. In the myopic eye, the image focuses in front of the retina, so the resultant image on the retina itself is blurry. The blurry image creates problems with a variety of distance viewing tasks (e.g., reading the blackboard,

Table 1.12 Code Sheet for the Variables in the Polypharmacy Data Set

Variable	Description	Codes/Values	Name
1	Subject ID	ID number 1–500	ID
2	Outcome; taking drugs from more than three different classes	0 = Not taking drugs from more than three classes 1 = Taking drugs from more than three classes	POLYPHARMACY
3	Number of outpatient mental health visits (MHV)	0 = None 1 = One to five 2 = Six to fourteen 3 = Greater than 14	MHV4
4	Number of inpatient mental health visits (MHV)	0 = None 1 = One 2 = More than one	INPTMHV3
5	Year	2002–2008	YEAR
6	Group	1 = Covered families and children (CFC) 2 = Aged, blind or disabled (ABD) 3 = Foster care (FOS)	GROUP
7	Location	0 = Urban 1 = Rural	URBAN
8	Comorbidity	0 = No 1 = Yes	COMORBID
9	Any primary diagnosis (bipolar, depression, etc.)	0 = No 1 = Yes	ANYPRIM
10	Number of primary diagnosis	0 = None 1 = One 2 = More than one	NUMPRIMRC
11	Gender	0 = Female 1 = Male	GENDER
12	Race	0 = White 1 = Black 2 = Other	RACE
13	Ethnic category	0 = NonHispanic 1 = Hispanic	ETHNIC
14	Age	Years and months (two decimal places)	AGE

doing homework, driving, playing sports) and requires wearing glasses or contact lenses to correct the problem. Myopia onset is typically between the ages of 8 and 12 years with cessation of the underlying eye growth that causes it by age 15–16 years.

The risk factors for the development of myopia have been debated for a long time and include genetic factors (e.g., family history of myopia) and the amount

and type of visual activity that a child performs (e.g., studying, reading, TV watching, computer or video game playing, and sports/outdoor activity). There is strong evidence that having myopic parents increases the chance that a child will become myopic, and weaker evidence that certain types of visual activities (called *near work*, e.g., reading) increase the chance that a child will become myopic.

These data are a subset of data from the Orinda Longitudinal Study of Myopia (OLSM), a cohort study of ocular component development and risk factors for the onset of myopia in children, which evolved into the Collaborative Longitudinal Evaluation of Ethnicity and Refractive Error (CLEERE) Study, and both OLSM and CLEERE were funded by the National Institutes of Health/National Eye Institute. OLSM was based at the University of California, Berkeley [see Zadnik et al. (1993, 1994)]. Data collection began in the 1989–1990 school year and continued annually through the 2000–2001 school year. All data about the parts that make up the eye (the ocular components) were collected during an examination during the school day. Data on family history and visual activities were collected yearly in a survey completed by a parent or guardian.

The dataset used in this text is from 618 of the subjects who had at least five years of followup and were not myopic when they entered the study. All data are from their initial exam and includes 17 variables. In addition to the ocular data there is information on age at entry, year of entry, family history of myopia and hours of various visual activities. The ocular data come from a subject's right eye. A subject was coded as myopic if they became myopic at any time during the first five years of followup. We refer to this data set, in Table 1.10, as the MYOPIA data.

1.6.7 The NHANES Study

The National Health and Nutrition Examination Survey (NHANES), a major effort of the National Center for Health Statistics, was conceived in the early 1960s to provide nationally representative and reliable data on the health and nutritional status of adults and children in the United States. NHANES has since evolved into a ongoing survey program that provides the best available national estimates of the prevalence of, and risk factors for, targeted diseases in the United States population. The survey collects interview and physical exam data on a nationally representative, multistage probability sample of about 5,000 persons each year, who are chosen to be representative of the civilian, non-institutionalized, population in the US.

For purposes of illustrating fitting logistic regression models to sample survey data in Section 6.4 we chose selected variables, shown in Table 1.11, from the 2009–2010 cycle of the National Health and Nutrition Examination Study [NHANES III Reference Manuals and Reports (2012)] and made some modifications to the data. We refer to this data set as the NHANES data.

1.6.8 The Polypharmacy Study

In Chapter 9, we illustrate model building with correlated data using data on polypharmacy described in Table 1.12. The outcome of interest is whether the

patient is taking drugs from three or more different classes (POLYPHARMACY), and researchers were interested in identifying factors associated with this outcome. We selected a sample of 500 subjects from among only those subjects with observations in each of the seven years data were collected. Based on the suggestions of the principal investigator, we initially treated the covariates for number of inpatient and outpatient mental health visits (MHVs) with categories described in Table 1.12. In addition we added a random number of months to the age, which was recorded only in terms of the year in the original data set. As our data set is a sample, the results in this section do not apply to the original study. We refer to this data set as the POLYPHARM data.

EXERCISES

1. In the ICU data described in Section 1.6.1 the primary outcome variable is vital status at hospital discharge, STA. Clinicians associated with the study felt that a key determinant of survival was the patient's age at admission, AGE.

(a) Write down the equation for the logistic regression model of STA on AGE. Write down the equation for the logit transformation of this logistic regression model. What characteristic of the outcome variable, STA, leads us to consider the logistic regression model as opposed to the usual linear regression model to describe the relationship between STA and AGE?

(b) Form a scatterplot of STA versus AGE.

(c) Using the intervals (15, 24), (25, 34), (35, 44), (45, 54), (55, 64), (65, 74), (75, 84), (85, 94) for age, compute the STA mean over subjects within each age interval. Plot these values of mean STA versus the midpoint of the age interval using the same set of axes as was used in 1(b). Note: this plot may done "by hand" on a printed copy of the plot from 1(b).

(d) Write down an expression for the likelihood and log-likelihood for the logistic regression model in Exercise 1(a) using the ungrouped, $n = 200$, data. Obtain expressions for the two likelihood equations.

(e) Using a logistic regression package of your choice obtain the maximum likelihood estimates of the parameters of the logistic regression model in Exercise 1(a). These estimates should be based on the ungrouped, $n = 200$, data. Using these estimates, write down the equation for the fitted values, that is, the estimated logistic probabilities. Plot the equation for the fitted values on the axes used in the scatterplots in 1(b) and 1(c).

(f) Using the results of the output from the logistic regression package used for 1(e), assess the significance of the slope coefficient for AGE using the likelihood ratio test, the Wald test, and if possible, the score test. What assumptions are needed for the p-values computed for each of these tests to be valid? Are the results of these tests consistent with one another? What is the value of the deviance for the fitted model?

(g) Using the results from 1(e) compute 95 percent confidence intervals for the slope coefficient for AGE. Write a sentence interpreting this confidence.

(h) Obtain from the package used to fit the model in 1(e) the estimated covariance matrix. Compute the logit and estimated logistic probability for a 60-year-old subject. Evaluate the endpoints of the 95 percent confidence intervals for the logit and estimated logistic probability. Write a sentence interpreting the estimated probability and its confidence interval.

2. In the Myopia Study described in Section 1.6.2, one variable that is clearly important is the initial value of spherical equivalent refraction (SPHREQ). Repeat steps (a)–(g) of Exercise 1, but for 2(c) use eight intervals containing approximately equal numbers of subjects (i.e., cut points at 12.5%, 25%, ..., etc.).

3. Using the data from the ICU study create a dichotomous variable NONWHITE (NONWHITE = 1 if RACE = 2 or 3 and NONWHITE = 0 if RACE = 1). Fit the logistic regression of STA on NONWHITE and show that the 95 percent profile likelihood confidence interval for the coefficient for nonwhite has asymmetry of -13% and that this interval is 26% wider than the Wald-based interval. This example points out that even when the sample size and number of events are large $n = 200$, and $n_1 = 40$ there can be substantial asymmetry and differences between the two interval estimators. Explain why this is the case in this example.