Applied Logistic Regression, Second Edition. By David W. Hosmer and Stanley Lemeshow
 ISBN: 0-471-72214-6

CHAPTER 1

Introduction to the Logistic Regression Model

1.1 INTRODUCTION

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

Before beginning a study of logistic regression it is important to understand that the goal of an analysis using this method is the same as that of any model-building technique used in statistics: to find the best fitting and most parsimonious, yet biologically reasonable model to describe the relationship between an outcome (dependent or response) variable and a set of independent (predictor or explanatory) variables. These independent variables are often called *covariates*. The most common example of modeling, and one assumed to be familiar to the readers of this text, is the usual linear regression model where the outcome variable is assumed to be continuous.

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

Example

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

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

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

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

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

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

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

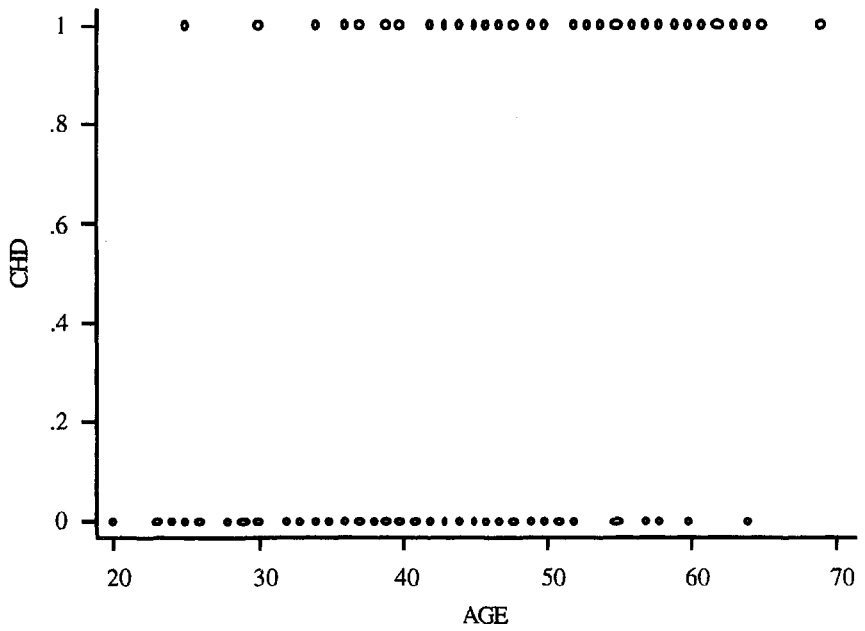

Figure 1.1 Scatterplot of CHD by AGE for 100 subjects.

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

The first difference concerns the nature of the relationship between the outcome and independent variables. In any regression problem the key quantity is the mean value of the outcome variable, given the value of the independent variable. This quantity is called the *conditional mean* and will be expressed as "$E(Y \mid x)$" where Y denotes the outcome

Table 1.2 Frequency Table of Age Group by CHD

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

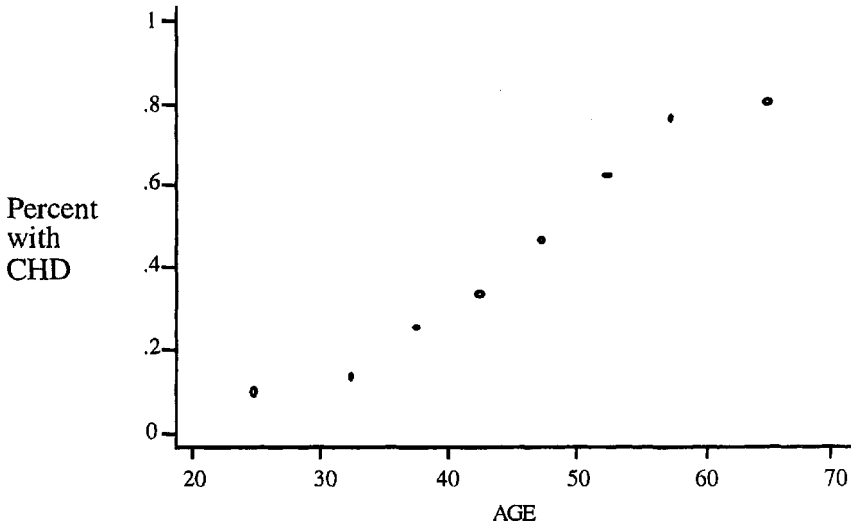

Figure 1.2 Plot of the percentage of subjects with CHD in each age group.

variable and x denotes a value of the independent variable. The quantity $E(Y \mid x)$ is read "the expected value of Y, given the value x." In linear regression we assume that this mean may be expressed as an equation linear in x (or some transformation of x or Y), such as

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

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

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

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

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

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

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

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

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

The importance of this transformation is that $g(x)$ has many of the desirable properties of a linear regression model. The logit, $g(x)$, is linear in its parameters, may be continuous, and may range from $-\infty$ to $+\infty$, depending on the range of x.

The second important difference between the linear and logistic regression models concerns the conditional distribution of the outcome variable. In the linear regression model we assume that an observation of the outcome variable may be expressed as $y = E(Y \mid x) + \varepsilon$. The quantity ε is called the *error* and expresses an observation's deviation from the conditional mean. The most common assumption is that ε follows a normal distribution with mean zero and some variance that is constant across levels of the independent variable. It follows that the

conditional distribution of the outcome variable given x will be normal with mean $E(Y \mid x)$, and a variance that is constant. This is not the case with a dichotomous outcome variable. In this situation we may express the value of the outcome variable given x as $y = \pi(x) + \varepsilon$. Here the quantity ε may assume one of two possible values. If $y = 1$ then $\varepsilon = 1 - \pi(x)$ with probability $\pi(x)$, and if $y = 0$ then $\varepsilon = -\pi(x)$ with probability $1 - \pi(x)$. Thus, ε has a distribution with mean zero and variance equal to $\pi(x)[1 - \pi(x)]$. That is, the conditional distribution of the outcome variable follows a binomial distribution with probability given by the conditional mean, $\pi(x)$.

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

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

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

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

1.2 FITTING THE LOGISTIC REGRESSION MODEL

Suppose we have a sample of n independent observations of the pair (x_i, y_i), $i = 1, 2, \ldots, n$, where y_i denotes the value of a dichotomous outcome variable and x_i is the value of the independent variable for the i^{th} subject. Furthermore, assume that the outcome variable has been coded as 0 or 1, representing the absence or the presence of the characteristic, respectively. This coding for a dichotomous outcome is used throughout the text. To fit the logistic regression model in equation (1.1) to a set of data requires that we estimate the values of β_0 and β_1, the unknown parameters.

In linear regression, the method used most often for estimating unknown parameters is *least squares*. In that method we choose those values of β_0 and β_1 which minimize the sum of squared deviations of the observed values of Y from the predicted values based upon the model. Under the usual assumptions for linear regression the method of least squares yields estimators with a number of desirable statistical proper-

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 $P(Y=1|x)$. It follows that the quantity $1-\pi(x)$ gives the conditional probability that Y is equal to zero given x, $P(Y=0|x)$. 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

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

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:

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

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

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

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:

$$\sum [y_i - \pi(x_i)] = 0 \qquad (1.5)$$

and

$$\sum x_i [y_i - \pi(x_i)] = 0. \qquad (1.6)$$

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|
AGE	0.111	0.0241	4.61	<0.001
Constant	−5.309	1.1337	−4.68	<0.001

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

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

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 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 multivariate case is discussed 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 would be called *goodness-of-fit*). Instead, our question is posed in a relative sense. The assessment of goodness-of-fit is a more complex question which 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 will motivate the approach used for logistic regression. A comparison of the two approaches will highlight the differences between modeling continuous and dichotomous response variables.

In linear regression, the assessment of the significance of the slope coefficient is approached by forming what is referred to as an *analysis of variance table*. This table partitions the total sum of squared devia-

tions 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 i^{th} 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 $\hat{\beta}_0 = \bar{y}$, the mean of the response variable. In this case, $\hat{y}_i = \bar{y}$ and SSE is equal to the total variance. When we include the independent variable in the model any decrease in SSE will be due to the fact that the slope coefficient for the independent variable is not zero. The change in the value of SSE is the due to the regression source of variability, denoted SSR. That is,

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

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* by some authors [see, for example, McCullagh and Nelder (1983)], and plays a central role in some approaches to assessing goodness-of-fit. The deviance for logistic regression 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 such as the one shown in Table 1.1, where the values of the outcome variable are either 0 or 1, the likelihood of the saturated model is 1. Specifically, it follows from the definition of a saturated model that $\hat{\pi}_i = y_i$ and the likelihood is

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

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, such as SAS, report the value of the deviance in (1.11) rather than the log likelihood for the fitted model. We discuss

the deviance in more detail in Chapter 5 in the context of evaluating model goodness-of-fit. At this stage we want to emphasize that we think of the deviance in the same terms that we think of the residual sum of squares in linear regression in the context of testing for the significance of a fitted model.

For purposes of assessing 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 obtained as:

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

This statistic plays the same role in logistic regression as the numerator of the partial F test does in linear regression. Because the likelihood of the saturated model is common to both values of D being differenced to compute G, it 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 value is constant, n_1/n. In this case, the value of G is:

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

or

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

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 also needed; however, for the above case they

are rather nonrestrictive and involve having a sufficiently large sample size, n.

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, $n_1 = 43$ and $n_0 = 57$; thus, 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.

The calculation of the log likelihood and the likelihood ratio test are standard features of all logistic regression software. This makes it easy to check for the significance of the addition of new terms to the model. In the simple case of a single independent variable, we first fit a model containing only the constant term. We then fit a model containing the independent variable along with the constant. This gives rise to a new log likelihood. The likelihood ratio test is obtained by multiplying the difference between 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, may be obtained from most software packages.

Two other similar, statistically equivalent tests have been suggested. These are the Wald test and the Score test. The assumptions needed for these tests are the same as those of the likelihood ratio test in equation (1.13). A more complete discussion of these tests and their assumptions may be found in Rao (1973).

The Wald test is obtained by comparing the maximum likelihood estimate of the slope parameter, $\hat{\beta}_1$, to an estimate of its standard error. The resulting ratio, under the hypothesis that $\beta_1 = 0$, will follow 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 logistic regression model in Table 1.3 is provided in the column headed z and is

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

and the two tailed p-value, provided in the last column of Table 1.3, is $P(|z| > 4.61)$, where z denotes a random variable following the standard normal distribution. 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. They recommended that the likelihood ratio test be used.

Jennings (1986a) has also looked at the adequacy of inferences in logistic regression based on Wald statistics. His conclusions are similar to those of Hauck and Donner. Both the likelihood ratio test, G, and the Wald test, W, require the computation of the maximum likelihood estimate for β_1.

A test for the significance of a variable which does not require these computations 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 cannot be obtained from some 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 which 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, $\hat{\pi} = n_1/n = \bar{y}$. Thus, the left-hand side of equation (1.6) becomes $\sum x_i(y_i - \bar{y})$. It may be shown that the estimated variance is $\bar{y}(1-\bar{y})\sum(x_i - \bar{x})^2$. The test statistic for the Score test (ST) is

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

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

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

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 uses the likelihood function for a dichotomous outcome variable.

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 based on their respective Wald tests. The endpoints of a $100(1-\alpha)\%$ confidence interval for the slope coefficient are

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

and for the intercept they are

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

where $z_{1-\alpha/2}$ is the upper $100(1-\alpha/2)\%$ point from the standard normal distribution and $\hat{\mathrm{SE}}(\cdot)$ 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 (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 in the absence of x 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 expression (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 constant is important when considering point and interval estimators of the logit.

The logit is the linear part of the logistic regression model and, as such, is most like 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

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

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

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

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

The estimated logit for the fitted model in Table 1.3 is shown in (1.8). In order to evaluate (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 (1.8) for a subject of age 50 is

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

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

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

and the estimated standard error is $\hat{\mathrm{SE}}[\hat{g}(50)] = 0.2549$. 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 (1.7) at age 50 the estimated logistic probability is

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

	AGE	Constant
AGE	0.000579	
Constant	−0.026677	1.28517

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

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

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

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 persons new to logistic regression modeling is to try and apply estimates on the probability scale to individual subjects. The fitted value computed in (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 (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. Each 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 the use of model-based fitted values like the one in (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 discuss discriminant analysis within the context of a method for obtaining estimators of the coefficients in the next section.

1.5 OTHER METHODS OF ESTIMATION

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 was proposed by Grizzle, Starmer, and Koch (1969), which 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 with their methods. We should add that the maximum likelihood estimators are usually calculated using an iterative reweighted least squares algorithm, and thus are also "least squares" estimators. The approach suggested by Grizzle et al. uses only one iteration in the process.

A major limitation of this method is that we must have an estimate of $\pi(x)$ which is not zero or 1 for most values of x. An example where we could use both maximum likelihood and 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 estimators become identical.

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 having different means and the same variance, then the conditional distribution of Y given $X = x$ is the logistic regression model. That is, if

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

then $P(Y = 1 \mid x) = \pi(x)$. The symbol "$\sim$" is read "is distributed" and the "$\mathrm{N}\left(\mu, \sigma^2\right)$" 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

$$\beta_0 = \ln\left(\frac{\theta_1}{\theta_0}\right) - 0.5\left(\mu_1^2 - \mu_0^2\right)/\sigma^2 \tag{1.22}$$

and

$$\beta_1 = \left(\mu_1 - \mu_0\right)/\sigma^2, \tag{1.23}$$

where $\theta_j = P(Y = j)$, $j = 0, 1$. The discriminant function estimators of β_0 and β_1 are found by substituting estimators for μ_j, θ_j, $j = 0, 1$ and σ^2 into the above equations. The estimators usually used are $\hat{\mu}_j = \bar{x}_j$, 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

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

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; the multivariable case is presented in Chapter 2.

It is natural to ask why, if the discriminant function estimators are so easy to compute, are they not used in place of the maximum likelihood estimators? Halpern, Blackwelder, and Verter (1971) and Hosmer, Hosmer, and Fisher (1983) have 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 coef-

ficients for nonnormally 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 (which occur in many situations), the discriminant function estimators will overestimate the magnitude of the coefficient.

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 (1.22) and (1.15) are maximum likelihood estimators. Those obtained from equations (1.5) and (1.6) are based on the conditional distribution of Y given X and, as such, are actually "conditional maximum likelihood estimators." Because discriminant function estimators are rarely used anymore, the word conditional has been dropped when describing the estimators given in equations (1.5) and (1.6). 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 method of maximum likelihood described in Section 1.2 throughout the rest of this text.

1.6 DATA SETS

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. Four data sets used throughout the text are described below. Other data sets will be introduced as needed in later chapters. All data sets used in this text may be obtained from the text web sites at John Wiley & Sons Inc. and the University of Massachusetts as described 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 sur-

Table 1.5 Code Sheet for the ICU Data

Variable	Description	Codes/Values	Name
1	Identification Code	ID Number	ID
2	Vital Status	0 = Lived 1 = Died	STA
3	Age	Years	AGE
4	Sex	0 = Male 1 = Female	SEX
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	TYP
15	Long Bone, Multiple, Neck, Single Area, or Hip Fracture	0 = No 1 = Yes	FRA
16	PO2 from Initial Blood Gases	0 = >60 1 = ≤ 60	PO2
17	PH from Initial Blood Gases	0 = ≥ 7.25 1 = <7.25	PH
18	PCO2 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

vival to hospital discharge of these patients. A number of publications have appeared which 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, Teres, Avrunin, and Pastides (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, Teres, Klar, Avrunin, Gehlbach and Rapoport (1993). 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.

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 due to the fact that infant mortality rates and birth defect rates are very high 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.

1.6.3 The Prostate Cancer Study

A third data set involves a study of patients with cancer of the prostate. These data have been provided to us by Dr. Donn Young at The Ohio State University Comprehensive Cancer Center. The goal of the analysis is to determine whether variables measured at a baseline exam can be used to predict whether the tumor has penetrated the prostatic capsule. The data presented are a subset of variables from the main study. Of

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

Variable	Description	Codes/Values	Name
1	Identification Code	ID Number	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	Birth Weight	Grams	BWT

the 380 subjects considered here, 153 had a cancer that penetrated the prostatic capsule. Actual observed variable values have been modified to protect subject confidentiality. These data will be used primarily for exercises. A code sheet for the variables to be considered in this text is shown in Table 1.7.

1.6.4 The UMARU IMPACT Study

Our colleagues, Drs. Jane McCusker, Carol Bigelow, and Anne Stoddard, have provided us with a subset of data from the University of Massachusetts Aids Research Unit (UMARU) IMPACT Study (UIS). This was a 5-year (1989–1994) collaborative research project (Benjamin F. Lewis, P.I., National Institute on Drug Abuse Grant #R18-DA06151) com-

Table 1.7 Code Sheet for the Prostate Cancer Study

Variable	Description	Codes/Values	Name
1	Identification Code	1 - 380	ID
2	Tumor Penetration of Prostatic Capsule	0 = No Penetration 1 = Penetration	CAPSULE
3	Age	Years	AGE
4	Race	1= White 2 = Black	RACE
5	Results of the Digital Rectal Exam	1 = No Nodule 2 = Unilobar Nodule (Left) 3 = Unilobar Nodule (Right) 4 = Bilobar Nodule	DPROS
6	Detection of Capsular Involvement in Rectal Exam	1 = No 2 = Yes	DCAPS
7	Prostatic Specific Antigen Value	mg/ml	PSA
8	Tumor Volume Obtained from Ultrasound	cm^3	VOL
9	Total Gleason Score	0 – 10	GLEASON

prised of two concurrent randomized trials of residential treatment for drug abuse. The purpose of the study was to compare treatment programs of different planned durations designed to reduce drug abuse and to prevent high-risk HIV behavior. The UIS sought to determine whether alternative residential treatment approaches are variable in effectiveness and whether efficacy depends on planned program duration.

We refer to the two treatment program sites as A and B in this text. The trial at site A randomized 444 participants and was a comparison of 3- and 6-month modified therapeutic communities which incorporated elements of health education and relapse prevention. Clients in the relapse prevention/health education program (site A) were taught to recognize "high-risk" situations that are triggers to relapse and were taught the skills to enable them to cope with these situations without using drugs. In the trial at site B, 184 clients were randomized to receive either a 6- or 12-month therapeutic community program involving a highly structured life-style in a communal living setting. Our colleagues have published a number of papers reporting the results of this study, see McCusker et. al. (1995, 1997a, 1997b).

Table 1.8 Description of Variables in the UMARU IMPACT Study

Variable	Description	Codes/Values	Name
1	Identification Code	1–575	ID
2	Age at Enrollment	Years	AGE
3	Beck Depression Score at Admission	0.000–54.000	BECK
4	IV Drug Use History at Admission	1 = Never 2 = Previous 3 = Recent	IVHX
5	Number of Prior Drug Treatments	0–40	NDRUGTX
6	Subject's Race	0 = White 1 = Other	RACE
7	Treatment Randomization Assignment	0 = Short 1 = Long	TREAT
8	Treatment Site	0 = A 1 = B	SITE
9	Returned to Drug Use Prior to the Scheduled End of the Treatment Program	1 = Remained Drug Free 0 = Otherwise	DFREE

As is shown in the coming chapters, the data from the UIS provide a rich setting for illustrating methods for logistic regression modeling. The data presented here are a subset of both variables and subjects of the data used to demonstrate methods for survival analysis in Hosmer and Lemeshow (1999). The small subset of variables from the main study we use in this text is described in Table 1.8. Since the analyses we report in this text are based on this small subset of variables and subjects, the results reported here should not be thought of as being in any way comparable to results of the main study. In addition we have taken the liberty in this text of simplifying the study design by representing the planned duration as short versus long. Thus, short versus long represents 3 months versus 6 months planned duration at site A, and 6 months versus 12 months planned duration at site B. The dichotomous outcome variable considered in this text is defined as having returned to drug use prior to the scheduled completion of the treatment program. The original data have been modified in such a way as to preserve subject confidentiality.

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 Exercise 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 Exercises 1(b) and 1(c).
   (f) Summarize (describe in words) the results presented in the plot obtained from Exercises 1(b), 1(c), and 1(e).
   (g) Using the results of the output from the logistic regression package used for Exercise 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?

(h) Using the results from Exercise 1(e) compute 95 percent confidence intervals for the slope and constant term. Write a sentence interpreting the confidence interval for the slope.

(i) Obtain the estimated covariance matrix for the model fit in Exercise 1(e). Compute the logit and estimated logistic probability for a 60-year old subject. Compute a 95 percent confidence intervals for the logit and estimated logistic probability. Write a sentence or two interpreting the estimated probability and its confidence interval.

(j) Use the logistic regression package to obtain the estimated logit and its standard error for each subject in the ICU study. Graph the estimated logit and the pointwise 95 percent confidence limits versus AGE for each subject. Explain (in words) the similarities and differences between the appearance of this graph and a graph of a fitted linear regression model and its pointwise 95 percent confidence bands.

2. Use the ICU Study and repeat Exercises 1(a), 1(b), 1(d), 1(e) and 1(g) using the variable "type of admission," TYP, as the covariate.

3. In the Low Birth Weight Study described in Section 1.6.2, one variable that physicians felt was important to control for was the weight of the mother at the last menstrual period, LWT. Repeat steps (a) – (g) of Exercise 1, but for Exercise 3(c) use intervals [80, 99], [100, 109], [110, 114], [115, 119], [120, 124], [125, 129], [130, 250].

   (h) The graph in Exercises 3(c) does not look "S-Shaped". The primary reason is that the range of plotted values is from approximately 0.2 to 0.56. Explain why a model for the probability of low birth weight as a function of LWT could still be the logistic regression model.

4. In the Prostate Cancer Study described in Section 1.6.3, one variable thought to be particularly predictive of capsule penetration is the prostate specific antigen level, PSA. Repeat steps (a) – (g) and (j) of Exercise 1 using CAPSULE as the outcome variable and PSA as the covariate. For Exercises 4(c) use intervals for PSA of [0, 2.4], [2.5, 4.4], [4.5, 6.4], [6.5, 8.4], [8.5, 10.4], [10.5, 12.4], [12.5, 20.4], [20.5, 140].