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

CHAPTER 3

Interpretation of the Fitted Logistic Regression Model

3.1 INTRODUCTION

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

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

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

For a linear regression model recall that the slope coefficient, β_1, is equal to the difference between the value of the dependent variable at $x+1$ and the value of the dependent variable at x, for any value of x. For example, if $y(x) = \beta_0 + \beta_1 x$, it follows that $\beta_1 = y(x+1) - y(x)$. In this case, the interpretation of the coefficient is relatively straightforward as it expresses the resulting change in the measurement scale of the dependent variable for a unit change in the independent variable. For example, if in a regression of weight on height of male adolescents the slope is 5, then we would conclude that an increase of 1 inch in height is associated with an increase of 5 pounds in weight.

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

3.2 DICHOTOMOUS INDEPENDENT VARIABLE

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

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

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

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

The possible values of the logistic probabilities may be conveniently displayed in a 2×2 table as shown in Table 3.1. The *odds* of the outcome being present among individuals with $x=1$ is defined as $\pi(1)/[1-\pi(1)]$. Similarly, the odds of the outcome being present among individuals with $x=0$ is defined as $\pi(0)/[1-\pi(0)]$. The *odds ratio*, denoted OR, is defined as the ratio of the odds for $x=1$ to the odds for $x=0$, and is given by the equation

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Log likelihood = −58.9795

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.3 POLYCHOTOMOUS INDEPENDENT VARIABLE

Suppose that instead of two categories the independent variable has $k > 2$ distinct values. For example, we may have variables that denote the county of residence within a state, the clinic used for primary health care within a city, or race. Each of these variables has a fixed number of discrete values and the scale of measurement is nominal. We saw in Chapter 2 that it is inappropriate to model a nominal scale variable as if it were an interval scale variable. Therefore, we must form a set of design variables to represent the categories of the variable. In this section we present methods for creating design variables for polychotomous independent variables. The choice of a particular method depends to some extent on the goals of the analysis and the stage of model development.

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

Table 3.5 Cross-Classification of Hypothetical Data on RACE and CHD Status for 100 Subjects

CHD Status	White	Black	Hispanic	Other	Total
Present	5	20	15	10	50
Absent	20	10	10	10	50
Total	25	30	25	20	100
Odds Ratio	1	8	6	4	
95 % CI		(2.3, 27.6)	(1.7, 21.3)	(1.1, 14.9)	
$\ln(\widehat{\text{OR}})$	0.0	2.08	1.79	1.39	

Table 3.6 Specification of the Design Variables for RACE Using Reference Cell Coding with White as the Reference Group

	Design Variables		
RACE(Code)	RACE_2	RACE_3	RACE_4
White (1)	0	0	0
Black (2)	1	0	0
Hispanic (3)	0	1	0
Other (4)	0	0	1

RACE by CHD status yields the data in Table 3.5. These data are hypothetical and have been formulated for ease of computation. The extension to a situation where the variable has more than four levels is not conceptually different, so all the examples in this section use $k = 4$.

At the bottom of Table 3.5, the odds ratio is given for each race, using White as the reference group. For example, for Hispanic the estimated odds ratio is $15 \times 20/5 \times 10$. The log of each odds ratio is given in the last row of Table 3.5. This table is typical of what is found in the literature. The reference group is indicated by a value of 1 for the odds ratio. These same estimates of the odds ratio may be obtained from a logistic regression program with an appropriate choice of design variables. The method for specifying the design variables involves setting all of them equal to zero for the reference group, and then setting a single design variable equal to 1 for each of the other groups. This is illustrated in Table 3.6. As noted in Section 3.2 this method is usually referred to as *reference cell* coding and is the default method in many packages.

Use of any logistic regression program with design variables coded as shown in Table 3.6 yields the estimated logistic regression coefficients given in Table 3.7.

A comparison of the estimated coefficients in Table 3.7 to the log odds ratios in Table 3.5 shows that

$$\ln\left[\hat{\text{OR}}(\text{Black, White})\right] = \hat{\beta}_1 = 2.079,$$

$$\ln\left[\hat{\text{OR}}(\text{Hispanic, White})\right] = \hat{\beta}_2 = 1.792,$$

and

Table 3.7 Results of Fitting the Logistic Regression Model to the Data in Table 3.5 Using the Design Variables in Table 3.6

Variable	Coeff.	Std. Err.	z	P>|z|
RACE_2	2.079	0.6325	3.29	0.001
RACE_3	1.792	0.6466	2.78	0.006
RACE_4	1.386	0.6708	2.07	0.039
Constant	−1.386	0.5000	−2.77	0.006

Log likelihood = −62.2937

$$\ln\left[\hat{\mathrm{OR}}(\text{Other, White})\right] = \hat{\beta}_3 = 1.386.$$

Did this happen by chance? Calculation of the logit difference shows that it is by design. The comparison of Black to White is as follows:

$$\begin{aligned}\ln\left[\hat{\mathrm{OR}}(\text{Black, White})\right] &= \hat{g}(\text{Black}) - \hat{g}(\text{White}) \\ &= \left[\begin{array}{c}\hat{\beta}_0 + \hat{\beta}_1 \times (RACE\_2 = 1) + \hat{\beta}_2 \times (RACE\_3 = 0) \\ + \hat{\beta}_3 \times (RACE\_4 = 0)\end{array}\right] \\ &\quad - \left[\begin{array}{c}\hat{\beta}_0 + \hat{\beta}_1 \times (\text{RACE\_2} = 0) + \hat{\beta}_2 \times (RACE\_3 = 0) \\ + \hat{\beta}_3 \times (RACE\_4 = 0)\end{array}\right] \\ &= \hat{\beta}_1.\end{aligned}$$

Similar calculations would demonstrate that the other coefficients estimated using logistic regression are also equal to the log of odds ratios computed from the data in Table 3.5.

A comment about the estimated standard errors may be helpful at this point. In the univariate case the estimates of the standard errors found in the logistic regression output are identical to the estimates obtained using the cell frequencies from the contingency table. For example, the estimated standard error of the estimated coefficient for the design variable RACE_2 is

$$\hat{\mathrm{SE}}\left(\hat{\beta}_1\right) = \left[\frac{1}{5} + \frac{1}{20} + \frac{1}{20} + \frac{1}{10}\right]^{0.5} = 0.6325.$$

A derivation of this result may be found in Bishop, Feinberg, and Holland (1975).

Confidence limits for odds ratios are obtained using the same approach used in Section 3.2 for a dichotomous variable. We begin by computing the confidence limits for the log odds ratio (the logistic regression coefficient) and then exponentiate these limits to obtain limits for the odds ratio. In general, the limits for a $100(1-\alpha)\%$ CIE for the coefficient are of the form

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

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

$$\exp\left[\hat{\beta}_j \pm z_{1-\alpha/2} \times \hat{\mathrm{SE}}\left(\hat{\beta}_j\right)\right]. \tag{3.6}$$

The confidence limits given in Table 3.5 in the row beneath the estimated odds ratios were obtained using the estimated coefficients and standard errors in Table 3.7 with (3.6) for $j = 1,2,3$ with $\alpha = 0.05$.

Reference cell coding is the most commonly employed coding method appearing in the literature. The primary reason for the widespread use of this method is the interest in estimating the risk of an "exposed" group relative to that of a "control" or "unexposed" group.

As discussed in Section 3.2 a second method of coding design variables is called *deviation from means* coding. This coding expresses effect as the deviation of the "group mean" from the "overall mean." In the case of logistic regression, the "group mean" is the logit for the

Table 3.8 Specification of the Design Variables for RACE Using Deviation from Means Coding

	Design Variables		
RACE(Code)	RACE_2	RACE_3	RACE_4
White (1)	−1	−1	−1
Black (2)	1	0	0
Hispanic (3)	0	1	0
Other (4)	0	0	1

Table 3.9 Results of Fitting the Logistic Regression Model to the Data in Table 3.5 Using the Design Variables in Table 3.8

Variable	Coeff.	Std. Err.	z	P>|z|
RACE_2	0.765	0.3506	2.18	0.029
RACE_3	0.477	0.3623	1.32	0.188
RACE_4	0.072	0.3846	0.19	0.852
Constant	−0.072	0.2189	−0.33	0.742

Log likelihood = −62.2937

group and the "overall mean" is the average logit over all groups. This method of coding is obtained by setting the value of all the design variables equal to −1 for one of the categories, and then using the 0, 1 coding for the remainder of the categories. Use of the deviation from means coding for race shown in Table 3.8 yields the estimated logistic regression coefficients in Table 3.9.

In order to interpret the estimated coefficients in Table 3.9 we need to refer to Table 3.5 and calculate the logit for each of the four categories of RACE. These are

$$\hat{g}_1 = \ln\left(\frac{5/25}{20/25}\right) = \ln\left(\frac{5}{20}\right) = -1.386$$

$\hat{g}_2 = \ln(20/10) = 0.693$, $\hat{g}_3 = \ln(15/10) = 0.405$, $\hat{g}_4 = \ln(10/10) = 0$, and their average is $\bar{g} = \sum \hat{g}_i/4 - 0.072$. The estimated coefficient for design variable RACE_2 in Table 3.9 is $\hat{g}_2 - \bar{g} = 0.693 - (-0.072) = 0.765$. The general relationship for the estimated coefficient for design variable RACE_j is $\hat{g}_j - \bar{g}$, for $j = 2,3,4$.

The interpretation of the estimated coefficients is not as easy or clear as in the situation when a reference group is used. Exponentiation of the estimated coefficients yields the ratio of the odds for the particular group to the geometric mean of the odds. Specifically, for RACE_2 in Table 3.9 we have

$$\begin{aligned}\exp(0.765) &= \exp(\hat{g}_2 - \bar{g}) \\ &= \exp(\hat{g}_2)/\exp\left(\sum \hat{g}_j/4\right) \\ &= (20/10)/\left[(5/20)\times(20/10)\times(15/10)\times(10/10)\right]^{0.25} \\ &= 2.15\,.\end{aligned}$$

This number, 2.15, is not a true odds ratio because the quantities in the numerator and denominator do not represent the odds for two distinct categories. The exponentiation of the estimated coefficient expresses the odds relative to an "average" odds, the geometric mean. The interpretation of this value depends on whether the "average" odds is in fact meaningful.

The estimated coefficients obtained using deviation from means coding may be used to estimate the odds ratio for one category relative to a reference category. The equation for the estimate is more complicated than the one obtained using the reference cell coding. However, it provides an excellent example of the basic principle of using the logit difference to compute an odds ratio.

To illustrate this we calculate the log odds ratio of Black versus White using the coding for design variables given in Table 3.8. The logit difference is as follows:

$$\begin{aligned}\ln\left[\hat{\mathrm{OR}}(\text{Black},\text{White})\right] &= \hat{g}(\text{Black}) - \hat{g}(\text{White}) \\ &= \left[\begin{array}{c}\hat{\beta}_0 + \hat{\beta}_1\times(\text{RACE\_2}=1) + \hat{\beta}_2\times(RACE\_3=0) \\ + \hat{\beta}_3\times(RACE\_4=0)\end{array}\right] \\ &\quad - \left[\begin{array}{c}\hat{\beta}_0 + \hat{\beta}_1\times(\text{RACE\_2}=-1) + \hat{\beta}_2\times(RACE\_3=-1) \\ + \hat{\beta}_3\times(RACE\_4=-1)\end{array}\right] \\ &= 2\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3. \qquad (3.7)\end{aligned}$$

To obtain a confidence interval we must estimate the variance of the sum of the coefficients in (3.7). In this example, the estimator is

$$\begin{aligned}\hat{\mathrm{Var}}\left\{\ln\left[\hat{\mathrm{OR}}(\text{Black},\text{White})\right]\right\} &= 4\times\hat{\mathrm{Var}}\left(\hat{\beta}_1\right) + \hat{\mathrm{Var}}\left(\hat{\beta}_2\right) \\ &\quad + \hat{\mathrm{Var}}\left(\hat{\beta}_3\right) + 4\times\hat{\mathrm{Cov}}\left(\hat{\beta}_1,\hat{\beta}_2\right) \\ &\quad + 4\times\hat{\mathrm{Cov}}\left(\hat{\beta}_1,\hat{\beta}_3\right) + 2\times\hat{\mathrm{Cov}}\left(\hat{\beta}_2,\hat{\beta}_3\right). \qquad (3.8)\end{aligned}$$

Values for each of the estimators in (3.8) may be obtained from output that is available from logistic regression software. Confidence intervals for the odds ratio are obtained by exponentiating the endpoints of the confidence limits for the sum of the coefficients in (3.7). Evaluation of (3.7) for the current example gives

$$\ln\left[\hat{\mathrm{OR}}(\text{Black, White})\right] = 2(0.765) + 0.477 + 0.072 = 2.079.$$

The estimate of the variance is obtained by evaluating (3.8) which, for the current example, yields

$$\hat{\mathrm{Var}}\left\{\ln\left[\hat{\mathrm{OR}}(\text{Black, White})\right]\right\} = 4(0.351)^2 + (0.362)^2 + (0.385)^2 + 4(-0.031) + 4(-0.040) + 2(-0.044) = 0.400$$

and the estimated standard error is

$$\hat{\mathrm{SE}}\left\{\ln\left[\hat{\mathrm{OR}}(\text{Black, White})\right]\right\} = 0.6325.$$

We note that the values of the estimated log odds ratio, 2.079, and the estimated standard error, 0.6325, are identical to the values of the estimated coefficient and standard error for the first design variable in Table 3.7. This is expected, since the design variables used to obtain the estimated coefficients in Table 3.7 were formulated specifically to yield the log odds ratio relative to the White race category.

It should be apparent that, if the objective is to obtain odds ratios, use of deviation from means coding for design variables is computationally much more complex than reference cell coding.

In summary, we have shown that discrete nominal scale variables are included properly into the analysis only when they have been recoded into design variables. The particular choice of design variables depends on the application, though the reference cell coding is the easiest to interpret, and thus is the one we use in the remainder of this text.

3.4 CONTINUOUS INDEPENDENT VARIABLE

When a logistic regression model contains a continuous independent variable, interpretation of the estimated coefficient depends on how it is entered into the model and the particular units of the variable. For purposes of developing the method to interpret the coefficient for a continuous variable, we assume that the logit is linear in the variable. Other modeling strategies that examine this assumption are presented in Chapter 4.

Under the assumption that the logit is linear in the continuous covariate, x, the equation for the logit is $g(x) = \beta_0 + \beta_1 x$. It follows that the slope coefficient, β_1, gives the change in the log odds for an increase of "1" unit in x, that is, $\beta_1 = g(x+1) - g(x)$ for any value of x. Most often the value of "1" is not clinically interesting. For example, a 1 year increase in age or a 1 mm Hg increase in systolic blood pressure may be too small to be considered important. A change of 10 years or 10 mm Hg might be considered more useful. On the other hand, if the range of x is from zero to 1, then a change of 1 is too large and a change of 0.01 may be more realistic. Hence, to provide a useful interpretation for continuous scale covariates we need to develop a method for point and interval estimation for an arbitrary change of "c" units in the covariate.

The log odds ratio for a change of c units in x is obtained from the logit difference $g(x+c) - g(x) = c\beta_1$ and the associated odds ratio is obtained by exponentiating this logit difference, $\mathrm{OR}(c) = \mathrm{OR}(x+c, x) = \exp(c\beta_1)$. An estimate may be obtained by replacing β_1 with its maximum likelihood estimate $\hat{\beta}_1$. An estimate of the standard error needed for confidence interval estimation is obtained by multiplying the estimated standard error of $\hat{\beta}_1$ by c. Hence the endpoints of the $100(1-\alpha)\%$ CI estimate of $\mathrm{OR}(c)$ are

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

Since both the point estimate and endpoints of the confidence interval depend on the choice of c, the particular value of c should be clearly specified in all tables and calculations. The rather arbitrary nature of the choice of c may be troublesome to some. For example, why use a change of 10 years when 5 or 15 or even 20 years may be equally good? We, of course, could use any reasonable value; but the goal must be kept in mind: to provide the reader of your analysis with a clear indi-

cation of how the risk of the outcome being present changes with the variable in question. Changes in multiples of 5 or 10 may be most meaningful and easily understood.

As an example, consider the univariate model in Table 1.3. In that example a logistic regression of AGE on CHD status using the data of Table 1.1 was reported. The resulting estimated logit was $\hat{g}(\text{AGE}) = -5.310 + 0.111 \times \text{AGE}$. The estimated odds ratio for an increase of 10 years in age is $\widehat{\text{OR}}(10) = \exp(10 \times 0.111) = 3.03$. This indicates that for every increase of 10 years in age, the risk of CHD increases 3.03 times. The validity of such a statement is questionable in this example, since the additional risk of CHD for a 40 year-old compared to a 30 year-old may be quite different from the additional risk of CHD for a 60 year-old compared to a 50 year-old. This is an unavoidable dilemma when continuous covariates are modeled linearly in the logit. If it is believed that the logit is not linear in the covariate, then grouping and use of dummy variables should be considered. Alternatively, use of higher order terms (e.g., $x^2, x^3, \ldots$) or other nonlinear scaling in the covariate (e.g., $\log(x)$) could be considered. Thus, we see that an important modeling consideration for continuous covariates is their scale in the logit. We consider this in considerable detail in Chapter 4. The endpoints of a 95% confidence interval for this odds ratio are

$$\exp(10 \times 0.111 \pm 1.96 \times 10 \times 0.024) = (1.90, 4.86).$$

Results similar to these may be placed in tables displaying the results of a fitted logistic regression model.

In summary, the interpretation of the estimated coefficient for a continuous variable is similar to that of nominal scale variables: an estimated log odds ratio. The primary difference is that a meaningful change must be defined for the continuous variable.

3.5 THE MULTIVARIABLE MODEL

In the previous sections in this chapter we discussed the interpretation of an estimated logistic regression coefficient in the case when there is a single variable in the fitted model. Fitting a series of univariate models rarely provides an adequate analysis of the data in a study since the independent variables are usually associated with one another and may

have different distributions within levels of the outcome variable. Thus, one generally considers a multivariable analysis for a more comprehensive modeling of the data. One goal of such an analysis is to *statistically adjust* the estimated effect of each variable in the model for differences in the distributions of and associations among the other independent variables. Applying this concept to a multivariable logistic regression model, we may surmise that each estimated coefficient provides an estimate of the log odds adjusting for all other variables included in the model.

A full understanding of the estimates of the coefficients from a multivariable logistic regression model requires that we have a clear understanding of what is actually meant by the term *adjusting, statistically, for other variables.* We begin by examining adjustment in the context of a linear regression model, and then extend the concept to logistic regression.

The multivariable situation we examine is one in which the model contains two independent variables — one dichotomous and one continuous — but primary interest is focused on the effect of the dichotomous variable. This situation is frequently encountered in epidemiologic research when an exposure to a risk factor is recorded as being either present or absent, and we wish to adjust for a variable such as age. The analogous situation in linear regression is called *analysis of covariance.*

Suppose we wish to compare the mean weight of two groups of boys. It is known that weight is associated with many characteristics, one of which is age. Assume that on all characteristics except age the two groups have nearly identical distributions. If the age distribution is also the same for the two groups, then a univariate analysis would suffice and we could compare the mean weight of the two groups. This comparison would provide us with a correct estimate of the difference in weight between the two groups. However, if one group was much younger than the other group, then a comparison of the two groups would be meaningless, since at least a portion of any difference observed would likely be due to the difference in age. It would not be possible to determine the effect of group without first eliminating the discrepancy in ages between the groups.

This situation is described graphically in Figure 3.1. In this figure it is assumed that the relationship between age and weight is linear, with the same significant nonzero slope in each group. Both of these assumptions would usually be tested in an analysis of covariance before making any inferences about group differences. We defer a discussion

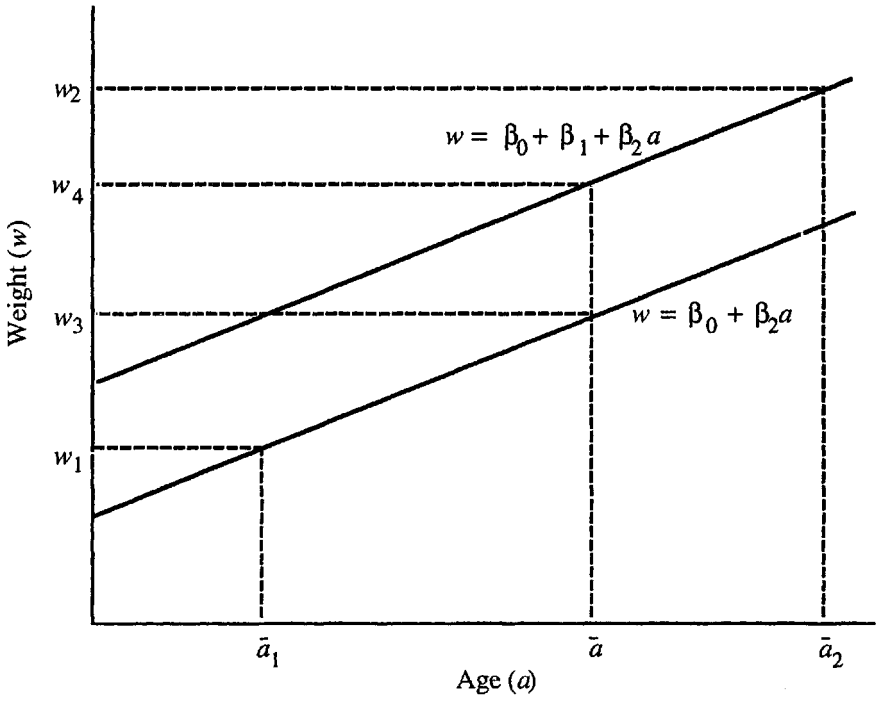

Figure 3.1 Comparison of the weight of two groups of boys with different distributions of age.

of this until Chapter 4, as it gets to the heart of modeling with logistic regression. We proceed as if these assumptions have been checked and are supported by the data.

The statistical model that describes the situation in Figure 3.1 states that the value of weight, w, may be expressed as $w = \beta_0 + \beta_1 x + \beta_2 a$, where $x = 0$ for group 1 and $x = 1$ for group 2 and "a" denotes age. In this model the parameter β_1 represents the true difference in weight between the two groups and β_2 is the rate of change in weight per year of age. Suppose that the mean age of group 1 is $\bar{a}_1$ and the mean age of group 2 is $\bar{a}_2$. These values are indicated in Figure 3.1. Comparison of the mean weight of group 1 to the mean weight of group 2 amounts to a comparison of w_1 to w_2. In terms of the model this difference is $(w_2 - w_1) = \beta_1 + \beta_2(\bar{a}_2 - \bar{a}_1)$. Thus the comparison involves not only the

Table 3.10 Descriptive Statistics for Two Groups of 50 Men on AGE and Whether They Had Seen a Physician (PHY) (1 = Yes, 0 = No) Within the Last Six Months

	Group 1		Group 2	
Variable	Mean	Std. Dev.	Mean	Std. Dev
PHY	0.36	0.485	0.80	0.404
AGE	39.60	5.272	47.34	5.259

true difference between the groups, β_1, but a component, $\beta_2(\bar{a}_2 - \bar{a}_1)$, which reflects the difference between the ages of the groups.

The process of statistically adjusting for age involves comparing the two groups at some common value of age. The value usually used is the mean of the two groups which, for the example, is denoted by $\bar{a}$ in Figure 3.1. In terms of the model this yields a comparison of w_4 to w_3, $(w_4 - w_3) = \beta_1 + \beta_2(\bar{a} - \bar{a}) = \beta_1$, the true difference between the two groups. In theory any common value of age could be used, as it would yield the same difference between the two lines. The choice of the overall mean makes sense for two reasons: it is biologically reasonable and lies within the range for which we believe that the association between age and weight is linear and constant within each group.

Consider the same situation shown in Figure 3.1, but instead of weight being the dependent variable, assume it is a dichotomous variable and that the vertical axis denotes the logit. That is, under the model the logit is given by the equation $g(x,a) = \beta_0 + \beta_1 x + \beta_2 a$. A univariate comparison obtained from the 2 × 2 table cross-classifying outcome and group would yield a log odds ratio approximately equal to $\beta_1 + \beta_2(\bar{a}_2 - \bar{a}_1)$. This would incorrectly estimate the effect of group due to the difference in the distribution of age. To account or adjust for this difference, we include age in the model and calculate the logit difference at a common value of age, such as the combined mean, $\bar{a}$. This logit difference is $g(x=1,\bar{a}) - g(x=0,\bar{a}) = \beta_1$. Thus, the coefficient β_1 is the log odds ratio that we would expect to obtain from a univariate comparison if the two groups had the same distribution of age.

The data summarized in Table 3.10 provide the basis for an example of interpreting the estimated logistic regression coefficient for a dichotomous variable when the coefficient is adjusted for a continuous variable.

It follows from the descriptive statistics in Table 3.10 that the univariate log odds ratio for group 2 versus group 1 is

$$\ln\left(\hat{\text{OR}}\right) = \ln(0.8/0.2) - \ln(0.36/0.64) = 1.962,$$

and the unadjusted estimated odds ratio is $\hat{\text{OR}} = 7.11$. We can also see that there is a considerable difference in the age distribution of the two groups, the men in group 2 being on average more than 7 years older than those in group 1. We would guess that much of the apparent difference in the proportion of men seeing a physician might be due to age. Analyzing the data with a bivariate model using a coding of GROUP = 0 for group 1, and GROUP = 1 for group 2, yields the estimated logistic regression coefficients shown in Table 3.11. The age-adjusted log odds ratio is given by the estimated coefficient for group in Table 3.11 and is $\hat{\beta}_1 = 1.263$. The age adjusted odds ratio is $\hat{\text{OR}} = \exp(1.263) = 3.54$. Thus, much of the apparent difference between the two groups is, in fact, due to differences in age.

Let us examine this adjustment in more detail using Figure 3.1. An approximation to the unadjusted odds ratio is obtained by exponentiating the difference $w_2 - w_1$. In terms of the fitted logistic regression model shown in Table 3.11 this difference is

$$[-4.866 + 1.263 + 0.107(47.34)] - [-4.866 + 0.107(39.60)] = 1.263 + 0.107(47.34 - 39.60).$$

The value of this odds ratio is

$$e^{[1.263 + 0.107(47.34 - 39.60)]} = 8.09.$$

The discrepancy between 8.09 and the actual unadjusted odds ratio, 7.11, is due to the fact that the above comparison is based on the difference in the average logit, while the crude odds ratio is approximately equal to a calculation based on the average estimated logistic probability for the two groups. The age adjusted odds ratio is obtained by exponentiating the difference $w_4 - w_3$, which is equal to the estimated coefficient for GROUP. In the example the difference is

$$[-4.866 + 1.263 + 0.107(43.47)] - [-4.866 + 0.107(43.47)] = 1.263.$$

Table 3.11 Results of Fitting the Logistic Regression Model to the Data Summarized in Table 3.10

Variable	Coeff.	Std. Err.	z	P>|z|
GROUP	1.263	0.5361	2.36	0.018
AGE	0.107	0.0465	2.31	0.021
Constant	−4.866	1.9020	−2.56	0.011

Log likelihood = −54.8292

Bachand and Hosmer (1999) compare two different sets of criteria for defining a covariate to be a confounder. They show that the numeric approach used in this Section, examining the change in the magnitude of the coefficient for the risk factor from logistic regression models fit with and without the potential confounder, is appropriate when the logistic regression model containing both risk factor and confounder is not fully S-shaped. A more detailed evaluation is needed when the fitted model yields fitted values producing a full S-shaped function within the levels of the risk factor. This is discussed in greater detail in Chapter 4.

The method of adjustment when the variables are all dichotomous, polychotomous, continuous or a mixture of these is identical to that just described for the dichotomous-continuous variable case. For example, suppose that instead of treating age as continuous it was dichotomized using a cutpoint of 45 years. To obtain the age-adjusted effect of group we fit the bivariate model containing the two dichotomous variables and calculate a logit difference at the two levels of group and a common value of the dichotomous variable for age. The procedure is similar for any number and mix of variables. Adjusted odds ratios are obtained by comparing individuals who differ only in the characteristic of interest and have the values of all other variables constant. The adjustment is statistical as it only estimates what might be expected to be observed had the subjects indeed differed only on the particular characteristic being examined, with all other variables having identical distributions within the two levels of outcome.

One point should be kept clearly in mind when interpreting statistically adjusted log odds ratios and odds ratios. The effectiveness of the adjustment is entirely dependent on the adequacy of the assumptions of the model: linearity and constant slope. Departures from these may render the adjustment useless. One such departure, where the relationship is linear but the slopes differ, is called *interaction*. Modeling interactions is discussed in Section 3.6 and again in Chapter 4.

3.6 INTERACTION AND CONFOUNDING

In the last section we saw how the inclusion of additional variables in a model provides a way of statistically adjusting for potential differences in their distributions. The term confounder is used by epidemiologists to describe a covariate that is associated with both the outcome variable of interest and a primary independent variable or risk factor. When both associations are present then the relationship between the risk factor and the outcome variable is said to be *confounded.* The procedure for adjusting for confounding, described in Section 3.5, is appropriate when there is no interaction. In this section we introduce the concept of interaction and show how we can control for its effect in the logistic regression model. In addition, we illustrate with an example how confounding and interaction may affect the estimated coefficients in the model.

Interaction can take many different forms, so we begin by describing the situation when it is absent. Consider a model containing a dichotomous risk factor variable and a continuous covariate, such as in the example discussed in Section 3.5. If the association between the covariate (i.e., age) and the outcome variable is the same within each level of the risk factor (i.e., group), then there is no interaction between the covariate and the risk factor. Graphically, the absence of interaction yields a model with two parallel lines, one for each level of the risk factor variable. In general, the absence of interaction is characterized by a model that contains no second or higher order terms involving two or more variables.

When interaction is present, the association between the risk factor and the outcome variable differs, or depends in some way on the level of the covariate. That is, the covariate modifies the effect of the risk factor. Epidemiologists use the term *effect modifier* to describe a variable that interacts with a risk factor. In the previous example, if the logit is linear in age for the men in group 1, then interaction implies that the logit does not follow a line with the same slope for the second group. In theory, the association in group 2 could be described by almost any model except one with the same slope as the logit for group 1.

The simplest and most commonly used model for including interaction is one in which the logit is also linear in the confounder for the second group, but with a different slope. Alternative models can be formulated which would allow for a relationship that is non-linear between the logit and the variables in the model within each group. In any

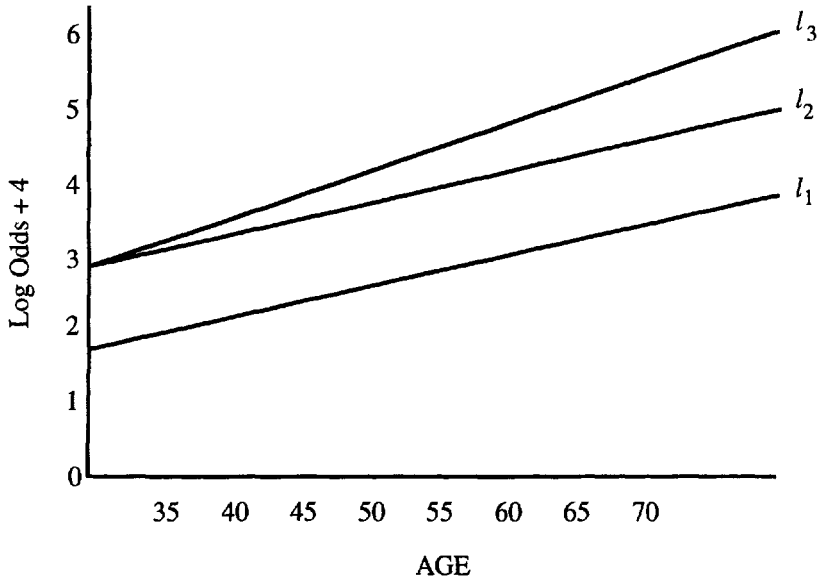

Figure 3.2 Plot of the logits under three different models showing the presence and absence of interaction.

model, interaction is incorporated by the inclusion of appropriate higher order terms.

An important step in the process of modeling a set of data is determining whether there is evidence of interaction in the data. This aspect of modeling is discussed in Chapter 4. In this section we assume that when interaction is present it can be modeled by nonparallel straight lines.

Figure 3.2 presents the graphs of three different logits. In this graph, 4 has been added to each of the logits to make plotting more convenient. The graphs of these logits are used to explain what is meant by interaction. Consider an example where the outcome variable is the presence or absence of CHD, the risk factor is sex, and the covariate is age. Suppose that the line labeled l_1 corresponds to the logit for females as a function of age. Line l_2 represents the logit for males. These two lines are parallel to each other, indicating that the relationship between age and CHD is the same for males and females. In this situation there is no interaction and the log odds ratios for sex (male versus female), controlling for age, is given by the difference between line l_2

Table 3.12 Estimated Logistic Regression Coefficients, Deviance, and the Likelihood Ratio Test Statistic (G) for an Example Showing Evidence of Confounding but No Interaction (n = 400)

Model	Constant	SEX	AGE	SEX×AGE	Deviance	G
1	0.060	1.981			419.816	
2	−3.374	1.356	0.082		407.780	12.036
3	−4.216	4.239	0.103	−0.062	406.392	1.388

and l_1, $l_2 - l_1$. This difference is equal to the vertical distance between the two lines, which is the same for all ages.

Suppose instead that the logit for males is given by the line l_3. This line is steeper than the line l_1, for females, indicating that the relationship between age and CHD among males is different from that among females. When this occurs we say there is an interaction between age and sex. The estimate of the log-odds ratios for sex (males versus females) controlling for age is still given by the vertical distance between the lines, $l_3 - l_1$, but this difference now depends on the age at which the comparison is made. Thus, we cannot estimate the odds ratio for sex without first specifying the age at which the comparison is being made. In other words, age is an effect modifier.

Tables 3.12 and 3.13 present the results of fitting a series of logistic regression models to two different sets of hypothetical data. The variables in each of the data sets are the same: SEX, AGE, and the outcome variable CHD. In addition to the estimated coefficients, the deviance for each model is given. Recall that the change in the deviance may be used to test for the significance of coefficients for variables added to the model. An interaction is added to the model by creating a variable that is equal to the product of the value of the SEX and the value of AGE. Some programs have syntax that automatically creates interaction variables in a statistical model, while others require the user to create them through a data modification step.

Examining the results in Table 3.12 we see that the estimated coefficient for the variable SEX changed from 1.981 in model 1 to 1.356, a 46 percent decrease, when AGE was added in model 2. Hence, there is clear evidence of a confounding effect due to age. When the interaction term "SEX×AGE" is added in model 3 we see that the change in the deviance is only 1.388 which, when compared to the chi-square distribution with 1 degree of freedom, yields a p-value of 0.24, which is clearly not significant. Note that the coefficient for sex changed from

Table 3.13 Estimated Logistic Regression Coefficients, Deviance, and the Likelihood Ratio Test Statistic (G) for an Example Showing Evidence of Confounding and Interaction (n = 400)

Model	Constant	SEX	AGE	SEX×AGE	Deviance	G
1	0.201	2.386			376.712	
2	−6.672	1.274	0.166		338.688	38.024
3	−4.825	−7.838	0.121	0.205	330.654	8.034

1.356 to 4.239. This is not surprising since the inclusion of an interaction term, especially when it involves a continuous variable, usually produces fairly marked changes in the estimated coefficients of dichotomous variables involved in the interaction. Thus, when an interaction term is present in the model we cannot assess confounding via the change in a coefficient. For these data we would prefer to use model 2 that suggests age is a confounder but not an effect modifier.

The results in Table 3:13 show evidence of both confounding and interaction due to age. Comparing model 1 to model 2 we see that the coefficient for sex changes from 2.386 to 1.274, an 87 percent decrease. When the age by sex interaction is added to the model we see that the change in the deviance is 8.034 with a p-value of 0.005. Since the change in the deviance is significant, we prefer model 3 to model 2, and should regard age as both a confounder and an effect modifier. The net result is that any estimate of the odds ratio for sex should be made with reference to a specific age.

Hence, we see that determining whether a covariate, X, is an effect modifier and/or a confounder involves several issues. The plots of the logits shown in Figure 3.2 show us that determining effect modification status involves the parametric structure of the logit, while determination of confounder status involves two things. First the covariate must be associated with the outcome variable. This implies that the logit must have a nonzero slope in the covariate. Second the covariate must be associated with the risk factor. In our example this is characterized by having a difference in the mean age for males and females. However, the association may be more complex than a simple difference in means. The essence is that we have incomparability in our risk factor groups. This incomparability must be accounted for in the model if we are to obtain a correct, unconfounded, estimate of effect for the risk factor.

In practice, one method to check for the confounder status of a covariate is to compare the estimated coefficient for the risk factor variable from models containing and not containing the covariate. Any "clinically important" change in the estimated coefficient for the risk factor suggests that the covariate is a confounder and should be included in the model, regardless of the statistical significance of its estimated coefficient. As noted above, Bachand and Hosmer (1999) show that the change in coefficient method does not always provide evidence that a variable is a confounder and a more detailed evaluation may be required. We return to this point in Chapter 4.

On the other hand, we believe that a covariate is an effect modifier only when the interaction term added to the model is both clinically meaningful and statistically significant. When a covariate is an effect modifier, its status as a confounder is of secondary importance since the estimate of the effect of the risk factor depends on the specific value of the covariate.

The concepts of adjustment, confounding, interaction, and effect modification, may be extended to cover the situations involving any number of variables on any measurement scale(s). The dichotomous-continuous variables example illustrated in this section has the advantage that the results are easily shown graphically. This is not the case with more complicated models. The principles for identification and inclusion of confounder and interaction variables in the model are the same regardless of the number of variables and their measurement scales.

3.7 ESTIMATION OF ODDS RATIOS IN THE PRESENCE OF INTERACTION

In Section 3.6 we showed that when there was interaction between a risk factor and another variable, the estimate of the odds ratio for the risk factor depends on the value of the variable that is interacting with it. In this situation we may not be able to estimate the odds ratio by simply exponentiating an estimated coefficient. One approach that will always yield the correct model-based estimate is to (1) write down the expressions for the logit at the two levels of the risk factor being compared, (2) algebraically simplify the difference between the two logits and compute its value (3) exponentiate the value obtained in step 2.

As a first example, we develop the method for a model containing only two variables and their interaction. In this model, denote the risk

factor as F, the covariate as X and their interaction as $F \times X$. The logit for this model evaluated at $F = f$ and $X = x$ is

$$g(f,x) = \beta_0 + \beta_1 f + \beta_2 x + \beta_3 f \times x \ . \tag{3.9}$$

Assume we want the odds ratio comparing two levels of F, $F = f_1$ versus and $F = f_0$, at $X = x$. Following the three step procedure first we evaluate the expressions for the two logits yielding

$$g(f_1,x) = \beta_0 + \beta_1 f_1 + \beta_2 x + \beta_3 f_1 \times x$$

and

$$g(f_0,x) = \beta_0 + \beta_1 f_0 + \beta_2 x + \beta_3 f_0 \times x \ .$$

Second we compute and simplify their difference to obtain the log-odds ratio yielding

$$\begin{aligned} \ln\left[\mathrm{OR}(F = f_1, F = f_0, X = x)\right] &= g(f_1,x) - g(f_0,x) \\ &= (\beta_0 + \beta_1 f_1 + \beta_2 x + \beta_3 f_1 \times x) \\ &\qquad - (\beta_0 + \beta_1 f_0 + \beta_2 x + \beta_3 f_0 \times x) \\ &= \beta_1 (f_1 - f_0) + \beta_3 x (f_1 - f_0). \end{aligned} \tag{3.10}$$

Third we obtain the odds ratio by exponentiating the difference obtained at step 2 yielding

$$\mathrm{OR} = \exp\left[\beta_1 (f_1 - f_0) + \beta_3 x (f_1 - f_0)\right] \ . \tag{3.11}$$

Note that the expression for the log-odds ratio in (3.10) does not simplify to a single coefficient. Instead, it involves two coefficients, the difference in the values of the risk factor and the interaction variable. The estimator of the log-odds ratio is obtained by replacing the parameters in (3.10) and (3.11) with their estimators.

We obtain the endpoints of the confidence interval estimator using the same approach used for models without interactions. We calculate the endpoints for the confidence interval for the log-odds ratio and then exponentiate the end points. The basic building block of the endpoints is the estimator of the variance of the estimator of the log-

odds ratio in (3.10). Using methods for calculating the variance of a sum we obtain the following estimator,

$$\begin{aligned}\widehat{\mathrm{Var}}\left\{\ln\left[\widehat{\mathrm{OR}}(F=f_1,F=f_0,X=x)\right]\right\} &= (f_1-f_0)^2\times\widehat{\mathrm{Var}}\left(\hat{\beta}_1\right)\\ &+\left[x(f_1-f_0)\right]^2\times\widehat{\mathrm{Var}}\left(\hat{\beta}_3\right)+2x(f_1-f_0)^2\times\widehat{\mathrm{Cov}}\left(\hat{\beta}_1,\hat{\beta}_3\right).\end{aligned} \tag{3.12}$$

Most logistic regression computer packages have the option to provide output showing estimates of the variances and covariances of the estimated parameters in the model. Substitution of these estimates into (3.12) obtains an estimate of the variance of the estimated log-odds ratio. The endpoints of a $100\times(1-\alpha)\%$ confidence interval estimator for the log-odds ratio are:

$$\begin{aligned}&\left[\hat{\beta}_1(f_1-f_0)+\hat{\beta}_3x(f_1-f_0)\right]\\ &\qquad\pm z_{1-\alpha/2}\widehat{\mathrm{SE}}\left\{\ln\left[\widehat{\mathrm{OR}}(F=f_1,F=f_0,X=x)\right]\right\},\end{aligned} \tag{3.13}$$

where the standard error in (3.13) is the positive square root of the variance estimator in (3.12). We obtain the endpoints of the confidence interval estimator for the odds ratio by exponentiating the endpoints in (3.13).

The estimators for the log-odds and its variance simplify in the case when F is a dichotomous risk factor. If we let $f_1=1$ and $f_0=0$ then the estimator of the log-odds ratio is

$$\ln\left[\widehat{\mathrm{OR}}(F=1,F=0,X=x)\right]=\hat{\beta}_1+\hat{\beta}_3x, \tag{3.14}$$

the estimator of the variance is

$$\begin{aligned}&\widehat{\mathrm{Var}}\left\{\ln\left[\widehat{\mathrm{OR}}(F=1,F=0,X=x)\right]\right\}\\ &\qquad=\widehat{\mathrm{Var}}\left(\hat{\beta}_1\right)+x^2\widehat{\mathrm{Var}}\left(\hat{\beta}_3\right)+2x\widehat{\mathrm{Cov}}\left(\hat{\beta}_1,\hat{\beta}_3\right)\end{aligned} \tag{3.15}$$

and the endpoints of the confidence interval are

$$\left(\hat{\beta}_1+\hat{\beta}_3x\right)\pm z_{1-\alpha/2}\widehat{\mathrm{SE}}\left\{\ln\left[\widehat{\mathrm{OR}}(F=1,F=0,X=x)\right]\right\}. \tag{3.16}$$

Table 3.14 Estimated Logistic Regression Coefficients, Deviance, the Likelihood Ratio Test Statistic (G), and the p-value for the Change for Models Containing LWD and AGE from the Low Birthweight Data (n = 189)

Model	Constant	LWD	AGE	LWD×AGE	$\ln[l(\beta)]$	G	p
0	–0.790				–117.34		
1	–1.054	1.054			–113.12	8.44	0.004
2	–0.027	1.010	–0.044		–112.14	1.96	0.160
3	0.774	–1.944	–0.080	0.132	–110.57	3.14	0.076

As an example, we consider a logistic regression model using the low birth weight data described in Section 1.6 containing the variables AGE and a dichotomous variable, LWD, based on the weight of the mother at the last menstrual period. This variable takes on the value 1 if LWT < 110 pounds, and is zero otherwise. The results of fitting a series of logistic regression models are given in Table 3.14.

Using the estimated coefficient for LWD in model 1 we estimate the odds ratio as $\exp(1.054) = 2.87$. The results shown in Table 3.14 indicate that AGE is not a strong confounder, $\Delta\hat{\beta}\% = 4.2$, but it does interact with LWD, $p = 0.076$. Thus, to assess the risk of low weight at the last menstrual period correctly we must include the interaction of this variable with the women's age because the odds ratio is not constant over age.

An effective way to see the presence of interaction is via a graph of the estimated logit under model 3 in Table 3.14. This is shown in Fig ure 3.3. The upper line in Figure 3.3 corresponds to the estimated logit for women with $LWD = 1$ and the lower line is for women with $LWD = 0$. Separate plotting symbols have been used for the two LWD groups. The estimated log-odds ratio for $LWD = 1$ versus $LWD = 0$ at $AGE = x$ from (3.14) is equal to the vertical distance between the two lines at $AGE = x$. We can see in Figure 3.3 that this distance is nearly zero at 15 years and progressively increases. Since the vertical distance is not constant we must choose a few specific ages for estimating the effect of low weight at the last menstrual period. We can see in Figure 3.3 that none of the women in the low weight group, $LWD = 1$, are older than about 33 years. Thus we should restrict our estimates of the effect of low weight to the range of 14 to 33 years. Based on these observations we estimate the effect of low weight at 15, 20, 25 and 30 years of age.

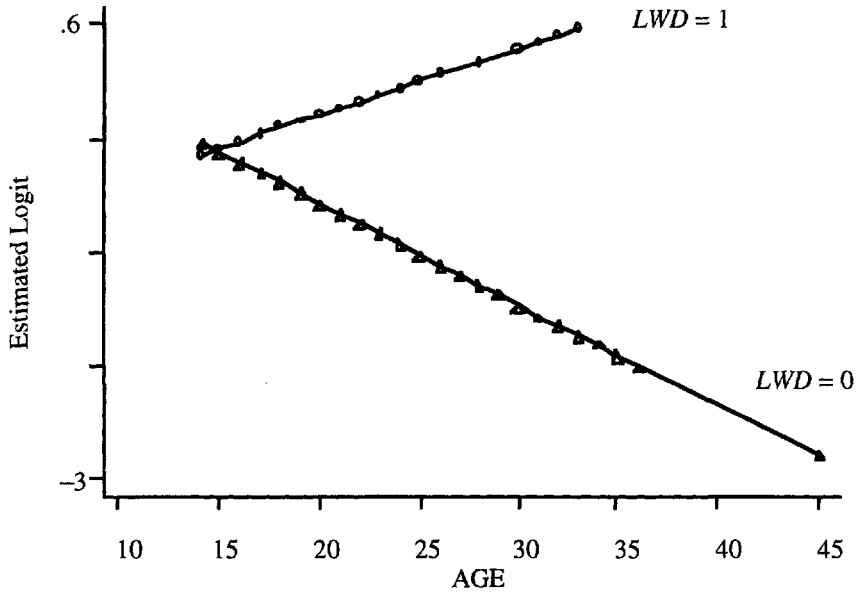

Figure 3.3 Plot of the estimated logit for women with $LWD = 1$ and for women with $LWD = 0$ from Model 3 in Table 3.17.

Using (3.14) and the results for model 3 the estimated log-odds ratio for low weight at the last menstrual period for a women of AGE = a is

$$\ln\left[\hat{\text{OR}}(\text{LWD}=1,\text{LWD}=0,\text{AGE}=a)\right]=-1.944+0.132a. \quad (3.17)$$

In order to obtain the estimated variance we must first obtain the estimated covariance matrix for the estimated parameters. Since this matrix is symmetric most logistic regression software packages print the

Table 3.15 Estimated Covariance Matrix for the Estimated Parameters in Model 3 of Table 3.14

Constant	0.828			
LWD	–0.828	2.975		
AGE	–0.353-02	–0.353-01	0.157-02	
LWD×AGE	–0.353-01	–0.128	–0.157-02	0.573-02
	Constant	LWD	AGE	LWD×AGE

Table 3.16 Estimated Odds Ratios and 95% Confidence Intervals for LWD, Controlling for AGE

Age	15	20	25	30
OR	1.04	2.01	3.90	7.55
95 % CIE	0.29, 3.79	0.91, 4.44	1.71, 8.88	1.95, 29.19

results in the form similar to that shown in Table 3.15.

The estimated variance of the log-odds ratio given (3.16) is obtained from (3.14) and is

$$\hat{\text{var}}\left\{\ln\left[\hat{\text{OR}}(\text{LWD}=1,\text{LWD}=0,\text{AGE}=a)\right]\right\} = 2.975 + a^2 \times 0.0057 + 2 \times a \times (-0.128). \quad (3.19)$$

Values of the estimated odds ratio and 95% CI computed using (3.16) and (3.19) for several ages are given in Table 3.16. The results shown in Table 3.16 demonstrate that the effect of LWD on the odds of having a low birth weight baby increase exponentially with age. The results also show that the increase in risk is significant for low weight women 25 years and older. In particular low weight women of age 30 are estimated to have a risk that is about 7.5 times that of women of the same age who are not low weight. The increase in risk could be as little as two times or as much as 29 times with 95 percent confidence.

3.8 A COMPARISON OF LOGISTIC REGRESSION AND STRATIFIED ANALYSIS FOR 2 × 2 TABLES

Many users of logistic regression, especially those coming from a background in epidemiology, have performed stratified analyses of 2×2 tables to assess interaction and to control confounding. The essential objective of such analyses is to determine whether the odds ratios are constant, or homogeneous, over the strata. If the odds ratios are constant, then a stratified odds ratio estimator such as the Mantel-Haenszel estimator or the weighted logit-based estimator is computed. This same analysis may also be performed using the logistic regression modeling techniques discussed in Sections 3.6 and 3.7. In this section we compare these two approaches. An example from the low birth weight data illustrates the similarities and differences in the two approaches.

Table 3.17 Cross-Classification of Low Birth Weight by Smoking Status

		SMOKE		
		1	0	Total
LOW	1	30	29	59
	0	44	86	130
	Total	74	115	189

Consider an analysis of the risk factor smoking on low birth weight. The crude (or unadjusted) odds ratio computed from the 2×2 table shown in Table 3.17, cross-classifying the outcome variable LOW with SMOKE, is $\hat{\text{OR}} = 2.02$.

Table 3.18 presents these data stratifying by the race of the mother. We can use these tables as the basis for computing either the Mantel-Haenszel estimate or the logit-based estimate of the odds ratio.

The Mantel-Haenszel estimator is a weighted average of the stratum specific odds ratios, $\hat{\text{OR}}_i = (a_i \times d_i)/(b_i \times c_i)$, where a_i, b_i, c_i, and d_i are the observed cell frequencies in the 2×2 table for stratum i. For example, in stratum 1 $a_1 = 19$, $b_1 = 4$, $c_1 = 33$, and $d_1 = 40$ and the total number of subjects is $N_1 = 96$. The Mantel-Haenszel estimator of the odds ratio is defined in this case as follows:

$$\hat{\text{OR}}_{\text{MH}} = \frac{\sum a_i \times d_i / N_i}{\sum b_i \times c_i / N_i}. \tag{3.20}$$

Evaluating (3.20) using the data in Table 3.18 yields the Mantel-Haenszel estimate

$$\hat{\text{OR}}_{\text{MH}} = \frac{13.067}{4.234} = 3.09.$$

The logit-based summary estimator of the odds ratio is a weighted average of the stratum specific log-odds ratios where each weight is the inverse of the variance of the stratum specific log-odds ratio,

$$\hat{\mathrm{OR}}_{\mathrm{L}} = \exp\left[\sum w_i \ln\left(\hat{\mathrm{OR}}_i\right) \Big/ \sum w_i\right]. \tag{3.21}$$

Table 3.19 presents the estimated odds ratio, log-odds ratio, estimate of the variance of the log-odds ratio and the weight, w.

The logit-based estimator based on the data in Table 3.18 is

$$\hat{\mathrm{OR}}_{\mathrm{L}} = \exp(7.109/6.582) = 2.95,$$

Table 3.18 Cross-Classification of Low Birth Weight by Smoking Status Stratified by RACE

White

		SMOKE		
		1	0	Total
LOW	1	19	4	23
	0	33	40	73
	Total	52	44	96

Black

		SMOKE		
		1	0	Total
LOW	1	6	5	11
	0	4	11	15
	Total	10	16	26

Other

		SMOKE		
		1	0	Total
LOW	1	5	20	25
	0	7	35	42
	Total	12	55	67

Table 3.19 Tabulation of the Estimated Odds Ratios, ln(Estimated Odds Ratios), Estimated Variance of the ln(Estimated Odds Ratios), and the Inverse of the Estimated Variance, w, for Smoking Status Within Each Stratum of RACE

	White	Black	Other
$\widehat{\text{OR}}$	5.758	3.300	1.250
$\ln(\widehat{\text{OR}})$	1.751	1.194	0.223
$\hat{\text{var}}[\ln(\widehat{\text{OR}})]$	0.358	0.708	0.421
w	2.794	1.413	2.375

which is slightly smaller than the Mantel-Haenszel estimate. The high fluctuation in the odds ratio across the race strata suggests that there may be either confounding or effect modification due to RACE, or both. In general, the Mantel-Haenszel estimator and the logit based estimator are similar when the data are not too sparse within the strata. One considerable advantage of the Mantel-Haenszel estimator is that it may be computed when some of the cell entries are zero.

It is important to note that these estimators provide a correct estimate of the effect of the risk factor only when the odds ratio is constant across the strata. Thus, a crucial step in the stratified analysis is to assess the validity of this assumption. Statistical tests of this assumption are based on a comparison of the stratum specific estimates to an overall estimate computed under the assumption that the odds ratio is, in fact, constant. The simplest and most easily computed test of the homogeneity of the odds ratios across strata is based on a weighted sum of the squared deviations of the stratum specific log-odds ratios from their weighted mean. This test statistic, in terms of the current notation, is

$$X_H^2 = \sum \left\{ w_i \left[\ln\left(\widehat{\text{OR}}_i\right) - \ln\left(\widehat{\text{OR}}_L\right) \right]^2 \right\} . \qquad (3.22)$$

Under the hypothesis that the odds ratios are constant, X_H^2 has a chi-square distribution with degrees-of-freedom equal to the number of strata minus 1. Thus, we would reject the homogeneity assumption when X_H^2 is large.

Using the data in Table 3.19 we have $X^2_H = 3.017$ which, with 2 degrees-of-freedom, yields a p-value of 0.221. Thus, in spite of the apparent differences in the odds ratios seen in Table 3.19, the logit-based test of homogeneity indicates that they are within sampling variation of each other. It should be noted that the p-value calculated from the chi-square distribution is accurate only when the sample sizes are not too small within each stratum. This condition holds in this example.

Another test that also may be calculated by hand, but not as easily, is discussed in Breslow and Day (1980) and is corrected by Tarone (1985). This test compares the value of a_i to an estimated expected value, $\hat{e}_i$, if the odds ratio is constant. As noted by Breslow (1996) the correct formula for the test statistic is

$$X^2_{BD} = \sum \frac{(a_i - \hat{e}_i)^2}{\hat{v}_i} - \frac{\left[\sum(a_i) - \sum(\hat{e}_i)\right]^2}{\sum(\hat{v}_i)}. \tag{3.23}$$

The quantity $\hat{e}_i$ is obtained as one of the solutions to a quadratic equation given by the following formula

$$\hat{e}_i = \frac{1}{2\left(\hat{\mathrm{OR}} - 1\right)}\left(\begin{array}{l} \hat{\mathrm{OR}}(n_{1i} + m_{1i}) + (n_{0i} - m_{1i}) \pm \\ \left\{\left[\hat{\mathrm{OR}}(n_{1i} + m_{1i}) + (n_{0i} - m_{1i})\right]^2 - \left[4(\hat{\mathrm{OR}} - 1)\,\hat{\mathrm{OR}}\, n_{1i} m_{1i}\right]\right\}^{1/2} \end{array} \right), \tag{3.24}$$

where $n_{1i} = a_i + b_i$, $m_{1i} = a_i + c_i$ and $n_{0i} = c_i + d_i$. The quantity $\hat{\mathrm{OR}}$ in (3.24) is an estimate of the common odds ratio and either $\hat{\mathrm{OR}}_L$ or $\hat{\mathrm{OR}}_{\mathrm{MH}}$ may be used. The quantity $\hat{v}_i$ is an estimate of the variance of a_i computed under the assumption of a common odds ratio and is

$$\hat{v}_i = \left(\frac{1}{\hat{e}_i} + \frac{1}{n_{1i} - \hat{e}_i} + \frac{1}{m_{1i} - \hat{e}_i} + \frac{1}{n_{0i} - m_{1i} + \hat{e}_i}\right)^{-1}. \tag{3.25}$$

If we use the value of the Mantel-Haenszel estimate, $\hat{\mathrm{OR}}_{\mathrm{MH}} = 3.086$ in (3.23) then the resulting values of $\hat{e}$ and $\hat{v}$ are: $\hat{e}_1 = 17.01$, $\hat{v}_1 = 3.56$, $\hat{e}_2 = 5.91$, $\hat{v}_2 = 1.43$, $\hat{e}_3 = 7.16$, and $\hat{v}_3 = 2.33$. The value of the Breslow-Day statistic obtained is $X^2_{BD} = 3.11 - 0.0081 = 3.10$, which is similar to

Table 3.20 Estimated Logistic Regression Coefficients for the Variable SMOKE, Log-Likelihood, the Likelihood Ratio Test Statistic (G), and Resulting p-Value for Estimation of the Stratified Odds Ratio and Assessment of Homogeneity of Odds Ratios Across Strata Defined by RACE

Model	SMOKE	Log-Likelihood	G	df	p
1	0.704	−114.90			
2	1.116	−109.99	9.83	2	0.007
3	1.751	−108.41	3.16	2	0.206

the value of the logit-based test. Some packages, for example SAS, report the value of the first term in (3.23) as the Breslow-Day test

The same analysis may be performed much more easily by fitting three logistic regression models. In model 1 we include only the variable SMOKE. We then add the two design variables for RACE to obtain model 2. For model 3 we add the two RACE×SMOKE interaction terms. The results of fitting these models are shown in Table 3.20. Since we are primarily interested in the estimates of the coefficient for SMOKE, the estimates of the coefficients for RACE and the RACE ×SMOKE interactions are not shown in Table 3.20.

Using the estimated coefficients in Table 3.20 we have the following estimated odds ratios. The crude odds ratio is $\widehat{\text{OR}} = \exp(0.704) = 2.02$. Adjusting for RACE, the stratified estimate is $\widehat{\text{OR}} = \exp(1.116) = 3.05$. This value is the maximum likelihood estimate of the estimated odds ratio, and it is similar in value to both the Mantel-Haenszel estimate, $\widehat{\text{OR}}_{\text{MH}} = 3.086$, and the logit-based estimate, $\widehat{\text{OR}}_{L} = 2.95$. The change in the estimate of the odds ratio from the crude to the adjusted is 2.02 to 3.05, indicating considerable confounding due to RACE.

Assessment of the homogeneity of the odds ratios across the strata is based on the likelihood ratio test of model 2 versus model 3. The value of this statistic from Table 3.20 is $G = 3.156$. This statistic is compared to a chi-square distribution with 2 degrees-of-freedom, since two interaction terms were added to model 2 to obtain model 3. This test statistic is comparable to the ones from the logit-based test, X^2_H, and the Breslow-Day test, X^2_{BD}. If we had used the maximum likelihood estimate of the stratified odds ratio, exp(1.116), in computing the Breslow-Day test, then the resulting statistic would have been equal to

the Pearson chi-square goodness-of-fit test of model 2, since model 3 is the saturated model.

The previously described analysis based on likelihood ratio tests may be used when the data have either been grouped into contingency tables in advance of the analysis, such as those shown in Table 3.17, or have remained in casewise form. When the data have been grouped it is possible to point out other similarities between classical analysis of stratified 2×2 tables and an analysis using logistic regression. Day and Byar (1979) have shown that the 1 degree of freedom Mantel-Haenszel test of the hypothesis that the stratum specific odds ratios are 1 is identical to the Score test for the exposure variable when added to a logistic regression model already containing the stratification variable. This test statistic may be easily obtained from a logistic regression package with the capability to perform Score tests such as the EGRET or SAS packages.

Thus, use of the logistic regression model provides a fast and effective way to obtain a stratified odds ratio estimator and to assess easily the assumption of homogeneity of odds ratios across strata.

3.9 INTERPRETATION OF THE FITTED VALUES

In previous sections in this chapter we discussed use of estimated coefficients to construct estimated odds in a number of settings typically encountered in practice. In our experience this accounts for the vast majority of the use of logistic regression modeling in applied settings. However there are situations where the fitted values from the model are equally, if not more, important. For example, Lemeshow, Teres, Klar, Avrunin, Gehlbach and Rapoport (1993) used logistic regression modeling methods to estimate a patient's probability of hospital mortality after admission to an intensive care unit.. We discussed in Section 1.4 and Section 2.5 the basic methods for computing the fitted values and confidence interval estimates. In this section, we expand on this work and include graphical presentation of fitted values and confidence bands. In addition we discuss prediction of the outcome for a subject not in the estimation sample.

As an example consider the model fit to the low birth weight data shown in Table 2.3. In Section 2.5 we illustrated the computations for a 150 pound white woman. A subject with these values was among the 189 subjects in the data set; thus estimates of the fitted value, logit and standard error of the logit are readily available from standard output.

Suppose instead that we wanted to present a graph illustrating the effect of weight of the mother at the last menstrual period on birth weight

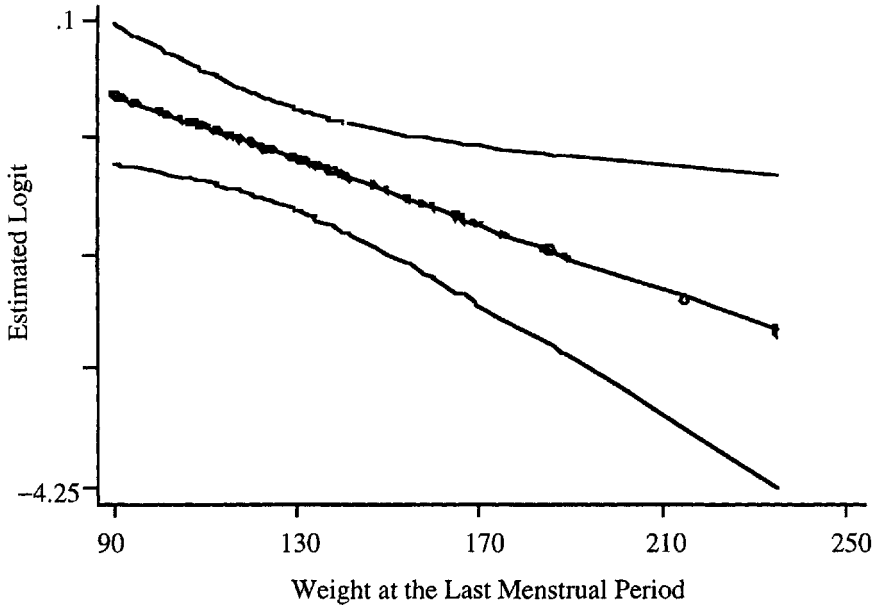

Figure 3.4 Graph of the estimated logit of low weight birth and 95 percent confidence intervals as a function of weight at the last menstrual period for white women.

holding race constant and equal to white. To accomplish this we take advantage of the fact that we can obtain the values of (2.6) and (2.7) for all subjects in the data set used to fit the model from standard logistic regression software. The graph for the estimated logit and its confidence bands is presented in Figure 3.4. The point and interval estimates for the logit are easily transformed to corresponding point and interval estimates for the logistic probability using the fundamental relationship between the two, see (1.19) and (1.21). These are presented in Figure 3.5. Note that we could have presented graphs for any of the three racial groups or for all three racial groups on the same graph. We arbitrarily chose the white mothers in order to keep the graph from getting unnecessarily complicated. The estimates in the figures are plotted at each observed value of LWT for the 96 white mothers. The estimated logit and probability decrease due to the fact that the estimated coefficient for LWT in Table 2.3 is negative. Note that the confidence bands in Figure 3.4 are narrowest near the mean value of LWT, approximately 130 pounds. The width increases in the

same hyperbolic manner seen in similar plots from fitted linear regression models. The same pattern, transformed, can be seen in Figure 3.5.

Each point, and associated confidence interval, in Figure 3.5 is an estimate of the mean of the outcome, low birthweight, among white mothers of the specified value of LWT. Using the results in Section 2.5 at 150 pounds the point and interval estimates are 0.191 and (0.120, 0.289) respectively. The interpretation is that estimated proportion of low weight births among 150 pound white women is 0.191 and it could be as low as 0.12 or as high as 0.289 with 95 percent confidence. We would interpret estimates and confidence intervals at other values of LWT in a similar manner.

Suppose we wanted to use our fitted model to estimate the probability of low birthweight for a population of women not represented in the 189 in the estimation sample. As an example, suppose 150-pound black women. We obtain the value of the estimated logit from (2.6) using the estimated coefficients in Table 2.3 as follows

$$\begin{aligned}\hat{g}(LWT = 150, RACE = Black) &= 0.806 - 0.015 \times 150 + 1.081 \times 1 + 0.481 \times 0 \\ &= -0.363\end{aligned}$$

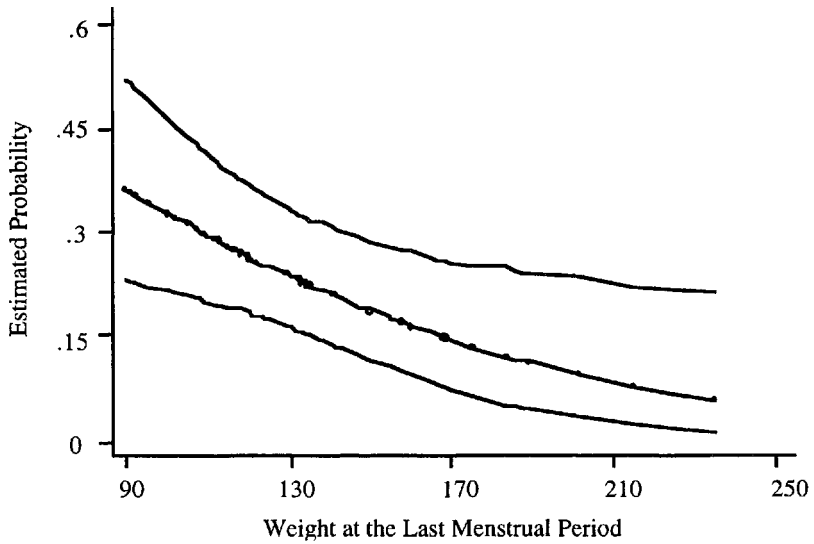

Figure 3.5 Graph of the estimated probability of low weight birth and 95 percent confidence intervals as a function of weight at the last menstrual period for white women.

and the estimated logistic probability is

$$\hat{\pi}(LWT = 150, RACE = Black) = \frac{e^{-0.363}}{1 + e^{-0.363}} = 0.410 .$$

The interpretation is the same as for patterns of data seen in the estimation sample. Namely, the model estimates that the 41 percent of 150 pound black women will have a low birthweight baby.

In order to obtain the confidence interval for this estimate we need to evaluate (2.7) or (2.9) using the covariance matrix in Table 2.4 with the data vector $\mathbf{x}' = (1,150,1,0)$. The resulting standard error from this computation is

$$\hat{\mathrm{SE}}[\hat{g}(LWT = 150, RACE = Black)] = .1725,$$

yielding a 95 percent confidence interval for the probability of (0.331, 0.494). The interpretation of this interval is that the proportion of 150 pound black women who give birth to a low weight baby could be as little as 0.331 or as high as 0.494 (with 95 percent confidence).

As is the case with any regression model we must take care not to extend model-based inferences out of the observed range of the data. The range of weight at the last menstrual period among the 26 black mothers is 98 to 241 pounds. We note that 150 pounds is well within this range. It is also important to keep in mind that any estimate is only as good as the model it is based on. In this section we did not attend to many of the important model building details that are discussed in Chapter 4. We have implicitly assumed that these steps have been performed.

EXERCISES

1. Consider the ICU data described in Section 1.6.1 and use as the outcome variable vital status (STA) and CPR prior to ICU admission (CPR) as a covariate.

   (a) Demonstrate that the value of the log-odds ratio obtained from the cross-classification of STA by CPR is identical to the estimated slope coefficient from the logistic regression of STA on CPR. Verify that the estimated standard error of the estimated slope coefficient for CPR obtained from the logistic regression package is identical to the square root of the sum of the inverse of the cell fre-

quencies from the cross-classification of STA by CPR. Use either set of computations to obtain 95% CI for the odds ratio. What aspect concerning the coding of the variable CPR makes the calculations for the two methods equivalent?

(b) For purposes of illustration, use a data transformation statement to recode, for this problem only, the variable CPR as follows: 4 = no and 2 = yes. Perform the logistic regression of STA on CPR (recoded). Demonstrate how the calculation of the logit difference of CPR = yes versus CPR = no is equivalent to the value of the log-odds ratio obtained in problem 1(a). Use the results from the logistic regression to obtain the 95% CI for the odds ratio and verify that they are the same limits as obtained in Exercise 1(a).

(c) Consider the ICU data and use as the outcome variable vital status (STA) and race (RACE) as a covariate. Prepare a table showing the coding of the two design variables for RACE using the value RACE = 1, white, as the reference group. Show that the estimated log-odds ratios obtained from the cross-classification of STA by RACE, using RACE = 1 as the reference group, are identical to estimated slope coefficients for the two design variables from the logistic regression of STA on RACE. Verify that the estimated standard errors of the estimated slope coefficients for the two design variables for RACE are identical to the square root of the sum of the inverse of the cell frequencies from the cross-classification of STA by RACE used to calculate the odds ratio. Use either set of computations to compute the 95% CI for the odds ratios.

(d) Create design variables for RACE using the method typically employed in ANOVA. Perform the logistic regression of STA on RACE. Show by calculation that the estimated logit differences of RACE = 2 versus RACE = 1 and RACE = 3 versus RACE = 1 are equivalent to the values of the log-odds ratio obtained in problem 1(c). Use the results of the logistic regression to obtain the 95% CI for the odds ratios and verify that they are the same limits as obtained in Exercise 1(c). Note that the estimated covariance matrix for the estimated coefficients is needed to obtain the estimated variances of the logit differences.

(e) Consider the variable AGE in the ICU data set. Prepare a table showing the coding of three design variables based on the empirical quartiles of AGE using the first quartile as the reference group. Fit the logistic regression of STA on AGE as recoded into these design variables and plot the three estimated slope coefficients versus the midpoint of the respective age quartile. Plot as a fourth

point a value of zero at the midpoint of the first quartile of age. Does this plot suggest that the logit is linear in age?

(f) Consider the logistic regression of STA on CRN and AGE. Consider CRN to be the risk factor and show that AGE is a confounder of the association of CRN with STA. Addition of the interaction of AGE by CRN presents an interesting modeling dilemma. Examine the main effects only and interaction models graphically. Using the graphical results and any significance tests you feel are needed, select the best model (main effects or interaction) and justify your choice. Estimate relevant odds ratios. Repeat this analysis of confounding and interaction for a model that includes CPR as the risk factor and AGE as the potential confounding variable.

(g) Consider an analysis for confounding and interaction for the model with STA as the outcome, CAN as the risk factor, and TYP as the potential confounding variable. Perform this analysis using logistic regression modeling and Mantel-Haenszel analysis. Compare the results of the two approaches.

2. Use the data from the Prostatic Cancer Study described in Section 1.6.3 to answer the following questions:

(a) By fitting a series of logistic regression models show that RACE is not a confounder of the PSA CAPSULE odds ratio but is an effect modifier (at the 10 percent level).

(b) Graph the estimated logits from the interactions model versus PSA and interpret the two lines that appear on the graph. Use the graph to illustrate the log-odds of Black versus White for a subject with PSA = 7. Use the graph to illustrate the log-odds for a 5-unit increase in PSA for Whites and for Blacks.

(c) Estimate the point and 95 percent confidence interval estimates of the odds ratios corresponding to each of the log-odds illustrated in problem 2(b). Add the 95 percent confidence bands to the graph of the estimated logits from the interactions model in Exercise 2(b). Transform the lines and bands in this plot to obtain a plot of the estimated probability with its 95 percent confidence bands. Use the graph to estimate, point and interval, the probability of penetration for both a White and Black with PSA = 7. Interpret the two point and interval estimates.