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 statistical models this involves the estimated coefficients for the independent variables in the model. In most instances, the intercept coefficient is of little interest. 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

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

McCullagh and Nelder (1989), or Dobson (2002)]. In the case of a linear regression model, the link function is the identity function as 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, the linear regression model at x is $y(x) = \beta_0 + \beta_1 x$. It follows that the slope coefficient is $\beta_1 = y(x + 1) - y(x)$. In this case, the interpretation of the slope coefficient is that it is the change in the outcome variable corresponding to a one-unit change in the independent variable. For example, in a regression of weight on height of male adolescents if 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 is 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 values of the logit function. 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 univariable logistic regression model for each of the possible measurement scales of the independent variable. We discuss interpretation of the coefficients from multivariable models and the probabilities from a fitted logistic model. We also compare the results of a logistic regression analysis to a stratified contingency table analysis that is common in epidemiological research. We conclude the chapter with a discussion of the construction, use and interpretation of the propensity score.

3.2 DICHOTOMOUS INDEPENDENT VARIABLE

We begin by discussing the interpretation of logistic regression coefficients in the situation where the independent variable is nominal scaled 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 0 or 1. The difference in the logit for a subject with $x = 1$ and $x = 0$ is

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

The algebra shown in this equation is rather straightforward. The rationale for presenting it in this level of detail is to emphasize that four steps are required to obtain the correct expression of the coefficient(s) and hence, the correct interpretation of the coefficient(s). In some settings, like the current one, these steps are quite straightforward, but in the examples in Section 3.3 they are not so obvious.

The first three of the four steps are: (1) define the two values of the covariate to be compared (e.g., $x = 1$ and $x = 0$); (2) substitute these two values into the equation for the logit [e.g., $g(1)$ and $g(0)$], and (3) calculate the difference in the two equations [e.g., $g(1) - g(0)$]. As shown, for a dichotomous covariate coded 0 and 1 the result at the end of step 3 is equal to β_1. Thus, the slope coefficient, or logit difference, is the difference between the log of the odds when $x = 1$ and the log of the odds when $x = 0$. The practical problem is that change on the scale of the log-odds is hard to explain and it may not be especially meaningful to a subject-matter audience. In order to provide a more meaningful interpretation we need to introduce the *odds ratio* as a measure of association.

The possible values of the logistic probabilities from a model containing a single dichotomous covariate coded 0 and 1 are displayed in the 2×2 table, shown in Table 3.1. The *odds* of the outcome being present among individuals with $x = 1$ is $\pi(1)/[1 - \pi(1)]$. Similarly, the odds of the outcome being present among individuals with $x = 0$ is $\pi(0)/[1 - \pi(0)]$. The *odds ratio*, denoted OR, is the ratio of the odds for $x = 1$ to the odds for $x = 0$, and is given by the equation

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

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

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

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

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

This illustrates the fourth step in interpreting the effect of a covariate, namely exponentiate the logit difference computed in step 3 to obtain an odds ratio.

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

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

The odds ratio is widely used as a measure of association as it approximates how much more likely or unlikely (in terms of odds) it is for the outcome to be present among those subjects with $x = 1$ as compared to those subjects with $x = 0$. For example, if the outcome, Y, denotes the presence or absence of lung cancer and if X denotes whether the subject is a smoker, then an OR = 2 is interpreted to mean that the odds of lung cancer among smokers is two times greater than the odds of lung cancer among the nonsmokers in this study population. As another example, suppose that the outcome, Y, is the presence or absence of heart disease and X denotes whether or not the person engages in regular strenuous physical exercise. If the odds ratio is OR = 0.5, then the odds of heart disease among those subjects who exercise is one-half the odds of heart disease for those subjects who do not exercise in the study population. This simple relationship between the coefficient and the odds ratio is the fundamental reason logistic regression has proven to be such a powerful analytic research tool.

In certain settings, the odds ratio can approximate another measure of association called the relative risk, which is the ratio of the two outcome probabilities, RR = $\pi(1)/\pi(0)$. It follows from equation (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 $x = 0$ and $x = 1$, often referred to in medical/epidemiological research as the *rare disease assumption*.

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), Rothman et al. (2008), Aschengrau and Seage (2008), Lilienfeld and Stolley (1994), and Oleckno (2008).

An example from the GLOW study described in Section 1.6.3 and used in Chapter 2 may help clarify how the odds ratio is estimated from the results of a fitted logistic regression model and from a 2 × 2 contingency table. To review, the outcome variable is having a fracture (FRACTURE) in the first year of follow-up. Here we use having had a fracture between the age of 45 and enrollment in the study (PRIORFRAC) as the dichotomous independent variable. The result of cross-classifying fracture during follow-up by prior fracture is presented in Table 3.2.

The frequencies in Table 3.2 tell us that there were 52 subjects with values $(x = 1, y = 1)$, 73 with $(x = 0, y = 1)$, 74 with $(x = 1, y = 0)$, and 301 with

Table 3.2 Cross-Classification of Prior Fracture and Fracture During Follow-Up in the GLOW Study, $n = 500$

Fracture During Follow-Up (y)	Prior Fracture (x)		
	Yes (1)	No (0)	Total
Present (1)	52	73	125
Absent (0)	74	301	375
Total	126	374	500

Table 3.3 Results of Fitting the Logistic Regression Model of Fracture (FRACTURE) on Prior Fracture (PRIORFRAC) Using the Data in Table 3.2

Variable	Coeff.	Std. Err.	z	p	95% CI
PRIORFRAC	1.064	0.2231	4.77	<0.001	0.627, 1.501
Constant	−1.417	0.1305	−10.86	<0.001	−1.672, −1.161

Log-likelihood = −270.03397.

($x = 0, y = 0$). The results of fitting a logistic regression model containing the dichotomous covariate PRIORFRAC are shown in Table 3.3.

The estimate of the odds ratio using equation (3.2) and the estimated coefficient for PRIORFRAC in Table 3.3 is $\widehat{\text{OR}} = e^{1.064} = 2.9$. Readers who have had some previous experience with the odds ratio undoubtedly wonder why we used a logistic regression package to estimate the odds ratio, when we easily could have computed it directly as the cross-product ratio from the frequencies in Table 3.2, namely,

$$\widehat{\text{OR}} = \frac{52 \times 301}{74 \times 73} = 2.897.$$

The tremendous advantage of using logistic regression will surface when additional independent variables are included in the logistic regression model.

Thus, we see that the slope coefficient from the fitted logistic regression model is $\hat{\beta}_1 = \ln[(52 \times 301)/(74 \times 73)] = 1.0638$. This emphasizes the fact that logistic regression is, even in the simplest possible case, a regression analysis. The fact that the data may be presented in terms of a contingency table just aids in the interpretation of the estimated coefficients as the log of the odds ratio.

Along with the point estimate of a parameter, it is always a good idea to use a confidence interval estimate to provide additional information about the parameter value. In the case of the odds ratio from a 2 × 2 table (corresponding to a fitted logistic regression model with a single dichotomous covariate) there is an extensive literature focused on the problem of confidence interval estimation for the odds ratio 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). Breslow and Day (1980), Kleinbaum et al. (1982),

and Rothman et al. (2008) discuss inference with small samples. We discuss the small sample setting in Section 10.3.

As we noted earlier, the odds ratio is usually the parameter of interest derived from a fitted logistic regression due to its ease of interpretation. However, its estimator, $\widehat{\text{OR}}$, tends to have a distribution that is highly skewed to the right. This is due to the fact that its range is between 0 and ∞, with the null value equaling 1. In theory, for extremely large sample sizes, the distribution of $\widehat{\text{OR}}$ would be 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. We obtain a $100 \times (1 - \alpha)\%$ confidence interval estimator for the odds ratio by first calculating the endpoints of a confidence interval estimator for the log-odds ratio (i.e., β_1) and then exponentiating the endpoints of this interval. In general, the endpoints are given by the expression

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

As an example, consider the estimation of the odds ratio for the dichotomized variable PRIORFRAC. Using the results in Table 3.3 the point estimate is $\widehat{\text{OR}} = 2.9$ and the 95% confidence interval is

$$\exp(1.064 \pm 1.96 \times 0.2231) = (1.87, 4.49).$$

This interval is typical of many confidence intervals for odds ratios when the point estimate exceeds 1, in that it is skewed to the right from the point estimate. This confidence interval suggests that the odds of a fracture during follow-up among women with a prior fracture could be as little as 1.9 times or much as 4.5 times the odds for women without a prior fracture, at the 95% level of confidence.

We discussed the profile likelihood confidence interval estimator for a logistic regression coefficient in Section 1.4. The resulting profile likelihood confidence interval for the odds ratio in this example is nearly identical to the Wald based interval given earlier and thus is not presented. There is an exercise at the end of the chapter where this is not the case.

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. The user must be aware that these automatically reported quantities provide estimates of odds ratios of interest in only a few special cases (e.g., a dichotomous variable coded 0 and 1 that is not involved in any interactions with other variables), a point we return to in the next section. One major goal of this chapter is to show, using the four steps noted earlier, that one may obtain point and confidence interval estimates of odds ratios, regardless of the complexity of the fitted model.

Before concluding the dichotomous variable case, it is important to consider the effect that coding has on computing the estimator of odds ratios. In the previous discussion we noted that the estimator is $\widehat{\text{OR}} = \exp(\hat{\beta}_1)$ and that this is correct as long as one codes the independent variable as 0 or 1 (or any two values

that differ by one). Any other coding requires that one calculate the value of the logit difference for the specific coding used and then exponentiate this difference, essentially following the four steps, not just blindly exponentiating the estimator of the coefficient.

We illustrate the setting of alternate coding in detail, as it helps emphasize the four steps in the general method for computing estimators of odds ratios from a fitted logistic regression model. Suppose that our dichotomous covariate is coded using values a and b and that, at Step 1, we would like to estimate the odds ratio for the covariate at level a versus b. Next, at Step 2, we substitute the two values of the covariate into the equation for the logit to obtain $\hat{g}(a) = \hat{\beta}_0 + \hat{\beta}_1 a$ and $\hat{g}(b) = \hat{\beta}_0 + \hat{\beta}_1 b$. For Step 3, we compute the difference in the two equations and algebraically simplify to obtain the expression for the log-odds as

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

At Step 4 we exponentiate the equation obtained in Step 3, shown in this case in equation (3.3), to obtain our estimator of the odds ratio, namely

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

In equations (3.3) and (3.4) the notation $\widehat{\mathrm{OR}}(a,b)$ denotes the specific odds ratio

$$\widehat{\mathrm{OR}}(a,b) = \frac{\dfrac{\hat{\pi}(x=a)}{[1-\hat{\pi}(x=a)]}}{\dfrac{\hat{\pi}(x=b)}{[1-\hat{\pi}(x=b)]}}. \tag{3.5}$$

In the usual case when $a = 1$ and $b = 0$ we suppress a and b and simply use $\widehat{\mathrm{OR}}$.

Some software packages offer a choice of methods for coding design variables. The "0–1 coding" is the one most often used and is referred to as *reference cell* coding. The reference cell method typically assigns the value of 0 to the lower code for x and 1 to the higher code. For example, if gender 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 gender 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 gender discussed earlier is shown in Table 3.4. 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 results of the general

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

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

four-step method that results in equations (3.3) and (3.4),

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

and the estimated odds ratio is $\widehat{\text{OR}}(\text{female,male}) = \exp(2\hat{\beta}_1)$. 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 example using deviation from means coding, the estimated standard error needed for confidence interval estimation is $\widehat{\text{SE}}(2\hat{\beta}_1) = 2\widehat{\text{SE}}(\hat{\beta}_1)$. Thus the endpoints of the confidence interval are

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

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

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

where $|a-b|$ is the absolute value of $(a-b)$. (This is necessary because a might be less than b.) As we have control of how we code our dichotomous variables, we recommend that, when interest focuses on the odds ratio, they be coded as 0 or 1 for analysis purposes.

In summary, for a dichotomous variable the parameter of interest in most, if not all, applied settings is the odds ratio. An estimate of this parameter may be obtained from a fitted logistic regression model by exponentiating the estimated coefficient. In a setting where the coding is not 0 or 1, the estimate may be found by simply following the four steps described in this section. The 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 shown in Section 3.2 for a dichotomous variable. In the GLOW study the covariate self-reported risk is coded at three levels (less, same, and more). The cross tabulation of it with fracture during follow-up (FRACTURE) is shown in Table 3.5. In addition we show the estimated odds ratio, its 95% confidence interval and log-odds ratio for same and more versus less risk. The extension to a situation where the variable has more than three levels is not conceptually different so all the examples in this section use $k = 3$.

At the bottom of Table 3.5, the odds ratio is given for the groups "same risk" and "more risk," as compared to the reference group, "less risk." For example, for the "same risk" group the estimated odds ratio is $\widehat{\text{OR}}(\text{Same, Less}) = (48 \times 139)/(28 \times 138) = 1.73$. The log of each odds ratio is given in the last row of Table 3.5. The example in this table is typical of what is found in the literature presenting univariable results for a nominal scaled variable. Note that the reference group is indicated by a value of 1 for the odds ratio. These same estimates and confidence intervals for the odds ratio are also easily 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 0 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 statistical software packages. However, not all packages use the lowest code as the referent group. In particular, the SPSS [SPSS for Windows, Release 20.0 (2012)] package's default coding is to use the highest code as the referent value.

Table 3.5 Cross-Classification of Fracture During Follow-Up (FRACTURE) by Self-Reported Rate of Risk (RATERISK) from the GLOW Study, $n = 500$

	RATERISK			
FRACTURE	Less	Same	More	Total
Yes	28	48	49	125
No	139	138	98	375
Total	167	186	147	500
Odds Ratio	1	1.73	2.48	
95% CI		(1.02, 2.91)	(1.46, 4.22)	
$\ln(\widehat{\text{OR}})$	0.0	0.55	0.91	

Table 3.6 Specification of the Design Variables for RATERISK Using Reference Cell Coding with Less as the Reference Group

RATERISK (Code)	RATERISK2	RATERISK3
Less (1)	0	0
Same (2)	1	0
More (3)	0	1

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	95% CI
RATERISK2	0.546	0.2664	2.05	0.040	0.024, 1.068
RATERISK3	0.909	0.2711	3.35	0.001	0.378, 1.441
Constant	−1.602	0.2071	−7.74	<0.001	−2.008, −1.196

Log-likelihood = −275.28917

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.

When we compare the estimated coefficients in Table 3.7 to the log-odds ratios in Table 3.5 we find that

$$\ln[\widehat{\text{OR}}(\text{Same, Less})] = \hat{\beta}_1 = 0.546,$$

and

$$\ln[\widehat{\text{OR}}(\text{More, Less})] = \hat{\beta}_2 = 0.909.$$

Did this happen by chance? We can check this by using the first three of the four-step procedure, described in Section 3.2, as follows:

Step 1: We want to compare levels Same to Less;

Step 2: The logit for Same is

$$\hat{g}(\text{Same}) = \hat{\beta}_0 + \hat{\beta}_1 \times (\text{RATERISK2} = 1) + \hat{\beta}_2 \times (\text{RATERISK3} = 0),$$

and the logit for Less is

$$\hat{g}(\text{Less}) = \hat{\beta}_0 + \hat{\beta}_1 \times (\text{RATERISK2} = 0) + \hat{\beta}_2 \times (\text{RATERISK3} = 0);$$

Step 3: The logit difference is

$$\begin{aligned}\ln[\widehat{\text{OR}}(\text{Same, Less})] &= \hat{g}(\text{Same}) - \hat{g}(\text{Less}) \\ &= [\hat{\beta}_0 + \hat{\beta}_1 \times 1 + \hat{\beta}_2 \times 0] - [\hat{\beta}_0 + \hat{\beta}_1 \times 0 + \hat{\beta}_2 \times 0] \\ &= \hat{\beta}_1.\end{aligned}$$

Similar calculations demonstrate that the estimated coefficient for RATERISK3 from the logistic regression in Table 3.7 is equal to the log-odds ratio computed from the data in Table 3.5.

A comment about the estimated standard errors may be helpful at this point. In the univariable 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 RATERISK2 is

$$\widehat{\mathrm{SE}}(\hat{\beta}_1) = \left[\frac{1}{48} + \frac{1}{139} + \frac{1}{28} + \frac{1}{138}\right]^{0.5} = 0.2664.$$

Confidence limits for the 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 $100(1-\alpha)\%$ confidence interval for the jth coefficient, β_j, are of the form

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

These are shown in the right most column of Table 3.7. The corresponding limits for the odds ratio, obtained by exponentiating these limits, are as follows:

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

The confidence limits given in Table 3.5 in the row beneath the estimated odds ratios are obtained using equation (3.6) with the estimated coefficients and standard errors in Table 3.7 for $j = 1, 2$ 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 odds 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 an 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 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 RATERISK 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 three categories of RATERISK. These are:

$$\hat{g}_1 = \ln\left(\frac{\frac{28}{167}}{\frac{139}{167}}\right) = \ln\left(\frac{28}{139}\right) = -1.602,$$

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

	Design Variables	
Rate Risk (Code)	RATERISK2D	RATERISK3D
Less (1)	−1	−1
Same (2)	1	0
More (3)	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	95% CI
RATERISK2D	0.061	0.1437	0.43	0.671	−0.221, 0.343
RATERISK3D	0.424	0.1466	2.89	0.004	0.137, 0.711
Constant	−1.117	0.1062	−10.51	<0.001	−1.325, −0.909

Log-likelihood = −275.28917

$$\hat{g}_2 = \ln\left(\frac{\frac{48}{186}}{\frac{138}{186}}\right) = \ln\left(\frac{48}{138}\right) = -1.056,$$

$$\hat{g}_3 = \ln\left(\frac{\frac{49}{147}}{\frac{98}{147}}\right) = \ln\left(\frac{49}{98}\right) = -1.056,$$

and the average of these three logits

$$\overline{g} = \sum_{i=1}^{3} \frac{\hat{g}_i}{3} = -1.117.$$

The estimated coefficient for design variable RATERISK2D in Table 3.9 is $\hat{g}_2 - \overline{g} = (-1.056) - (-1.117) = 0.061$ and for RATERISK3D it is $\hat{g}_3 - \overline{g} = (-0.693) - (-1.117) = 0.424$. The general expression for the estimated coefficient for the jth design variable using deviation from means coding is $\hat{g}_j - \overline{g}$.

The interpretation of the estimated coefficients from deviation from means coding is not as easy or clear as when reference cell coding 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 RATERISK2D in Table 3.9 we have

$$\exp(0.061) = \exp(\hat{g}_2 - \overline{g})$$

$$= \frac{\exp(\hat{g}_2)}{\exp\left(\sum \hat{g}_j/3\right)}$$
$$= \frac{(48/138)}{[(28/139) \times (49/138) \times (49/98)]^{0.333}}$$
$$= 1.06.$$

This number, 1.06, 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 the geometric mean odds. The interpretation of this value depends on whether the geometric mean odds is at all meaningful in the context of the study.

The estimated coefficients obtained using deviation from means coding can 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 how application of the four-step method can always yield the odds ratio of interest.

Step 1: Suppose we want to estimate the odds ratio of RATERISK = 2 (Same) versus RATERISK = 1 (Less).

Step 2: Using the coding for design variables given in Table 3.8 the logit at RATERISK = 2 is

$$\hat{g}(\text{RATERISK} = 2) = \hat{\beta}_0 + \hat{\beta}_1 \times (\text{RATERISK2}D = 1) + \hat{\beta}_2 \times (\text{RATERISK3}D = 0),$$

and the logit at RATERISK = 1 is

$$\hat{g}(\text{RATERISK=1}) = \hat{\beta}_0 + \hat{\beta}_1 \times (\text{RATERISK2}D = -1) + \hat{\beta}_2 \times (\text{RATERISK3}D = -1).$$

Step 3: The difference between the two logit functions is

$$\begin{aligned}\hat{g}(\text{RATERISK} = 2) - \hat{g}(\text{RATERISK} = 1) = {} & \\ & [\hat{\beta}_0 + \hat{\beta}_1 \times (\text{RATERISK2}D = 1) + \hat{\beta}_2 \times (\text{RATERISK3}D = 0)] \\ & - [\hat{\beta}_0 + \hat{\beta}_1 \times (\text{RATERISK2}D = -1) + \hat{\beta}_2 \times (\text{RATERISK3}D = -1)] \\ & = 2\hat{\beta}_1 + \hat{\beta}_2. \end{aligned} \tag{3.7}$$

Step 4: The estimator of the odds ratio is obtained as the exponentiation of the logit difference calculated in Step 3 and is

$$\widehat{\text{OR}}(\text{Same, Less}) = e^{2\hat{\beta}_1 + \hat{\beta}_2}.$$

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

$$\begin{aligned}\widehat{\text{Var}}[\hat{g}(\text{RATERISK} = 2) - \hat{g}(\text{RATERISK} = 1)] \\ = 4 \times \widehat{\text{Var}}(\hat{\beta}_1) + \widehat{\text{Var}}(\hat{\beta}_2) + 4 \times \widehat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_2). \end{aligned} \tag{3.8}$$

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

$$\begin{aligned}\hat{g}(\text{RATERISK} = 2) - \hat{g}(\text{RATERISK} = 1) &= 2 \times 0.061 + 0.424 \\ &= 0.546.\end{aligned}$$

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

$$\begin{aligned}\widehat{\text{Var}}[\hat{g}(\text{RATERISK} = 2) - \hat{g}(\text{RATERISK} = 1)] \\ = 4 \times 0.02065 + 0.02149 - 4 \times 0.00828 = 0.07097,\end{aligned}$$

and the estimated standard error is

$$\widehat{\text{SE}}[\hat{g}(\text{RATERISK} = 2) - \hat{g}(\text{RATERISK} = 1)] = 0.2664.$$

We note that the values of the estimated logit difference (i.e., the log-odds ratio), 0.546, and the estimated standard error, 0.2664, are identical to the values of the estimated coefficient and standard error for RATERISK2D in Table 3.7. This is expected, as the design variables used to obtain the estimated coefficients in Table 3.7 were formulated specifically to yield the log-odds ratio of Same versus Less.

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. However, if the objective is to flag (through the Wald tests) which of the subgroups differ from the average, the deviation from means strategy can be extremely effective.

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, has a direct relationship to the odds ratio, and thus is the one used 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. We note that this linearity assumption is key and methods for examining 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 $g(x) = \beta_0 + \beta_1 x$. Application of the four steps to obtain the estimator of the odds ratio yields the following: (1) suppose that we are interested in the odds ratio for a one-unit increment in the covariate, i.e., $x + 1$ versus x; (2) it follows from the equation for the logit at x that the logit at $x + 1$ is $g(x + 1) = \beta_0 + \beta_1(x + 1)$; (3) hence the estimator of the logit difference is

$$\hat{g}(x + 1) - \hat{g}(x) = \hat{\beta}_1;$$

and (4) the estimator of odds ratio is $\widehat{\text{OR}} = \exp(\hat{\beta}_1)$. This estimator has exactly the same form as the estimator in equation (3.2) for a dichotomous covariate. The problem is that a value of "1" is not likely to be clinically interesting for a continuous covariate. For example, a 1-year increase in age or a 1-pound increase in body weight for adults is probably too small to be considered an important change. A change of 10 years or 10 pounds might be more interesting. On the other hand, if the range of a covariate is only from 0 to 1, then a change of 1 is too large and a change of 0.01 or 0.05 is more realistic. Hence, to provide a useful interpretation for continuous covariates we need to develop a method for point and interval estimation of the odds ratio for an arbitrary change of "c" units in the covariate.

Following the first three steps, we find that the estimator of the log-odds ratio for a change of c units in x is $\hat{g}(x + c) - \hat{g}(x) = c\hat{\beta}_1$ and (5) the estimator odds ratio is $\widehat{\text{OR}}(x + c, x) = \exp(c\hat{\beta}_1)$, more concisely denoted as $\widehat{\text{OR}}(c)$. The estimator of the standard error of $c\hat{\beta}_1$ is $\widehat{\text{SE}}(c\hat{\beta}_1) = |c|\widehat{\text{SE}}(\hat{\beta}_1)$, where "$|c|$" denotes the absolute value of c. We need to use the absolute value as c could be negative. Hence, the endpoints of the $100(1 - \alpha)\%$ confidence interval estimate are

$$\exp[c\hat{\beta}_1 \pm z_{1-\alpha/2}|c|\widehat{\text{SE}}(\hat{\beta}_1)].$$

As 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 indication of how the odds of the outcome change with the variable in question. For most continuous covariates changes in multiples of 2, 5, or 10 may be most meaningful and easily understood.

As an example, we show the results in Table 1.3 of a logistic regression of AGE on CHD status using the data in Table 1.1. The estimated logit is $\hat{g}(\text{AGE}) = -5.310 + 0.111 \times \text{AGE}$. The estimated odds ratio for an increase of 10 years in age is $\widehat{\text{OR}}(10) = \exp(10 \times 0.111) = 3.03$. Thus, for every increase of 10 years in age, the odds of CHD being present is estimated to increase 3.03 times. The

validity of this statement is questionable, because the increase in the odds of CHD for a 40-year-old compared to a 30-year-old may be quite different from the odds for a 60-year-old compared to a 50-year-old. This is the unavoidable dilemma when a continuous covariate is modeled linearly in the logit and motivates the importance of examining the linearity assumption for continuous covariates. As already noted, we consider this in detail in Chapter 4. The endpoints of a 95% confidence interval for this odds ratio are

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

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

3.5 MULTIVARIABLE MODELS

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 univariable models, although useful for a preliminary analysis, rarely provides an adequate or complete analysis of the data in a study because the independent variables are usually associated with one another and may have different distributions within levels of the outcome variable. Thus, one generally uses 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 in the model. 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 the other variables in the model.

Another important aspect of multivariable modeling is to assess to what extent, if at all, the estimate of the log-odds of one independent variable changes, depending on the value of another independent variable. When the odds ratio for one variable is not constant over the levels of another variable, the two variables are said to have a *statistical interaction*. In some applied disciplines statistical interaction is referred to as *effect modification*. This terminology is used to describe the fact that the log-odds of one variable is modified or changed by values of the other variable. In this section we consider, in considerable detail, the concepts of statistical adjustment and interaction and illustrate estimation of odds ratios under each case with examples.

A full understanding of estimating the log-odds or 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 in the model*. In some fields variables that are used to adjust the effects of others are called *confounders* and adjustment for them is called *controlling for confounding*. We begin by examining statistical adjustment in the context of the usual linear regression model, and then extend the concept to the logistic regression model.

To begin, we consider a multivariable model that contains two independent variables: one dichotomous and one continuous, but primary interest is focused on estimating the effect of the dichotomous variable on the outcome variable. This situation is frequently encountered in epidemiological and medical research when an exposure to a risk factor is recorded as being either present or absent, and we wish to adjust for a continuous 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 univariable analysis of group comparing the mean weight of the two groups would suffice. This analysis would provide us with a correct estimate of the difference in the mean weight of the two groups. However, if one group was, on average, much younger than the other group, then a comparison of the two groups would be meaningless, because a portion of any difference observed would be due to the differences in mean age. It would not be possible to determine the effect of group without first eliminating the discrepancy in the distribution of age in the two groups.

This situation is described graphically in Figure 3.1. In the figure it is assumed that the true relationship between age and mean 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 of methods to examine these assumptions until Chapter 4, as they are an integral part of modeling with logistic regression. Here, we proceed as if these assumptions have been checked and are supported by the data.

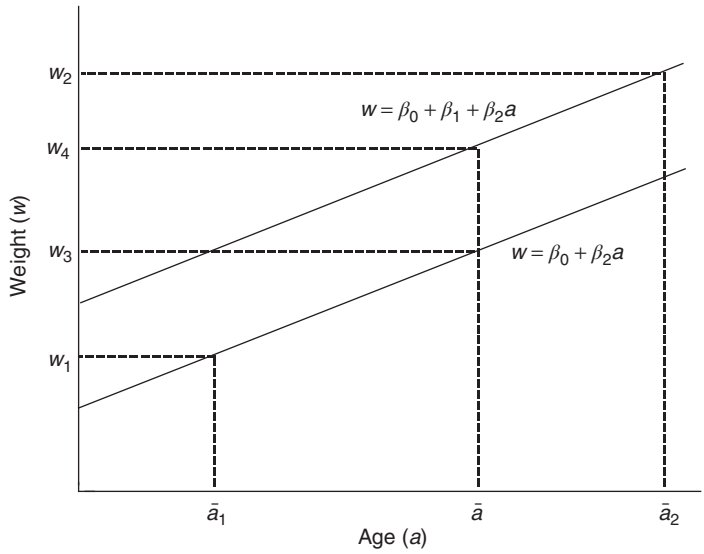

Figure 3.1 Figure describing the model for mean weight of two groups of boys as a function of age.

The statistical model that describes the situation in Figure 3.1 states that the value of mean 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, as it is the vertical distance between the two lines at any age. The coefficient β_2 is the change in weight per year of age. Suppose, as shown in Figure 3.1, that the mean age of group 1 is $\overline{a}_1$ and the mean age of group 2 is $\overline{a}_2$. 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

$$\begin{aligned}(w_2 - w_1) &= (\beta_0 + \beta_1 + \beta_2\overline{a}_1) - (\beta_0 + \beta_2\overline{a}_0)\\ &= \beta_1 + \beta_2(\overline{a}_1 - \overline{a}_0).\end{aligned}$$

This comparison involves not only the true difference between the groups, β_1, but a component, $\beta_2(\overline{a}_2 - \overline{a}_1)$, which reflects the difference between the mean ages of the groups and the association of age and weight.

The process of statistically adjusting for age involves comparing the two groups at some common value of age. The value usually used is the overall mean of the two groups, which, for the example, is denoted by $\overline{a}$ in Figure 3.1. Hence, comparing group 2 to group 1 at the mean age is, in terms of the model, a comparison of w_4 to w_3. This difference is

$$\begin{aligned}(w_4 - w_3) &= (\beta_0 + \beta_1 + \beta_2\overline{a}) - (\beta_0 + \beta_2\overline{a})\\ &= \beta_1 + \beta_2(\overline{a} - \overline{a})\\ &= \beta_1,\end{aligned}$$

which is the true difference between the mean weight of two groups. In theory any common value of age could be used, as it would yield the same difference, β_1. The choice of the overall mean makes sense for two reasons: it is clinically reasonable and lies within the range where we believe 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 outcome variable, assume it is a dichotomous variable and that the vertical axis now denotes the logit or log-odds of the outcome (i.e., in the figure w denotes the log-odds). That is, the logit of the outcome is given by the equation $g(x, a) = \beta_0 + \beta_1 x + \beta_2 a$. Under this logit model the univariable comparison of the log-odds of the two groups is approximately $(w_2 - w_1) = \beta_1 + \beta_2(\overline{a}_2 - \overline{a}_1)$. This would incorrectly estimate the effect of group on the log-odds 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, $\overline{a}$. This logit difference is, using the figure,

$$\begin{aligned}(w_4 - w_3) &= g(x = 1, a = \overline{a}) - g(x = 0, a = \overline{a})\\ &= (\beta_0 + \beta_1 + \beta_2\overline{a}) - (\beta_0 + \beta_2\overline{a})\\ &= \beta_1.\end{aligned}$$

The natural question to ask is: What conditions are required for the unadjusted difference $(w_2 - w_1)$ to be the same as the adjusted difference $(w_4 - w_3)$? Stated in terms of the model, the question is, under what conditions is $\beta_1 + \beta_2(\overline{a}_2 - \overline{a}_1) = \beta_1$? This is true if age is not associated with the outcome, $\beta_2 = 0$, or if the mean age of the two groups is the same, $(\overline{a}_2 - \overline{a}_1) = 0$. Conversely, the unadjusted and adjusted logit differences are not the same if $\beta_2(\overline{a}_2 - \overline{a}_1) \neq 0$, which only happens when both β_2 and $(\overline{a}_2 - \overline{a}_1)$ are nonzero.

As the amount of statistical adjustment or control for confounding is a function of two quantities β_2 and $(\overline{a}_2 - \overline{a}_1)$ we cannot determine whether x is a confounder simply by using a significance test of β_2. Also, in an applied setting it is impractical to calculate $(\overline{a}_2 - \overline{a}_1)$ or its equivalent for every possible pair of variables. Instead, we use some approximations. First, we fit a model containing only d (i.e., the model excludes the adjustment covariate, a. Denote the estimate of the coefficient of d from this model as $\hat{\theta}_1$. Next, we fit a model containing d along with the adjustment covariate, a. Denote the estimates of the coefficients from this model as $\hat{\beta}_1$ and $\hat{\beta}_2$ respectively. Under the model the estimate $\hat{\theta}_1$ should be approximately equal to $\hat{\beta}_1 + \hat{\beta}_2(\overline{a}_2 - \overline{a}_1)$. Hence the difference between the unadjusted and adjusted estimates of the effect of d, $(\hat{\theta}_1 - \hat{\beta}_1)$, should approximate the theoretical amount of adjustment, $\beta_2(\overline{a}_2 - \overline{a}_1)$. As the amount of adjustment is more of a relative than an absolute quantity, we scale it by dividing by $\hat{\beta}_1$ to obtain a measure we call *delta-beta-hat-percent*, defined as

$$\Delta\hat{\beta}\% = 100\frac{(\hat{\theta}_1 - \hat{\beta}_1)}{\hat{\beta}_1}. \tag{3.9}$$

Thus, the amount of adjustment is expressed as a percentage of the adjusted log-odds ratio. Some colleagues we have worked with scale differently, preferring to divide by $\hat{\theta}_1$. The disadvantage of this scaling is that both numerator and denominator now contain the amount of adjustment. The rule of thumb that we use in practice to conclude that a covariate is needed in the model to adjust the effect of another covariate is $\Delta\hat{\beta}\% > 20$. Some of our colleagues prefer to use 10% whereas others use 25%. What is important is that one calculate $\Delta\hat{\beta}\%$ and make some sort of assessment as to whether it is large enough to make a practical difference in the estimate of the log-odds ratio. Examples of the calculation and interpretation of $\Delta\hat{\beta}\%$ may be found at the end of this section and in Chapter 4.

Statistical adjustment when the variables are all dichotomous, polychotomous, continuous, or a mixture of these is identical to that just described for the case of one dichotomous and one continuous variable. The advantage of the setting we described is that it lends itself nicely to the graphical description shown in Figure 3.1.

One point must 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 assumptions of linearity in each covariate and constant slopes. Departures from either or both of these assumptions may render the adjustment useless. One commonly occurring departure is the setting where there is a statistical interaction.

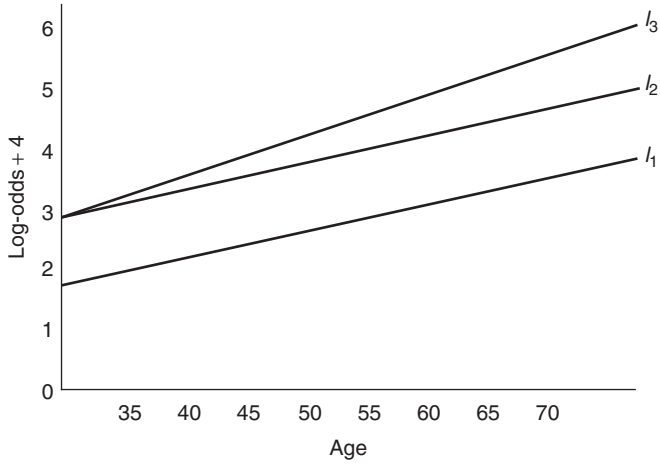

Figure 3.2 Plot of the logit under models showing the presence and absence of statistical interaction.

The simplest and most commonly used model for including a statistical interaction is one in which the logit is also linear in the second group, but with a different slope. Alternative models can be formulated that allow for a nonlinear relationship within each group. Regardless, a statistical interaction is incorporated by the inclusion of product terms of the general form "$d \times x$."

In order to more easily explain statistical interaction we plot three different logit functions in Figure 3.2, where 4 has been added to make the plotting more convenient. Suppose the plotted functions come from a setting where the outcome variable is the presence or absence of CHD, the risk factor is GENDER, and the covariate is AGE. Suppose that the line labeled l_1 corresponds to the logit for females as a function of age and 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 ratio for GENDER (male versus female), controlling for AGE, is given by the difference between line l_2 and l_1, $l_2 - l_1$. This difference is equal to the vertical distance between the two lines, which, because the lines are parallel, is the same for all ages.

Suppose instead that the logit for males is the line l_3. This line is steeper than the line l_1, for females, indicating that the relationship between AGE and CHD for males is different from that of females. When this occurs we say that there is an interaction between AGE and GENDER. The estimate of the log-odds ratios for GENDER (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 GENDER without first specifying the AGE at which the comparison is being made. In other words, age *modifies the effect* of gender, so in this terminology age is called an *effect modifier*.

Suppose we continue to consider models with the pair of independent variables d (dichotomous) and x (continuous). The role of x with respect to the effect of d in the model can be one of three possibilities.

1. There is no statistical adjustment or interaction. The covariate, x, is not a confounder or an effect modifier.
2. There is statistical adjustment but no statistical interaction. The covariate, x, is a confounder but not an effect modifier.
3. There is statistical interaction. The covariate, x, is an effect modifier.

We present an example of each of the three possibilities using data from the studies described in Section 1.6. In each example we fit three models: (i) a model that contains only d; (ii) a model that contains d and x; and (iii) a model that contains d, x and their statistical interaction, $d \times x$. We use the results of the three fitted models to decide which model is the best one to use in practice.

We begin with an example where there is neither statistical adjustment nor statistical interaction. The data we use come from the GLOW study described in Section 1.6.3. The outcome variable is having a fracture during the first year of follow up (FRACTURE). For the dichotomous variable, we use variable history of prior fracture (PRIORFRAC) and for the continuous covariate, we use height in centimeters (HEIGHT). The results from the three fitted models are presented in Table 3.10. In discussing the results from the examples we use significance levels from the Wald statistics. In all cases the same conclusions would be reached had we used likelihood ratio tests.

The Wald Statistic for the coefficient of PRIORFRAC in Model 1 is significant with $p < 0.001$. When we add HEIGHT to the model the Wald statistics are significant at the 1% level for both covariates. Note that there is little change in the

Table 3.10 Estimated Logistic Regression Coefficients, Standard Errors, Wald Statistics, *p*-Values and 95% CIs from Three Models Showing No Statistical Adjustment and No Statistical Interaction from the GLOW Study, *n* = 500

Model	Variable	Coeff.	Std. Err.	z	p	95% CI	
1	PRIORFRAC	1.064	0.2231	4.77	<0.001	0.627,	1.501
	Constant	−1.417	0.1305	−10.86	<0.001	−1.672,	−1.161
2	PRIORFRAC	1.012	0.2254	4.49	<0.001	0.570,	1.454
	HEIGHT	−0.045	0.0174	−2.61	0.009	−0.079,	−0.011
	Constant	5.785	2.7980	2.07	0.039	0.301,	11.269
3	PRIORFRAC	−3.055	5.7904	−0.53	0.598	−14.404,	8.294
	HEIGHT	−0.054	0.0219	−2.49	0.013	−0.097,	−0.012
	PRIORFRAC × HEIGHT	0.025	0.0361	0.70	0.482	−0.045,	0.096
	Constant	7.361	3.5102	2.10	0.036	0.481,	14.241

estimate of the coefficient for PRIORFRAC as

$$\Delta\hat{\beta}\% = 100 \times \frac{(1.064 - 1.012)}{1.012}$$
$$= 5.1,$$

indicating that inclusion of HEIGHT does not statistically adjust the coefficient of PRIORFRAC. Thus we conclude that, in these data, height it is not a confounder of prior fracture. The fact that the coefficient for HEIGHT is significant, $\hat{\beta}_2 \neq 0$, implies that the mean HEIGHT for the two PRIORFRAC groups must be similar. In fact they are with values of 161.7 and 161.2 cm. Under the dichotomous–continuous covariate model we showed that the univariable model coefficient should be approximately $\hat{\beta}_1 + \hat{\beta}_2(\overline{x}_1 - \overline{x}_0)$. Evaluating this expression we obtain a value of

$$1.08 = 1.012 - 0.045(160.2 - 161.7),$$

which is quite close to the value of the estimate from the univariable model of 1.064.

The statistical interaction of prior fracture (PRIORFRAC) and height (HEIGHT) is added to Model 2 to obtain Model 3. The Wald statistic for the added product term has $p = 0.482$, and thus is not significant. In these data height is not an effect modifier of prior fracture. Hence, the choice is between Model 1 and Model 2. Even though the estimate of the effect of prior fracture is basically the same for the two models, we would choose Model 2 as height (HEIGHT) is not only statistically significant in Model 2, but is an important clinical covariate as well. One would estimate the odds ratio for prior fracture using the results from Model 2 and follow the methods discussed in Section 3.2 for a dichotomous covariate coded 0 or 1.

In the next example we illustrate a setting where there is statistical adjustment but no statistical interaction. The data come from the Myopia study described in Section 1.6.6. The outcome variable is becoming myopic in the first 5 years of follow-up (MYOPIC). We use gender (GENDER) as the dichotomous variable and spherical equivalent refraction at enrollment (SPHEQ) as the continuous covariate. The results of the three fitted models are presented in Table 3.11.

The Wald test for the coefficient of GENDER in Model 1 is not significant with $p = 0.127$, which presents an interesting dilemma that occurs reasonably often in practice. We know that gender can be an important covariate, but it is not significant in the univariable model. Thus, under some model building methods it might not be considered for a multivariable model. We address this situation explicitly in Chapter 4 where we discuss *purposeful selection* of covariates. For this example, suppose we proceed on to Model 2 where we add SPHEQ. The Wald test for SPHEQ is significant with $p < 0.001$. We note that the value of the estimated coefficient for GENDER has increased from 0.366 to 0.558. In addition, it is now significant with a Wald statistic significance level of $p = 0.050$. The percentage

Table 3.11 Estimated Logistic Regression Coefficients, Standard Errors, Wald Statistics, *p*-Values and 95% CIs from Three Models Showing Statistical Adjustment and No Statistical Interaction from the Myopia Study, *n* = 618

Model	Variable	Coeff.	Std. Err.	*z*	*p*	95% CI
1	GENDER	0.366	0.2404	1.52	0.127	−0.105, 0.838
	Constant	−2.083	0.1792	−11.62	<0.001	−2.434, −1.732
2	GENDER	0.558	0.2851	1.96	0.050	−0.001, 1.117
	SPHEQ	−3.845	0.4171	−9.22	<0.001	−4.662, −3.027
	Constant	−0.226	0.2527	−0.89	0.371	−0.721, 0.269
3	GENDER	0.492	0.4157	1.18	0.237	−0.323, 1.306
	SPHEQ	−3.948	0.6353	−6.21	<0.001	−5.193, −2.703
	GENDER × SPHEQ	0.185	0.8422	0.22	0.826	−1.466, 1.836
	Constant	−0.191	0.2999	−0.64	0.524	−0.779, 0.397

difference in the two estimated coefficients is

$$\Delta\hat{\beta}\% = 100\frac{(0.366 - 0.588)}{0.588}$$
$$= -37.6.$$

Why did this happen? The mean value of SPHEQ for males is 0.781 and the mean for females is 0.821, which, although not identical, are similar in value. However, the estimated coefficient for SPHEQ is quite large and negative. Putting the two parts together under the dichotomous–continuous covariate model, the univariable estimated coefficient for GENDER is approximately

$$0.430 = 0.588 - 3.845(0.822 - 0.781),$$

which is larger than the actual univariable estimated value of 0.366. Thus we conclude that the univariable coefficient underestimates the effect of GENDER due to the fact that females tended to have larger values of spherical equivalent refraction and it is strongly negatively related to myopia.

When we add the statistical interaction of gender and spherical equivalent refraction, GENDER × SPHEQ, to the model, its estimated coefficient by Wald test is not significant with $p = 0.185$, as shown in the last row of Table 3.11. We use the estimated coefficient from Model 2 and the methods from Section 3.2 to estimate the odds ratio of gender. In this case we use Model 2 as it adjusts for SPHEQ.

In the third example we illustrate a setting where there is statistical interaction. The data we use come from the GLOW study, used earlier in the first example. Again, we use as the dichotomous variable history of prior fracture (PRIORFRAC). In this example the continuous covariate is age (AGE). The results of the three fitted models are presented in Table 3.12.

In this example we are going to see that age could be described as being both a confounder and an effect modifier. To describe age in this way is somewhat

Table 3.12 Estimated Logistic Regression Coefficients, Standard Errors, Wald Statistics, *p*-Values and 95% CIs from Three Models Showing Statistical Adjustment and Statistical Interaction from the GLOW Study, *n* = 500

Model	Variable	Coeff.	Std. Err.	*z*	*p*	95% CI
1	PRIORFRAC	1.064	0.2231	4.77	<0.001	0.627, 1.501
	Constant	−1.417	0.1305	−10.86	<0.001	−1.672, −1.161
2	PRIORFRAC	0.839	0.2342	3.58	<0.001	0.380, 1.298
	AGE	0.041	0.0122	3.38	0.001	0.017, 0.065
	Constant	−4.214	0.8478	−4.97	<0.001	−5.876, −2.553
3	PRIORFRAC	4.961	1.8102	2.74	0.006	1.413, 8.509
	AGE	0.063	0.0155	4.04	<0.001	0.032, 0.093
	PRIORFRAC × AGE	−0.057	0.0250	−2.29	0.022	−0.106, −0.008
	Constant	−5.689	1.0841	−5.25	<0.001	−7.814, −3.565

misleading and some would argue it is incorrect. So we qualify this by noting that if the analysis stopped, incorrectly, at Model 2 there is evidence of statistical adjustment. However, in Model 3 we show that there is a significant statistical interaction, whereupon Model 2 is no longer relevant. In many, if not all, practical analyses of data a vital step in modeling is deciding which model is best: the adjustment model, Model 2, or the interaction model, Model 3. We discuss this in detail in Chapter 4.

The Wald test for prior fracture (PRIORFRAC) in Model 1 is highly significant with $p < 0.001$. When we add age (AGE), Model 2, the coefficient for PRIORFRAC continues to be highly significant with $p = 0.006$. The percentage change in the coefficient for PRIORFRAC from Model 1 to Model 2 is

$$\Delta\hat{\beta}\% = 100\frac{(1.064 - 0.839)}{0.839}$$
$$= 26.8.$$

Thus the coefficient from the univariable model overestimates the effect by 26.8%. Hence, at this point, we could conclude that adding age (AGE) to the model provides an important statistical adjustment to the effect of prior fracture (PRIORFRAC). Looking at the two factors required for adjustment, the mean age of those without a prior fracture is 67.0 whereas the mean for those with a prior fracture is 73.1 and the coefficient for AGE is statistically significant. Under the dichotomous–continuous covariate model, the univariable estimated coefficient for PRIORFRAC is approximately

$$1.089 = 0.839 + 0.041(73.1 - 67.0),$$

which is quite close to the univariable value from Model 1 of 1.064.

When the interaction term, PRIORFRAC×AGE, is added to Model 2 to obtain Model 3 we see that the Wald statistic for its coefficient is statistically significant

with $p = 0.022$. Thus, there is considerable evidence of a statistical interaction between these two covariates. We are commonly asked if it is appropriate to drop the main effects and include only the interaction term in the model. In our opinion this is not appropriate, as the model must contain both main effects when the interaction is significant in order to correctly estimate odds ratios of interest.

Another commonly asked question is whether the change in the main effect coefficient for the dichotomous covariate from Model 2 to Model 3 is evidence of confounding or statistical adjustment? The answer is "no," because once an interaction is included in a model one is no longer able to estimate an adjusted odds ratio that applies to all values of the adjusting covariate. Odds ratios, as we show shortly, are estimated at specific values of the interacting covariate. In the interaction model the main effect coefficient provides an estimate of the log-odds at the value of 0 for the other covariate.

When a model contains an interaction term the only sure way to obtain the correct expression of model coefficients to estimate an odds ratio is to carefully follow the four-step method. As the estimates depend on the values of the adjusting covariate one has the option to present them graphically or in a table. We illustrate both. Although the four-step process is more complicated when an interaction term is present than it is for an adjustment model (i.e., Model 2), the results may be more interesting for subject-matter scientists.

To start, suppose we would like to estimate the odds ratio for prior fracture at some arbitrary choice of age, say "a." The four steps are as follows:

Step 1: The two sets of values of the covariates are (PRIORFRAC $= 1$, AGE $= a$) compared to (PRIORFRAC $= 0$, AGE $= a$).

Step 2: Substituting these values into the general expression for the estimated logit under Model 3 we obtain:

$$\hat{g}(\text{PRIORFRAC} = 1, \text{AGE} = a) = \hat{\beta}_0 + \hat{\beta}_1 \times 1 + \hat{\beta}_2 \times a + \hat{\beta}_3 \times 1 \times a,$$

and

$$\begin{aligned}\hat{g}(\text{PRIORFRAC} = 0, \text{AGE} = a) &= \hat{\beta}_0 + \hat{\beta}_1 \times 0 + \hat{\beta}_2 \times a + \hat{\beta}_3 \times 0 \times a \\ &= \hat{\beta}_0 + \hat{\beta}_2 \times a.\end{aligned}$$

Step 3: Taking the difference in the two functions in Step 2 and algebraically simplifying we obtain:

$$\begin{aligned}&[\hat{g}(\text{PRIORFRAC} = 1, \text{AGE} = a) - \hat{g}(\text{PRIORFRAC} = 0, \text{AGE} = a)] \\ &\quad = [(\hat{\beta}_0 + \hat{\beta}_1 \times 1 + \hat{\beta}_2 \times a + \hat{\beta}_3 \times 1 \times a) - (\hat{\beta}_0 + \hat{\beta}_2 \times a)] \\ &\quad = \hat{\beta}_1 + \hat{\beta}_3 \times a. \qquad (3.10)\end{aligned}$$

This is the correct function of the coefficients to exponentiate in Step 4 to estimate the odds ratio for prior fracture, specifically at AGE $= a$. Note

that this expression involves the coefficients for both the main effect and interaction terms.

Step 4: Exponentiating the result of Steps 3 we obtain

$$\widehat{\text{OR}}[(\text{PRIORFRAC} = 1, \text{AGE} = a), (\text{PRIORFRAC} = 0, \text{AGE} = a)] = \exp[\hat{\beta}_1 + \hat{\beta}_3 \times a]. \tag{3.11}$$

Following a point made earlier, we see from equation (3.11) that $\exp(\hat{\beta}_1)$ is the AGE = 0 estimate of the odds ratio for PRIORFRAC (a quantity that is obviously not clinically relevant in a study of women 55 and older).

As noted earlier, we have the choice at this point to either tabulate or graph the results. As this is our first example of a model with an interaction we delve into it in more detail than one might typically do in practice. The easiest way to see the nature of the interaction is to plot the two logit functions in Step 2 as a function of age, which we present in Figure 3.3.

The upper line in Figure 3.3 is a plot of the log-odds for subjects with a prior fracture, which from Model 3 in Table 3.12 and equation (3.10) is

$$\hat{g}(\text{PRIORFRAC} = 1, \text{AGE} = a) = -5.689 + 4.961 \times 1 + 0.063 \times a - 0.057 \times a \times 1 = -0.728 + 0.006 \times a.$$

The lower line is the logit for subjects without a prior fracture and is

$$\hat{g}(\text{PRIORFRAC} = 0, \text{AGE} = a) = -5.689 + 4.961 \times 0 + 0.063 \times a - 0.057 \times a \times 0 = -5.689 + 0.063 \times a.$$

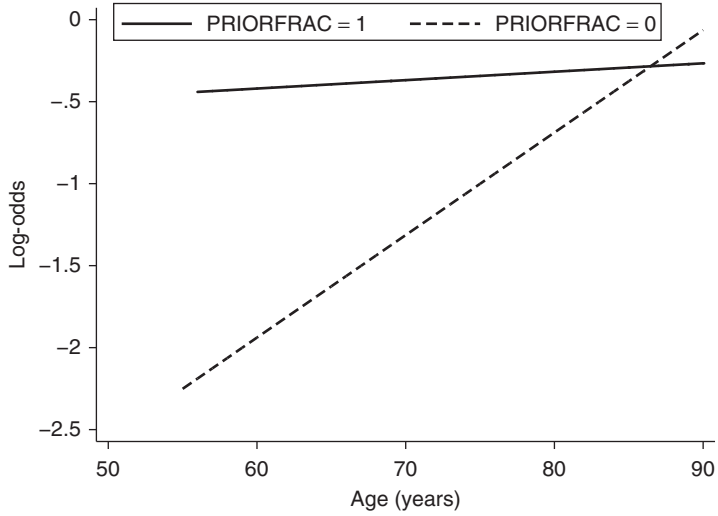

Figure 3.3 Plot of the estimated logit as a function of age for subjects with PRIORFRAC = 1 and PRIORFRAC = 0.

The log-odds ratio is given in equation (3.10) and is the vertical distance between the two lines in Figure 3.3 at a specific age, a, and is equal to

$$\ln\{\widehat{\text{OR}}[(\text{PRIORFRAC} = 1, \text{AGE} = a), (\text{PRIORFRAC} = 0, \text{AGE} = a)]\}$$
$$= 4.961 - 0.057 \times a.$$

This function is plotted in Figure 3.4.

We have included the 95% confidence bands in Figure 3.4. These are calculated, using equation (3.10), at each observed value of age as

$$(\hat{\beta}_1 + \hat{\beta}_3 \times a) \pm 1.96 \times \widehat{\text{SE}}(\hat{\beta}_1 + \hat{\beta}_3 \times a), \tag{3.12}$$

where

$$\widehat{\text{SE}}(\hat{\beta}_1 + \hat{\beta}_3 \times a) = [\widehat{\text{Var}}(\hat{\beta}_1) + a^2\widehat{\text{Var}}(\hat{\beta}_3) + 2a\widehat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_3)]^{0.5}. \tag{3.13}$$

The actual plotted values in Figure 3.4 are obtained by substituting the estimates of the coefficients from Model 3 in Table 3.12 and the values of the estimated covariance matrix of the estimated coefficients (not shown) into equations (3.12) and (3.13). One should note that the form of the plot in Figure 3.4 is similar to the plot of a linear regression model with its confidence bands being narrower in the middle at about the mean age, 68.6, and wider at the extremes. We added a line at log-odds ratio of 0 to the figure to aid interpretation. We see that the lower confidence limit crosses the 0-line at 78 years. This means that the log-odds ratio is not significantly different from 0 for ages greater than or equal to 78.

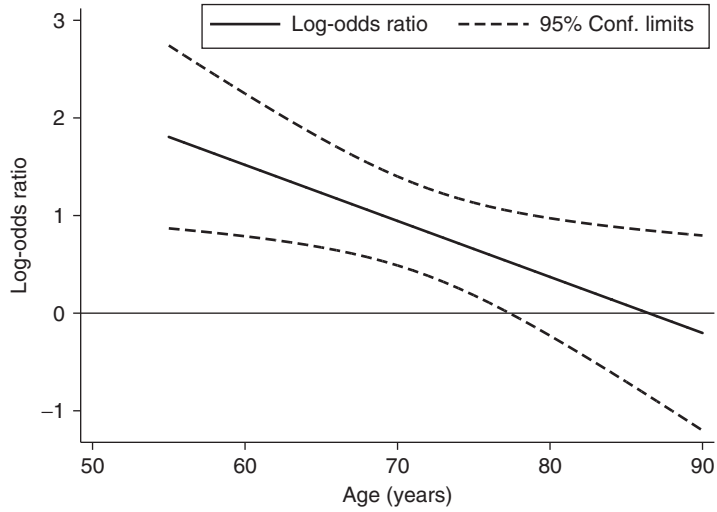

Figure 3.4 Plot of the estimated log-odds ratio for PRIORFRAC = 1 versus PRIORFRAC = 0 as a function of age, with 95% confidence bands.

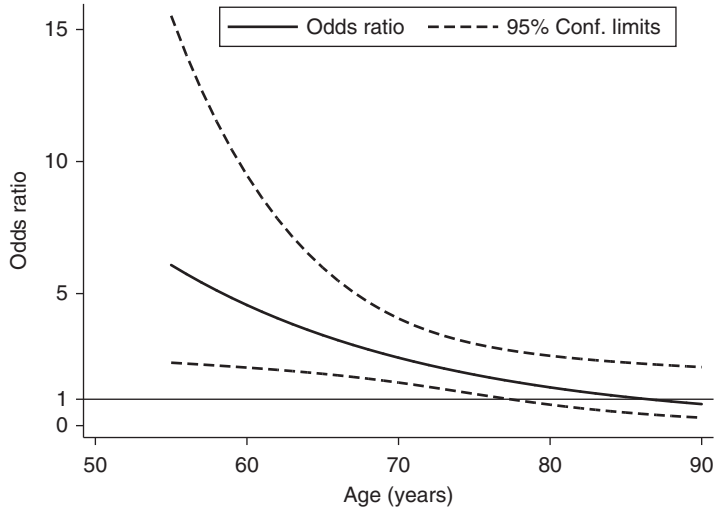

Figure 3.5 Plot of the estimated odds ratio for PRIORFRAC = 1 versus PRIORFRAC = 0 as a function of age, with 95% confidence bands.

Figure 3.5 presents the plot of the estimated odds ratio and its confidence limits. This is accomplished by exponentiating the values on the three lines in Figure 3.4. A horizontal line at 1.0 is added to aid interpretation. We see that the estimated odds ratio decreases from a value of about 6 at 55 years and becomes insignificant at 78 years, where the lower confidence limit line drops below 1.0. The problem with the plot is that the upper confidence limits are so large at ages below about 60 that the rest of the lines become compressed and are difficult to read with any accuracy. This is often the case in an interaction model with a continuous covariate and for this reason, in practice, we prefer a plot of the log-odds ratio.

The advantage of a plot is that it describes, in a general way, how the estimated log-odds ratios or odds ratios change as a function of the plotted covariate. It is, however, not as useful as a table for obtaining specific values. We show in Table 3.13 estimates of the odds ratio and confidence intervals at ages 55, 60, 65, 70, and 80 years of age. These values are obtained by first evaluating equations (3.12) and (3.13) at each of the ages and then exponentiating the values.

The values in Table 3.13 provide more detail on how the odds ratio for prior fracture decreases as a function of age from 6.1 at age 55 to the point where it becomes not statistically significant at age 80 (actually 78). Note that had we incorrectly used Model 2 we would have stated that the "age-adjusted" odds ratio is $2.3 = \exp(0.839)$, implying that this estimate was valid for all ages. In fact, we can see from the results in Table 3.13, this is only true for age approximately equal to 72.

In summary, the examples in this section demonstrate that evidence for a covariate being necessary in a model to adjust for the effect of another variable cannot be determined by a statistical test. It is a judgmental decision based on the change

Table 3.13 Estimated Odds Ratios for Prior Fracture as a Function of Age from Model 3 in Table 3.12

Age	Odds Ratio	95% CI
55	6.1	2.38, 15.53
60	4.6	2.20, 9.49
65	3.4	1.96, 5.99
70	2.6	1.63, 4.06
75	1.9	1.20, 3.11
80	1.4	0.79, 2.65

in the estimate of a coefficient. When there is a statistically significant interaction statistical adjustment is no longer an issue as one must estimate the odds ratio at specific values of the covariate. These are most easily calculated by carefully following the four-step method.

3.6 PRESENTATION AND INTERPRETATION OF THE FITTED VALUES

In previous sections of this chapter we discussed using the logistic regression model coefficients to estimate odds ratios and construct confidence intervals 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 (i.e., the estimated probabilities) from the model are equally, if not more, important. For example, Groeger et al. (1998) used logistic regression modeling methods to estimate a patient's probability of hospital mortality after admission to an intensive care unit. We discussed in Sections 1.4 and 2.5 the methods for computing a fitted value and its confidence interval estimate. In this section, we expand this work to include graphical presentations of fitted values. In addition we discuss prediction of the outcome for a subject not in the estimation sample.

Settings where predicted probabilities are of interest tend be those where there is a reasonably wide range in the values. Conversely, if the range is too narrow graphs of fitted values tend to look like straight lines and thus are not much different, though on a different scale, than plots of fitted logits shown in Section 3.5 and add little to the analysis. Among the data sets described in Section 1.6 the Burn Study (Section 1.6.5) has the widest range of fitted values and we use it for the example in this section.

Suppose we fit a model containing the total burn surface area (TBSA) and burn involved an inhalation injury (INH_INJ). Furthermore, suppose we are interested in describing, graphically, the effect of these two covariates on the estimated probabilities. We encourage the reader to review the details in Section 1.6.5 on how these data were sampled from a much larger data set. The results of the fit are shown in Table 3.14.

Table 3.14 Fitted Multiple Logistic Regression Model of Death from a Burn Injury (DEATH) on TBSA and Inhalation Injury Involved (INH_INJ) from the Burn Study, $n = 1000$

Variable	Coeff.	Std. Err.	z	p	95% CI
TBSA	0.073	0.0072	10.11	<0.001	0.059, 0.087
INH_INJ	1.290	0.2926	4.41	<0.001	0.716, 1.863
Constant	−3.380	0.1776	−19.03	<0.001	−3.728, −3.031

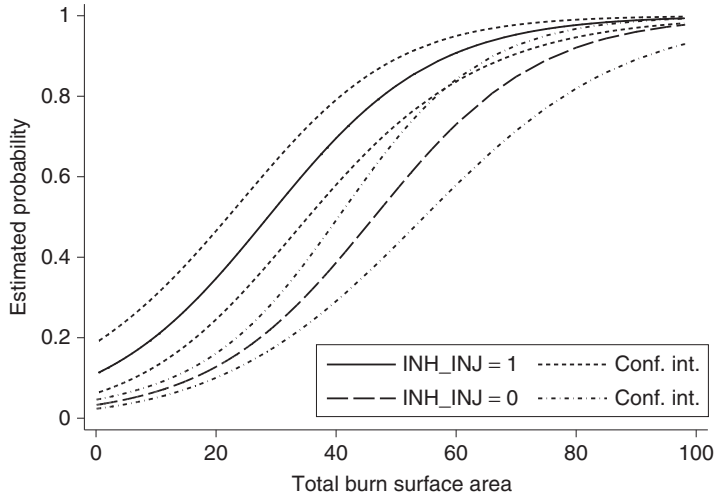

Figure 3.6 Plot of the fitted values from the model in Table 3.14 and their 95% confidence bands.

Both burn area and inhalation injury are highly significant and, after controlling for TBSA, the estimated odds ratio for inhalation injury is 3.63 (95% confidence interval: 2.05, 6.45). Clearly, involvement of an inhalation injury greatly increases the odds of dying, but how could we express the effect of this variable on the probability of dying? In this example, the model is not complicated and contains the continuous covariate TBSA, thus a plot of the fitted values versus burn area for those with and without inhalation injury involvement provides a simple graphical summary of the effect of the two covariates. This is shown in Figure 3.6 along with the 95% confidence bands.

The plotted curves in Figure 3.6 show that the estimated probability of death from a burn injury ranges from about 0.03 (3%) for a small burn with no inhalation injury involvement to almost 1.0 (100%) for subjects with a burn area of more than 95%. The plotted confidence bands for two fitted value curves show that inhalation injury involvement greatly increases the estimated probability of death, particularly when the burn area is between 20% and 60%. For burn area greater than 70% the estimated probability curves converge and approach 1.0. In this range the estimated

probabilities are so large that inhalation injury cannot add much. Note that we have described the difference in the two curves as an additive difference rather than a relative difference. Under the fitted model the relative difference (i.e., the estimated odds ratio) for inhalation injury involvement is 18.7 at all estimated probabilities. We discuss summary measures that use fitted values to describe model performance in Chapter 5.

Using the methods described at the end of Section 2.5 and the results in Table 3.14 the required calculations to obtain the values plotted in Figure 3.6 are as follows. First, we calculate the two fitted logit functions:

$$\begin{aligned}\hat{g}_1(a) &= \hat{g}(\text{TBSA} = a, \text{INH\_INJ} = 1)\\ &= -3.380 + 0.073 \times a + 1.290 \times 1\\ &= -2.090 + 0.073 \times a,\end{aligned}$$

$$\begin{aligned}\hat{g}_0(a) &= \hat{g}(\text{TBSA} = a, \text{INH\_INJ} = 0)\\ &= -3.380 + 0.073 \times a + 1.290 \times 0\\ &= -3.380 + 0.073 \times a.\end{aligned}$$

Next, we compute the estimator of the variance of each estimated logit:

$$\begin{aligned}\widehat{\text{Var}}[\hat{g}_1(a)] &= 0.03154 + (a^2) \times 0.00005199 + (1^2) \times 0.08561 - 2 \times a\\ &\quad \times 0.0008339 - 2 \times 1 \times 0.007801 \times 1 - 2 \times a \times 1 \times 0.0006466\\ &= 0.1093 - a \times 0.01727 + (a)^2 \times 0.00005199,\end{aligned}$$

and

$$\begin{aligned}\widehat{\text{Var}}[\hat{g}_0(a)] &= 0.03154 + (a^2) \times 0.00005199 + (0^2) \times 0.08561 - 2 \times a \times 0.0008339\\ &\quad - 2 \times 0 \times 007801 \times 0 - 2 \times a \times 0 \times 0.0006466\\ &= 0.03154 - a \times 0.00001667 + (a)^2 \times 0.00005199,\end{aligned}$$

where the values of the various estimated variances and covariances are obtained from the estimated covariance matrix of the estimated parameters in the fitted model (not shown but available from all software packages). Hence the two sets of fitted values as a function of burn area are

$$\hat{\pi}_j(a) = \frac{e^{\hat{g}_j(a)}}{1 + e^{\hat{g}_j(a)}}, \qquad j = 0, 1, \tag{3.14}$$

and their lower (l) and upper (u) confidence bands are obtained from

$$\hat{\pi}_j^{\text{l}}(a), \hat{\pi}_j^{\text{u}}(a) = \frac{e^{\hat{g}_j(a) \pm 1.96\widehat{\text{SE}}(\hat{g}_j(a))}}{1 + e^{\hat{g}_j(a) \pm 1.96\widehat{\text{SE}}(\hat{g}_j(a))}}, \qquad j = 0, 1, \tag{3.15}$$

where $\widehat{\text{SE}}(g) = \sqrt{\widehat{\text{Var}}(g)}$.

We now focus our attention on the fitted value and confidence interval for the single set of values (TBSA = 30, INH_INJ = 1). As these two values are among those in the data set we can obtain them from those used in the plot in Figure 3.6. In fact, the estimated probability is 0.52 with 95% confidence interval (0.41, 0.64). The interpretation is that among patients admitted with a burn area of 40% and inhalation injury involvement the model estimates that 52% would die and it could be between 41% and 64% with 95% confidence.

If the covariate values that we would like estimates for are within the range of those in the observed data set but not specifically present (e.g., TBSA = 64, INH_INJ = 1) then we use the expressions for $\hat{g}_1(a)$ and $\widehat{\text{Var}}[\hat{g}_1(a)]$ with these values and use the results to evaluate equations (3.14) and (3.15).

The fitted model in Table 3.14 is much simpler than one typically uses to model data in a practical multivariable data set. To extend the bivariable example in Table 3.14 we show in Table 3.15 the fit of a model that adds age (AGE), gender (GENDER), race (RACE) and flame involved in the burn injury (FLAME) to the model. Suppose that all variables are kept in the model for either clinical or statistical reasons. We would like to plot the estimated probability of death as a function of burn area (same as Figure 3.6) and inhalation injury but now controlling for the other four covariates in the model.

In order to control for the additional covariates we could choose "typical" values for each (e.g., median age and 0 for the three dichotomous covariates). However, these values may not provide a logit that is, in some sense, at the median or middle of the log-odds of death for these covariates. What we propose is to calculate a modified logit that subtracts the contribution of burn area and inhalation injury from the logit and uses its median value as a way to control for the additional model covariates. Specially, the logit for the fitted model in Table 3.15 is

$$\hat{g}(\mathbf{x}) = -7.695 + 0.089 \times \text{TBSA} + 1.365 \times \text{INH\_INJ} + 0.083 \times \text{AGE} - 0.201 \times \text{GENDER} + 0.583 \times \text{FLAME} - 0.701 \times \text{RACE}.$$

Table 3.15 Fitted Multiple Logistic Regression Model of Death from a Burn Injury (DEATH) on Total Body Surface Area (TBSA), Inhalation Injury (INH_INJ), Age (AGE), Gender (GENDER), Race (RACE), and Flame Involved (FLAME) from the Burn Study, n = 1000

Variable	Coeff.	Std. Err.	z	p	95% CI
TBSA	0.089	0.0091	9.83	<0.001	0.072, 0.107
INH_INJ	1.365	0.3618	3.77	<0.001	0.656, 2.074
AGE	0.083	0.0086	9.61	<0.001	0.066, 0.100
GENDER	−0.201	0.3078	−0.65	0.513	−0.805, 0.402
FLAME	0.583	0.3545	1.64	0.100	−0.112, 1.277
RACE	−0.701	0.3098	−2.26	0.024	−1.309, −0.094
Constant	−7.695	0.6912	−11.13	<0.001	−9.050, −6.341

Our proposed modified logit is

$$\widehat{gm}(\mathbf{x}) = \hat{g}(\mathbf{x}) - (+0.089 \times \text{TBSA} + 1.365 \times \text{INH\_INJ}),$$

and the median value of $\widehat{gm}(\mathbf{x})$ over the 1000 subjects is $\widehat{gm}_{50} = -5.349$. Here we use $\mathbf{x}$ to generically denote the covariates. Next we calculate the adjusted logit for the two inhalation injury groups as a function of burn area as

$$\begin{aligned}\hat{g}_1(a) &= \widehat{gm}_{50} + 0.089 \times \text{TBSA} + 1.365 \times 1 \\ &= -5.349 + 0.089 \times \text{TBSA} + 1.365 \\ &= -3.984 + 0.089 \times \text{TBSA},\end{aligned}$$

and

$$\begin{aligned}\hat{g}_0(a) &= \widehat{gm}_{50} + 0.089 \times \text{TBSA} + 1.365 \times 0 \\ &= -5.349 + 0.089 \times \text{TBSA} + 0 \\ &= -5.349 + 0.089 \times \text{TBSA}.\end{aligned}$$

The estimated probabilities are computed using equation (3.14) and are plotted in Figure 3.7.

The plot in Figure 3.7 shows, quite clearly, how having an inhalation injury increases the probability of death over the range of burn area. A specific value is easily obtained by substituting in values for TBSA and INH_INJ into the equation for the logit. For example, for the pair (TBSA = 40, INH_INJ = 0) the value of the logit is

$$\begin{aligned}\hat{g}_0(a) &= -5.349 + 0.089 \times 40 \\ &= -1.789,\end{aligned}$$

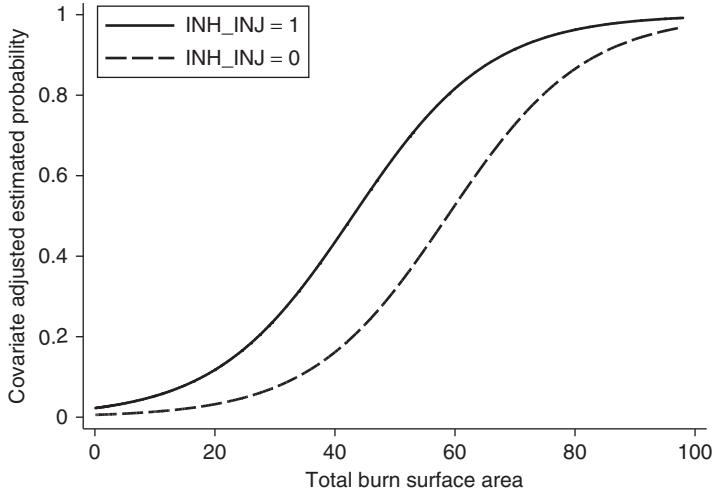

Figure 3.7 Plot of the covariate adjusted fitted values from the model in Table 3.15.

and the covariate adjusted probability is

$$\hat{\pi}_0(40) = \frac{e^{-1.789}}{1 + e^{-1.789}} = 0.143.$$

Using the modified logit $\widehat{gm}(\mathbf{x})$ avoids having to choose specific values for the covariates. Using its median value adjusts at a middle level of log-odds for these covariates. However, as we have not used specific covariate values, an extension of the expression for $\widehat{\text{Var}}[\hat{g}_j(a)]$ can no longer be evaluated and thus confidence bands based on it are not possible to compute. Confidence bands can be obtained with some additional programming and using a resampling method called bootstrapping. As this topic is beyond the technical level of this text, we do not consider it further here.

As is the case with any regression model we must take care not to extend model-based inferences beyond the observed range of the data. It is also important to keep in mind that any estimate is only as good as the model upon which it is based. 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.

3.7 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 produce an adjusted odds ratio. This is accomplished by first determining whether the odds ratios are constant, or homogeneous, over a number of 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 3.5 and 3.6. In this section we compare these two approaches. An example from the Burn Study illustrates the similarities and differences in the two approaches.

Consider an analysis of the risk factor whether a flame was involved in the burn injury (FLAME) on the outcome variable vital status at hospital discharge (DEATH). The crude (or unadjusted) odds ratio computed from the 2 × 2 table shown in Table 3.16, cross-classifying the outcome variable DEATH with FLAME, is $\widehat{\text{OR}} = 7.35$.

As we have seen earlier in this chapter, total body surface area burned (TBSA) is an important determinant of patient survival. Examination of the distribution of TBSA shows that the 25th, 50th, and 75th percentiles of body surface area are 2.5%, 6%, and 16%, respectively. Using these quartiles, Table 3.17 presents the cross tabulation of DEATH by FLAME within each of the four quartiles of TBSA.

Table 3.16 Cross-Classification of Vital Status at Hospital Discharge (DEATH) by Whether a Flame Was Involved in the Burn Injury (FLAME)

		FLAME		
		0	1	Total
DEATH	0	451	399	850
	1	20	130	150
Total		471	529	1000

Table 3.17 Cross-Classification of DEATH by FLAME Stratified by TBSA Quartile Groups

			FLAME		
TBSA			0	1	Total
TBSA $<$ 2.5%	DEATH	0	168	73	241
		1	3	2	5
		Total	171	75	246
2.5% $\leq$ TBSA $<$ 6%	DEATH	0	124	101	225
		1	2	6	8
		Total	126	107	233
6% $\leq$ TBSA $<$ 16%	DEATH	0	117	134	251
		1	5	14	19
		Total	122	148	270
TBSA $\geq$ 16%	DEATH	0	42	91	133
		1	10	108	118
		Total	52	199	251

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, $\widehat{\text{OR}}_i = (a_i \times d_i)/(b_i \times c_i)$, 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, $a_1 = 168$, $b_1 = 73$, $c_1 = 3$, and $d_1 = 2$, and the total number of subjects is $N_1 = 246$. The Mantel–Haenszel estimator of the odds ratio is defined in this case as follows:

$$\widehat{\text{OR}}_{\text{MH}} = \frac{\sum \dfrac{a_i \times d_i}{N_i}}{\sum \dfrac{b_i \times c_i}{N_i}}. \tag{3.16}$$

Evaluating equation (3.16) using the data in Table 3.17 yields the Mantel–Haenszel estimate

$$\widehat{\text{OR}}_{\text{MH}} = \frac{28.697}{7.864} = 3.65.$$

Table 3.18 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 FLAME Within Each Quartile of TBSA

	Quartile of TBSA			
	1	2	3	4
$\widehat{\text{OR}}$	1.534	3.683	2.445	4.985
$\ln(\widehat{\text{OR}})$	0.428	1.304	0.894	1.606
$\widehat{\text{Var}}[\ln(\widehat{\text{OR}})]$	0.853	0.685	0.287	0.144
w	1.172	1.461	3.479	6.942

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,

$$\widehat{\text{OR}}_{\text{L}} = \exp\left[\frac{\sum w_i \ln\left(\widehat{\text{OR}}_i\right)}{\sum w_i}\right]. \tag{3.17}$$

Table 3.18 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

$$\widehat{\text{OR}}_{\text{L}} = \exp\left(\frac{16.667}{13.054}\right) = 3.585,$$

which is slightly smaller than the Mantel–Haenszel estimate. 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 0.

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

$$X^2_{\text{H}} = \sum\{w_i[\ln(\widehat{\text{OR}}_i) - \ln(\widehat{\text{OR}}_{\text{L}})]^2\}. \tag{3.18}$$

Under the hypothesis that the odds ratios are constant, X^2_{H} 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^2_{H} is large.

Using the data in Table 3.18 we have $X^2_{\mathrm{H}} = 2.11$ which, with 3 degrees of freedom, yields a p-value of 0.5492. Thus, the logit-based test of homogeneity indicates that the four groups, based on the quartiles of the distribution of TBSA, 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, $\hat{e}_i$, calculated under the assumption that the odds ratio is constant in all strata. As noted by Breslow (1996) the correct formula for the test statistic is

$$X^2_{\mathrm{BD}} = \sum \frac{(a_i - \hat{e}_i)^2}{\hat{v}_i} - \frac{\left[\sum (a_i) - \sum (\hat{e}_i)\right]^2}{\sum (\hat{v}_i)}. \tag{3.19}$$

We note that some packages, for example, STATA, calculate the first part of equation (3.19) as the Breslow–Day test and the entire expression in equation (3.19) as the Tarone test. The quantity $\hat{e}_i$ is one of the two solutions to the following quadratic equation:

$$\widehat{\mathrm{OR}} = \frac{(\hat{e}_i)(n_{1i} - m_{0i} + \hat{e}_i)}{(n_{0i} - \hat{e}_i)(m_{0i} - \hat{e}_i)}, \tag{3.20}$$

where $n_{0i} = a_i + b_i$, $m_{0i} = a_i + c_i$, and $n_{1i} = c_i + d_i$. The two solutions for $\hat{e}_i$ in equation (3.20) are found by evaluating the following expressions

$$\frac{-s_i + \sqrt{s_i^2 - 4 \times r \times t_i}}{2 \times r} \text{ and } \frac{-s_i - \sqrt{s_i^2 - 4 \times r \times t_i}}{2 \times r}, \tag{3.21}$$

where $r = 1 - \widehat{\mathrm{OR}}$, $s_i = (n_{1i} - m_{0i}) + (\widehat{\mathrm{OR}})(m_{0i} + n_{0i})$, and $t_i = -(\widehat{\mathrm{OR}})(n_{0i} m_{0i})$, but only one of them yields an estimated frequency that is positive and less than both n_{0i} and m_{0i}.

The quantity $\widehat{\mathrm{OR}}$ in equation (3.20) is an estimate of the common odds ratio and either $\widehat{\mathrm{OR}}_{\mathrm{L}}$ or $\widehat{\mathrm{OR}}_{\mathrm{MH}}$ may be used, but the default used in most packages is the Mantel–Haenszel estimator. 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

$$\hat{v}_i = \left(\frac{1}{\hat{e}_i} + \frac{1}{n_{0i} - \hat{e}_i} + \frac{1}{m_{0i} - \hat{e}_i} + \frac{1}{n_{1i} - m_{0i} + \hat{e}_i}\right)^{-1}. \tag{3.22}$$

If we use the value of the Mantel–Haenszel estimate, $\widehat{\mathrm{OR}}_{\mathrm{MH}} = 3.65$ to compute the Breslow–Day test in equation (3.19) then $X^2_{\mathrm{BD}} = 2.18$ ($p = 0.5366$), which is similar to the value of the logit-based test.

The same analysis may be performed much more easily by fitting three logistic regression models. In model 1 we include only the variable FLAME. We then

Table 3.19 Estimated Logistic Regression Coefficients for the Variable FLAME, 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 Quartiles of TBSA

Model	FLAME	Log-Likelihood	G	df	p
1	1.994	−258.34			
2	1.296	−288.64	178.17	3	<0.001
3		−287.57	2.14	3	0.545

add the three design variables representing the four quartiles of TBSA to obtain model 2. For model 3 we add the three TBSA × FLAME interaction terms. The results of fitting these models are shown in Table 3.19. As we are primarily interested in the estimates of the coefficient for FLAME, the estimates of the coefficients for TBSA and the FLAME × TBSA interactions are not shown in Table 3.19.

Using the estimated coefficients in Table 3.19 we have the following estimated odds ratios. The crude odds ratio is $\widehat{\text{OR}} = \exp(1.994) = 7.35$. Adjusting for TBSA, the stratified estimate is $\widehat{\text{OR}} = \exp(1.2958) = 3.65$. This value is the maximum likelihood estimate of the estimated odds ratio, and it is similar in value to both the Mantel–Haenszel estimate, $\widehat{\text{OR}}_{\text{MH}} = 3.65$, and the logit-based estimate, $\widehat{\text{OR}}_{\text{L}} = 3.59$. The change in the estimate of the odds ratio from the crude to the adjusted is 7.35 to 3.65, indicating considerable confounding due to TBSA.

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 3.19 is $G = 2.14$. This statistic is compared to a chi-square distribution with 3 degrees of freedom, as three 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} (=2.11), and the Breslow–Day test, X^2_{BD} (=2.18), each with 3 degrees of freedom.

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, as we did in the example from the burn data, 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 SAS.

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.

EXERCISES

1. Consider the ICU data described in Section 1.6.1 and use as the outcome variable vital status (STA) and infection probable at ICU admission (INF) as a covariate.

(a) Demonstrate that the value of the log-odds ratio obtained from the cross-classification of STA by INF is identical to the estimated slope coefficient from the logistic regression of STA on INF. Verify that the estimated standard error of the estimated slope coefficient for INF obtained from the logistic regression package is identical to the square root of the sum of the inverse of the cell frequencies from the cross-classification of STA by INF. Use either set of computations, contingency table, or logistic regression, to obtain the 95% confidence interval for the odds ratio.

(b) For purposes of illustration, use a data transformation statement to recode, for this problem only, the variable INF as follows: 4 = No and 2 = Yes. Perform the logistic regression of STA on INF (recoded). Use the four-step method to calculate the estimate of the odds ratio of INF = Yes versus INF = No. Use the results from the fitted logistic regression model to obtain the 95% confidence interval for the odds ratio. Note that they are the same limits as obtained in Exercise 1(a).

2. Consider data from the Low Birth Weight Study described in Section 1.6.2 and use as the outcome variable low birth weight (LOW) and race of the mother (RACE) as the covariate.

(a) 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 LOW 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 LOW 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 LOW by RACE used to calculate the odds ratio. Use either set of computations to compute the 95% confidence interval for the odds ratios. Note that in this example the results are significant at the 10 but not 5% level of significance. Explain circumstances under which you would choose to keep RACE in a statistical model and ones when you might not keep it.

(b) Create design variables for RACE using the deviation from means coding typically employed in ANOVA. Perform the logistic regression of LOW on RACE. Use the four-step method to compute the estimate of the odds ratio RACE = 2 versus RACE = 1 and RACE = 3 versus RACE = 1. Are these estimates the same as those computed in 2(a)? Use the results of the logistic regression to obtain the 95% confidence interval for the odds ratios and verify that they are the same limits as obtained in 2(a). In this example

you need the estimated covariance matrix for the estimated coefficients to obtain the estimated variances of the two log-odds ratios.

3. In the ICU data vital status at discharge (STA) is the outcome variable and consider history of chronic renal failure (CRN) as the factor of interest. Using logistic regression, demonstrate and then explain why age (AGE) is needed to adjust the effect of CRN. Using logistic regression modeling, demonstrate that there is no statistical interaction between age (AGE) and history of chronic renal failure (CRN).

4. Repeat problem 3 using cancer part of the present problem (CAN) as the factor of interest and type of admission (TYP) as a potential adjustment and interaction variable.

5. In the Burn Injury Data described in Section 1.6.5 vital status at hospital discharge (DEATH) is the outcome variable.

(a) Show that age (AGE) is not a confounder of the effect of inhalation injury (INH_INJ) but is an effect modifier.

(b) Using the interactions model from part 5(a) and the four-step method prepare a table with estimates of the odds ratio and 95% confidence interval for inhalation injury for ages 20, 40, 60, and 80.

(c) Using the interaction model from part 5(a) prepare a graph of the estimate of the odds ratio for inhalation injury as a function of age.

(d) Add 95% confidence bands to the graph in part 5(c).

6. The outcome variable in the Myopia Study described in Section 1.6.6 is becoming myopic during the first five years of follow up (MYOPIC). Consider a logistic regression model containing spherical equivalent refraction (SHPQ), gender (GENDER), sports hours (SPORTHR), reading hours (RESDHR), computer hours (COMPHR), study hours (STUDYHR) and television hours (TVHR). Graph the fitted logistic probability of becoming myopic for males and females as a function of spherical equivalent refraction (SHPQ) adjusted for all other variables in the model.

7. In the Low Birth Weight Study described in Section 1.6.2, determine the crude odds ratio of smoking (SMOKE) on the outcome low birthweight (LOW). Stratify on RACE and note the odds ratios within the three strata. Do the odds ratios appear to be homogeneous across strata? Compute the Mantel–Haenszel and logit-based estimates of the odds ratio. How do these compare to the crude estimate? Determine whether homogeneity of the odds ratios across strata holds through the use of the chi-square test of homogeneity and the Breslow–Day test. Finally, use a logistic regression analysis to compute the adjusted odds ratio and to determine whether the odds ratios were homogeneous across strata. How do these results compare to the ones you obtained using the more classical categorical data approach?