CHAPTER 2

The Multiple Logistic Regression Model

2.1 INTRODUCTION

In Chapter 1 we introduced the logistic regression model in the context of a model containing a single variable. As in the case of linear regression, the strength of the logistic regression model is its ability to handle many variables, some of which may be on different measurement scales. In this chapter, we generalize the model to one with more than one independent variable (i.e., the multivariable or multiple logistic regression model). Central to the consideration of the multiple logistic models is estimating the coefficients and testing for their significance. We use the same approach discussed in Chapter 1 for the univariable setting. An additional modeling consideration, which is introduced in this chapter, is using design variables for modeling discrete, nominal scale, independent variables. In all cases, we assume that there is a predetermined collection of variables to be examined. We consider statistical methods for selecting variables in Chapter 4.

2.2 THE MULTIPLE LOGISTIC REGRESSION MODEL

Consider a collection of p independent variables denoted by the vector $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$. For the moment we assume that each of these variables is at least interval scaled. Let the conditional probability that the outcome is present be denoted by $\Pr(Y = 1|\mathbf{x}) = \pi(\mathbf{x})$. The logit of the multiple logistic regression model is given by the equation

$$g(\mathbf{x}) = \ln\left(\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad (2.1)$$

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

where, for the multiple logistic regression model,

$$\pi(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}}. \tag{2.2}$$

If some of the independent variables are discrete, nominal scale variables such as race, sex, treatment group, and so forth, it is inappropriate to include them in the model as if they were interval scale variables. The numbers used to represent the various levels of these nominal scale variables are merely identifiers, and have no numeric significance. In this situation, the method of choice is to use a collection of *design variables* (or *dummy variables*). Suppose, for example, that one of the independent variables is race, which has been coded as "white," "black," and "other." In this case, two design variables are necessary. One possible coding strategy is that when the respondent is "white," the two design variables, D_1 and D_2, would both be set equal to zero; when the respondent is "black," D_1 would be set equal to 1 while D_2 would still equal 0; when the race of the respondent is "other," we would use $D_1 = 0$ and $D_2 = 1$. Table 2.1 illustrates this coding of the design variables.

Every logistic regression software package we use has the capability to generate design variables, and some provide a choice of several different methods. We discuss different strategies for creation and interpretation of the coefficients for the design variables in detail in Chapter 3.

In general, if a nominal scaled variable has k possible values, then $k - 1$ design variables are needed. The reason for using one less than the number of values is that, unless stated otherwise, our models have a constant term. To illustrate the notation used for design variables in this text, suppose that the jth independent variable x_j has k_j levels. The $k_j - 1$ design variables will be denoted as D_{jl} and the coefficients for these design variables will be denoted as $\beta_{jl}, l = 1, 2, \ldots, k_j - 1$. Thus, the logit for a model with p variables, with the jth variable being discrete is

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \sum_{l=1}^{k_j - 1} \beta_{jl} D_{jl} + \beta_p x_p.$$

With a few exceptions, we suppress the summation and double subscripting needed to indicate when design variables are being used when discussing the multiple logistic regression model.

Table 2.1 An Example of the Coding of the Design Variables for Race, Coded at Three Levels

RACE	D_1	D_2
White	0	0
Black	1	0
Other	0	1

2.3 FITTING THE MULTIPLE LOGISTIC REGRESSION MODEL

Assume that we have a sample of n independent observations $(\mathbf{x}_i, y_i), i = 1, 2, \ldots, n$. As in the univariable case, fitting the model requires that we obtain estimates of the vector $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_p)$. The method of estimation used in the multivariable case is the same as in the univariable situation – maximum likelihood. The likelihood function is nearly identical to that given in equation (1.3) with the only change being that $\pi(\mathbf{x})$ is now defined as in equation (2.1). There will be $p + 1$ likelihood equations that are obtained by differentiating the log-likelihood function with respect to the $p + 1$ coefficients. The likelihood equations that result may be expressed as follows:

$$\sum_{i=1}^{n}[y_i - \pi(\mathbf{x}_i)] = 0$$

and

$$\sum_{i=1}^{n} x_{ij}[y_i - \pi(\mathbf{x}_i)] = 0$$

for $j = 1, 2, \ldots, p$.

As in the univariable model, the solution of the likelihood equations requires software that is available in virtually every statistical software package. Let $\hat{\boldsymbol{\beta}}$ denote the solution to these equations. Thus, the fitted values for the multiple logistic regression model are $\hat{\pi}(\mathbf{x}_i)$, the value of the expression in equation (2.2) computed using $\hat{\boldsymbol{\beta}}$ and $\mathbf{x}_i$.

In the previous chapter only a brief mention was made of the method for estimating the standard errors of the estimated coefficients. Now that the logistic regression model has been generalized, both in concept and notation to the multivariable case, we consider estimation of standard errors in more detail.

The method of estimating the variances and covariances of the estimated coefficients follows from well-developed theory of maximum likelihood estimation [see, e.g., Rao, (1973)]. This theory states that the estimators are obtained from the matrix of second partial derivatives of the log-likelihood function. These partial derivatives have the following general form

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i(1 - \pi_i) \tag{2.3}$$

and

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} \pi_i(1 - \pi_i) \tag{2.4}$$

for $j, l = 0, 1, 2, \ldots, p$ where π_i denotes $\pi(\mathbf{x}_i)$. Let the $(p + 1) \times (p + 1)$ matrix containing the negative of the terms given in equations (2.3) and (2.4) be denoted as $\mathbf{I}(\boldsymbol{\beta})$. This matrix is called the *observed information matrix*. The variances and

covariances of the estimated coefficients are obtained from the inverse of this matrix, which we denote as $\text{Var}(\boldsymbol{\beta}) = \mathbf{I}^{-1}(\boldsymbol{\beta})$. Except in very special cases it is not possible to write down an explicit expression for the elements in this matrix. Hence, we will use the notation $\text{Var}(\beta_j)$ to denote the jth diagonal element of this matrix, which is the variance of $\hat{\beta}_j$, and $\text{Cov}(\beta_j, \beta_l)$ to denote an arbitrary off-diagonal element, which is the covariance of $\hat{\beta}_j$ and $\hat{\beta}_l$. The estimators of the variances and covariances, which will be denoted by $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})$, are obtained by evaluating $\text{Var}(\boldsymbol{\beta})$ at $\hat{\boldsymbol{\beta}}$. We use $\widehat{\text{Var}}(\hat{\beta}_j)$ and $\widehat{\text{Cov}}(\hat{\beta}_j, \hat{\beta}_l)$, $j, l = 0, 1, 2, \ldots, p$ to denote the values in this matrix. For the most part, we only use the estimated standard errors of the estimated coefficients, which we denote as

$$\widehat{\text{SE}}(\hat{\beta}_j) = [\widehat{\text{Var}}(\hat{\beta}_j)]^{1/2} \tag{2.5}$$

for $j = 0, 1, 2, \ldots, p$. We use this notation in developing methods for coefficient testing and confidence interval estimation.

A formulation of the information matrix that is useful when discussing model fitting and assessment of fit is $\hat{\mathbf{I}}(\hat{\boldsymbol{\beta}}) = \mathbf{X}'\hat{\mathbf{V}}\mathbf{X}$ where $\mathbf{X}$ is an n by $p+1$ matrix containing the data for each subject and $\mathbf{V}$ is an n by n diagonal matrix with general element $\hat{\pi}_i(1-\hat{\pi}_i)$. That is, the matrix $\mathbf{X}$ is

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

and the matrix $\mathbf{V}$ is

$$\hat{\mathbf{V}} = \begin{bmatrix} \hat{\pi}_1\left(1-\hat{\pi}_1\right) & 0 & \cdots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & 0 & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix},$$

where $\hat{\pi}_i = \hat{\pi}(\mathbf{x}_i)$ is value of equation (2.2) using $\hat{\boldsymbol{\beta}}$ and the covariates of subject i, $\mathbf{x}_i$.

Before proceeding further, we present an example that illustrates the formulation of a multiple logistic regression model and the estimation of its coefficients using a subset of the variables from the data for the Global Longitudinal Study of Osteoporosis in Women (GLOW) study described in Section 1.6.3. The code sheet for the full data set is given in Table 1.7. As discussed in Section 1.6.3, one goal of this study is to evaluate risk factors for fracture during follow up.

The GLOW data set used in this text has information on 500 women, $n_1 = 125$ of whom had a fracture during the first year of follow up and $n_0 = 375$ who did not have a fracture. As an example, we consider five variables thought to be of importance that are age at enrollment (AGE), weight at enrollment (WEIGHT), history of a previous fracture (PRIORFRAC), whether or not the woman experienced

Table 2.2 Fitted Multiple Logistic Regression Model of Fracture in the First Year of Follow Up (FRACTURE) on Age, Weight, Prior Fracture (PRIORFRAC), Early Menopause (PREMENO), and Self-Reported Risk of Fracture (RATERISK) from the GLOW Study, $n = 500$

Variable	Coeff.	Std. Err.	z	p	95% CI
AGE	0.050	0.0134	3.74	<0.001	0.024, 0.076
WEIGHT	0.004	0.0069	0.59	0.556	−0.009, 0.018
PRIORFRAC	0.679	0.2424	2.80	0.005	0.204, 1.155
PREMENO	0.187	0.2767	0.68	0.499	−0.355, 0.729
RATERISK2	0.534	0.2759	1.94	0.053	−0.006, 1.075
RATERISK3	0.874	0.2892	3.02	0.003	0.307, 1.441
Constant	−5.606	1.2207	−4.59	<0.001	−7.998, −3.213

Log-Likelihood = −259.03768

menopause before or after age 45 (PREMENO) and self-reported risk of fracture relative to women of the same age (RATERISK) coded at three levels: less, same or more risk. In this example, the variable RATERISK is modeled using the two design variables in Table 2.1. The results of fitting the multiple logistic regression model to these data are shown in Table 2.2.

In Table 2.2 the estimated coefficients for the two design variables for RATERISK are indicated by RATERISK2 and RATERISK3. The estimated logit is given in the following equation:

$$\begin{aligned}\hat{g}(\mathbf{x}) = &-5.606 + 0.050 \times \text{AGE} + 0.004 \times \text{WEIGHT} \\ &+ 0.679 \times \text{PRIORFRAC} + 0.187 \times \text{PREMENO} \\ &+ 0.534 \times \text{RATERISK 2} + 0.874 \times \text{RATERISK 3}\end{aligned}$$

and the associated estimated logistic probabilities are found by using equation (2.2).

2.4 TESTING FOR THE SIGNIFICANCE OF THE MODEL

Once we have fit a particular multiple (multivariable) logistic regression model, we begin the process of model assessment. As in the univariable case presented in Chapter 1, the first step in this process is usually to assess the significance of the variables in the model. The likelihood ratio test for overall significance of the p coefficients for the independent variables in the model is performed in exactly the same manner as in the univariable case. The test is based on the statistic G given in equation (1.12). The only difference is that the fitted values, $\hat{\pi}$, under the model are based on the fitted model containing $p + 1$ parameters, $\hat{\boldsymbol{\beta}}$. Under the null hypothesis that the p "slope" coefficients for the covariates in the model are equal to zero, the distribution of G is chi-square with p degrees of freedom.

Consider the fitted model whose estimated coefficients are given in Table 2.2. For that model, the value of the log-likelihood, shown at the bottom of the table, is $L = -259.0377$. The log-likelihood for the constant only model may be obtained by evaluating the numerator of equation (1.13) or by fitting the constant only model. Either method yields the log-likelihood $L = -281.1676$. Thus the value of the likelihood ratio test is, from equation (1.12),

$$G = -2[-281.1676 - (-259.0377)] = 44.2598$$

and the p-value for the test is $P[\chi^2(6) > 44.2598] \leq 0.0001$, which is significant at well beyond the $\alpha = 0.05$ level. We reject the null hypothesis in this case and conclude that at least one or more of the p coefficients are different from zero, an interpretation analogous to the F-test used in multiple linear regression.

Before concluding that any or all of the coefficients are nonzero, we may look at the univariable Wald test statistics,

$$W_j = \frac{\hat{\beta}_j}{\widehat{\text{SE}}(\hat{\beta}_j)}.$$

These are shown in the fourth column, labeled z, in Table 2.2. Under the hypothesis that an individual coefficient is zero, these statistics will follow the standard normal distribution. The p-values computed under this hypothesis are shown in the fifth column of Table 2.2. If we use a level of significance of 0.05, then we would conclude that the variables AGE, history of prior fracture (PRIORFRAC) and self-reported rate of risk (RATERISK) are statistically significant, while WEIGHT and early menopause (PREMENO) are not significant.

As our goal is to obtain the best fitting model while minimizing the number of parameters, the next logical step is to fit a reduced model containing only those variables thought to be significant and compare that reduced model to the full model containing all of the variables. The results of fitting the reduced model are given in Table 2.3.

The difference between the two models is the exclusion of the variables WEIGHT and early menopause (PREMENO) from the full model. The likelihood

Table 2.3 Fitted Multiple Logistic Regression Model of Fracture in the First Year of Follow Up (FRACTURE) on AGE, Prior Fracture (PRIORFRAC), and Self-Reported Risk of Fracture (RATERISK) from the GLOW Study, $n = 500$

Variable	Coeff.	Std. Err.	z	p	95% CI
AGE	0.046	0.0124	3.69	<0.001	0.022, 0.070
PRIORFRAC	0.700	0.2412	2.90	0.004	0.228, 1.173
RATERISK2	0.549	0.2750	1.99	0.046	0.010, 1.088
RATERISK3	0.866	0.2862	3.02	0.002	0.305, 1.427
Constant	−4.991	0.9027	−5.53	<0.001	−6.760, −3.221

Log-Likelihood = −259.4494

ratio test comparing these two models is obtained using the definition of G given in equation (1.12). It has a distribution that is chi-square with 2 degrees of freedom under the hypothesis that the coefficients for both excluded variables are equal to zero. The value of the test statistic comparing the model in Table 2.3 to the one in Table 2.2 is

$$G = -2[-259.4494 - (-259.0377)] = 0.8324$$

which, with 2 degrees of freedom, has a p-value of $P[\chi^2(2) > 0.8324] = 0.663$. As the p-value is large, exceeding 0.05, we conclude that the full model is no better than the reduced model. That is, there is little statistical justification for including WEIGHT and PREMENO in the model. However, we must not base our models entirely on tests of statistical significance. As we discuss in Chapters 4 and 5, there are numerous other considerations that influence our decision to include or exclude variables from a model.

Whenever a categorical independent variable is included (or excluded) from a model, all of its design variables should be included (or excluded); to do otherwise implies that we have recoded the variable. For example, if we only include design variable D_1 as defined in Table 2.1, then the self-reported risk of fracture is entered into the model as a dichotomous variable coded as 0 (for less risk than others of the same age) and 1 (for the same or more risk than others of the same age). If k is the number of levels of a categorical variable, then the contribution to the degrees of freedom for the likelihood ratio test for the exclusion of this variable is $k - 1$. For example, if we exclude self-reported risk from the model and it is coded at three levels using the design variables shown in Table 2.1, then there are 2 degrees of freedom for the test, one for each design variable.

Because of the multiple degrees of freedom we must be careful in our use of the Wald (W) statistics to assess the significance of the coefficients. For example, if the W statistics for both coefficients exceed 2, then we could reasonably conclude that the design variables are significant. Alternatively, if one coefficient has a W statistic of 3.0 and the other a value of 0.1, then we cannot be sure about the contribution of the variable to the model. As both design variables for RATERISK are significant we can be fairly certain that the 2 degree of freedom test is also significant. We leave the details as an exercise, but for now it suffices to report that the $p < 0.001$ for the likelihood ratio test for the removal of RATERISK from the model in Table 2.3.

In the previous chapter we described, for the univariable model, two other tests equivalent to the likelihood ratio test for assessing the significance of the model: the Wald test and the Score test. At this point, we briefly discuss the multivariable versions of these tests, as their use appears occasionally in the literature. These tests are available in some software packages. For example, SAS computes both the likelihood ratio and score tests for a fitted model and STATA has the capability to easily perform the Wald test. For the most part we use likelihood ratio tests in this text because, as noted earlier, the quantities needed to carry it out may be obtained from all computer packages.

The multivariable analog of the Wald test is obtained from the following vector-matrix calculation:

$$\begin{aligned} W &= \hat{\boldsymbol{\beta}}'[\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})]^{-1}\hat{\boldsymbol{\beta}} \\ &= \hat{\boldsymbol{\beta}}'(\mathbf{X}\hat{\mathbf{V}}\mathbf{X})\hat{\boldsymbol{\beta}}, \end{aligned}$$

which is distributed as chi-square with $p + 1$ degrees of freedom under the hypothesis that each of the $p + 1$ coefficients is equal to zero. The multivariable Wald test, equivalent to the likelihood ratio test for the significance of the fitted model, is based on just the p slope coefficients and is obtained by eliminating $\hat{\beta}_0$ from $\hat{\boldsymbol{\beta}}$ and the relevant row (first or last) and column (first or last) from $(\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})$. As the evaluation of this test requires an extra step to perform vector-matrix operations and to obtain $\hat{\boldsymbol{\beta}}$, there is no gain over the likelihood ratio test for determining the significance of the model. Extensions of the Wald test that can be used to examine functions of the coefficients are quite useful and are illustrated in subsequent chapters. The value of the multivariable Wald test for the fitted model in Table 2.3 is $W = 39.88$, which, with 4 degrees of freedom, corresponds to $p < 0.001$. Hence, both the likelihood ratio test and the Wald test reject the hypothesis that the model is not significant. In this particular example the value of the multivariable Wald test is smaller than the likelihood ratio test, but this is not always the case.

The multivariable analog of the Score test for the significance of the model is based on the distribution of the p derivatives of $L(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. The computation of this test is of the same order of complication as the Wald test. To define it in detail would require introduction of additional notation that would find little use in the remainder of this text. Thus, we refer the interested reader to Cox and Hinkley (1974) or Dobson (2002). We do note that the score test is computed by some statistical packages (e.g., the logistic procedure in SAS).

2.5 CONFIDENCE INTERVAL ESTIMATION

We discussed confidence interval estimators for the coefficients, the logit and the logistic probabilities for the univariable logistic regression model in Section 1.4. The methods used for confidence interval estimators for a multivariable model are essentially the same.

The endpoints for a $100(1 - \alpha)\%$ Wald-based confidence interval for the coefficients are obtained from equation (1.15) for slope coefficients and from equation (1.16) for the constant term. For example, using the fitted model presented in Table 2.3, the 95 percent confidence interval for the coefficient of AGE is

$$0.046 \pm 1.96 \times 0.0124 = (0.022, 0.070),$$

which are exactly the values in the last column of Table 2.3, labeled "95% Conf. Int.". The interpretation of this interval is that we are 95 percent confident that the

increase in the log-odds per one-year increase in age is between 0.022 and 0.070. As we noted in Section 1.4 many software packages (e.g., STATA) automatically provide confidence intervals for all model coefficients in the output. Confidence intervals for the other coefficients shown in Table 2.3 are calculated in a similar manner. We also calculated the profile likelihood confidence interval estimator discussed at the end of Section 1.4 for each of the variables in Table 2.3, and they differed from their respective Wald-based confidence intervals at most by 0.3 percent and as a result, are not shown.

The confidence interval estimator for the logit is a bit more complicated for the multiple variable model than the result presented in equation (1.19). The basic idea is the same; only there are now more terms involved in the summation. It follows that a general expression for the estimator of the logit for a model containing p covariates is

$$\hat{g}(\mathbf{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p. \tag{2.6}$$

An alternative way to express the estimator of the logit in equation (2.6) is through the use of vector notation as $\hat{g}(\mathbf{x}) = \mathbf{x}'\hat{\boldsymbol{\beta}}$, where the vector $\hat{\boldsymbol{\beta}}' = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)$ denotes the estimator of the $p+1$ coefficients and the vector $\mathbf{x}' = (x_0, x_1, x_2, \ldots, x_p)$ represents a set of values of the p-covariates in the model and the constant, $x_0 = 1$.

It follows from equation (1.18) that an expression for the estimator of the variance of the estimator of the logit in equation (2.6) is

$$\widehat{\text{Var}}[\hat{g}(\mathbf{x})] = \sum_{j=0}^{p} x_j^2 \widehat{\text{Var}}(\hat{\beta}_j) + \sum_{j=0}^{p} \sum_{k=j+1}^{p} 2x_j x_k \widehat{\text{Cov}}(\hat{\beta}_j, \hat{\beta}_k). \tag{2.7}$$

We can express this result much more concisely by using the matrix expression for the estimator of the variance of the estimator of the coefficients. From the expression for the observed information matrix, we have that

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})^{-1}. \tag{2.8}$$

It follows from equation (2.8) that an equivalent expression for the estimator in equation (2.7) is

$$\begin{aligned} \widehat{\text{Var}}[\hat{g}(\mathbf{x})] &= \mathbf{x}'\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})\mathbf{x} \\ &= \mathbf{x}'(\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})^{-1}\mathbf{x}. \end{aligned} \tag{2.9}$$

Fortunately, all good logistic regression software packages provide the option for the user to create a new variable containing the estimated values of equation (2.9) or the standard error for all observed values of the covariates of subjects in the data set. This feature eliminates the computational burden associated with the matrix calculations in equation (2.9) and allows the user to routinely calculate fitted values and confidence interval estimates. However, it is useful to illustrate the details of the calculations.

Using the model in Table 2.3, the estimated logit for a 65-year-old woman with a prior fracture (PRIORFRAC = 1) who thinks that her risk is the same as other women of her age is

$$\begin{aligned}\hat{g}(\text{AGE} = 65, \text{PRIORFRAC} = 1, \text{RATERISK} = 2) &= -4.991 + 0.046 \times 65 + 0.700 \times 1 + 0.549 \times 1 + 0.866 \times 0 \\ &= -0.752\end{aligned}$$

and the estimated logistic probability is

$$\hat{\pi}(\text{AGE} = 65, \text{PRIORFRAC} = 1, \text{RATERISK} = 2) = \frac{e^{-0.752}}{1 + e^{-0.752}} = 0.320.$$

The interpretation of this fitted value is that the estimated proportion of 65-year-old women with a prior fracture, who rate their risk of fracture as the same as women of their age having a facture in the next year is 0.320.

In order to use equation (2.7) to estimate the variance of this estimated logit we need the estimated covariance matrix, which is shown in Table 2.4. The expression for the estimated variance of the logit is

$$\begin{aligned}&\widehat{\text{Var}}[\hat{g}(\text{AGE} = 65, \text{PRIORFRAC} = 1, \text{RATERISK} = 2)] \\ &= \widehat{\text{Var}}(\hat{\beta}_0) + (65)^2 \times \widehat{\text{Var}}(\hat{\beta}_1) + (1)^2 \times \widehat{\text{Var}}(\hat{\beta}_2) + (1)^2 \times \widehat{\text{Var}}(\hat{\beta}_3) + 2 \times 65 \\ &\quad \times \widehat{\text{Cov}}(\hat{\beta}_0, \hat{\beta}_1) + 2 \times 1 \times \widehat{\text{Cov}}(\hat{\beta}_0, \hat{\beta}_2) + 2 \times 1 \times \widehat{\text{Cov}}(\hat{\beta}_0, \hat{\beta}_3) + 2 \times 65 \times 1 \\ &\quad \times \widehat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_2) + 2 \times 65 \times 1 \times \widehat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_3) + 2 \times 1 \times 1 \times \widehat{\text{Cov}}(\hat{\beta}_2, \hat{\beta}_3),\end{aligned}$$

which when evaluated using the values in Table 2.4 is

$$\begin{aligned}&\widehat{\text{Var}}[\hat{g}(\text{AGE} = 65, \text{PRIORFRAC} = 1, \text{RATERISK} = 2)] \\ &= 0.81487 + (65)^2 \times 0.00015 + 1 \times 0.05816 + 1 \times 0.07563 \\ &\quad + 2 \times 65(-0.01089) + 2 \times 1 \times 0.04450 + 2 \times 1 \times (-0.06039) + 2 \times 65 \\ &\quad \times 1 \times (-0.00083) + 2 \times 65 \times 1 \times 0.00022 + 2 \times 1 \times 1 \times (-0.00313) \\ &= 0.04937.\end{aligned}$$

Table 2.4 Estimated Covariance Matrix of the Estimated Coefficients in Table 2.3

	AGE	PRIORFRAC	RATERISK2	RATERISK3	Constant
AGE	0.00015				
PRIROFRAC	−0.00083	0.05816			
RATERISK2	0.00022	−0.00313	0.07563		
RATERISK3	0.00054	−0.01184	0.04624	0.08191	
Constant	−0.01089	0.04450	−0.06039	−0.08055	0.81487

The standard error is

$$\widehat{\text{SE}}[\hat{g}(\text{AGE} = 65, \text{PRIORFRAC} = 1, \text{RATERISK} = 2)] = \sqrt{0.04937} = 0.22220$$

and the 95 percent confidence interval for the estimated logit is

$$-0.752 \pm 1.96 \times 0.22220 = (-1.18751, -0.31648).$$

The associated confidence interval for the fitted value is (0.234, 0.422). We defer discussion and interpretation of the estimated logit, fitted values and their respective confidence intervals until Chapter 3.

2.6 OTHER ESTIMATION METHODS

In Section 1.5, we discussed the discriminant function estimators of the coefficients of the logistic regression model and note here that it may also be employed in the multivariable case. This approach to estimation of the logistic regression coefficients is based on the assumption that the distribution of the independent variables, given the value of the outcome variable, is multivariate normal. Two points should be kept in mind: (i) the assumption of multivariate normality is rarely, if ever, satisfied in practice because of the frequent occurrence of categorical independent variables, and (ii) the discriminant function estimators of the coefficients for non-normally distributed independent variables, especially dichotomous variables, will be biased away from zero when the true coefficient is nonzero. For these reasons, in general, we do not recommend the use of this method. However, these estimators are of historical importance as a number of the classic papers in the applied literature [such as Truett et al. (1967)] used them. These estimators are easily computed and in the absence of a logistic regression program, could be used as a first approximation to parameter estimates. Thus, it seems worthwhile to include the relevant formulae for their computation. An exception to the general recommendation is when the focus is on the effect of a single continuous variable and all other variables in the model are there for adjustment, a concept we discuss in the next chapter. In this special setting Lyles et al. (2009) show how one may compute the discriminant function estimator of this single coefficient through an easily performed linear regression.

Specifically, the assumptions for the discriminant function approach are that the conditional distribution of $\mathbf{X}$ (the vector of p covariate random variables) given the outcome variable, $Y = y$, is multivariate normal with a mean vector that depends on y, but a covariance matrix that does not. Using notation defined in Section 1.5 we have that $(\mathbf{X}|y = j) \sim N(\mu_j, \Sigma)$ where μ_j contains the means of the p independent variables for the subpopulation defined by $y = j$ and Σ is the $p \times p$ covariance matrix of these variables. Under these assumptions, $\Pr(Y = 1|\mathbf{x}) = \pi(\mathbf{x})$, where the coefficients are given by:

$$\beta_0 = \ln\left(\frac{\theta_1}{\theta_0}\right) - 0.5(\mu_1 - \mu_0)'\Sigma^{-1}(\mu_1 + \mu_0) \tag{2.10}$$

and

$$\boldsymbol{\beta} = (\mu_1 - \mu_0)' \Sigma^{-1}, \tag{2.11}$$

where $\theta_1 = \Pr(Y = 1)$ and $\theta_0 = 1 - \theta_1$ denote the proportion of the population with y equal to 1 or 0, respectively. Equations (2.10) and (2.11) are the multivariable analogs of equations (1.23) and (1.24).

The discriminant function estimators of β_0 and $\boldsymbol{\beta}$ are found by substituting estimators for μ_j, $j = 0, 1$, Σ, and θ_1 into equations (2.10) and (2.11). The estimators most often used are the maximum likelihood estimators under the multivariate normal model. That is, we let

$$\hat{\mu}_j = \overline{\mathbf{x}}_j,$$

the mean of $\mathbf{x}$ in the subgroup of the sample with $y = j$, $j = 0, 1$.

The estimator of the covariance matrix, Σ, is the multivariable extension of the pooled sample variance given in Section 1.5. This may be represented as

$$\mathbf{S} = \frac{(n_0 - 1)\mathbf{S}_0 + (n_1 - 1)\mathbf{S}_1}{(n_0 + n_1 - 2)},$$

where $\mathbf{S}_j$, $j = 0, 1$ is the $p \times p$ matrix of the usual unbiased estimators of the variances and covariances computed within the subgroup defined by $y = j$, $j = 0, 1$.

Because of the bias in the discriminant function estimators when normality does not hold, they should be used only when logistic regression software is not available, and then only in preliminary analyses. Any final analyses should be based on the maximum likelihood estimators of the coefficients.

EXERCISES

1. In Section 2.4 we stated, but did not provide details for, the likelihood ratio test for the addition of weight and early menopause to the model containing AGE, prior fracture (PRIORFRAC) and self-reported risk (RATERISK).

(a) Using the GLOW500 data and a logistic regression package verify the values of the coefficients for the models shown in Table 2.2 and Table 2.3.

(b) Perform the likelihood ratio test comparing these two models [i.e., the test for the contribution of WEIGHT and early menopause (PREMENO) to a model containing AGE, prior fracture (PRIORFRAC) and self-reported risk (RATERISK)].

2. Use the ICU data described in Section 1.6.1 and consider the multiple logistic regression model of vital status, STA, on age (AGE), cancer part of the present problem (CAN), CPR prior to ICU admission (CPR), infection probable at ICU admission (INF), and race (RACE).

(a) The variable race is coded at three levels. Prepare a table showing the coding of the two design variables necessary for including this variable in a logistic regression model.

(b) Write down the equation for the logistic regression model of STA on AGE, CAN, CPR, INF, and RACE. Write down the equation for the logit transformation of this logistic regression model. How many parameters does this model contain?

(c) Write down an expression for the likelihood and log-likelihood for the logistic regression model in Exercise 2(b). How many likelihood equations are there? Write down an expression for a typical likelihood equation for this problem.

(d) Using a logistic regression package, obtain the maximum likelihood estimates of the parameters of the logistic regression model in Exercise 2(b). Using these estimates write down the equation for the fitted values (i.e., the estimated logistic probabilities).

(e) Using the results of the output from the logistic regression package used in Exercise 2(d), assess the significance of the slope coefficients for the variables in the model using the likelihood ratio test. What assumptions are needed for the p-values computed for this test to be valid? What is the value of the deviance for the fitted model?

(f) Use the Wald statistics to obtain an approximation to the significance of the individual slope coefficients for the variables in the model. Fit a reduced model that eliminates those variables with nonsignificant Wald statistics. Assess the joint (conditional) significance of the variables excluded from the model. Present the results of fitting the reduced model in a table.

(g) Using the results from Exercise 2(f), compute 95 percent confidence intervals for all coefficients in the model. Write a sentence interpreting the confidence intervals for the nonconstant covariates.

(h) Obtain the estimated covariance matrix for the final model fit in Exercise 2(f). Choose a set of values for the covariates in that model and estimate the logit and logistic probability for a subject with these characteristics. Compute 95 percent confidence intervals for the logit and estimated logistic probability. Write a sentence or two interpreting the estimated probability and its confidence interval.

3. Use the Myopia Study data described in Section 1.6.6 and use MYOPIC as the outcome and as possible variables for a model: AGE, GENDER, family history of myopia (MOMMY and DADMY), number of hours playing sports (SPORTHR) and number of hours watching television (TVHR).

(a) Repeat parts 2(b)–2(h) of Exercise 2.

(b) Verify that there is little difference between the Wald-based and profile likelihood intervals for the variables in the model in part 3(a).