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

CHAPTER 2

Multiple Logistic Regression

2.1 INTRODUCTION

In the previous chapter we introduced the logistic regression model in the univariate context. As in the case of linear regression, the strength of a modeling technique lies in its ability to model many variables, some of which may be on different measurement scales. In this chapter we will generalize the logistic model to the case of more than one independent variable. This will be referred to as the "multivariable case." Central to the consideration of multiple logistic models will be estimation of the coefficients in the model and testing for their significance. This will follow along the same lines as the univariate model. An additional modeling consideration which will be introduced in this chapter is the use of design variables for modeling discrete, nominal scale independent variables. In all cases it will be assumed that there is a predetermined collection of variables to be examined. The question of variable selection is dealt with in Chapter 4.

2.2 THE MULTIPLE LOGISTIC REGRESSION MODEL

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

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p, \tag{2.1}$$

in which case the logistic regression model is

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

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.

Most logistic regression software will generate design variables, and some programs have a choice of several different methods. The different strategies for creation and interpretation of design variables are discussed in detail in Chapter 3.

In general, if a nominal scaled variable has k possible values, then $k-1$ design variables will be needed. This is true since, unless stated otherwise, all of our models have a constant term. To illustrate the notation used for design variables in this text, suppose that the j^{th} 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 $\beta_{jl}, l = 1,2,...,k_j - 1$. Thus, the logit for a model with p vari-

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

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

ables and the j^{th} variable being discrete would be

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

When discussing the multiple logistic regression model we will, in general, suppress the summation and double subscripting needed to indicate when design variables are being used. The exception to this will be the discussion of modeling strategies when we need to use the specific value of the coefficients for any design variables in the model.

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 univariate 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 will be the same as in the univariate 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 (2.2). 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:

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

and

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

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

As in the univariate model, the solution of the likelihood equations requires special software that is available in most, if not all, statistical packages. 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, for example, 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 $j, l = 0,1,2,...,p$ 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 j^{th} 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 $\hat{\text{Var}}(\hat{\boldsymbol{\beta}})$, are obtained by evaluating $\text{Var}(\boldsymbol{\beta})$ at $\hat{\boldsymbol{\beta}}$. We will use $\hat{\text{Var}}(\hat{\beta}_j)$ and $\hat{\text{Cov}}(\hat{\beta}_j, \hat{\beta}_l)$, $j, l = 0, 1, 2,..., p$ to denote the values in this matrix.

For the most part, we will have occasion to use only the estimated standard errors of the estimated coefficients, which we will denote as

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

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

A formulation of the information matrix which will be useful when discussing model fitting and assessment of fit is $\hat{\mathbf{I}}\left(\hat{\boldsymbol{\beta}}\right) = \mathbf{X}'\mathbf{V}\mathbf{X}$ 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 $\hat{\pi}_i\left(1-\hat{\pi}_i\right)$. That is, the matrix $\mathbf{X}$ is

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

and the matrix $\mathbf{V}$ is

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

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 low birth weight study described in Section 1.6.2. The code sheet for the full data set is given in Table 1.6. As discussed in Section 1.6.2, the goal of this study was to identify risk factors associated with giving birth to a low birth weight baby (weighing less than 2500 grams). Data were collected on 189 women, $n_1 = 59$ of whom had low birth weight babies and $n_0 = 130$ of whom had normal birth weight babies. Four variables thought to be of importance were age, weight of the mother at her last menstrual period, race, and number of physician visits during the first trimester of the pregnancy. In this example, the variable race has been

Table 2.2 Estimated Coefficients for a Multiple Logistic Regression Model Using the Variables AGE, Weight at Last Menstrual Period (LWT), RACE, and Number of First Trimester Physician Visits (FTV) from the Low Birth Weight Study

Variable	Coeff.	Std. Err.	z	P>|z|
AGE	–0.024	0.0337	–0.71	0.480
LWT	–0.014	0.0065	–2.18	0.029
RACE_2	1.004	0.4979	2.02	0.044
RACE_3	0.433	0.3622	1.20	0.232
FTV	–0.049	0.1672	–0.30	0.768
Constant	1.295	1.0714	1.21	0.227

Log likelihood = –111.286

recoded using the two design variables in Table 2.1. The results of fitting the 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 race are indicated by RACE_2 and RACE_3. The estimated logit is given by the following expression:

$$\hat{g}(\mathbf{x}) = 1.295 - 0.024 \times AGE - 0.014 \times LWT + 1.004 \times RACE\_2 + 0.433 \times RACE\_3 - 0.049 \times FTV.$$

The fitted values are obtained using the estimated logit, $\hat{g}(\mathbf{x})$.

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 univariate 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 univariate 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 vector 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 will be 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 $L = -111.286$. 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 $L = -117.336$. Thus the value of the likelihood ratio test is, from equation (1.12),

$$G = -2\left[(-117.336) - (-111.286)\right] = 12.099$$

and the p-value for the test is $P\left[\chi^2(5) > 12.099\right] = 0.034$ which is significant at the $\alpha = 0.05$ level. We reject the null hypothesis in this case and conclude that at least one and perhaps all p coefficients are different from zero, an interpretation analogous to that in multiple linear regression.

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

$$W_j = \hat{\beta}_j / \hat{\mathrm{SE}}\left(\hat{\beta}_j\right).$$

These are given in the fourth column in Table 2.2. Under the hypothesis that an individual coefficient is zero, these statistics will follow the standard normal distribution. The p-values are given 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 LWT and possibly RACE are significant, while AGE and FTV are not significant.

If 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 it to the full model containing all 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 AGE and FTV from the full model. The likelihood ratio test comparing these two models is obtained using the definition of G given in equation (1.12). It will have a distribution that is chi-square with 2 degrees-of-freedom under the hypothesis that the coefficients for the

Table 2.3 Estimated Coefficients for a Multiple Logistic Regression Model Using the Variables LWT and RACE from the Low Birth Weight Study

Variable	Coeff.	Std. Err.	z	P>|z|
LWT	−0.015	0.0064	−2.36	0.018
RACE_2	1.081	0.4881	2.22	0.027
RACE_3	0.481	0.3567	1.35	0.178
Constant	0.806	0.8452	0.95	0.340

Log likelihood = −111.630

variables excluded are equal to zero. The value of the test statistic comparing the models in Tables 2.2 and 2.3 is

$$G = -2[(-111.630) - (-111.286)] = 0.688,$$

which, with 2 degrees-of-freedom, has a p-value of $P[\chi^2(2) > 0.688] = 0.709$. Since the p-value is large, exceeding 0.05, we conclude that the reduced model is as good as the full model. Thus there is no advantage to including AGE and FTV in the model. However, we must not base our models entirely on tests of statistical significance. As we will see in Chapter 5, there are numerous other considerations that will 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 race is entered into the model as a dichotomous variable coded as black or not black. 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 will be $k-1$. For example, if we exclude race from the model, and race is coded at three levels using the design variables shown in Table 2.1, then there would be 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 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. The estimated coefficients for the variable RACE in Table 2.3 provide a good example. The Wald statistic for the coefficient for the first design variable is 2.22, and 1.35 for the second. The likelihood ratio test comparing the model containing LWT and RACE to the one containing only LWT yields

$$G = -2\left[-(114.345) - (-111.630)\right] = 5.43,$$

which, with 2 degrees-of-freedom, yields a p-value of 0.066. Strict adherence to the $\alpha = 0.05$ level of significance would justify excluding RACE from the model. However, RACE is known to be a "clinically important" variable. In this case the decision to include or exclude RACE should be made in conjunction with subject matter experts.

In the previous chapter we described, for the univariate model, two other tests equivalent to the likelihood ratio test for assessing the significance of the model, the Wald and Score tests. We will briefly discuss the multivariable versions of these tests, as their use appears occasionally in the literature. These tests are available in some software packages. SAS computes both the likelihood ratio and score tests for a fitted model and STATA has the capability to perform the Wald test easily. For the most part we will use likelihood ratio tests in this text. As noted earlier, we favor the likelihood ratio test as 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:

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

which will be distributed as chi-square with $p+1$ degrees-of-freedom under the hypothesis that each of the $p+1$ coefficients is equal to zero. Tests for just the p slope coefficients are obtained by eliminating $\hat{\beta}_0$ from $\hat{\boldsymbol{\beta}}$ and the relevant row (first or last) and column (first or last) from $(\mathbf{X}'\mathbf{V}\mathbf{X})$. Since evaluation of this test requires the capability to perform vector-matrix operations and to obtain $\hat{\boldsymbol{\beta}}$, there is no gain over the likelihood ratio test of the significance of the model. Extensions of the Wald test which can be used to examine functions of the coefficients

are quite useful and are illustrated in subsequent chapters. In addition, the modeling approach of Grizzle, Starmer, and Koch (1969), noted earlier, contains many such examples.

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 which would find little use in the remainder of this text. Thus, we refer the interested reader to Cox and Hinkley (1974) or Dobson (1990).

2.5 CONFIDENCE INTERVAL ESTIMATION

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

The endpoints for a $100(1-\alpha)\%$ confidence interval for the coefficients are obtained from (1.4.1) for slope coefficients and from (1.4.2) for the constant term. For example, using the fitted model presented in Table 2.3, the 95 percent confidence interval for LWT is

$$-0.015 \pm 1.96 \times 0.0064 = (-0.028,\ -0.002).$$

The interpretation of this interval is that we are 95 percent confident that the decrease in the log-odds per one pound increase in weight of the mother is between −0.028 and −0.002. As we noted in Section 1.4 many software packages automatically provide confidence intervals for all model coefficients in the output.

The confidence interval estimator for the logit is a bit more complicated for the multiple variable model than the result presented in (1.19). The basic idea is the same, only there are now more terms involved in the summation. It follows from (2.1) 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 (2.6) is through the use of vector notation as $\hat{g}(\mathbf{x}) = \mathbf{x}'\hat{\boldsymbol{\beta}}$, where the vector

$\hat{\boldsymbol{\beta}}' = \left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p\right)$ denotes the estimator of the $p+1$ coefficients and the vector $\mathbf{x}' = \left(x_0, x_1, x_2, \ldots, x_p\right)$ represents the constant and a set of values of the p-covariates in the model, where $x_0 = 1$.

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

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

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

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

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

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

Fortunately, all good logistic regression software packages provide the option for the user to create a new variable containing the estimated values of (2.9) or the standard error for all subjects in the data set. This feature eliminates the computational burden associated with the matrix calculations in (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 150 pound white woman is

$$\begin{aligned} \hat{g}(LWT = 150, RACE = White) &= 0.806 - 0.015 \times 150 + 1.081 \times 0 + 0.481 \times 0 \\ &= -1.444 \end{aligned}$$

and the estimated logistic probability is

$$\hat{\pi}(LWT = 150, RACE = White) = \frac{e^{-1.444}}{1 + e^{-1.444}} = 0.191.$$

The interpretation of the fitted value is that the estimated proportion of low birthweight babies among 150 pound white women is 0.191.

In order to use (2.7) to estimate the variance of this estimated logit we need to obtain the estimated covariance matrix shown in Table 2.4. Thus the estimated variance of the logit is

$$\begin{aligned}
\hat{\text{Var}}[\hat{g}(LWT = 150, RACE = White)] &= \hat{\text{Var}}\left(\hat{\beta}_0\right) + (150)^2 \times \hat{\text{Var}}\left(\hat{\beta}_1\right) + \\
&(0)^2 \times \hat{\text{Var}}\left(\hat{\beta}_2\right) + (0)^2 \times \hat{\text{Var}}\left(\hat{\beta}_3\right) + 2 \times 150 \times \hat{\text{Cov}}\left(\hat{\beta}_0, \hat{\beta}_1\right) \\
&+ 2 \times 0 \times \hat{\text{Cov}}\left(\hat{\beta}_0, \hat{\beta}_2\right) + 2 \times 0 \times \hat{\text{Cov}}\left(\hat{\beta}_0, \hat{\beta}_3\right) + 2 \times 150 \times 0 \times \hat{\text{Cov}}\left(\hat{\beta}_1, \hat{\beta}_2\right) \\
&+ 2 \times 150 \times 0 \times \hat{\text{Cov}}\left(\hat{\beta}_1, \hat{\beta}_3\right) + 2 \times 0 \times 0 \times \hat{\text{Cov}}\left(\hat{\beta}_2, \hat{\beta}_3\right) \\
&= 0.7143 + (150)^2 \times 0.000041 + 0 \times 0.2382 + 0 \times 0.1272 \\
&+ 2 \times 150 \times (-0.0052) + 2 \times 0 \times 0.0226 + 2 \times 0 \times (-0.1035) \\
&+ 2 \times 150 \times 0 \times (-0.000647) + 2 \times 150 \times 0 \times 0.000036 \\
&+ 2 \times 0 \times 0 \times 0.0532 = 0.0768
\end{aligned}$$

and the standard error is $\hat{\text{SE}}[\hat{g}(LWT = 150, RACE = White)] = 0.2771$. The 95 percent confidence interval for the estimated logit is

$$-1.444 \pm 1.96 \times 0.2771 = (-1.988, \ -0.901).$$

The associated confidence interval for the fitted value is (0.120, 0.289). We defer further discussion and interpretation of the estimated logit, fitted values and their respective confidence intervals until Chapter 3.

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

	LWT	RACE_2	RACE_3	Constant
LWT	0.000041			
RACE_2	−0.000647	0.2382		
RACE_3	0.000036	0.0532	0.1272	
Constant	−0.005211	0.0226	−0.1035	0.7143

2.6 OTHER METHODS OF ESTIMATION

In Section 1.5, two alternative methods of estimating the parameters of the logistic regression model were discussed. These were the methods of noniteratively weighted least squares and discriminant function. Each may also be employed in the multivariable case, though application of the noniteratively weighted least squares estimators is limited by the need for nonzero estimates of $\pi(\mathbf{x})$ for most values of $\mathbf{x}$ in the data set. With a large number of independent variables, or even a few continuous variables, this condition is not likely to hold. The discriminant function estimators do not have this limitation and may be easily extended to the multivariable case.

The discriminant function approach to estimation of the logistic 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: (1) the assumption of multivariate normality will rarely if ever be satisfied because of the frequent occurrence of dichotomous independent variables, and (2) the discriminant function estimators of the coefficients for nonnormally distributed independent variables, especially dichotomous variables, will be biased away from zero when the true coefficient is nonzero. For these reasons we, in general, do not recommend its use. However, these estimators are of some historical importance as a number of the classic papers in the applied literature, such as Truett, Cornfield, and Kannel (1967), have used them. These estimators are easily computed and, in the absence of a logistic regression program, should be adequate for a preliminary examination of your data. Thus, it seems worthwhile to include the relevant formulae for their computation.

The assumptions necessary to employ the discriminant function approach to estimating the logistic regression coefficients state that the conditional distribution of $\boldsymbol{X}$ (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 say $\mathbf{X} \mid y = j \sim N(\mu_j, \Sigma_j)$ where μ_j contains the means of the p independent variables for the subpopulation defined by $y = j$ and $\boldsymbol{\Sigma}$ is the $p \times p$ covariance matrix of these variables. Under these assumptions, $P(Y = 1 \mid \mathbf{x}) = \pi(\mathbf{x})$, 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 $\theta_1 = P(Y = 1)$ 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 (1.22) and (1.23).

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

$$\hat{\mu}_j = \bar{\mathbf{x}}_j$$

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

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

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. 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 1(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 1(b). Using these estimates write down the equation for the fitted values, that is, the estimated logistic probabilities.

(e) Using the results of the output from the logistic regression package used in Exercise 1(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 1(f), compute 95 percent confidence intervals for all coefficients in the model. Write a sentence interpreting the confidence intervals for the non-constant covariates.

(h) Obtain the estimated covariance matrix for the final model fit in Exercise 1(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.

2. Use the Prostate Cancer data described in Section 1.6.3 and consider the multiple logistic regression model of capsule penetration (CAPSULE),

on AGE, RACE, results of the digital rectal exam (DPROS and DCAPS), prostate specific antigen (PSA), Gleason score (GLEASON) and tumor volume (VOL).

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

(b) The variable DCAPS is coded 1 and 2. Can this variable be used in its original coding or must a design variable be created? Explore this question by comparing the estimated coefficients obtained from fitting a model containing DCAPS as originally coded with those obtained from one using a 0–1 coded design variable, $DCAPSnew = DCAPS - 1$.

(c) Repeat parts 1(b) – 1(h) of Exercise 1.