Chapter 10

Logistic Regression

- Dichotomous outcomes
- Problems with use of multiple regression
- Logistic regression model
- Statistical inference
- Model fit
- Classification of individuals

The modeling of outcomes with multiple regression in previous chapters has assumed that the outcome of interest is measured on an interval scale. However, many outcomes of interest in the behavioral and social sciences are dichotomous or binary in nature. Some common examples include pass/fail results on mastery testing, voting for or against the incumbent in an election, a choice between two questionnaire options, survival or failure of a business, completion or noncompletion of school, and recovery or nonrecovery from a potentially fatal disease. The most common modeling approach for such dichotomous outcomes is *logistic regression* analysis.

Logistic regression is very similar to multiple regression in many ways. For example, both techniques are potentially used for two purposes, to gain better understanding of phenomena and/or to make predictions (classifications) for individuals. Also, the basic steps in logistic regression analysis are very similar to those in multiple regression. To review, these steps are:

- Case analysis for individual observations,
- Assessment of the validity of assumptions,
- Test of the overall relationship,
- Description of the effect of each IV on the outcome, and
- Assessment of the accuracy of classification

There are also some fundamental differences between logistic regression and multiple regression. Most of these differences result directly from the consideration of a dichotomous rather than an interval outcome. For example, in this chapter we will be modeling the probability of each of the two possible outcomes rather than describing the relationship between IV's and changes in an interval outcome. Some of the familiar indices from multiple regression will also disappear. For example, there will be no R^2-like index provided to characterize the overall strength of the relationship in logistic regression.

After presenting an example for illustration in Section 10.1, the problems associated from attempting to use standard multiple regression for dichotomous outcomes will be described in Section 10.2. Logistic regression models for different types of IV's are discussed in Section 10.3, followed by consideration of the associated statistical inference in Section 10.4. The assessment of model fit, including case analysis and testing of the model fit, is presented in Section 10.5.

Finally, use of the logistic regression equation for the classification of individuals is illustrated in Section 10.6.

10.1 An Example

Consider a hypothetical eight-week summer remedial program with individualized, computer-based instruction in reading and mathematics for middle school students. Program administrators have been concerned about the relatively large number of students who fail to complete the program. As part of the effort to better understand and address the problem, the program administrators hired a graduate student from a local university to conduct a study identifying factors that predict whether a student will complete the voluntary remedial program. Such information might be useful for two purposes. Greater insight into why students in general complete the program would inform attempts to improve the program. Also, an equation that would accurately predict risk of failing to complete the program could be used to identify individuals who might benefit from special attention at the start of the program.

The sample of students studied by the graduate student researcher consisted of all students who started in the remedial program last summer. Available evidence supported the assumption that this group of students was representative of students who will enter the program in future years. The measured outcome was whether a student completed the eight-week program. For an initial model of the probability of completion of the program, the researcher included the following independent variables

- student academic aptitude as measured by a standardized test given in the schools,
- a dichotomous indicator of whether or not the student received a "good citizen" award in school the previous year,
- age at the beginning of the remedial program, and
- socio-economic status of the student's family (SES assessed as low, medium, or high).

Basic sample statistics for the hypothetical sample of 160 students included the following results. Of the 160 students who entered last summer's

program, 90 completed the program, while 70 dropped out at some point before completion. One hundred and fifteen students of the 160 received the citizenship award, and frequencies in the low, medium, and high SES categories were 59, 51, and 50, respectively. The mean (and standard deviation) for aptitude were 50.1 (9.87), and the same statistics for age were 10.9 (0.28). Inspection of scatterplots and cross tabulation tables for the various variables did not suggest any obvious data problems.

10.2 Problems with Multiple Regression for Dichotomous Outcomes

At first glance, it may appear that with multiple regression we already have the appropriate tool to model the probability of a student completing the program for the current example. We would code the dichotomous outcome variable of program completion and regress this coded variable on the independent variables of interest. Unfortunately, this approach would result in some serious difficulties. To illustrate the problems associated with the use of multiple regression for dichotomous dependent variables, we will briefly consider here an analysis of the current example data based on the single IV of aptitude (labeled X), assuming that the outcome, Y, is coded 1 for "completed" and 0 for "not completed."

The scatterplot of Y versus X is shown in Figure 10.1. The appearance of the plot has changed from that in previous scatterplots for continuous dependent variables since there are now only two possible values for Y. Use of multiple regression to regress Y on X results in the equation $\hat{Y} = -0.385 + 0.0189\ X$. When the corresponding prediction line is superimposed on the scatterplot in Figure 10.1, a problem becomes immediately apparent. We would like to interpret a predicted value of Y for a given value of X as the estimated probability of a student completing the program. This is not possible in the current case though because some of the predicted Y values are not legitimate probabilities, i.e. some are negative (for small X values) and some are larger than one (for large X values). Moreover, consideration of the conditional distributions of Y for specific values of X in Figure 10.1 indicates clear violations of the normality and constant variance assumptions associated

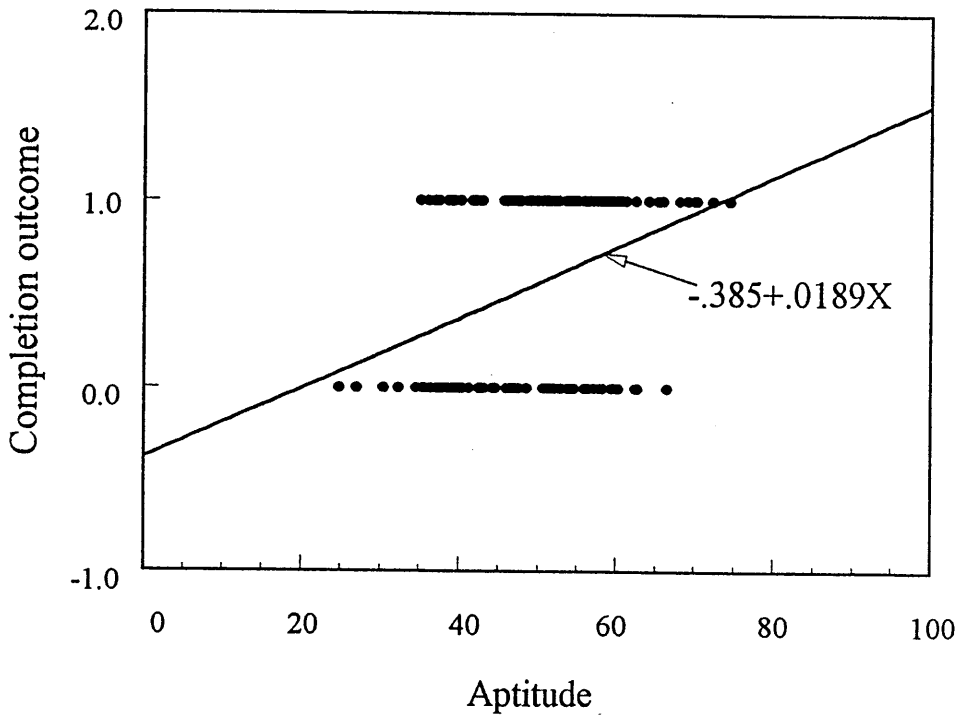

Figure 10.1: Scatterplot of the program completion outcome (Y=1 for completed and Y=0 for not completed) versus student aptitude. The line shown is the best fitting simple linear regression line.

---

with multiple regression. These violations would result in estimates that are not maximally efficient (i.e., have larger standard errors than necessary).

These difficulties associated with the use of multiple regression for dichotomous outcomes are resolved when logistic regression is used to model the outcome.

10.3 Logistic Regression Models

Before introducing the logistic regression model, it is useful to define some basic concepts for dichotomous outcomes using only hand computations. A nonparametric approach to the determination of the probability of program completion is illustrated in the next section.

10.3.1 Basic concepts with a nonparametric approach

In our search for an alternative to multiple regression for modeling the probability of a student completing the program, consider the following possible common-sense approach. Table 10.1 was constructed by first grouping the students in the sample on the basis of the X variable, aptitude. The proportion of students in each aptitude group completing the program is presented in the table as a function of the midpoints of the aptitude groupings. For a given aptitude group with midpoint X_i, the proportion of students who complete the program is an estimate of the probability of completing the program for a single random student with aptitude X_i , a probability denoted $\pi(X_i)$,. For example, a student with a relatively low aptitude of 34 is estimated to have a probability of only 0.125 of completing the program, while a student with the higher aptitude of 54 is estimated to have a probability of 0.562 of completion.

The Y = 1 probability results in Table 10.1 also imply the probability of a student *not* completing the program (i.e., the probability of Y = 0), because the latter probability is equal to 1 - $\pi(X_i)$. For example, the probability of not completing for the aptitude group with midpoint of 54 is 1-0.562 or 0.438. These probabilities of failure to complete the program are also shown in Table 10.1.

Table 10.1
Frequencies and Proportions of Program Completions as a Function of Aptitude[a]

Aptitude Midpoint	Program not completed (Y = 0)	Program completed (Y = 1)	Total frequency	Odds of completion
26	2 (1.000)	0 (0.000)	2	0
30	2 (1.000)	0 (0.000)	2	0
34	7 (0.875)	1 (0.125)	8	0.143
38	9 (0.562)	7 (0.438)	16	0.779
42	9 (0.600)	6 (0.400)	15	0.667
46	15 (0.556)	12 (0.444)	27	0.799
50	6 (0.273)	16 (0.727)	22	2.663
54	7 (0.438)	9 (0.562)	16	1.283
58	9 (0.310)	20 (0.690)	29	2.226
62	3 (0.333)	6 (0.667)	9	2.003
66	1 (0.143)	6 (0.857)	7	5.993
70	0 (0.000)	5 (1.000)	5	infinity
74	0 (0.000)	2 (1.000)	2	infinity

[a] Proportions are in parentheses.

The *odds* of program completion for a given aptitude are defined as the ratio of the probability of program completion to the probability of noncompletion ($\pi/[1-\pi]$). For example, for the aptitude group with midpoint of 54, the odds of program completion are 0.562/0.438 = 1.283. The odds results for all aptitude groups using the nonparametric approach are shown in Table 10.1. As would be expected from the earlier results, the odds of program completion tend to increase with increasing aptitude.

The effect of an IV on a dichotomous outcome is usually represented with an *odds ratio* (ψ) that is defined as a ratio of the odds of Y = 1 for two different values of the IV. For the current example with an interval IV, we would be interested in the odds ratio for some given increase c in the aptitude. Consider, for example, the aptitude groups in Table 10.1 with aptitude midpoints of 34 and 54, two groups that have a difference in aptitude of 20 units. The effect of this aptitude change of 20 units on the odds of program completion is then expressed with the ratio of the appropriate odds in Table 10.1, i.e., with the ratio of 1.283/0.143 = 8.97. Thus, the odds of completion increase by a factor of 8.97 when the aptitude increases from 34 to 54. For the current nonparametric approach, this estimate of the effect of a 20 unit aptitude change would vary somewhat depending on which aptitude groups were chosen for the computation, but under the logistic model described later the odds ratio for a given change c is constant across the range of the IV.

In summary, a nonparametric approach based on grouping of subjects with respect to the interval IV can be used to estimate how the probability of Y = 1 and the corresponding odds vary with the interval IV. The ratio of odds for two different values of the IV is used to represent the effect of the IV on the outcome.

10.3.2 Logistic regression models with one interval IV

A weakness of the nonparametric approach described above is the relatively large number of parameters that must be estimated from the data (a proportion for every aptitude category). If one assumes that the underlying population value of the probability of Y = 1 changes in a regular and continuous fashion as a function of aptitude, a parametric model with a small number of parameters offers a more parsimonious alternative approach. The logistic model, defined below, provides such an approach.

The logistic model for the probability of Y = 1 for the ith subject ($\pi[X_i]$) is

$$\pi(X_i) = \frac{\exp(g[X_i])}{1 + \exp(g[X_i])} \qquad (10.1)$$

where $g[X_i]$ is the *logit* for the ith subject having a value of $X = X_i$. The expression $\exp(g[X_i])$ represents the constant e (e = 2.718) raised to the power $g[X_i]$ and varies from zero for very large negative values of g to plus infinity for very large positive values of g. Exponentiation (i.e., the raising of e to a power) can be accomplished with most pocket calculators. For example, if the logit for a subject is 1.5, a pocket calculator indicates that exp(1.5) = 4.48, and the probability of Y = 1 is then 4.48/(1+4.48) = 0.82.

The logistic model for π as a function of the logit is represented graphically in Figure 10.2. This figure can be viewed as "universal." Every model that we consider in later sections, no matter how complex, results in a single logit for each individual, a logit that in turn can be used with (10.1) or Figure 10.2 to indicate the probability of Y = 1 for that individual. Note that a logit of zero will always result in a probability of 0.5 that Y = 1, with negative logit values giving lower probabilities and positive logits producing higher probabilities.

For a simple logistic regression model with only one IV, X, the logit $g[X_i]$ for the ith subject is expressed as a linear function of X, i.e.,

$$g[X_i] = \beta_0 + \beta_1 X_i \qquad (10.2)$$

Thus, the model for the logit in logistic regression is identical in appearance to the multiple regression model. The combination of (10.1) and (10.2) represents the logistic regression model for the probability of Y = 1 as a function of the IV, X.

Interval IV example. Using the data described in Section 10.1 and assuming a model with the single IV of aptitude, a logistic

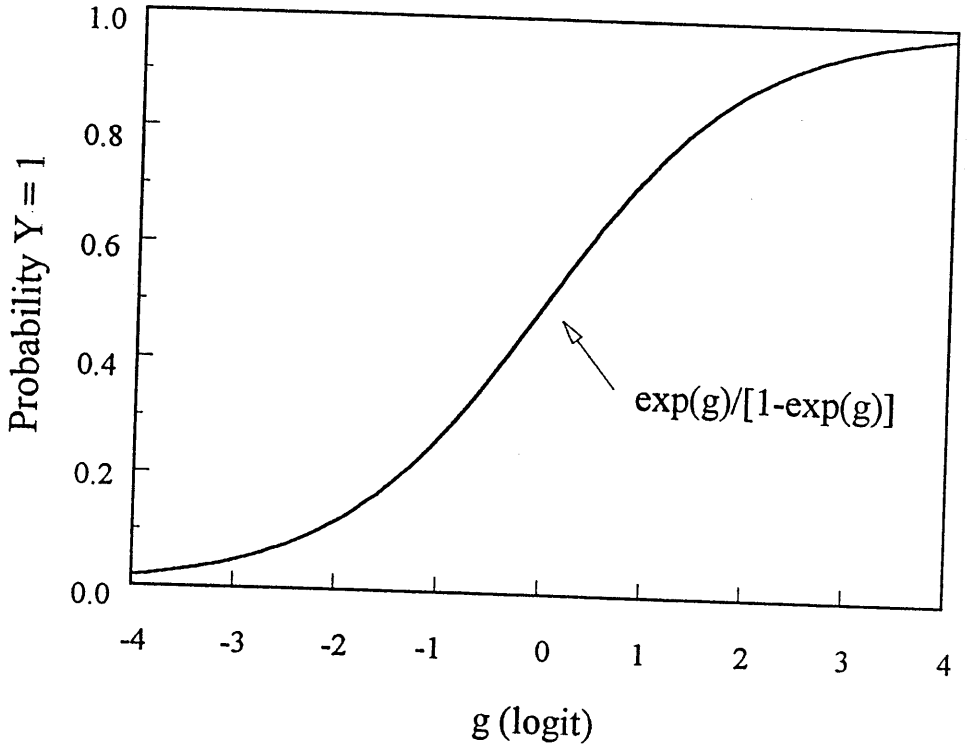

Figure 10.2: The logistic function (Equation 10.1) shows the variation of the probability of Y = 1 as a function of the logit.

regression computer program gives the estimated parameters of (10.2) shown below.

$$\hat{g} = -4.106 + 0.0878\, X_i$$

Substitution of this expression into (10.1) results in the final estimated logistic model. For example, a student with an aptitude of 38 will have an estimated logit of

$$\begin{aligned}\hat{g} &= -4.106 + 0.0878\ (38) \\ &= -0.770\end{aligned}$$

which is then substituted into (10.1) to obtain

$$\hat{\pi} = \frac{\exp(-0.770)}{1 + \exp(-0.770)}$$
$$= \frac{0.463}{1 + 0.463}$$
$$= 0.316$$

This computation can be repeated for various X values to obtain the continuous representation of the logistic model shown in Figure 10.3. It is seen that a 0.5 probability of program completion ($\pi = 0.5$) occurs at an aptitude of approximately 47, with higher aptitudes resulting in greater probability.

As noted in the previous section, the *odds* of Y = 1 for a given value of X are defined as the ratio of the probability of Y = 1 to the probability of Y = 0, i.e.,

$$odds\ (X_i) = \frac{\pi(X_i)}{1 - \pi(X_i)} \qquad (10.3)$$

The odds for a given X_i can also be determined by simply exponentiating the associated logit, i.e.,

$$odds\ (X_i) = \exp(g[X_i]) \qquad (10.4)$$

This latter expression can be derived by substituting (10.1) into (10.3) and simplifying. If one takes the log of both sides of (10.4), it can be seen why the logit, g, is sometimes called the "log-odds."

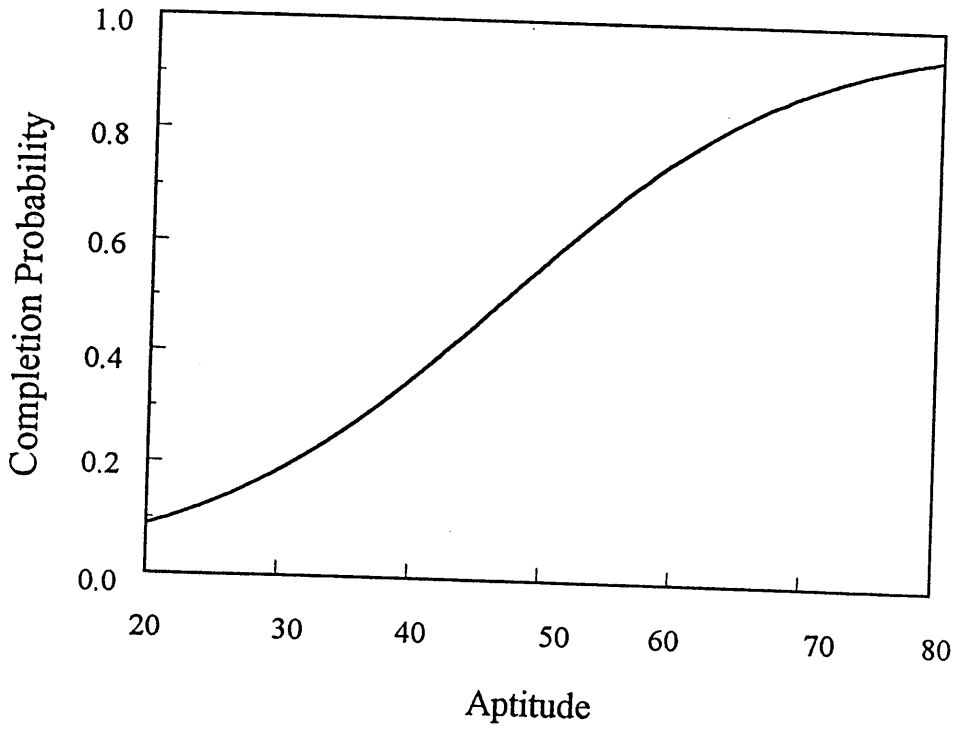

Figure 10.3: The probability of program completion is expressed as a function of student aptitude.

Example continued. For the student with an aptitude of 38 and a logit, as noted above, of -0.770, the odds of program completion are exp(-0.770) = 0.463. The same result, within rounding error, can also be found with the definition in (10.3).

Computation of the odds for different X values results in the continuous odds function shown in Figure 10.4. Note that odds of 1 (a 0.5 probability of program completion) will always occur when the logit is equal to zero. From (10.2), it is seen that in general a logit of zero occurs when the X is equal to $-\beta_0/\beta_1$. For the current example, this aptitude value is -(-4.106)/0.0878 = 46.8.

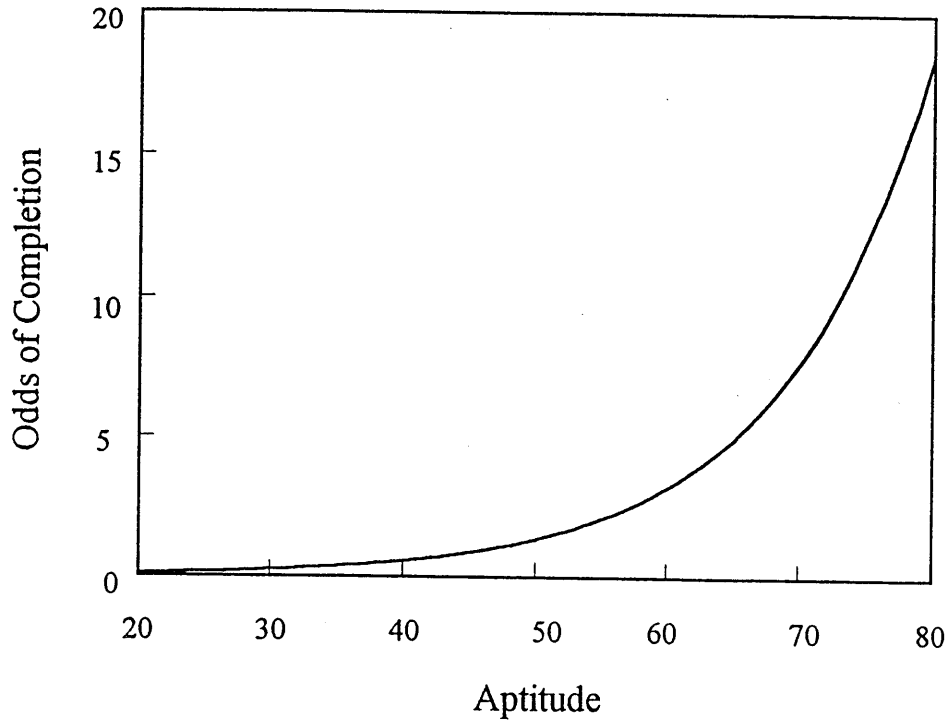

Figure 10.4: The odds of program completion (i.e., the probability of completing the program divided by the probability of not completing) increase sharply with increasing aptitude.

The effect of an interval IV, X, on Y in logistic regression is represented by the *odds ratio*, ψ(c), defined as the change in the odds of Y = 1, expressed in ratio form, which is associated with an increase in X of some specified c units, i.e.,

$$\psi(c) = \frac{odds\ (X+c)}{odds\ (X)} \qquad (10.5)$$

The odds ratio can also be computed by exponentiating the product of c and the coefficient of X, i.e.,

$$\psi(c) = \exp(c\,\beta_1) \qquad (10.6)$$

There are several features of the expression in (10.6) to be noted. First, when the β_1 coefficient is zero, the odds ratio is equal to one, reflecting no change in the odds associated with a change in X. If the coefficient is positive, as with the current example, the odds ratio is larger than one, indicating an increase in the odds. A negative coefficient, on the other hand, results in a ratio less than one, reflecting a decrease in the odds of Y = 1. The odds ratio is never negative. Finally, note that the change in odds due to an increase of c units in X is constant across the range of X.

Example continued. For the IV of aptitude, a choice of 20 for c would correspond to a change of aptitude of approximately two standard deviations. The associated odds ratio is exp([20][.0878]) = 5.79. That is, an increase in aptitude of 20 units is associated with an increase of a factor of 5.79 in the odds of program completion.

10.3.3 Logistic regression models with one dichotomous IV

The modeling of a dichotomous outcome with IVs that are also dichotomous in logistic regression is very similar to the approach discussed previously for multiple regression. For each dichotomous IV, the X is coded 1 for all in the first level of the dichotomous IV and 0 for those in the second level. Once estimates of the coefficients in (10.2) are obtained, substitution of this equation with X equal to either 1 or 0 into (10.1) results in two probabilities of Y = 1, one for each of the two groups. The odds of Y = 1 for the two groups are obtained either with the basic definition of odds in (10.3) or with the exponentiation of the logit shown in (10.4). Finally, the ratio of the odds of Y = 1 for those in the first level of the dichotomous IV (X = 1) to the odds for those in the second level (X = 0) is

$$\psi(\textit{level 1 v. level 2}) = \exp(\beta_1) \qquad (10.7)$$

This equation is a special case of (10.5) where c = 1 for the change in X from 0 to 1.

Dichotomous IV example. In the current example with program completion as the outcome, consider a model consisting only of the single IV of citizenship. Assume that X is coded 1 for those who received a good citizenship award for the previous school year and 0 for those who did not receive the grade. The frequencies of students in the sample for different combinations of the two values of Y and X are shown in Table 10.2. Using logistic regression software, the estimated equation for the logit in the example is

$$\hat{g}[X_i] = -1.253 + 2.079\,X_i$$

For students who received the citizenship award (X = 1), the logit is -1.253 + (2.079)(1) = 0.826 and the corresponding estimated probability of program completion is 0.70. A similar computation for X = 0 indicates that the logit for students not receiving the citizenship award is the intercept in the equation above, -1.253, and the resulting probability of program completion is 0.22. Using (10.4), the odds of program completion for those receiving the citizenship award is exp(0.826) = 2.28, while the odds for those not receiving the award is exp(-1.253) = 0.286. The odds ratio reflecting the effect of the citizenship award, using (10.6), is ψ = exp(2.079) = 8.00. Thus, the odds of program completion for students receiving the citizenship award are 8.00 times greater than the odds for those not receiving the award.

Table 10.2
Dichotomous IV Example:
Program Completion and Citizenship[a]

	Citizenship award		
Program completion	No award	Award	Total
Not completed	35 (77.8)	35 (30.4)	70
Completed	10 (22.2)	80 (69.6)	90
Total	45	115	160

[a] Percentages shown in parentheses are based on column totals.

If one wishes to reverse the direction of the contrast associated with a dichotomous IV for the purpose of interpretation (i.e., consider the contrast of level 2 versus level 1), the corresponding odds ratio is equal to the reciprocal of the original odds ratio. That is, ψ(level 2 v. level 1) = 1/[ψ(level 1 v. level 2)]. In the example above, the odds of program completion for students *not* receiving the citizenship award are 1/8.0 = 0.105 times those for students who did receive the award.

In the current example with a single dichotomous IV all of the logistic regression results discussed above could also be obtained with simple hand computations involving the frequencies and proportions shown in Table 10.2. For example, the probability of Y = 1 for those receiving the citizenship award (X = 1) is 80/115 = 0.70. Computation of the other results is left to the reader as an exercise. Once multiple IVs are included in the model, hand computation of the logistic regression results for a dichotomous IV is no longer be possible.

10.3.4 Logistic regression models with one categorical IV

Representation of one or more categorical IVs in a logistic model is also identical to that in multiple regression. Each categorical IV with m levels is represented by a set of m-1 coded variables. Many different coding schemes can be used, but the most convenient for consideration of all pairwise comparisons is the 1,0 coding used in previous chapters. Thus, the first variable in the set is coded 1 for those in the first level and zero otherwise, the second variable in the set is coded 1 for those in the second level and zero otherwise, etc.. After the coefficients for the m-1 IVs are estimated, the logit, probability of Y = 1, and odds of Y = 1 can be obtained for each level of the categorical IV by substitution of the appropriate coding. The odds ratios for the control group comparisons (i.e., the comparisons of each level with the last or "control" level) are then given by

$$\begin{aligned} \psi(\textit{level } 1 \textit{ v. level } m) &= \exp(\beta_1) \\ \psi(\textit{level } 2 \textit{ v. level } m) &= \exp(\beta_2) \\ \ldots &= \ldots \end{aligned} \qquad (10.8)$$

The odds ratios for noncontrol group contrasts can be found with the general expression below for the contrast of the jth group versus the kth group.

$$\psi(\textit{level } j \textit{ v. level } k) = \exp(\beta_j - \beta_k) \qquad (10.9)$$

Odds ratios for all pairwise comparisons of the levels of each categorical IV would be determined with (10.8) and (10.9).

Categorical IV example. Consider the program completion outcome modeled with the single categorical IV of SES having the three levels of low, medium, and high. Frequencies for different combinations of Y and SES are shown in Table 10.3. Two coded variables are created, X_1 coded 1 for students with families in the low SES category and X_2 coded 1 for students with families in the

medium SES category. Using logistic regression software, the estimated equation for the logit is

$$\hat{g} = 0.406 - 0.783 X_1 + 0.470 X_2$$

Substituting appropriate coding into this equation results in logits for low SES ($X_1 = 1$, $X_2 = 0$), medium SES ($X_1 = 0$, $X_2 = 1$), and high SES ($X_1 = X_2 = 0$) of -0.377, 0.876, and 0.406, respectively. Substitution of these logits into (10.1) results in probabilities of program completion of 0.407, 0.706, and 0.600 for the three SES levels. Thus, students in the medium and high SES categories have the highest probabilities of program completion. Exponentiating the logits (Equation 10.4), the odds of program completion for the three SES levels are 0.686, 2.401, and 1.501, indicating that the probability of program completion is less than 0.5 for students in the low SES category.

Using (10.8), the odds ratios for the two control group comparisons are:

$$\begin{aligned} \psi(\textit{low SES v. high SES}) &= \exp(-0.783) \\ &= 0.457 \\ \psi(\textit{medium SES v. high SES}) &= \exp(0.470) \\ &= 1.600 \end{aligned}$$

The odds ratio associated with the single non-control group contrast is, using (10.9),

$$\begin{aligned} \psi(\textit{low SES v. medium SES}) &= \exp([-0.783] - [0.470]) \\ &= 0.286 \end{aligned}$$

It is seen that the odds of completion for low SES students is only 0.457 times the odds for the high SES students and 0.286 times the odds for medium SES students. The odds of completion for

the medium SES students appear to be greater by a factor of 1.600 than the odds for the high SES students, but this difference will be shown later to be nonsignificant.

Table 10.3
Categorical IV Example:
Program Completion and SES[a]

	SES			
Program completion	Low	Medium	High	Total
Not completed	35 (59.3)	15 (29.4)	20 (40.0)	70
Completed	24 (40.7)	36 (70.6)	30 (60.0)	90
Total	59	51	50	160

[a] Percentages shown in parentheses are based on column totals.

If one wishes to reverse any of the contrasts above, the corresponding odds ratio is inverted. For example, the contrast of high SES versus low SES would have an odds ratio of 1/0.457 = 2.188, i.e., the odds of completion for high SES students is 2.188 times those for low SES students.

As with the dichotomous IV example, these logistic regression results for a *single* categorical IV can also be obtained with simple hand computations using the frequencies in Table 10.3. This is left again as an exercise for the reader.

10.3.5 Logistic regression models with multiple IVs

Multiple IVs can be included in a logistic regression model for the purposes discussed earlier for multiple regression, i.e., to provide statistical control in estimating the effect of each IV on the outcome and to improve the

precision of the prediction of the outcome. The general model for k IV's consists of (10.1) and the following equation for the logit.

$$g_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} \qquad (10.10)$$

As with multiple regression, the IV's can also be used to represent functional complexities. For example, the use of powers of an interval X in (10.10) would allow the probability of passing a test to first increase with increasing levels of anxiety and then to decrease as anxiety levels become very high (see the example of polynomial regression in Chapter 5). Also, interactive relationships in which the odds ratio associated with one IV depends on the level of another IV could be fit using product terms in (10.10).

To illustrate, consider the complete example first discussed in Section 10.1 with the following variables:

- Y - Remedial program completion (yes = 1, no = 0)
- X_1 - Aptitude (interval)
- X_2 - Citizenship ("good citizen" award = 1, no award = 0)
- X_3 - Age in years at program entry (interval)
- SES of student family (low, medium, and high)
  - X_4 (coded 1 for low SES, 0 otherwise)
  - X_5 (coded 1 for medium SES, 0 otherwise)

The coefficient estimates, with other results, are given in Table 10.4. The estimates for a given IV will in general be different from those obtained in earlier parts of Section 10.3 because of the presence of control variables. The general procedures and interpretations however remain the same. Suppose, for example, that we wish to predict the odds of program completion for a student with the following values of the IVs: aptitude of 55, a good citizenship award, age of 10.5 years at program entry, and a family in the medium SES category. The estimated logit, using the coefficients from Table 10.4, is

$$\hat{g} = -22.41 + (.138)(55) + (3.061)(1) + (1.302)(10.5) + (-1.9889)(0) + (.496)(1)$$
$$= 2.408$$

resulting in an estimated probability of program completion of π = exp[2.408]/(1+exp[2.408]) = 0.917. The associated odds of completion are exp(2.408) = 11.10.

Table 10.4
Selected Logistic Regression Example Results

Variable	Coefficient	S.E.	z statistic	p-value
X_1	0.138	0.028	4.9	<0.001
X_2	3.061	0.572	5.4	<0.001
X_3	1.302	0.793	1.6	0.101
X_4	-1.989	0.579	-3.4	0.001
X_5	0.496	0.565	0.9	0.380
constant	-22.41			

The interpretation of the effect of each of the IVs on Y is the same as before, with only the addition of the qualifier "controlling for all other IVs." To illustrate for each of the variables:

- *Aptitude:* When aptitude is increased by 20 units (approximately two standard deviations), the odds of program completion increase by a factor of exp([20][.138]) = 15.8, controlling for other IVs.
- *Citizenship*: The odds of program completion for students receiving the citizenship award are exp(3.061) = 21.3 times greater than the odds for those not receiving the award, controlling for other IVs.
- *Age:* When age is increased by one year, the odds of program completion increase by a factor of exp([1][1.302]) = 3.7, controlling for other IVs.

- *SES:* The following odds ratios describe the pairwise comparisons of interest:
  - The odds of program completion for those in the low SES group are exp(-1.989) = 0.14 times the odds for those in the high SES group, controlling for other IVs.
  - The odds of program completion for those in the medium SES group are exp(0.496) = 1.64 times the odds for those in the high SES group, controlling for other IVs.
  - The odds of program completion for those in the low SES group are exp([-1.989]-[0.496]) = 0.08 times the odds for those in the medium group, controlling for other IVs.

10.3.6 Logistic regression assumptions

There are three formal assumptions associated with logistic regression. Two are identical to assumptions in multiple regression. Specifically, the observations are assumed to be independent and the independent variables are assumed to be measured exactly. Conclusions about robustness to violation of these assumptions are similar to those given in Chapter 2 for multiple regression, i.e., the logistic regression procedure is not robust to violations of the independence assumption caused by, say, nonrandom sampling, contextual effects, or repeated measures, and the bias introduced by measurement error in the IV's is a matter of degree.

Logistic regression also assumes that the combination of the logistic function in (10.1) and the specified equation for the logit in (10.10) correctly describes the true functional form of the relationship between the probability of Y = 1 and the IV's. This assumption is analogous to the correct fit assumption in multiple regression. The robustness of the logistic technique to violations of this assumption will be a matter of degree, with increasingly problematic distortions of reality resulting from increasing departures from the correct functional form. The test of goodness of model fit described in Section 10.5 is used to attempt to identify violations of this assumption.

There is also, in effect, a fourth assumption in logistic regression. The inferential procedures have been derived as "asymptotic theory," which means that the derivation assumed that sample sizes are very large. Since the validity of the inferential procedure is only ensured for very large sample sizes, in practice one should be aware that nominal error rates may not be accurate for analyses based on small sample sizes.

10.4 Statistical Inference

The rationale underlying the estimation of the logistic regression coefficients is briefly described in Section 10.4.1, followed by presentation of a procedure for testing the overall relationship and sets of IV's in Section 10.4.2. Statistical inference for the effects of individual IV's is discussed in Section 10.4.3.

10.4.1 Model parameter estimation

The principle of *maximum likelihood* is used for the estimation of the coefficients in logistic regression. Although the mathematical details of this approach are beyond the scope of this discussion, we will briefly consider the conceptual nature of maximum likelihood estimation. An expression called the *likelihood function* states the probability or likelihood of the observed data in a study as a function of the unknown logistic model parameters. This is a continuous function with a different likelihood of observed data for each of the infinite number of possible combinations of model parameters. The *maximum likelihood estimates* (MLE's) of the model parameters (e.g., those reported for our example in Table 10.4) are those optimal values that maximize the likelihood of the observed data.

In general, it is not possible to obtain a "closed form" solution for the MLE's, but instead an iterative solution is required. Computer programs will typically provide some information on the number of iterative cycles required to achieve convergence and the final solution.

10.4.2 Testing the overall relationship and sets of IVs

The test of the overall logistic relationship and the test of the effect of a subset of IVs can both be conducted with the *likelihood ratio (LR) test*. Since the LR test is analogous to the partial F test of ΔR^2 for a subset of IVs in multiple regression, a brief review of the multiple regression procedure may be useful. The hypothesis tested by the partial F test is H_o: $\Delta R^2(X_h) = 0$, where X_h represents a set of hypothesis variables added to the reduced model with control variables. Since an increase in R^2 can also be expressed as a decrease in the residual sums of squares (SSE) due to the hypothesis variables, this null hypothesis can also be stated H_o: $\Delta SSE(X_h) = 0$. A rejection of this null hypothesis indicates that addition of the hypothesis variables to a model already containing the control variables has produced a significant reduction in the lack of fit of the model.

For logistic regression, the quantity representing model lack of fit is the model *Deviance* (D), with decreasing values of D reflecting improving fit. The Deviance is inversely related to the likelihood (LH) that was maximized in obtaining the MLEs, with increasing values of LH (between 0 and 1) resulting in decreasing values of D. The quantity -2lnLH provided in most computer programs (-2 times the natural log of LH) also reflects lack of model fit, with larger values indicating poorer fit.

The null hypothesis tested in the likelihood ratio (LR) test for logistic regression is that the decrease in model Deviance due to adding the hypothesis variables is equal to zero, i.e.,

$$H_0 : \Delta D(X_h) = 0 \qquad (10.11)$$

An equivalent form of this hypothesis states that all coefficients of the h hypothesis IVs are simultaneously equal to zero, i.e.,

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_h = 0$$

The LR test statistic, G^2, for this hypothesis is the difference between the Deviance for the reduced model ($D[X_r]$) and that for the full model ($D[X_f]$), which is also equal to the corresponding difference in the -2ln LH result provided in most computer output, i.e.,

$$\begin{aligned} G^2 &= D[X_r] - D[X_f] \\ &= [-2 \ln LH(X_r)] - [-2 \ln LH(X_f)] \end{aligned} \qquad (10.12)$$

The G^2 statistic is distributed as a chi-square variate with degrees of freedom h, where h is the number of hypothesis variables. Critical values are given in Table C-6 of Appendix C.

Examples. In using the LR test to test the *overall relationship*, the reduced model consists only of an intercept with no IVs, while the full model contains all of the IVs in the final model. The values of D or -2lnLH obtained for each of these models from computer results is then used for the test. For the current program completion example, -2lnLH for the model with intercept only is 219.3, that for the full model is 135.4, and the difference is $G^2 = 83.9$. Comparison of this computed statistic with a critical value of 11.07 for degrees of freedom 5 and $\alpha = 0.05$ indicates a significant overall relationship.

The *test of a subset of variables* with the LR test will now be illustrated. The categorical IV of SES (low, medium, and high) has been represented in our current example with two coded IVs, X_4 and X_5. The test of the effect of SES then is a test of the variable set of X_4 and X_5, controlling for X_1 through X_3. Using a forced order of entry approach, the control variables are first entered into the model for the reduced model. Then the hypothesis variables, X_4 and X_5, are entered, and the decrease in -2lnLH determined and tested. For the current example, -2lnLH

for the reduced model is 160.3 and that for the full model, as noted above, is 135.4. The difference of $G^2 = 24.9$, when compared to the critical value of 5.99 for degrees of freedom 2 and $\alpha = 0.05$, indicates a significant effect of SES on the probability of program completion. The associated p value from the printout is < 0.0001.

In principle, the LR test can also be used to test the effects of each of the multiple individual IVs. Multiple runs would be required, with each one forcing a different variable in last. In practice, however, the more convenient test discussed in the next section is often used instead.

Indices of overall strength of relationship. Unfortunately, in logistic regression there is no universally accepted index of overall strength of relationship similar to the R^2 index in multiple regression. "R^2- like" indices have been proposed but are argued to be problematic. As another possible alternative, the use of classification hit rates (see Section 10.6) is also potentially misleading except when the classification of individuals is an intended purpose. At this time, it appears that the description of the strength of the relationship represented by a logistic regression equation must be based primarily on the collective description of the effects of each of the individual IV's.

10.4.3 Tests and interval estimates for individual effects

Hypothesis tests. The standard error of the MLE for the β_j coefficient in the logistic regression model, $s_{\hat{\beta}_j}$, can be used to compute the following z statistic:

$$z = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}} \qquad (10.13)$$

For the test of H_o: $\beta_j = 0$, this statistic is compared to the appropriate percentile in the standard normal distribution. Thus, for example, the critical value, $z(\alpha)$, for $\alpha = 0.05$ would be 1.96. This test is equivalent to the Wald test provided by many computer packages. The Wald statistic is distributed as a chi-square variate with one degree of freedom and is equal to the square of the z statistic in (10.13).

The test of an individual effect with the z statistic in (10.13) or the Wald statistic is not as powerful as the test of the same effect using the LR test discussed in Section 10.4.2. If the test of a specific effect in the model is of particular importance in the study, it would be advisable to use a forced order of entry and LR test for that effect.

For those individual effects represented by a difference of two coefficients (e.g., a noncontrol group contrast for a categorical IV), the test procedure is similar to that used in multiple regression. For example, for the noncontrol group contrast of the jth level versus the kth level in a categorical IV, the difference of the two coefficients (β_j-β_k) must be tested. The z statistic is obtained with

$$z = \frac{(\hat{\beta}_j - \hat{\beta}_k)}{S_{\hat{d}}} \qquad (10.14)$$

where the standard error of the difference is

$$S_{\hat{d}} = \sqrt{S^2_{\hat{\beta}_j} + S^2_{\hat{\beta}_k} - 2S_{\hat{\beta}_j \hat{\beta}_k}} \qquad (10.15)$$

Either the required covariance of the two coefficients or the associated correlation must be given by the computer program. If the correlation is provided, the covariance is determined by multiplying the correlation by the two associated standard errors. To test H_o: β_j-$\beta_k = 0$, the z statistic is again compared with $z(\alpha)$.

Example computations. For the program completion example, the standard errors for all of the coefficients are given in Table

10.4. To illustrate, the test for the effect of aptitude (β_1) at $\alpha = 0.05$ compares $z = 0.138/0.028 = 4.93$ with 1.96, resulting in the decision to reject the null hypothesis. The corresponding p values shown in Table 10.4 are from the same computer results. Other test results in the table indicate that the effect of the dichotomous IV of citizenship (β_2) is also significant but that the effect of student age (β_3) is not significant. For the categorical IV of SES, the low versus high contrast (β_4) is significant but the medium versus high contrast (β_5) is not significant.

To illustrate the test of a noncontrol group contrast, consider the low versus medium contrast in the SES categorical IV represented by β_4-β_5. A computer printout gives the correlation between the two associated coefficients as 0.463. Using the standard errors shown in Table 10.4, the covariance for the same coefficients is computed to be (0.463)(0.579)(0.565) = 0.151. Substituting into (10.15), the standard error for the coefficient difference is

$$S_d = \sqrt{0.579^2 + 0.565^2 - (2)(0.151)}$$
$$= 0.594$$

and the z statistic is $z = (-1.989-0.496)/0.594 = 4.18$. Comparison with $z(0.05) = 1.96$ indicates that the low versus medium SES contrast is statistically significant.

Interval estimates. Since odds ratios (ψ's) are used to represent the effects of IV's in logistic regression, it is important to compute interval estimates of these odds ratios. For an *interval IV*, the confidence interval (with 1-α confidence level) on the odds ratio associated with an increase of c units in X_j is

$$CI(\psi) = \exp\left(c\,\hat{\beta}_j \pm z(\alpha)\,c\,S_{\hat{\beta}_j}\right) \qquad (10.16)$$

For *dichotomous IVs* or *control group contrasts* in categorical IV's, the confidence interval is

$$CI(\psi) = \exp(\hat{\beta}_j \pm z(\alpha) S_{\hat{\beta}_j}) \qquad (10.17)$$

For a *noncontrol group contrast* between the jth and kth levels in a categorical IV's, the confidence interval is

$$CI(\psi) = \exp([\hat{\beta}_j - \hat{\beta}_k] \pm z(\alpha) S_d) \qquad (10.18)$$

where the standard error is computed with (10.15). The usual interpretation of an interval applies, i.e., in the long run the interval will have (1-α) probability of capturing the true odds ratio. The resulting interval can also be used to test the effect. If the interval does not contain 1 (the odds ratio reflecting no effect), the null hypothesis is rejected. If the interval does contain 1, the decision is fail to reject.

Example computations. To illustrate the computation of a confidence interval for an *interval IV*, consider the effect of a 20 unit increase in the aptitude IV. Results from Table 10.4 can be substituted as follows into (10.16) to obtain the 95% interval

$$\begin{aligned} CI(\psi) &= \exp([20][.138]) \pm [1.96][20][.0285]) \\ &= \exp(1.64, 3.88) \\ &= 5.16, 48.4 \end{aligned}$$

This interval does not contain 1, indicating a statistically significant effect of aptitude, a result consistent with that found above with the use of the z statistic.

Illustrating the computation for a *dichotomous IV* with (10.17), the 95% interval for the contrast of award versus no award reflecting the effect of citizenship (X_2) is

$$\begin{aligned} CI(\psi) &= \exp(3.06 \pm [1.96][.572]) \\ &= \exp(1.94, 4.18) \\ &= 6.96, 65.4 \end{aligned}$$

For a *control group contrast* example, the 95% confidence interval for the medium versus high SES contrast (second versus last levels represented with β_4) is computed with (10.17) to be

$$\begin{aligned} CI(\psi) &= \exp(.496 \pm [1.96][.565]) \\ &= 0.543, 4.97 \end{aligned}$$

Finally, to illustrate the computation for a *noncontrol group contrast* with (10.18), the interval for the low versus medium SES contrast, using results from the corresponding test above, is

$$\begin{aligned} CI(\psi) &= \exp([-1.989 - .496] \pm [1.96][.594]) \\ &= 0.026, 0.266 \end{aligned}$$

A summary of the test and interval results for the example is given in Table 10.5.

Table 10.5
Logistic Regression Example Summary

Variable	Effect	S.E.	Odds ratio	
			Estim	95% Interval
Aptitude	0.138*	0.028	15.80[a]	5.16, 48.4
Citizenship	3.061*	0.572	21.35	6.96, 65.4
Age	1.303	0.793	3.68[b]	0.59, 13.3
SES[c]				
Low v. High	-1.989*	0.579	0.14	0.04, 0.42
Med v. High	0.495	0.565	1.64	0.54, 4.97
Low v. Med	-2.485*	0.593	0.08	0.03, 0.27
Constant	-22.4			

* Significant at the 0.05 level.
[a] Odds ratio associated with a 20 unit increase in aptitude.
[b] Odds ratio associated with a 1 year increase in age.
[c] The global effect of SES was significant at the 0.05 level.

10.5 Model Fit

The phrase "model fit" in the logistic regression literature typically covers two separate concerns. One, considered in Section 10.5.1, is whether there are any problematic *individual* observations, a concern addressed under the topic of case analysis in our discussion of multiple regression. A second aspect of logistic model fit is whether the assumed logistical functional form provides an adequate approximation of the underlying true functional form. This type of concern was previously addressed under the "correct fit assumption" for multiple regression and is discussed in Section 10.5.2.

10.5.1 Case analysis

All of the issues of interest in a case analysis for multiple regression (see Chapter 3) are, for the most part, also important in logistic regression. Here we focus on the question of possible *influential observations*, i.e., individual observations with excessive influence on the logistic model effects. Influential observations can be identified with $\Delta\beta$s that indicate the change in the coefficient for each IV that would result from the elimination of each observation. These case indices are available in some current computer packages

How does one determine if a given $\Delta\beta$ represents "excessive" influence? Let $\psi(c)$ represent the odds ratio associated with an increase of c units for an interval IV that is based on the complete sample, and let $\psi'(c)$ represent the same odds ratio from the sample with the individual observation of interest deleted. The ratio of these two odds ratios is equal to

$$\frac{\psi'(c)}{\psi(c)} = \exp(c\,\Delta\beta) \qquad (10.19)$$

For each IV, one would judge whether the departure of this ratio from 1 for the largest $\Delta\beta$ is excessive. A similar determination for effects of dichotomous and categorical IVs would be accomplished with c = 1 in (10.19).

Example. For the current example, the largest $\Delta\beta$ over all observations for the aptitude IV is approximately -0.01. Assuming an interest in the effect of a change of 20 units in aptitude, the ratio in (10.19) is exp([20][-0.01]) = 0.82. That is, the odds ratio for a 20 unit aptitude change would decrease by a factor of 0.82 if the observation were dropped. In the context of an interval estimate for this effect of (5.16, 18.4), it does not appear that deletion of the observation would result in any qualitative change in the study conclusion about the effect of

aptitude on program completion. When this procedure is repeated for the other IV's in the example, no evidence of individual influential observations is found.

10.5.2 Test of model goodness of fit

Logistic regression assumes that the shape of the actual variation of π (the probability of Y = 1) with change in one or more IVs can be represented with the logistic function defined in (10.1) and the logit equation in 10.10. This assumption is analogous to the correct fit assumption in multiple regression. We consider here a test useful for determining if there is any apparent violation of this assumption. The null hypothesis to be tested states that the true underlying relationship has the specified logistic functional form. Thus, rejection of this null would constitute evidence that there is a violation of the assumption. The null hypothesis can be tested with the Pearson chi-square statistic defined as

$$\chi^2 = \Sigma\ \hat{r}_i^2 \qquad (10.20)$$

where r_i is the Pearson residual for the ith observation defined as

$$r_i = \frac{(Y_i - \pi_i)}{\sqrt{\pi_i(1 - \pi_i)}} \qquad (10.21)$$

The χ^2 test statistic in (10.20) has a χ^2 distribution with degrees of freedom of df = n-k-1, where n is the sample size and k is the number of IVs in the model. For the larger degrees of freedom typically associated with this test, the critical value is equal to

$$C_v = \frac{(z[\alpha] + \sqrt{2df - 1})^2}{2} \qquad (10.22)$$

where $z(\alpha)$ is the critical value for α from a standard normal distribution.

Example. For the current example, the Pearson χ^2 statistic is 155.5. With a sample size of 160 and k = 5, the degrees of freedom are 154, resulting in a critical value of 189.9 from (10.22). Thus, the null hypothesis is not rejected, a decision consistent with correct fit of the model.

If the null hypothesis is rejected, indicating possible fit problems, the next step would involve a comparison of a nonparametric representation of the relationship with the fitted relationship. For the example, the required nonparametric representation would be obtained by plotting the probabilities of program completion as a function of aptitude (see Table 10.1). Such a comparison would suggest the degree and nature of the failure of the logistic model.

10.6 Classification of Individuals

The introduction of the current example in Section 10.1 mentioned two possible purposes of the logistic regression to be illustrated -- better understanding of why some children do not complete the summer remedial program and the identification of individual children at the beginning of the program who are at risk of dropping out. Thus far, the discussion of the effects of the different IV's in the model has addressed primarily the goal of understanding. We now turn our attention to the use of the model for decisions about individual children, an application analogous to the goal of individual predictions in multiple regression.

When the outcome of interest is dichotomous, a classification of individuals into one of two categories corresponding to the two possible outcomes is often of interest. In the current hypothetical example, assume that the goal at the beginning of the summer program is to classify students into one of two groups -- those at risk of dropping out of the program and those not at risk. Special attention could then be given to those identified as at risk in an attempt to raise the program completion rate.

A decision rule is required for the classification of individuals into two categories with logistic regression. The simplest approach is to use the estimated probability of program completion with a specified cutpoint. If the estimated probability of completion for an individual is lower than the cutpoint, the individual is assigned to the noncompletion group, i.e., the individual is classified as "at risk." If the estimated probability is higher than the cutpoint, the individual is assigned to the completion group, i.e., they are not identified as "at risk." The computation of the estimated probability of completion for an individual was illustrated in Section 10.3.5, where it was found, for example, that a student of age 10.5 years at program entry from a family with medium SES having an aptitude of 55 and a good citizenship award was estimated to have a probability 0.917 of completing the program.

The selection of the cutpoint for the decision rule can be based on several considerations. One possible intuitive choice could reason that any student having a probability of *not* completing the program that is greater than the student's probability of completing should be classified at risk. The cutpoint value in this case would be 0.5. Resource limitations and/or consequences of a student not completing the program, though, may make it necessary to modify this common-sense choice. For example, if the resources available for the special attention to be given to the identified at-risk students are very limited, a somewhat lower cutpoint may be chosen to select and focus on the students who are "most at risk." On the other hand, if the consequences of failing to complete the program are extremely serious. it may be decided that the group to receive special attention should be broadened by increasing the cutpoint above 0.5.

Once a decision rule with cutpoint is specified, it is essential to assess the accuracy of the resulting classifications. This is usually accomplished with a

classification table summarizing the number of correct and incorrect predictions or classifications. The classification table for the current example, assuming that the decision rule cutpoint is 0.5, is shown in Table 10.6.

Table 10.6
Example Classification Table

		Predicted outcome		
		Fail to complete	Complete	Total
Observed outcome	Fail to complete	50 (71.4%)	20 (28.6%)	70
	Complete	11 (12.2%)	79 (87.8%)	90

Overall correct assignment = 80.6%

Example results. Table 10.6 indicates that 71.4% of the 70 students who failed to complete the program were correctly classified by the decision rule, while 28.6% were incorrectly assigned. That is, if completing the program is designated as the "positive" outcome, the decision rule produced 28.6% "false positives." For the 90 students who completed the program, 87.8% were correctly predicted and 12.2% were incorrectly predicted to fail (i.e., 12.2% were "false negatives"). Of the total of 160 students in the sample, a proportion of (50+79)/160 or 80.6% were correctly assigned. Thus, the derived decision rule provides a substantially better "hit rate" than a baseline of 50% associated with a random assignment.

The *cross validation* procedure discussed in Chapter 7 for multiple regression would also be appropriate here when the logistic regression equation must be estimated with small to moderate samples. Capitalization on chance, which resulted in a positively biased R^2 estimate in multiple regression, produces here positively biased estimates of the classification hit rates. Approximately unbiased hit rates would be obtained by following the procedure described in Chapter 7. First, the sample of subjects with known outcomes would be randomly split into two samples, one a calibration sample and the other a cross validation sample. The logistic regression equation would be estimated with the calibration sample and then applied to the subjects in the cross validation sample. If the accuracy of the decision rule is judged to be acceptable based on the cross validation hit rates, the two samples would then be recombined to obtain the final estimated logistic regression equation.

10.7 Example Results Summary

The study sample was comprised of 160 students who entered the remedial program last summer. Ninety students completed the program, while 70 dropped out at some point before completion. One hundred and fifteen of the 160 received the school citizenship award, and frequencies in the low, medium, and high categories of family SES were 59, 51, and 50, respectively. For the entire sample, the means (and standard deviations) for aptitude and age were 50.1 (9.87) and 10.9 (0.28), respectively.

The estimated logistic regression model coefficients are shown in Table 10.5. The strength of the overall relationship was statistically significant with the likelihood ratio test at the 0.05 level ($\chi^2 = 83.9$, $\chi^2(0.05;5) = 11.1$). A search for any individual observations that exerted excessive influence on the model results did not reveal any problematic data. A test of the Pearson chi-square statistic for the model fit produced a fail to reject decision, a result consistent with the assumption that the specified logistic model is correct. Given the individualized nature of the program, there was no reason to believe there was any violation of the independence assumption.

The effects of three of the four hypothesized IV's -- aptitude, citizenship, and SES -- were statistically significant at the 0.05 level, as shown in Table 10.5. All of the significant effects were in the expected positive direction and quite strong. For the effect of aptitude, the odds of program completion were estimated to increase by a factor of 15.8 when aptitude is increased by 20 units (approximately two standard deviations), controlling for other variables. For the citizenship variable, the odds of program completion for students receiving the citizenship award were estimated to be 21.3 times greater than the odds for those not receiving the award. For the SES variable, there were statistically significant odds ratios for the contrasts of the low SES level with the medium and high levels (0.08 and 0.14, respectively), with the odds of program completion being greater at the higher SES levels. The odds ratio for the medium versus high SES contrast was not significant. Finally, the effect of age on the probability of program completion was not significant.

There was support for the use of the resulting logistic regression equation for the identification of students at risk of not completing the program, assuming a decision rule cutpoint for the estimated probability of completion of 0.5. The classification results shown in Table 10.6 indicate that 71.4% of those not completing the program were correctly identified by the decision rule. The cost of this classification was a 12.2% rate of false negative predictions for those who in reality completed the program.

10.8 Summary

When the outcome of interest is dichotomous, logistic regression provides a model of the probability of each of the two possible outcomes. The resulting model can be used for either or both of the two usual study purposes -- gaining better understanding of the outcome with descriptions of how each of the hypothesized IV's is related to the outcome, controlling for other IV's, and allowing classifications of individuals based on their predicted outcomes. The steps in the typical logistic regression analysis are very similar to those in multiple regression, i.e., any individual problematic observations exerting excessive influence are identified with a case analysis, the validity of the assumptions

associated with logistic regression is assessed, including the formal test of goodness of the model fit, the overall relationship between the outcome and the IV's is tested, the effect of each of the IV's, controlling for the others, is described with tests and interval estimates, and if the model is to be used to classify individuals, the accuracy of classification is assessed.

10.9 Related Topics and References

(to be added)