Meyers, L. S., Gamst, G., & Guarino, A. J. (2006). Applied multivariate research: Design and interpretation. Thousand Oaks, CA: Sage.

CHAPTER 6A

Logistic Regression

In the preceding chapters, we have seen that regression analysis involves a quantitatively measured dependent variable and independent variables that are quantitatively measured, dichotomous, or both. This chapter will address regression analysis where the dependent variable is categorical; the independent variables here may still be quantitative, dichotomous, or both. Although the categorical dependent variable could consist of more than two categories (called a *polytomous* or *multinominal* variable under that circumstance), this chapter will address logistic regression designs with a dichotomous (binary) dependent variable. Multinominal logistic regression is beyond the scope of this chapter.

Many research studies in the social and behavioral sciences investigate dependent (outcome) variables of a dichotomous nature. Pedhazur (1997) illustrates these designs:

> The ubiquity of such variables in social and behavioral research is exemplified by a yes or no response to diverse questions about behavior (e.g., voted in a given election), ownership (e.g., of a personal computer), educational attainment (e.g., graduated from college), status (e.g., employed), to name but some. Among other binary response modes are agree-disagree, success-failure, presence-absence, and pro-con. (p. 714)

The use of logistic regression is increasing of late because of the wide availability of sophisticated statistical software and high-speed computers. Hosmer and Lemeshow (2000) report that the use of logistic regression has "exploded during the past decade" (p. ix). Logistic regression analysis has expanded from its origins in biomedical research to fields such as business and finance, criminology, ecology, engineering, health policy, linguistics, and

wildlife biology. Logistic regression has become so popular that Huck (2004) predicts, "It may soon overtake multiple regression and become the most frequently used regression tool of all!" (p. 438).

The Variables in Logistic Regression Analysis

In a typical logistic regression analysis, there will always be one dependent variable (dichotomous for our purposes here) and usually a set of independent variables that may be either dichotomous or quantitative or some combination thereof. Furthermore, the dichotomous variables need not be truly binary; for example, researchers may transform a highly skewed quantitative dependent variable into a dichotomous variable with approximately equal numbers of cases in each category. And akin to what we have seen in multiple regression, some of the independent variables in logistic regression analysis may serve as covariates to allow researchers to hold constant or statistically control for these variable(s) to better assess the unique effects of the other independent variables.

Assumptions of Logistic Regression

Although logistic regression makes fewer assumptions than linear regression (e.g., homogeneity of variance and normality of errors are not assumed), logistic regression does require the following:

1. There must be an absence of perfect multicollinearity.
2. There must be no specification errors (i.e., all relevant predictors are included and irrelevant predictors are excluded).
3. The independent variables must be measured at the summative response scale, interval, or ratio level (although dichotomous variables are also allowed).

Logistic regression requires larger samples than does linear regression for valid interpretation of the results. Although statisticians disagree on the precise sample requirements, Pedhazur (1997) suggests using at least 30 times as many cases as parameters being estimated.

In this chapter, three examples of logistic regression will be presented. The first example is the simplest design possible with a single dichotomous independent (predictor) variable (gender) and a dichotomous dependent (criterion or outcome) variable of seeking therapy (yes or no). This example will introduce the basic concepts of logistic regression. The next example will present a single quantitative predictor variable (level of depression) with the same dichotomous dependent variable of seeking therapy. The final

example will use two predictors—one dichotomous independent variable (gender) and one quantitative variable (level of depression) to predict the dichotomous dependent variable of seeking therapy. This final analysis will allow us to illustrate how we can assess the unique effects of the independent variables in logistic regression.

Coding of the Dichotomous Variables

In logistic regression, as we discuss it here, the dependent variable will always be dichotomous. Although researchers could use any two arbitrarily selected numbers to label the two categories (e.g., 2 and 3 or 6 and 17), Hosmer and Lemeshow (2000) recommend coding dichotomous variables as 1 for the event occurring and 0 for the event not occurring. It is also possible to use dichotomous variables as predictors; for example, gender (female vs. male) may be predictive of (related to) an event occurring or not (e.g., pass or fail, select one program over another), or having a family history of a certain medical condition (yes vs. no) may be predictive of whether or not individuals are diagnosed with that condition themselves. Such predictors also need to be coded as 1 and 0. Here, those cases who are the focus of the study or who possess some attribute should be coded as 1 and the others coded as 0. In the former example, females (if they were the focus of the study) would be coded as 1 and males would be coded as 0; in the latter example, those having a family history of the condition would be coded as 1 and those with no such history would be coded as 0.

In a very general sense, cases or incidents coded as 1 are referred to (or are thought of) as the *response group, comparison group,* or *target group;* cases or incidents coded as 0 are sometimes called the *referent group, base group,* or *control group.* The ultimate objective of logistic regression is to predict a case's group membership on the dependent variable by calculating the probability that a case will belong to the 1 (event occurring) category.

There are certain advantages to using this 1 and 0 coding scheme. One advantage is that the mean of the distribution is equal to the proportion of 1s in the distribution, thus allowing researchers to immediately know this second piece of information from the first. Consider the situation where there are 100 people in the sample. If 30 of them are coded 1, then the mean of the distribution is .30, which is the proportion of 1s in the data set. The mean of the distribution is also the probability of drawing a person labeled as 1 at random from the sample. Therefore, the indexes of proportion and probability with respect to the value or code of 1 are the same.

So widespread is this type of coding scheme that these indexes have become institutionalized. The mean of a binary distribution coded as 1 and 0 is denoted as P, the proportion of 1s. The proportion of 0s is (1 – P), which

is sometimes denoted as Q. The variance of such a distribution is PQ. In the above example where there are thirty 1s in a set of 100 cases, the variance of the distribution is .21 (.3 × .7 = .21) and the standard deviation is the square root of PQ or .458.

Another reason to follow the 1 and 0 coding scheme is that SPSS will automatically recode the lower number assigned to a category to 0 and the higher number of the category to 1 in its logistic regression procedure. For example, if a researcher coded the dichotomous dependent variable as 1 and 2, SPSS will recode the 1 to 0 and the 2 to 1. This automatic recoding can cause no end of confusion, so beginning the procedure in the same way that SPSS finishes it makes keeping track of which group is which a whole lot easier.

The Logistic Regression Model

Conceptually, logistic regression and linear regression are analogous. Both methods produce prediction equations. Recall from Chapter 5A that in multiple regression analysis the ordinary least squares strategy is used to calculate the prediction of a quantitative dependent variable. The regression function is a straight line; that is, the prediction model is a linear combination of the independent variables. Things are a bit different in logistic regression.

Least Squares Is Not Appropriate in Logistic Regression

Because the dependent variable in logistic regression is dichotomous, using the least squares technique to calculate the prediction of a quantitative dependent variable is inappropriate for two reasons.

First, because the dependent variable is dichotomous, the equal variance assumption underlying linear multiple regression is violated. Recall that in multiple regression there is the assumption that the variance of Y is constant across values of X (this is an assumption of homoscedasticity). Equal variances are an unreasonable condition to meet in logistic regression for at least two reasons:

1. The variance is calculated by multiplying the proportion of 1s by the proportion of 0s. Thus, the notion of "equal variances" of presumably the 1s and 0s does not make sense because we have to work with both proportions to compute the variances in the first place.
2. If we take "equal variances" to mean an equal proportion of 1s and 0s (which must be .5 and .5), then we place an unnecessary restriction on the variables that can be used in the analysis. Most variables are probably going to show different numbers of 1s and 0s in the set, and we would certainly be interested in using some of these as dependent variables in a logistic regression analysis.

Second, using the least squares method can produce predicted values greater than 1 and less than 0, values that are theoretically inadmissible. This means that the linear function that we use in multiple regression is inappropriate here. Instead, the shape of the best-fit line in logistic regression is sigmoidal or S-shaped.

We have illustrated in Figure 6a.1 the difference between the S-shaped or sigmoidal logistic function and the linear regression function. Notice that if researchers were to use least squares for a dichotomous outcome, the equation would suggest that a change in one of the predictor variables (labeled generically as X and represented on the X axis) has a constant effect on the probability of the event occurring (shown as Y on the Y axis).

The Logistic Function

If a straight-line function is not workable, what sort of function is better? The answer provided by logistic regression is, perhaps not surprisingly, a logistic function as shown in Figure 6a.1. Variable X is the predictor here and, for the sake of exposition, assume that it is a quantitative variable. Y is the probability of the particular event of interest occurring.

This function is interesting in that it is pretty flat at both the low and high levels of X. Notice in Figure 6a.1 that the corresponding Y values for Xs in the range of 1 through 6 are all around 0. This suggests that differences in this range of X do not make much of a difference in outcome. Let X be

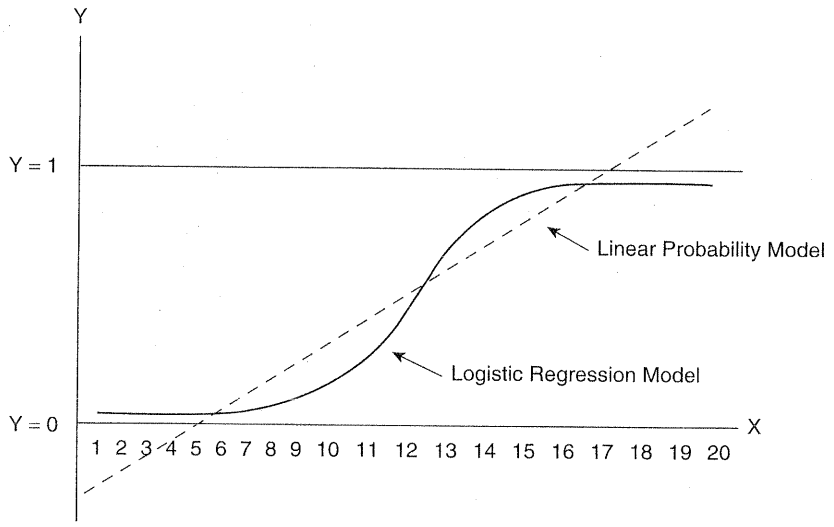

Figure 6a.1 Comparing the Linear Probability and Logistic Regression Models

the amount of distress individuals are experiencing at a given time in their lives, with lower values indicating less distress, and let Y be the probability of seeking therapy. Then the logistic function indicates that people experiencing little stress, whether at levels of 1 or 6 or anyplace in between, are not likely to enter therapy.

An analogous situation is seen at the upper end of the X range where again the function is quite flat. Individuals experiencing lots of distress, whether as low as 15 or as high as 20, are all very likely to seek therapy. Again, distress level within this range does not make much of a difference in prediction.

So where is the "real action" in prediction? It is in the midrange of X where different distress levels are associated with different probabilities of seeking therapy. In fact, the steeper the slope in this range, the more differentiating would be the different distress levels. And it makes sense that prediction of who will seek therapy is more uncertain in this range of distress.

In essence, what we have said is that distress level (X) does not have "a constant effect" on the probability of seeking therapy (Y). And that may make it apparent why least squares multiple regression is less applicable here.

The point is that a linear function defines a situation in which the predictor and the criterion bear a constant relationship to each other; that is, this much difference in the predictor results in that much difference in the criterion over the entire range of the predictor. When a constant relationship is not descriptive of the relationship, least squares, with its fit of a linear function, is not an acceptable strategy to use. This is where the logistic regression model, with its S-shaped function that relates predictors to probabilities of events occurring, takes on a considerable amount of predictive power.

Some Underlying Mathematical Issues

The sigmoidal function represents a nonlinear relationship between the predictor(s) and the binary outcome. Because of this, the mathematical operations underlying logistical regression are a bit more complex than those used in ordinary least squares multiple regression. We will take you through a simplified treatment of these. It is simplified because SPSS does all the mathematical work and our emphasis is not in that domain, but we treat it because what is done mathematically drives the way in which researchers interpret the result of a logistical regression analysis.

The Natural Log (ln) Transformation

Logistic regression requires a mathematical transformation of the original data. The mathematical transformation used in logistic regression is the

natural log (abbreviated ln) *transformation.* It "bends" the data to fit the sigmoidal curve. The ultimate objective of logistic regression is to predict a case's group membership. This translates to the probability or likelihood of an event occurring for a given value of a predictor variable—more specifically, the probability or likelihood of a case's membership in the response group. Making this prediction requires a sequence of two equations. At first, this set of equations might give you the impression that they are a Rube Goldberg parody (i.e., a structure to accomplish a simple task in the most complicated, elaborate, and ridiculous method). However, these transformations, accomplished over a series of two steps, are necessary to interpret the results of the logistic regression analysis.

Step 1: Forming the Logistic Regression Equation

For calculation purposes in logistic regression, the probability of an event is transformed to odds. Odds are computed by the formula (P / 1 – P), where P is the probability of an event occurring (of the outcome being a 1) and 1 – P is the probability of an event not occurring (of the outcome being a 0). It is possible, however, to create a linear relationship between the predictors and odds. Although a number of functions work, one of the most useful is the logit function. It is the ln of the odds that a case belongs to the response group (the group coded as 1).

The logistic regression equation that results is this:

$$\ln [\text{odds}] = a + b_1X_1 + b_2X_2 + \cdots + b_nX_n$$

The left side of the equation simply substitutes the ln odds for the predicted dependent variable in linear regression. The b coefficients indicate the change in log odds of membership for any 1-unit change in the independent variables. In this sense, logistic regression is in reality linear regression using the logit as the outcome variable.

Translating the Outcome Variable

To make the above equation more comprehensible, the logistic regression can be rewritten in such a manner that ln, the natural log, is the predicted group membership. We can symbolize predicted group membership as gpred. Thus,

$$g_{pred} = \ln [\text{odds}] = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

Step 2: Computing the Logit Outcome

Because the logit (i.e., the natural logarithm of an odds ratio) is difficult to interpret, the log odds are transformed into probabilities by taking the antilog (the number corresponding to a logarithm) of the above equation. This is accomplished as follows:

$$e^{gpred} / (1 + e^{gpred}) = \text{predicted probability}$$

That is, the log odds (now represented as g^{pred}) are now inserted into the antilog function where e (the exponential function) = 2.7182. This is the antilog equation that transforms the log odds to probabilities. We will talk you through worked numerical examples to show you that this may sound a bit more complicated than it really is.

Deriving the Constant (a) and b Weights

The values for the constant (a) and the b weights are calculated through maximum likelihood estimation (MLE) after transforming the dependent variable into a logit variable. MLE seeks to maximize the log likelihood, which indicates how likely the observed grouping can be predicted from the observed values of the independent variable(s). MLE is an iterative process that starts with an initial arbitrary "guesstimate" and then determines the direction and magnitude of the logit coefficients. SPSS provides the result of that process to you in its output.

Example 1: One Dichotomous Predictor Variable

The research question of interest in our hypothetical study is this: "Is there a gender difference in seeking psychotherapy for depression?" Gender will be the independent (predictor) variable in this study, and seeking psychotherapy will be the dependent (criterion) variable. The occurrence of the event is the outcome (usually positive) in which the researchers are most interested. Based on this reasoning, those individuals seeking therapy will be coded as 1 and those not seeking therapy will be coded as 0. As for gender, assume our focus is on the women in the sample. Because of this focus, females will be coded as 1; we will therefore code the males as 0.

Because the central mathematical concept of logistic regression is the logit, we are going to elaborate on the concepts of (a) probability, (b) odds, and (c) odds ratio. We will use the simple hypothetical study whose results are shown in Table 6a.1 to illustrate these important concepts. Table 6a.1 presents a count of 500 participants seeking psychotherapy by gender.

Table 6a.1

Overall, ther
50 males sou

Probability

Probabili
as a decimal
divide it by tl
rolling a 4 or
say that the
100 attempts
Probabilities
1 make it im
ties (ordinar

Notice t
example sou
the probabili
interest is, h
females. The
seeking ther

Odds

Odds in
group (seeki
group (not s
chotherapy a
This means
seek psychot
100 or 2. Thi
seek therapy

Unlike p
to infinity. If t

Table 6a.1 Hypothetical Data Illustrating Gender Differences for Seeking Psychotherapy

	Therapy	*No Therapy*
Female	200	100
Male	50	150
Total	250	250

Overall, there were 300 females and 200 males; of these, 200 females and 50 males sought therapy.

Probability

Probability is the likelihood that an event will occur. It is often expressed as a decimal value. To compute it, take the number of occurrences and divide it by the total number of possibilities. For example, the probability of rolling a 4 on a six-sided die is 1 divided by 6 or .167. One would therefore say that the probability of that event occurring is about .17. Thus, out of 100 attempts, we would expect about 17 of them to result in rolls of 4. Probabilities are constrained to lie between 0 and 1. The constraints at 0 and 1 make it impossible to construct a linear equation for predicting probabilities (ordinary least squares will cause these bounds to be exceeded).

Notice that an equal proportion (50%) of the participants in our example sought psychotherapy (250 sought therapy and 250 did not). Thus, the probability of a study participant seeking therapy is .50. The researchers' interest is, however, in whether this proportion is the same for males and females. The first step would be for the researcher to calculate the odds of seeking therapy for males, and then for females.

Odds

Odds in this example represent the probability of belonging to one group (seeking therapy) divided by the probability of not belonging to that group (not seeking therapy). In this study, the odds of a male seeking psychotherapy are .33 (50 males sought therapy divided by 150 who did not). This means that for every male who sought therapy, three males did not seek psychotherapy. The odds for females seeking psychotherapy are 200 / 100 or 2. This means that a female is twice as likely to seek therapy as not to seek therapy.

Unlike probability, which is bounded by 0 and 1, odds can range from 0 to infinity. If the chances of an event occurring are greater than not occurring,

then the odds will be greater than 1. If the chances of an event failing to occur are greater, then the odds will be less than 1. If there is an equal chance of the event occurring or not occurring, then the odds equal 1. Odds of 1 are considered the equivalent of the null hypothesis in logistic regression.

Odds Ratio

Obtaining the odds ratio is one of the important objectives in logistic regression. It is what researchers need to calculate to answer the research question in most logistic regression studies. In this study, the question of whether there is a gender difference in seeking therapy for depression is essentially answered by calculating an odds ratio.

The odds ratio is, as may be obvious from its name, a ratio of the odds for each group. It is a way of comparing whether the probability of a certain event is the same for two groups. The numerator represents the odds of the event occurring (seeking psychotherapy) for the response group (the females) and the denominator represents the odds of the event occurring (seeking psychotherapy) for the referent group (the males).

Analogous to odds, an odds ratio of 1 implies that the event is equally likely in both groups. An odds ratio greater than 1 indicates that the likelihood of an event occurring is more likely in the group coded 1 than in the group coded 0. An odds ratio less than 1 suggests that the event is less likely to occur in the group coded 1 than in the group coded 0. In our hypothetical example, the odds ratio is computed by dividing the odds of females seeking therapy by the odds of males seeking therapy.

Given that the female odds were calculated as 2 and the male odds were calculated as .33, we divide 2 by .33 to obtain 6. We interpret this outcome as follows: Women in this sample were six times more likely than males to seek psychotherapy for depression.

Women were coded as 1 because they were the focus of our study. Had men been our focus, we would have coded the males as 1 and the females as 0. The odds ratio would then have been .33 / 2 = .167, and we would then say that males were one sixth as likely to seek therapy.

The odds ratio just discussed would be considered "crude" or "unadjusted" because no other variables were considered in the analysis. Frequently, logistic regression would employ a set of independent variables to predict or explain the dichotomous dependent variable. When multiple variables are used as predictors in a logistic analysis, the odds ratio is referred to as an *adjusted odds ratio* to indicate the contribution of a particular variable when other variables are controlled or held constant. This controlling for another variable will be illustrated in our third example.

The Logistic Regression Equation

Just as in linear regression, logistic analysis will also produce a single constant (a) and a regression coefficient (b) for each predictor in the model. Recall that the b coefficient in linear regression is derived through the use of ordinary least squares estimation. The least squares strategy seeks to minimize the sum of squared distances of the data points to the regression line, as explained in our preceding chapters. In logistic regression the b coefficients are calculated through MLE after transforming the dependent variable into a logit variable.

MLE seeks to maximize the log likelihood, which reflects how likely it is (the odds) that the observed values of the dependent variable may be predicted from the observed values of the independent variable(s). MLE is an iterative process that starts with an initial arbitrary guesstimate and then determines the direction and magnitude of the logit coefficients. Thus, the b coefficient in logistic regression indicates the change in log odds of membership in the dependent variable code of 1 for any 1-unit change in the independent variable. This increase in log odds is not easily interpreted. The b coefficient will eventually allow the researcher to predict the probability of membership of the response group that is seeking therapy. But remember that the logistic regression equation does not directly predict the probability of an event occurring; rather, it predicts the log odds that an observation will have an outcome (a code) of 1. We therefore need to engage in a two-step process to calculate the probability for a case belonging to the response group.

Computing the Logistic Regression Equation

Recall that determining a case's group membership requires calculating the log odds and then transforming the log odds to probabilities. The first step in calculating the logistic equation for our hypothetical example is based on the linear equation, $g_{pred} = a + b_1X_1$. Note that there is only one predictor (X_1) in the equation because there is only one predictor (gender) in our example study. With more predictors, there would be more terms.

In this example, the values of X_1 will be either 1 for females or 0 for males, and g_{pred} is the ln of the odds of a case in the target group. The logistic regression, through MLE analysis that was performed by SPSS, yielded values of −1.099 for the constant and 1.792 for the b weight. The resulting formula demonstrates the relationship between the regression equations. This is a linear formula. The logistic regression equation to be addressed in the second step is nonlinear.

Thus, for a female,

$$g_{pred} = -1.099 + 1.792(1)$$

$$g_{pred} = 0.693$$

For the second step, the value of .693 is now inserted into the following formula, known as the antilog, to transform the log odds to probabilities:

$$e^{gpred} / (1 + e^{gpred}) = \text{predicted probability}$$

Recall that e is the exponential function and has a value of approximately 2.718. Thus, for a female, the logit equation would be $2.718^{.693} / (1 + 2.718^{.693}) = .667$.

The analogous computations for a male would be as follows:

$$g_{pred} = -1.099 + 1.792(0)$$

$$g_{pred} = -1.099$$

$$\text{predicted probability} = 2.718^{-1.099} / (1 + 2.718^{-1.099}) = .250.$$

Interpreting the Logit Outcome

The calculated probabilities from the logistic analysis can now be used to predict group membership. For this, we need to apply a decision rule for this prediction, a rule based on the predicted probability. The rule used is as follows: If the predicted probability is .5 or greater, then the outcome is to seek therapy (coded as 1; if the probability is less than .5, the case is classified as not seeking therapy (coded as 0).

In this example, females would be classified as seeking therapy because the calculated probability (.667) exceeds .5. Males, however, would be classified as not seeking therapy because their calculated probability (.250) is less than .5. Later, we will compare these predicted group memberships with the actual group memberships as a method to assess the accuracy of the prediction based on the independent variable.

The odds ratio can be calculated directly from e (which has a value of 2.718) and the b coefficient. We simply raise e to the b power. In the present example:

$$e^{b} = \text{odds ratio}$$

$$e^{1.792} = \text{odds ratio}$$

$$2.718^{1.792} = 6.0$$

As indicated, we interpret the odds ratio here (with females coded as 1) as indicating that women in this sample sought psychotherapy for depression six times more than males.

Example 2: One Quantitative Predictor Variable

This next example presents a single quantitative variable (level of depression) used as a predictor with the dichotomous dependent variable of seeking therapy. For this example, 46 students are assessed by a hypothetical self-report depression inventory. Assume that this inventory validly assesses depression. Further assume that the scores on this inventory can range from a low score of 0, indicating a low level of depression, to a high score of 100, indicating greater levels of depression.

Recall that odds are the probability of belonging to one group (seeking psychotherapy) divided by the probability of not belonging to that group (not seeking psychotherapy). In this example of depression and psychotherapy, the probability of seeking psychotherapy is contingent on depression level. The raw data for a small number of cases in this hypothetical study are shown in Table 6a.2. Those who sought therapy are shown as 1s, and those not seeking therapy are shown as 0s. Depression scores for the cases in the study range from a low of 35 to a high of 80.

Visual inspection of the data contained in Table 6a.2, even though it lists only 46 of the 500 cases in the data file, reveals a pattern we have discussed before in this chapter. In the low regions of depression scores shown in the table, 0s predominate under the therapy outcome, indicating that most of these cases do not seek therapy. In the high regions of depression scores shown in the table, 1s predominate under the therapy outcome, indicating that most of these cases do seek therapy. In the midrange region of depression we find a mixture of 0s and 1s. Although even this limited set of cases would resemble an S-shaped curve if plotted on a set of axes, the entire set of 500 would yield a very clear sigmoidal function.

We can now talk about the data in terms of odds. Assume that in the larger data set, 10 cases had depression scores of 15. Of these 10 individuals, 1 of them (10% or .10) sought therapy and 9 of them (90% or .90) did not. Thus, the odds of seeking psychotherapy for individuals with a low score of 15 are .11 (.10 / .90 = .11). This indicates that for every 1 person with a depression score of 15 who seeks psychotherapy, there are 9 who do not seek therapy.

We can examine the odds of seeking therapy for cases whose depression scores are more in the midrange of the distribution. At a depression score of 55, for example, we find that an equal number of cases did and did not

Table 6a.2 A Set of 46 Cases Drawn From a Larger Sample of 500 Cases Showing Whether or Not They Sought Psychotherapy and Their Corresponding Depression Score

Therapy	*Depression Score*	*Therapy*	*Depression Score*
0	35	1	60
0	35	1	60
0	40	1	60
0	40	1	60
1	40	0	65
1	40	0	65
0	45	1	65
0	45	1	65
0	45	1	65
0	45	1	70
0	50	1	70
0	50	1	70
0	50	1	70
0	50	1	70
0	50	1	75
0	50	1	75
1	50	1	75
1	50	1	75
0	55	1	80
0	55	1	80
0	60	1	80
0	60	1	80
1	60	1	80

Note: 0 = No therapy; 1 = therapy.

seek therapy. Therefore, the odds of seeking therapy with a depression score of 55 are 1 (.50 / .50 = 1), indicating that there is an equal chance of seeking and not seeking therapy. Using the same reasoning process, we find that the odds of seeking therapy with a depression score of 85 are 9 (.90 / .10 = 9). At this very high level of depression, for every 9 individuals who seek therapy, there is 1 who does not.

Applying a logistic function makes intuitive sense here. There is little change in the probability of seeking therapy at low or high depression levels. That is, there is little difference between those who are not at all depressed and those who are depressed just a little; these individuals, as a

general rule, simply do not seek therapy for depression. At the same time, there is also little difference between those who are very depressed and those who are very, very depressed; these individuals in our database will generally seek therapy. The decision to seek therapy is most volatile or uncertain in the midrange of the depression continuum; increasingly higher depression scores are associated with an increasing likelihood that individuals will seek therapy.

Given such a pattern across the depression range means that the relationship between the predictor and the predicted values is nonlinear. Note also that in our larger hypothetical data set, the odds of seeking therapy with a low depression score was .11 (.10 / .90 = .11), whereas the odds of seeking therapy with a high depression score was 9 (.90 / .10 = 9). These odds reveal an asymmetry (odds of .11 and 9) that may be reconciled through the logit transformation.

The logit transformation resolves both the asymmetrical issue concerning the odds as well as the calculations of different probabilities at different predictor values. Recall that once we calculate the constant (a) and the b coefficient through MLE, a probability can be eventually derived by using the two-step procedure previously described. The formula is this: $g_{pred} = \text{logit} = a + b_1X_1$, where X_1 is now the depression score. This logistic regression analysis yielded a constant value of −7.734 and a b weight of .139. Thus, for a depression score of, say, 35, the probability of an individual being in the response group is as follows:

$$g_{pred} = -7.734 + .139(35)$$

$$g_{pred} = -2.869$$

This value of −2.869 can now be inserted into the antilog equation to transform the log odds to probabilities, where $e = 2.718$.

$$\text{predicted probability} = 2.718 - 2.869/ (1 + 2.718 - 2.869) = .053.$$

For a depression score of 80, the results would be

$$g_{pred} = -7.734 + .139(80)$$

$$g_{pred} = 3.386$$

$$2.718^{3.386}/ (1 + 2.718^{3.386}) = .966$$

In this example, a depression score of 35 would result in an individual's being classified as someone who is not seeking therapy because the calculated probability (.053) is less than .5. However, a depression score of 80 would result in an individual's being classified as someone who is seeking therapy because that person's calculated probability (.966) exceeds .5.

To calculate the odds ratio in this example, we raise *e* to the power of b. For this example,

$$e^{.139} = 1.149$$

This tells us that the odds of seeking therapy are 1.149 times greater for a person who had a depression score of, say 35, than for a person with a depression score of 34. For a quantitative variable in general, the odds ratio indicates the odds of the target outcome occurring (the outcome coded as 1) when comparing one level of a predictor with another. For example, when comparing the therapy-seeking behavior of individuals with a depression score of 35 to those of 37, one multiplies the regression coefficient by the size of the difference in quantitative scores before raising *e* to the power of the coefficient. Thus, the difference between a depression score of 35 and 37 (a 2-unit increase) would be calculated as follows:

$$e^{(2 \times .139)} = e^{.278} = 1.320$$

We would therefore say that those scoring a 37 on depression are 1.32 times greater in seeking psychotherapy than those scoring 35. If there is a 10-point difference on scores, the equation would be

$$e^{(10 \times .139)} = e^{1.39} = 4.01$$

and our interpretation would be that those who scored 70 are 4.01 times greater in seeking psychotherapy than those scoring 60.

The key thing to remember in interpreting an odds ratio is that, by definition, you are comparing one set of cases with another on the outcome coded as 1 in your data file. The odds ratio changes as a function of the "distance" between the two sets. How much of a change is predicted is a function of b, the regression weight, which is the power to which we raise *e*. The only issue is how many bs we need, which is told to us by the distance between the sets. If the difference is 2 (depression scores of 35 and 37, depression scores of 63 and 65, and so on), then we multiply b by that difference in the power function (*e* raised to the power of twice b when comparing those who scored 37 with those who scored 35 or when comparing those who scored 65 with those who scored 63).

Example 3: One Continuous and One Dichotomous Variable Predictor

In this final example, the sample of 46 students now has one dichotomous independent variable (gender), with females coded as 1 and males coded as 0, and one continuous variable (level of depression). The dichotomous dependent variable is still the interest in obtaining therapy, coded as 1 for seeking therapy and 0 for not seeking therapy. This third example analysis is presented because logistic regression would generally employ a set of independent variables to predict or explain the dichotomous dependent variable. When more than one predictor variable is used in a logistic analysis, researchers can examine the contribution of one variable when other variables are controlled or held constant. Controlling for another variable (using a variable as a covariate) will be illustrated with this example.

The MLE method yielded a constant (a) of −8.35, a b_{gender} of 2.129, and a $b_{depression}$ score of .131. Thus, for a male (coded as 0) with a depression score of 40, the probability of seeking therapy (target group coded as 1) is

$$g_{pred} = -8.35 + 2.129(0) + .131(40) = -3.11$$

$$\text{predicted probability} = 2.718^{-3.11} / 1 + 2.718^{-3.11} = .042$$

Because the predicted probability of group membership is below .5, this respondent would be predicted not to seek therapy. However, for a female with a depression score of 75, the equation would be

$$g_{pred} = -8.35 + 2.129(1) + .131(75) = 3.604$$

$$\text{predicted probability} = 2.718^{3.604} / (1 + 2.718^{3.604}) = .973$$

Because the predicted probability of group membership is greater than .5, this respondent would be predicted to seek therapy.

Because multiple variables (two in this simple situation) were used as predictors in the logistic analysis, the odds ratio is now referred to as an *adjusted odds ratio* to indicate the contribution of a particular variable when other variables are controlled or held constant. The odds ratio for gender is now $e^{2.129} = 8.404$. This indicates that females (coded as 1) are 8.40 times more likely to seek therapy when we have statistically controlled for depression level. Another way to interpret this result would be to imagine that for the condition in which males and females had identical depression scores, females are about 8.4 times more likely to seek therapy than the males.

Evaluating the Logistic Model

Logistic regression produces a number of tests to assess the validity of the model (i.e., the regression equation). These tests can be characterized as either "absolute measures" (assessing the entire model) or "relative measures" (assessing the unique contribution from the individual independent variable).

Absolute Measures

Likelihood Ratio Test

If a researcher has no information other than the outcome, known as the *constant-only model,* then the researcher's "best guess" of an individual is that he or she will seek therapy because that was the more common outcome with this sample (56.5% in this sample). Thus, without considering any other information (gender or depression level), the likelihood or probability is that an individual will seek therapy.

The first absolute measure of the validity of the model is the likelihood ratio test, which evaluates whether or not the set of the independent variables improves prediction of the dependent variable better than chance. Because each case is independent of the others (one case's decision to seek or not to seek therapy in no way affects the decision of any other case), the probability of seeking therapy can be computed as the percentage of the sample seeking therapy (the percentage of 1s in the sample) raised to the power equal to the number of cases in the sample. For our small data set of 46 cases, 26 sought therapy. That makes up 56.5% of the sample. We thus raise .565 to the 46th power. The result is approximately .00000000000392. This is not an unusual magnitude to obtain because we are raising some decimal value to a power equal to the sample size.

Because these likelihood values are ordinarily very small, the ln of the likelihood is usually reported instead in the output. We calculate this value by multiplying the ln of the above value (the percentage of 1s raised to the power of the sample size) by –2. This outcome (i.e., –2 times the log likelihood) is referred to as –2LL. Taking the ln of the likelihood transforms a typically very small number into a "reasonably large" value that is more familiar to most of us. For example, the ln of .0000000000039 is –26.262. Multiplying this ln by –2 yields a value of 52.52. Why not just leave the ln of the likelihood alone (it's certainly a large enough value for most of us) instead of multiplying it by –2? The answer is that the distribution of –2LL is distributed as chi square, whereas LL by itself is not. Thus, it can be used for assessing the significance of the logistic regression model. Specifically, this test

assesses if at least one of the independent variables (covariates) is statistically significantly different from zero.

Omnibus Test of Model Coefficients

The omnibus test of model coefficients is another absolute measure of the validity of the model. The model chi square is a statistical test of the null hypothesis that all the coefficients are zero. It is equivalent to the overall *F* test in linear regression. The model chi-square value is the difference between the constant-only model and the full model (i.e., with constant and predictors). In this example, the null hypothesis is rejected because the significance is less than .05. The researcher would conclude that the set of independent variables improves prediction of the outcome over the situation where they are not used.

Model Summary

The third absolute measure of the validity of the model is the model summary, which is a goodness-of-fit statistic. This fit statistic is usually not interpreted directly, but is useful when comparing different logistic models.

Pseudo R^2

The Cox and Snell and the Nagelkerke tests are two alternative ways to compute a pseudo R^2 estimate and are thought of as absolute measures of the validity of the model. They are used to determine the percentage of variance in the dependent variable explained by the independent variables in logistic regression and are thus analogous to but not the same as the R^2 generated in multiple regression analysis. Although technically, an R^2 cannot be computed the same way in logistic regression as it is in least squares regression, the two tests are referred to as pseudo R^2. The pseudo R^2 in logistic regression is defined as $(1 - L_{full})/L_{reduced}$, where $L_{reduced}$ represents the log likelihood for the "constant-only" model and L_{full} is the log likelihood for the full model with constant and predictors. Usually, the Nagelkerke pseudo R^2 is preferred because it can achieve a maximum value of 1, unlike the Cox and Snell pseudo R^2.

Hosmer and Lemeshow Test

The Hosmer and Lemeshow test is another absolute measure to assess whether the predicted probabilities match the observed probabilities. A researcher is seeking a nonsignificant *p* value for this test because the goal of the research is to derive a set of independent variables (covariates) that will

accurately predict the actual probabilities. Thus, the researcher does not want to reject the null hypothesis. In this example, the goodness-of-fit statistic is 10.161, distributed as a chi-square value, with the p value of .180 indicating an acceptable match between predicted and observed probabilities.

Relative Measures

Wald Test

The relative measures are used to test the statistical significance of the unique contribution of each coefficient (b) in the model. These coefficients indicate the amount of change expected in the log odds when there is a 1-unit change in the predictor variable with all the other variables in the model held constant. A coefficient close to 0 suggests that there is no change due to the predictor variable. To assess if a coefficient is statistically significantly greater than zero, logistic regression uses the Wald test. This test is analogous to the t test in multiple regression. In this example, gender is statistically significantly associated with seeking psychotherapy exclusive of depression level. Likewise, depression level is statistically significantly associated with seeking psychotherapy exclusive of gender.

However, several authors have identified problems with the use of the Wald statistic. Menard (2002) warns that for large coefficients, standard error is inflated, lowering the Wald statistic (chi square) value. Agresti (1996) states that the likelihood ratio test is more reliable for small sample sizes than the Wald test.

The Issue of Standardized (Beta) Coefficients

Current statistical computer programs do not produce the standardized logit coefficients. However, a researcher could standardize the data first, and the logit coefficients would then be the standardized logit coefficients. Alternatively, researchers could multiply the unstandardized logit coefficients by the standard deviations of the corresponding variables, giving a result that is not the standardized logit coefficient but that can be used to rank the relative importance of the independent variables.

Recommended Readings

Allison, P. D. (1999). Comparing logit and probit coefficients across groups. *Sociological Methods and Research, 28,* 186–208.

Cox, D. R., & Snell, E. J. (1989). *Analysis of binary data* (2nd ed.). London: Chapman & Hall.

DeMaris, A. (1992). *Logit modeling: Practical applications.* Newbury Park, CA: Sage.

Estrella, A. (1998). A new measure of fit for equations with dichotomous dependent variables. *Journal of Business and Economic Statistics, 16,* 198–205.

Fox, J. (2000). *Multiple and generalized nonparametric regression.* Thousand Oaks, CA: Sage.

Hosmer, D. W., Jr., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.

Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). Thousand Oaks, CA: Sage.

Pampel, F. C. (2000). *Logistic regression: A primer.* Thousand Oaks, CA: Sage.

Rice, J. C. (1994). Logistic regression: An introduction. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 3, pp. 191–245). Greenwich, CT: JAI Press.

Wright, R. E. (1995). Logistic regression. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 217–244). Washington, DC: American Psychological Association.