13

Alternative Regression Models: Logistic, Poisson Regression, and the Generalized Linear Model

Throughout the text we have used the ordinary least squares regression (OLS) model. For statistical inference, OLS regression assumes that the residuals from our analysis are both normally distributed and exhibit homoscedasticity (see Section 4.3). But we are sometimes confronted with a dependent variable Y that does not result in our meeting these assumptions. For example, Y may be dichotomous, as when someone is diagnosed with a disease or not, referred to in the epidemiological literature as "case" versus "noncase" (e.g., Fleiss, 1981). Or Y may be in the form of counts of rare outcomes, for example, the number of bizarre behaviors exhibited by individuals in a given period of time. When the objects or events counted are rare (e.g., many people exhibit no bizarre behavior), many individuals have zero counts, so that the count variable Y is very positively skewed. By the nature of the dichotomous and count dependent variables, residuals from OLS regression of these dependent variables do not standardly meet OLS assumptions. In such instances the OLS regression model is not efficient and may well lead to i/naccuracies in inference. A class of statistical approaches subsumed under a broad model, the *generalized linear model* (Fahrmier & Tutz, 1994; Long, 1997; McCullagh & Nelder, 1989) has been developed to handle such dependent variables that lead to residuals that violate OLS assumptions.

In this chapter we present two statistical procedures that fall under the generalized linear model: logistic regression and Poisson regression. Having presented these procedures, we integrate them in an overview of the generalized linear model.

13.1 ORDINARY LEAST SQUARES REGRESSION REVISITED

A more formal characterization of the OLS regression model will help to frame the developments in this chapter. To reiterate, throughout the text we have used the OLS regression equation:

$$\hat{Y} = B_1X_1 + B_2X_2 + \cdots + B_kX_k + B_0. \tag{13.1.1}$$

A continuous dependent variable Y is written as a linear combination (weighted sum) of a set of predictors. The predictors may be categorical or continuous. Individual predictors may

be functions of other predictors, as in Chapter 6 for polynomial regression where we have predictors that are powers of other predictors such as X_i^2, or in Chapter 7 for interactions where we have predictors that are products of other predictors, such as X_iX_j. A broader term, the *general linear model*, encompasses OLS regression and other statistical procedures, among them ANOVA.

13.1.1 Three Characteristics of Ordinary Least Squares Regression

There are three important characteristics of OLS regression. The first characteristic is the algebraic form of the model. The model is referred to as a linear model because it is *linear in the parameters*, or, equivalently, *linear in the coefficients*, that is, each predictor is merely multiplied by its regression coefficient, as in Eq. (13.1.1) (see Section 6.1.1).

The second characteristic is the *error structure* or distribution of residuals. As we stated, a critical assumption for OLS regression from the point of view of inference is that the *residuals* $(Y - \hat{Y})$ *be normally distributed.* The normal distribution has the special property that the mean and the variance of the distribution are independent. In regression, we consider the *conditional distribution of Y* for any given value of $\hat{Y}$, that is, the distribution of Y scores associated with a single predicted score, where the predicted score is a linear combination (or weighted sum) of the predictors, $\hat{Y} = B_1X_1 + B_2X_2 + \cdots + B_kX_k + B_0$. We assume that all these conditional distributions, one for each value of $\hat{Y}$, are normally distributed. If this is so, then the value of conditional variance of the Y scores given any value of $\hat{Y}$ is independent of the value of $\hat{Y}$. Such independence is required if our data are to exhibit *homoscedasticity*, that the conditional variance of Y at each value of $\hat{Y}$ is constant over all values of $\hat{Y}$ (see Section 4.3.1). Our inferences in OLS regression depend on these assumptions of normality and homoscedasticity, as well as on the assumption that all observations are independent.

The third characteristic is the scale of the predicted score in relation to the scale of the observed Y score. In OLS regression, the scale of the predicted score is the same as the scale of the criterion; put another way, predicted scores are in the *same units* as the observed Y. For example, in familiar OLS regression, if the observed Y is in the units of number of "pounds of weight," then the predicted score will be in those same units of "pounds of weight." It is possible, however, to have forms of a regression equation in which the units of the observed Y differ from the units of the predicted score. For example, in Poisson regression used with count data, the observed Y entered into the regression equation might be the number of aggressive acts a child carries out in a period of time; the predicted score will be in the form of the logarithm of the number of aggressive acts.

In sum, then, OLS regression with which we have worked throughout the text has three characteristics: (1) it is linear in the parameters (i.e., there is a linear equation relating the set of Xs to Y in the form of Eq. 13.2.1), (2) the errors of prediction (residuals) are assumed to be normally distributed and to exhibit homoscedasticity, and (3) the units of the predicted scores are the same as the units of the observed Y scores.

13.1.2 The Generalized Linear Model

A broad class of regression models, collectively known as the *generalized linear model* (McCullagh & Nelder, 1989), has been developed to address multiple regression with a variety of dependent variables Y like dichotomies and counts. OLS regression is one special case of the generalized linear model. Like OLS regression, all these regression models can be expressed in a form that is linear in the parameters (Section 6.1.1). Statistical methods that fall under the generalized linear model include, in addition to OLS regression, other forms of regression

analyses for data that do not lead to normally distributed residuals exhibiting homoscedasticity. These methods allow for residuals, the variance of which depends on the predicted value of Y. In addition, in these methods of regression analysis, unlike OLS regression, the form of the predicted score is sometimes different from the form of the observed Y. In this chapter we focus on two examples of the generalized linear model—logistic regression for categorical outcome variables and Poisson regression for count variables that measure frequency of occurrence of rare events. We then characterize the class of generalized linear models, drawing upon logistic and Poisson regression as specific examples. We warn that the conventions for data analysis in the generalized linear model are not so well developed as for OLS regression, for example, measures of overall model fit or regression diagnostics. Where analogs to OLS regression exist, they are described, their limitations noted, and any lack of consensus about their use explained.

13.1.3 Relationship of Dichotomous and Count Dependent Variables Y to a Predictor

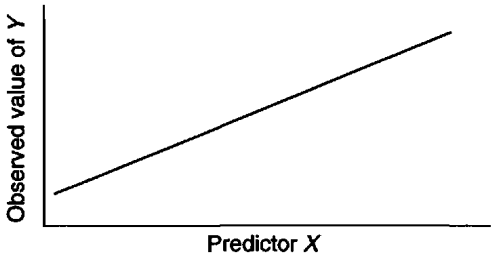
For any linear model, we require a dependent variable Y that is linearly related to the set of predictors. Figure 13.1.1 illustrates the relationship of a predictor to Y for three outcomes: (1) Fig. 13.1.1(A) for an ideal continuous Y linearly related to the predictor, as in OLS regression, (2) Fig. 13.1.1(B), for a binary (or dichotomous) Y, and (3) Fig. 13.1.1(C) for a count variable Y.

For a dichotomous dependent variable Y, we consider the score for one individual to be $Y = 1$ if the person exhibits a particular characteristic, $Y = 0$, otherwise; that is, we use 1 and 0 for case versus noncase, respectively. For example, consider whether an assistant professor is promoted to associate professor ($Y = 1$, for case) versus is not promoted ($Y = 0$, for noncase). It is also useful for illustrative purposes to imagine summarizing the dichotomous Y for a set of individuals as the proportion of individuals with $Y = 1$ at each value of some predictor. For example, if we examine a large pool of faculty who were considered for promotion from assistant professor to associate professor, we compute the proportion of those faculty with $Y = 1$ as a function of the number of publications (e.g., the proportion of faculty with seven publications promoted to associate professor). Figure 13.1.1(B) illustrates the likely relationship of the proportion of people classified as a case as a function of a single predictor X. The form of the relationship in Fig. 13.1.1(B), is not linear, but rather S-shaped, suggesting that probability of being a case first increases very slowly as X increases, then increases in a rather linear fashion over the midrange of X, and then reaches asymptote (flattens out) with high values of the predictor. A faculty member would hardly be promoted with none, one, or even several publications, and would very likely be promoted when the number of publications, assuming high quality, exceeds 20. In contrast to a linear model, the impact of adding a single publication on probability of promotion is not constant across the range of predictor X; in our example the effect is much stronger in the middle than at the ends of the distribution of number of publications.

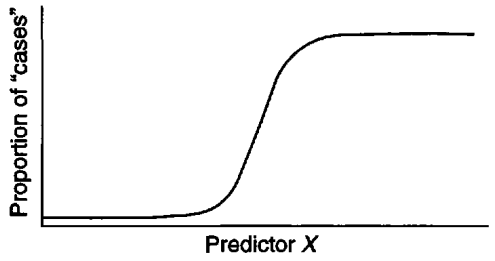
We expect still a different form of the relationship of the predictor to Y if the outcome is in the form of counts (e.g., the number of times a child acts aggressively on the playground), with the predictor being scores on a teacher rating measure of aggressiveness. Figure 13.1.1(C) illustrates a typical form of such count data. There are few, if any, episodes of aggressive behavior when aggressiveness scores are low (there are many zeros on the outcome measure); but the number of aggressive acts accelerates rapidly (i.e., at a faster than linear rate) as the level on the aggressiveness predictor increases.

It is clear from Fig. 13.1.1(B) and (C) that the relationship of the observed outcome to the predictor is not linear when we have dichotomous or count data. Yet, for any linear model, including the generalized linear model, we require a predicted outcome that is linearly related

(A) For a continuous outcome variable Y, the numerical value of Y at each value of X.



(B) For a binary outcome variable, the proportion of individuals who are "cases" (exhibit a particular outcome property) at each value of X.



(C) For an outcome in the form of a count variable, the average number of events exhibited at each value of predictor X.

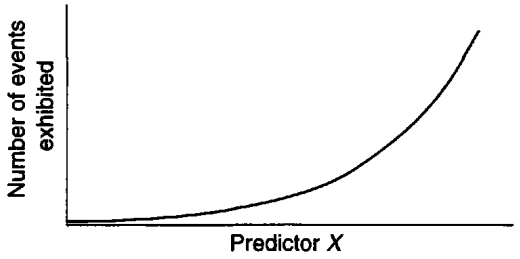


FIGURE 13.1.1 Typical form of relationship of continuous, binary, and count outcome variables to a predictor.

to the predictors. In generalized linear models we generate a predicted outcome that is, in fact, linearly related to the predictors; the linear relationship is achieved by creating a predicted score that is a monotonic but nonlinear transformation of the observed outcome.

13.2 DICHOTOMOUS OUTCOMES AND LOGISTIC REGRESSION

Beyond the continuous dependent variable Y, the most common form of Y is likely the dichotomous (binary, two category) outcome, the analysis of which we will consider in detail. The outcome for each case is dummy coded ($Y = 1$ for case; $Y = 0$ for noncase; see Chapter 8). The probability distribution associated with a dichotomous variable Y is the *binomial distribution*. The proportion P of scores with the characteristic (having values of 1) is the mean of

the distribution; the variance of the distribution is a function of (depends on) P, specifically, $\text{var}(Y) = P(1 - P)$. The variance of a binomial distribution is maximum when $P = .5$. That the variance of the distribution depends on the mean of the distribution is different from the familiar normal distribution, in which the mean and variance are independent.

We begin by exploring several statistical approaches to analyzing the dichotomous dependent variable Y as a function of one or more predictors: the linear probability model, discriminant analysis, and probit and logistic regression. Using the example of promotion to associate professor, we then develop a logistic regression model that describes the probability of promotion to associate professor as a function of a single predictor X, the number of publications. In the logistic regression model, the predicted score is not itself dichotomous; we are not predicting whether someone is a case versus a noncase. Rather we are predicting a value on an *underlying variable* that we associate with each individual, the *probability of membership in the case group* (π_i). What we actually observe is the group membership (case/noncase) of each individual, but what we predict is probability of being a case. This is easier to conceptualize for assistant professors who have not yet been considered for promotion—we want to develop a regression equation that predicts the probability that they will be promoted, based on their number of publications.

Throughout the presentation of the prediction of binary outcomes we will need to distinguish three entities. The first is the probability in the population of being a case, which we will note as π_i for person i. The second is the predicted probability of being a case, based on some regression model, which we will note as $\hat{p}_i$ for person i. Third is the proportion of individuals who are cases, which we will note as P, with $Q = (1 - P)$.

13.2.1 Extending Linear Regression: The Linear Probability Model and Discriminant Analysis

First, recall the familiar linear regression model in the one-predictor case, which yields predicted scores associated with a continuous predictor:

$$\hat{Y}_i = B_1 X_i + B_0. \tag{13.2.1}$$

All the assumptions of OLS regression operate here. In addition, in this equation, predicted scores have a range that is bounded by the actual observed scores; that is, there is no predicted score lower than the lowest observed score and no predicted score higher than the highest observed score.

The Linear Probability Model

One approach to the dichotomous criterion is merely the linear model in Eq. (13.2.1) but with the dichotomous criterion as the dependent variable, that is, OLS regression with a dichotomous criterion. This yields the *linear probability model*, a regression model in which the predicted probability that an individual is a member of the case category $\hat{p}_i$ bears a simple linear relations to the predictor:

$$\hat{p}_i = B_1 X_i + B_0. \tag{13.2.2}$$

Again, this model is the OLS regression model; thus all the requirements of OLS regression apply.

Since the promotion variable Y is dichotomous, it follows a binomial distribution, described earlier. The arithmetic mean of Y is the proportion P of individuals in the whole sample who

are cases (i.e., for whom $Y = 1$). For example, if we have 5 cases with dependent variable scores (1 0 0 1 0), the mean of these scores is .40, the proportion of "cases" in the sample. The variance of the scores in the sample is

$$P(1 - P) = .40(1 - .40) = .24.$$

There are difficulties with this model. First is that the predicted scores, which are supposed to be predicted probabilities of being a case given the value of the predictor, may fall outside the range of the observed criterion scores (i.e., may be less than zero or greater than one). Thus they cannot serve as appropriate estimates of π_i, the population probability of being a case.

Beyond this, there are complications with the residuals that may undermine inference in the linear probability model. An individual can have only one of two scores on Y, that is, $Y = 1$ or $Y = 0$. Thus, only two values of the residual r_i are possible for an individual i, with a given predicted probability $\hat{p}_i$:

Observed score	Predicted score	Residual (r_i)
1	$\hat{p}_i$	$(1 - \hat{p}_i)$
0	$\hat{p}_i$	$(0 - \hat{p}_i)$

In turn, there are two undesirable results of this constraint on the residuals:

1. The residuals exhibit heteroscedasticity. The variance of the residuals, $\text{var}(r_i)$, is not constant across the range of the criterion but depends on the value of the predicted score. The variance of the residuals for any value of $\hat{p}_i$ is given as

(13.2.3) $$\text{var}(r_i) = \hat{p}_i(1 - \hat{p}_i).$$

Although the OLS regression coefficients will be unbiased, they will have incorrect standard errors. This problem can be remedied through the use of weighted least squares regression (see Section 4.5.4).

2. The residuals are not normally distributed. This violates a required assumption for statistical tests and the estimation of confidence intervals for individual regression coefficients in OLS regression.

Discriminant Analysis

OLS regression with a dichotomous outcome Y (i.e, the linear probability model) is mathematically equivalent to another statistical procedure called *discriminant analysis* or *discriminant function analysis*. Two-group discriminant analysis was developed by Sir Ronald Fisher in 1936 as a statistical procedure for using a set of predictors to account for the membership of individuals in one of two groups, that is, to classify individuals into groups on the basis of scores on the predictors in a way that best matched their actual classification. (For example, we might have a clinician's diagnoses of a set of psychiatric patients as well as measures on a battery of test scores on these same individuals; we could explore whether we could account for the clinician's diagnosis of each individual based on scores on the test battery.) In discriminant analysis, a set of predictors is used to generate a prediction equation, called the *linear discriminant function*, that best distinguishes between the two groups. Discriminant function analysis yields estimates of coefficients for each predictor in the linear discriminant function, called *discriminant function coefficients*, and predicted scores, which can be used for statistical classification, called *discriminant function scores*. The equivalence of OLS regression predicting a dichotomous criterion reflecting group membership and *two* group discriminant analysis is manifested in several ways. First, the F test

for the significance of R^2 (the squared multiple correlation) in OLS regression yields the same value as the F test for the overall discrimination between groups in discriminant analysis. Second, the values of the OLS regression coefficients differ only by a multiplicative constant from the values of the corresponding discriminant function coefficients. Third, the tests of significance of individual regression coefficients in OLS regression are identical to the corresponding tests of significance of the discrimiant function coefficients in the discriminant function. Fourth, the predicted scores in OLS regression are correlated 1.0 with the discriminant function scores in discriminant function analysis. See Tatsuoka (1988) for a classic presentation of discriminant analysis, and Tabachnick and Fidell (2001) for a very accessible introduction.

The question may be raised as to why this chapter presents more recent statistical methods for dealing with dichotomous dependent variables Y, particularly logistic regression, when we have discriminant analysis, which is so closely related to OLS regression. The existence of discriminant analysis versus newer logistic regression reflects the ongoing evolution of statistical procedures over time, with efforts devoted to the development of statistical procedures that make assumptions that are more likely to hold true in observed data.

Discriminant analysis makes *strong* assumptions for inference that are not made in logistic regression, as outlined in a classic paper by Press and Wilson (1978). The assumptions are (a) that for each group on the dependent variable Y, the set of k predictor variables is multivariate normal, and (b) that the within-group covariance matrices are homogeneous across the groups. If these assumptions are met, then newer logistic regression is less powerful than discriminant analysis. However, only rarely are these assumptions met in practice. Violation of these assumptions may lead to a number of difficulties with inference in discriminant analysis. Thus, the current recommendation among statisticians is to use logistic regression rather than discriminant analysis in the two-group case. (As a practical rule of thumb, logistic regression and discriminant analysis will yield similar results when the split between groups is not more extreme than 80% in one group versus 20% in the other group.)

An Alternative Approach: Using a Nonlinear Model

Now consider Fig. 13.1.1(B) once again, which shows the proportion of individuals in the sample who are cases as a function of the predictor. This figure suggests that empirically the proportion of people who are cases is not expected to be linearly related to the value of X, but that the function is S-shaped; thus the linear probability model is not appropriate. Rather, Fig. 13.1.1(B) suggests that we should impose a monotonic but nonlinear function relating the predictor to the observed criterion, where the observed criterion is conceptualized as the proportion of cases at each value of the predictor. The function should be an S-shaped function that follows the form in Fig. 13.1.1(B). That is, we should employ a nonlinear model—a model in which the predicted score $\hat{p}_i$ bears a nonlinear relationship to the value of the predictor. It is this latter option that underlies the regression models that we apply to dichotomous outcomes.

13.2.2 The Nonlinear Transformation From Predictor to Predicted Scores: Probit and Logistic Transformation

To operationalize a regression model for dichotomous outcomes, we require a mathematical function that relates the predictor X to the predicted score $\hat{p}_i$ (i.e., predicted probability of being a case). A number of mathematical functions follow a form that highly resembles the S-shaped curve sketched in Fig. 13.1.1(B). Two commonly used functions are the *probit function* and the *logistic function*. The probit function is one in which the predicted probability of being a case,

given a value of X, is generated from the normal curve. The logistic function is developed in detail later. The use of these functions lead to *probit regression* and *logistic regression*, respectively, two special cases of the generalized linear model.

Both the probit and logistic functions are expressions for the relationship between the predictor X and the predicted probability $\hat{p}_i$. The logistic model predominates in use in psychology and sociology, and we will focus on logistic regression. The choice of logistic over probit regression is based on various factors. First is that the logistic regression model has advantages in interpretation of regression coefficients in terms of the *odds*, that is, the ratio of the probability that an individual is a case to the probability that the person is a noncase (odds are a familiar way of expressing probabilities for those who bet on races or other sporting events). A second advantage is the simplicity of interpretation in case-control studies, in which cases are systematically sampled based on their status on the dichotomous outcome; for example, individuals with a particular disease (cases) are matched with those not having the disease (noncases or controls). The proportion of cases in the sample is typically grossly different from that in the population. Nonetheless, with logistic regression strong inferences about the magnitude of effects in the population are appropriate if certain assumptions are made. The most common application of probit regression in psychology is in the context of structural equation modeling with binary variables (see Chapter 12).

Classic sources on the analysis of dichotomous data include Agresti (1990), Fleiss (1981), and Hosmer and Lemeshow (2000). Excellent sources also include Collett (1991) and Long (1997), and the very accessible introductions by Aldrich and Nelson (1984), Menard (2001), and Pampel (2000). We draw on all these sources here. It should be noted that when prediction of dichotomous outcomes is by categorical predictors only, logistic regression is equivalent to a logit model applied to contingency tables (Fox, 1997).

Boxed Material in the Text

We caution the reader at the outset that the form of regression equations, strategies for statistical inference, fit indices, and the like are somewhat more complex in logistic and Poisson regression than in now familiar OLS regression. (The reader may benefit from reviewing Section 6.4.3 on logarithms and exponents before proceeding.) Thus, in this chapter we again adopt a strategy of putting some material into boxes to ease the presentation. The material in the boxes is typically of interest to the more mathematically inclined reader (this same stategy was employed in Chapters 4 and 6). The boxes provide supplementation to the text; the text can be read without the boxes. Boxed material is set apart by bold lines; boxes appear in the section in which the boxed material is relevant. Readers not interested in boxed material should simply skip to the beginning of the next numbered section.

13.2.3 The Logistic Regression Equation

The logistic regression equation for predicting the probability of being a case $\hat{p}_i$ from a single predictor X is given as

$$\hat{p}_i = \frac{1}{1 + e^{-(B_1X_i + B_0)}} = \frac{e^{(B_1X_i + B_0)}}{1 + e^{(B_1X_i + B_0)}}. \tag{13.2.4}$$

The expression $(B_1X_i + B_0)$ is what we usually treat as the predicted score in a single-predictor OLS regression (see Chapter 2); it is a straightforward linear function of the value of the predictor. The logistic function given in Eq. (13.2.4) relates this score to the predicted probability of being a case $\hat{p}_i$; this is the first way in which the logistic regression equation is

expressed. A plot of $\hat{p}_i$ as a function of X using Eq. (13.2.4) would generate the S-shaped curve of Fig. 13.1.1(B). Equation (13.2.4) gives two equivalent algebraic expressions for the logistic regression equation to predict $\hat{p}_i$. Both of these expressions for $\hat{p}_i$ in Eq. (13.2.4) are unfamiliar forms for a regression equation; we are accustomed to seeing the right-hand side of the regression equation as $(B_1X_i + B_0)$.

Equation (13.2.4) is actually one of three ways in which the logistic regression equation is expressed. By algebraic manipulation we obtain the second form of the logistic regression:

$$\frac{\hat{p}_i}{1-\hat{p}_i} = e^{(B_1X_i+B_0)}. \tag{13.2.5}$$

Of particular note is that the form of the predicted score in Eq. (13.2.5) differs from that in Eq. (13.2.4). The predicted score in Eq. (13.2.5) is the odds of being a case, explained further later on.

The third form of the logistic regression is actually the natural logarithm of Eq. (13.2.5):

$$\ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = B_1X_i + B_0. \tag{13.2.6}$$

The right-hand side of the logistic regression equation is now linear in X; that is, it is identical to the predictor side of the one-predictor OLS regression equation presented in Chapter 2 and given as Eq. (13.2.1). Once again, the predicted score has changed form, this time to the *logit*, the logistic probability unit.

$$\text{logit} = \ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right). \tag{13.2.7}$$

The logit is the function of the predicted probability $\hat{p}_i$ that is linearly related to the predictor X, that is, that lets the predictor side of the regression equation be linear in the parameter estimates. As we will illustrate in detail later, the logit ranges from $-\infty$ to $+\infty$ as $\hat{p}_i$ ranges from 0 to 1. Box 13.2.1 provides the algebraic manipulations to develop the three forms of the logistic regression equation.

13.2.4 Numerical Example: Three Forms of the Logistic Regression Equation

Equations (13.2.4), (13.2.5), and (13.2.6) are the three algebraically equivalent forms of the logistic regression equation. They are illustrated with a fictitious numerical example.

For example, imagine predicting the probability that an assistant professor is promoted to associate professor as a function of the number of publications. The fictitious logistic regression equation of the form of Eq. (13.2.6) predicting the logit of promotion is given as

$$\begin{aligned}\text{logit(promotion)} &= B_1\,(\text{publications}) + B_0\\ &= .39\,(\text{publications}) - 6.00,\end{aligned}$$

where $B_1 = .39$ and $B_0 = -6.00$. Table 13.2.1 gives 31 cases who vary in number of publications from 0 to 30. In addition, the three predicted scores are given: the logit, the odds of being promoted, and the predicted probability of being promoted.The SPSS code to generate these values is provided. One additional entry shows the number of publications that would lead to a $\hat{p} = .50$ predicted probability of promotion, according to the regression equation.

BOX 13.2.1
Development of the Three Forms of the Logistic Regression Equation

We begin with the logistic function relating some variable z_i to the predicted probability $\hat{p}_i$; this function generates the S-shaped curve of Fig. 13.1.1(B):

$$\hat{p}_i = \frac{1}{1 + e^{-z_i}}.$$

To simplify by getting rid of the negative exponent in the denominator, we multiply numerator and denominator by e^{z_i}.

$$\hat{p}_i = \frac{e^{z_i}}{e^{z_i} + (e^{z_i}e^{-(z_i)})} = \frac{e^{z_i}}{e^{z_i} + 1} = \frac{e^{z_i}}{1 + e^{z_i}}.$$

To generate the equation for a one-predictor logistic regression, we substitute the predictor side of the one-predictor regression equation $(B_1X_i + B_0)$, for z_i, which yields Eq. (13.2.4).

Equation (13.2.4) is actually one of three ways in which the logistic regression equation is expressed. Some relatively simple algebraic manipulations of the left expression of the two in Eq. (13.2.4) will convert Eq. (13.2.4) to a more usual form in which the right-hand side of the regression equation is $(B_1X_i + B_0)$. These algebraic manipulations also lead us to the other two forms besides Eq. (13.2.4) in which the logistic regression may be expressed.

We take the reciprocal of Eq. (13.2.4)

$$\frac{1}{\hat{p}_i} = 1 + e^{-(B_1X_i + B_0)}.$$

Then we move the 1 from the right hand to the left-hand side of the equation:

$$\frac{1}{\hat{p}_i} - 1 = e^{-(B_1X_i + B_0)}.$$

Then we place the expression on the left-hand side over a common denominator:

$$\frac{1 - \hat{p}_i}{\hat{p}_i} = e^{-(B_1X_i + B_0)}.$$

We take the reciprocal of both sides of the equation, yielding Eq. (13.2.5) in the text, the second form of the logistic regression equation. Then, we take the natural logarithm of of Eq. (13.2.5) to yield Eq. (13.2.6), the third form of the logistic regression equation.

In the logistic regression equation, logit(promotion) = .39 (publications) − 6.00, the $B_1 = .39$ indicates that the predicted logit increases by .39 for each increase by one in the number of publications. This can be verified in Table 13.2.1 by examining the two columns Number of publications and Logit. $B_0 = -6.00$ is the value of the predicted logit at $X = 0$ publications, which can again be seen in Table 13.2.1. These interpretations of B_1 and B_0 are identical to the interpretations of the analogous coefficients in OLS regression.

TABLE 13.2.1
Fictitious Logistic Regression Example Predicting Probability of Promotion to Associate Professor as a Function of Number of Publications

The regression equation is

$$\text{logit(promotion)} = .39\ \text{(publications)} - 6.00.$$

Case	Number of publications	Logit	Odds	Probability	
100	0	−6.00	.00	.00	
101	1	−5.61	.00	.00	
102	2	−5.22	.01	.01	
103	3	−4.83	.01	.01	
104	4	−4.44	.01	.01	
105	5	−4.05	.02	.02	
106	6	−3.66	.03	.03	
107	7	−3.27	.04	.04	
108	8	−2.88	.06	.05	
109	9	−2.49	.08	.08	
110	10	−2.10	.12	.11	
111	11	−1.71	.18	.15	
112	12	−1.32	.27	.21	
113	13	−.93	.39	.28	
114	14	−.54	.58	.37	
115	15	−.15	.86	.46	
	15.38	**.00**	**1.00**	**.50**	hypothetical case with 15.38 publications and exactly .50 probability of promotion.
116	16	.24	1.27	.56	
117	17	.63	1.88	.65	
118	18	1.02	2.77	.73	
119	19	1.41	4.10	.80	
120	20	1.80	6.05	.86	
121	21	2.19	8.94	.90	
122	22	2.58	13.20	.93	
123	23	2.97	19.49	.95	
124	24	3.36	28.79	.97	
125	25	3.75	42.52	.98	
126	26	4.14	62.80	.98	
127	27	4.53	92.76	.99	
128	28	4.92	137.00	.99	
129	29	5.31	202.35	1.00	
130	30	5.70	298.87	1.00	

SPSS code

```
compute logit = .39*publications − 6.00.
compute odds = exp(logit).
compute prob = odds/(1 + odds).
```

Equivalently, the logistic regression equation may be written in the form of Eq. (13.2.5), predicting the odds of promotion:

$$\text{odds(promotion)} = e^{(.39 \text{ publications} - 6.00)}.$$

Finally, the equation may be written in the form of Eq. (13.2.4), predicting the probability of promotion:

$$\text{probability(promotion)} = \frac{1}{1 + e^{(.39 \text{ publications} - 6.00)}}.$$

Three Forms of the Predicted Score

Predicted probability. In Eq. (13.2.4) the predicted score is the predicted probability $\hat{p}_i$ of being a case. In general, the predicted probability ranges from 0.0 to 1.0. In Table 13.2.1, the predicted probability of promotion is zero for assistant professors with 0 publications, and 1.00 for 30 publications. A useful value is $(-B_0/B_1)$, which gives us the value of predictor X for which the predicted probability is .50. For our example, $(-B_0/B_1) = -(-6.00)/.39 = 15.38$ is the number of publications for which the predicted probability of being promoted $= .5$, as illustrated in Table 13.2.1.

Odds. Equation (13.2.5) has the predicted odds $[\hat{p}_i/(1 - \hat{p}_i)]$ of being a case as the predicted score. Odds are defined as the ratio of the predicted probability of being a case $\hat{p}_i$ to the predicted probability of not being a case $(1 - \hat{p}_i)$. Theoretically, the odds range from 0.0 to $+\infty$ as the probability $\hat{p}_i$ ranges from 0.0 to 1.0. If the probability of being a case is exactly .50, the odds of being a case versus not being a case are exactly 1.0. The odds exceed 1.0 when the probability exceeds .5; the odds are less than 1 (but never negative), when the probability is less than .5. In Table 13.2.1, the computed odds range from .00 to 298.87. The odds are 1.00 for $X = 15.38$ publications, when $\hat{p}_i = .50$. Note that for $\hat{p}_i < .50$, the odds are less than one, though never negative; as $\hat{p}_i$ ranges from .50 to 1.00, the odds accelerate rapidly in value.

Logit. Equation (13.2.6) is the expression of the logistic regression in which the predictor side is linear in the parameters, as in OLS regression. The predicted score in this form of regression equation, that is, the logit or natural logarithm of the odds, $\ln/[\hat{p}_i/(1 - \hat{p}_i)]$, is linearly related to the predictor X. The characteristics of the logit and the relationship of the probability $\hat{p}_i$ to the logit are illustrated in Fig. 13.2.1. As the probability of being a case $\hat{p}_i$ ranges from zero to one (on the abscissa of Fig. 13.2.1), the logit theoretically ranges from $-\infty$ to $+\infty$, that is, the logit is a predicted score that potentially ranges without bound, just as in OLS regression. (Note that Fig. 13.2.1 is cast in terms of population probability π). Hence the compression of probabilities close to zero and close to one in Fig. 13.1.1(B) is eliminated in the logit. (Computationally, the logit is well behaved as $\hat{p}_i$ ranges between zero and one, but offers some computational complexities at the boundaries of $\hat{p}_i$ at exactly zero and one). The logit equals zero when $\hat{p}_i = .50$; put another way, the logit is centered at zero. Table 13.2.1 illustrates the behavior of the logit in the numerical example. The logit ranges from -6.00 to $+5.70$ as $\hat{p}_i$ ranges from zero to one. (We do not see the logit go to $-\infty$, because the actual predicted probability for zero publications is .00246, not zero; we do not see the logit go to $+\infty$ for the same reason). Box 13.2.2 explains how the predicted score is transformed from the logit to the odds to the probability.

The behavior of the logit, odds, and probability are well displayed in Table 13.2.1. To summarize, the logit takes on both negative and positive values without bound. The odds range from zero upward without bound. The probabilities naturally range from 0 to 1. When probabilities are less than .50, the odds are less than one, and the logit is negative; for probabilities greater than .5, the logit is positive and the odds greater than one. The logit varies linearly with the value of the predictor (recall the .39 additive increment to the logit for each increase of one

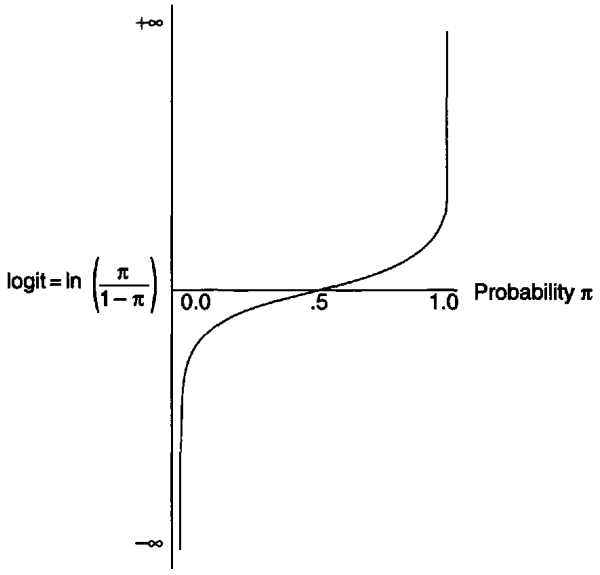


FIGURE 13.2.1 Logit $= \ln\left(\frac{\pi}{1-\pi}\right)$ as a function of the value of probability π. The logit ranges from $-\infty$ to $+\infty$ as probability ranges from 0 to 1. The logit $= 0$ when probability $= .5$.

publication). The probabilities, in contrast, do not. As the number of publications increases from zero to 10, the probability of promotion increases from only .00 to .11. As number of publications increases from 10 to 20, the probability of promotion increases dramatically from .11 to .86. Finally, the diminishing returns of publications above 23 is clearly noted, in that the probability of promotion increases from .95 to 1.00.

BOX 13.2.2
Unwinding the Logit: From Logit to Odds to Probability

The three different logistic regression equations (Eqs. 13.2.4, 13.2.5, and 13.2.6), are merely transformations of one another. Equation (13.2.6) has the most appeal from the predictor side in that it is linear in the coefficients; yet the predicted score is the unfamiliar logit.

Although the form of Eq. (13.2.6) is completely familiar on the predictor side, we might wish to couch the predicted score as the odds or the probability of being a case. We can easily compute the odds and probability from the logit. To find the odds from the logit, we simply exponentiate the logit (equivalently, find the antilog of the logit; see Section 6.4.3). This is straightforward to do. On a calculator, enter the value of the logit and hit the key marked e^x. In SPSS or other statistical packages a statement of the form COMPUTE ODDS = EXP(LOGIT) produces the odds. For example, in Table 13.2.1, with 12 publications, the logit is -1.32; the corresponding odds are $e^{-1.32} = .27$. Finally, to find the probability from the odds, we use the expression

$$\hat{p}_i = \frac{\text{odds}_i}{1 + \text{odds}_i}. \tag{13.2.8}$$

For example, if the odds are .27, the probability of being promoted are $.27/(1 + .27) = .21$.

13.2.5 Understanding the Coefficients for the Predictor in Logistic Regression

The coefficients for predictors in logistic regression analysis are presented in two forms in most software and in publication. First, they are presented as typical *regression coefficients* from Eq. (13.2.6). In the example of Table 13.2.1, $B_1 = .39$ and $B_0 = -.60$ are the familiar regression coefficient and regression constant. As we have shown, the B_1 coefficient indicates the linear increment in the logit for a one-unit increment in the predictor. Second, coefficients for the predictors are presented as *odds ratios:*

(13.2.9) odds ratio for predictor $= e^B$, or, equivalently, $\exp(B)$.

An *odds ratio* is the ratio of the odds of being a case for one value of the predictor X divided by the odds of being a case for a value of X one point lower than the value of X in the numerator (see Section 6.4.3 and Box 13.2.4 for e notation). The odds ratio tells us by what amount the odds of being in the case group are *multiplied* when the predictor is incremented by a value of one unit (e.g., by how much the odds of promotion are multiplied for each additional publication). An odds ratio of 1.0 is associated with a regression coefficient $B = 0$, indicating the absence of a relationship with Y; that is, the odds of being a case are equal for subjects with any given score on X and for those with a score one unit higher. Odds ratios greater than 1.0 correspond to positive B (regression) coefficients and reflect the increase in odds of being in the case category associated with each unit increase in X. Thus an odds ratio of 1.80 indicates that the odds of being a case are multiplied by 1.80 each time X is incremented by one unit. Because the relationship is multiplicative in the odds ratio, a two-unit increase in X would be associated with $1.8 \times 1.8 = 3.24$ times the odds of being a case. Odds ratios falling between 0.0 and just below 1.0 correspond to negative B coefficients and signify that the odds of being a case decrease as predictor X increases.

Epidemiologists most often report outcomes in terms of odds ratios for each predictor rather than the value of the regression coefficients themselves. Hence, in epidemiological literature in which the probability of contracting a disease is given as a function of some risk factor, such as exposure to some chemical, the results might be stated as follows: The odds are four times higher of getting a rare form of cancer if one has been exposed versus not exposed to the chemical.

We have rewritten the form of the logistic regression equation for the odds in a slightly different way and substituted in the values of the coefficients from the numerical example in Table 13.2.1 (recall that algebraically, $r^{(s+t)} = r^s r^t = r^t r^s$; see Table 6.4.1):

(13.2.10) $$\frac{\hat{p}_i}{1-\hat{p}_i} = e^{(B_1X_i+B_0)} = e^{B_1X_i}e^{B_0} = e^{B_0}e^{B_1X_i} = e^{-6.00}e^{.39X}$$

Suppose we examine the odds of promotion given 3 publications versus 2 publications: for 3 publications,

$$\frac{\hat{p}_i}{1-\hat{p}_i} = e^{-6.00}e^{.39\times 3},$$

and for 2 publications,

$$\frac{\hat{p}_i}{1-\hat{p}_i} = e^{-6.00}e^{.39\times 2},$$

for an odds ratio of

$$\frac{e^{-6.00}e^{.39\times 3}}{e^{-6.00}e^{.39\times 2}} = e^{.39} = 1.48 = \text{odds ratio.}$$

If we repeat this examination for 5 versus 4 publications, or 11 versus 10 publications, we find the value of the odds ratio to be the same. The odds of promotion are multiplied by 1.48 for each increment of 1 publication. In Table 13.2.1, for example, the odds of promotion with 10 publications are .12, and with 11 publications, .12(1.48) = .18. The odds of promotion with 16 publications are 1.27; with 17 publications, 1.27(1.48) = 1.88. In sum, when the logit is incremented by a constant additive amount, here $B_1 = .39$, the odds are multiplied by a constant amount, the odds ratio, here 1.48.

We can also consider increments in the odds when predictor X increases by more than one point. For example, an increase from 10 to 15 publications is associated with an increase in the estimated odds of promotion of $1.48^5 = 7.10$ times. Thus, since the odds of promotions with 10 publications are .12, the odds of promotion with 15 publications are $.12(1.48^5) = .12(7.10) = .85$. Note that increments in the predictor X (here, increasing in the number of publications from 10 to 15) are associated with a corresponding powering of the odds ratio (here, raising the odds ratio to the fifth power).

13.2.6 Multiple Logistic Regression

Multiple logistic regression is the straightforward extension of the univariate case, with the same three forms of the logistic regression equation:

(13.2.11) for the logit : $$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = (B_1X_1 + B_2X_2 + \cdots + B_kX_k + B_0);$$

(13.2.12) for the odds : $$\frac{\hat{p}}{1-\hat{p}} = e^{(B_1X_1+B_2X_2+\cdots+B_kX_k+B_0)}.$$

In Eq. (13.2.10), we learned that when the logistic regression equation is in the form predicting the odds, the coefficients are multiplicative. Extending this to multiple logistic regression, an alternative expression for the odds that shows the multiplicative nature of the coefficients in predicting the odds is given as

(13.2.13) for the odds : $$\frac{\hat{p}}{1-\hat{p}} = e^{B_1X_1}e^{B_2X_2}\ldots e^{B_kX_k}e^{B_0}.$$

Equations (13.2.11) and (13.2.13) illustrate the two commonly used forms of logistic regression. Equation (13.2.11) is the linear regression, which is expressed in log odds (logits). Just as in OLS regression, each of the regression coefficients in multiple logistic regression is a *partial regression coefficient*; each is interpreted adjusting for other effects in the model. Equation (13.2.13) is the multiplicative equation, in which the coefficients have been transformed to odds ratios (Section 13.2.5) and the predicted scores are odds. This reflects the mathematical relationship between original units and their logs—a relationship that is multiplicative in original units will be additive in their logs. Standard computer programs report both the linear regression coefficients from Eq. (13.2.11) and the odds ratios from Eq. (13.2.13) (equivalently Eq. 13.2.12).

Generalizing from Eq. (13.2.4) for the third form of the logistic regression equation, in which probabilities are predicted, we have the expression

(13.2.14) $$\hat{p}_i = \frac{1}{1+e^{-(B_1X_1+B_2X_2+\cdots+B_kX_k+B_0)}},$$

or equivalently,

$$\hat{p}_i = \frac{e^{(B_1X_1+B_2X_2+\cdots+B_kX_k+B_0)}}{1+e^{(B_1X_1+B_2X_2+\cdots+B_kX_k+B_0)}}.$$

Interactions and Higher Order Variables

Equation (13.2.11) shows the general form of the linear regression equation to predict the logit. Predictors in logistic regression may take the form of interactions formed as products of other predictors X_jX_j' or powers of other predictors as in polynomial regression X_j^k (Greenland, 1998; Jaccard, 2001). If we consider the multiple logistic regression equation in the form of Eq. (13.2.11), we could characterize the interaction between X_1 and X_2 in predicting the logit as

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = (B_1X_1 + B_2X_2 + B_3X_1X_2 + B_0).$$

In this form, we can think of the interaction as having an impact on the *logit*, as an additive amount over and above the prediction from X_1 and X_2 alone, following the interpretation of interactions in Chapter 7. That is, for a one-unit increase in X_1X_2, from which X_1 and X_2 have been partialed, the logit is increased additively by B_3 units. We can say that the regression of the logit on X_1 depends on the value of X_2 (or the converse), just as in OLS regression.

Our thinking about the interaction is different for the odds and follows our thinking about the meaning of the regression coefficients as amounts by which the odds are multiplied for a one-unit increment in a predictor. In the form of the logistic regression predicting the odds, as in Eq. (13.2.13), the interaction will appear as follows:

$$\frac{\hat{p}}{1-\hat{p}} = e^{B_1X_1}e^{B_2X_2}e^{B_3X_1X_2}e^{B_0}.$$

For a one-unit increase in X_1X_2, from which X_1 and X_2 have been partialed, the odds are multiplied by B_3 units. We can again think of the regression of the odds on X_1 as depending on the value of X_2 (or the converse); however, the model is multiplicative in the odds, and this holds for the interaction as well.

If we analyze the same data with a dichotomous outcome in both OLS and logistic regression, and the model contains an interaction, we may not find the same effects for the interaction in the two analyses. We may find an interaction in OLS but not in logistic regression or the converse. The existence of interactions depends on the scale of the dependent variable. Moving from OLS to logistic regression in essence involves changing the scale of the dependent variable as we move from a linear function shown in Fig. 13.1.1(A) to an S-shaped function as shown in Fig. 13.1.1(B). We encourage readers to trust the results of the logistic regression model, which is more suited to the properties and error structure of binary outcome data. It should be noted that discrepancies between OLS and logistic regression are an example of the broader issue of model consistency across linear versus nonlinear models (See Chapter 6). Jaccard (2001) provides an exceptionally clear explanation of the interpretation of interactions in logistic regression.

13.2.7 Numerical Example

In Table 13.2.2 we present a numerical example of the prediction of whether a woman is in compliance with mammography screening recommendations (1 = in compliance, 0 = not in compliance) from four predictors, one reflecting medical input and three reflecting a woman's psychological status with regard to screening: (1) PHYSREC, whether she has received a recommendation for mammography screening from a physician; (2) KNOWLEDG, her knowledge of breast cancer and mammography screening; (3) BENEFITS, her perception of the benefits of mammography screening for her health; and (4) BARRIERS, her perception

TABLE 13.2.2
Multiple Logistic Regression Predicting Compliance With Mammography Screening Guidelines

I. Logistic Regression

A. Initial Log Likelihood Function (intercept is included in the model).
−2 log likelihood 226.473 (D_{null}, the null deviance)

B. Prediction from four predictors.
Estimation terminated at iteration number 4 because log likelihood decreased by less than .01 percent.
−2 log likelihood 167.696 (D_k, the model deviance)

	Chi-square	*df*	significance	
Model chi-square	58.778	4	.001	$(D_{\text{null}} - D_k)$

C. Regression equation.

1. Regression coefficients.

						95% *CI*	
Variable	*B*	*SE*	Wald χ^2	*df*	Significance	Lower	Upper
PHYSREC	1.842	.488	14.230	1	.001	.88	2.80
KNOWLEDG	−.079	1.074	.001	1	.941	−2.18	2.02
BENEFITS	−.544	.243	5.020	1	.025	.07	1.02
BARRIERS	−.581	.166	12.252	1	.001	−.91	−.26
Constant	−3.051	1.369	4.967	1	.026		

2. Odds ratios.

		95% *CI*	
Variable	Exp(*B*) (odds ratio)	Lower	Upper
PHYSREC	6.311	2.42	16.44
KNOWLEDG	.924	.11	7.54
BENEFITS	1.722	1.07	2.77
BARRIERS	.559	.40	.77
Constant			

II. Discriminant Function Analysis

A. Prediction from four predictors.
$F(4, 159) = 16.294, p < .001$

B. Discriminant function coefficients.

Variable	Discriminant function coefficient	*t*	Significance
PHYSREC	1.419	4.034	.001
KNOWLEDG	−.312	−.373	.710
BENEFITS	−.295	1.817	.071
BARRIERS	−.416	−3.530	.001
Constant	−1.424	.684	.495

of the barriers to her being screened. The data are a random sample of 175 cases from a larger sample of 615 cases in Aiken, West, Woodward, and Reno (1994); 11 cases were eliminated due to missing data, yielding 164 complete cases for analysis. Of the 164 complete cases, 46% were in compliance with screening guidelines; 69% had received a recommendation for screening from a physician.

We initially considered the relationship of each separate predictor to screening compliance. There is a powerful bivariate relationship between physician recommendation and screening compliance: Of women who had received such a recommendation for screening, 61% were in compliance, as opposed to only 14% of those who had not received such a recommendation, $\chi^2(1) = 31.67, \phi = .44, p < .01$. Both perceived benefits and perceived barriers have strong bivariate correlations with compliance, $r(162) = .36, -.41$, respectively, $p < .01$ in both cases. However, knowledge does not correlate with compliance, $r(162) = -.07$.


CH13EX01

Table 13.2.2, Part I, summarizes the results of the logistic regression. We focus first on the coefficients for the individual variables, presented in two forms, as regression coefficients and odds ratios in Parts I (C1) and I (C2), respectively. In Part I (C1), the variables are listed, with the coefficients in the linear form of the regression analysis:

$$\begin{aligned}\text{logit(compliance)} = {} & 1.84\ \text{PHYSREC} - .08\ \text{KNOWLEDG} \\ & + .54\ \text{BENEFITS} - .58\ \text{BARRIERS} - 3.05.\end{aligned}$$

Each regression coefficient is a partial regression coefficient, as in OLS regression. For physician recommendation, holding knowledge, benefits, and barriers constant, the logit increases by $B = 1.84$, when a woman has received a screening recommendation from her physician. As previously explained, the odds ratio for physician recommendation is computed by exponentiating the regression coefficient: $e^{1.842} = 6.31$. Again, partialing out knowledge, benefits, and barriers, the odds of compliance with screening recommendations increase by a factor of over 6, if the women receives a physician recommendation for a mammogram. Perceived benefits are positively related to compliance ($B = .54$), with a corresponding odds ratio greater than one, odds $\text{ratio}_{\text{benefits}} = e^{.54} = 1.72$. Perceived barriers are negatively related ($B = -.58$), with a corresponding odds ratio less than one, odds $\text{ratio}_{\text{barriers}} = .56$. Finally, knowledge is unrelated to compliance ($B = -.08$), with a corresponding odds ratio very close to one, odds $\text{ratio}_{\text{knowledge}} = .92$. Formal significance tests for individual regression coefficients are developed in Section 13.2.12.

Since the split between cases and noncases in this data set is close to equal (i.e., .46/.54 for cases to noncases), we expect very similar results from a discriminant analysis applied to the same data. Results of the discriminant analysis are given in Table 13.2.2, Part II. The discriminant function (analogous to a regression equation) that best distinguished the two groups was as follows:

$$\begin{aligned}\hat{Y}_{\text{DISCRIMINANT}} = {} & 1.42\ \text{PHYSREC} - .31\ \text{KNOWLEDG} + .30\ \text{BENEFITS} \\ & - .42\ \text{BARRIERS} - 1.42.\end{aligned}$$

Overall, the groups were significantly differentiated by the discriminant function. The discriminant function coefficients for PHYSREC and BARRIERS predictors reached conventional significance levels, and the coefficient for BENEFITS approached significance. As we explore significance testing in logistic regression, we will see that the results of the discriminant analysis and the logistic regression converge. We are not surprised, since the groups are close to evenly divided into cases versus noncases.

13.2.8 Confidence Intervals on Regression Coefficients and Odds Ratios

Regression Coefficients

The estimated regression coefficients in logistic regression are asymptotically normally distributed (more about their estimation is given later). Thus, the structure of the confidence interval for a regression coefficient B_j is the same as for regression coefficients in OLS regression, as given in Chapter 2, Section 2.8.2, and in Chapter 3, Section 3.6.1.

(13.2.15) $$CI = [B_j - me \leq \beta_j^* \leq B_j + me],$$

where β_j^* is the population logistic regression coefficient. The margin of error $me = z_{1-\alpha/2}SE_{B_j}$. The value $z_{1-\alpha/2}$ is the familiar critical value from the z distribution, $z = 1.96$ for $\alpha = .05$, two tailed; $z = 2.58$ for $\alpha = .01$, two tailed; SE_{B_j} is the estimate of the standard error of the regression coefficient. This yields a lower limit and an upper limit of an interval within which we are $(1 - \alpha)$ percent confident that the population value β_j^* lies:

(13.2.16) $$lower = B_j - me; \quad upper = B_j + me.$$

The relationship of the ranges of the confidence intervals to significance of regression coefficients is as in OLS regression. A confidence interval for a nonsignificant coefficient will include zero; the confidence interval for a significant coefficient will not include zero. The 95% confidence intervals are given for the regression coefficients in Table 13.2.2, Part I (C1). For example, for BENEFITS, $me = z_{1-\alpha/2}SE_{B_j} = 1.96 \times .243 = .476$, so that $lower = B_j - me = .544 - .476 = .07$, and $upper = B_j + me = .544 + .476 = 1.02$. The confidence interval does not include zero; just as in OLS regression, a test of the difference of this coefficient from zero would be significant. For KNOWLEDG, $me = z_{1-\alpha/2}SE_{B_j} = 1.96 \times 1.074 = 2.104$ so that $lower = B_j - me = -.0794 - 2.104 = -2.18$, and $upper = B_j + me = -.0794 + 2.104 = 2.02$. This confidence interval includes zero, and, as in OLS regression, is associated with a nonsignificant regression coefficient.

Odds Ratios

The confidence intervals on odds ratios are not symmetric, since the odds ratio has a lower limit of zero. The upper and lower limits of the confidence interval on an odds ratio can be easily computed from the corresponding limits on the confidence interval for the regression coefficient. Each limit for an odds ratio is computed by exponentiating the limits from the confidence interval for the regression coefficient:

(13.2.17) $$odds\ lower = e^{B_j - me} \quad \text{and} \quad odds\ upper = e^{B_j + me}.$$

Again, for BENEFITS, $lower = e^{B_j - me} = e^{.07} = 1.07$, and $e^{B_j + me} = e^{1.02} = 2.77$.

Odds ratios close to 1.0 are associated with regression coefficients close to zero. The confidence interval on an odds ratio for a nonsignificant predictor will include the value of one, when the corresponding confidence interval for the regression coefficient includes the value of zero. For example, in Table 13.2.2, Part I(C2), the confidence interval on the odds ratio for knowledge is [.11 $\leq$ odds ratio $\leq$ 7.54], corresponding to a confidence interval on the knowledge regression coefficient itself, in Part I(C1), of [$-2.18 \leq \beta_j^* \leq 2.02$]. The confidence interval on an odds ratio for a negatively predicting variable will range below one when the corresponding confidence interval for the regression coefficient has negative limits. For example, in Table 13.2.2, Part I(C2), the confidence interval on the odds ratio for BARRIERS is [.40 $\leq$ odds ratio $\leq$ .77], corresponding to a confidence interval on the barriers regression

coefficient itself, in Part I(C1), of $[-.91 \leq \beta_j^* \leq -.26]$. Finally, the confidence interval on an odds ratio for a positively predicting variable will range above one when the corresponding confidence interval for the regression coefficient has positive limits. Once again, in Part I(C2) the confidence interval on the odds ratio for BENEFITS is [1.07 $\leq$ odds ratio $\leq$ 2.74], corresponding to an odds ratio on the benefits regression coefficient itself, in Part I(C1), of $[.07 \leq \beta_j^* \leq 1.02]$.

13.2.9 Estimation of the Regression Model: Maximum Likelihood

In OLS regression, estimated values of regression coefficients are selected that minimize the sum of squared residuals of prediction, the *least squares criterion*. The solution is an *analytic solution*; that is, there is a set of known equations from which the coefficients are calculated, the *normal equations* (see Appendix 1). In logistic regression and in other cases of the generalized linear model (e.g., Poisson regression), there is no analytic solution (i.e., there are not a set of equations from which the coefficients are derived directly). Instead, the solution to the regression coefficient estimates is *iterative*, that is, by trial and error, with each trial informed by the previous trial. A statistical criterion is specified for the coefficients to be chosen, and different values of the coefficients are tried until a set of coefficients is found that makes the solution as close to the statistical criterion as possible. The statistical criterion employed is *maximum likelihood*. We had an earlier encounter with this approach when we estimated a model for a sample for which some cases had missing data (Section 11.2.1). As in the current consideration, the estimation in that case concerned a dichotomous outcome.

The maximum likelihood concept begins with the concept of the *likelihood* of an individual or a sample. A *likelihood* for any person is a measure of how *typical* the person is of some population. The likelihood for a sample is a measure of how typical the sample is of the population. For example, we could quantify the likelihood that a woman 5′3″ occurs in a population of women, or the likelihood of drawing a sample of women with a mean height of 5′3″ from a population of women. Extended to regression analysis, the likelihoods under consideration are the likelihoods of individuals having particular scores on the dependent variable Y, given values on the predictors $X_1, \ldots, X_k$, and the *specific values of regression coefficients* chosen as the parameter estimates. The *maximum likelihood estimation* method provides *maximum likelihood estimates* of the regression coefficients (and their standard errors), that is, estimates that make a sample as likely or typical as possible, given values on the predictors and dependent variable Y. The computed likelihood of a sample given the maximum likelihood estimates is termed the *maximum likelihood of the sample*, typically denoted L.

In the course of maximum likelihood estimation, estimates of regression coefficients are tried, the likelihood of the sample, given the estimates, is calculated. Then the estimates are modified slightly according to a search procedure that guides the selection of regression estimates in a manner that increases the likelihood of the sample. This process is repeated, with each attempt referred to as an *iteration*. These iterations continue until the likelihood of the sample, given the set of regression coefficients, ceases to change by more than a small amount termed the *convergence* criterion. A solution has converged when the amount of change from iteration to iteration falls below the convergence criterion. Under some circumstances, convergence fails to be reached. Multicollinearity among predictors and a large number of predictors contribute to nonconvergence. A caution with maximum likelihood estimation is that estimates of the coefficients will not exist if there is *complete separation* on a predictor or set of predictors between the group coded 1 and the group coded 0 (e.g., if all cases in Table 13.2.1 with 15 or fewer publications were not promoted, and all those in with 16 or more publications were promoted).

The iterations in logistic regression (and other generalized linear models like Poisson regression) are accomplished by special mathematical algorithms. Different computer programs use different algorithms and thus may provide (usually slightly) different estimates and statistical test values. This is in contrast to OLS regression, which has a single analytic solution—any discrepancies between computer programs in OLS regression are attributable to differences in accuracy of the programs. Table 13.2.2., Part I(B) shows that the final solution is reached in four iterations for this example.

13.2.10 Deviances: Indices of Overall Fit of the Logistic Regression Model

Measures of model fit and tests of significance for logistic regression are not identical to those in OLS regression, though they are conceptually related. In familiar OLS regression, measures of variation (*sums of squares* or SS) are the building blocks of R^2 (the squared multiple correlation, index of overall fit) as well as of tests of significance of overall prediction and gain in prediction (see Section 3.6.4). For OLS regression, we have the total variation in the DV, that is $SS_Y = \Sigma(Y - M_Y)^2$; this value is a summary number of all the variation in the criterion that can potentially be accounted for by a set of predictors. We also have the predictable variation, the amount of variation in the criterion accounted for by the set of predictors, that is, $SS_{regression} = \Sigma(\hat{Y} - M_{\hat{Y}})^2$. Finally, in OLS regression we have the residual variation, or variation not accounted for by the set of predictors, that is,

$$SS_{residual} = SS_Y - SS_{regression}.$$

In logistic regression, measures of *deviance* replace the sums of squares of OLS regression as the building blocks of measures of fit and statistical tests. These measures can be thought of as analogous to sums of squares, though they do not arise from the same calculations. Each deviance measure in logistic regression is a measure of *lack of fit* of the data to a logistic regression model. Two measures of deviance are particularly useful. The first is the *null deviance*, D_{null}, which is the analog of SS_Y in OLS regression. D_{null} is a summary number of all the deviance that could potentially be accounted for. It can be thought of as a measure of lack of fit of data to a model containing an intercept but no predictors. It provides a baseline against which to compare prediction from other models that contain at least one predictor. The second is the *model deviance* from a model containing k predictors, D_k; it is the analog of $SS_{residual}$ in OLS regression. It is a summary number of all the deviance that remains to be predicted after prediction from a set of k predictors, a measure of lack of fit of the model containing k predictors. In logistic regression, if the model containing k predictors fits better than a model containing no predictors, then the model deviance should be smaller than the null deviance. This is the same idea as in OLS regression; if a set of predictors in OLS regression provides prediction, then $SS_{residual}$ after prediction should be smaller than SS_Y.

We caution here that although these analogies exist between deviance and variation (or variance), deviance is not measured in the same units as variation; thus deviances should not be referred to in writing in terms of variation or variance, a temptation into which we can easily fall when considering goodness of fit indices in logistic regression, presented in Section 13.2.11.

The deviance measures are actually built from maximum likelihoods under various logistic regression models (see Section 13.2.9 for a discussion of likelihoods). As we have said, measures of goodness of fit and test statistics in logistic regression are constructed from the deviance measures. Since the deviance measures are derived from ratios of maximum likelihoods under different models, the statistical tests built on deviances are referred to collectively

as *likelihood ratio tests*. A full explanation of the development of deviance measures from maximum likelihoods and likelihood ratios is given in Box 13.2.3. Because of the way in which deviances are structured from likelihoods, standard notation for deviance in many regression texts and computer output is $-2LL$ or -2 *log likelihood*.

An examination of the deviances associated with the mammography screening example in Table 13.2.2 provides some intuition about how we use deviances. In Table 13.2.2, Part I(A), the null deviance, $D_{\text{null}} = 226.47$, from a model containing only the intercept and no predictors. In Table 13.2.2, Part I(B), the model deviance, $D_k = 167.70$ when the four predictors are included in the regression equation. That D_k is smaller than D_{null} tells us that the four predictors collectively contributed to prediction of the DV. Again, model deviance is a measure of lack of fit, or what is left to predict after the inclusion of k predictors.

BOX 13.2.3
Maximum Likelihoods, Likelihood Ratios, and Deviances

Measures of deviance are developed from maximum likelihoods under various regression models. Maximum likelihoods from different models are formed into likelihood ratios. Deviances are then defined as a function of differences between likelihood ratios. The series of steps in the development of deviances is explained here.

Maximum Likelihoods for Varying Models

The likelihood of scores on the dependent variable Y, given scores on the predictors and the set of regression coefficients, varies as a function of the predictors included. For any regression model with a given set of predictors, there is a *maximum likelihood* that can be obtained, given the values of the regression coefficients. Three different maximum likelihoods are used in the development of measures of overall fit and statistical significance of fit in logistic regression.

1. ***Maximum likelihood of sample under a perfectly fitting model.*** A theoretical model with perfect fit forms the basis of comparison for the fit of other models. Conceptually, such a model has as many predictors as cases. The maximum likelihood under this perfectly fitting model is 1.0, the highest possible.

L_{perfect} = maximum likelihood of sample under a perfectly fitting model = 1.0.

2. ***Maximum likelihood of sample under model containing only an intercept.*** We define a maximum likelihood under the assumption that the outcomes on Y are randomly related to set of predictors X. We do so by defining a *null* model that contains *only an intercept*. The predicted probability for each individual is the base rate of cases in the sample; the predictors offer no differentiation among cases whatever, the worst possible fit.

L_{null} = maximum likelihood of sample, given null model containing only an intercept, lowest maximum likelihood under any possible model.

3. ***Maximum likelihood of sample under model containing intercept plus k predictors.*** We compute the maximum likelihood of a sample for any model

containing the intercept plus k predictors.

L_k = maximum likelihood of a sample under a model containing intercept plus k predictors.

We use this likelihood to assess the goodness of prediction from the model containing the intercept plus k predictors.

Likelihood Ratio

A *likelihood ratio* is a ratio of two maximum likelihoods, typically under one model versus under a more complete model (i.e., with more predictors):

$$\text{(13.2.18)} \quad \text{likelihood ratio} = \frac{\text{maximum likelihood under one model}}{\text{maximum likelihood under more complete model}}.$$

Deviance

The deviance is a measure of lack of fit of one model compared to another model. The deviance is defined as minus twice the value of the log of the likelihood ratio, and is abbreviated as $-2LL$ in various texts.

$$\text{(13.2.19)} \quad \text{deviance} = -2LL = -2\ln(\text{likelihood ratio})$$
$$= -2\ln\left(\frac{\text{maximum likelihood under one model}}{\text{maximum likelihood under more complete model}}\right).$$

Given that $\ln(a/b) = \ln(a) - \ln(b)$, the expression for deviance can also be written as

$$\text{(13.2.20)} \quad \text{deviance} = -2LL = -2\ln(\text{likelihood ratio})$$
$$= -2[\ln(\text{maximum likelihood under model})$$
$$- \ln(\text{maximum likelihood under more complete model})].$$

Deviances contrast maximum likelihoods under various models. The larger the value of deviance for a particular model, the worse the model; that is, deviances are measures of "badness of fit." The specific deviance calculations we present here have direct analogies to familiar measures of total variation SS_Y and residual variation SS_{residual} in OLS regression.

D_{null}, the Null Deviance

The null deviance D_{null} contrasts the maximum likelihood L_{null} under the model containing only the intercept with the maximum likelihood L_{perfect} under the theoretically perfectly fitting model:

$$\text{(13.2.21)} \quad \text{null deviance:} \quad D_{\text{null}} = -2[\ln(L_{\text{null}}) - \ln(L_{\text{perfect}})].$$

This null deviance is analogous to SS_Y, the total variation in the dependent variable Y, from OLS regression. The null deviance measures the discrepancy from the worst possible to the best possible model, all the discrepancy for which it is possible that a model account.

(*Continued*)

D_k, the Model Deviance

The model deviance D_k contrasts the maximum likelihood L_k under the model containing a set of k predictors with the maximum likelihood L_{perfect} under the theoretical perfectly fitting model:

(13.2.22) model deviance: $D_k = -2[\ln(L_k) - \ln(L_{\text{perfect}})]$.

The model deviance is analogous to SS_{residual} from ordinary least squares regression. This deviance measures the amount of the lack of fit that remains after modeling with k predictors, a measure of badness of fit. We expect this value to decrease as we include useful predictors in the regression equation.

13.2.11 Multiple R^2 Analogs in Logistic Regression

In OLS regression we have the squared multiple correlation, R^2 as a single agreed upon measure of goodness of fit of the model, the proportion of total variation in the criterion accounted for by a set of predictors. No single agreed upon index of goodness of fit exists in logistic regression. Instead a number have been defined (see reviews in Estrella, 1998, and Long, 1997; Hosmer and Lemeshow (2000) present a current review). These indices are sometimes referred to as *Pseudo-R^2s*. None of the measures is without limitations, yielding no clear choice for logistic regression. None of these indices is a goodness of fit measure in the sense of having an interpretation as "proportion of variance accounted for," as in OLS regression (more about this later). We present three such indices, the first of which is in common use, the second and third of which will enjoy increasing use now that they are part of standard computer output. Additional information about the relationship of these measures to the likelihoods defined in Box 13.2.3 is given in Box 13.2.4.

R_L^2

A commonly used index in logistic regression (Menard, 2000) follows the form of R^2 from OLS regression, that is, $R^2 = (\text{SS}_{\text{total}} - \text{SS}_{\text{residual}})/\text{SS}_{\text{total}}$, and employs the deviance measures based on measures of likelihood,

(13.2.23) $$R_L^2 = \frac{D_{\text{null}} - D_k}{D_{\text{null}}}$$

R_L^2 ranges between zero and one.[1] The measure is easily calculated from the deviance measures ($-2LL$ measures) from the null model and the model containing k predictors. Simulation work by Estrella (1998) suggests that this measure does not increase monotonically with increases in the odds ratio in the single-predictor case.

Cox and Snell Index

Cox and Snell (1989) offered a second index of overall goodness of model fit that is related to R^2 from OLS regression. The Cox and Snell index is problematic however, in that it does not have a maximum value of one, but rather reaches a maximum value of .75 when the proportion of cases in the sample equals .5.

---

[1] R_L^2 has the same in structure as the normed fit index (NFI) proposed by Bentler and Bonett (1980) in the structural equation modeling context.

Nagelkerke Index

To ameliorate the difficulties with the Cox and Snell index, Nagelkerke (1991) proposed a third measure of overall goodness of fit. The Nagelkerke index corrects the Cox and Snell index by dividing the Cox and Snell index by the maximum possible value that it can attain for a given proportion of cases. Both the Cox and Snell and the Nagelkerke measures are reported in SPSS.

Table 13.2.3 summarizes three R^2 analogs (R_L^2, Cox and Snell, and Nagelkerke) using the example from Table 13.2.2. The null deviance for this model was $D_{\text{null}} = 226.473$. The model deviance for a model containing only the PHYSREC predictor (not given in Table 13.2.2), was $D_{\text{PHYSREC}} = 191.869$, and that containing the four predictors (PHYSREC, KNOWLEDG, BENEFITS, BARRIERS) was $D_4 = 167.696$. R_L^2 for the four predictor model from Eq. (13.2.23) = (226.473 − 167.696)/226.473 = .26, or 26% of the null deviance accounted for by the set of predictors (notice that we are careful to avoid referring to this as a "variance accounted for"). We are tempted to think of this as an effect size measure, scaled in the same manner as R^2 from OLS regression, but the two are not directly the same measure. An inspection of Table 13.2.3 shows the substantial differences in the values of the R_L^2, Cox and Snell, and Nagelkerke indices. The R_L^2 and Cox and Snell measures show much closer agreement with one another than either does with the Nagelkerke index. The Nagelkerke index will always be larger than Cox and Snell, because, as explained earlier, the Nagelkerke index corrects for the fact that Cox and Snell does not reach a theoretical maximum of 1.0. Publications employing these measures should clearly indicate which is being used. If the Nagelkerke index is reported, it is important to explain that the index is adjusted so that the maximum value it can attain is 1.00, an appropriate adjustment relative to Cox and Snell.

Why These Aren't "Variance Accounted For" Measures

Again, we caution that all these indices are not goodness of fit indices in the sense of "proportion of variance accounted for," in contrast to R^2 in OLS regression. This seems puzzling, perhaps, but the explanation is straightforward. Reflect for a moment on the OLS regression model, which assumes homoscedasticity—the same error variance for every value of the criterion. Given homoscedasticity, we are able to think of the total proportion of variance that is error variance in a universal sense, across the full range of Y. In contrast, in logistic regression, we have inherent heteroscedasticity, with a different error variance for each different value of the predicted score $\hat{p}_i$ (recall Eq. 13.2.3). For each value of $\hat{p}_i$, then, we would have a different measure of variance accounted for if we were to apply the R^2 analogs to different portions of

TABLE 13.2.3
Measures of Fit for the Example in Table 13.2.2. Predicting Mammography Compliance

Variables in equation	Measure of fit		
	R_L^2	Cox and Snell	Nagelkerke
PHYSREC alone	.15	.19	.25
PHYSREC, KNOWLEDG	.15	.19	.26
PHYSREC, KNOWLEDG, BENEFITS, BARRIERS	.26	.30	.40

the range of $\hat{p}_i$. Thus, we cannot talk about variance accounted for in a universal sense for logistic regression.

Use of goodness of fit indices in logistic regression is by no means universal as it is in OLS regression, where reporting of R^2 is standard. Traditional users of logistic regression focus on odds ratios for individual predictors. For example, epidemiologists use logistic regression to develop models of specific risk factors for disease. In contrast, psychologists seek overall fit indices based on their grounding in OLS regression. The logistic R^2 analogs are generally not so well behaved statistically as is R^2 in OLS regression. The logistic analogs may fail to reach a maximum of 1; they may fail to track the odds ratios as indices of strength of prediction from individual predictors. According to Hosmer and Lemeshow (2000) the logistic R^2 measures for good logistic regression models are generally smaller than R^2 for good models in OLS regression; this may lead to misperception of logistic regression results as indicating poor models.

13.2.12 Testing Significance of Overall Model Fit: The Likelihood Ratio Test and the Test of Model Deviance

Likelihood Ratio Test of Contribution of the Predictor Set

In OLS regression we have an overall F test for the significance of prediction from the set of k predictors, given in Eq. (3.6.7). The analog to this test in logistic regression is a likelihood ratio χ^2 test for overall model fit.

Recall that in OLS regression, $SS_{regression} = SS_Y - SS_{residual}$, with k degrees of freedom. In logistic regression, we compute a difference between the null and model deviances. This difference is a measure of amount of the null deviance (total deviance that might be accounted for) that is accounted for by a model containing k predictors. The difference is frequently noted as G, for goodness of fit or model prediction:

(13.2.24) $$G = \text{model } \chi^2 = D_{\text{null}} - D_k.$$

BOX 13.2.4
Fit Indices in Terms of Likelihoods

The R_L^2 and Cox and Snell indices of overall model goodness of fit discussed in Section 13.2.11 can be expressed in terms of likelihoods defined in Box 13.2.3. R_L^2 is expressed in terms of likelihoods as follows:

(13.2.25) $$R_L^2 = \frac{\ln L_k - \ln L_{\text{null}}}{\ln L_k - \ln L_{\text{perfect}}}$$

This expression is algebraically equivalent to $R_L^2 = (D_{\text{null}} - D_k)/D_{\text{null}}$ given in Eq. (13.2.23) and is presented in lieu of Eq. (13.2.23) in some texts (e.g., Hosmer and Lemeshow, 2000, which refers to the index as R_{LS}^2).

The Cox and Snell (1989) index of goodness of fit reflects the exact relationship between R^2 and the likelihood ratio statistic in a linear model with normally distributed errors and is given as

(13.2.26) $$R_{\text{Cox Snell}}^2 = 1 - (L_{\text{null}}/L_k)^{2/n}.$$

This measure is distributed as χ^2 with k degrees of freedom, where k is the number of predictors, or, equivalently, the difference in degrees of freedom of the null deviance versus the model deviance. G is a test of the simultaneous contribution of the set of k predictors to the prediction of the dichotomous DV. It can be thought of as a measure of "goodness of contribution from the predictor set." In Table 13.2.2, Part I(B), the model chi square is $G = 226.473 - 167.696 = 58.778$, with $k = 4$ degrees of freedom for the four predictors, and is significant at beyond conventional levels.

As we showed in Box 13.2.3, the null and model deviance are calculated from likelihood ratios. In general, tests that involve likelihood ratios in their calculation are referred to as *likelihood ratio tests* (standardly abbreviated LR); the G statistic in Eq. (13.2.24) is a likelihood ratio test. This test is not the familiar *Pearson* χ^2 test based on contingency tables. Both the likelihood ratio χ^2 test and the Pearson χ^2 test can be computed for logistic regression; both are compared to the same χ^2 critical values for significance. Both are often reported in standard computer output. These two measures depend on different mathematical formulations of the residuals from a logistic regression, as explained in Box 13.2.7. Reporting of the LR tests of model fit in publication is standard practice.[2] Other tests of overall model fit in logistic regression are described in Box 13.2.5.

Is There More Deviance to Be Accounted For: A Test of Model Deviance

The likelihood ratio test we have just considered assesses the contribution to prediction from a set of k predictors, a test of goodness of fit of the k-predictor model. It leaves open the question of whether there is still more deviance that can be accounted for after the inclusion of the k predictors. In work in model testing—for example, in structural equation modeling, introduced in Chapter 12—there is a focus on testing whether models are adequate or whether they leave significant proportions of deviance unaccounted for. Analogous testing of failure of model fit can be carried out in logistic regression. The G statistic developed earlier tells us whether our model containing a set of predictors is better than the null model; here we learn whether the model provides less than perfect fit.

Model deviance, D_k, is a measure of lack of fit to a model including k predictors. In the numerical example in Table 13.2.2, Part I(B), the model deviance = 167.696 with all four predictors in the model. We may test the *null* hypothesis that this model deviance does not differ from that expected by chance alone. The corresponding alternate hypothesis is that the model deviance is systematically larger than expected by chance alone, indicating failure of the predictors to account completely for the criterion (i.e., there is room for improvement in prediction). Here, failure to reject the null hypothesis is the desired outcome to support the adequacy of the regression model. (Note that this is the opposite of classic hypothesis

---

[2]Deviances are labeled "−2 log likelihood" and "−2 log L" in SPSS and SAS, respectively. Deviances carry these labels wherever they appear in output. Both SPSS and SAS begin with a model that contains only the intercept and provide the value of D_{null}, the deviance with the intercept only. Both SPSS and SAS for any particular logistic regression equation containing k predictors provide the value of D_k. For each regression equation, both SPSS and SAS provide a LR χ^2 test of the significance of contribution of the set of predictors to prediction. These tests are labeled "Model chi square" in SPSS and "chi square for covariates" in SAS. SAS also provides the Akaike Information Criterion (labeled AIC) and the Score test (so labeled) as well. SPSS refers to the regression coefficients for predicting the logit from Eq. (13.2.6) as "B"; SAS, as "parameter estimates." SPSS refers to odds ratios as "Exp(*B*)"; SAS, as "Odds Ratio." There is an important discrepancy between SPSS and SAS. If one codes case = 1, noncase = 0, then SPSS by default will predict being a case, whereas SAS will predict being a noncase. Hence, all the coefficients will be of opposite sign in the SPSS versus SAS output, and odds ratios will be inverted; the keyword *descending* in SAS causes SAS to predict being the probability of being a case ($Y = 1$), consistent with SPSS.

testing, in which rejection of the null hypothesis supports the research hypothesis.) The actual value of the model deviance value is tested for significance against a χ^2 distribution with $[n - (k + 1)]$ *df*, where k is the number of predictors, not including the intercept.

For our numerical example, the critical value for the model deviance has $[n - (k + 1)] = [164 - (4 + 1)] = 159\,df$ and is $\chi^2_{.95}(159) = 189.42$; the model deviance $D_k = 167.696$. We do not reject the null hypothesis, and we interpret this as indicating that the four predictors are adequate to account for mammography screening compliance. There is not a significant amount of unaccounted for deviance remaining after prediction from the four predictors.

Sparseness of Data and Tests of Model Adequacy

There is concern that statistical tests in logistic regression may encounter difficulties if data are sparse. To understand sparseness, conceptualize the data of logistic regression as falling into cells defined by a combination of the dependent variable and values of the predictor. In the mammography screening example PHYSREC takes on two values (1, 0) and BENEFITS takes on six values (0, 5) as predictors. Compliance (1, 0) taking on two values as the DV. Thus we have $2 \times 6 \times 2 = 24$ cells. Sparseness refers to having zero frequencies or very small frequencies in some of these cells. With regard to sparseness of data, likelihood ratio tests like G in Eq. (13.2.6) that are based on differences in deviances are not affected when data are sparse. However, with sparse data, the χ^2 test of model deviance just described is no longer distributed as χ^2 and p values from the χ^2 distribution are no longer accurate. In fact, the test of model deviance for the mammography example is subject to the problem of sparseness and should not be trusted.

BOX 13.2.5
The Wald Test, the Score Test, and the Hosmer-Lemeshow Index of Fit

In addition to the likelihood ratio test G described in Section 13.2.12 for overall model fit, there are two other tests, the Wald test and the Score test, that may be applied to testing whether a set of predictors contributes to prediction of an outcome. The Score test is also known as the LaGrange multiplier (LM) test. Both the Score and Wald tests are based on the distribution of likelihoods as a function of the values of estimates of the regression coefficients. Long (1997) provides an extended discussion of model testing in generalized linear models.

Hosmer and Lemeshow (2000) provide an additional goodness of fit test that examines whether the S-shaped function of the logistic regression is appropriate for the observed data. It is based on the familiar Pearson χ^2 in which observed frequencies (f_o) are compared to expected frequencies (f_e) qunder a model. The basis of the test is the predicted probabilities $\hat{p}_i$ of being a case. Data are broken into g categories, and the expected frequency of cases versus noncases in each category based on the $\hat{p}_i$s are computed. The Hosmer and Lemeshow goodness of fit statistic is the Pearson χ^2 for the 2 (case, noncase) by g (categories) table, with $g - 2\,df$. Nonsignificance indicates the fit of observed frequencies of cases in the categories compared to those expected based on the logistic regression. The validity of the test of fit depends on there being large expected frequences in all cells; the power of the test is not high for sample sizes less than 400 (Hosmer and Lemeshow, 2000).

13.2.13 χ^2 Test for the Significance of a Single Predictor in a Multiple Logistic Regression Equation

In OLS regression, we test whether each individual predictor contributes to overall prediction. The test of contribution of an individual predictor in OLS regression is actually an F test of the increment in $SS_{regression}$ by the inclusion of that variable, over and above all other variables. This F test (with degrees of freedom = 1, $df_{residual}$) is the square of the t test for an individual predictor (with degrees of freedom = $df_{residual}$). The t test is defined as the ratio of the predictor to the estimate of its standard error.

Likelihood Ratio Test

In logistic regression, the direct analogy to the OLS F test of gain of prediction for a single predictor is defined on the basis of the difference in model deviances for the model containing k predictors and one containing $(k - 1)$ predictors, with the predictor in question eliminated. This yields a likelihood ratio χ^2 test with 1 df.

(13.2.27) contribution of individual predictor $= D_{(k-1)} - D_k$, with 1 degree of freedom to multiple logistic regression.

Suppose we wished to compute such measures for each of the four predictors in Table 13.2.2, Part I(C). We would require four further regression analyses, each containing only three predictors. These three predictor equations are compared to the four-predictor equation to test for the increment in prediction from the addition of a single predictor. These three predictor regressions, each eliminating a different predictor, are actually not shown in Table 13.2.2; results of these analyses are reported here. For example, with PHYSREC eliminated, the model deviance from the remaining three predictors (BENEFITS, BARRIERS, KNOWLEDG) was $D_{(k-1)} = 184.368$. With all four predictors including PHYSREC in the model, the model deviance was $D_k = 167.696$, as before, so that $\chi^2(1) = 184.368 - 167.696 = 16.672, p < .01$. The χ^2 values for BENEFITS and BARRIERS are 5.287 ($p < .05$), and 13.770 ($p < .01$), respectively. The test for KNOWLEDG does not reach a conventional significance level, $\chi^2(1) = .005$.

Wald Tests

The likelihood ratio χ^2 test just described is the preferred test for the impact of individual predictors in a set of predictors. However, standard computer programs, among them SPSS and SAS, report Wald tests instead for individual predictors. The Wald statistic reported in SPSS and SAS is the ratio of square of the estimate of the regression coefficient B_j to the square of the estimate of its standard error SE_{B_j}

(13.2.28) $$\text{Wald statistic} = \frac{B_j^2}{SE_{B_j}^2}.$$

The test is distributed as χ^2 with 1 degree of freedom under the null hypothesis. The Wald tests for individual predictors are given in Table 13.2.2, Part I(C), and may be compared to the likelihood ratio tests reported earlier for the individual predictors, since both tests are distributed as χ^2 with 1 degree of freedom. In all cases except the zero value of the test for the KNOWLEDG predictor, the likelihood ratio tests exceed the corresponding Wald tests in value. This is consistent with findings that the Wald test is less powerful than the likelihood

ratio test. Wald tests are also biased when data are sparse. Again, the likelihood ratio test is preferred.[3]

13.2.14 Hierarchical Logistic Regression: Likelihood Ratio χ^2 Test for the Significance of a Set of Predictors Above and Beyond Another Set

A common strategy in OLS regression, developed in Section 5.5, is to examine whether a set $\boldsymbol{B}$ of m predictors contributes significant prediction over and above another set $\boldsymbol{A}$ of k predictors. Likelihood ratio (LR) χ^2 tests in logistic regression can be formulated for the same purpose. Hierarchical LR tests of the contribution of a set of m predictors over and above another set of k predictors follow the same structure of differences between deviances. Deviances are computed for the k predictor model, D_k, and the $(m+k)$ predictor model, $D_{(m+k)}$. The difference between these deviances is an LR test for the significance of contribution of the set of m predictors over and above the set of k predictors, with m degrees of freedom.

(13.2.30) contribution of set of m predictors $= D_k - D_{(m+k)}$, with m degrees of freedom over and above another k predictor

In the numerical example of Table 13.2.2, we are most interested in whether psychological factors contribute to screening compliance beyond physician recommendation. Thus PHYSREC constitutes set $\boldsymbol{A}$ with $k = 1$ predictor. The deviance with only PHYSREC as a predictor is 191.869. The second set $\boldsymbol{B}$ consists of $m = 3$ predictors, KNOWLEDG, BENEFITS, and BARRIERS, the psychological factors. The deviance from the four-predictor equation is 167.696. The LR χ^2 test with $m = 3$ degrees of freedom $= 191.869 - 167.696 = 24.173, p < .01$. The psychological factors do add predictability over and above physician recommendation.

Revisiting the Indices of Goodness of Fit

In Section 13.2.11 we reviewed R^2 analogs in logistic regression. We saw (Table 13.2.3) that the indices differed in magnitude from one another for a single model. On the other hand, if we inspect these indices in hierarchical models, they tell a consistent story about the gain in prediction from adding sets of variables. In Table 13.2.3, we present a series of three models: prediction of mammography screening (a) from PHYSREC alone, (b) from PHYSREC plus KNOWLEDG, and (c) from PHYSREC and KNOWLEDG, plus BENEFITS and BARRIERS. All three indices in Table 13.2.3 tell the same story: There is no increment in prediction by the addition of KNOWLEDG to PHYSREC, but the addition of BENEFITS and BARRIERS contributes substantial incremental prediction. Hosmer and Lemeshow (2000) point out the

---

[3]The Wald test for the contribution of an individual predictor is defined in two ways. First is as given in Eq. (13.2.28). Second is as the square root of Eq. (13.2.28) (e.g., Hosmer & Lemeshow, 2000, p. 16):

(13.2.29) $$\text{Wald statistic} = \frac{B_j}{SE_{B_j}}$$

At asymptote, maximum likelihood estimators, including the estimates of the regression coefficients, are normally distributed, meaning that as sample size increases, the distribution of the estimators becomes more and more normal in form. Hence the Wald statistic, as given in Eq. (13.2.29) is distributed as a z test for large samples. The user of statistical software for logistic regression should take care to determine whether the Wald test is given in the χ^2 form of Eq. (13.2.28) or the z test form of Eq. (13.2.29).

utility of the logistic R^2 analogs in the course of model building. Table 13.2.3 illustrates their utility as relative measures for comparison across models.

13.2.15 Akaike's Information Criterion and the Bayesian Information Criterion for Model Comparison

The comparison of models using LR tests described in Section 13.2.14 requires that one model be *nested* within the other. By nested is meant that all the predictors in the smaller model are included among the predictors in the larger model and the identical cases are included in both analysis. Two indices, *Akaike's Information Criterion* (AIC, Akaike, 1973) and the *Bayesian Information Criterion* (*BIC*) provide comparison of model fit in models that are not nested. These two indices also take into account the number of regression coefficients being tested; given equal fit of two models, the more parsimonious model (i.e., having fewer predictors) will have a better AIC fit index. Values of the AIC will be smallest for a model that exhibits good fit with a small number of predictors. (See Box 13.2.6 for computation of the AIC.) The AIC is used by comparing AIC values across estimated models; there is no statistical test of the AIC. The Bayesian Information Criterion (BIC) is a second measure of fit that takes into account the number of predictors. The BIC may be negative or positive in value; the more negative the value of the BIC, the better the fit.

13.2.16 Some Treachery in Variable Scaling and Interpretation of the Odds Ratio

To this point our numerical example has been presented with only unstandardized logistic regression coefficients. The PHYSREC predictor is a dichotomous predictor that ranges from 0 to 1; we now call this $\text{PHYSREC}_{(1,0)}$. The BENEFITS and BARRIERS psychological predictors range from 0 to 5; we call them $\text{BENEFITS}_{(5,0)}$ and $\text{BARRIERS}_{(5,0)}$. Thus, of course, a 1-unit change on $\text{PHYSREC}_{(1,0)}$, which covers the full range of the $\text{PHYSREC}_{(1,0)}$, scale, is not comparable to a 1-unit change on $\text{BENEFITS}_{(5,0)}$, which covers one-fifth of the $\text{BENEFITS}_{(5,0)}$ scale.

Consider the regression equation for predicting mammography screening from PHYSREC and BENEFITS in Table 13.2.4. The same regression equation is shown with four different predictor scalings. Table 13.2.4A gives the analysis of predictors in the original scaling.

BOX 13.2.6
Computation of the Akaike's Information Criterion

Computation of the AIC is based on the likelihood under the model containing $m = k+1$ predictors (including the intercept, L_k)

$$\text{AIC} = \frac{-2 \ln L_k + 2m}{n} \tag{13.2.31}$$

where n is the number of cases. Note the penalty in the numerator for the number of predictors in the model; for two models yielding the same maximum likelihood L_k, the one with the smaller number of predictors will have a smaller AIC.

TABLE 13.2.4
Impact of Predictor Scaling on Regression Coefficients and Odds Ratios

A. Original predictor scaling: physician recommendation (1, 0); benefits (5, 0).

Variable	*B*	*SE*	Wald χ^2	*df*	Significance	Exp(*B*) (odds ratio)
PHYSREC (1, 0)	1.934	.467	17.164	1	.000	6.920
BENEFITS (5, 0)	.694	.229	9.157	1	.002	2.002
Constant	−4.550	1.053	18.687	1	.001	

B. Revised predictor scaling: physician recommendation (1, −1); benefits (5, 0).

Variable	*B*	*SE*	Wald χ^2	*df*	Significance	Exp(*B*) (odds ratio)
PHYSREC (1, −1)	.967	.234	17.164	1	.000	2.631
BENEFITS (5, 0)	.694	.229	9.157	1	.002	2.002
Constant	−3.583	1.015	12.454	1	.001	

C. Revised predictor scaling: physician recommendation (1, 0); benefits (1, 0).

Variable	*B*	*SE*	Wald χ^2	*df*	Significance	Exp(*B*) (odds ratio)
PHYSREC (1, 0)	1.934	.467	17.164	1	.000	6.920
BENEFITS (1, 0)	3.470	1.147	9.157	1	.002	32.129
Constant	−4.550	1.053	18.687	1	.000	

D. Revised predictor scaling: physician recommendation and benefits standardized (z scores), and criterion of compliance unstandardized.

Variable	*B*	*SE*	Wald χ^2	*df*	Significance	Exp(*B*) (odds ratio)
ZPHYSREC	.898	.2168	17.164	1	.000	2.455
ZBENEFIT	.718	.2373	9.157	1	.002	2.051
Constant	−.327	.1941	2.843	1	.092	

Note: Dependent variable is COMPLY (1, 0).

Scaling a Dichotomous Predictor

For physician recommendation, the dummy-variable coding (see Section 8.2) is 1 = recommendation and 0 = no recommendation, which means that a *1-unit* change in the value of the predictor goes from not having a recommendation to having a recommendation. Note that the regression coefficient for $\text{PHYSREC}_{(1,0)}$ is $B_{\text{PHYSREC}(1,0)} = 1.934$ and the odds ratio is $e^{B_{\text{PHYSREC}(1,0)}} = 6.92$. The logit for obtaining a mammogram increases by additive amount of 1.934 when a woman receives a recommmendation for screening from her physician, and the odds of her obtaining a mammogram are multiplied by 6.92. Recall that in general the odds ratio is the amount by which the odds are multiplied for a 1-unit increase in the predictor (here, of receiving a recommendation for a mammogram).

Now we repeat the analysis, but with an unweighted effects code form of the PHYSREC predictor (see Section 8.3), that is, 1 = recommendation, −1 = no recommendation. The change in interpretation of regression coefficients for unweighted effects versus dummy coding

is the same as in OLS regression. The results of an analysis with unweighted effects coded $\text{PHYSREC}_{(1,-1)}$, that is, 1 = yes; −1 = no, are given in Table 13.2.4B. First, as in OLS regression, the regression coefficient for $\text{PHYSREC}_{(1,-1)}$ is .967, exactly half of the value of the corresponding coefficient in the first dummy-coded analysis. What value is the odds ratio for $\text{PHYSREC}_{(1,-1)}$ relative to that for $\text{PHYSREC}_{(1,0)}$?

For $\text{PHYSREC}_{(1,0)}$, $B_{(1/0)} = 1.934$ and e^B = odds ratio = $e^{1.934} = e^{2(.967)} = 6.92$.
For $\text{PHYSREC}_{(1,-1)}$, $B_{(1/-1)} = .967$ and e^B = odds ratio = $e^{.967} = 2.63$.

Note that $2.63 = \sqrt{6.92}$. Halving the regression coefficient corresponds to taking the square root of the odds.

The odds ratio based on effects coding (1, −1) does not inform us directly of odds that a woman will receive a mammogram if she does versus does not receive a physician's recommendation. To get this odds ratio directly we must use the (1, 0) coding of physician recommendation. With the (1, −1) effects codes, a 1-unit increase in the $\text{PHYSREC}_{(1,-1)}$ predictor is only half the distance from no recommendation (−1) to recommendation (1). The regression coefficient from the effects coded predictor can be converted to the odds ratio for the impact of recommendation on odds of screening. First, double the regression coefficient from the effects coded analysis (since $B_{(1/-1)} = .967, 2 \times B_{(1/-1)} = 1.934$). Then *exponentiate the doubled coefficient* to get the proper odds ratio for the increase in odds of mammography screening when one has received a physician recommendation ($e^{1.934} = 6.92$).

Treachery in Scaling a Continuous Predictor

We often combine medical or demographic variables that are dichotomous (male, female; African American, Caucasian; family history, no family history; physician recommendation, no physician recommendation) with continuous variables such as psychological variables (e.g., perceived benefits, barriers) that are scaled and cover a range well beyond (0, 1). If we ignore the difference in scaling, we may misinterpret the smaller regression coefficients that result from prediction from the psychological variables with larger ranges as indicating weaker prediction from the psychological variables. The differences in coefficient magnitude are accentuated when we move to odds ratios. To the uninitiated or casual consumer of logistic regression, who quickly scans a column of odds ratios, the binary variables may appear much more powerful than the psychological variables. The benefits predictor $\text{BENEFITS}_{(5,0)}$ has a 5-point range, so a 1-unit change in benefits covers only a fifth of the scale. A 1-unit change in $\text{PHYSREC}_{(1,0)}$ is from no recommendation to a recommendation. We rescale the benefits scale to have the range from 0 to 1 by dividing each benefits score by 5, yielding $\text{BENEFITS}_{(1,0)}$. Having divided the benefits scale by 5, the regression coefficient for $\text{BENEFITS}_{(1,0)}$ is multiplied by 5; $B = .694 \times 5 = 3.47$, as shown in Table 13.2.4C. Then we rescale the odds ratio: $e^{B_{\text{rescaled}}} = e^{3.4698} = 32.13$. If a woman moves from the lowest to highest score on BENEFITS (perhaps a goal for an intervention to increase screening rates), her odds of being screened increase by a factor of 32; perceived benefit appears to be a very powerful psychological variable.

This example illustrates the importance of addressing predictor scaling when comparing odds ratios. Hosmer and Lemeshow (2000) suggest that when working with a continuous predictor, one should consider the magnitude of change in units on that predictor that would be meaningful and report coefficients and odds ratios associated with that change. For example, if a 2-unit change seemed meaningful for the BENEFITS scale, then one would report the rescaled B coefficient and odds ratio for a 2-unit change. For a 1-unit change on the 5-point BENEFIT scale (Table 13.2.4A), the logit increases by $\text{B}_{\text{BENEFITS(5,0)}} = .694$. For a w-unit increase in benefits, recall that the amount of change in the logit is simply wB. Here, for a

2-unit increase in $\text{BENEFITS}_{(5,0)}$, the increase in the logit is $2 \times .694 = 1.388$. The odds are multiplied by the value e^{wB_j} for a w-unit increase in the predictor. For a 2-point increase in $\text{BENEFITS}_{(5,0)}$, the odds ratio is $e^{2\times.694} = 4.01$, or, equivalently, the odds of compliance are multiplied by 4.01. If one reports coefficients and odds for greater than a 1-unit change on a scale, what is being reported should be clearly explained to the reader.

Standardized Regression Coefficients

The use of standardized regression coefficients is the familiar way in OLS regression to address the issue of differential scaling of predictors. However, standardized regression coefficients are a matter of some complexity in logistic regression. In OLS regression, we compute the standardized regression coefficient β_j from the corresponding unstandardized coefficient B_j as follows (rearranged from Eq. 3.2.5):

(13.2.32)
$$\beta_j = B_j \frac{sd_X}{sd_Y},$$

where sd_X is the standard deviation of the predictor, and sd_Y is the standard deviation of Y. Using this equation in logistic regression poses problems, because in the linear form of logistic regression, the variable being predicted is the logit of the underlying probability of being a case and not the observed Y (case, noncase). Thus, to standardize coefficients, we would require the standard deviation of this underlying logit. Although some software packages do report standardized coefficients, it may be unclear precisely how standardization is accomplished. If the analyst wishes to report a standardized solution, then a simple strategy exists: Standardize the predictors and estimate the unstandardized logistic regression (Pampel, 2000). The resulting coefficients give the change in the logit for a one standard deviation change in the predictors. The coefficients are "semistandardized," that is, standardized only on the predictor side. Use of this approach should be explained in publication, due to the unusual semistandardization. In Table 13.2.4D, the data are reanalyzed with the dependent variable retained as COMPLY (1, 0), and both predictors first converted to standardized scores (z scores, i.e., ZPHYSREC and ZBENEFIT for the z scores associated with PHYSREC and BENEFITS, respectively). The resulting regression coefficients and odds ratios are approximately equal, suggesting relatively equal strength of the predictors. There is a downside to this approach: The coefficients and odds ratios for the dichotomous predictors lose their straightforward interpretation because the values of 0 and 1 in the z score scale metric no longer represent the two categories of the scale. This leads some analysts to object to the approach. Menard (1995) and Pampel (2000) provide clear discussions of standardization in logistic regression, including standardization of predicted logits to come closer to full standardization.

Another way of thinking about the contributions of two predictors, as in OLS regression, is to ask whether each contributes prediction over and above the other. Thus we may ask whether BENEFITS contributes to reduction in deviance over and above PHYSREC and vice versa, according to Eq. (13.2.27).

13.2.17 Regression Diagnostics in Logistic Regression

Section 10.3 provides a full exposition of regression diagnostics and their use in OLS regression. These regression diagnostics are based on the assumption that the residuals in an analysis are normally distributed, according to the general linear model. Regression diagnostic measures of leverage, distance, and influence have been generalized to logistic regression in classic work by Pregibon (1981). The reader is cautioned that the generalizations are not complete, due to

the complexity of the logistic regression model. Graphical diagnostics are more difficult to interpret because of the dichotomous distribution of the criterion. Informative discussions of diagnostics in the logistic regression context are found in Collett (1991), Fox (1997), Hosmer and Lemeshow (2000), Long (1997), Menard (1995), and Ryan (1997).

The present section does not give a full explication of diagnostics in logistic regression. Rather, since regression diagnostics are regularly reported in logistic regression software, this section aims to highlight divergences between OLS and logistic regression diagnostics and to caution analysts concerning issues in the use of diagnostics in logistic regression. A review of Section 10.3 is recommended to set this section in context.

Leverage in Logistic Regression

Recall that in OLS regression, the leverage of a point, h_{ii}, is a measure of the potential of a case to influence regression results, (i.e., to change the regression coefficients by its presence in the data set). Leverage in OLS regression is based solely on scores on the predictors. In OLS regression the farther a case is from the centroid of the points on the predictors (the means of all the predictors), the greater is the leverage (see Section 10.3.1, Eq. 10.3.1). In OLS regression the value of the dependent variable (DV) for the case has no effect on the leverage measure. Pregibon (1981) provided a generalization of the measure of leverage to logistic regression. In this generalization, the leverage values h_{ii} depend on the DV scores as well as on the predictors, yielding a discontinuity between the definitions of leverage in OLS and logistic regression. A further discontinuity exists between leverage measures in the two cases. Leverage in OLS regression is greatest for those cases most extreme in the predictor space (i.e., farthest from the centroid of points). However, leverage in logistic regression increases as the extremeness of cases increases up to a point and then diminishes rapidly for the most extreme cases. In other words, a very extreme case can have a lower leverage score than a less extreme case! Hosmer and Lemeshow (2000) recommend that one should examine the predicted probability $\hat{p}_i$ for a case before interpreting leverage measures; only for $\hat{p}_i$ between .10 and .90 are leverages assured to increase with increasing distance of the point from the centroid of the predictor space.

Residuals in Logistic Regression

Residuals play a central role in regression diagnostics. The fact that the residuals in OLS regression are theoretically normally distributed yields great simplification in regression diagnostics. The distribution of residuals in OLS is expected to be independent of the predicted score $\hat{Y}$, so the size of residuals can be interpreted in the same manner across the range of the predictor. Further, if the homoscedasticity assumed in OLS regression holds, the variance of the residuals is constant for all values of the predicted score; residuals associated with different predicted scores may be directly compared. In logistic regression, the size of residuals and their variance is dependent upon the predicted probability, $\hat{p}_i$; the residuals are non-normal and heteroscedastic. This adds a layer of complexity to the analysis of residuals.

Deviance residuals. In OLS regression the squared residual of each case $(Y - \hat{Y})^2$ contributes to $SS_{residual}$, the overall measure of lack of fit of the OLS regression model. The residuals from individual cases form the basis of a number of regression diagnostic measures in OLS regression.

In logistic regression, a *deviance residual* is computed for each case; it measures the numerical contribution of the case to the overall model deviance D_k, the overall measure of lack of fit of a logistic regression model. Adding to the complexity of residual diagnostics is the fact that there is a second type of residual in logistic regression, the *Pearson residual.* Both the

deviance and Pearson residuals are used in the computation of diagnostic measures in logistic regression that are analogs of residual diagnostics in OLS regression. There is a preference in the literature for the use of deviance residuals over Pearson residuals for two reasons: (1) deviance residuals are closer to normally distributed, and (2) Pearson residuals are unstable when $\hat{p}_i$ is close to zero or one. Deviance residuals pose problems for interpretation, however, in that the expected value of the deviance residual (the average deviance residual) depends on the value of $\hat{p}$, the overall probability of a case; the value of the residual cannot be considered independent of this overall probability. Details of the computation of the deviance and Pearson residuals are given in Box 13.2.7.

Externally studentized residuals. In OLS regression externally studentized residuals are useful in identifying outliers. The externally studentized residual for each case is based on a regression analysis in which the case in question has been omitted (see Section 10.3.2). Externally studentized residuals in logistic regression have been defined for both deviance and Pearson residuals. Those based on deviance residuals are asymptotically normally distributed; the difficulty for psychology is that we have small sample sizes—we can hardly assume asymptotic distributions.

In Section 10.3.2 t tests were provided for externally studentized residuals (see Eq. 10.3.4). In addition, suggestions were made for cut scores, beyond which a residual is seen as signaling a potentially problematic case. We do not see these same t tests and suggestions for cut scores in the logistic regression context. The definition of cut scores for residuals is typically justified by normal theory, based on the number of standard deviations on a normal curve. Recall again that residuals in logistic regression are not normally distributed; in fact, the residuals follow a binomial distribution for each value of $\hat{p}_i$.

Influence in Logistic Regression

Influence diagnostics, that is, measures of the extent to which individual cases affect the regression coefficient estimates (Section 10.3.3) have been extended to logistic regression; these include an analog of Cook's distance for impact of a case on the overall fit of the regression model and *DFBETA* for the impact of a case on individual regression coefficients. *DFBETA*s are useful in logistic regression for identifying cases that may have an undue impact on particular regression coefficients.

Graphical Approaches to Diagnostics

A number of graphical displays, among them index plots of residuals, normal probability plots of residuals (q-q plots) and added variable plots have been suggested for use in diagnostics in logistic regression (for a discussion of the use of these graphs in OLS regression, see Chapter 4). Collett (1991) and Cook and Weisberg (1999) provide an extensive review of graphical displays applied to logistic regression. These plots may also be extended to model checking, that is, to examining whether the fitted model is an adequate representation of the data. As Collett (1991) pointed out, however, since residuals in logistic regression are generally not normally distributed, a correct logistic regression model may yield plots that suggest difficulties with the model. For example, a normal probability plot of deviance residuals may well show residuals deviating from a straight line even with a well-fitting model.

How to Proceed With Diagnostics in Logistic Regression

In light of these caveats, the reader must be cautious in drawing conclusions from diagnostics in logistic regression. One may examine measures of leverage for cases for which $\hat{p}_i$ falls between .10 and .90, remembering that beyond these values leverage no longer reflects distance

from the centroid of the X space. Studentized residuals that are very large relative to the rest of the sample may reflect cases that are problematic. Cases that will have the largest residuals are those with extreme predicted probabilities, either close to 0 or close to 1. They will be cases that do not follow the model (e.g., a student who by virtue of exceptionally strong scores on a set of academic ability predictors in a model should succeed in college but who, unlike other students with these same scores, fails miserably). *DFBETA*s may also flag potentially problematic cases, which may well be the same cases that simply do not follow the model. Graphical displays of diagnostics will aid detection of potentially problematic points. The reader should take the view, however, that diagnostics in logistic regression are not so straightforward as in OLS regression. Even greater caution should be applied before cases are deleted based on diagnostic measures in logistic regression than in OLS regression.[4]

BOX 13.2.7
Deviance and Pearson Residuals in Logistic Regression

The Pearson residual for case i is given as (Long, 1997, p. 98)

$$r_i = \frac{Y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}. \tag{13.2.33}$$

Since the residuals exhibit heteroscedasticity, the residual $(Y_i - \hat{p}_i)$ is divided by its binomial standard error, which depends upon $\hat{p}_i$.

The *deviance residual* is defined as

$$d_i = \text{sign}\,(Y_i - \hat{p}_i)\sqrt{-2[(-Y_i \ln(\hat{p}_i) - (1-Y_i)\ln(1-\hat{p}_i)]}, \tag{13.2.34}$$

where sign $(Y_i - \hat{p}_i)$ is the sign of the discrepancy between the observed score Y (1, 0) and the predicted probability. The expression compares the observed Y_i to the predicted $\hat{p}_i$ score.

Neither the Pearson nor the deviance residual given here is standardized; that is, neither has a standard deviation of 1. Both may be standardized by dividing each value by the value $\sqrt{1-h_{ii}}$, yielding *standardized Pearson and deviance residuals*, respectively.

[4]The naming of the diagnostics identified here is inconsistent across software packages. These are the terms used in SPSS and SAS: (a) leverage h_{ii} (LEVER in SPSS, H in SAS); (b) Pearson residual, Eq. (13.2.33), (ZRESID in SPSS, RESCHI in SAS); (c) deviance residual, Eq. (13.2.34), (DEV in SPSS; RESDEV in SAS); (d) externally studentized residual (SRESID in SPSS); (e) *DFBETA* (DFBETA in SPSS; DFBETAS in SAS); (f) analog of Cook's distance (COOK in SPSS; CBAR in SAS).

A final complication in the computation of diagnostics in logistic regression is that they are computed in one of two ways: (1) based on individual cases in the data set, as they are presented here, or, (2) based on aggregated data (see Hosmer and Lemeshow, 2000 for a discussion). For the aggregated measures, data are broken into categories, where a category consists of all those cases that have the same values on the predictors. With continuous predictors, categories will be sparse (i.e., contain few cases). Basic diagnostic building blocks, specifically residuals and leverage values, are defined somewhat differently depending on whether aggregation is or is not used; further, the asymptotic distributions of the measures differ depending on aggregation strategy. Numerical results of diagnostics differ depending on aggregation. It may be unclear what strategy is used for computation in any particular software package, adding a layer of uncertainty to the meaning of specific values of the measures.

13.2.18 Sparseness of Data

We defined sparseness of data in Section 13.2.12 and indicated how sparseness biases tests of model deviance and Wald tests, as well. Sparseness (having numerous cells with zero counts) also decreases the power of statistical tests in logistic regression. In addition, sparseness may cause difficulties in estimation. The analysis may not converge (see Section 13.2.9). Or, estimates of regression coefficients and their standard errors may "blow up" (i.e., become huge), signaling estimation difficulties.

13.2.19 Classification of Cases

Once a logistic regression has been accomplished, the predicted probabilities $\hat{p}_i$ for each individual may be used to generate a predicted case status (i.e., whether the individual is predicted to be a case or a noncase). This is accomplished by choosing a cut score on the $\hat{p}_i$ continuum, above which an individual is classified as a case; otherwise, noncase. Then the predicted status (case, noncase) can be compared with the observed case status to determine how well the logistic regression model recovers case status. The classification of cases is ancillary to logistic regression and is also carried out following other statistical procedures, particularly discriminant analysis, which is described in Section 13.2.1. Classification is useful when statistical models are developed to make decisions among individuals (e.g., hiring decisions based on a battery of test scores). Classification provides another way of characterizing the goodness of fit of a logistic regression model. A description of classification is given in Box 13.2.8, along with a numerical example. The critical issue of selection of a cutoff score for classification is discussed.

BOX 13.2.8
How Accurately Does a Logistic Regression Model Identify Cases?

Suppose we compute for each case the predicted probability of being a case, $\hat{p}_i$. Then we classify each individual as case versus noncase based on whether the $\hat{p}_i$ score exceeds some cutoff. These classifications are *statistical classifications*, based on logistic regression model. We then compare the statistical classifications to the actual classifications in a 2×2 *classification table*, shown in Table 13.2.5. The number of correct statistical classifications is the sum of the *correct rejections* (classifying a noncase as a noncase) plus *hits* (classifying a case as a case) in the parlance of statistical decision theory. In epidemiological terms, one can examine *sensitivity*, the proportion of actual cases who are classified as cases, and *specificity*, the proportion of noncases who are classified as noncases (Fleiss, 1981). Such an analysis may be informative if the goal of the logistic regression analysis is, in fact, classification of cases, as in computerized medical diagnosis, rather than the derivation of a model of "caseness" based on a set of predictors.

A most critical issue in classification is the choice of cutoff on the $\hat{p}_i$ continuum. Neter, Kutner, Nachtsheim, and Wasserman (1996) have suggested three criteria: (1) use a cutoff of .5, such that if the predicted probability of being a case is greater than .5, the individual is classified as a case; (2) select the cutoff that leads to the most accurate classification, through a process of trial and error; and (3) use some a priori information about the proportion of cases versus noncases in the population (e.g. the actual proportion of women who are in compliance with mammography screening guidelines in the population). The choice of cutoff will change the sensitivity versus specificity of the classification scheme (this is analogous to the inverse relationship between Type I and Type II error in hypothesis testing as one changes the critical value of a statistical test).

(*Continued*)

TABLE 13.2.5
Classification Results From the Mammography Example

A. Classification table.

		Predicted class membership	
		0	1
Observed class membership	0	Correct rejections	False alarms
	1	Misses	Hits

B. Classifications under various cut scores.

1. Cut $= \hat{p}_i = .50$.

Observed class membership	Predicted class membership 0	1	Total	Number correct	R^2_{Count}	R^2_{AdjCount}
0	60	28	88			
1	23	53	76	113	.68	.33
	83	81	164			

2. Cut $= \hat{p}_i = .20$.

Observed class membership	Predicted class membership 0	1	Total	Number correct	R^2_{Count}	R^2_{AdjCount}
0	40	48	88			
1	4	72	76	112	.68	.32
	44	120	164			

3. Cut $= \hat{p}_i = .80$.

Observed class membership	Predicted class membership 0	1	Total	Number correct	R^2_{Count}	R^2_{AdjCount}
0	81	7	88			
1	50	26	76	107	.65	.25
	131	33	164			

Note: R^2_{Count} is the unadjusted proportion of correct classifications, the sum of the main diagonal elements divided by the total $n = 164$.
R^2_{AdjCount} is the adjusted proportion of correct classifications, "the proportion of correct guesses beyond the number that would be correctly guessed by choosing the largest marginal" (Long, 1997, p. 108).

A number of measures of the agreement of two classifications have been developed, among them the phi coefficient, weighted kappa (Cohen, 1968a), and Goodman and Kruskal's λ (Goodman & Kruskal, 1979). Treatments of these measures are given in Fleiss (1981), Kraemer (1985; 1988), and Menard (2001). Kraemer, Kazdin, Offord, Kessler, Jensen, and Kupfer (1999) provide a useful explication of such measures and their interrelationships.

(Continued)

Classification accuracy depends on the *base rate* of a phenomenon in the population, that is, the proportion of individuals in the population who are cases. If, for example, a base rate of 80% of the adults in a community suffer from allergies during a particular month, then a physician has an 80% chance of being correct in diagnosing a new patient from that community as having allergies without ever seeing the patient!

Two simple measures of classification accuracy (Long, 1997) are the *proportion of correct classifications* (hits plus correct rejections) and the *proportion of additional classification accuracy* gained by the logistic regression scheme, over and above classification accuracy based on the distribution of the outcome alone (e.g., above the base rate of 80% in the above example). The former measure does not handle the base-rate issue; the latter does. These are given as follows, where n is the total number of cases, and $n$max is the total number of cases in the larger observed category:

$$R^2_{\text{Count}} = \frac{\text{hits} + \text{correct rejections}}{n} \tag{13.2.35}$$

for the uncorrected proportion correct, and

$$R^2_{\text{AdjCount}} = \frac{\text{hits} + \text{correct rejections} - n\text{max}}{n - n\text{max}} \tag{13.2.36}$$

for the proportion gain in prediction accuracy over and above that provided by classification based on marginals alone. This is Goodman and Kruskal's λ (Long, 1997).

Computations are given for the mammography data in Table 13.2.5B. Part B1 gives the classification table for a cut score of $\hat{p}_i = .50$. There are 88 cases observed to be noncases, and 76 cases, of the 164 cases in all. In all 60 of the noncases are correctly classified, along with 53 of the cases. The proportion of correct classifications with a .50 criterion is .68, that is, $R^2_{\text{Count}} = (53 + 60)/164$. The adjusted count is $R^2_{\text{AdjCount}} = (53 + 60 - 88)/(164 - 88) = .33$, where 88 is the number of cases in the larger observed class. This .33 indicates that the prediction scheme produces classification accuracy that is 33% higher than by merely guessing that all cases arise from the more frequent category.

Table 13.2.5 shows how insensitive measures of overall classification accuracy are to the choice of cutoff, but how dramatically measures of sensitivity (proportion of actual cases classified as cases) and specificity (proportion of actual noncases classified as noncases) are affected by cutoff choice. Part B2 gives the classification results for a cut score of $\hat{p}_i = .20$; Part B3, for a cut score of $\hat{p}_i = .80$. R^2_{Count} essentially does not change as the cut score is moved from .20 to .80. However, sensitivity and specificity change dramatically as the cutoff is moved. The sensitivity, or proportion of actual cases classified as cases, decreases from sensitivity of $72/76 = .95$ when the cut score is .20 (Part 2) and fully 120 of the 164 individuals are classified as in compliance to sensitivity $= 26/76 = .34$, when the cut score is .80 (Part 3) and only 33 of 164 individuals are classified as in compliance. Conversely, specificity increases from $40/88 = .45$ when the cut score is .20 (Part 2) to 81/88 when the cut score is .80 (Part 3). The issue of cutoff is well illuminated by considering the use of medical diagnostic tests; a change in cutoff may well determine whether an individual is diagnosed or not as having a disease.

Classification results reflect the adequacy of the model in distinguishing cases from noncases (once accuracy that can be achieved from the base rate by just predicting the larger category is taken into account). These results provide a useful adjunct to other meaures of fit in logistic regression. However, sometimes we may have a well-fitting

(*Continued*)

model in terms of predicted probabilities of being in a category and simultaneously low classification accuracy, above and beyond the base rate. For example, suppose our model is very accurate in predicting that a person with a certain profile has a predicted probability of being a case of $\hat{p}_i = .50$. For this person, we have a 50/50 chance of being wrong in classification based on the model because we can only classify this person as a case versus a noncase. Poor classification results in the face of a well-fitting model may particularly occur when we are predicting rare events. A mathematically based science of classification has been developed in which classification rules take into account prior odds of class membership and the costs of misclassification. Tatsuoka (1988) provides an introduction to misclassification models.

13.3 EXTENSIONS OF LOGISTIC REGRESSION TO MULTIPLE RESPONSE CATEGORIES: POLYTOMOUS LOGISTIC REGRESSION AND ORDINAL LOGISTIC REGRESSION

We may encounter dependent variables for which the outcomes fall into several nonordered categories. For example, we may wish to account for the college (business, engineering, liberal arts) into which a student matriculates as a function of ability and interest scores. Alternatively, the outcome categories may be ordered, as when students express one of three levels of interest in being liberal arts majors (low, moderate, and high). *Polytomous logistic regression* (also refered to as *multinomial logistic regression*) is used to examine an outcome variable consisting of nonordered responses. A second approach is the analysis of *nested categories*, in which contrasts among categories, like familiar contrasts in ANOVA and OLS regression (Chapter 8), are accomplished in a series of dichotomous outcome logistic regressions. Categories may be either ordered or not. Third, *ordinal logistic regression* is used to examine an outcome variable consisting of ordered categories.

13.3.1 Polytomous Logistic Regression

Expositions of polytomous logistic regression of increasing mathematical detail are given in Menard (2001), Hosmer and Lemeshow (2000), who use the term *multinomial logistic regression*, and Fox (1997), along with numerical examples. Assume that as a dependent variable we are comparing a group of college students who are undecided about a major (major = 0) to those who have elected a humanities (major = 1) or a science (major = 2) major. We wish to distinguish these $g = 3$ nonordered groups on the basis of a series of interest test measures $X_1, X_2, \ldots, X_k$. We begin by choosing one group to serve as a baseline group, here those students who are undecided about a major. The data from all three groups are entered into a single polytomous logistic regression analysis. In the course of the analysis, $(g - 1)$ distinct logistic regression functions, all with the same k predictors, are computed for the g groups (here, $g - 1 = 2$ for the three student groups). The first contrasts the humanities majors with the undecided students; the second contrasts the science majors with the undecided students. The logistic regression functions are combined into one overall polytomous regression equation that includes the intercepts from the $(g - 1)$ logistic regression functions plus the $(g - 1)k$ regression coefficients for the k predictors in the $(g - 1)$ regression functions. Testing of model fit of this combined regression equation proceeds along the lines previously described for likelihood ratio tests (see Section 13.2.12, Eq. 13.2.26). There is one overall likelihood ratio χ^2 test (G test) of fit of the model, with $(g - 1)k$ degrees of freedom. Tests for the impact of individual

predictors proceed as before. Indices of fit such as R_L^2 in Eq. (13.2.23) can be computed for the full model. Numerical examples of polytomous logistic regression are provided in Menard (2001), Hosmer and Lemeshow (2000), and Fox (1997).[5]

13.3.2 Nested Dichotomies

An alternative to polytomous logistic regressions is a series of dichotomous logistic regressions (Fox, 1997). The particular dichotomous logistic regressions are a series of nested regressions based on a series of *nested partitions* of the multiple categories represented by the dependent variable. The partitions follow the patterns of sets of orthogonal contrasts, described in Section 8.5. For the example of majors (undecided = 0; humanities = 1; science = 2), we might first consider a partition of the undecided versus the other students (0 versus 1 + 2) and then a second partition in which we contrast the humanities majors versus the science majors. The patterns of nested contrasts of the categories follow from the logic or theory underlying the research. In the case of predicting choice of major from a number of academic interest measures, we might predict that undecided students have a lower overall level of interest in academic subjects than those who have declared any major (contrast 1); we might then predict that differential patterns of interest predict the choice of science versus humanities major (contrast 2). As Fox (1997) shows, a variety of partitions can be generated from a set of categories. Some do not imply an order among the full set of categories (e.g., for four categories, 1 + 3 versus 2 + 4, followed by 1 versus 3 and 2 versus 4). Other series imply an underlying order (e.g., for a series, 1 versus 2 + 3 + 4, followed by 2 versus 3 + 4, followed by 3 versus 4, referred to as *continuation dichotomies* (Fox, 1997) or as following a *continuation-ratio model* (Greenland, 1998). The individual contrasts are treated in separate dichotomous logistic regressions. For the example of majors, the first logistic regression would be of the group coded 0 versus the group formed by combining the groups coded 1 and 2. The second would contain only the groups coded 1 versus 2. Each analysis yields a likelihood ratio χ^2 test (G test). By virtue of the fact that the contrasts are orthogonal, the likelihood ratio tests from the two analyses may be pooled into an overall fit statistic by adding the likelihood ratio χ^2 values and the corresponding degrees of freedom.


CH13EX02

Table 13.3.1 presents the analysis of an ordinal outcome variable Y, the steps (STEPS) women have taken to obtaining a mammogram following an intervention (versus no-intervention control) to increase mammography screening. Data are a subset of those presented in Aiken, West, Woodward, Reno, and Reynolds (1994). Four ordered categories of the outcome are (1) to do nothing about getting a mammogram, (2) to contact a health professional about mammograms, (3) to make an appointment for a mammogram, (4) to actually obtain a mammogram. The single predictor is whether the woman participated in a psychosocial intervention to increase mammography screening or served as a control subject (INTERVEN), with participants coded 1 and control subjects coded 0. Three continuation dichotomies were created from the STEPS outcome: (a) S123V4, which measures whether women *obtained* a mammogram [category 4] versus not [categories 1, 2, 3]; (b) S12V3, which measures whether a woman made an appointment [category 3] versus did nothing or contacted a health professional [categories 1, 2]; and (c) S1V2, which measures whether a woman contacted a health professional [category 2] versus did nothing [category 1]. Three separate dichotomous logistic regression analyses are presented in Table 13.3.1A, B, and C for S123V4, S12V3, S1V2, respectively. From the analysis of S123V4 (Table 13.3.1A), we see that the odds of obtaining a mammogram were increased by a factor of 4 (odds ratio = 4.15) if women participated in

[5]SPSS 10.0 has a procedure for polytomous regression. Epidemiologists recommend STATA for handling polytomous data.

TABLE 13.3.1
Three Approaches to Analysis of Ordinal Outcome Variable of STEPS to Compliance as a Function of Intervention (INTERVEN).

A. Dichotomous logistic regression predicting S123V4 from INTERVEN, i.e., obtaining a mammogram (4) versus all other categories (1, 2, 3)

Null deviance	170.326	$R_L^2 = .07$
Model deviance	159.081	
Model chi square	11.245	1 *df* $p < .001$

	Variables in the Equation					
Variable	*B*	*SE*	Wald	*df*	Significance	Exp(*B*)
INTERVEN	1.423	.461	9.504	1	.002	4.148
Constant	−1.838	.407	20.403	1	.001	

B. Dichotomous logistic regression predicting S12V3 from INTERVEN, i.e., having an appointment for a mammogram (3) versus taking no action or contacting a health professional (1, 2)

Null deviance	39.391	$R_L^2 = .04$
Model deviance	37.908	
Model chi square	1.483	1 *df* $p = .223$

	Variables in the Equation					
Variable	*B*	*SE*	Wald	*df*	Significance	Exp(*B*)
INTERVEN	1.255	1.137	1.218	1	.270	3.508
Constant	−3.761	1.011	13.829	1	.001	

C. Dichotomous logistic regression predicting S1V2 from INTERVEN, i.e., contacting a health professional concerning mammograms (2) versus doing nothing (1)

Null deviance	120.090	$R_L^2 = .05$
Model deviance	114.387	
Model chi square	5.704	1 *df* $p = .017$

	Variables in the Equation					
Variable	*B*	*SE*	Wald	*df*	Significance	Exp(*B*)
INTERVEN	1.071	.461	5.408	1	.020	2.919
Constant	−1.194	.361	10.940	1	.001	

D. Ordinal Regression Analysis predicting STEPS continuum from INTERVEN

Null deviance	329.806	$R_L^2 = .05$
Model deviance	311.406	
Model chi square	17.479	1 *df* $p < .001$

TABLE 13.3.1
(Continued)

	Variables in the Equation					
Variable	*B*	*SE*	Wald	*df*	Significance	Exp(*B*)
Threshold-1	−1.872	.324	33.295	1	.001	
Threshold-2	−1.692	.318	28.274	1	.001	
Threshold-3	−.601	.288	4.365	1	.037	
INTERVEN	1.464	.354	17.106	1	.021	4.324

Score Test of Proportional Odds Assumption
Chi-square = .032 with 2 *df* ($p = .984$)

E. OLS Regression, predicting STEPS continuum from INTERVEN
(1 = intervention, 0 = control)
$R^2 = .12$, $F(1, 137) = 17.899$, $p < .001$

	Variables in the Equation				
Variable	*B*	*SE*	Beta	*t*	Significance of *t*
INTERVEN	.898	.212	.340	4.231	.001
(Constant)	1.647	.169		9.748	.001

Note: Analyses A, B, and C are continuation category analyses. Analysis D is an ordinal logistic regression. Analysis E is an ordinary least squares regression.

an intervention, model $\chi^2(1) = 11.24, p < .01$. From the analysis of S12V3 (Table 13.3.1B), we see that the odds of making an appointment were increased by a factor of between 3 and 4 (odds ratio = 3.51) if women participated in an intervention, model $\chi^2(1) = 1.48$, *ns*. (Lack of model significance is attributable to the very small group ($n = 5$) who made an appointment for a mammogram but did not actually obtain the mammogram by the time of data collection). Finally, from the analysis of S1V2 (Table 13.3.1C), we see that the odds of contacting a health provider were increased by a factor of 3 (odds ratio = 2.92) if women participated in an intervention, model $\chi^2(1) = 5.70, p < .05$. The model χ^2 values are summed to yield an overall test of the impact of the intervention on the propensity of a woman to obtain a mammogram, $\chi^2 = 11.25 + 1.48 + 5.7046 = 18.43$ with 3 *df*, $p < .01$. From these analyses, we may conclude that the intervention does have a strong impact on propensity to obtain a mammogram. Further, we may conclude that the odds of moving up the steps toward a mammogram are similarly impacted by the intervention across the continuum, since all three odds ratios were within close range of one another. This last result is clearly not a necessary outcome. It might have been the case, for example, that the intervention was powerful in stimulating women who had already contacted a health professional about a mammogram to make an appointment for a mammogram or to obtain a mammogram (S12V3) and (S123V4), respectively, but that the intervention did nothing to stimulate women to take the first step of contacting a health professional (S1V2).

13.3.3 Ordinal Logistic Regression

Suppose we believe that the steps are, in fact, a logically ordered behavioral progression of movement toward obtaining a mammogram. Conceptually, the four categories that will serve as the DV are thought of as reflecting an *underlying continuum of propensity to obtain a*

mammogram (ψ). Movement from step to step indicates that the woman has passed a threshold along the *underlying continuum*. The thresholds need not be equally spaced, since only ordinality is assumed. With four different ordered behaviors, as in the STEPS outcome, there are three hypothesized latent thresholds τ_j. Categories 1 versus 2, 3, and 4 are separated by threshold τ_1; categories 1 and 2 versus 3 and 4, by threshold τ_2; categories 1, 2, and 3 versus 4 by threshold τ_3.

Movement along the latent continuum as a function of predictor(s) can be modeled in a form of ordinal regression model, also referred to as a *proportional odds model* or *parallel regression model*. These names are informative about the assumptions of the model. The structure of the model can be cast in terms of the odds of transition across thresholds, given values on the predictors. In the ordinal logistic regression model, it is assumed that these odds are equal across the continuum, given values of the predictors, hence the term *proportional odds model*. Put another way, we assume that the predictors have the same impact on crossing all the thresholds. For the mammography intervention, this amounts to saying that the intervention has the same impact on moving a woman from doing nothing to contacting a health professional (i.e., crossing the first threshold τ_1) as on moving a woman from contacting a health professional to making an appointment for a mammogram (i.e., crossing the second threshold τ_2) as on moving a woman from making an appointment for a mammogram to obtaining a mammogram (i.e., crossing the third threshold τ_3).

Estimation of an ordinal logistic model involves estimation of a model for the probability of membership in a particular category, given values on the predictors and values of the thresholds. Consider a 3-category scale—disagree (D), undecided (U), and agree (A). Calling p_{ij} the probability that case$_i$ is in category$_j$, we have three predicted probabilities for each case: $\hat{p}_{iD}, \hat{p}_{iU}$, and $\hat{p}_{iA}$. Threshold τ_1 is between D and U; and threshold τ_2 is between U and A. Then in ordinal logistic regression the following equations are estimated in the one predictor case:

(13.3.1)
$$\ln\left(\frac{\hat{p}_{iU}+\hat{p}_{iA}}{\hat{p}_{iD}}\right) = t_1 + BX_i;$$

(13.3.2)
$$\ln\left(\frac{\hat{p}_{iA}}{\hat{p}_{iD}+\hat{p}_{iU}}\right) = t_2 + BX_i.$$

where t_1 and t_2 are the sample estimates of the population thresholds τ_1 and τ_2, respectively. Note that the regression coefficient B is constant across equations, following the parallel response assumption. The estimated thresholds t_1 and t_2 differ across equations (for simplicity, an overall regression intercept is omitted).

Operationally, a single analysis is performed in which the dependent variable entered is the ordered step variable in which each individual is assigned the value of the category in which he/she falls. In *ordinal logistic regression*, the logistic regression model is used to predict the probability of membership in a category. A single regression coefficient B_j and corresponding odds ratio for predictor X_j are estimated for the full data set, corresponding to the overall impact of the predictor on the probability of membership in a category. The regression coefficient and odds ratio are assumed to apply equally across the continuum of categories. Estimates of the latent thresholds are also given. The scale of the latent variable is arbitrary; hence the scale of the thresholds τ_j is arbitrary (not in any particular units). The progression of thresholds along a continuum, however, can be seen. Long (1997) provides a discussion of standardization and interpretation of values on the latent continuum.[6]

[6]Ordinal logistic regression is implemented in SAS PROC LOGISTIC by simply entering an ordered categorical variable as the DV, and in SPSS PLUM as well.

An ordinal logistic regression of the STEPS variable is given in Table 13.3.1D. Overall, the odds of moving along the STEPS continuum from category to category as a function of the intervention are 4.32, an estimate of the overall impact of the treatment across the continuum.

The critical assumption of the model of proportional odds (or, equivalently, parallel slopes) is tested by a Score test. This test compares the fit of a model in which a single slope is applied to the whole continuum (the ordinal regression model) versus an unconstrained model in which a different slope is permitted for cases below versus above each threshold. The null hypothesis is that the parallel slopes model applies, that is, that the predictors have the same impact on crossing all the thresholds or, equivalently, that the odds ratios for crossing the thresholds, given the predictors, are equal. For the data set in Table 13.3.1D, the model is not rejected, $\chi^2(2) = .03$, *ns*. The assumption of parallel slopes is met. This is consistent with our conclusions from the analysis of the continuation thresholds in Table 13.3.1A, B, and C, that the odds ratios are quite similar across the continuum.

What if this test of proportional odds were significant, signifying that the ordinal regression model is not appropriate? The continuation category approach might be applied to develop separate models of the transition across each of the thresholds. In the ordinal regression model, a single set of predictors with identical regression coefficients must apply across the whole continuum. With the continuation category approach, there is the opportunity to develop different models of each transition, each containing its own set of predictors.

Ordinal Logistic Regression Versus the Nested Dichotomies Approach

If the researcher has reason to believe that a single model describes movement along the full latent propensity continuum across all thresholds, it is useful to begin with ordinal logistic regression. The associated Score test provides useful information about whether the researcher's hunch is correct or not. If the researcher is correct (i.e., the Score test is nonsignificant for an adequate sized sample), then a simple and parsimonious model of movement along the propensity continuum has been established, the overall ordinal logistic regression equation. In addition, the researcher obtains estimates of the thresholds. Should the Score test be rejected for ordinal logistic regression, then the researcher can move to a more complex representation of crossing each threshold, using the nested dichotomies approach for continuation categories.

Ordered Category Methods Versus Ordinary Least Squares Regression

Finally, an alternative approach to the analysis of the STEPS continuum is a usual OLS regression in which we assume that the four steps constitute an *interval scale* of propensity to obtain a mammogram, with the three thresholds *equally* spaced across the continuum. If this is so, then the OLS and ordinal regression results converge. An OLS analysis of the STEPS variable is reported in Table 13.3.1E. Of course, the INTERVEN predictor does predict STEPS, as we would expect from the three dichotomous logistic regressions and the ordinal logistic regression. The use of the OLS model for an ordinal DV that lacks equally spaced scale intervals may result in the same difficulties we saw with OLS regression applied to a dichotomous variable: non-normality of residuals and heteroscedasticity.

The analysis strategies of ordinal logistic regression and nested dichotomous logistic regressions with continuation categories may be useful when the outcome is measured on a Likert scale (e.g., strongly disagree . . . strongly agree), as well as with behavioral continua. In psychology, at least, we use Likert scales frequently for attitude measurement, for example. Yet in attitude change experiments, we typically examine arithmetic mean shifts. We do not ask

whether our experimental manipulations are equally effective across the attitude continuum, a question that is answered with ordinal logistic regression and the nested dichotomies strategy. Use of OLS regression with Likert scales or other ordered category scales does not inform this question. In OLS, an overall regression coefficient is assumed to apply across the continuum; no test of this assumption like the Score test for parallel slopes in ordinal logistic regression is supplied in the OLS framework. Neither are measures of the odds of crossing thresholds on the Likert scale supplied in OLS regression as they are in both ordinal logistic regression and nested dichotomous logistic regressions with continuation categories. Ordinal logistic regression also offers measures of thresholds; these threshold estimates provide information about category width. We believe that ordinal logistic regression and nested category approaches merit more frequent use than they currently enjoy.

13.4 MODELS FOR COUNT DATA: POISSON REGRESSION AND ALTERNATIVES

Count data in many substantive areas provide an informative dependent variable. As pointed out earlier, counts might include the number of aggressive acts a child commits during a 30-minute playground period, the number of cigarettes an individual smokes in an hour, the number of scholarly articles a faculty member publishes in a year. In all cases, the measure is characterized as the number of events that occur in a particular time period—a count. *Poisson regression analysis* predicts the number of events that occur in a specific time period from one or more independent variables. The assumption is that the number of events generated in a period of time depends on an *underlying rate* parameter. Because both logistic regression and Poisson regression are cases of the generalized linear model, there are direct parallels between logistic regression analysis and Poisson regression analysis. We present the Poisson regression model and then draw parallels to the logistic regression model. Finally, we use the characteristics of logistic regression and Poisson regression as examples of the generalized linear model to provide a more complete characterization of the generalized linear model.

Poisson regression is appropriate when we examine a phenomenon of very rare events, so that if we count the number of events for each of a sample of people, there are numerous people with scores of zero. To return to the example of number of cigarettes smoked, if we count the number of cigarettes smoked in an hour by attendees at a large party, there will many zeros, since many people do not smoke at all. We warn the reader at the outset that there are relatively few examples in behavioral science of the use of Poisson regression. Further, less effort has been devoted to data-analytic conventions (e.g., diagnostics) and interpretation of results than in logistic regression.

13.4.1 Linear Regression Applied to Count Data

The problems that we encounter in applying OLS regression to count data mirror those encountered in applying OLS regression to dichotomous outcomes. The residuals are not normally distributed, and they exhibit heteroscedasticity; therefore inferences about individual predictors and overall prediction may well be biased if OLS regression is applied. Moreover predicted scores can be out of range, specifically, below zero, which is impossible for counts. Further, the regression coefficients from OLS regression applied to count data may be biased and inconsistent, meaning that they do not become more accurate as sample size increases (Long, 1997). Moreover, the standard errors from the OLS regression may underestimate true standard errors, leading to inflated t tests for individual regression coefficients; significance of these coefficients is thus overestimated (Gardner, Mulvey, & Shaw, 1995).

Transforming Y Versus Poisson Regression

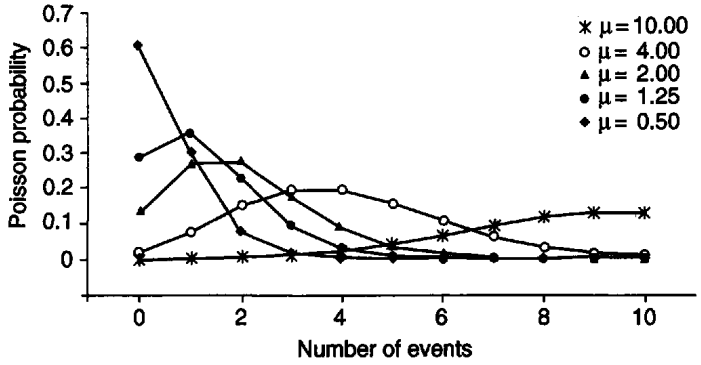
In Chapter 6 we discussed transformations that can be applied to dependent (and independent) variables in order to render the data better behaved for us in OLS regression. By better behaved we mean closer to meeting assumptions of normality of residuals and homoscedasticity. For a count dependent variable, the square root is suggested as a potentially useful transformation (see Section 6.4.13). The question may be raised as to whether one need use Poisson regression. Might it not be possible to simply take the square root of each count and apply OLS regression to the square roots of the counts? When we are dealing with rare events for which there are many zero counts, the OLS approach does not handle the excess of very low scores relative to the rates predicted by OLS regression; it also does not handle the heteroscedasticity problem (Gardner, Mulvey, & Shaw, 1995). Poisson regression is preferred. This question of the use of transformations plus OLS regression versus the use of newer methodologies is reflective of the evolution of statistics, a point made in Section 13.2.1 in the discussion of discriminant analysis versus logistic regression. The work on transformations historically precedes some of the developments of numerical methods for the generalized linear model. If one takes the approach of transformation of the count DV followed by OLS regression, it is important to examine residuals for non-normality and heteroscedasticity before accepting conclusions based on the analysis.

13.4.2 Poisson Probability Distribution

An understanding of Poisson regression is facilitated by a consideration of the Poisson distribution. The Poisson distribution is a probability distribution, like the normal, binomial, or *t* distribution. The Poisson distribution is used to represent the error structure in Poisson regression. Five examples of the Poisson distribution are plotted in Fig. 13.4.1. Each distribution shows the probability that an individual will produce a specific number of events in a given time period. The *x* axis shows number of events and the *y* axis, the probability of each number of events. Each of the distributions is generated by the same equation, but a characteristic known as the *rate parameter* differs across the distributions. The *rate parameter* is the *average number of events expected in the time period.* The rate parameter is typically denoted as μ, as in Fig. 13.4.1. The numerical values of the probabilities plotted in each distribution are also given above the graphical representation. Imagine that the probabilities for $\mu = .50$ are probabilities of achieving various numbers of scholarly publications in a year in a work environment that does not emphasize publication for merit and promotion; the probability of no publications is .61; of one publication, .30, of four or more publications, 0.00. Those probabilities for $\mu = 4.00$ might represent probabilities of different numbers of publications in a year for an institution that heavily emphasizes scholarly publication for merit and promotion. Note that the probability of no publications is only .02; of four or more publications, .57. The source of these probabilities in Fig. 13.4.1 is the *Poisson distribution*, given in Box 13.4.1.

The Poisson distribution is the most basic probability distribution applied to regression when the outcomes are count data. The five examples in Fig. 13.4.1 highlight important characteristics of the Poisson distribution. First, the probabilities only apply to counts of events; counts are whole numbers that range from 0 to $+\infty$, as represented on the *x* axis of Fig. 13.4.1. The rate parameter, however, can have decimal values, because it represents the average or expected number of counts. Second, when the rate parameter or mean (μ) is very small (e.g., when the expected number of events is $\mu = .50$ in Fig. 13.4.1), then many cases with zero events are expected, and the distribution is very positively skewed. As the mean number of events increases, the distribution becomes increasingly symmetrical, approaching a normal distribution. With $\mu = 10$, the distribution is almost normal in form; what we see in Fig. 13.4.1

		Expected number of events (μ)				
Case	Y	$\mu = .50$	$\mu = 1.25$	$\mu = 2.00$	$\mu = 4.00$	$\mu = 10.00$
100	0	.61	.29	.14	.02	.00
101	1	.30	.36	.27	.07	.00
102	2	.08	.22	.27	.15	.00
103	3	.01	.09	.18	.20	.01
104	4	.00	.03	.09	.20	.02
105	5	.00	.01	.04	.16	.04
106	6	.00	.00	.01	.10	.06
107	7	.00	.00	.00	.06	.09
108	8	.00	.00	.00	.03	.11
109	9	.00	.00	.00	.01	.13
110	10	.00	.00	.00	.01	.13



Note: Each curve reflects a different rate parameter (μ). The illustration of the curve for $\mu = 10$ shows only the left half of this distribution. The curve would decline almost symmetrically to the right if the number of events from 11 to 20 were provided.

FIGURE 13.4.1 Probability of Y events according to Poisson distribution as a function of the expected (mean) number of events. The expected number of events (or the rate parameter) is noted as μ.

for $\mu = 10$ is the left half of such a distribution that declines almost symmetrically to zero as the number of events continues to increase from 11 to 20 (the right half of the distribution is not shown in Fig. 13.4.1). Third, as the mean number of events increases, the variance of the number of events across the population also increases. When the mean number of events is $\mu = .50$, the majority of cases have counts of 0, and the highest count is 3. When the mean is $\mu = 4.00$, in contrast, the counts given in Fig. 13.4.1 range from 0 to 10. In fact, for the Poisson probability distribution, the mean and the variance of the distribution are equal:

$$\text{variance}\,(Y) = \mu. \tag{13.4.1}$$

This last property is important; it states that the variance of the Poisson distribution is completely determined by the mean of the distribution.

When the mean count in a distribution is large so that symmetry of the distribution of events is approached, then OLS regression may be tried. Careful checking of residuals for heteroscedasticity and nonnormality is advised, in justifying the appropriateness of OLS regression.

BOX 13.4.1
The Poisson Probability Distribution

The expression for the *Poisson distribution* is as follows:

(13.4.2) $$P(Y) = \frac{e^{(-\mu)}\mu^Y}{Y!}$$

where Y = the number of events and μ = the expected or mean number of events, and $Y!$ is Y factorial $= Y(Y-1)(Y-2)\ldots 1$.

This expression generates all the curves displayed in Fig. 13.4.1. For example, for $\mu = 4.00$ and $Y = 3$ publications,

$$P(Y) = \frac{e^{(-4.00)}\mu^3}{3!} = \frac{.0183 \times 4^3}{3 \times 2 \times 1} = .20.$$

(As before, to find $e^{-4.00}$, enter -4.00 into a calculator, and press the e^x button; see Box 13.2.2.)

13.4.3 Poisson Regression Analysis

Exactly paralleling logistic regression, we have three forms of the regression equation in Poisson regression. First, we predict the expected number of events ($\hat{\mu}$) from values on a set of predictors $X_1, X_2, \ldots, X_k$.

(13.4.3) $$\hat{\mu} = e^{(B_1X_1+B_2X_2+\cdots+B_kX_k+B_0)}.$$

Equation (13.4.3) is not in a form that is linear in the coefficients. If we take the logarithm of both sides, we have a second regression equation that is linear in the coefficients and in which the logarithm of the predicted expected number of events is the predicted score:

(13.4.4) $$\ln(\hat{\mu}) = B_1X_1 + B_2X_2 + \cdots + B_kX_k + B_0.$$

Third, we can write an equation in which we predict the probability of each specific number of events, given the expected average number of events $\hat{\mu}$. For the predicted probability of a count of c events ($\hat{p}_c$) we have

(13.4.5) $$\hat{p}_c = \frac{e^{(-\hat{\mu})}\hat{\mu}^c}{c!}.$$

Let us focus on Eq. (13.4.3). This is a simple exponential equation. The equation represents a curve like that in Fig. 13.1.1(C), which shows the predicted number of events as a function of values of a predictor X. The curve rises faster than a straight line as X increases. Equation (13.4.3) predicts more events when X is high than would be predicted from a linear equation.

A natural question, given the curve in Fig. 13.1.1(C) is, whether a quadratic polynomial regression equation $\hat{Y} = B_1X + B_2X^2$ (a form of OLS regression) would fit the curve. In fact, a quadratic polynomial might closely approximate the curve. However, the increasing variance of Y with increasing X, a characteristic of count data, would still exist in the data and would be ignored in polynomial regression, potentially causing difficulties in inference that are handled in Poisson regression.

Fictitious Example of Poisson Regression and Interpretation of Coefficients

A fictitious example illustrates the form of a Poisson regression for count data. Assume we are predicting the expected number of aggressive acts that a young child will exhibit on a playground during a 30-minute recess period. The single predictor X is a 0–10 rating of each child's aggressiveness by the teacher. We have a one-predictor Poisson regression equation; Eq. (13.4.3) simplifies to

$$\hat{\mu} = e^{(B_1X+B_0)}. \tag{13.4.6}$$

Assume that the appropriate Poisson regression equation is $\hat{\mu} = e^{(.35X-1.68)}$. Figure 13.4.2 provides the aggressiveness ratings for 11 children, one at each value of rated aggressiveness. The expected rate (predicted number of aggressive acts) for each child according to the Poisson regression equation and the predicted number of aggressive acts from a linear regression analysis (OLS) are also presented. The predicted rate ($\hat{\mu}$) from Poisson regression indicates that aggressive acts are rare if the child is rated 5 or lower in aggressiveness (at a rating of 5, only 1.07 acts are expected). Thereafter, as the rating rises from 6 to 10, the number of expected aggressive acts increases rapidly to 6. Poisson predicted rates are noted with dots,

Poisson regression equation: $\hat{\mu} = e^{(.35X-1.68)}$

Case	Aggressiveness rating	Predicted rate ($\hat{\mu}$) from Poisson regression	Predicted number of acts from linear regression	ln ($\hat{\mu}$) from Poisson regression
100	0	0.19	−0.79	−1.68
101	1	0.26	−0.26	−1.33
102	2	0.38	0.27	−0.98
103	3	0.53	0.80	−0.63
104	4	0.76	1.33	−0.28
105	5	1.07	1.86	0.07
106	6	1.52	2.39	0.42
107	7	2.16	2.92	0.77
108	8	3.06	3.45	1.12
109	9	4.35	3.98	1.47
110	10	6.17	4.51	1.82

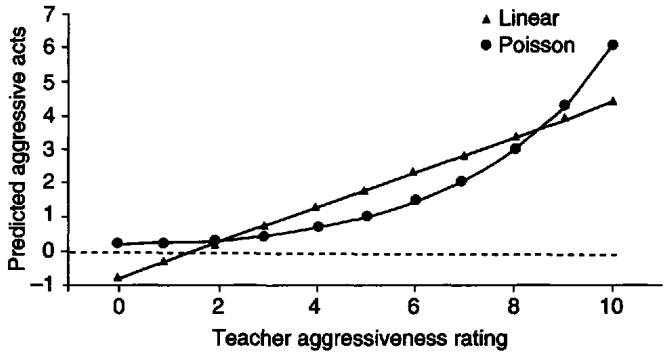


FIGURE 13.4.2 Poisson regression versus linear regression for a fictitious example predicting number of aggressive acts in a 30-minute recess as a function of teacher rating of aggressiveness according to a Poisson regression equation.

linear regression predicted scores with filled triangles. Note that when the rating is either very low or very high, the Poisson regression analysis predicts a higher number of aggressive acts than the linear regression equation, whereas the linear regression equation predicts higher scores in the midrange of the rating scale. The Poisson model accounts for the cases with zero counts and with very high counts, whereas OLS regression misses these aspects of the count data.

The B coefficient in the Poisson equation, $B = .35$, can be interpreted in two ways. First, for a 1-unit increase in X (the aggressiveness rating), the predicted rate ($\hat{\mu}$, expected number of aggressive acts) is *multiplied* by the value e^B. For $B = .35$, $e^{.35} = 1.42$. For example, for a child with a rating of 6, we expect $\hat{\mu} = e^{(.35X-1.68)} = \hat{\mu} = e^{[.35(6)-1.68]} = 1.52$ aggressive acts; for a child with a rating of 7, we expect 2.16 aggressive acts, $(1.52)(1.42) = 2.16$. Second, for a unit-1 increase in X, the predicted score in the linear equation, $\ln(\hat{\mu})$, the natural logarithm of the expected number of aggressive acts, increases by .35. For a rating of 7, $\ln(\hat{\mu}) = .77$; for a rating of 8, $\ln(\hat{\mu}) = .77 + .35 = 1.12$. (Interpreting the meaning of the Poisson regression equation is clear in the e^B form, which yields a predicted count, whereas the value of the logarithm of the expected count cannot be interpreted directly in terms of counts). It should be noted that all the predicted rates from the Poisson regression equation are positive, corresponding to numbers of events, which cannot fall below zero. In contrast, the predicted number of acts from the linear regression equation is negative for aggressiveness scores of 0 and 1, an impossible situation.

Heteroscedasticity of Residuals

Refer again to Fig. 13.4.1, and assume that the values μ represent values of predicted scores from Eq. (13.4.3). For each curve based on a particular value of μ we see a distribution of Y scores (the number of events); each curve is a conditional distribution of Y scores, (i.e., the Y scores of all individuals with a particular value of μ.) There is clear heteroscedasticity of the conditional distributions of Y. As the value of μ increases, the spread of the expected distribution of Y scores becomes broader. For example, for $\mu = .50$, Y scores between 0 and 3 have expected probabilities greater than zero; for $\mu = 4.00$, Y scores between 0 and 10 have expected probabilities greater than zero.

13.4.4 Overdispersion and Alternative Models

The Poisson regression model makes the very restrictive assumption that we have already encountered: The variance $\sigma^2 = \hat{\mu}$, or the variance of the residuals around each predicted rate equals the predicted rate. In order that this assumption be met, all of the systematic variation in rates $\hat{\mu}_i$ across individuals must be accounted for completely by the set of predictors; no other potential predictor could account for additional variance. Put another way, all individuals with the same values on the predictors should have the same observed rate parameter. In the fictitious example, this means that all the systematic variation in the observed number of aggressive acts would be accounted for by scores on the teacher aggressiveness rating scale. If there is systematic variation in rates that is not accounted for by the predictor set, then there will be greater variation in the residuals around each predicted rate than is permitted by the Poisson regression model (i.e., $\sigma^2 > \hat{\mu}$). This condition is termed *overdispersion* (see Gardner, Mulvey & Shaw, 1995, and Land, McCall, and Nagin, 1996 for discussions of overdispersion). Overdispersion is frequently found in count data. Overdispersion leads to inflation of the goodness of fit χ^2 test. In addition, with overdispersion, the standard errors of Poisson regression coefficients are too small, so the significance of predictors is overestimated. The level of dispersion relative to a Poisson distribution is often characterized by a *dispersion*

parameter ϕ, which equals 1 if the Poisson variance assumption is met, is greater than 1 for overdispersion, and is less than 1 for *underdispersion.*

There are a number of statistical approaches for analysis of count data that exhibit overdispersion. First, is the *overdispersed Poisson model,* which belongs to a class of models called *quasi-likelihood regression models* (Fahrmier & Tutz, 1994). A second approach to overdispersion is the use of an alternative regression model, the *negative binomial regression model.* Both approaches are described in Box 13.4.2.

BOX 13.4.2
Alternative Models for Count Data With Overdispersion: The Overdispersed Poisson Model and the Negative Binomial Regression Model

In the overdispersed Poisson model, the dispersion parameter ϕ is calculated from the data themselves; the standard errors of the regression coefficients for overdispersed data are adjusted by multiplying them by ϕ. If there is overdispersion, and thus $\phi > 1$, the values of the standard errors are increased, thereby decreasing the excess significance in the statistical tests of the regresison coefficients. If $\phi = 1$, the Poisson model holds. No special model is specified of the distribution of the excess variance relative to the Poisson variance. That is, no probability distribution is assumed for how the individual rate parameters μ_i vary around the expected rate parameter μ, given values on the set of predictors. This is the hallmark of quasi-likelihood models, that a portion of the variance is not assumed to follow a particular probability model.

The negative binomial model assumes that for each individual, a Poisson distribution applies, but that the rates for individuals μ_i, given specific values on the predictors, vary across individuals. A new probability distribution known as the negative binomial distribution is used to characterize the variance of the residuals. The variance of the negative binomial distribution is comprised of two components: (1) the expected rate μ (as in Poisson regression) plus (2) a second amount that characterizes the additional variance in the rate parameter across individuals, not accounted for by the Poisson distribution (see Gardner, Mulvey, and Shaw, 1995, p. 399; Land, McCall, and Nagin, 1996, p. 397; Long, 1997, p. 233). As a result, the negative binomial variance for each value of μ_i is greater than μ_i. Put another way, the negative binomial model of the errors allows greater variance than is permitted by Poisson regression, thereby accounting for overdispersion in count data. Negative binomial regression may still result in inflated t values.

The negative binomial regression model is one of a class of *mixed Poisson models* that mix a second source of variance with the Poisson variance to account for overdispersion (Land, McCall, & Nagin, 1996, p. 397). In contrast to quasi-likelihood models, which specify no particular probability distribution for the excess variance, mixture models specify a second probability distribution for the second source of variance, over and above the Poisson distribution, which characterizes the first source of variance. In the negative binomial regression model, the second probability distribution is another discrete probability distribution, the gamma distribution. The combination of the Poisson distribution with the gamma distribution yields the negative binomial distribution. It is this mixture of probability distributions that is the hallmark of mixture models. Newer approaches to overdispersion include the *semiparametric mixed Poisson regression* characterized by Land, McCall, and Nagin (1996).

13.4.5 Independence of Observations

Poisson regression also assumes that the observations are *independent* of one another, just as in OLS regression and logistic regression. However, the basic datum in Poisson regression is an event (a publication, an aggressive act, etc.) exhibited by one individual. Such events emitted by one individual tend to be correlated; that is, the fact that one event has occurred may increase the probability of subsequent events. This correlation between events is referred to in economics as *state dependence*, or in biometric and sociological literature as the *contagion model* (see Land, McCall, & Nagin, 1996, p. 395). Considering our example of number of aggressive acts, it is easy to imagine how a single aggressive act of a child on the playground can lead to still other aggressive acts as a fight ensues. Such nonindependence of events leads to clusters of events, or higher numbers of events in particular individuals than would be expected from the Poisson distribution. In addition, it leads to an excess of zeros, that is, more cases in which there are zero events than would be expected by the Poisson model. In the case of state dependence, the distribution of the counts observed for individuals in the sample does not follow Poisson distributions, and other count models must be employed.

13.4.6 Sources on Poisson Regression

Accessible sources on Poisson regression are less readily available than are sources on logistic regression. Neter, Kutner, Nachtsheim, and Wasserman (1996, pp. 609–614) provide an introduction; Long (1997, pp. 217–249) provides a more extensive treatment. Gardner, Mulvey, and Shaw (1995) and Land, McCall, and Nagin (1996) provide discussions of limitations of Poisson regression and alternative models.

13.5 FULL CIRCLE: PARALLELS BETWEEN LOGISTIC AND POISSON REGRESSION, AND THE GENERALIZED LINEAR MODEL

13.5.1 Parallels Between Poisson and Logistic Regression

There are many parallels between Poisson regression and logistic regression. Having reviewed the presentation of logistic regression in detail, the reader should find these parallels facilitate an understanding of Poisson regression. These parallels will also facilitate a characterization of the generalized linear model, of which logistic regression and Poisson regression are two special cases.

Parallels between logistic and Poisson regression exist in six areas: (1) three forms of regression equation, (2) the interpretation of coefficients, (3) the relationship of the form of observed Y scores to the form of predicted scores, (4) the concept of error structure, (5) nature of the estimates and estimation procedures, and (6) the nature of significance tests.

Three Equations

Three forms of the logistic regression equation predict the odds of a case, the log of the odds (the logit), and the probability of being a case, given in Eqs. (13.2.12), (13.2.11), and (13.2.14), respectively. In Poisson regression, the three forms of the Poisson regression equation predict the expected number of events ($\hat{\mu}$), the log of the expected number of events, $\ln(\hat{\mu})$, and the predicted probability of observing c events ($\hat{p}_c$), in Eqs. (13.4.3), (13.4.4), and (13.4.5), respectively. In logistic regression we move from the predicted score in the form of the logit

to the odds by taking the antilog; we then compute the probability from the odds. In Poisson regression, we move from the predicted score in the linear equation form, that is, $\ln(\hat{\mu})$, to the predicted number of events, by taking the antilog of $\ln(\hat{\mu})$, yielding $(\hat{\mu})$. Then we can substitute $(\hat{\mu})$ into Eq. (13.4.5) to obtain predicted probabilities.

Interpretation of Coefficients

Consider the one-predictor logistic regression equation Eq. (13.2.6) and the one-predictor Poisson regression equation, written in linear form as

$$\ln(\hat{\mu}) = B_1X + B_0. \tag{13.5.1}$$

In logistic regression, the value of the logit, $\ln[\hat{p}/(1-\hat{p})]$, increases linearly with the value of B_1, the regression coefficient. In Poisson regression, the value of $\ln(\hat{\mu})$ increases linearly with the value of B_1, the regression coefficient. In logistic regression the odds, $[\hat{p}/(1-\hat{p})]$, are multiplied by the value e^{B_1} for each one unit increase in X, as in Eq. (13.2.5). In Poisson regression, written as Eq. (13.4.3), the expected rate $(\hat{\mu})$ is multiplied by the value e^{B_1} for each one unit increase in X. For both logistic and Poisson regression, these interpretations generalize to the multiple predictor case.

Form of Observed Y Versus Predicted Score

In both logistic regression and Poisson regression, the predicted score in the linear form of the regression equation is not in the same units as the observed Y score. In logistic regression, the observed Y score indicates group membership (1 = case; 0 = noncase). However, in the linear form of the logistic regression equation, the predicted score is in the form of a logit, as shown in Eq. (13.2.11). In Poisson regression, the predicted score in the linear form of the regression equation is again not in the same units as the observed Y score. The observed Y score is a count of the number of events in a specific time period. However, in the linear form of the Poisson regression equation, the predicted score is the logarithm of the count, as shown in Eq. (13.4.4).

Error Structure

Both Poisson regression and logistic regression have non-normally distributed residuals as inherent in the model. In both cases the variance of the errors around the predicted score is determined by the value of the predicted score, leading to heteroscedasticity in both models.

Estimates and Estimation Procedures

Both Poisson and logistic regression employ maximum likelihood estimation, with iterative solutions for the regression coefficients.

Significance Tests

Likelihood ratios, deviances (null and model), and likelihood ratio χ^2 tests proceed in the same fashion for Poisson as for logistic regression for testing overall model fit, contribution of predictor sets and individual predictors. As with all maximum likelihood estimates, the distributions of estimates of the regression coefficients in Poisson regression approach normality as sample size approaches infinity; Wald tests apply to individual coefficients and combinations of coefficients.

13.5.2 The Generalized Linear Model Revisited

The generalized linear model is a highly flexible approach to characterizing the relationship of predictors to a dependent variable Y that subsumes a variety of specific regression models. Logistic and Poisson regression are two such specific instances of the generalized linear model; they serve to illustrate the characteristics of this broad class of models of prediction, which also includes OLS regression. The regression models included in the generalized linear model can all be expressed in a form that is *linear in the parameters*. For OLS regression, logistic regression, and Poisson regression, these forms are given as Eqs. (13.1.1), (13.2.11), and (13.4.4), respectively. All instances of the generalized linear model assume that observations are independent. The varieties of the generalized linear model are characterized in two ways, explained next: the *variance function* and the *link function*.

Variance Function

The generalized linear model gains part of its great flexibility by extending the assumption of the distribution of the residuals from normality to a *family of probability distributions*, the *exponential family*. If the dependent variable is dichotomous, then the residuals follow a binomial distribution (see Section 13.2). If the dependent variable consists of counts of the number of events in a period of time, then the residuals follow a Poisson distribution (see Section 13.4.2). The binomial and Poisson distributions, and other distributions that are members of the exponential family (e.g. gamma, inverse Gaussian) in general have the property that the mean and the variance of the distribution are not independent, that is, the variance depends on the mean. (The Gaussian or normal distribution, which is also a member of the exponential family, is an exception in which the mean and variance are independent). The lack of independence of the mean and variance of the distribution of residuals leads to heteroscedasticity, since the conditional variance of the criterion around a predicted value $\hat{Y}$ depends on the value of $\hat{Y}$. We saw this dependence in Eq. (13.2.3) for the variance of residuals associated with dichotomous Y and Eq. (13.4.1) for the variance of residuals for a count variable Y. When we find residuals that are not normally distributed, we require a model of the *variance function*, that is, a model of how the conditional variance of Y varies as a function of Y. For example, the Poisson distribution, which gives Poisson regression its name, comes into play in the assumed *variance function* for the residuals, a central aspect of the generalized linear model. In Poisson regression, it is assumed that the residuals at each value of $\hat{\mu}$, the predicted rate, are distributed as a Poisson distribution, with variance also equal to $\hat{\mu}$.

Link Function

In each form of regression we have encountered—OLS regression, logistic regression, and Poisson regression—we have considered the relationship of the form of the observed Y score versus the predicted score in the regression equation that is linear in the coefficients. We have noted that in OLS regression, observed Y and predicted scores are in the same units. Once again, in logistic regression we have dichotomous Y versus the predicted logit, in Poisson regression, the count versus the predicted log (count). The *link function* in the generalized linear model is the transformation that relates the predicted outcome to the observed dependent variable Y. A second source of flexibility in generalized linear models is the variety of link functions that are possible. For OLS regression, the link function is the *identity* function, since observed and predicted scores are on the same scale. For logistic regression, the link function is the *logit*; for Poisson regression, the link function is the *logarithm*.

Regression models that are linear in the coefficients, whose residuals are assumed to follow a variance function from the exponential family, with one of a variety of link functions are

members of the generalized linear model. McCullagh and Nelder (1989) is the classic reference work on generalized linear models. Fahrmier and Tutz (1994) is a second complete source.

13.6 SUMMARY

Ordinary least squares regression assumes that the dependent variable has normally distributed errors that exhibit homoscedasticity. Categorical dependent variables and count variables do not exhibit these properties. If these conditions are not met, OLS regression may be inefficient and lead to inaccurate conclusions (Section 13.1).

Binary dependent variables traditionally have been examined using two-group discriminant analysis. The logistic regression model for binary dependent variables is presented as an appropriate alternative, first for the single-predictor case, and then for the multiple-predictor case (Section 13.2). Three forms of the logistic regression model—predicting the logit (log odds), the odds, and the probability of being a case—are developed and explained (Sections 13.2.3 and 13.2.4). The characterization of regression coefficients in the form of odds ratios is explained (Section 13.2.5). Confidence intervals for regression coefficients and odds ratios are presented (Section 13.2.8). Maximum likelihood estimates and maximum likelihood estimation are characterized (Section 13.2.9). Likelihood ratios, deviances, and statistical tests for overall model fit based on likelihood ratios and deviances are introduced (Section 13.2.10, 13.2.12). Indices of overall model fit are introduced (Section 13.2.11). Wald and likelihood ratio tests for significance of individual predictors are presented (Section 13.2.13), and likelihood ratio tests of gain in prediction from sets of predictors are explained (Section 13.2.14). Difficulties in predictor scaling are addressed in logistic regression (Section 13.2.16). Issues in the use of regression diagnostics in logistic regression are explained (Section 13.2.17). The application of logistic regression to statistical classification is introduced (Section 13.2.19).

The logistic regression model is extended to multiple response categories with presentation of polytomous logistic regression. The analysis of ordered categories by nested dichotomies and ordinal logistic regression is illustrated (Section 13.3).

Poisson regression is developed for count data, that is, dependent variables that are counts of rare events, such that there are many scores of zero and the count dependent variable is highly positively skewed (Section 13.4).

Logistic regression and Poisson regression are used to illustrate the characteristics of the generalized linear model, with explication of the variance function and the link function (Section 13.5).