SAGE UNIVERSITY PAPERS

Series: Quantitative Applications
in the Social Sciences

Series/Number 07-106

APPLIED LOGISTIC REGRESSION ANALYSIS
Second Edition

SCOTT MENARD
Institute of Behavioral Science
University of Colorado

Copyright
2002

SAGE PUBLICATIONS
International Educational and Professional Publisher
Thousand Oaks London New Delhi

2. SUMMARY STATISTICS FOR EVALUATING THE LOGISTIC REGRESSION MODEL

When we evaluate a linear regression model, the evaluation typically has three parts. First, how well does the overall model work? Can we be confident that there is a relationship between all of the independent variables, taken together, and the dependent variable, above and beyond what we might expect as a coincidence, attributable to random variation in the sample we analyze? If there is a relationship, how strong is it? Second, if the overall model works well, how important is each of the independent variables? Is the relationship between any of the variables attributable to random sample variation? If not, how much does each independent variable contribute to our ability to predict the dependent variable? Which variables are stronger or weaker, better or worse predictors of the dependent variable? Third and finally, does the form of the model appear to be correct? Do the assumptions of the model appear to be satisfied? In this chapter, we deal with the first question, the overall adequacy of the model. Chapter 3 deals with the contributions of each of the independent variables, and Chapter 4 focuses on testing the assumptions of the model.

In linear regression analysis, we need to know (a) whether knowing the values of all of the independent variables put together allows us to predict the dependent variable any better than if we had no information on any of the independent variables and, if so, (b) how well the independent variables as a group explain the dependent variable. For logistic regression, we also may be interested in the frequency of correct as opposed to incorrect predictions of the exact value of the dependent variable, in addition to how well the model minimizes errors of prediction. In linear regression, when the dependent variable is assumed to be measured on an interval or ratio scale, it would be neither alarming nor unusual to find that none of the predicted values of the dependent variable exactly matched the observed value of the dependent variable. In logistic regression, with a finite number (usually only two) of possible values of the dependent variable, we may sometimes be more concerned with whether the predictions are correct or incorrect than with how close the predicted values (the predicted conditional means, which are equal to the predicted conditional probabilities) are to the observed (0 or 1) values of the dependent variable.

2.1. R^2, F, and Sums of Squared Errors

In linear regression analysis, evaluation of the overall model is based on two sums of squares. If we were concerned with minimizing the sum of the squared errors of prediction and if we knew only

the values of the dependent variable (but not the cases to which those values belonged), we could minimize the sum of the squared errors of prediction by using $\bar{Y}$, the mean of Y, as the predicted value of Y for all cases. The sum of squared errors based on this prediction would be $\sum(Y_j - \bar{Y})^2$, the *Total Sum of Squares* (SST). If the independent variables are useful in predicting Y, then $\hat{Y}_j$, the value of Y predicted by the regression equation (the conditional mean of Y) will be a better predictor than $\bar{Y}$ of the values of Y, and the sum of squared errors $\sum(Y_j - \hat{Y}_j)^2$ will be smaller than the sum of squared errors $\sum(Y_j - \bar{Y})^2$. $\sum(Y_j - \hat{Y}_j)^2$ is called the *Error Sum of Squares* (SSE) and is the quantity OLS selects parameters $(\beta_1, \beta_2, \ldots, \beta_k)$ to minimize. A third sum of squares, the *Regression Sum of Squares* (SSR) is simply the difference between SST and SSE: SSR = SST − SSE.

It is possible in a sample of cases to get an apparent reduction in error of prediction by using the regression equation instead of $\bar{Y}$ to predict the values of Y_j, even when the independent variables are really unrelated to Y. This occurs as a result of sampling variation, that is, random fluctuations in sample values that may make it appear as though a relationship exists between two variables when there really is no relationship. The multivariate F ratio is used to test whether the improvement in prediction using $\hat{Y}$ instead of $\bar{Y}$ is attributable to random sampling variation. Specifically, the multivariate F ratio tests two equivalent hypotheses: $H_0 : R^2 = 0$ and $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$. For OLS linear regression, the F ratio with N cases and k independent variables can be calculated as

$$F = [\mathrm{SSR}/k]/[\mathrm{SSE}/(N - k - 1)] = (N - k - 1)\mathrm{SSR}/(k)\mathrm{SSE}.$$

The *attained statistical significance* (p) associated with the F ratio indicates the probability of obtaining an R^2 as large as the observed R^2, or β coefficients as large as the observed β coefficients, *if the null hypothesis is true*. If p is small (usually less than .05, but other values of p may be chosen), then we reject the null hypothesis and conclude that there is a relationship between the independent variables and the dependent variable that cannot be attributed to chance. If p is large, then we "fail to reject the null hypothesis" and conclude that there is insufficient evidence to be sure that the variance explained by the model is not attributable to random sample variation. This does not mean that we conclude that there is no relationship, only that if there is a relationship, we have insufficient evidence to be confident that it exists.

The coefficient of determination, or R^2, or "explained variance" (really, the proportion of the variance that is explained) is an indicator of *substantive* significance, that is, whether the relationship is "big enough" or "strong enough" for us to be concerned about it. R^2 is a *proportional reduction in error* statistic. It measures the proportion (or, multiplied by 100, the percentage) by which use of the regression equation reduces the error of prediction relative to predicting the mean, $\bar{Y}$. R^2 ranges from 0 (the independent variables are no help at all) to 1 (the independent variables allow us to predict the individual values Y_j perfectly). R^2 is calculated as $R^2 = \text{SSR/SST} = (\text{SST} - \text{SSE})/\text{SST} = 1-(\text{SSE/SST})$. The F ratio and R^2 also can be expressed as functions of one another: $F = [R^2/k]/[(1 - R^2)/(N - k - 1)]$ and $R^2 = kF/(kF + N - k - 1)$.

It is possible for a relationship to be statistically significant ($p \leq .0001$) but for R^2 not to be substantively significant (for example, $R^2 \leq .005$) for a large sample. If the independent variables explain less than one-half of 1% of the variance in the dependent variable, we are unlikely to be very concerned with them, even if we are relatively confident that the explained variance cannot be attributed to random sample variation. It is also possible for a relationship to be substantively significant (for example, $R^2 \geq .4$), but not statistically significant for a small sample. Even though the relationship appears to be moderately strong (an explained variance of .40 or, equivalently, a 40% reduction in errors of prediction), there may not be enough cases for us to be confident that this result cannot be attributed to random sampling variation.

2.2. Goodness of Fit: G_M, R^2_L, and the Log Likelihood

Close parallels to F and R^2 exist for the logistic regression model. Just as the sum of squared errors is the criterion for selecting parameters in the linear regression model, the *log likelihood* is the criterion for selecting parameters in the logistic regression model. In presenting information on the log likelihood, however, statistical packages usually present not the log likelihood itself, but the log likelihood multiplied by −2, for reasons noted subsequently. For convenience, the log likelihood multiplied by −2 will be abbreviated as −2LL. Whereas the log likelihood is negative, −2LL is positive, and larger values indicate worse prediction of the dependent variable. The value of −2LL for the logistic regression model with only the intercept included can be calculated in SPSS LOGISTIC REGRESSION by adding the chi-square for the model in the Omnibus Tests of Model Coefficients table plus the −2 log likelihood in the Model Summary table (see Figure 1.4). In SPSS NOMREG and PLUM, to be discussed in more detail later, it is the −2 log likelihood for intercept only in the Model Fitting Information table. In SAS, it is designated as −2 LOG L in the column "Intercept Only" in the output from SAS PROC LOGISTIC. The intercept-only or initial −2LL, hereafter designated D_0 to indicate that is the −2 log-likelihood statistic with none (zero) of the independent variables in the equation, is analogous to the total sum of squares (SST) in linear regression analysis. For a dichotomous dependent variable (coded as 0 or 1), if $n_{Y=1}$ is the number of cases for which $Y = 1, N$ is the total number of cases, and $\text{P}(Y = 1) = n_{Y=1}/N$ is the probability that Y is equal to 1, then

$$\begin{aligned} D_0 &= -2\{n_{Y=1}\ln[\text{P}(Y = 1)] + (N - n_{Y=1})\ln[1 - \text{P}(Y = 1)]\} \\ &= -2\{(n_{Y=1})\ln[\text{P}(Y = 1)] + (n_{Y=0})\ln[\text{P}(Y = 0)]\}. \end{aligned}$$

The value of −2LL for the logistic regression model that includes the independent variables as well as the intercept is designated as −2 log likelihood in the Model Summary table in the output for SPSS LOGISTIC REGRESSION, as −2 log likelihood for the final model in the Model Fitting Information table in SPSS NOMREG and PLUM, and as −2 LOG L in the "Intercept and Covariates" column in SAS PROC LOGISTIC. Hereafter, this −2LL statistic will be referred to as D_M for the full model. D_M is analogous to the error sum of squares (SSE) in linear regression analysis. The most direct analogue in logistic regression analysis to the regression sum of squares (SSR in linear regression) is the difference between D_0 and D_M, that is, $(D_0 - D_M)$. This difference is called the Model chi-square (in the Omnibus Tests table) in SPSS LOGISTIC REGRESSION, or the chi-square for the final model (in the Model Fitting Information table) in SPSS NOMREG and PLUM, or −2 LOG L in the column "Chi-Square for Covariates" in SAS PROC LOGISTIC. Hereafter, it will be referred to as G_M, or the model χ^2.

In logistic regression (and in other general linear models), the difference between two log likelihoods, when multiplied by −2, can be interpreted as a χ^2 statistic if they come from two different models, one of which is *nested* within the other (McCullagh & Nelder, 1989). One model is nested within another if the first model contains some, but not all, of the predictors in the second model and contains no predictors that are not included in the second model. In other words, the predictors in the first model are a proper subset of the predictors in the second. G_M can be straightforwardly interpreted as the difference between a first model that contains only an intercept and a second model that contains the intercept plus one or more variables as predictors. Treated as a chi-square statistic, G_M provides a test of the null hypothesis that $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ for the logistic regression model. If G_M is statistically significant ($p \leq .05$), then we reject the null hypothesis and conclude that information about the independent variables allows us to make better predictions of $P(Y = h)$ (where h is some specific value, usually 1, usually for a dichotomous dependent variable) than we could make without the independent variables. G_M is thus analogous to the multivariate F test for linear regression as well as the regression sum of squares.

Designated the "deviance" by McCullagh and Nelder (1989) and others (a term with, at best, mixed meanings when the substantive example is marijuana use and that I will avoid to the extent possible hereafter), D_M has historically been used as a measure of "goodness of fit," which is essentially a test for the statistical significance of the variation *unexplained* by the logistic regression model and is akin to testing for the statistical significance of unexplained variance in an OLS regression model. If a cliche will help, G_M asks how full the cup is (how much improvement the predictors make in predicting the dependent variable), while D_M asks how empty the cup is (how much improvement is needed before the predictors provide the best possible prediction of the dependent variable). While G_M compares the intercept-only model with the full model (the model that includes all the predictors), D_M compares the full model with a *saturated* model (a model that includes all predictors plus all possible interactions among them). In previous versions of SPSS LOGISTIC REGRESSION (and in the first edition of this monograph), D_M was assumed to have an approximately χ^2 distribution and was assigned a level of statistical significance. The problem with using D_M as a χ^2 statistic lies in the fact that there are different ways to define a saturated model, resulting in different values for D_M and different degrees of freedom (Simonoff, 1998).

Briefly (and bypassing some detail), as explained by Simonoff (1998), one approach (the one taken in SPSS LOGISTIC REGRESSION and SAS PROC LOGISTIC) is to consider each case as independent (casewise approach), and contributing 1 degree of freedom. The alternative is to consider each combination of values of the predictors, or each *covariate pattern*, as a separate cell in a crosstabulation (contingency table approach), and to calculate degrees of freedom based on the number of covariate patterns (cells in the table) rather than the number of individuals. In either approach, if the number of cases per covariate pattern (cell) is too small or if there are many empty cells, D_M will not generally have a χ^2 distribution and it would be inappropriate to use it as a χ^2 statistic to test goodness of fit (McCullagh & Nelder, 1989; Simonoff, 1998). If there is a large number of cases relative to the number of covariate patterns and sufficient cases per covariate pattern, it is possible to define an appropriate saturated model and to calculate a deviance statistic that will have a χ^2 distribution and the correct degrees of freedom based on the contingency table approach. This is done in SPSS NOMREG and PLUM, both of which can be used to analyze dichotomous as well as nominal or ordinal variables with more than two categories. In NOMREG and PLUM, the Goodness-of-Fit table provides Pearson and deviance χ^2 statistics, the latter based on the -2 log likelihood.

For casewise data, it is still possible to construct a goodness-of-fit index. One commonly available index for dichotomous dependent variables is Hosmer and Lemeshow's (1989) goodness-of-fit index $\hat{C}$, which can be included in the output for SPSS LOGISTIC REGRESSION or SAS PROC LOGISTIC. Hosmer and Lemeshow's goodness-of-fit index was designed primarily as an alternative to avoid the problems associated with using D_M as a goodness-of-fit index for casewise data, and it proceeds by collapsing the data into deciles based on the probability of having the characteristic of interest (for example, being a marijuana user). Other possible goodness-of-fit indices include the score statistic, the Akaike information criterion (AIC), and the Schwartz criterion (a modification of the AIC), all of which are provided in SAS PROC LOGISTIC. The score statistic is, like G_M, a test of the statistical significance of the combined effects of the independent variables in the model. The AIC and the Schwartz criterion, which are briefly discussed in Bollen (1989), are two related indices used to compare models, rather than to provide absolute tests of adequacy of fit. It is possible to compare the AIC or the Schwartz criterion for the fitted model with the AIC or the Schwartz criterion for the model with only the intercept, but this provides little more information than G_{M}.

For some researchers, particularly those who have a strong background in log-linear models or general linear models, or a perspective that is more theoretical than applied, goodness of fit will be an important consideration. Given the goal of the logistic regression model (prediction of a single dependent variable), and consistent both with an applied focus and with the analogy between linear and logistic regression, it seems advisable for most purposes to focus here primarily on G_{M}.

2.2.1. Measures of Multiple Association Between the Independent Variables and the Dependent Variable

Several analogues to the linear regression R^2 have been proposed for logistic regression. For general reviews, see Hagle and Mitchell (1992), Menard (2000), and Veall and Zimmerman (1996). Here the focus is on R^2 analogues that are commonly used in general purpose statistical packages such as SAS and SPSS, and on some general categories of coefficients of determination with which they may be compared. If we maintain the analogy between the −2LL statistics for logistic regression and the sums of squares for linear regression analysis, the most natural choice, directly analogous to SSR/SST, is the likelihood ratio R^2, $R^2_{\mathrm{L}} = G_{\mathrm{M}}/(D_0) = G_{\mathrm{M}}/(G_{\mathrm{M}}+D_{\mathrm{M}})$ (McFadden, 1974; see also Agresti, 1990, pp. 110–111; DeMaris, 1992, p. 53; Hosmer & Lemeshow, 1989, p. 148; Knoke & Burke, 1980, p. 41; Menard, 2000). R^2_{L} is a *proportional reduction in −2LL* or a *proportional reduction in the absolute value of the log-likelihood* measure, where the −2LL or the absolute value of the log likelihood—the quantity being minimized to select the model parameters—is taken as a measure of "variation" (Nagelkerke, 1991), not identical but analogous to the variance in OLS regression. R^2_{L} indicates how much inclusion of the independent variables in the model reduces the variation, as measured by D_0. The variation is between 0 (for a model in which $G_{\mathrm{M}} = 0$, $D_{\mathrm{M}} = D_0$, and the independent variables are useless in predicting the dependent variable) and 1 (for a model in which $G_{\mathrm{M}} = D_0$, $D_{\mathrm{M}} = 0$, and the model predicts the dependent variable with perfect accuracy). R^2_{L} can be obtained directly from the output for SPSS NOMREG and PLUM, where it is presented as the McFadden R^2 in the Pseudo-R^2 table. Curiously it is not included (as of this writing) in SPSS LOGISTIC REGRESSION. Instead, in SPSS LOGISTIC REGRESSION and SAS PROC LOGISTIC, it must be computed by hand from the information provided (as described previously) on D_0 (or D_{M}) and G_{M}.[7]

Two measures used in the current versions of SPSS and SAS are (1) the geometric mean squared improvement per observation $R^2_{\mathrm{M}} = 1 - (L_0/L_{\mathrm{M}})^{2/N}$, where L_0 is the likelihood function for the model that contains only the intercept, L_{M} is the likelihood function that contains all the predictors, and N is the total number of cases (Cox & Snell, 1989; Maddala, 1983, pp. 39–40), and (2) an adjusted geometric mean squared improvement per observation R^2_N (Cragg & Uhler, 1970; Maddala, 1983, p. 40; Nagelkerke, 1991). The unadjusted measure cannot have a value of 1, even for a model that fits the data perfectly. The adjusted measure permits a value of 1 by dividing by the maximum possible value of R^2_{M} for a particular dependent variable in a particular data set: $R^2_{\mathrm{N}} = [1 - (L_0/L_{\mathrm{M}})^{2/N}]/[1 - (L_0)^{2/N}] = R_{\mathrm{M}}{}^2$/(maximum possible R^2_M). In SPSS LOGISTIC REGRESSION, R^2_{M} and R^2_{N} are presented, respectively, as the Cox–Snell and Nagelkerke R^2 measures in the Model Summary table or as the Cox–Snell and Nagelkerke pseudo-R^2 measures in the Pseudo-R^2 table in SPSS NOMREG and PLUM. In SAS PROC LOGISTIC, they are simply referred to as the R^2 and adjusted R^2.

A family of alternatives to R^2_{L} includes the pseudo-R^2 or contingency coefficient R^2_{C}, which was proposed by Aldrich and Nelson (1984) in their discussion of logit and probit models, the Wald R^2_{W} (Magee, 1990), and the McKelvey and Zavoina (1975) R^2_{MZ}. In the notation used in this monograph, if N is the number of cases, $R^2_{\mathrm{C}} = G_{\mathrm{M}}/(G_{\mathrm{M}}+N)$. Similarly, the Wald $R^2_{\mathrm{W}} = W/(W+N)$, where W is the multivariate Wald statistic. The McKelvey–Zavoina $R^2_{\mathrm{MZ}} = s^2_{\hat{Y}}/(s^2_{\hat{Y}}+1)$ for the probit model (the context in which it was originally developed) or $R^2_{\mathrm{MZ}} = s^2_{\hat{Y}}/(s^2_{\hat{Y}} + \pi^2/3)$ for a logit or logistic regression model, where $s^2_{\hat{Y}}$ is the variance in $\widehat{Y}$ (the predicted value of Y), and 1 and $\pi^2/3$ are the standard deviations for the standard normal and logistic distributions, respectively. These measures share the common feature that they cannot attain a value of 1, even for a perfect model fit. Hagle and Mitchell (1992) suggested a correction for Aldrich and Nelson's pseudo-R^2 that allows it to vary from 0 to 1; in principle, this approach could also be applied to the Wald and McKelvey–Zavoina measures.

Hagle and Mitchell also noted that the corrected R^2_C provided a good approximation for the OLS regression R^2, and Veall and Zimmerman noted the same with respect to the McKelvey–Zavoina R^2_{MZ}, *when the dichotomous dependent variable represents a latent interval scale*. In this instance, however, there are several other alternatives, including the possibility of using a linear probability model (because the restriction of values to a dichotomy is really artificial for a latent interval scale), using polychoric correlation and weighted least-squares estimation in the context of a more complex structural equation model (Jöreskog & Sörbom, 1993), and using R^2 itself to measure the strength of the association between the observed and predicted values of the dependent variable.

The use of R^2, the familiar coefficient of determination from OLS linear regression analysis, has received relatively little attention in the literature on logistic regression analysis. (For an exception, see Agresti, 1990, pp. 111–112.) Its utility in logistic regression has been questioned because, unlike R^2_L and Aldrich and Nelson's pseudo-R^2, it is not based on the criteria used to select the model parameters. Also, if the dichotomous dependent variable is assumed to be an indicator for an unmeasured latent variable, R^2 provides a biased estimate of the explained variance. There are certain advantages to the use of R^2, not instead of R^2_L, but as a supplemental measure of association between the independent variables and the dependent variable. First, using R^2 permits direct comparison of logistic regression models with linear probability, analysis of variance, and discriminant analysis models when predicting the observed value (instead of predicting the observed probability that the dependent variable is equal to that value) is of interest. Second, R^2 is useful in calculating standardized logistic regression coefficients, a topic to be covered in the next chapter. Third, R^2 is relatively easy to calculate using existing statistical software.

To calculate R^2 for logistic regression, assume that the dependent variable is Y and that you want to name the variable that represents the value of Y predicted by the logistic regression model LPREDY. In SPSS and SAS, to obtain R^2, it is necessary to save the predicted values of the dependent variable from SPSS LOGISTIC REGRESSION [using SAVE = PRED(LPREDY)] or from SAS PROC LOGISTIC [using OUTPUT PRED = LPREDY]. Next, use a bivariate or multiple regression routine (such as SPSS REGRESSION or SAS PROC REG) to calculate R^2. Alternatively, use any analysis of variance routine that calculates η^2 or η (SPSS MEANS or ANOVA; SAS PROC GLM or ANOVA) with the *observed* value of the dependent variable, Y, as the *independent* variable and the *predicted* value of the dependent variable, LPREDY, as the *dependent* variable. Because there are only two variables (the observed values of Y as one variable, the predicted values of Y as the other), $\eta^2 = R^2$ and the two variables may be used interchangeably. Although for η^2 this role switching between the dependent variable and its predicted value (which is based on the values of the independent variables) may seem strange for η^2, it exactly parallels the method for calculating canonical correlation coefficients in discriminant analysis (Klecka, 1980).

Based on research on the properties of the different proposed measures, I have suggested (Menard, 2000) that R^2_L is the most appropriate for logistic regression, based on several considerations.[8] First and most importantly, R^2_L is conceptually closest to the OLS R^2 insofar as it reflects a proportional reduction in the quantity actually being minimized (−2LL; equivalently, the log likelihood is being maximized), in contrast to R^2, R^2_W, and R^2_{MZ}. Also, unlike measures that depend on the sample size as well as the log likelihood or −2LL (R^2_M, R^2_N, R^2_C), R^2_L depends *only* on the quantity being maximized or minimized. Second, R^2_L is not sensitive to the *base rate*, the proportion of cases that have the attribute (for example, being or not being a marijuana user) being studied. Evidence indicates that R^2_M, R^2_N, R^2_C, and R^2 all have the undesirable property that their value increases as the base rate (whichever is smaller, $n_{Y=1}/N$ or $n_{Y=0}/N$) increases from 0 to .50, absurdly suggesting that one could, in effect, substitute the sample size for one of these coefficients of determination as a measure of explained variation (Menard, 2000, p. 23). Third, R^2_L, unlike the unadjusted versions of R^2_W, R^2_C, and R^2_{MZ}, varies between 0 and 1, where 0 represents no predictive utility for the independent variables and 1 represents perfect prediction. Fourth, as noted by Veall and Zimmerman (1996), R^2_L works as well for polytomous nominal or ordinal dependent variables as for dichotomous dependent variables, in contrast to the variance-based measures R^2_{MZ} and R^2.

2.3. Predictive Efficiency: λ_p, τ_p, ϕ_p, and the Binomial Test

In addition to statistics regarding goodness of fit, logistic regression programs commonly print classification tables that indicate the predicted and observed values of the dependent variable for the cases in the analysis. These tables resemble the contingency tables produced by SPSS CROSSTABS and SAS PROC FREQ. In most instances, we will be more interested in how well the model predicts probabilities, $P(Y_j = 1)$. In other cases, however, we may be more interested in the accurate prediction of group membership, so the classification tables may be of as much or more interest than the overall fit of the model. There is no consensus at present on how to measure the association between the observed and predicted classification of cases based on logistic regression or related methods such as discriminant analysis. There are, however, several good suggestions that can easily be implemented to provide summary measures for classification tables. The best options for analyzing the prediction tables provided by logistic regression packages involve *proportional change in error* measures of the form

$$\text{predictive efficiency} = \frac{(\text{errors without model}) - (\text{errors with model})}{(\text{errors without model})}, \qquad [2.1]$$

which is a *proportional change in error* formula. If the model improves our prediction of the dependent variable, this formula is the same as a *proportional reduction in error* (PRE) formula. It is possible under some circumstances, however, that a model actually will do worse than chance at predicting the values of the dependent variable. When that occurs, the predictive efficiency is negative and we have a proportional *increase* in error. The errors with the model are simply the number of cases for which the predicted value of the dependent variable is incorrect. The errors without the model differ for the three indices and depend on whether we are using a prediction, classification, or selection model.

2.3.1. *Prediction, Classification, and Selection Models*

In prediction models, the attempt is made to classify cases according to whether they satisfy some criterion, such as success in college, absence of behavioral or emotional problems in the military, or involvement in illegal behavior after release from prison. In prediction models, there are no *a priori* constraints on the number or proportion of cases predicted to have or not have the specified behavior or characteristic. In principle, it is possible (but not necessary) to have the

same number of cases *predicted* to be "positive" (having the behavior or characteristic, e.g., "successes") and "negative" (not having the behavior or characteristics, e.g., "failures") as are *observed* to be positive and negative. That is, there is nothing that constrains the *marginal distributions* (the number or proportion of cases in each category, positive or negative) of predicted and observed frequencies to be equal or unequal. In particular, all cases may be predicted to belong to the same category, that is, the sample or population may be *homogeneous*. In practical terms, prediction models are appropriate when identical treatment of all groups ("lock 'em all up" or "let 'em all go") is a viable option.

In classification models, the goal is similar to that of prediction models, but there is the added assumption that the cases are truly heterogeneous. Correspondingly, the evaluation of a classification model imposes the constraint that the model should classify as many cases into each category as are actually observed in each category. The proportion or number of cases observed to be in each category (the *base rate*) should be the same as the proportion or number of cases predicted to be in each category. To the extent that a model fails to meet this criterion, it fails as a classification model. Complete homogeneity is an unacceptable solution for a classification model. Practically speaking, classification models are appropriate when heterogeneity is assumed, and identical treatment of all groups is not a viable option.

In selection models (Wiggins, 1973), the concern is with "accepting" or "rejecting" cases for inclusion in a group, based both on whether they will satisfy some criterion for success in the group and on the minimum required, maximum allowable, or specified number of cases that may (or must) be included in the group. In selection models, the proportion of cases observed to be successful (the *base rate* again) may or may not be equal to the proportion of cases accepted or selected for inclusion in the group (the *selection ratio*). For example, a company may need to fill 20 positions from a pool of 200 applicants. The selection ratio will be $20/200 = .10$ (10%) regardless of whether the base rate (the observed probability of success on the job) is 5% or 20%, half or twice the selection ratio. The classification tables provided in logistic regression packages may naturally be regarded as prediction or classification models. They may be used to construct selection models, but they must be altered (unless, purely by coincidence, the selection ratio turns out to be equal to the base rate) so that the correct number of cases is selected.