05-Spicer.qxd 7/12/04 9:06 AM Page 123

FIVE

LOGISTIC REGRESSION AND DISCRIMINANT ANALYSIS

In the previous chapter, multiple regression was presented as a flexible technique for analyzing the relationships between multiple independent variables and a single dependent variable. Much of its flexibility is due to the way in which all sorts of independent variables can be accommodated. However, this flexibility stops short of allowing a *dependent* variable consisting of categories. How then can the analyst deal with data representing multiple independent variables and a categorical dependent variable? How can independent variables be used to account for differences in categories?

This chapter introduces two techniques for accomplishing this aim: logistic regression and discriminant analysis. Even though the two techniques often reveal the same patterns in a set of data, they do so in different ways and require different assumptions. As the name implies, logistic regression draws on much of the same logic as ordinary least squares regression, so it is helpful to discuss it first, immediately after Chapter 4. Discriminant analysis sits alongside multivariate analysis of variance, the topic of Chapter 6, so discussing it second will help to build a bridge across the present chapter and the next. That said, the multivariate strategy of forming a composite of weighted independent variables remains central, despite differences in the ways in which it is accomplished.

In Subsection 5.1.1 we explore the nature of the weighted composite variable in logistic regression with a dichotomous dependent variable and introduce the main statistical tools that accompany it. Subsection 5.1.2 shows two-group, or "binary," logistic regression in action, first with further analyses

05-Spicer.qxd 7/12/04 9:06 AM Page 124

of the nurses' data introduced in Chapter 4 and then with examples from the research literature. In Subsection 5.1.3 the usual questions of trustworthiness will be raised with specific reference to logistic regression. Then in Subsection 5.1.4 extensions to the basic technique are discussed, including how to deal with different types of independent variables and with a dependent variable that has more than two categories. Subsections 5.1.3 and 5.1.4 will be relatively brief since they will draw heavily on material we have already covered in Chapter 4 on multiple regression. The second half of this chapter, comprising Subsections 5.2.1–5.2.4, follows the same sequence of topics for discriminant analysis.

5.1 LOGISTIC REGRESSION

5.1.1 The Composite Variable in Logistic Regression

Although it is inappropriate to use ordinary least squares (OLS) regression when the dependent variable is categorical, it is instructive to begin by asking how the composite variable would function if OLS regression were used. In its most general form the relationship between multiple independent variables (IVs) and a single dependent variable (DV) is:

DV = [coefficient 1(effect 1) + coefficient 2(effect 2) + ... + constant] + residual

For OLS regression, this general expression becomes:

DV = [slope 1(IV1) + slope 2(IV2) + ... + Y intercept] + residual

The composite variable in the square brackets generates predicted scores on the dependent or Y variable. Values for the slopes and Y intercept are chosen that maximize the correlation between the actual and predicted Y scores or, equivalently, minimize the gap or residual between them.

What happens if this strategy is applied to data in which the dependent variable consists of two categories, labeled 0 and 1 (i.e., a dummy variable)?

05-Spicer.qxd 7/12/04 9:06 AM Page 125

The composite cannot be used to generate predicted *scores* on the dependent variable since there are no scores to predict, only categories. Instead the composite now generates the predicted *probability* of a case being in the category labeled 1. These predicted probability values should lie between 0 and 1 and can be subtracted from the actual 0 and 1 values to obtain residuals. The regression slope will have the usual interpretation, except that it will be in probability terms: for every 1-unit change in a given independent variable there will be a change in probability of being in category 1, which is equivalent to the slope value. All of this makes it sound as if OLS regression is well suited to a categorical dependent variable, so where is the problem?

Actually, there are several problems that have been detailed with great clarity by Pampel (2000), to whose primer on logistic regression the present account is much indebted. In summary, using OLS regression to generate predicted probabilities can produce values outside the 0 to 1 range, forces linearity on what is more likely an S-shaped relationship, violates the assumption that the components of the composite variable are additive, and violates the assumptions of normality and homoscedasticity required for statistical tests. After such a list of charges, there seems little option but to seek an alternative strategy. The logistic regression strategy retains the goal of generating predicted probabilities but achieves it indirectly by using another probability index and a different criterion to choose the coefficients in the composite variable. These two "moves" make for a convoluted and abstract journey from the data to the results. We will just highlight the landmarks along the way and as usual emphasize the familiarity of the big road map.

In the present context the probability of being in one of two groups is provided by the relative frequency, that is, the number of cases in one group divided by the number of cases in both groups. If group 1 contained 80 cases and group 0 contained 20 cases, the probability of being in group 1 would be 80/100 = .8 or 80%. This is the type of probability that we are trying to predict, but that is inadequately predicted using OLS regression. To obtain more accurate predicted probabilities, the first step is to focus on another type of probability index that we encountered in Chapter 1: the odds. The odds of being in group 1 for our imaginary 100 cases would be 80/20 = 4. A case is four times more likely to be in group 1 than in group 0. Since the probability and the odds combine the same frequencies in different ways, they are obviously closely related (the probability is just the odds divided by the odds plus 1). But this simple move opens the door to a solution to the problem of predicting probabilities.

05-Spicer.qxd 7/12/04 9:06 AM Page 126

The next step within the first "move" is to change the scale of the odds by transforming it, for reasons that will become apparent shortly. The specific transformation is to replace the odds with its natural log. The natural log of a number is the power to which 2.718 has to be raised to produce that number. So now, instead of dealing with odds, we are confronting log odds, also known as **logits.** What is the payoff for this mind-numbing shift into mathematical abstraction? It can be summed up in the following:

predicted log odds of a DV = [logistic coefficient 1(IV1)
+ logistic coefficient 2(IV2)
+ . . . . . + constant]

Working with log odds rather than probabilities as such means that the familiar composite of independent variables is applicable and retains all its usual properties. In terms of the problems raised earlier, the composite will capture an S-shaped relationship between the independent and dependent variables, it will be additive, and the predicted probabilities that can be derived from it will fall between 0 and 1. The logistic coefficients will be interpretable as statistically controlled effects as usual although, since they are on a log odds scale, they will require some massaging to be useful. But before we delve into this sort of detail, we need to ask how the logistic coefficients including the constant are obtained: the second move in the overall strategy.

As we just noted, a predicted probability for each case can be derived from the log odds and consequently so can a residual—the difference between the prediction for that case and their actual 1 or 0 status. However, the regression coefficients that minimize the residuals' sum of squares for all the cases, that is, that meet the ordinary least squares criterion, will not necessarily maximize predictive power. Moreover, any statistical tests that are based on this way of choosing coefficients will violate the assumptions of normality and homoscedasticity and produce inaccurate *p* values. To avoid these problems, a different criterion for selecting coefficients is adopted: the criterion of **maximum likelihood.**

Under this criterion, the aim is still to minimize the difference between a case's predicted probability of being in a category and its actual category. The search is for the coefficients that will produce the log odds that in turn produce the predicted probabilities that will most accurately place cases in their actual category. So the maximum likelihood criterion produces the logistic coefficients that will most closely reproduce the actual categories in which cases

05-Spicer.qxd 7/21/04 11:56 AM Page 127

appear. The predicted probabilities and actual categories for each case are bundled up, not into a sum of squares package for all cases, but into a statistic called the **log likelihood function.** So, in an opaque nutshell, the aim is to find the coefficients that maximize the value of the log likelihood function. To make matters even more opaque, the log likelihood function is often multiplied by −2 to turn it into the **log likelihood chi²** statistic, as we will see shortly. Multiplying by −2 also means that the log likelihood values range from 0 to positive infinity and that the strategic aim is now to find coefficients that *minimize* the value of this function. These rapid turnabouts should become less dizzying when we look at logistic regression in action, below.

To summarize, the relationships between multiple independent variables and a categorical dependent variable can be analyzed using a technique called logistic regression. This involves forming the independent variables into the usual weighted, additive composite, which is then used to predict the probability of cases appearing in a particular category of the dependent variable. However, in order to achieve this legitimately, two moves are made. First, the predicted probabilities are derived indirectly through logged odds, or logits. Second, the coefficients in the composite are calculated using a procedure called maximum likelihood estimation. A set of coefficients is chosen provisionally that, through log odds, generates the probability of each case being in a given category. These probabilities and the actual category memberships are fed into the log likelihood function, which produces a particular log likelihood value. Then different sets of coefficients are tried and those that produce the maximum log likelihood value are the ones that are finally selected as the logistic coefficients. To make all of this more concrete, and to see what other statistics result, we now turn to some actual logistic regression analyses.

5.1.2 Binary Logistic Regression in Action

A large part of Chapter 4 was spent in exploring the use of multiple regression to analyze the effects of workplace characteristics on the mental health of a group of nurses (Budge, Carryer & Wood, 2003). To highlight the similarities and differences between ordinary least squares and logistic regression, we will now return to this data set. We will reanalyze the relationships between workplace characteristics and mental health, but with the latter now treated as a dichotomous categorical variable. So for present purposes, the data

05-Spicer.qxd 7/12/04 9:06 AM Page 128

Table 5.1 Mean Scores and *F* Tests on Three Workplace Characteristics for Nurses in Good and Poor Mental Health

Variable	*Good Health Mean*	*Poor Health Mean*	*F*	*p*
Professional relations	14.70	13.40	8.35	.004
Autonomy	13.60	13.10	1.23	.269
Control	17.90	17.20	0.82	.368

set consists of three independent variables: professional relations, autonomy, and control; and a dichotomous dependent variable: mental health coded 1 for good health and 0 for poor health. Of the 163 nurses in the sample, 91 reported good mental health and 72 reported poor health. The earlier multiple regression analyses also included age as an independent variable, but since it had no effect on mental health and adds distracting detail, it has been omitted from the present analyses.

Before we embark on the logistic regression, it is helpful to gain a bivariate overview of the data, just as an inspection of bivariate correlations is advisable in multiple regression. Since we are interested in the relationships between interval and categorical variables, we can make use of mean differences and ANOVAs (or *t* tests) to achieve this. Table 5.1 summarizes the bivariate relationships between the independent and dependent variables.

The mean differences indicate that nurses in good mental health report better relations, better autonomy, and better control than those in poor health. However, the *F* tests suggest that only in the case of professional relations is the difference statistically significant. These separate bivariate analyses are informative, but they fail to take into account the correlations among the three work variables. Since professional relations correlates .56 with autonomy and .48 with control, and autonomy correlates .69 with control, it is quite possible that the conclusions we have just drawn are distorted by confounding. This is one of the fundamental reasons for turning to a multivariate analysis, in this case logistic regression analysis. As usual, we take a top-down approach, beginning with the performance of the composite variable overall and then proceeding to examine particular coefficients, if this appears justified.

The –2 log likelihood statistic of 215.15, which is the lowest value that emerged from trying out different sets of coefficients, reflects the multivariate relationship. Since it is significant by chi^2 test at $p = .035$, we can conclude that

05-Spicer.qxd 7/12/04 9:06 AM Page 129

there is a statistically significant relationship between the set of independent variables and the dependent variable, if we adopt an alpha of .05. But what exactly is the null hypothesis under test here? Testing omnibus hypotheses in logistic regression is an inherently comparative exercise. In fact, this is usually the case in statistical analysis, but here it becomes more explicit. As we have noted, a set of independent variables or effects is referred to as a "model," so more precisely we are engaged in a comparison of models. In the present case the model containing the three independent variables is being compared with a model that contains only the constant. In other words, we are testing whether knowledge of the workplace characteristics improves our ability to predict mental health status. If we were to do a logistic regression *without* the independent variables, the "baseline" –2 log likelihood would be 223.75. Including the independent variables reduces the –2 log likelihood to 215.15: an improvement of 8.6. It is this *change* that is indexed and tested by the log likelihood chi^2 statistic, the null hypothesis being that there is no change in the population. This is conceptually parallel to the statistical testing of R^2 change in sequential regression. At the first step of a stepwise multiple regression, for example, the statistical test can be seen as being of the *change* in R^2 from zero, when there are no independent variables present, to whatever value it achieves when the first independent variable enters.

This reference to R^2 in multiple regression highlights the fact that, in logistic regression, there is no straightforward index of the strength of the multivariate relationship between the independent and dependent variables. This should not be surprising given the grounding of R^2 in sums of squares, which are now notable by their absence. The log likelihood statistics are useful for hypothesis testing but do not offer an interpretable measure of association. Various attempts have been made to develop "pseudo" R^2 statistics for logistic regression. For example, the SPSS program provides the Cox & Snell and the Nagelkerke pseudo R^2 statistics, which are .051 and .069, respectively, in the present analysis. So we could tentatively conclude that the three independent variables explain between 5.1% and 6.9% of the variance in mental health. However, there are a variety of such pseudo statistics, all giving different estimates, and none regarded as superior to all the others, so they are best treated with caution if not actually avoided. Note this means that, aside from these pseudo statistics, logistic regression statistics therefore inherently focus on group differences rather than individual differences.

There is another method of expressing the strength of the multivariate relationship, which is not only less contentious but also more intuitively

05-Spicer.qxd 7/12/04 9:06 AM Page 130

appealing and potentially more practicable. This method uses the predicted probabilities to assign cases into the categories of the dependent variable and then compares the results with their actual categories. Cross-classifying cases according to their assigned and actual categories provides another picture of how well the independent variables predict the dependent variable. Since the predicted probabilities are decimal values between 0 and 1, they need to be dichotomized so that they can be compared with the actual 0 and 1 categories in a 2×2 table. In the present analysis, cases with a predicted probability below .5 were assigned to the 0 category, and those with a value above .5 were assigned to the 1 category.

In the nurses' sample about 82% of those in good mental health were correctly classified, while only 33% of those in poor health were accurately predicted. This gives an overall "hit rate" of about 61%. This sounds impressive, especially in the good health group, but these figures have to be compared with what could be achieved even in the absence of any knowledge about the nurses' workplaces. In such a situation one prediction strategy would be to assign all cases to the modal category, the one that actually contains most of the cases. This is the good health category that contains 91 of the 163 cases. Following this prediction rule would result in a 100% hit rate for the good health category and 0% for the poor health category. Overall this would give an average hit rate of about 56%. So, on this strategy, using the workplace predictors increases the overall hit rate from 56% to 61%: a gain of only 5%, but at the cost of a disastrous hit rate for the poor health group.

A less draconian strategy would simply be to use the relative frequencies in the sample as a basis for assignment: 56% for the good health group and 44% for the bad health group. If this defined the baseline or chance expectation, the gain in hit rate for the good health group would be 26% (82% – 56%) and 11% for the poor health group. Details aside, the key points to note in classification analysis are that hit rates need to be compared with chance and that chance can be interpreted in more than one way. We will return to this issue in Subsection 5.2.2 as classification analysis is also used as part of discriminant analysis.

At this point the results indicate that the set of work characteristics is related to mental health to a degree that is unlikely to be due to chance. The magnitude of this relationship is hard to specify precisely, but the pseudo R^2 statistics in the range 5%–7%, and the gain in predictive hit rate, suggest that the relationship is probably weak. Given that there is some relationship, which independent variables are contributing to it? The answer to this question can

05-Spicer.qxd 7/12/04 9:06 AM Page 131

be found in the logistic coefficients and their associated statistical tests, which may be *z* tests or the Wald tests shown in Table 5.2.

Table 5.2 Logistic Regression Statistics Showing the Effects of Three Workplace Characteristics on Mental Health

Independent Variable	*Coefficient*	*Wald Statistic*	*p*	*Odds*
Professional relations	.193	6.753	.009	1.212
Autonomy	−.027	.116	.734	.973
Control	−.012	.057	.812	.988
Constant	−1.907			

The SPSS program uses the Wald statistic to test the null hypothesis that a coefficient value is zero in the population. As Table 5.2 shows, only in the case of the professional relations variable should this null hypothesis be rejected with an associated *p* value of .009. Under some circumstances the Wald statistic can produce misleading results, and so it is wise to check the pattern of results by comparing models. In the present situation the question of interest would be whether the model containing professional relations and the constant would perform better than the constant-only model. In other words, does the professional relations variable truly have predictive power in the absence of the other two predictors?

As we saw earlier, the −2 log likelihood for the constant-only model is 223.75. Running a logistic regression that includes professional relations reduces this figure to 215.55, a reduction and chi^2 of 8.2 that is statistically significant with a *p* value of .004. Moreover, this professional relations model has pseudo R^2 statistics in the range 4.9%–6.6% and a similar gain in hit rate. The significant chi^2 test for the professional relations model, and the similarity of the magnitude statistics in the models containing one or three independent variables, strongly suggest that only the professional relations variable is contributing to differences in mental health.

The logistic coefficient for professional relations of .193 indicates its impact on mental health when the other two independent variables are statistically controlled. However, the fact that it is on a log odds scale means that it is not easy to interpret. The coefficient says that nurses who are higher by one unit on the professional relations scale have a .193 increase in their log odds of being in the good health group. As usual, the coefficients can be positive, as

05-Spicer.qxd 7/12/04 9:06 AM Page 132

in this case, or negative if the relationship is inverse. The coefficients can be turned into probabilities, but these are even more difficult to interpret because the impact is not uniform across the independent variable scale. Instead the most common strategy is to convert the coefficients into odds, and these appear in the last column of Table 5.2. Remember that an odds of 1 indicates no relationship, a value greater than 1 indicates a positive relationship, and a value less than 1 indicates a negative relationship. The odds statistics can be interpreted in terms of percentage change by subtracting 1 and multiplying by 100. So the odds of 1.212 for professional relations mean that for every 1-unit increase in that independent variable, the odds of being in the good mental health group increase by 21.2%. Every unit increase in autonomy produces a 2.7% *decrease* in the odds [100(.973 − 1) = −2.7] and for control the decrease in odds is 1.2%. Bear in mind, though, that the coefficients for autonomy and control are not statistically significant, so we should be treating them as effectively zero and their corresponding odds as 1. The description here is purely for illustrative purposes.

As in OLS regression, confidence intervals may be calculated around logistic coefficients and around the odds. For example, the 95% confidence interval for the professional relations odds ranges from 1.048 to 1.402. So, while the best single odds estimate is 1.212, we can be 95% confident that the population value lies within this range. Finally, it is important to appreciate that the logistic coefficients are unstandardized, and therefore not directly comparable with each other unless the independent variables happen to share the same unit of measurement. According to Pampel (2000), while there are various ways to calculate standardized coefficients, none are truly equivalent to the betas found in ordinary least squares regression. A partial solution is to standardize the independent variables, either before they are entered into the analysis or by multiplying the coefficient for a variable by its standard deviation. However, since the dependent variable remains in its original form, this semistandardization is only a semisolution.

Now that we have discussed the most commonly used statistics in logistic regression, two examples from the research literature may help to consolidate understanding. Kirschenbaum, Oigenblick, and Goldberg (2000) used binary logistic regression to examine differences between two groups of Israeli workers: 77 who had suffered a first-time work injury and 123 who had suffered injuries on multiple occasions. The independent variables were a variety of sociodemographic, work environment, and well-being indicators. For a

05-Spicer.qxd 7/12/04 9:06 AM Page 133

logistic regression containing 11 independent variables, they report a chi^2 of 15.9 with a p value of $< .01$. Clearly, there was a multivariate relationship between the independent variables and the work injury categories that was unlikely to be due to chance. A classification analysis showed that 140 of the 200 cases were correctly classified by the probabilities derived from the model: an overall hit rate of 70%.

Turning to the independent variables, we find that 11 of the 17 logistic coefficients were statistically significant at $p < .05$. Three of these were well-being variables: a feeling that things are going wrong, unhappy with family life, and unhappy with housing. Thus it appears that some aspects of well-being had an influence on proneness to multiple work injuries. However, inspection of the *signs* of the coefficients revealed an anomalous pattern. Multiple injuries were more likely for workers who were unhappy with their housing (coefficient = 3.396), but *less* likely for those who felt that things were going wrong (coefficient = −1.657) or who were unhappy with family life (coefficient = −2.66). The authors then provide some interesting suggestions on how this apparent anomaly might be resolved.

Many reports of logistic regression analyses omit information about the chi^2 for the model, classification results, and the logistic coefficients. Instead they focus on the odds for each independent variable, often including the 95% confidence interval rather than p values. Natvig, Albrektsen, and Qvamstrom (2003), for example, analyzed the predictors of happiness in a sample of 887 Norwegian school adolescents. In one logistic regression the dichotomous dependent variable was very or quite happy versus not happy, and the independent variables were school alienation, school distress, general self-efficacy, school self-efficacy, support from teacher, support from pupils, and decision control. The logistic regression results showed that several of these variables had odds with a 95% confidence interval that did not include 1. These are the variables whose effects would be statistically significant if null hypothesis testing were used with an alpha of .05. The school alienation and general self-efficacy variables exemplify this and can be used to reiterate how odds statistics are interpreted.

The school alienation odds of .47 were less than 1, indicating a negative relationship with happiness. The dependent variable was coded such that the odds are those of being in the very or quite happy group. Subtracting 1 from .47 and multiplying by 100 indicates that with every 1-unit increase in school alienation the odds of being happy decreased by 53%. The confidence interval

05-Spicer.qxd 7/12/04 9:06 AM Page 134

revealed that in the population this decrease could be as great as 69% or as little as 27%. Turning to general self-efficacy, we find that the odds of 1.7 mean that for every 1-unit increase in that variable, the odds of being happy increased by 70%. Again, the confidence interval suggests that we can be 95% confident that the population value for this increase lies between 10% and 163%. Since this is a multivariate logistic analysis, the odds have been adjusted to take account of any associations among the independent variables and, in this example, they have also been adjusted to control for age, gender, and school.

5.1.3 Trustworthiness in Logistic Regression

The issues that bear on the trustworthiness of logistic regression results can be discussed briefly since the details of most of them have already been explored with respect to OLS regression in Section 4.3 of Chapter 4 and more generally in Chapter 2. We will follow the usual sequence of first considering sampling and measurement issues, then the assumptions required for the legitimate use of the technique, and finish with some other general concerns.

The sample size required for logistic regression is typically greater than that needed for OLS regression. Statistical tests of coefficients obtained by maximum likelihood estimation may give misleading results for samples under 100 (Pampel, 2000, p. 30). More independent variables require more cases, and a minimum of 50 cases per independent variable is recommended (Wright, 1995, p. 221). As usual, the appropriate sample size for a given analysis is also dependent on the acceptable levels of Type I and II error, the expected magnitude of the relationships between the independent and dependent variables, the reliability of measurement, and the frequency distribution of the dependent variable. In the logistic regression context, the more unequal the numbers in the categories, the more cases are needed. Add to all this the problem of missing data because of listwise deletion, and the desirability of having enough cases to cross-validate results on a holdout sample, and it becomes painfully clear that logistic regression typically requires cases in the hundreds to guarantee trustworthy results.

Regarding measurement, the independent variables may be on any type of scale, and they are dealt with as in OLS regression, using dummy coding where necessary. The dependent variable is usually categorical and may have two or more categories, as we will see in the next section. Also in the next section, it will become apparent that the dependent variable may be on an

05-Spicer.qxd 7/12/04 9:06 AM Page 135

ordinal scale, again by using dummy coding. However the dependent variable is scaled, it is required that the categories are mutually exclusive and jointly comprehensive. So each case must be locatable in one and only one category. As usual, it is assumed that the data have been produced by reliable and valid measurement procedures. It is important to emphasize that assigning cases to categories may be a highly complex and error-prone process. A simple outcome of a measurement procedure does not mean that the procedure itself is simple.

The assumptions required for statistical tests in logistic regression are far less restrictive than those for OLS regression. There is no formal requirement for multivariate normality, homoscedasticity, or linearity of the independent variables within each category of the dependent variable. However, as Tabachnick and Fidell (2001, p. 521) note, satisfying these conditions among the independent variables for the whole sample may enhance power. The problem of multicollinearity—very high correlations among the independent variables—does apply to logistic regression. All of these assumptions about the independent variables may be evaluated by treating one of the independent variables as a pseudodependent variable and regressing it on all the other independent variables using OLS regression. The tenability of the assumptions can then be examined with the usual OLS diagnostic tools. The assumption of independence of cases remains in place. In other words, each case can appear in the data set only once, and their data must be uncorrelated with the data of any other case. Casewise exploration of the residuals—the difference between the predicted probability and the actual category—may reveal patterns suggesting nonindependence and may identify outliers for whom the model provides notably poor predictions.

To this point we have concentrated on trustworthiness issues that arise in one form or another in regression generally. Two further issues present themselves in logistic but not in OLS regression. The first issue is that the maximum likelihood procedure for estimating the logistic coefficients is an **iterative procedure.** This means that the coefficient values are calculated in a series of steps or iterations rather than in one hit, as in OLS regression. The aim at each iteration is to produce a log likelihood that is greater than that at the preceding iteration. This process continues until a convergence criterion is satisfied, that is, the amount of increase between two iterations is small enough for the solution to be regarded as stable. In some circumstances the procedure may fail to converge on estimates of the coefficients, either because the

05-Spicer.qxd 7/12/04 9:06 AM Page 136

convergence criterion could not be met or the number of permitted iterations was exceeded. Or less dramatically, convergence may be achieved, but only at the cost of a large number of iterations. In all of these situations a warning signal is being given that the data are problematic in some way, and the results may not be trustworthy. So information about the iteration history of a maximum likelihood analysis can be another useful diagnostic tool.

The second issue concerns the trustworthiness of classification analyses. Classification results are never definitive because they depend on at least two decisions made by the analyst that may be questionable. The first is the cut-point used to translate predicted probabilities into predicted categories. The usual default is .5, but this may not be the optimum choice. The second decision concerns the best choice of baseline hit rate against which the achieved hit rates should be judged: an issue we noted earlier. This could be the simple probability (50% in a two category analysis), the relative frequency in the sample, the relative frequency taken from available population data, or a figure based on some other criterion. A different choice of baseline hit rate can give a very different sense of the predictive power of a given model. Even when appropriate decisions on these two issues have been made, the concreteness of classification analysis can also distract from the point that the hit rates are generated from and tested on the same data. This capitalization on chance means that hit rates in any replication are almost inevitably going to be less impressive. Accordingly, when possible it is advisable to generate predictions with one subgroup from the sample and to test their predictive power on another "holdout" subgroup. Failing this, other cross-validation techniques can be used *within* one sample to test the stability of the results across different subsets of the sample.

5.1.4 Extending the Scope of Logistic Regression

Like OLS regression, logistic regression can accommodate independent variables on any measurement scale with the use of dummy coding. For example, Hintikka (2001) examined the relationship between religious attendance and life satisfaction in a random sample of 1,642 adults in Finland, using almost entirely categorical variables. Both of these variables had two categories, while the control variables of sex, employment status, household category, and adequate social support had two, three, or four categories. Age was the only independent variable that was not categorical. A binary logistic

05-Spicer.qxd 7/12/04 9:06 AM Page 137

regression was conducted to assess the relationship between religious attendance and life satisfaction, while controlling for all of the other independent variables. Hintikka reports an adjusted odds for religious attendance of 1.7 with a 95% confidence interval of 1.2–2.4. This means that religious attenders were 70% more likely than nonattenders to be satisfied with their lives. Alternatively, we could say that religious attenders were 1.7 times more likely than nonattenders to be satisfied with their lives.

Logistic regression can also be used to evaluate interaction or moderating effects, using the products of independent variables, as in OLS regression. The study by Natvig et al. (2003) of happiness in school adolescents, discussed earlier, included such interaction terms. Thus their logistic model included not only the independent variables described earlier, but also the product of each separately with age and sex. Since none of these interaction variables were statistically significant, it could be concluded that the predictors of happiness that were found were not moderated by age or sex.

All of the sequential techniques used in OLS regression are also available in logistic regression. Kirschenbaum et al.'s (2000) analysis of the predictors of work accident proneness was actually more complex than the description given earlier. As noted then, the independent variables fell into three groups: sociodemographic, work environment, and well-being characteristics. The analytic strategy was to build a hierarchical model where these blocks of variables were entered in three cumulative steps. This allowed the analysts to examine the predictive gain at each step and to note changes in the coefficient for a particular variable at each step. For example, the coefficient for sex changed from a statistically significant 1.201 at step 1 to a nonsignificant .742 at step 2 and then dropped further to .418 at step 3. Such a pattern suggests confounding or possibly mediation if other conditions were met. In fact, even this description understates the complexity of the analysis because the selection of particular variables into the blocks was guided by earlier forward and backward logistic regressions. That is, variables were selected for inclusion in the blocks according to statistical rules rather than by the analysts.

The final extension of logistic regression in this section concerns the structure of the dependent variable. To this point we have focused on binary logistic regression, which allows for a two-category dependent variable. More than two categories can be accommodated with the technique of **multinomial or polytomous logistic regression.** To achieve this, the categories are converted into a set of dummy variables, one less than the number of categories.

05-Spicer.qxd 7/12/04 9:06 AM Page 138

Each dummy variable represents a particular difference between particular categories, either singly or in sets. As we saw in Chapter 4, two systems of particular interest are reference and ordinal coding. In the first, a particular category is chosen as a reference, and each dummy variable represents a difference between that category and each of the others. In ordinal coding the ordinal scaling of the dependent variable is represented by a set of dummy variables, each representing a comparison of the sets of categories above and below each scale point.

In multinomial logistic regression, a logistic model is estimated for each dummy dependent variable. This means that no new interpretive issues arise, and the only concern is to be clear about the particular difference that a given model is estimating. Returning to the study of the predictors of happiness in adolescent schoolchildren (Natvig et al., 2003), the researchers conducted both binary and multinomial logistic regressions. In the former, as we saw earlier, the dependent variable was very or quite happy versus not happy. For the multinomial analysis, they created two dummy variables: very happy versus not happy, and quite happy versus not happy, to represent three categories of happiness. This is an example of reference dummy coding with not happy as the reference category.

Earlier in the binary logistic regression we saw that one of the successful predictors, school alienation, differentiated between the very or quite happy and not happy categories with an odds of .47. The multinomial regression produced odds of .53 and .35 for the very happy versus not happy and quite happy versus not happy contrasts, respectively, and neither of their 95% confidence intervals included 1. This means that school alienation differentiates the not happy category from the other two, both singly and jointly, and this pattern is unlikely to be due to chance. However, a different pattern emerged for the other successful predictor—general self-efficacy. For this variable, the binary odds for the very or quite happy versus not happy comparison were 1.7. The multinomial regression produced odds of 1.39 and 2.89 for the very happy versus not happy and quite happy versus not happy contrasts, respectively, but the 95% confidence interval for the former did not include 1. Accordingly, it appears that general self-efficacy did not predict the difference between the quite happy and not happy categories. Details aside, it should be apparent that multinomial logistic regression provides the capacity not only to accommodate a variety of categorical and ordinal dependent variables, but also to detect specific differences between categories within these variables.

05-Spicer.qxd 7/12/04 9:06 AM Page 139

This completes our introduction to logistic regression. We now turn to an alternative multivariate technique for analyzing the relationships between multiple independent variables and a single categorical variable: discriminant analysis.

5.2 DISCRIMINANT ANALYSIS

5.2.1 The Composite Variable in Discriminant Analysis

Discriminant analysis captures the relationship between multiple independent variables and a categorical dependent variable in the usual multivariate way, by forming a composite of the independent variables. So, discriminant analysis and logistic regression can be used to address the same types of research question. As in logistic regression, the variable generated by the composite cannot be a predicted score on the dependent variable. Instead it is a **discriminant function score** that then feeds into calculations that produce the predicted probability of a case being in a particular category of the dependent variable. This predicted probability is then used to generate a predicted category for each case. So, in broad terms the strategy is very similar to logistic regression in which the composite variable generates logits, which produce predicted probabilities, which produce predicted categories. The composite variable in two-group discriminant analysis is:

discriminant score = [discriminant coefficient 1(IV1)
+ discriminant coefficient 2(IV2) . . . . . .
+ constant]

The coefficients are now called discriminant function coefficients. For each case, the coefficient for an independent variable is multiplied by the case's score on that variable; these products are summed and added to the constant; and the result is a composite score for that case—their discriminant score. From these scores can be derived predicted probabilities and predicted group membership on the dependent variable.

Before we look more closely at the coefficients, it would be helpful to discuss the principle by which they are calculated. This principle will be clearer if we first pause to appreciate the hybrid nature of discriminant analysis and to review briefly some material from Chapter 1. When we consider the typical interpretation and application of the technique, it is convenient to frame it in

05-Spicer.qxd 7/12/04 9:06 AM Page 140

regression terms: the prediction of a categorical dependent variable using multiple independent variables. However, it is easier to appreciate *how* the technique works if we frame it as a form of multivariate analysis of variance (MANOVA), which indeed it is. MANOVA is the multivariate form of ANOVA in which there are multiple *dependent* variables. (This is discussed in detail in Chapter 6.) The unsettling consequence of this shift in perspective is that we need to reverse the status of the independent and dependent variables temporarily. So from a MANOVA perspective we are now asking how well a categorical variable accounts for differences in a set of *dependent variables.* To make this more concrete, in the next section we will return to the nurses' data and the relationship that autonomy, control, and professional relations have with good versus poor mental health. From a regression perspective, there are three independent variables and one dependent variable. But from the MANOVA perspective that we now adopt temporarily, we have one independent categorical variable and three dependent variables. This may sound like cheating, but in fact it just highlights the way in which independent and dependent variable status is something imposed by the analyst rather than embedded in the statistics.

Imagine that we want to analyze the bivariate relationship between the nurses' mental health, treated as a dichotomous independent variable, and their professional relations treated as a dependent variable. In Chapter 1 we saw how analysis of variance can be used to analyze the relationship between a categorical independent variable and an interval-level dependent variable. In fact, we quickly carried out an ANOVA on the relationship between mental health and professional relations in Subsection 5.1.2 of this chapter. At the heart of the ANOVA strategy is the idea of capturing group differences on the dependent variable with a between-groups sum of squares and individual differences with a within-group sum of squares. The between-group and within-group sums of squares add up to the total sum of squares, which represents all of the individual differences on the dependent variable, regardless of group. The basic rationale of this approach is that the bigger the between-groups sum of squares is relative to the within-group sum of squares, the more likely it is that the independent and dependent variables are related. In Chapter 1 we saw how this relationship can be indexed with the ratio of between-group/total sum of squares (eta^2 = explained variability), or of within-group/total sum of squares (Wilks's lambda = unexplained variability). Further, the ratio of between-group/within-group sum of squares can be changed into a ratio of variances that then becomes the test statistic known as the *F* ratio.

05-Spicer.qxd 7/12/04 9:06 AM Page 141

For the nurses' data, those in good mental health have a mean professional relations score of 14.7, while those in poor mental health have a mean score of 13.4: a mean difference of 1.3, as we saw in Table 5.1. Analysis of variance gives an eta^2 of .049, a Wilks's lambda of .951, and an F of 8.35 (p = .004). The F statistic reassures that the relationship is unlikely to be due to chance, and the other statistics indicate that mental health status accounts for 4.9% of variance in professional relations or, conversely, that it leaves 95.1% of variance unexplained. As we move into discriminant analysis, we carry forward from this bivariate analysis two particular perspectives. The first focuses on the *distance* between the two means. The second focuses on the ratio of the between-groups/within-groups sum of squares, which lies at the heart of the F ratio. This sum of squares ratio is known as the **eigenvalue,** and it is the statistic on which discriminant analysis pivots.

In discriminant analysis the ANOVA logic we have been reviewing is applied to the composite variable: the discriminant score. If we return to the example in which there are three quasi-dependent variables (autonomy, control, and professional relations), each nurse will have a discriminant score that combines their weighted scores on these three variables. These discriminant scores can be divided into the good and poor mental health groups, and the mean discriminant score can be calculated for each group. The group means on the composite variable are known as **centroids.** Now we are finally in a position to state the principle by which the discriminant coefficients or weights are selected. They are chosen so that the distance between the centroids is maximized, within certain constraints that need not concern us. So coefficients are chosen that push the group means on the composite variable as far apart as possible, that is, that maximally discriminate between the two groups.

The principle can also be stated in terms of eigenvalues. Discriminant coefficients are chosen that maximize the eigenvalue for the composite variable, that is, the ratio of between-group to within-group sums of squares. A critical feature of these composite sums of squares is that they encapsulate, not only the variability of each variable, but also their *covariability.* This means that the coefficients are partial, just as in multiple and logistic regression, so each indicates the contribution of a particular variable while statistically controlling for all of the others. Further, the coefficients can again be calculated in unstandardized or standardized form, as in multiple regression. That said, we will see in the next section that discriminant coefficients are less informative than those in regression, whatever their form. After so many abstractions,

05-Spicer.qxd 7/21/04 11:56 AM Page 142

it is more than time to return to an example where these ideas are made more concrete.

5.2.2 Two-Group Discriminant Analysis in Action

In order to appreciate similarities and differences in the techniques, it will be helpful to begin this section with a discriminant analysis that parallels the logistic regression carried out in Subsection 5.1.2. As a reminder, the data set consists of three independent variables: professional relations, autonomy, and control; and a dichotomous dependent variable: good mental health versus poor mental health. Bear in mind, though, that the independent/dependent status of these variables will flip occasionally in our discussion. This somersault is potentially confusing, but it does avoid deeper confusions that can arise in a nonstatistical account.

We begin with the question of whether the composite variable or discriminant function discriminates between the two groups to a degree that is unlikely to be due to chance. This is equivalent to asking whether the multivariate association between the independent and dependent variables is statistically significant. The null hypothesis that there is no multivariate association in the population can be tested using chi^2, which in the present case is 8.45 with a reassuring p value of .038. Since it is statistically significant it is meaningful to ask about the magnitude of the relationship. Discriminant analysis produces a multivariate version of Wilks's lambda (see Chapter 1), which has a value of .948 in this case. This means that the discriminant function or composite variable fails to account for 94.8% of the variance in mental health status. Conversely, by subtraction the function does account for 5.2% of variance. In the bivariate context this explained variance is indexed by the eta^2 statistic, but in this multivariate context it becomes known as the **canonical correlation^2**. So in the present analysis SPSS reports a canonical correlation of .228, that is the square root of .052. In summary, the discriminant analysis suggests that the three work variables considered as a set are related to mental health status and explain just over 5% of its variance. This outcome opens the way to an inspection of the coefficients in the composite variable to discover which variables are contributing to its discriminating power.

The contributions of individual variables can be shown in a variety of ways, and the most common appear in Table 5.3. The **unstandardized discriminant coefficients** in the first column are the weights used to generate the discriminant score. However, since they do not take account of any differences

05-Spicer.qxd 7/12/04 9:06 AM Page 143

Table 5.3 Three Types of Discriminant Coefficients Showing the Contributions of Three Workplace Variables to the Discriminant Function

Variable	*Unstandardized Coefficients*	*Standardized Coefficients*	*Structure Coefficients*
Professional relations	.407	1.126	.976
Autonomy	−.057	−.174	.375
Control	−.025	−.110	.305
Constant	−4.567		

in the measurement scales of the variable, they are not usually comparable, as in multiple and logistic regression. The **standardized discriminant coefficients** in the middle column are comparable, but only in a limited sense. Their rank order, ignoring the signs, provides an indication of the relative contribution made by variables to the discriminant function. Thus it is clear that the professional relations variable takes the lions' share of the credit in this case. Note that unlike beta coefficients in multiple regression, these coefficients cannot be interpreted in rate of change terms, nor do they have associated statistical tests. Since the standardized coefficients have been adjusted to take account of correlations among the variables, it is also helpful to have an unadjusted view of their contributions for comparison. This is provided by the **discriminant structure coefficients** in the last column. The structure coefficient is the simple correlation between scores on a particular variable and the discriminant scores. It thereby gives an uncluttered view of a variable's contribution and is favored by many analysts because of this. Ideally, as in the present case, the standardized and structure coefficients provide a similar message, although sometimes the differences can be instructive. The message is clearly that the professional relations variable is the key discriminator, as we found in the logistic regression analysis.

Another similarity to logistic regression is the availability of classification analysis: the prediction of group membership and the assessment of its success. In fact, in some discriminant analyses, particularly in applied settings, this is of more interest than the inspection of coefficients. There are a variety of ways to conduct classification analyses that can be pursued in the readings at the end of this chapter. In SPSS the discriminant scores are used to calculate what is called each case's **posterior probability:** their probability of being in a particular category given their discriminant score. This is then adjusted by the case's **prior probability:** the probability of their being in a

05-Spicer.qxd 7/12/04 9:06 AM Page 144

category regardless of their discriminant score. The result of all this is a predicted group membership for each case, that is, the category that is their most probable location given their attributes. This predicted category can then be compared with their actual category to calculate various indices of classification success for the sample as a whole.

In the nurses' sample, about 84% of those in good mental health were correctly classified, and about 32% of those actually in poor health were so classified, giving an overall hit rate of nearly 61%. As we discussed in the context of logistic regression, these figures can and should be compared with what could be achieved by chance. However, as before, the notion of "chance" can take on a variety of meanings, and it is important to choose the most appropriate for the research context. In this form of classification analysis the choice of chance level is equivalent to the choice of prior probabilities. The simplest choice would be the tossed coin model, which gives a prior probability of 50% for either category. Using this would lead to the conclusion that the overall hit rate was 11% better than chance, but that this hid a gain of 34% in the good health group and a loss of 18% in the poor health group. Since the actual groups were not equal in size, a better choice of prior probabilities would be the sample relative frequencies: 56% in the good health group and 44% in the poor health group. Using these as base rates gives an overall hit rate gain of about 10% with a gain of 28% in the good health group and a loss of 12% in the poor health group.

This change in the choice of prior probabilities results in a relatively small shift in the pattern of classification success. However, there may be grounds for choosing quite different prior probabilities that could alter the rates considerably. For example, population figures may be available for the prevalence of mental health in nurses that are notably different from the sample figures and that might provide more accurate prior probabilities. Or there may be good reason to set the prior probabilities in a way that favors the accurate detection of mental health problems at the expense of detecting those in good mental health. Details aside, the two general points to reiterate from the earlier discussion on classification analysis in logistic regression are that the results hinge on the analyst's choice of prior probability, and that until the results are cross-validated in some way they should be regarded as how good classification can get.

The results of the discriminant analysis suggest that work characteristics can discriminate among nurses in good versus poor mental health, albeit to

05-Spicer.qxd 7/12/04 9:06 AM Page 145

a modest degree. However, they also suggest that most if not all of this discriminative power is due to the professional relations variable. Since there are no individual variable tests in discriminant analysis, it is wise to check their contributions by running sets of analyses with them present or absent. If we run a discriminant analysis including only the professional relations variable, we find very similar figures to those resulting from the three variable model. The figures for the single variable model, with those from the three variable model in brackets, are $chi^2 = 8.12$ (8.45), $p = .004$ (.038), Wilks's lambda = .948 (.951), canonical correlation = .222 (.228), and overall classification success 62% (61%). The similarities here make it clear that the autonomy and control variables are quite redundant. Of course, all we have really done here is to repeat the ANOVA we conducted earlier. When there is only one independent variable, discriminant analysis collapses into an analysis of variance: a further demonstration of the cumulative nature of multivariate statistics. The example nonetheless exemplifies the value of comparing models with different subsets of variables to clarify their individual contributions.

To complete this section and to help consolidate understandings of the main features of discriminant analysis, we can turn to an example of a two-group analysis from the research literature on well-being. Philips and Murrell (1994) compared a group of 120 older adults who sought help for their mental health with another similar group of 120 who did not seek help, to see if they differed in terms of their well-being, experience of undesirable events, social integration, social support, and physical health. The discriminating power of these independent variables, accompanied by 10 sociodemographic control variables, was analyzed with a two-group discriminant analysis. The authors report the following statistics for the discriminant function: Wilks's lambda = .5861; $chi^2 = 122.89$, $p < .0001$; canonical correlation = .643. From the chi^2 and associated p value it is clear that the independent variable composite's capacity for discriminating between the two groups was highly unlikely to be due to chance. The extent of this capacity can be quantified by squaring the canonical correlation and concluding that the variables explained about 41% of the variance in help-seeking status. This same figure can be arrived at by subtracting the Wilks's lambda figure from 1 and converting to a percentage, since Wilks's lambda reflects unexplained variance.

To evaluate the contribution of individual variables, the authors present the standardized and structure coefficients for each. The rank orders of these two types of coefficient were strikingly different in some respects, reflecting

05-Spicer.qxd 7/12/04 9:06 AM Page 146

the adjustments made to the standardized coefficients to account for correlations among the independent variables. Like many analysts, the authors chose to base their interpretations on the structure coefficients: the simple, unadjusted correlations between variable and discriminant scores. From these they concluded that the helpseekers had poorer well-being (structure coefficient = .77) and physical health (.63), experienced more undesirable events (.50), and reported less social support (.33). The social integration variable had a structure coefficient of .08, well below the conventional .30 threshold for interpretation (Hair, Anderson, Tatham & Black, 1998). Interestingly, they did not proceed to analyze a model without the sociodemographic variables. The full model suggested that these variables were contributing little, so it would have been interesting to evaluate the discriminating power of the five variables that were apparently making the sole contribution to predicting group membership. It is also interesting to note that the authors chose not to proceed to a classification analysis, despite the implications this might have had for mental health services for older adults.

5.2.3 Trustworthiness in Discriminant Analysis

In this section we review concerns about sampling, measurement, and the statistical assumptions that can influence the trustworthiness of discriminant analysis results. As in the case of logistic regression, the review can be relatively brief since it draws heavily on more extensive discussions in Section 4.3 of Chapter 4 and in Chapter 2.

It is generally recommended that the sample size in a discriminant analysis should provide at least 20 cases for each independent variable. A sample size smaller than this can result in discriminant coefficients that are not stable across samples and therefore not trustworthy (Stevens, 2002, p. 289; Hair et al., 1998, p. 258). It is also recommended that the smallest group size in the dependent variable categories be at least 20, with an absolute minimum greater than the number of independent variables. If these conditions are met, unequal sample sizes across the categories are not problematic in themselves, though they may have implications for the choice of prior probabilities in a classification analysis and may contribute to assumption violation as discussed below. As usual, the more general issues of acquiring enough cases to achieve appropriate levels of Type I and II error for the expected effect size, to compensate for unreliable measurement, to allow for missing data, and to create a holdout

05-Spicer.qxd 7/12/04 9:06 AM Page 147

sample if desired, should all be considered in determining the optimum sample size.

The dependent variable in a discriminant analysis should be categorical and may have any number of categories. The categories should be mutually exclusive and jointly comprehensive, allowing each case to be assigned to a single category. It is assumed that all of the independent variables are measured on at least an interval scale. Including other types of variables using dummy coding will produce meaningful results. However, the more noninterval variables that are included, the less trustworthy the results will be in terms of finding the optimum separation of the groups. In this situation it is usually wiser to resort to logistic regression, which can accommodate any mix of independent variable types. No new issues of measurement quality arise in discriminant analysis; as usual, reliable and valid measurement of all variables is the order of the day.

The statistical assumptions required for discriminant analysis are essentially the same as for OLS regression, though some of them take on a slightly different form. The independent variables are assumed to have a multivariate normal distribution in each population from which the category samples are drawn. As in OLS regression, the consequences of violating this assumption are not usually serious if the sample size requirements above are met. The assumption of multivariate homoscedasticity found in OLS regression takes on a more elaborate form in the present context. Discriminant analysis requires that the population variances *and covariances* for all independent variables are equal across the dependent variable groups. This is known as the **homogeneity of variance-covariance matrices** assumption. The status of the assumption can be explored by inspecting the group variances and covariances, examining appropriate plots, and testing with statistics such as **Box's M.** If the sample sizes in each category are reasonably large and approximately equal, violation of this assumption has little effect on statistical tests, but classification analyses may be distorted. If there is a clear violation, remedies may be found in transformations of variables, or possibly in an alternative approach to classification called quadratic discrimination.

Discriminant analysis also assumes independence of cases and multivariate linearity of relationships among the independent variables in each category of the dependent variable. As in OLS regression, multicollinearity or high correlations among independent variables can be a problem to which the analyst should be alert. Also, outliers on the independent variables may distort the

05-Spicer.qxd 7/12/04 9:06 AM Page 148

results. Examination of univariate and bivariate statistics and plots for the independent variables is important to check for these potential problems and for nonnormality. As we noted earlier, it is also possible to explore the multivariate structure of the independent variables by treating one of them as a pseudodependent variable and conducting an OLS multiple regression. This will generate all of the diagnostic tools that were discussed in Chapter 4 that can now be used to evaluate many of the assumptions required by discriminant analysis.

5.2.4 Extending the Scope of Discriminant Analysis

In Subsection 5.1.4 we discussed how the scope of binary logistic regression could be extended by accommodating various types of independent variables, conducting sequential analyses, and analyzing dependent variables with more than two categories or groups. As we noted earlier, it is not advisable in discriminant analysis to include independent variables that have less than interval scaling, so there is no need to pursue that topic here. All of the sequential strategies, both hierarchical and statistical, can be used in discriminant analysis, though the statistical approach using such techniques as stepwise analysis is the most common application. No new general issues arise when sequential strategies are used in discriminant analysis, so the earlier discussion in Chapter 4 on OLS regression should suffice as an introduction. This then leaves the topic of how discriminant analysis can be applied to a dependent variable with more than two groups.

When discriminant analysis is applied to more than two groups, the major consequence is that more than one discriminant function can be calculated. Each function will have its own set of coefficients and each will generate a discriminant score for every case. Mathematically, it is possible to derive as many functions as there are groups minus 1. So for a four-group analysis, there will be a maximum of three functions, and each case will potentially have three discriminant scores. However, the fact that three functions can be derived does not mean that all are necessary in order to achieve maximum discrimination between the groups. This may be achievable with only one, or perhaps two, of the available functions. Not surprisingly then, the major new issue that arises when the dependent variable has more than two categories is how many functions are worth retaining from those that are available.

The broad strategy for deriving and testing multiple discriminant functions can be confusing, so we will first describe the logic in broad outline and

05-Spicer.qxd 7/12/04 9:06 AM Page 149

then make it less abstract with a closing research example. When there two groups, there is only one dimension along which their two means on the composite variable (centroids) can be pushed apart as far as possible. However, each time a group is added, another dimension emerges along which the centroids can be separated. A four-group discriminant analysis, for example, first calculates the set of coefficients that maximally separates the four centroids: the first discriminant function. It then calculates another set of coefficients that separates the centroids in a completely different way. This is the second discriminant function, whose discriminatory power is unrelated to that found in the first function. The process then continues to derive the third function. Since the objective is to maximize the discriminating power of a function, the result is a series of functions (three in this case) that have decreasing discriminating power and that are uncorrelated with each other.

Another way of thinking about multiple discriminant functions is in terms of explained variance. All of the available functions in an analysis are jointly responsible for any explained variance that is achieved. Deriving separate functions can be seen as assigning portions of this explained variance to each function. The first function will be awarded the largest portion, and the succeeding functions will receive diminishing portions. Moreover, the portions will be mutually exclusive, so that they add up to the total explained variance. In general then, a discriminant analysis will produce a series of functions one less in number than the number of groups or categories in the dependent variable. These functions will be ordered such that they have decreasing discriminating or explanatory power, and each will achieve this power in different ways.

As we noted, the question this creates is whether all of the functions are worth retaining in the analysis. The usual approach to this question is to rely on the statistical significance of functions. Unfortunately, this is not simply a matter of testing the significance of each function. The first step is to test the significance of all the available functions considered jointly. This makes sense, as it is equivalent to testing whether the functions jointly capture more explained variance than would be expected by chance. If this test achieves statistical significance, the way is open to testing for superfluous functions. (If not, the analysis is not worth pursuing at all.) The superfluity tests proceed by testing the significance of subsets of functions, each time omitting the next largest function. So, in the three-function case, the first test would evaluate the joint significance of the second and third functions; and the second test would

05-Spicer.qxd 7/12/04 9:06 AM Page 150

evaluate the significance of the third function alone. The occurrence of a statistically *non*significant result suggests that none of the functions under test at that point are worth retaining. To see this process in action, we can turn to a study involving four groups, which meant that three functions were available in principle.

Diehl, Elnick, Bourbeau, and Labouvie-Vief (1998) conducted a study to examine how adult attachment styles are related to a range of family context and personality variables. They identified 304 cases as exhibiting one of four attachment styles: secure (154), dismissing (77), preoccupied (25), or fearful (48). One of their research objectives was to discover how well a wide range of well-being, family, and personality variables would predict membership of these four groups. To find out, they conducted a four-group discriminant analysis with 12 independent variables. Since there were four groups, it was possible to derive three functions. Their first analytic task was to decide whether three functions were needed to account for any discriminating power of the independent variables, or whether a smaller number would suffice.

All three functions had a joint chi^2 of 109.96, with an associated p value $< .001$. This demonstrated an overall level of discriminatory power that was unlikely to be due to chance. The reported Wilks's lambda of .69 indicates that the functions accounted for $100(1 - .69) = 31\%$ of the variance in attachment styles. The next test of functions 2 and 3 was also significant with a chi^2 value of 47.24 and a $p < .01$. This outcome, and the fact that the second function accounted for about 13% of the variance in attachment styles, suggested that it was worth retaining for its discriminating power. However, when the third function was tested it did not achieve statistical significance and accounted for a minuscule amount of variance in the dependent variable. Moreover, functions 1 and 2 together accounted for over 95% of the explained variance (not of the total variance). All of this suggested that only two functions were required rather than the three that were available in principle.

Since two functions were retained, two sets of coefficients resulted, and the authors present the two sets of structure coefficients as the basis for their interpretations of how the independent variables contribute to the functions. Based on the patterning of the coefficients, the researchers labeled the first function as a "self-model" and the second function as an "other-model." They also examined the group means (centroids) on the two discriminant scores and found that the first function discriminated the secure and dismissing styles from the preoccupied and fearful styles. In contrast, the second function

05-Spicer.qxd 7/12/04 9:06 AM Page 151

discriminated the secure and preoccupied styles from the dismissing and fearful styles. This differential outcome demonstrates nicely how the two functions captured different aspects of the discriminating power of the super function.

Finally, Diehl et al. (1998) conducted a classification analysis. Given the very different numbers of cases in each category, they wisely chose the relative frequencies as their prior probabilities rather than a uniform 25% for each category. The success rates for each category with the prior probability in brackets were secure 50% (51%), dismissing 55.8% (25%), preoccupied 44% (8%), and fearful 50% (16%). The overall success rate for the classification analysis was 51%. As the authors note, the prediction gains from using the discriminating power in the independent variables were over 30%, but only in the groups with insecure styles.

5.3 FURTHER READING

Pampel's (2000) "primer" on logistic regression is exactly that—a model of clear exposition for the novice, while Tabachnick and Fidell (2001, Chapter 12) provide a more extensive, computer-analysis-oriented account. In the present context, the chapter by Hair et al. (1998, Chapter 5) is particularly germane because it discusses logistic regression and discriminant analysis in parallel. After more than 20 years, Klecka's (1980) brief introduction to discriminant analysis remains a valuable source. More extensive treatments of discriminant analysis can be found in Tabachnick and Fidell (2001, Chapter 11), Stevens (2002, Chapter 7), and Huberty (1984, 1994).

05-Spicer.qxd 7/12/04 9:06 AM Page 152