26

MULTINOMIAL LOGISTIC REGRESSION

CAROLYN J. ANDERSON

LESLIE RUTKOWSKI

Chapter 24 presented logistic regression models for dichotomous response variables; however, many discrete response variables have three or more categories (e.g., political view, candidate voted for in an election, preferred mode of transportation, or response options on survey items). Multicategory response variables are found in a wide range of experiments and studies in a variety of different fields. A detailed example presented in this chapter uses data from 600 students from the High School and Beyond study (Tatsuoka & Lohnes, 1988) to look at the differences among high school students who attended academic, general, or vocational programs. The students' socioeconomic status (ordinal), achievement test scores (numerical), and type of school (nominal) are all examined as possible explanatory variables. An example left to the interested reader using the same data set is to model the students' intended career where the possibilities consist of 15 general job types (e.g., school, manager, clerical, sales, military, service, etc.). Possible explanatory variables include gender, achievement test scores, and other variables in the data set.

Many of the concepts used in binary logistic regression, such as the interpretation of parameters in terms of odds ratios and modeling probabilities, carry over to multicategory logistic regression models; however, two major modifications are needed to deal with multiple categories of the response variable. One difference is that with three or more levels of the response variable, there are multiple ways to dichotomize the response variable. If J equals the number of categories of the response variable, then $J(J-1)/2$ different ways exist to dichotomize the categories. In the High School and Beyond study, the three program types can be dichotomized into pairs of programs (i.e., academic and general, vocational and general, and academic and vocational).

How the response variable is dichotomized depends, in part, on the nature of the variable. If there is a baseline or control category, then the analysis could focus on comparing each of the other categories to the baseline. With three or more categories, whether the response variable is nominal or ordinal is an important consideration. Since models for nominal responses can be applied to both nominal and ordinal response variables, the emphasis in this chapter

is on extensions of binary logistic regression to models designed for nominal response variables. Furthermore, a solid understanding of the models for nominal responses facilitates mastering models for ordinal data. A brief overview of models for ordinal variables is given toward the end of the chapter.

A second modification to extend binary logistic regression to the polytomous case is the need for a more complex distribution for the response variable. In the binary case, the distribution of the response is assumed to be binomial; however, with multicategory responses, the natural choice is the multinomial distribution, a special case of which is the binomial distribution. The parameters of the multinomial distribution are the probabilities of the categories of the response variable.

The *baseline logit model,* which is sometimes also called the *generalized logit model,* is the starting point for this chapter because it is a well-known model, it is a direct extension of binary logistic regression, it can be used with ordinal response variables, and it includes explanatory variables that are attributes of individuals, which is common in the social sciences. The baseline model is a special case of the *conditional multinomial logit model,* which can include explanatory variables that are characteristics of the response categories, as well as attributes of individuals.

A word of caution is warranted here. In the literature, the term *multinomial logit model* sometimes refers to the baseline model, and sometimes it refers to the conditional multinomial logit model. An additional potential source of confusion lies in the fact that the baseline model is a special case of the conditional model, which in turn is a special case of Poisson (log-linear) regression.[1] These connections enable researchers to tailor models in useful ways and test interesting hypotheses that could not otherwise be tested.

Multinomial Regression Models

One Explanatory Variable Model

The most natural interpretation of logistic regression models is in terms of odds and odds ratios; therefore, the baseline model is first presented as a model for odds and then presented as a model for probabilities.

Odds

The baseline model can be viewed as the set of binary logistic regression models fit simultaneously to all pairs of response categories. With three or more categories, a binary logistic regression model is needed for each (nonredundant) dichotomy of the categories of the response variable. As an example, consider high school program types from the High School and Beyond data set (Tatsuoka & Lohnes, 1988). There are three possible program types: academic, general, and vocational. Let $P(Y_i = \text{academic})$, $P(Y_i = \text{general})$, and $P(Y_i = \text{vocational})$ be the probabilities of each of the program types for individual *i.* Recall from Chapters 24 and 25 that odds equal ratios of probabilities. In our example, only two of the three possible pairs of program types are needed because the third can be found by taking the product of the other two. Choosing the general program as the reference, the odds of academic versus general and the odds of vocational versus general equal

$$\frac{P(Y_i = \text{academic})}{P(Y_i = \text{general})} \tag{1a}$$

$$\frac{P(Y_i = \text{vocational})}{P(Y_i = \text{general})}. \tag{1b}$$

The third odds, academic versus vocational, equals the product of the two odds in (1a) and (1b)—namely,

$$\frac{P(Y_i = \text{academic})}{P(Y_i = \text{vocational})} = \frac{P(Y_i = \text{academic})/P(Y_i = \text{general})}{P(Y_i = \text{vocational})/P(Y_i = \text{general})}. \tag{2}$$

More generally, let J equal the number of categories or levels of the response variable. Of the $J(J-1)/2$ possible pairs of categories, only $(J-1)$ of them are needed. If the same category is used in the denominator of the $(J-1)$ odds, then the set of odds will be nonredundant, and all other possible odds can be formed from this set. In the baseline model, one response category is chosen as the *baseline* against which all other response categories are compared. When a natural baseline exists, that category is the best choice in terms of convenience of interpretation. If there is not a natural baseline, then the choice is arbitrary.

As a Model for Odds

Continuing our example from the High School and Beyond data, where the general program is chosen as the baseline category, the first model contains a single explanatory variable, the mean of five achievement test scores for each student (i.e., math, science, reading, writing, and civics). The baseline model is simply two binary logistic regression models applied to each pair of program types; that is,

$$\frac{P(Y_i = \text{academic}|x_i)}{P(Y_i = \text{general}|x_i)} = \exp[\alpha_1 + \beta_1 x_i] \quad (3)$$

and

$$\frac{P(Y_i = \text{vocational}|x_i)}{P(Y_i = \text{general}|x_i)} = \exp[\alpha_2 + \beta_2 x_i], \quad (4)$$

where $P(Y_i = \text{academic}|x_i)$, $P(Y_i = \text{general}|x_i)$, and $P(Y_i = \text{vocational}|x_i)$ are the probabilities for each program type given mean achievement test score x_i for student i, the α_js are intercepts, and the β_js are regression coefficients. The odds of academic versus vocational are found by taking the ratio of (3) and (4),

$$\begin{aligned}\frac{P(Y_j = \text{academic}|x_i)}{P(Y_j = \text{vocational}|x_i)} &= \frac{\exp[\alpha_1 + \beta_1 x_i]}{\exp[\alpha_2 + \beta_2 x_i]} \\ &= \exp[(\alpha_1 - \alpha_2) + \\ &\quad (\beta_1 - \beta_2)x_i] \\ &= \exp[\alpha_3 + \beta_3 x_i],\end{aligned} \quad (5)$$

where $\alpha_3 = (\alpha_1 - \alpha_2)$ and $\beta_3 = (\beta_1 - \beta_2)$.

For generality, let $j = 1, \ldots, J$ represent categories of the response variable. The numerical values of j are just labels for the categories of the response variable. The probability that individual i is in category j given a value of x_i on the explanatory variable is represented by $P(Y_i = j|x_i)$. Taking the Jth category as the baseline, the model is

$$\frac{P(Y_i = j|x_i)}{P(Y_i = J|x_i)} = \exp[\alpha_j + \beta_j x_i] \quad \text{for } j = 1, \ldots, (J-1). \quad (6)$$

When fitting the baseline model to data, the binary logistic regressions for the $(J - 1)$ odds must be estimated simultaneously to ensure that intercepts and coefficients for all other odds equal the differences of the corresponding intercepts and coefficients (e.g., $\alpha_3 = (\alpha_1 - \alpha_2)$ and $\beta_3 = (\beta_1 - \beta_2)$ in Equation 5). To demonstrate this, three separate binary logistic regression models were fit to the High School and Beyond data, as well as the baseline regression model, which simultaneously estimates the models for all the odds. The estimated parameters and their standard errors are reported in Table 26.1. Although the parameters for the separate and simultaneous cases are quite similar, the logical relationships between the parameters when the models are fit separately are not met (e.g., $\hat{\beta}_1 - \hat{\beta}_2 = 0.1133 + 0.0163 = 0.1746 \neq 0.1618$); however, the relationships hold for simultaneous estimation (e.g., $\hat{\beta}_1 - \hat{\beta}_2 = 0.1099 + 0.0599 = 0.1698$).

Besides ensuring that the logical relationships between parameters are met, a second advantage of simultaneous estimation is that it is a more efficient use of the data, which in turn leads to more powerful statistical hypothesis tests and more precise estimates of parameters. Notice that the parameter estimates in Table 26.1 from the baseline model have smaller standard errors than those in the estimation of separate regressions. When the model is fit simultaneously, all 600 observations go into the estimation of the parameters; however, in the separately fit models, only a subset of the observations is used to estimate the parameters (e.g., 453 for academic and general, 455 for academic and vocational, and only 292 for vocational and general).

A third advantage of the simultaneous estimation, which is illustrated later in this chapter, is the ability to place equality restrictions on parameters across odds. For example, if two βs are very similar, they could be forced to be equal. The complexity of the baseline model increases as the number of response options increases, and any means of reducing the number of parameters that must be interpreted can be a great savings in terms of interpreting and summarizing the results. For example, if we modeled career choice with 15 possible choices, there would be 14 nonredundant odds and 14 different βs to interpret for

Table 26.1 Estimated Parameters (and Standard Errors) From Separate Binary Logistic Regressions and From the Simultaneously Estimated Baseline Model

		Separate Models		*Baseline Model*	
Odds	*Parameter*	*Estimate*	*SE*	*Estimate*	*SE*
$\frac{P(Y_i = \text{academic}\|x_i)}{P(Y_i = \text{general}\|x_i)}$	α_1	–5.2159	0.8139	–5.0391	0.7835
	β_1	0.1133	0.0156	0.1099	0.0150
$\frac{P(Y_i = \text{vocational}\|x_i)}{P(Y_i = \text{general}\|x_i)}$	α_2	2.9651	0.8342	2.8996	0.8156
	β_2	–0.0613	0.0172	–0.0599	0.0168
$\frac{P(Y_i = \text{academic}\|x_i)}{P(Y_i = \text{vocational}\|x_i)}$	α_3	–7.5331	0.8572	–7.9387	0.8439
	β_3	0.1618	0.0170	0.1698	0.0168

each (numerical) explanatory variable in the model.

Turning to interpretation, the regression coefficients provide estimates of odds ratios. Using the parameter estimates of the baseline model (column 5 of Table 26.1), the estimated odds that a student is from an academic program versus a general program given achievement score x equals

$$\frac{\hat{P}(Y_i = \text{academic}|x)}{\hat{P}(Y_i = \text{general}|x)} = \exp[-5.0391 + 0.1099x], \quad (7)$$

and the estimated odds of an academic versus a general program for a student with achievement score $x + 1$ equals

$$\frac{\hat{P}(Y_i = \text{academic}|x+1)}{\hat{P}(Y_i = \text{general}|x+1)} = \exp[-5.0391 + 0.1099(x+1)]. \quad (8)$$

The ratio of the two odds in Equations 8 and 7 is an odds ratio, which equals

$$\frac{\hat{P}(Y_i = \text{academic}|x+1)\,\hat{P}(Y_i = \text{academic}|x)}{\hat{P}(Y_i = \text{general}|x+1)\;\hat{P}(Y_i = \text{general}|x)} = \frac{\exp[-5.0391 + 0.1099(x+1)]}{\exp[-5.0391 + 0.1099x]} = \exp(0.1099) = 1.12.$$

This odds ratio is interpreted as follows: For a one-unit increase in achievement, the odds of a student attending an academic versus a general program are 1.12 times larger. For example, the odds of a student with $x = 50$ attending an academic program versus a general one is 1.12 times the odds for a student with $x = 49$. Given the scale of the achievement variable (i.e., $\bar{x} = 51.99$, $s = 8.09$, min = 32.94, and max = 70.00), it may be advantageous to report the odds ratio for an increase of one standard deviation of the explanatory variable rather than a one-unit increase. Generally speaking, $\exp(\beta c)$, where c is a constant, equals the odds ratio for an increase of c units. For example, for an increase of one standard deviation in mean achievement, the odds ratio for academic versus general equals $\exp(0.1099(8.09)) = 2.42$. Likewise, for a one standard deviation increase in achievement, the odds of an academic versus a vocational program are $\exp(0.1698(8.09)) = 3.95$ times larger, but the odds of a vocational program versus a general program are only $\exp(-0.0599(8.09)) = 0.62$ times as large.

Odds ratios convey the multiple one odds is relative to another odds. The parameters also provide information about the probabilities; however, the effects of explanatory variables on probabilities are not necessarily straightforward, as illustrated in the "Multiple Explanatory Variables" section.

As a Model of Probabilities

Probabilities are generally a more intuitively understood concept than odds and odds ratios. The baseline model can also be written as a model for probabilities. There is a one-to-one relationship between odds and probabilities. Using Equation 6, the model for probabilities is

$$P(Y_i = j|x_i) = \frac{\exp[\alpha_j + \beta_j x_i]}{\sum_{h=1}^{J} \exp[\alpha_h + \beta_h x_i]}, \quad (9)$$

where $j = 1, \ldots, J$. The sum in the numerator ensures that the sum of the probabilities over the response categories equals 1.

When estimating the model, identification constraints are required on the parameters. These constraints do not influence the goodness of model fit, odds ratios, estimated probabilities, interpretations, or conclusions. Identification constraints do affect the specific values of parameter estimates. The typical constraints are to set the parameter values of the baseline category equal to zero (e.g., $a_j = \beta_j = 0$) or to set the sum of the parameters equal to zero (e.g., $\sum_j a_j = \sum_j \beta_j = 0$). In practice, it is very important to know what constraints a computer program uses when writing the model as a model for probabilities. In our example, the software program (by default) set $a_J = \beta_J = 0$. Since $\exp(0) = 1$ for the general program, the estimated models for probabilities equal that shown in equation 10.

The estimated probabilities are plotted in Figure 26.1. The baseline model will always have one curve that monotonically decreases (e.g., $P(Y_i = \text{vocational}|x_i)$) and one that monotonically increases (e.g., $P(Y_i = \text{academic}|x_i)$). All others will increase and at some point start to decrease (e.g., $P(Y_i = \text{general}|x_i)$). At any point along the horizontal axis, the sum of the three probabilities equals 1.

Multiple Explanatory Variables

Multiple explanatory variables are typically available in most studies. Models with multiple explanatory variables are illustrated here by adding to our model a nominal (i.e., whether the school a student attends is public or private) and an ordinal variable (i.e., socioeconomic status reported as low, middle, or high).

Discrete variables are added using either dummy or effect coding. For example, school type could be coded either as a dummy variable (Equation 11a) or as an effect code (Equation 11b):

$$p_i = \begin{cases} 1 & \text{if public} \\ 0 & \text{if private} \end{cases} \quad (11a)$$

or

$$p_i = \begin{cases} 1 & \text{if public} \\ -1 & \text{if private} \end{cases} \quad (11b)$$

Most computer programs will automatically create the codes for discrete variables; however, proper interpretation requires that the user know how variables are coded.

The model presented and developed here has main effects for achievement, school type, and socioeconomic status (SES). Effect codes for school type, which are given in Equation 11b, are used to add school type to the model. The effects codes used to add SES, which has three levels, to the model are as follows:

$$\begin{aligned} \hat{P}(Y_i = \text{academic}) &= \frac{\exp[-5.0391 + 0.1099x_i]}{1 + \exp[-5.0391 + 0.1099x_i] + \exp[2.8996 - 0.0599x_i]} \\ \hat{P}(Y_i = \text{vocational}) &= \frac{\exp[2.8996 - 0.0599x_i]}{1 + \exp[-5.0391 + 0.1099x_i] + \exp[2.8996 - 0.0599x_i]} \\ \hat{P}(Y_i = \text{general}) &= \frac{1}{1 + \exp[-5.0391 + 0.1099x_i] + \exp[2.8996 - 0.0599x_i]} \end{aligned} \quad (10)$$

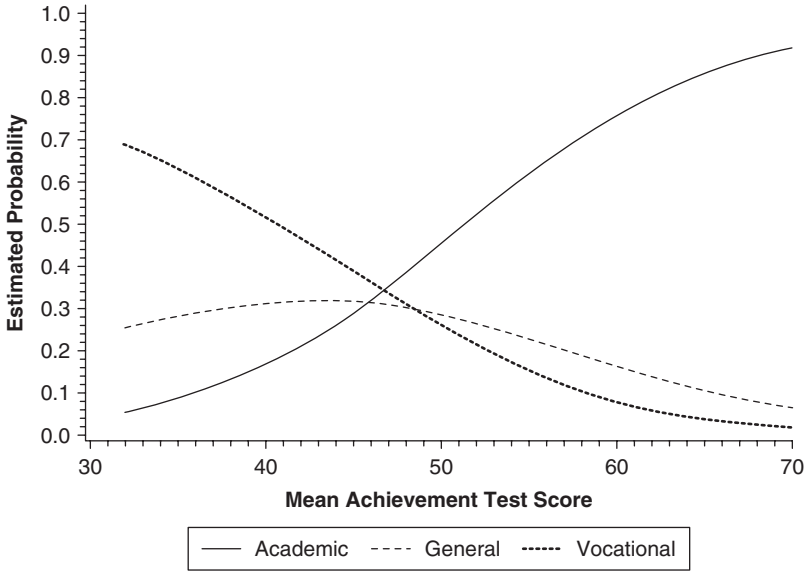

Figure 26.1 Estimated probabilities of attending different high school programs as a function of mean achievement.

$$s_{1i} = \begin{cases} 1 & \text{for low SES} \\ 0 & \text{for middle SES} \\ -1 & \text{for high SES} \end{cases} \tag{12a}$$

and

$$s_{2i} = \begin{cases} 0 & \text{for low SES} \\ 1 & \text{for middle SES} \\ -1 & \text{for high SES} \end{cases} \tag{12b}$$

Defining $j = 1$ for academic, $j = 2$ for vocational, and $j = 3 = J$ for general program, the first model with multiple explanatory variables examined here is

$$\frac{P(Y_i = j|x_i,\ p_i,\ s_{1i},\ s_{2i})}{P(Y_i = J|x_i,\ p_i,\ s_{1i},\ s_{2i})} = \exp[\alpha_j + \beta_{j1}x_i + \beta_{j2}p_i + \beta_{j3}s_{1i} + \beta_{j4}s_{2i}] \tag{13}$$

for $j = 1, \ldots, J - 1$. Additional regression coefficients (i.e., β_js) are estimated, one for each explanatory variable or code in the model. The same model expressed in terms of probabilities is

$$P(Y_i = j|x_i,\ p_i,\ s_{1i},\ s_{2i}) = \frac{\exp[\alpha_j + \beta_{j1}x_i + \beta_{j2}p_i + \beta_{j3}s_{1i} + \beta_{j4}s_{2i}]}{\sum_{h=1}^{J} \exp[\alpha_h + \beta_{h1}x_i + \beta_{h2}p_i + \beta_{h3}s_{1i} + \beta_{h4}s_{2i}]}. \tag{13}$$

One model has main effects of each of the variables. The estimated parameters and their standard errors are reported in Table 26.2.

The parameter estimates and statistics given in Table 26.2 have been rounded to two decimal places, which reflects a reasonable level of precision and is in line with American Psychological Association (APA) guidelines for reporting results. It should be noted that more decimal places were used in computations. For simplicity, in Table 26.2 and the remainder of the chapter, the explanatory variables are dropped from the symbols for the modeled probabilities; that is, $P(Y_i = j)$ will be used instead of $P(Y_i = j \mid x_i, p_i, s_{1i}, s_{2i})$. The values of the parameter estimates for high SES and private schools are included explicitly in the table to aid in the proper interpretation of the model.

The interpretation in terms of odds ratios is the same as binary logistic regression; however, the number of parameters to interpret is much larger than in binary logistic regression. Using the parameters reported in Table 26.2, for a one-unit increase in mean achievement, the odds of an academic versus a general program are 1.10 times larger, and for a one standard deviation increase, the odds are $\exp(0.10(0.809)) = 2.25$ times larger. The odds of academic versus general programs are larger for higher levels of

Table 26.2 Estimated Parameters, Standard Errors, and Wald Test Statistics for All Main Effects Model

Odds	*Effect*	*Parameter*	*Estimate*	*SE*	*exp(β)*	*Wald*	*p Value*
$\frac{P(Y_i = \text{academic})}{P(Y_i = \text{general})}$	Intercept	α_1	–3.92	0.83		22.06	< .01
	Achievement	β_{11}	0.10	0.02	1.10	37.80	< .01
	School type (public)	β_{12}	–0.61	0.18	0.54	12.01	< .01
	School type (private)	$-\beta_{12}$	0.61		1.84		
	SES (low)	β_{13}	–0.46	0.18	0.63	6.83	.01
	SES (middle)	β_{14}	–0.07	0.15	0.94	0.19	.66
	SES (high)	$-(\beta_{13} + \beta_{14})$	0.53		1.70		
$\frac{P(Y_i = \text{vocational})}{P(Y_i = \text{general})}$	Intercept	α_2	2.88	0.88		10.61	< .01
	Achievement	β_{13}	–0.06	0.02	0.94	13.28	< .01
	School type (public)	β_{22}	0.13	0.24	1.94	0.27	.60
	School type (private)	$-\beta_{22}$	–0.13		0.88		
	SES (low)	β_{23}	–0.23	0.19	0.80	1.45	.23
	SES (middle)	β_{24}	0.24	0.17	1.28	2.16	.14
	SES (high)	$-(\beta_{23} + \beta_{24})$	–0.02		0.98		

NOTE: SES is treated as a nominal variable.

achievement, private schools, and higher SES levels. It appears that the odds of vocational versus general programs are larger for public schools and lower SES levels; however, this conclusion is not warranted. These parameters are not significantly different from zero (see the last two columns of Table 26.2). Statistical inference and nonsignificant parameters are issues that are returned to later in this chapter.

To illustrate the effect of the extra variables on probabilities, the estimated probabilities of attending an academic program are plotted against mean achievement scores in Figure 26.2 with a separate curve for each combination of SES. The figure on the left is for students at public schools, and the figure on the right is for private schools. The discrete variables shift the curves horizontally. In particular, the horizontal distance between the curves for high and middle SES for either public or private schools is the same regardless of the mean achievement value.

When there is no interaction, the curves for the estimated probabilities of the response category that is monotonically increasing (e.g., academic programs) will be parallel; that is, they will have the same shape and are just shifted horizontally. The curves for the response category where the probabilities monotonically decrease (e.g., vocational programs) will also be parallel; however, this is not true for the other categories. In fact, the curves for responses whose probabilities increase and then decrease are not parallel and may even cross within the range for which there are data. To illustrate this, the probabilities of students attending a general program for private schools are plotted against mean achievement scores with a separate curve for each SES level at the top of Figure 26.3. The curves for different SES levels cross. In normal linear regression, this would indicate an interaction between achievement and SES; however, this is not the case in logistic regression. Crossing curves can occur in the baseline model with only main effects because the relative size of the numerator and denominator changes as values of the explanatory variables change. If there are no interactions in the model, then curves of odds ratios and logarithms of odds ratios will not cross. This is illustrated at the bottom of Figure 26.3, where the logarithms of the odds are plotted against mean achievement. The logarithms of the odds are linear, and the lines are parallel. Although probabilities are more intuitively understandable, figures of estimated probabilities may be misleading to researchers unfamiliar with multicategory logistic regression models. Care must be taken when presenting results such as these.

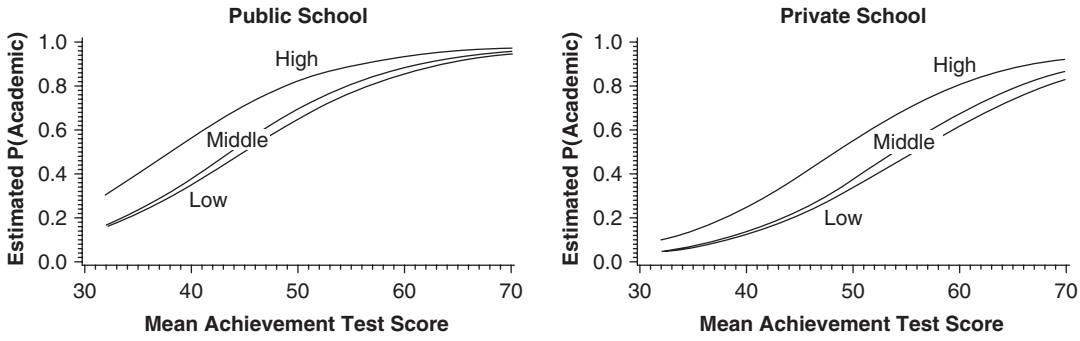

Figure 26.2 Estimated probabilities of attending an academic high school program as a function of mean achievement with a separate curve for each SES level.

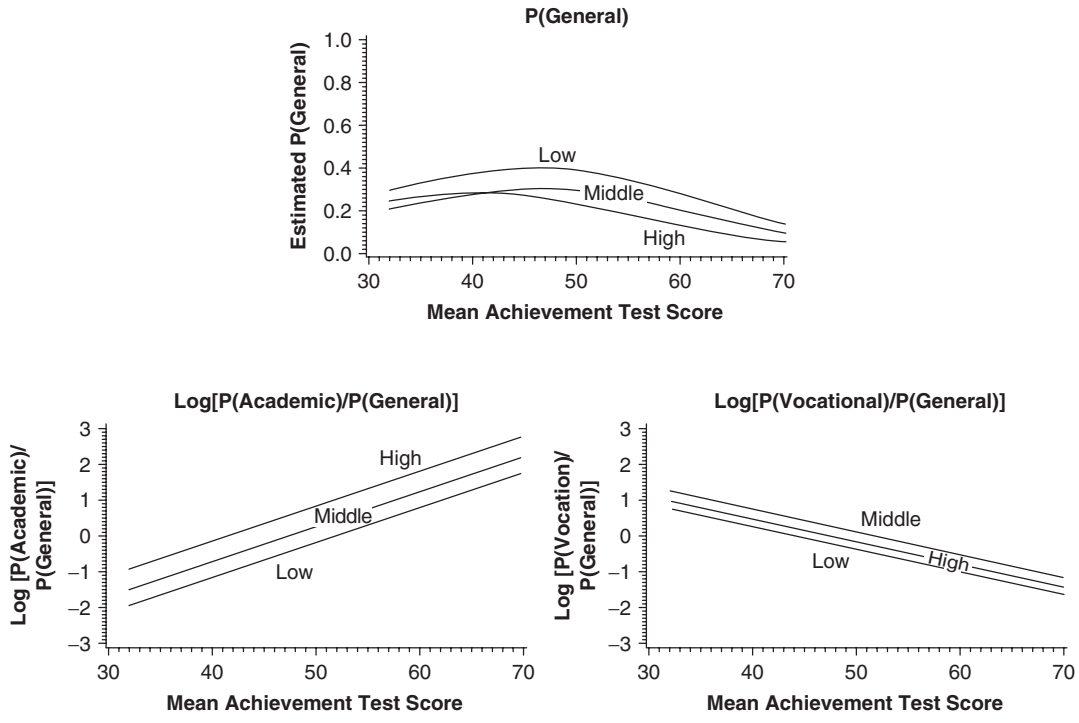

Figure 26.3 At the top is a plot of estimated probabilities of attending a general high school program for private schools as a function of mean achievement with a separate curve for each SES level from the model with only main effects. The two lower figures are plots of the estimated logarithm of odds ratios using parameters from the same model.

With ordinal explanatory variables such as SES, one way to use the ordinal information is by assigning scores or numbers to the categories and treating the variables as numerical variables in the model (e.g., like mean achievement). Often, equally spaced integers are used, which amounts to putting equality restrictions on the βs for the variable. In our example, suppose we assign 1 to low SES, 2 to middle SES, and 3 to high SES and refit the model. In this model, only one β parameter is estimated for SES rather than two for each of the $(J - 1)$ odds. Using SES as a numerical variable with equally spaced scores in the model imposes restrictions

on the βs. In our example, $\beta_{j3} = \beta_{j4}$ for the odds for $j = 1$ and 2 (i.e., academic and vocational, respectively). The parameter estimates for the model where SES is treated numerically are reported in Table 26.3.

Placing the restrictions on the βs for ordinal variables is often a good way to reduce the complexity of a model. For example, the estimated odds ratio of academic versus general for middle versus low SES equals $\exp(\hat{\beta}_{13}(2-1)) = \exp(\hat{\beta}_{13}) = 1.70$, which is the same as the odds ratio of high versus middle SES, $\exp(\hat{\beta}_{13}(3-2)) = 1.70$.

In our example, putting in equally spaced scores for SES is not warranted and is misleading. When no restrictions were imposed on the parameters (see Figure 26.3 and Table 26.2), the order of the SES levels for the odds of academic (versus general) schools is in the expected order (i.e., the odds of an academic program are larger the higher the student's SES level), and the parameter estimates are approximately equally spaced. On the other hand, the parameter estimates of SES for odds of vocational schools do not follow the natural ordering of low to high, are relatively close together, and are not significantly different from zero. The numerical scores could be used for the SES effect on the odds of academic programs but the scores are inappropriate for the odds of vocational programs. There may not even be a difference between vocational and general programs in terms of SES. Furthermore, there may not be a difference between students who attended vocational and general programs with respect to school type (Wald = 0.27, $df = 1$, $p = .60$). In the following section, these conjectures are incorporated into the model, which permits statistical testing of these hypotheses.

Conditional Multinomial Logistic Regression

The conjectures described above regarding possible equalities, nonsignificant effects, and restrictions on parameters can be imposed (and tested) by reexpressing the baseline model as a conditional multinomial logistic regression model, which is a more general model. The conditional multinomial logistic regression model is also known as the "discrete choice model," "McFadden's model," and "Luce's choice model." Some sources for more complete introductions to this model include Long (1997), Agresti (2002), and Powers and Xie (2000). Unlike the baseline model, the conditional model permits explanatory variables that are attributes of response categories.

The general form of the conditional multinomial logistic model is

$$P(Y_i = j|\mathbf{x}^*_{ij}) = \frac{\exp(\boldsymbol{\beta}^{*\prime}\mathbf{x}^*_{ij})}{\sum_h \exp(\boldsymbol{\beta}^{*\prime}\mathbf{x}^*_{ih})}, \qquad (15)$$

where $\boldsymbol{\beta}^\star$ is a vector of coefficients and $\mathbf{x}^\star_{ij}$ is a vector of explanatory variables. The explanatory variables $\mathbf{x}^\star_{ij}$ may depend on the attributes of an individual (i), the response category (j), or both, but the coefficients $\boldsymbol{\beta}^\star$ do not depend on the response categories. To reexpress the baseline

Table 26.3 Estimated Parameters, Standard Errors, and Wald Statistics for All Main Effects Model

Odds	*Effect*	*Parameter*	*Estimate*	*SE*	*exp(β)*	*Wald*	*p Value*
$\frac{P(Y_i = \text{academic})}{P(Y_i = \text{general})}$	Intercept	α_1	–4.97	0.83	—	35.73	< .01
	Achievement	β_{11}	0.10	0.02	1.10	37.48	< .01
	School type	β_{12}	–0.61	0.18	0.55	11.80	< .01
	SES	β_{13}	0.53	0.18	1.70	11.80	< .01
$\frac{P(Y_i = \text{vocational})}{P(Y_i = \text{general})}$	Intercept	α_2	2.57	0.87	—	8.78	< .01
	Achievement	β_{13}	–0.06	0.02	0.95	12.96	< .01
	School type	β_{22}	0.12	0.24	1.13	0.26	.61
	SES	β_{23}	0.17	0.19	1.19	0.92	.34

NOTE: SES is treated as a numerical variable with scores of 1 = low, 2 = middle, and 3 = high.

model in the general form of the conditional logistic model given in Equation 15, data need to be reformatted from having one line in the data file for each individual to multiple lines of data for each individual. When fitting conditional multinomial models, the data must have one line of data for each response category for each individual to incorporate explanatory variables that are characteristics of the response categories.

The format of the data file for the baseline model where mean achievement is the only explanatory variable in the model is given in Table 26.4. In the data, Y is the response variable that indicates a student's high school program. The indicators, d_{ij}, are the key to putting the baseline model into the form of the conditional multinomial model, as well as to specifying different models for the different program types (i.e., levels of the response variable). In our example, the two indicator variables, d_{i1} and d_{i2}, indicate the category of the response variable corresponding to particular line in the data file. They are defined as

$$d_{i1} = \begin{cases} 1 & \text{when program type is academic} \\ 0 & \text{otherwise} \end{cases}$$

$$d_{i2} = \begin{cases} 1 & \text{when program type is vocational} \\ 0 & \text{otherwise} \end{cases}$$

For general programs, $d_{i1} = d_{i2} = 0$.

For our simple baseline model with only achievement test scores in the model, $\beta^{\star\prime} = (\alpha_1, \alpha_2, \beta_{11}, \beta_{21})$ and $x^{\star\prime}_{ij} = (d_{i1}, d_{i2}, d_{i1}x_i, d_{i2}x_i)$. Using the definitions of d_{ij}, $\beta^{\star}$, and $x^{\star}_{ij}$, we obtain our familiar form of the baseline model,

$$P(Y_i = j) = \frac{\exp[\alpha_j + \beta_{j1}x_i]}{(1 + \exp[\alpha_1 + \beta_{11}x_i] + \exp[\alpha_2 + \beta_{21}x_i])}, \quad (16)$$

which, for our example, corresponds to

$$P(Y_i = \text{academic}) = \frac{\exp[\alpha_1 + \beta_{11}x_i]}{(1 + \exp[\alpha_1 + \beta_{11}x_i] + \exp[\alpha_2 + \beta_{21}x_i])}, \quad (17)$$

$$P(Y_i = \text{vocational}) = \frac{\exp[\alpha_2 + \beta_{21}x_i]}{(1 + \exp[\alpha_1 + \beta_{11}x_i] + \exp[\alpha_2 + \beta_{21}x_i])}, \quad \text{and} \quad (18)$$

Table 26.4 The Format of the Data File Needed to Fit the Baseline Multinomial Model as Conditional Multinomial Model With Mean Achievement as the Explanatory Variable

Student ID	*Achievement* x_i	*Program Type*	Y	d_{i1}	d_{i2}	$d_{i1}x_i$	$d_{i2}x_i$
1	32.94	General	0	0	0	0	0
1	32.94	Academic	0	1	0	32.94	0
1	32.94	Vocational	1	0	1	0	32.94
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
102	43.74	General	1	0	0	0	0
102	43.74	Academic	0	1	0	43.74	0
102	43.74	Vocational	0	0	1	0	43.74
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
600	70.00	General	0	0	0	0	0
600	70.00	Academic	1	1	0	70.00	0
600	70.00	Vocational	0	0	1	0	70.00

$$P(Y_i = \text{general}) = \frac{1}{(1 + \exp[\alpha_1 + \beta_{11}x_i] + \exp[\alpha_2 + \beta_{21}x_i])}. \quad (19)$$

Note that the parameters for the last category, which in this case is general programs, were set equal to zero for identification (i.e., $\alpha_{31} = 0$ for the intercept, and $\beta_{31} = 0$ for achievement test scores).

The conditional model was introduced here as a means to refine and reduce the number of parameters of the baseline model. First, the complex model will be represented as a conditional multinomial model, and then, by appropriately defining $x^{\star}_{ij}$, restrictions will be imposed on the parameters. To express the complex baseline model as a conditional multinomial logistic model, we define $\beta^{\star}$ and $x^{\star}_{ij}$ as

$$\boldsymbol{\beta}^{\star\prime} = (\alpha_1, \alpha_2, \beta_{11}, \beta_{21}, \beta_{12}, \beta_{22}, \beta_{13}, \beta_{23}, \beta_{14}, \beta_{24})$$

and

$$\mathbf{x}^{\star\prime}_{ij} = (d_{i1}, d_{i2}, d_{i1}x_i, d_{i2}x_i, d_{i1}p_i, d_{i2}p_i, d_{i1}s_{1i}, d_{i2}s_{1i}, d_{i1}s_{2i}, d_{i2}s_{2i}).$$

Using these $\beta^{\star}$ and $x^{\star}_{ij}$ in Equation 15, the complex baseline model with main effects for achievement (x_i), school type (p_i), and SES (s_{1i}, s_{2i}) expressed as a conditional multinomial model is

$$P(Y_i = j) = \exp[\alpha_j d_{ij} + \beta_{j1} d_{ij} x_i + \beta_{j2} d_{ij} p_i + \beta_{j3} d_{ij} s_{1i} + \beta_{j4} d_{ij} s_{2i}] k_i, \quad (20)$$

$$\text{where } k_i = \sum\nolimits_h \exp[\alpha_h d_{ih} + \beta_{h1} d_{ih} x_i + \beta_{j2} d_{ih} p_i + \beta_{h3} d_{ih} s_{1i} + \beta_{h4} d_{ih} s_{2i}].$$

When reexpressing the baseline model as a conditional multinomial model, a good practice is to refit the baseline models that have already been fit to the data as conditional models. If the data matrix and conditional models have been correctly specified, then the results obtained from the baseline and conditional models will be the same (i.e., parameters estimates, standard errors, etc.).

One simplification of Equation 20 is to treat SES numerically for academic programs (i.e., use $s_i = 1, 2, 3$ instead of s_{1i} and s_{2i}). The parameters for SES and school type for vocational and general programs can be set to zero by deleting the terms $d_{i2}p_i$, $d_{i2}s_{1i}$, and $d_{i2}s_{2i}$ from the data matrix, which implicitly sets $\beta_{22} = \beta_{23} = \beta_{24} = 0$. Making these changes, the models for the probabilities for each high school program type become

$$P(Y_i = \text{academic}) = \frac{\exp[\alpha_1 + \beta_{11}x_i + \beta_{12}p_i + \beta_{13}s_i]}{1 + \exp[\alpha_1 + \beta_{11}x_i + \beta_{12}p_i + \beta_{13}s_i] + \exp[\alpha_2 + \beta_{21}x_i]},$$ and

$$P(Y_i = \text{vocational}) = \frac{\exp[\alpha_2 + \beta_{21}x_i]}{1 + \exp[\alpha_1 + \beta_{11}x_i + \beta_{12}p_i + \beta_{13}s_i] + \exp[\alpha_2 + \beta_{21}x_i]},$$

$$P(Y_i = \text{general}) = \frac{1}{1 + \exp[\alpha_1 + \beta_{11}x_i + \beta_{12}p_i + \beta_{13}s_i] + \exp[\alpha_2 + \beta_{21}x_i]}.$$

The parameter estimates for this final model are given in Table 26.5 and are similar to those in the baseline model with no restrictions on the parameters (i.e., Table 26.2). The model with restrictions fit as a conditional multinomial model is more parsimonious than the baseline model (i.e., 6 vs. 10 nonredundant parameters). The models for the probabilities of different high school programs do not have the same effects. As stated earlier, by fitting all of the odds simultaneously, the logical restrictions on the parameters are maintained, and the standard errors of the parameters are smaller (i.e., there is greater precision).

Switching to the conditional multinomial model emphasizes that users of multinomial logistic regression are not restricted to the standard models that computer programs fit by default, which often fail to address specific research questions. By creating new variables, using indicator variables, and using a slightly more general model, researchers can tailor their models to best match their research questions. Imposing restrictions on parameters when they are warranted can greatly simplify a complex model. This was illustrated in our example. Compare Tables 26.2 and 26.5. In the model reported in Table 26.2, there were 10

Table 26.5 Estimated Parameters From the Conditional Multinomial Logistic Regression Model

Odds	*Effect*	*Parameter*	*Estimate*	*SE*	*exp(β)*	*Wald*	*p Value*
$\frac{P(Y_i = \text{academic})}{P(Y_i = \text{general})}$	Intercept	β_{11}	–4.87	0.82		35.31	< .01
	Achievement	β_{12}	0.10	0.02	1.10	39.81	< .01
	School type (public)	β_{13}	–0.66	0.15	0.52	20.61	< .01
	School type (private)	$-\beta_{13}$	0.66		1.93		
	SES (low)	β_{14}	0.46	0.14	1.58	10.22	< .01
$\frac{P(Y_i = \text{vocational})}{P(Y_i = \text{general})}$	Intercept	β_{21}	2.83	0.81		12.33	< .01
	Achievement	β_{22}	–0.06	0.02	0.94	1.42	< .01

nonredundant parameters, some of which are not significant and others that are. An additional reason for going the extra step of using the conditional model is that novice users often succumb to the temptation of interpreting nonsignificant parameters such as those in Table 26.2. The better approach is to avoid the temptation. Alternatively, researchers may simply not report the nonsignificant effects even though they were in the model. This also is a poor (and misleading) practice. Table 26.5 contains only 6 nonredundant parameters, all of which are significant. What is statistically important stands out, and the interpretation is much simpler. In our example, it is readily apparent that students in general and vocational programs only differ with respect to achievement test scores, whereas students in academic programs and those in one of the other programs differ with respect to achievement test scores, SES, and school type.

Statistical Inference

Simple statistical tests were used informally in the previous section to examine the significance of effects in the models discussed. The topic of statistical inference is explicitly taken up in detail in this section. There are two basic types of statistical inference that are of interest: the effect of explanatory variables and whether response categories are indistinguishable with respect to the explanatory variables.

Tests of Effects

If $\beta = 0$, then an effect is unrelated to the response variable, given all the other effects in the model. The two test statistics discussed here for assessing whether $\beta = 0$ are the Wald statistic and the likelihood ratio statistic.

Maximum likelihood estimates of the model parameters are asymptotically normal with mean equal to the value of the parameter in the population—that is, the sampling distribution of $\hat{\beta} \sim N(\beta, \sigma^2_{\hat{\beta}})$. This fact can be used to test the hypothesis $\beta = 0$ and to form confidence intervals for β and odds ratios.

If the null hypothesis $H_0: \beta = 0$ is true, then the statistic

$$z = \frac{\hat{\beta}}{SE},$$

where *SE* is the estimate of $\sigma^2_{\hat{\beta}}$, has an approximate standard normal distribution. This statistic can be used for directional or nondirectional tests. Often, the statistic is squared, $X^2 = z^2$, and is known as a Wald statistic, which has a sampling distribution that is approximately chi-square with 1 degree of freedom.

When reporting results in papers, many journals require confidence intervals for effects. Although many statistical software packages automatically provide confidence intervals, we show where these intervals come from, which points to the relationship between the confidence intervals and the Wald tests. A $(1 - \alpha)\%$ Wald confidence interval for β is

$$\hat{\beta} \pm z_{(1-\alpha)/2}(SE),$$

where $z_{(1-\alpha)/2}$ is the $(1 - \alpha)/2$th percentile of the standard normal distribution. A $(1 - \alpha)\%$ confidence interval for the odds ratio is found by exponentiating the end points of the interval for β.

As an example, consider the parameter estimates given in Table 26.5 of our final model from

the previous section. The Wald statistics and corresponding p values are reported in the seventh and eighth columns. For example, the Wald statistic for achievement for academic programs equals

$$X^2 = \left(\frac{0.0978}{0.0155}\right)^2 = 39.81.$$

The 95% confidence interval of β for achievement equals

$$0.0987 \pm 1.96(0.0155) \rightarrow (0.06742, 0.12818),$$

and the 95% confidence interval for the odds ratio equals

$$(\exp(0.06742), \exp(0.12818)) \rightarrow (1.07, 1.14).$$

The ratio of a parameter estimate to its standard errors provides a test for single parameters; however, Wald statistics can be used for testing whether any and multiple linear combinations of the βs equal zero (Agresti, 2002; Long, 1997). Such Wald statistics can be used to test simultaneously whether an explanatory variable or variables have an effect on any of the models for the odds. Rather than present the general formula for Wald tests (Agresti, 2002; Long, 1997), the likelihood ratio statistic provides a more powerful alternative to test more complex hypotheses.

The likelihood ratio test statistic compares the maximum of the likelihood functions of two models, a complex and a simpler one. The simpler model must be a special case of the complex model where restrictions have been placed on the parameters of the more complex model. Let $\ln(L(M_0))$ and $\ln(L(M_1))$ equal the natural logarithms of the maximum of the likelihood functions for the simple and the complex models, respectively. The likelihood ratio test statistic equals

$$G^2 = -2(\ln(L(M_0)) - \ln(L(M_1))).$$

Computer programs typically provide either the value of the maximum of the likelihood function, the logarithm of the maximum, or –2 times the logarithm of the maximum. If the null hypothesis is true, then G^2 has an approximate chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the two models.

The most common type of restriction on parameters is setting them equal to zero. For example, in the baseline model with all main effects (nominal SES), we could set all of the parameters for SES and school type equal to zero; that is, $\beta_{j2} = \beta_{j3} = \beta_{j4} = 0$ for $j = 1$ and 2. The complex model is

$$\frac{P(Y_i = j)}{P(Y_i = J)} = \exp[\alpha_j + \beta_{j1}x_i + \beta_{j2}p_i + \beta_{j3}s_{1i} + \beta_{j4}s_{2i}],$$

and the simpler model is

$$\frac{P(Y_i = j)}{P(Y_i = J)} = \exp[\alpha_j + \beta_{j1}x_i].$$

The maximum of the likelihoods of all the models estimated for this chapter are reported in Table 26.6. The difference between the likelihood for these two models equals

$$G^2 = -2(-541.8917 + 520.26080) = 43.26.$$

The degrees of freedom for this test equal 6, because 6 parameters have been set equal to zero. Comparing 43.26 to the chi-square distribution with 6 degrees of freedom, the hypothesis that SES and/or school type have significant effects is supported.

Equality restrictions on parameters can also be tested using the likelihood ratio test. Such tests include determining whether an ordinal explanatory variable can be treated numerically. For example, when SES was treated as a nominal variable in the baseline model, SES was represented in the model by $\beta_{j3}s_{1i} + \beta_{j4}s_{2i}$, where s_{1i} and s_{2i} were effect codes. When SES was treated as a numeric variable (i.e., $s_i = 1, 2, 3$ for low, middle, high), implicitly the restriction that $\beta_{j3} = \beta_{j4}$ was imposed, and SES was represented in the model by $\beta_{j3}s_i$. The likelihood ratio test statistic equals 3.99, with $df = 2$ and $p = .14$ (see Table 26.6). Even though SES should not be treated numerically for vocational and general programs, the test indicates otherwise. It is best to treat categorical explanatory variables nominally, examine the results, and, if warranted, test whether they can be used as numerical variables.

Tests Over Regressions

A second type of test that is of interest in multinomial logistic regression modeling is whether two or more of the response categories are indistinguishable in the sense that they have

Table 26.6 Summary of All Models Fit to the High School and Beyond Data Where Mean Achievement, School Type, and SES Are Explanatory Variables

			Likelihood Ratio Tests				
Model	*Number of Parameters*	*Ln(Likelihood)*	*Models Compared*	G^2	*df*	*p*	*AIC*
M_1: All main (nominal SES)	10	–520.2608	—				1061
M_2: Achievement only	4	–541.8917	M_2 & M_1	43.26	6	< .01	1092
M_3: All main (ordinal SES)	8	–522.2579	M_3 & M_1	3.99	2	.14	1061
M_4: General and vocational indistinguishable?	6	–528.3721	M_4 & M_1	16.22	4	< .01	1069
M_5: Restrictions on parameters	6	–522.8114	M_5 & M_1	5.10	4	.28	1058

the same parameter values. If there is no difference between two responses—say, j and j^*—then

$$(\beta_{j1} - \beta_{j^*1}) = \ldots = (\beta_{jK} - \beta_{j^*K}) = 0, \quad (21)$$

where K equals the number of explanatory variables. If two response levels are indistinguishable, they can be combined (Long, 1997). In our example, the parameter estimates for the two main effect models have nonsignificant odds for vocational versus general (see Tables 26.2 and 26.3). This suggests that vocational and general programs may be indistinguishable.

The indistinguishability hypothesis represented in Equation 21 can be tested in two ways. The simple method is to create a data set that only includes the two response variables, fit a binary logistic regression model to the data, and use a likelihood ratio test to assess whether the explanatory variables are significant (Long, 1997). In our example, the model in Equation 13 was fit to the subset of data that only includes students from general and vocational programs. According to the likelihood ratio test, the explanatory variables are significant ($G^2 = 17.53$, $df = 4$, $p < .01$).

The second and preferable way to test the indistinguishability hypothesis makes use of the fact that the baseline model is a special case of the conditional model. Besides using all of the data, this second method has the advantage that it can be used to simultaneously test whether three or more of the responses are indistinguishable or to place restrictions on subsets of parameters. To test the indistinguishability hypotheses for vocational and general programs relative to the baseline model with all main effects, we fit the conditional multinomial model with $\beta^{*\prime} = (\alpha_1, \alpha_2, \beta_{11}, \beta_{12}, \beta_{13}, \beta_{14})$ and $x_{ij}^{*\prime} = (d_{i1}, d_{i2}, d_{i1}x_i, d_{i1}p_i, d_{i1}s_{1i}, d_{i1}s_{2i})$; that is, the terms with d_{i2} indicators were all dropped from the model except for the intercept α_2. The reduced model is M_4 in Table 26.6, and $G^2 = 16.22$, $df = 4$, and $p < .01$. The null hypothesis is again rejected, and the general and vocational programs are distinguishable on at least one of the three explanatory variables.

Model Assessment

Before drawing conclusions, the adequacy of the model or subset of models must be assessed. Goodness of fit, model comparisons, and regression diagnostics are discussed below.

Goodness of Fit

Two typical tests of goodness of fit are Pearson's chi-square statistic and the likelihood ratio chi-square statistic; however, for tests of goodness of fit to be valid, these statistics should have approximate chi-square distributions. For the sampling distribution of the goodness-of-fit statistics to be chi-square requires two conditions (Agresti, 2002, 2007). First, most fitted values of "cells" of the cross-classification of the response variable by all of the explanatory variables should be greater than 5. Second, as more observations are added to the data set, the size of the cross-classification does not increase. In other words, as observations are added, the number of observations per cell gets larger.

The High School and Beyond data set, even with 600 students, fails both of the requirements. Consider the model with only achievement. The cross-classification of students by high school program type and achievement is given on the right side of Table 26.7. There are 545 unique values of the mean achievement scores, so most cells equal 0. If a new student is added to the data set, it is possible that his or her mean achievement level will be different from the 545 levels already in the data set, and adding a new student would add another row to Table 26.7. Even if only school type and SES are included in the model, the closeness of the approximation of the sampling distribution of goodness-of-fit statistics is uncertain. The cross-classification of programs by SES by school type is given on the left side of Table 26.7, and 5 of the 18 (28%) of the cells are less than 5. The sampling distributions of Pearson's chi-square and the likelihood ratio statistics may not be close to chi-square. The solution is not to throw away information by collapsing data. Alternatives exist.

In the case of binary logistic regression, the Hosmer-Lemeshow statistic is often used to assess model goodness of fit (Hosmer & Lemeshow, 2000); however, no such statistic is readily computed for the multinomial case. Goodness-of-fit statistics for large, sparse contingency tables, including multinomial logistic models, is an active area of research (e.g., Maydeu-Olivares & Joe, 2005). Until more suitable procedures become available in standard statistical packages, one suggestion is to perform dichotomous logistic regressions and compute the Hosmer-Lemeshow statistic for each of them (Hosmer & Lemeshow, 2000).

The Hosmer-Lemeshow statistic is basically Pearson's chi-square statistic,

$$X^2 = \sum_{\text{cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}.$$

The observed and expected values are frequencies found by ordering the predicted values from a binary logistic regression model from smallest to largest and then partitioning the cases into approximately equal groups. Both the data and the expected values from the regression are cross-classified into tables of group by response category. Even though the sampling distribution of the Hosmer-Lemeshow statistic is not chi-square, comparing the Hosmer-Lemeshow statistic to a chi-square distribution performs reasonably well.

A dichotomous logistic regression model for academic versus general was fit with achievement, school type, and SES as an ordinal variable and yielded a Hosmer-Lemeshow statistic = 8.84, $df = 8$, and $p = .36$. For the binary logistic regression model of vocational and general programs with only achievement as an explanatory variable, the model had a Hosmer-Lemeshow statistic = 9.00, $df = 8$, and $p = .34$. In both cases, the models appear adequate.

Model Comparisons

When a subset of models is available and a user wishes to select the "best" one to report and interpret, various measures and statistics are

Table 26.7 Cross-Classifications of the 600 Students in the High School and Beyond Data Set by School Type, SES, and High School Program (Left) and by Mean Achievement and High School Program Type (Right)

School Type	*SES*	*High School Program*			*Achievement*	*High School Program*		
		General	*Academic*	*Vocational*		*General*	*Academic*	*Vocational*
Public	Low	40	40	44	32.94	0	0	1
	Middle	63	111	75	33.74	0	0	1
	High	24	82	20	⋮	⋮	⋮	⋮
Private	Low	3	4	1	43.74	2	0	0
	Middle	7	36	7	⋮	⋮	⋮	⋮
	High	1	35	0	70.00	0	1	0

available for model comparisons. If the models are nested, then the conditional likelihood ratio tests discussed in the "Statistical Inference" section are possible. Although the goodness-of-fit statistics may not have good approximations by chi-square distributions, the conditional likelihood ratio tests often are well approximated by chi-square distributions. One strategy is to start with an overly complex model that gives an adequate representation of the data. Various effects could be considered for removal by performing conditional likelihood ratio tests following the procedure described earlier for testing effects in the model.

When the subset of models is not nested, information criteria can help to choose the "best" model in the set. Information criteria weight goodness of model fit and complexity. A common choice is Akaike's information criterion (AIC), which equals

–2(maximum log likelihood – number of parameters)

(Agresti, 2002). The model with the smallest value of AIC is the best. The AIC values for all the models fit in this chapter are reported in Table 26.6. The best model among those fit to the High School and Beyond data is model M_5, which is the conditional multinomial model with different explanatory variables for the program types.

When AIC statistics are very close (e.g., M_1 and M_3), then the choice between models must be based on other considerations such as interpretation, parsimony, and expectations. An additional consideration is whether statistically significant effects are significant in a practical sense (Agresti, 2007). Although many statistics can be reported, which make model selection appear to be an objective decision, model selection is in the end a subjective decision.

Regression Diagnostics

Before any model is reported, regression diagnostics should be performed. Lesaffre and Albert (1989) extended diagnostic procedures for dichotomous responses to the multicategory case of multinomial logistic regression. Unfortunately, these have not been implemented in standard statistical packages. One recommendation is to dichotomize the categories of the response, fit binary logistic regression models, and use the regression diagnostics that are available for binary logistic regression (Hosmer & Lemeshow, 2000).

Observations may have too much influence on the parameter estimates and the goodness of fit of the model to data. Influential observations tend to be those that are extreme in terms of their values on the explanatory variables. Most diagnostics for logistic regression are generalizations of those for normal linear regression (Pregibon, 1981; see also Agresti, 2002, 2007; Hosmer & Lemeshow, 2000), but there are some exceptions (Fahrmeir & Tutz, 2001; Fay, 2002). One exception is "range-of-influence" statistics. Rather than focusing on whether observations are outliers in the design or exert great influence on results, range-of-influence statistics are designed to check for possible misclassifications of a binary response (Fay, 2002). They are particularly useful if the correctness of classifications into the response categories is either very costly or impossible to check.

Problems With Multinomial Regression Models

Two common problems encountered in multinomial regression models when there are multiple explanatory variables are multicollinearity and "quasi" or "complete separation." As in normal multiple linear regression, multicollinearity occurs when explanatory variables are highly correlated. In such cases, the results can change drastically when an explanatory variable that is correlated with other explanatory variables is added to the model. Effects that were statistically significant may no longer be significant.

To illustrate multicollinearity, two-way interactions between all three main effects in our model were added (i.e., between mean achievement, SES as a numerical variable [1, 2, and 3], and school type). Table 26.8 contains the Wald chi-square test statistics for testing whether the βs for each effect equal zero, the degrees of freedom, and p value for each test. The third and fourth columns contain the results for the model with only main effects and show that the three effects are statistically significant. When all two-way interactions are added to the model, nothing is significant (i.e., the fifth and sixth columns). To reveal the culprit, correlations between

Table 26.8 Statistical Test for Effects When There Are Only Main Effects in the Model, All Two-Way Interactions (Highly Correlated Effects) and All Two-Way Interactions Where Mean Achievement and SES Are Standardized

		Main Effects		*Unstandardized*		*Standardized*	
Effect	*df*	*Wald*	*p Value*	*Wald*	*p Value*	*Wald*	*p Value*
Achieve	2	87.62	< .01	5.42	.07	18.06	< .01
SES	2	11.08	< .01	1.10	.58	11.44	< .01
School type	2	20.72	< .01	0.96	.62	16.72	< .01
Achieve × SES	2			0.29	.87	0.29	.87
Achieve × School type	2			0.93	.63	0.93	.63
SES × School type	2			4.07	.13	4.07	.13

achievement, SES, school type, and interactions containing them are reported in Table 26.9. Four of the nine correlations are greater than 0.70, and two of these are greater than 0.90.

Dealing with multicollinearity is the same as that for normal multiple linear regression. If correlations between main effects and interactions lead to multicollinearity, then the variables can be standardized. Table 26.9 also contains the correlations between the standardized explanatory variables, and these are all quite small. In our example, standardization solved the multicollinearity problem. As can be seen in the right side of Table 26.8, the main effects are significant when the explanatory variables are standardized.

A problem more unique to multinomial logistic regression modeling is "quasi-complete separation" or "complete separation of data points." Separation means that the pattern in the data is such that there is no overlap between two or more response categories for some pattern(s) of the values on the explanatory variables (Albert & Anderson, 1984). In other words, using the model to classify individuals into response levels performs perfectly for one or more of the response categories. In a very simple case, if only men attend vocational programs, then including gender in the model would lead to separation. Typically, the situation is not quite so simple when specific combinations among the explanatory variables occur such that one or more categories of the response variable can be predicted perfectly. When data exhibit quasi-complete or complete separation, maximum likelihood estimates of the parameters may not exist. Some computer programs will issue a warning message that quasi-complete separation has occurred, and others may not. The estimated standard errors for the parameter estimates get very large or "blow up." For example, separation is a problem in the High School and Beyond data set if the variable race is added to the model that includes all two-way interactions for achievement, SES, and school type. Most of the standard errors are less than 1; however, there are many that are between 10 and 64. Separation occurs when the model is too

Table 26.9 Correlations Between Unstandardized Explanatory Variables and Standardized Explanatory Variables

	Unstandardized			*Standardized*		
	SES	*Achieve*	*School Type*	*SES*	*Achieve*	*School Type*
SES × School type	0.74	0.28	0.76	–0.05	–0.04	–0.05
Achieve × School type	0.27	0.49	0.92	–0.04	–0.18	–0.04
SES × Achieve	0.93	0.64	0.17	–0.05	0.01	–0.06

complex for the data. The solution is to get more data or simplify the model.

Software

An incomplete list of programs that can fit the models reported in this chapter is given here. The models were fit to data for this chapter using SAS Version 9.1 (SAS Institute, Inc., 2003). The baseline models were fit using PROC LOGISTIC under the STAT package, and the conditional multinomial models were fit using PROC MDC in the econometrics package, ETS. Input files for all analyses reported in this chapter as well as how to compute regression diagnostics are available from the author's Web site at http://faculty.ed.uiuc.edu/cja/BestPractices/index.html. Also available from this Web site is a SAS MACRO that will compute range-of-influence statistics. Another commercial package that can fit both kinds of models is STATA (StataCorp LP, 2007; see Long, 1997). In the program R (R Development Core Team, 2007), which is an open-source version of SPLUS (Insightful Corp., 2007), the baseline model can be fit to data using the *glm* function, and the conditional multinomial models can be fit using the *coxph* function in the *survival* package. Finally, SPSS (SPSS, 2006) can fit the baseline model via the multinomial logistic regression function, and the conditional multinomial model can be fit using COXREG.

Models for Ordinal Responses

Ordinal response variables are often found in survey questions with response options such as *strongly agree, agree, disagree,* and *strongly disagree* (or *never, sometimes, often, all the time*). A number of extensions of the binary logistic regression model exist for ordinal variables, the most common being the proportional odds model (also known as the cumulative logit model), the continuation ratios model, and the adjacent categories model. Each of these is a bit different in terms of how the response categories are dichotomized as well as other specifics. Descriptions of these models can be found in Agresti (2002, 2007), Fahrmeir and Tutz (2001), Hosmer and Lemeshow (2000), Long (1997), Powers and Xie (2000), and elsewhere.

MANOVA, Discriminant Analysis, and Logistic Regression

Before concluding this chapter, a discussion of the relationship between multinomial logistic regression models, multivariate analysis of variance (MANOVA), and linear discriminant analysis is warranted. As an alternative to multinomial logistic regression with multiple explanatory variables, a researcher may choose MANOVA to test for group differences or linear discriminant analysis for either classification into groups or description of group differences. The relationship between MANOVA and linear discriminant analysis is well documented (e.g., Dillon & Goldstein, 1984; Johnson & Wichern, 1998); however, these models are in fact very closely related to logistic regression models.

MANOVA, discriminant analysis, and multinomial logistic regression all are applicable to the situation where there is a single discrete variable Y with J categories (e.g., high school program type) and a set of K random continuous variables denoted by $\boldsymbol{X} = (X_1, X_2, \ldots, X_K)'$ (e.g., achievement test scores on different subjects). Before performing discriminant analysis, it is the recommended practice to first perform a MANOVA to test whether differences over the J categories or groups exist (i.e., $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_J$, where $\boldsymbol{\mu}_j$ is a vector of means on the K variables). The assumptions for MANOVA are that vectors of random variables from each group follow a multivariate normal distribution with mean equal to $\boldsymbol{\mu}_j$ for group j, the covariance matrices for the groups are all equal, and observations over groups are independent (i.e., $\boldsymbol{X}_j \sim N_K(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ and independent).

If the assumptions for MANOVA are met, then a multinomial logistic regression model *must necessarily* fit the data. This result is based on statistical graphical models for discrete and continuous variables (Lauritzen, 1996; Lauritzen & Wermuth, 1989). Logistic regression is just the "flip side of the same coin." If the assumptions for MANOVA are not met, then the statistical tests performed in MANOVA and/or discriminant analysis are not valid; however, statistical tests in logistic regression will likely still be valid. In the example in this chapter, SES and school type are discrete and clearly are not normal. This poses no problem for logistic regression.

A slight advantage of discriminant analysis and MANOVA over multinomial logistic regression

may be a greater familiarity with these methods by both researchers and readers and the ease of implementation of these methods. However, when the goal is classification or description, these advantages may come at the expense of classification accuracy, invalid statistical tests, and difficulty describing differences between groups. As an illustration of classification using the High School and Beyond data, discriminant analysis with achievement as the only feature was used to classify students, as well as a discriminant analysis with all main effects. Frequencies of students in each program type and the predicted frequencies under discriminant analysis and multinomial logistic regression are located in Table 26.10. In this example, logistic regression performed better than discriminant analysis in terms of overall classification accuracy. The correct classification rates from the multinomial logistic regression models were .58 (achievement only) and .60 (main effects of achievement, SES, and school type), and those for linear discriminant analysis were .52 in both cases.

If a researcher's interest is in describing differences between groups, multinomial logistic regression is superior for a number of reasons. The flexibility of logistic regression allows for a number of different models. In our example, we were able to test whether SES should be treated as an ordinal or nominal variable. Furthermore, we were also able to discern that students who attended general and vocational programs were indistinguishable in terms of SES and school type, but students who attended academic programs versus one of the other programs were distinguishable with respect to achievement test scores, SES, and school type. We were also able to describe the nature and amount of differences between students who attended general and vocational programs in a relatively simple way. The differences between students in academic programs and either a general or vocational program were slightly more complex but still straightforward to describe.

Exercises

1. Use the High School and Beyond data set (Tatsuoka & Lohnes, 1988) to model students' career choice as the response variable. Consider gender, achievement test scores, and self-concept as possible explanatory variables. Do any of the careers have the same regression parameters? Note that some careers have very low numbers of observations; these may have to be deleted from the analysis.

2. The English as a second language (ESL) data come from a study that investigated the validity and generalizability of an ESL placement test (Lee & Anderson, in press). The test is administrated to international students at a large midwestern university, and the results are used to place students in an appropriate ESL course sequence. The data set contains the test results (scores of 2, 3, 4), self-reported Test of English as a Foreign Language (TOEFL) scores, field of study (business, humanities, technology, life science), and topic used in the placement exam (language acquisition, ethics, trade barriers) for 1,125 international students. Controlling for general English-language ability as measured by the TOEFL, use this data set to investigate whether the topic of the placement test influences the test results. In particular, do students who receive a topic in their major field of study have an advantage over others? Majors in business were thought to have an advantage when

Table 26.10 Predicted Frequencies of Program Types for Linear Discriminant Analysis (DA) and Multinomial Logistic Regression (LR)

		Predicted Frequency			
Program Type	*Observed Frequency*	*DA—Achieve*	*DA—All Main*	*LR—Achieve*	*LR—All Main*
General	145	99	132	0	40
Academic	308	292	284	429	398
Vocational	147	209	184	171	162
Correction classification rate:		.52	.52	.58	.60

the topic was trade, and those in humanities might have had an advantage when the topic was language acquisition or ethics. Create a variable that tests this specific hypothesis.

NOTE

1. See Chapter 21 (this volume) for details on Poisson regression.

REFERENCES

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: John Wiley.

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). New York: John Wiley.

Albert, A., & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika, 71,* 1–10.

Dillon, W. R., & Goldstein, M. (1984). *Multivariate analysis: Methods and applications.* New York: John Wiley.

Fahrmeir, L., & Tutz, G. (2001). *Multivariate statistical modeling based on generalized linear models.* New York: Springer.

Fay, M. P. (2002). Measuring a binary response's range of influence in logistic regression. *The American Statistician, 56,* 5–9.

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

Insightful Corp. (2007). S-Plus, Version 8 [Computer software]. Seattle, WA: Author. Available from http://www.insightful.com

Johnson, R. A., & Wichern, D. W. (1998). *Applied multivariate statistical analysis* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Lauritzen, S. L. (1996). *Graphical models.* New York: Oxford University Press.

Lauritzen, S. L., & Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some of which are quantitative. *Annals of Statistics, 17,* 31–57.

Lee, H. K., & Anderson, C. J. (in press). Validity and topic generality of a writing performance test. *Language Testing.*

Lesaffe, E., & Albert, A. (1989). Multiple-group logistic regression diagnostics. *Applied Statistics, 38,* 425–440.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables.* Thousand Oaks, CA: Sage.

Maydeu-Olivares, A., & Joe, J. (2005). Limited and full-information estimation and goodness-of-fit testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association, 100,* 1009–1020.

Powers, D. A., & Xie, Y. (2000). *Statistical methods for categorical data analysis.* San Diego: Academic Press.

Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics, 9,* 705–724.

R Development Core Team. (2007). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: Author. Available from http://www.R-project.org

SAS Institute, Inc. (2003). SAS/STAT software, Version 9.1 [Computer software]. Cary, NC: Author. Available from http://www.sas.com

SPSS. (2006). SPSS for Windows, Release 15 [Computer software]. Chicago: Author. Available from http://www.spss.com

StataCorp LP. (2007). Stata Version 10 [Computer software]. College Station, TX: Author. Available from http://www.stata.com/products

Tatsuoka, M. M., & Lohnes, P. R. (1988). *Multivariate analysis: Techniques for educational and psychological research* (2nd ed.). New York: Macmillan.