A Tutorial in Logistic Regression

Author(s): Alfred DeMaris

Source: *Journal of Marriage and Family*, Nov., 1995, Vol. 57, No. 4 (Nov., 1995), pp. 956-968

Published by: National Council on Family Relations

Stable URL: https://www.jstor.org/stable/353415

ALFRED DEMARIS Bowling Green State University

A Tutorial in Logistic Regression

*This article discusses some major uses of the logistic regression model in social data analysis. Using the example of personal happiness, a trichotomous variable from the 1993 General Social Survey (*n *= 1,601), properties of the technique are illustrated by attempting to predict the odds of individuals being less, rather than more, happy with their lives. The exercise begins by treating happiness as dichotomous, distinguishing those who are not too happy from everyone else. Later in the article, all three categories of happiness are modeled via both polytomous and ordered logit models.*

Logistic regression has, in recent years, become the analytic technique of choice for the multivariate modeling of categorical dependent variables. Nevertheless, for many potential users this procedure is still relatively arcane. This article is therefore designed to render this technique more accessible to practicing researchers by comparing it, where possible, to linear regression. I will begin by discussing the modeling of a binary dependent variable. Then I will show the modeling of polytomous dependent variables, considering cases in which the values are alternately unordered, then ordered. Techniques are illustrated throughout using data from the 1993 General Social Survey (GSS). Because these data are widely available, the reader is encouraged to replicate the analyses shown so that he or she can receive a "hands on" tutorial in the techniques. The Appendix presents coding instructions for an exact replication of all analyses in the paper.

BINARY DEPENDENT VARIABLES

A topic that has intrigued several family researchers is the relationship of marital status to subjective well-being. One indicator of well-being is reported happiness, which will be the focus of our analyses. In the GSS happiness is assessed by a question asking, "Taken all together, how would you say things are these days—would you say that you are very happy, pretty happy, or not too happy?" Of the total of 1,606 respondents in the 1993 survey, five did not answer the question. Hence, all analyses in this article are based on the 1,601 respondents providing valid answers to this question. Because this item has only three values, it would not really be appropriate to treat it as interval. Suppose instead, then, that we treat it as dichotomous, coding the variable 1 for those who are not too happy, and 0 otherwise. The mean of this binary variable is the proportion of those in the sample who are "unhappy," which is 178/1,601, or .111. The corresponding proportion of unhappy people in the population, denoted by π, can also be thought of as the probability that a randomly selected person will be unhappy. My focus will be on modeling the probability of un-

*Department of Sociology, Bowling Green State University, Bowling Green, OH 43403.

Key Words: logistic regression, logit models, odds ratios, ordered logit models, polytomous logistic regression, probability models.

happiness as a function of marital status, as well as other social characteristics.

Why Not OLS?

One's first impulse would probably be to use linear regression, with $E(Y) = \pi$ as the dependent variable. The equation would be $\pi = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K$. (There is no error term here because the equation is for the expected value of Y, which, of course, is π.) However, the problems incurred in using OLS have been amply documented (Aldrich & Nelson, 1984; Hanushek & Jackson, 1977; Maddala, 1983). Three difficulties are paramount: the use of a linear function, the assumption of independence between the predictors and the error term, and error heteroskedasticity, or nonconstant variance of the errors across combinations of predictor values. Briefly, the use of a linear function is problematic because it leads to predicted probabilities outside the range of 0 to 1. The reason for this is that the right-hand side of the regression equation, $\alpha + \Sigma\beta_k X_k$, is not restricted to fall between 0 and 1, whereas the left-hand side, π, is. The pseudo-isolation condition (Bollen, 1989), requiring the error term to be uncorrelated with the predictors, is violated in OLS when a binary dependent variable is used (see Hanushek & Jackson, 1977, or McKelvey & Zavoina, 1975, for a detailed exposition of why this happens). Finally, the error term is inherently heteroskedastic because the error variance is $\pi(1-\pi)$. In that π varies with the values of the predictors, so does the error variance.

The Logistic Regression Model

Motivation to use the logistic regression model can be generated in one of two ways. The first is through a latent variable approach. This is particularly relevant for understanding standardized coefficients and one of the R^2 analogs in logistic regression. Using unhappiness as an example, we suppose that the true amount of unhappiness felt by an individual is a continuous variable, which we will denote as Y^*. Let us further suppose that positive values for Y^* index degrees of unhappiness, while zero or negative values for Y^* index degrees of happiness. (If the dependent variable is an event, so that correspondence to an underlying continuous version of the variable is not reasonable, we can think of Y^* as a propensity for the event to occur.) Our observed measure is very crude; we have recorded only whether or not an individual is unhappy. The observed Y is a reflection of the latent variable Y^*. More specifically, people report being not too happy whenever $Y^* > 0$. That is, $Y = 1$ if $Y^* > 0$, and $Y = 0$ otherwise. Now Y^* is assumed to be a linear function of the explanatory variables. That is, $Y^* = \alpha + \Sigma\beta_k X_k + \varepsilon$. Therefore, π, the probability that $Y = 1$, can be derived as:

$$\pi = P(Y = 1) = P(Y^* > 0) = P(\alpha + \Sigma\beta_k X_k + \varepsilon > 0)$$
$$= P(\varepsilon > -[\alpha + \Sigma\beta_k X_k]) = P(\varepsilon < \alpha + \Sigma\beta_k X_k).$$

The last term in this expression follows from the assumption that the errors have a symmetric distribution. In fact, it is the choice of distribution for the error term, ε, that determines the type of analysis that will be used. If, for example, one assumes that ε is normally distributed, we are led to using probit analysis (Aldrich & Nelson, 1984; Maddala, 1983). In this case, the last term in the above expression is the probability of a normally distributed variable (ε) being less than the value $\alpha + \Sigma\beta_k X_k$, which we shall call the *linear predictor.* There is no closed-form expression (i.e., simple algebraic formula) for evaluating this probability. If, instead, one assumes that the errors have a logistic distribution, the appropriate analytic technique is logistic regression. Practically, the normal and logistic distributions are sufficiently similar in shape that the choice of distribution is not really critical. For this reason, the substantive conclusions reached using probit analysis or logistic regression should be identical. However, the logistic distribution is advantageous for reasons of both mathematical tractability and interpretability, as will be made evident below.

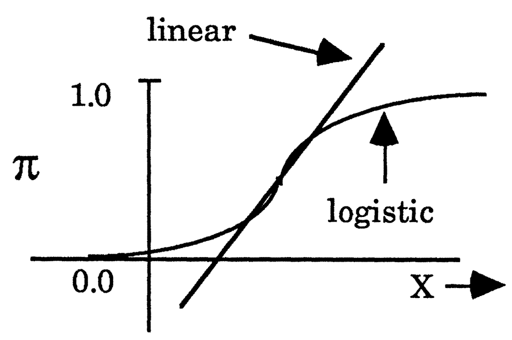
The mathematical advantage of the logit formulation is evident in the ability to express the probability that $Y = 1$ as a closed-form expression:

$$P(Y = 1) = \pi = \frac{\exp(\alpha + \Sigma\beta_k X_k)}{1 + \exp(\alpha + \Sigma\beta_k X_k).} \qquad (1)$$

In that the exponential function (exp) always results in a number between 0 and infinity, it is evident that the right-hand side of Equation 1 above is always bounded between 0 and 1. The difference between linear and logistic functions involving one predictor (X) is shown in Figure 1. The logistic function is an S-shaped or sigmoid curve that is approximately linear in the middle, but curved at either end, as X approaches either very small or very large values.

Indeed, we need not resort to a latent-variable formulation to be motivated to use the logistic

FIGURE 1. LINEAR VERSUS LOGISTIC FUNCTIONS OF *X*



curve to model probabilities. The second avenue leading to the logistic function involves observing that the linear function needs to "bend" a little at the ends in order to remain within the 0, 1 bounds. Just as the linear function is a natural choice for an interval response, so a sigmoid curve—such as the logistic function—is also a natural choice for modeling a probability.

Linearizing the Model

To write the right-hand side of Equation 1 as an additive function of the predictors, we use a logit transformation on the probability π. The logit transformation is log[π/(1−π)], where *log* refers to the natural logarithm (as it does throughout this article). The term π/(1−π) is called the *odds*, and is a ratio of probabilities. In the current example, it is the probability of being unhappy, divided by the probability of not being unhappy, and for the sample as a whole it is .111/.889 = .125. One would interpret these odds to mean that individuals are, overall, one-eighth (.125 = 1/8) as likely to be unhappy as they are to be happy. The log of this value is −2.079. The log odds can be any number between minus and plus infinity. It can therefore be modeled as a linear function of our predictor set. That is, the logistic regression model becomes:

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K. \quad (2)$$

This model is now analogous to the linear regression model, except that the dependent variable is a log odds. The estimation of the model proceeds via maximum likelihood, a technique that is too complex to devote space to here, but that is amply covered in other sources (e.g., Hosmer & Lemeshow, 1989). The user need not be concerned with these details except for certain numerical problems unique to this estimation method, which will be touched on later. Maximum likelihood estimates have desirable properties, one of which is that, in large samples, the regression coefficients are approximately normally distributed. This makes it possible to test each coefficient for significance using a *z* test.

Predicting Unhappiness

Let us attempt to predict unhappiness using several variables in the 1993 GSS. The predictors chosen for this exercise are marital status, age, self-assessed health (ranging from *poor* = 1 to *excellent* = 4), income, education, gender, race, and trauma in the past year. This last factor is a dummy variable representing whether the respondent experienced any "traumatic" events, such as deaths, divorces, periods of unemployment, hospitalizations, or disabilities, in the previous year. Table 1 presents descriptive statistics for variables used in the analysis. Four of the independent variables are treated as interval: age, health, income, and education. The other variables are dummies. Notice that marital status is coded as four dummies with married as the omitted category. Race is coded as two dummies with White as the omitted category.

To begin, we examine the relationship between unhappiness and marital status. Previous research suggests that married individuals are happier than those in other marital statuses (see especially Glenn & Weaver, 1988; Lee, Seccombe, & Shehan, 1991; Mastekaasa, 1992). The log odds of being unhappy are regressed on the four dummy variables representing marital status, and the result is shown in Model 1 in Table 2. In

TABLE 1. DESCRIPTIVE STATISTICS FOR VARIABLES IN THE ANALYSES

VARIABLE	MEAN	*SD*
Unhappy	.111	.314
Widowed	.107	.309
Divorced	.143	.350
Separated	.027	.164
Never married	.187	.390
Age	46.012	17.342
Health	3.017	.700
Income	12.330	4.261
Education	13.056	3.046
Male	.427	.495
Black	.111	.314
Other race	.050	.218
Trauma in past year	.235	.424

Note: *n* = 1,601.

linear regression, we use an F test to determine whether the predictor set is "globally" significant. If it is, then at least one beta is nonzero, and we conduct t tests to determine which betas are nonzero. Similarly, there is a global test in logistic regression, often called the *model chi-square* test. The test statistic is –2Log(L0) – [–2Log (L1)], where L0 is the likelihood function for a model containing only an intercept, and L1 is the likelihood function for the hypothesized model, evaluated at the maximum likelihood estimates. (The likelihood function gives the joint probability of observing the current sample values for Y, given the parameters in the model.) It appears on SPSS printouts as the "Model Chi-Square," and on SAS printouts in the column headed "Chi-Square for Covariates." The degrees of freedom for the test is the same as the number of predictors (excluding the intercept) in the equation, which, for Model 1, is 4. If this chi-square is significant, then at least one beta in the model is nonzero. We see that, with a value of 47.248, this statistic is highly significant ($p < .0001$).

Tests for each of the dummy coefficients reveal which marital statuses are different from being married in the log odds of being unhappy. (These show up on both SPSS and SAS printouts as the "Wald" statistics, and are just the squares of the z tests formed by dividing the coefficients by their standard errors.) Apparently, widowed, divorced, and separated respondents all have significantly higher log odds of being unhappy, compared to married respondents. Never-married people, on the other hand, do not. Because there are monotonic relationships among the log odds, the odds, and the probability, any variable that is positively related to the log odds is also positively related to the odds and to the probability. We can therefore immediately see that the odds or the probability of being unhappy are greater for widowed, divorced, and separated people, compared with married people.

Interpretation in terms of odds ratios. The coefficient values themselves can be rendered more interpretable in two ways. First, the estimated log odds are actually an estimate of the conditional mean of the latent unhappiness measure, Y^*. Thus, the betas are maximum likelihood estimates for the linear regression of Y^* on the predictors. This interpretation, however, is typically eschewed in logistic regression in favor of the "odds-ratio" approach. It turns out that $\exp(b_k)$ is the estimated odds ratio for those who are a unit apart on X_k, net of other predictors in the model. For dummy coefficients, a unit difference in X_k is the difference between membership in category X_k and membership in the omitted category. In this case, $\exp(b_k)$ is the odds ratio for those in the membership category versus those in the omitted category. For example, the odds ratio for widowed versus married respondents is exp(1.248) = 3.483. This implies that the odds of being unhap-

TABLE 2. COEFFICIENTS FOR VARIOUS LOGISTIC REGRESSION MODELS OF THE LOG ODDS OF BEING UNHAPPY

	Model				
Variable	1	2	3	Beta	4
Intercept	–2.570***	.225	.405	—	–.461
Widowed	1.248***	.967***	.907**	.155	.915**
Divorced	.955***	.839***	.819***	.158	.820***
Separated	1.701***	1.586***	1.486***	.134	1.487***
Never married	.376	.259	.262	.056	.268
Age		–.005	–.006	–.060	–.006
Health		–.823***	–.793***	–.306	
Income		–.011	–.001	–.002	–.001
Education			–.038	–.064	–.038
Male			.052	.014	.052
Black			–.152	–.026	–.144
Other race			.210	.025	.209
Trauma in past year			.535**	.125	.519**
Excellent health					–2.451***
Good health					–1.486***
Fair health					–.719*
Model chi-square[a]	47.248	104.540	116.036	—	116.419
Degrees of freedom	4	7	12	—	14

[a]All model chi-squares are significant at $p < .001$.
*$p < .05$. **$p < .01$. ***$p < .001$.

py are about 3.5 times as large for widowed people as they are for married people.

It is important to note that it would not be correct to say that widowed people are 3.5 times as likely to be unhappy, meaning that their probability of unhappiness is 3.5 times higher than for married people. If p_1 is the estimated probability of unhappiness for widowed people and p_0 is the estimated probability of unhappiness for married people, then the odds (of being unhappy) ratio for widowed, versus married people is:

$$\frac{p_1/1-p_1}{p_0/1-p_0} = \left(\frac{p_1}{p_0}\right)\left(\frac{1-p_0}{1-p_1}\right). \qquad (3)$$

The aforesaid (incorrect) interpretation applies only to the first term on the right-hand side of Equation 3, $\left(\frac{p_1}{p_0}\right)$, which would be called the *relative risk* of unhappiness (Hosmer & Lemeshow, 1989). However, this is not equivalent to the odds ratio unless both p_1 and p_0 are very small, in which case the second term on the right-hand side of Equation 3, $\left(\frac{1-p_0}{1-p_1}\right)$, approaches 1.

Testing hierarchical models. To what extent are marital status differences in unhappiness explained by other characteristics associated with marital status? For example, widowhood is associated with older age, and possibly poorer health. Divorce and separation are typically accompanied by a drop in financial status. Might these variables account for some of the effect of marital status? Model 2 in Table 2 controls for three additional predictors: age, health, and income.

First, we can examine whether the addition of these variables makes a significant contribution to the prediction of unhappiness. The test is the difference in model chi-squares between models with and without these extra terms. Under the null hypothesis that the coefficients for the three additional variables are all zero, this statistic is itself distributed as chi-square with degrees of freedom equal to the number of terms added. In this case, the test is 104.540 – 47.248 = 57.292. With 3 degrees of freedom, it is a highly significant result ($p < .001$), suggesting that at least one of the additional terms is important.

Tests for the additional coefficients reveal that the significance of the added variable set is due primarily to the strong impact of health on unhappiness. Each unit increment in self-assessed health (e.g., from being in fair health to being in good health) decreases the odds of unhappiness by a factor of exp(–.823) = .439. Another way of expressing this result is to observe that $100(e^b - 1)$ is the percentage change in the odds for each unit increase in X. Therefore, each unit increase in health changes the odds of unhappiness by $100(e^{-.823} - 1) = -56.1\%$, or effects a 56.1% reduction in these odds. The effects for dummies representing marital status have all been reduced somewhat, indicating that the additional variables account for some of the marital status differences. However, widowed, divorced, and separated respondents are still significantly more unhappy than married respondents, even after these additional variables are controlled.

Model 3 in Table 2 includes the remaining variables of interest: education, gender, race, and trauma. Addition of this variable set also makes a significant contribution to the model ($\chi^2 = 116.036 - 104.540 = 11.496$, $df = 5$, $p = .042$). Of the added variables, trauma alone is significant. Its coefficient suggests that having experienced a traumatic event in the past year enhances the odds of being unhappy by a factor of exp(.535) = 1.707.

Standardized coefficients. In linear regression, the standardized coefficient is $b_k(s_{xk}/s_y)$, and is interpreted as the standard deviation change in the mean of Y for a standard deviation increase in X_k, holding other variables constant. In logistic regression, calculating a standardized coefficient is less straightforward. The regression coefficient, b_k, is the "unit impact" (i.e., the change in the dependent variable for a unit increase in X_k) on Y^*, which is unobserved. But we require the standard deviation of Y^* to compute a standardized coefficient. One possibility would be to estimate the standard deviation of Y^* as follows.

Recall from above that $Y^* = \alpha + \Sigma\beta_k X_k + \varepsilon$, where ε is assumed to follow the logistic distribution. Like the standard normal, this distribution has a mean of zero, but its variance is $\pi^2/3$, where π is approximately 3.1416. Moreover, $V(Y^*) = V(\alpha + \Sigma\beta_k X_k + \varepsilon) = V(\alpha + \Sigma\beta_k X_k) + V(\varepsilon)$. The simplicity of the last expression is due to the assumption that the error term is uncorrelated with the predictors (hence, there are no covariance terms involving the errors). Now the variance of the errors is assumed to be $\pi^2/3$ (or about 3.290), so if we add to that the sample variance of the linear predictor, $\hat{V}(\alpha + \Sigma\beta_k X_k)$, we have an estimate of the variance of Y^*. The sample variance of the linear predictor can easily be found in SAS by using an output statement after PROC LOGISTIC to add the SAS variable XBETA to the data set, and then employing PROC MEANS to find its variance. (SPSS does not make the linear predictor available as a variable.) For Model 3, this variance is .673, implying that the variance of Y^* is .673 + 3.290 = 3.963. Adjusting our coefficients by the factor (s_{xk}/s_{y*}) would then result in standardized coefficients with the usual interpretation, except that it would apply to the mean of the latent variable, Y^*.

The major drawback to this approach is that, because the variance of the linear predictor varies according to which model is estimated, so does our estimate of the variance of Y^*. This means that it would be possible for a variable's standardized effect to change across equations even though its unstandardized effect remained the same. This would not be a desirable property in a standardized estimate.

One solution to this problem, utilized by SAS's PROC LOGISTIC, is to "partially" standardize the coefficients by the factor $(s_{xk}/\sigma_\varepsilon)$, where σ_ε is the standard deviation of the errors, or $\pi/\sqrt{3}$. In that σ_ε is a constant, such a "standardized" coefficient will change across equations only if the unstandardized coefficient changes. Although these partially standardized coefficients no longer have the same interpretation as the standardized coefficients in linear regression, they perform the same function of indicating the relative magnitude of each variable's effect in the equation. The column headed *beta* in Table 2 shows the "standardized coefficients" produced by SAS for Model 3. From these coefficients, we can see that self-assessed health has the largest impact on the log odds of unhappiness, with a coefficient of –.306. This is followed by marital status, with being widowed, divorced, or separated all having the next largest effects. In that the standardized coefficients are estimates of the standardized effects of predictors that would be obtained from a linear regression involving Y^*, the latent continuous variable, all caveats for the interpretation of standardized coefficients in ordinary regression apply (see, e.g., McClendon, 1994, for cautions in the interpretation of standardized coefficients).

Nonlinear effects of predictors. Until now, we have assumed that the effects of interval-level predictors are all linear. This assumption can be checked. For example, health has only four levels. Is it reasonable to assume that it bears a linear relationship to the log odds of being unhappy? To examine this, I have dummied the categories of health, with poor health as the reference category, and rerun Model 3 with the health dummies in place of the quantitative version of the variable. The results are shown in Model 4 in Table 2. The coefficients for the dummies indeed suggest a linear trend, with increasingly negative effects for groups with better, as opposed to poorer, health. In that Model 3 is nested within Model 4, the test for linearity is the difference in model chi-squares for these two models. The result is 116.419 – 116.036 = .383, which, with 2 degrees of freedom, is very nonsignificant ($p > .8$). We conclude that a linear specification for health is reasonable.

What about potential nonlinear effects of age, income, and education? To examine these, I added three quadratic terms to Model 3: age-squared, income-squared, and education-squared. The chi-square difference test is not quite significant ($\chi^2 = 6.978$, $df = 3$, $p = .073$), suggesting that these variables make no significant contribution to the model. Nonetheless, age-squared is significant at $p = .02$. For didactic purposes, therefore, I will include it in the final model. Sometimes predictors have nonlinear effects that cannot be fitted by using a quadratic term. To check for these, one could group the variable's values into four or five categories (usually based on quartiles or quintiles of the variable's distribution) and then use dummies to represent the categories (leaving one category out, of course) in the model. This technique is illustrated in other treatments of logistic regression (DeMaris, 1992; Hosmer & Lemeshow, 1989).

Predicted probabilities. We may be interested in couching our results in terms of the probability, rather than the odds, of being unhappy. Using the equation for the estimated log odds, it is a simple matter to estimate probabilities. We substitute the sample estimates of parameters and the values of our variables into the logistic regression equation shown in Equation 2, which provides the estimated log odds, or logit. Then exp(logit) is the estimated odds, and the probability = odds/(1 + odds). As an example, suppose that we wish to estimate the probability of unhappiness for 65-year-old widowed individuals in excellent health with incomes between $25,000 and $30,000 per year (RINCOM91 = 15), based on Model 2 in Table 2. The estimated logit is .225 + .967 – .005(65) – .823(4) –.011(15) = –2.59. The estimated odds is exp(–2.59) = .075, and the estimated probability is therefore .075/(1 + .075) = .07.

Probabilities versus odds. Predicted probabilities are perhaps most useful when the purpose of the analysis is to forecast the probability of an event, given a set of respondent characteristics. If, as is more often the case, one is merely interested in the impact of independent variables, controlling for other effects in the model, the odds ratio is the preferred measure. It is instructive to consider why. In linear regression, the partial derivative, β_k, is synonymous with the unit impact of X_k: Each unit increase in X_k adds β_k to the expected value of Y. In logistic regression $\exp(\beta_k)$ is its multiplicative analog: Each unit increase in X_k multiplies the odds by $\exp(\beta_k)$. There is no such summary measure for the impact of X_k on the probability. The reason for this is that an artifact of the probability model, shown in Equation 1, is that it is interactive in the X values. This can be seen by taking the partial derivative of π with respect to X_k. The result is $\beta_k(\pi)(1\text{-}\pi)$. In that π varies with each X in the model, so does the value of the partial slope in this model. That, of course, means that the impact of X_k on π is not constant, as in linear regression, but rather depends on the values of the other variables, as well as on the value of X_k itself. Therefore, a value computed for the partial slope applies only to a specific set of values for the Xs, and by implication, a specific level of π. (See DeMaris, 1993a, 1993b, for an extended discussion of this problem. Another artifact of the model is that the partial slope is greatest when $\pi = .5$; see Nagler, 1994, for a description of the "scobit" model, in which the π value at which X_k has maximum effect can be estimated from the data.)

Suppose that we are interested in examining the estimated impact of a given variable, say education, on the probability of being unhappy. The final model for unhappiness is shown in the first two columns of Table 3. The antilogs of the coefficients are shown in the column headed *Exp(b)*. As noted, these indicate the impact on the odds of being unhappy for a unit increase in each predictor, controlling for the others. To estimate the unit impact of education on the probability, we must choose a set of values for the Xs. Hence, suppose that we set the quantitative variables, age, health, income, and education, at the mean values for the sample, and we set the qualitative variables, mari-

TABLE 3. FINAL LOGISTIC REGRESSION MODEL FOR THE LOG ODDS OF BEING UNHAPPY AND INTERACTION MODEL ILLUSTRATING PROBLEMS WITH ZERO CELLS

	Final Model		Interaction Model	
Variable	*b*	Exp(*b*)	*b*	*SE*(*b*)
Intercept	–1.022	.360	–1.044	.911
Widowed	1.094***	2.985	1.164***	.327
Divorced	.772***	2.163	.880***	.250
Separated	1.476***	4.376	1.394**	.503
Never married	.430	1.537	.624*	.288
Age	.061	1.062	.058	.032
Age-squared	–.0007*	.999	–.0006*	.0003
Health	–.785***	.456	–.789***	.116
Income	–.006	.994	–.006	.022
Education	–.043	.958	–.042	.031
Male	.065	1.067	.052	.181
Black	–.179	.836	.380	.467
Other race	.197	1.217	.561	.513
Trauma in past year	.535**	1.708	.558**	.180
Black × widowed			–.651	.743
Black × divorced			–1.010	.818
Black × separated			–.364	.862
Black × never married			–.904	.745
Other race × widowed			–.340	1.063
Other race × divorced			–.292	.989
Other race × separated			5.972	22.252
Other race × never married			–1.695	1.183
Model chi-square	120.910***		126.945***	
Degrees of freedom	13		21	

$^*p < .05.$ $^{**}p < .01.$ $^{***}p < .001.$

tal status, gender, race, and trauma, to their modes. We are therefore estimating the probability of unhappiness for 46.012-year-old White married females with 13.056 years of education, earning between \$17,500 and \$20,000 per year (RINCOM91 = 12.33), who are in good health (health = 3.017) and who have not experienced any trauma in the past year. Substituting these values for the predictors into the final model for the log odds produces an estimated probability of unhappiness of .063.

The simplest way to find the change in the probability of unhappiness for a year increase in education is to examine the impact on the odds for a unit change in education, then translate the new odds into a new probability. The current odds are .063/(1 – .063) = .067. The new odds are .067(.958) = .064, which implies a new probability of .06. The change in probability (new probability – old probability) is therefore –.003. What if we are interested, more generally, in the impact on the odds or the probability of a *c*-unit change, where *c* can be any value? The estimated impact on the odds for a *c*-unit increase in X_k is just $\exp(cb_k)$. For example, what would be the change in the odds and the probability of unhappiness in the current example for a 4-year increase in educational level? The new odds would be .067[exp(–.043 x 4)] = .0564, implying a new probability of .053, or a change in probability of –.01.

Nonlinear effects, such as those for age, involve the coefficients for both *X* and *X*-squared. In general, if b_1 is the coefficient for *X*, and b_2 is the coefficient for *X*-squared, then the impact on the log odds for a unit increase in *X* is $b_1 + b_2 + 2(b_2X)$. The impact on the odds therefore is $\exp(b1+b2)\ \exp(2b_2X)$. For age, this amounts to $\exp(.061 - .0007)\ \exp(2[-.0007 \times \text{age}])$. It is evident from this expression that the impact on the odds of unhappiness of growing a year older depends upon current age. For 18-year-olds, the impact is 1.036, for 40-year-olds, it is 1.004, and for 65-year-olds it is .97. This means that the effect of age changes direction, so to speak, after about age 40. Before age 40, growing older slightly increases the odds of unhappiness; after age 40, growing older slightly decreases those odds. This trend might be called *convex curvilinear*. In that I have devoted considerable attention to interpretation issues in logistic regression elsewhere (DeMaris 1990, 1991, 1993a, 1993b), the interested reader may wish to consult those sources for further discussion.

Evaluating predictive efficacy. Predictive efficacy refers to the degree to which prediction error is reduced when using, as opposed to ignoring, the predictor set. In linear regression, this is assessed with R^2. There is no commonly accepted analog in logistic regression, although several have been proposed. I shall discuss three such measures, all of which are readily obtained from existing software. Additional means of evaluating the fit of the model are discussed in Hosmer and Lemeshow (1989).

In logistic regression, –2LogL0 is analogous to the total sum of squares in linear regression, and –2LogL1 is analogous to the residual sum of squares in linear regression. Therefore, the first measure, denoted R^2_L (Hosmer & Lemeshow, 1989) is:

$$R^2_L = \frac{-2\text{LogL0} - (-2\text{LogL1})}{-2\text{LogL0}}$$

The log likelihood is not really a sum of squares, so this measure does not have an "explained variance" interpretation. Rather it indicates the relative improvement in the likelihood of observing the sample data under the hypothesized model, compared with a model with the intercept alone. Like R^2 in linear regression, it ranges from 0 to 1, with 1 indicating perfect predictive efficacy. For the final model in Table 3, $R^2_L = .108$.

A second measure is a modified version of the Aldrich-Nelson "pseudo-R^2" (Aldrich & Nelson, 1984). This measure, proposed by Hagle and Mitchell (1992), is:

$$R^{2*}_{AN} = \frac{R^2_{AN}}{\max(R^2_{AN} \mid \hat{\pi})}$$

where R^2_{AN}, the original Aldrich-Nelson measure, is: model chi-square/(model chi-square + sample size), and $\max(R^2_{AN} \mid \hat{\pi})$ is the maximum value attainable by R^2_{AN}, given $\hat{\pi}$, the sample proportion with *Y* equal to 1. The formula for $\max(R^2_{AN} \mid \hat{\pi})$is:

$$\frac{-2[\hat{\pi}\log\hat{\pi} + (1-\hat{\pi})\log(1-\hat{\pi})]}{1 - 2[\hat{\pi}\log\hat{\pi} + (1-\hat{\pi})\log(1-\hat{\pi})].}$$

This modification to R^2_{AN} ensures that the measure will be bounded by 0 and 1. Other than the fact that it is so bounded, with 1 again indicating perfect predictive efficacy, the value is not really interpretable. For the final model, $\max(R^2_{AN} \mid \hat{\pi})$ is .4108, and R^{2*}_{AN} is .171.

A third measure of predictive efficacy, proposed by McKelvey and Zavoina (1975), is an estimate of R^2 for the linear regression of Y^* on the predictor set. Its value is:

$$R^2_{MZ} = \frac{\hat{V}(\alpha + \Sigma\beta_k X_k)}{\hat{V}(\alpha + \Sigma\beta_k X_k) + \pi^2/3}$$

where, as before, $\hat{V}(\alpha + \Sigma\beta_k X_k)$ is the sample variance of the linear predictor, or the variance accounted for by the predictor set, and the denominator is, once again, an estimate of the variance of Y^*. This measure is .175 in the current example (the variance of the linear predictor is .7 for the final model), suggesting that about 17.5% of the variance in the underlying continuous measure of unhappiness is accounted for by the model. (For a comparison of the performance of these three measures, based on simulation results, see Hagle & Mitchell, 1992.)

Tests for contrasts on a qualitative variable. To assess the global impact of marital status in the final model, we must test the significance of the set of dummies representing this variable when they are added after the other predictors. This test proves to be quite significant ($\chi^2 = 22.156$, $df = 4$, $p = .00019$). We see that widowed, separated, and divorced individuals have higher odds of being unhappy than those who are currently married. What about other comparisons among categories of marital status? In all, there are 10 possible contrasts among these categories. Exponentiating the difference between pairs of dummy coefficients provides the odds ratios for the comparisons involving the dummied groups. For example, the odds ratio for widowed versus divorced individuals is exp(1.094 – .772) = 1.38. This ratio is significantly different from 1 only if the difference between coefficients for widowed and divorced respondents is significant. The test is the difference between estimated coefficients divided by the estimated standard error of this difference. If b_1 and b_2 are the coefficients for two categories, the estimated standard error is:

$$\textit{estimated } SE(b_1 - b_2) = \sqrt{\hat{V}(b_1) + \hat{V}(b_2) - 2\text{cov}(b_1, b_2)}.$$

The necessary variances and covariances for these computations can be obtained in SAS by requesting the covariance matrix of parameter estimates.

When making multiple comparisons, it is customary to adjust for the concomitant increase in the probability of Type I error. This can be done using a modified version of the Bonferroni procedure, proposed by Holm (Holland & Copenhaver, 1988). The usual Bonferroni approach would be to divide the desired alpha level for the entire series of contrasts, typically .05, by the total number of contrasts, denoted by m, which in this case is 10. Hence each test would be made at an alpha level of .005. Although this technique ensures that the overall Type I error rate for all contrasts is less than or equal to .05, it tends to be too conservative, and therefore not very powerful.

The Bonferroni-Holm procedure instead uses a graduated series of alpha levels for the contrasts. First, one orders the p values for the contrasts from smallest to largest. Then, the smallest p value is compared to alpha/m. If significant, the next smallest p value is compared to alpha/(m – 1). If that contrast is significant, the next smallest p value is compared to alpha/(m – 2). We continue in this manner, until, provided that each contrast is declared significant, the last contrast is tested at alpha. If at any point, a contrast is not significant, the testing is stopped, and all contrasts with larger p values are also declared nonsignificant. The results of all marital-status contrasts, using this approach, are shown in Table 4, in which α' is the Bonferroni-Holm alpha level for each contrast. It appears that the only significant contrasts are, as already noted, between widowed, divorced, and separated people, on the one hand, and married individuals on the other. Notice that, without adjusting for capitalization on chance, we would also have declared separated individuals to have greater odds of being unhappy than the never-married respondents.

TABLE 4. BONFERRONI-HOLM ADJUSTED TESTS FOR MARITAL STATUS CONTRASTS ON THE LOG ODDS OF BEING UNHAPPY (BASED ON THE FINAL MODEL IN TABLE 3)

Contract	p	α'	Conclusion
Separated vs. married	.0002	.0050	Reject H0
Widowed vs. married	.0003	.0056	Reject H0
Divorced vs. married	.0008	.0063	Reject H0
Separated vs. never married	.0133	.0071	Do not reject H0
Widowed vs. never married	.0732	.0083	Do not reject H0
Divorced vs. separated	.0823	.0100	Do not reject H0
Married vs. never married	.1073	.0125	Do not reject H0
Divorced vs. never married	.2542	.0167	Do not reject H0
Widowed vs. divorced	.3310	.0250	Do not reject H0
Widowed vs. separated	.4017	.0500	Do not reject H0

Numerical problems. As in linear regression, multicollinearity is also a problem in logistic regression. If the researcher suspects collinearity problems, these can be checked by running the model using OLS and requesting collinearity diagnostics. Two other numerical problems are unique to maximum likelihood estimation. One is a problem called *complete separation.* This refers to the relatively rare situation in which one or more of the predictors perfectly discriminates the outcome groups (Hosmer & Lemeshow, 1989). Under this condition the maximum likelihood estimates do not exist for the predictors involved. The tip-off in the estimation procedure will usually be estimated coefficients reported to be infinite, or unreasonably large coefficients along with huge standard errors. As Hosmer and Lemeshow noted (1989, p. 131), this is a problem the researcher has to work around.

A more common problem with a simpler solution is the case of zero cell counts. For example, suppose that we wish to test for an interaction between race and marital status in their effects on the odds of unhappiness. If one forms the three-way contingency table of unhappy by marital by race, one will discover a zero cell: Among those in the other race category who are separated, there are no happy individuals. This causes a problem in the estimation of the coefficient for the interaction between the statuses other race and separated. The interaction model in Table 3 shows the results. We are alerted that there is a problem by the standard error for the other race × separated term: It is almost 20 times larger than the next largest standard error, and clearly stands out as problematic. To remedy the situation, we need only collapse categories of either marital status or race. Taking the latter approach, I have coded both Blacks and those of other races as non-White, and retested the interaction. It proves to be nonsignificant ($\chi^2 = 3.882$, $df = 4$, $p > .4$).

Interpreting interaction effects. Although the interaction effect is not significant here, it is worthwhile to consider briefly the interpretation of first-order interaction in logistic regression (the interested reader will find more complete expositions in DeMaris, 1991, or Hosmer & Lemeshow, 1989). Suppose that, in general, we wish to explore the interaction between X_1 and X_2 in their effects on the log odds. We shall also assume that these variables are involved only in an interaction with each other, and not with any other predictors. We will further designate X_1 as the focus variable and X_2 as the moderator variable. The logistic regression equation, with $W_1, W_2, \ldots, W_K$ as the other predictors in the model, is:

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \Sigma\lambda_k W_k + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$
$$= \alpha + \Sigma\lambda_k W_k + \beta_2 X_2 + (\beta_1 + \beta_3 X_2) X_1.$$

The partial slope of the impact of X_1 on the log odds is therefore $(\beta_1 + \beta_3 X_2)$, implying that the impact of X_1 depends upon the level of X_2. The multiplicative impact of X_1 on the odds is, correspondingly, $\exp(\beta_1 + \beta_3 X_2)$. This can also be interpreted as the odds ratio for those who are a unit apart on X_1, controlling for other predictors. This odds ratio changes across levels of X_2, however. When X_1 and X_2 are each dummy variables, the interpretation is simplified further. For instance, in the model discussed above in which the variable non-White interacts with marital status (results not shown), the coefficient for being separated is 1.396, whereas the coefficient for the cross-product of separated status (X_1) by non-White status (X_2) is –.290. The multiplicative impact on the odds of being separated is therefore exp(1.396 – .290*non-White). For Whites, the impact of being separated is exp(1.396) = 4.039. This suggests that among Whites, separated individuals have odds of unhappiness that are about 4 times those of married individuals. Among non-Whites, on the other hand, the odds of being unhappy are only exp(1.396 – .290) = 3.022 times as great for separated, as opposed to married, individuals.

Polytomous Dependent Variables

The Qualitative Case

We now turn to the case in which we wish to model a response with three or more categories. This actually describes the original uncollapsed variable, happy, in the GSS. Although we would most probably consider this variable ordinal, we will begin by ignoring the ordering of levels of happiness, and treat the variable as though it were purely qualitative. We will see that the nature of the associations between happiness and the predictors will itself suggest treating the variable as ordinal.

To model a polytomous dependent variable, we must form as many log odds as there are categories of the variable, minus 1. Each log odds contrasts a category of the variable with a baseline category. In that the modal category of happiness was being pretty happy, this will be the baseline category. My interest will be in examining how the predictor set affects the log odds of (a) being very happy, as opposed to being pretty happy, and (b) being not too happy, as opposed to being pretty happy. That is, I will predict departures in either direction from the modal condition of saying that one is pretty happy. Each log odds is modeled as a linear function of the predictor set, just as with a binary dependent variable. Notice again that each odds is the ratio of two probabilities, and that in both cases the denominator of the odds is the probability of being pretty happy. The third contrast, involving the odds of being very happy versus not too happy, is not estimated, because the coefficients for this equation are simply the differences between coefficients from the first and second equations.

To run this analysis, I use PROC CATMOD in SAS. (As of this writing, SPSS has no procedure for running polytomous logistic regression; an approximation, however, can be obtained using the method outlined by Begg & Gray, 1984.) However, I must first recode the variable happy so that 1 is very happy, 2 is not too happy, and 3 is pretty happy. The reason for this is that the highest valued category of the variable is automatically used by SAS as the baseline category. The results are shown in the first two columns of Table 5.

Once again there is a global test for the impact of the predictor set on the dependent variable, of the form –2Log(L0) – [–2Log(L1)]. This test, however, is not a standard part of SAS output. Nevertheless, because SAS always prints out minus twice the log likelihood for the current model, it is easy to compute. We simply request a model without any predictors (e.g., MODEL POLYHAP = /ML NOPROFILE NOGLS;) to get –2LogL0 (it is the value for the last iteration under "Maximum-Likelihood Analysis"). With –2LogL1 reported in the analysis for the hypothesized model, we can then calculate the model chi-square by hand. In this example it is 235.514. With 26 degrees of freedom, it is highly significant ($p < .0001$).

In polytomous logistic regression, each predictor has as many effects as there are equations estimated. Therefore there is also a global test for each predictor, and these are reported in SAS. From these results, we find that variables with significant effects overall are: being widowed, being divorced, being separated, being never-married, health, being male, trauma, and age-squared. Tests for coefficients in each equation reveal which log odds are significantly affected by the predictor. Hence, we see that being widowed or divorced significantly reduces the odds of being very happy, and significantly enhances the odds of being not too happy, compared with being married. Exponentiating the coefficients moreover provides odds ratios to facilitate interpretation, as before. Divorced individuals, for example, have odds of being very happy that are exp(–.831) = .436 times those for married indi-

TABLE 5. COEFFICIENTS FOR POLYTOMOUS AND ORDERED LOGIT MODELS OF GENERAL HAPPINESS

Variable	Very Happy Vs. Pretty Happy	Not Too Happy Vs. Pretty Happy	Not Too Happy Vs. Pretty or Very Happy	Not Too or Pretty Vs. Very Happy	Less, Rather Than More, Happy
Intercept	–2.665	–1.125	–1.022	2.946	—
Widowed[a]	–.942***	.823**	1.094***	1.079***	1.155***
Divorced[a]	–.831***	.542*	.772***	.906***	.849***
Separated[a]	–.213	1.393***	1.476***	.509	1.090***
Never married[a]	–.395*	.312	.430	.435*	.441**
Age	–.006	.061	.061	.014	.030
Age-squared[a]	.0002	–.0006*	–.0007*	–.0003	–.0004*
Health[a]	.683***	–.614***	–.785***	–.771***	–.793***
Income	–.0002	–.008	–.006	–.0003	–.0005
Education	.011	–.039	–.043	–.017	–.025
Male[a]	–.361**	–.030	.065	.359**	.272*
Black	–.353	–.260	–.179	.321	.135
Other race	–.127	.159	.197	.153	.173
Trauma in past year[a]	–.100	.500**	.535**	.173	.283*
Model chi-square	235.514***	235.514***	120.910***	154.356***	215.295***
Degrees of freedom	26	26	13	13	13

[a]Global effect on both log odds is significant in polytomous model.
*$p < .05$. **$p < .01$. ***$p < .001$.

viduals. That is, divorced people's odds of being very happy are less than half of those experienced by married people.

Health is also a variable that enhances the odds of being very happy, while reducing the odds of being not too happy. Other factors, on the other hand, affect only one or the other log odds. Thus being separated increases the odds of being not too happy, as opposed to being pretty happy, but does not affect the odds of being very happy. Trauma similarly affects only the odds of being not too happy. And age apparently has significant nonlinear effects only on the log odds of being unhappy. Interestingly, when all three categories of happiness are involved, men and never-married individuals are less likely than women and married people to say they are very happy. There are no significant differences between these groups, however, in the likelihood of being not too happy.

Ordinal Dependent Variables

The direction of effects of the predictors on each log odds provides considerable support for treating happiness as an ordinal variable. All significant predictors, and most of the nonsignificant ones as well, have effects that are of opposite signs on each log odds. (As an example, health is positively related to the odds of being very happy, and negatively related to the odds of being unhappy.) This suggests that unit increases in each predictor are related to monotone shifts in the odds of being more, rather than less, happy. To take advantage of the ordered nature of the levels of happiness, we use the ordered logit model.

With ordinal variables, the logits are formed in a manner that takes ordering of the values into account. One method is to form "cumulative logits," in which odds are formed by successively cumulating probabilities in the numerator of the odds (Agresti, 1990). Once again we estimate equations for two log odds. However, this time the odds are (a) the probability of being not too happy, divided by the probability of being pretty happy, or very happy and (b) the probability of being not too happy, or pretty happy, divided by the probability of being very happy. The first log odds is exactly the same as was estimated in Table 3. Those coefficients are reproduced in the third column of Table 5. The equation for the second log odds is shown in the fourth column of the Table.

Because the coefficients for each predictor across these two equations are, for the most part, similar, we might test whether, in fact, only one equation is necessary instead of two. That is, whether the predictor set has the same impact on the log odds of being less, rather than more, happy, regardless of how less happy and more happy are defined, can be tested via a chi-square statistic. This is reported in SAS as the "Score Test for the Proportional Odds Assumption." In the current example, its value is 22.21, which, with 13 degrees of freedom, is not quite significant (p = .052). This suggests that we cannot reject the hypothesis that the coefficients are the same across equations, thus implying that one equation is sufficient to model the log odds of being less, rather than more, happy. This equation is shown in the last column of Table 5. (To run this analysis in SAS, the variable happy must be reverse-coded.) To summarize, we can say that all nonmarried statuses are associated with enhanced odds of being less (rather than more) happy, compared with married people. Men and those experiencing trauma in the past year also have enhanced odds of being less happy. Those in better health have lower odds of being less happy. Age, as before, shows a convex curvilinear relationship with the log odds of being less happy.

CONCLUSION

Because of its flexibility in handling ordinal as well as qualitative response variables, logistic regression is a particularly useful technique. In this article I have touched on what I feel are its most salient features. Space limitations have precluded the discussion of many other important topics, such as logistic regression diagnostics, or the use of logistic regression in event history analysis. The interested reader should consult the references below for more in-depth coverage of these topics.

REFERENCES

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

Aldrich, J. H., & Nelson, F. D. (1984). *Linear probability, logit, and probit models.* Thousand Oaks, CA: Sage.

Begg, C. B., & Gray, R. (1984). Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrica, 71,* 11-18.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

DeMaris, A. (1990). Interpreting logistic regression results: A critical commentary. *Journal of Marriage and the Family, 52,* 271-277.

DeMaris, A. (1991). A framework for the interpretation of first-order interaction in logit modeling. *Psychological Bulletin, 110,* 557-570.

DeMaris, A. (1992). *Logit modeling: Practical applications.* Thousand Oaks, CA: Sage.

DeMaris, A. (1993a). Odds versus probabilities in logit equations: A reply to Roncek. *Social Forces, 71,* 1057-1065.

DeMaris, A. (1993b). *Sense and nonsense in logit analysis: A comment on Roncek.* Unpublished manuscript.

Glenn, N. D., & Weaver, C. N. (1988). The changing relationship of marital status to reported happiness. *Journal of Marriage and the Family, 50,* 317-324.

Hagle, T. M., & Mitchell, G. E. (1992). Goodness-of-fit measures for probit and logit. *American Journal of Political Science, 36,* 762-784.

Hanushek, E. A., & Jackson, J. E. (1977). *Statistical methods for social scientists.* New York: Academic Press.

Holland, B. S., & Copenhaver, M. D. (1988). Improved Bonferroni-type multiple-testing procedures. *Psychological Bulletin, 104,* 145-149.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression.* New York: Wiley.

Lee, G. R., Seccombe, K., & Shehan, C. L. (1991). Marital status and personal happiness: An analysis of trend data. *Journal of Marriage and the Family, 53,* 839-844.

Maddala, G. S. (1983). *Limited dependent and qualitative variables in econometrics.* Cambridge, England: Cambridge University Press.

Mastekaasa, A. (1992). Marriage and psychological well-being: Some evidence on selection into marriage. *Journal of Marriage and the Family, 54,* 901-911.

McClendon, M. J. (1994). *Multiple regression and causal analysis.* Itasca, IL: Peacock.

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal dependent variables. *Journal of Mathematical Sociology, 4,* 103-120.

Nagler, J. (1994). Scobit: An alternative estimator to logit and probit. *American Journal of Political Science, 38,* 230-255.

Appendix

Coding Instructions for Creating the Data File Used in the Analyses

Variable	Value Declared Missing	Replaced With	Comment
HAPPY	9	Left missing	Reverse coded for ordered logit
SEX	None	—	Dummy coded with 1 = male
MARITAL	9	1	Four dummies, married is left out
AGE	99	46.04	
EDUC	98	13.05	
RINCOM91	22,98,99	12.33	
RACE	None	—	Two dummies, White is left out
HEALTH	8,9	2	Reverse coded for analysis
TRAUMA1	9	0	Dummied; 1 = 1 through 3

Note: Data are from the 1993 General Social Survey.