SAGE UNIVERSITY PAPERS

Series: Quantitative Applications in the Social Sciences

Series/Number 07-106

APPLIED LOGISTIC REGRESSION ANALYSIS
Second Edition

SCOTT MENARD
Institute of Behavioral Science
University of Colorado

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

5. POLYTOMOUS LOGISTIC REGRESSION AND ALTERNATIVES TO LOGISTIC REGRESSION

Logistic regression analysis may be extended beyond the analysis of dichotomous variables to the analysis of categorical (nominal or ordinal) dependent variables with more than two categories. In the literature on logistic regression, the resulting models have been called polytomous, polychotomous, or multinomial logistic regression models. Here, the terms dichotomous and polytomous will be used to refer to logistic regression models, and the terms binomial and multinomial will be used to refer to logit models from which polytomous logistic regression models may be derived. For polytomous dependent variables, the logistic regression model may be calculated as a special case of the multinomial logit model (Agresti, 1990; Aldrich & Nelson, 1984; DeMaris, 1992; Knoke & Burke, 1980).

Mathematically, the extension of the dichotomous logistic regression model to polytomous dependent variables is straightforward. One value (typically the first or last) of the dependent variable is designated as the reference category, $Y = h_0$, and the probability of membership in other categories is compared to the probability of membership in the reference category. For nominal variables, this may be a direct comparison, like the indicator contrasts for independent

variables in the logistic regression model for dichotomous variables. For an ordinal variable, contrasts may be made with successive categories, in a manner similar to repeated or Helmert contrasts for independent variables in dichotomous logistic regression models. For dependent variables with some number of categories M, this requires the calculation of $M-1$ equations, one for each category relative to the reference category, to describe the relationship between the dependent variable and the independent variables. For each category of the dependent variable except the reference category, we may write the equation

$$g_h(X_1, X_2, \ldots, X_k) = e^{(a_h + b_{h1}X_1 + b_{h2}X_2 + \cdots + b_{hk}X_k)}, \quad h = 1, 2, \ldots, M-1, \qquad [5.1]$$

where the subscript k refers, as usual, to specific independent variables X and the subscript h refers to specific values of the dependent variable Y. For the reference category, $g_0(X_1, X_2, \ldots, X_k) = 1$. The probability that Y is equal to any value h other than the excluded value h_0 is

$$P(Y = h|X_1, X_2, \ldots, X_k) = \frac{e^{(a_h + b_{h1}X_1 + b_{h2}X_2 + \cdots + b_{hk}X_k)}}{1 + \sum_{h=1}^{M-1} e^{(a_h + b_{h1}X_1 + b_{h2}X_2 + \cdots + b_{hk}X_k)}}, \quad h = 1, 2, \ldots, M-1, \qquad [5.2]$$

and for the excluded category $h_0 = M$ or 0,

$$P(Y = h_0|X_1, X_2, \ldots, X_k) = \frac{1}{1 + \sum_{h=1}^{M-1} e^{(a_h + b_{h1}X_1 + b_{h2}X_2 + \cdots + b_{hk}X_k)}} \quad h = 1, 2, \ldots, M-1. \qquad [5.3]$$

Note that when $M = 2$, we have the logistic regression model for the dichotomous dependent variable, the reference category is the first category, $h_0 = 0$, and we have a total of $M - 1 = 1$ equations to describe the relationship. Logistic regression models for polytomous nominal dependent variables can be calculated in SAS using

CATMOD and in SPSS prior to version 10 using LOGLINEAR, both general log-linear analysis routines in which the calculation of polytomous logistic regression models is rather cumbersome. In SPSS as of version 10, however, NOMREG provides a more user-friendly approach to logistic regression models for nominal dependent variables. SAS LOGISTIC and SPSS PLUM provide similarly user-friendly routines for ordered polytomous dependent variables. Although the focus of this monograph is on SAS and SPSS, it is also worth noting that STATA (1999) provides a broad range of routines for logistic regression, including MLOGIT for nominal dependent variables and OLOGIT for ordinal dependent variables.

To illustrate the use of polytomous logistic regression, the dependent variable from previous examples, prevalence of marijuana use, is replaced by drug user type. Drug user type has four categories.

1. Nonusers report that they have not used alcohol, marijuana, heroin, cocaine, amphetamines, barbiturates, or hallucinogens in the past year.
2. Alcohol users report having used alcohol, but no illicit drugs, in the past year.
3. Marijuana users report having used marijuana (and, except in one case, using alcohol as well).
4. Polydrug users report illicit use of one or more of the "hard" drugs (heroin, cocaine, amphetamines, barbiturates, hallucinogens). Polydrug users also report using alcohol and, except in one case (a respondent who reported a single incident of hard drug use), marijuana as well.

The four categories can reasonably be regarded as being ordered from least serious to most serious drugs, in terms of legal consequences. Alternatively, with respect to the nonlegal consequences of the drugs, the scale could arguably be regarded as nominal. Both ordinal and nominal models of this variable will be considered. One additional change is made from previous models. Because the dependent variable has four categories and because of the small number of cases in the category "other" on the variable ethnicity (ETHN), ethnicity was recoded into two categories, white and nonwhite, for the following analyses. Failure to do this would have resulted in problems with zero cells, and instability in estimates of coefficients and their standard errors.

5.1. Polytomous Nominal Dependent Variables

Figure 5.1 presents the output from SPSS NOMREG[22] with DRGTYPE as a dependent variable, using a contrast for DRGTYPE that compares, in succession, (a) nonusers with alcohol users, (b) nonusers with marijuana users, and (c) nonusers with poly-drug users. The resulting functions, $g_1(X)$, $g_2(X)$, and $g_3(X)$ may be defined as

g_1 = logit (probability of using some alcohol versus nonuse of drugs),

g_2 = logit (probability of using marijuana versus nonuse of drugs),

and

g_3 = logit (probability of using other illicit drugs versus nonuse of drugs).

The equations for g_1, g_2, and g_3 using unstandardized coefficients are, from Figure 5.1,

$$g_1 = .165(\text{EDF5}) - .271(\text{BELIEF4}) + .505(\text{SEX}) + 1.616(\text{WHITE}) + 5.085,$$

$$g_2 = .506(\text{EDF5}) - .285(\text{BELIEF4}) - .920(\text{SEX}) + .357(\text{WHITE}) + 2.503,$$

and

$$g_3 = .633(\text{EDF5}) - .360(\text{BELIEF4}) - 2.224(\text{SEX}) + 2.209(\text{WHITE}) + .768.$$

The calculation of R^2 or η^2 and the standardized logistic regression coefficients is done separately for each logistic function, g_1, g_2, and g_3. (This is similar to the calculation of separate canonical correlation coefficients and standardized discriminant function coefficients for each linear discriminant function in discriminant analysis; see Klecka, 1980.) R^2 for the full model is calculated based on the predicted probabilities and observed classification for all four categories. Prediction tables are included in SPSS NOMREG, and can

nomreg drgtyp5 by sex ethn with edf5 belief4/model=edf5 belief4 sex ethn/print=fit lrt parameter summary classtable/scale=deviance.

Warnings: There are 437 (70.9%) cells (i.e., dependent variable levels by subpopulations) with zero frequencies.

Case Processing Summary

		N
DRGTYP5	1.000 alcohol	87
	2.000 marijuana	50
	3.000 drugs	31
	4.000 nonuser	59
SEX	1 male	110
	2 female	117
ETHN	1 white	175
	2 nonwhite	52
Valid		227
Missing		30
Total		257

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	549.126			
Final	379.778	169.348	12	.000

Likelihood Ratio Tests

Effect	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	379.778	.000	0	.
EDF5	444.950	65.172	3	.000
BELIEF4	396.795	17.017	3	.001
SEX	404.765	24.987	3	.000
ETHN	399.995	20.217	3	.000

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	479.072	447	.142
Deviance	341.094	447	1.000

Pseudo R-Square

Cox and Snell	.526
Nagelkerke	.566
McFadden	.282

R^2 = .303
R_1^2 = .189
R_2^2 = .149
R_3^3 = .337
λ_p = .300 p = .000
τ_p = .399 p = .000

Classification

	Predicted				
Observed	1.000 alcohol	2.000 marijuana	3.000 drugs	4.000 nonuser	Percent Correct
1.000 alcohol	61	7	2	17	70.1%
2.000 marijuana	21	16	8	5	32.0%
3.000 drugs	6	7	18	0	58.1%
4.000 nonuser	20	5	0	34	57.6%
Overall Percentage	47.6%	15.4%	12.3%	24.7%	56.8%

Figure 5.1. Polytomous Nominal Logistic Regression

be constructed for SAS CATMOD by calculating the probability of classification for each value of Y, including the reference category, using Equations 5.2 and 5.3, then classifying each case into the category of Y for which it has the highest probability. The table itself can then be constructed using SAS PROC FREQ. Once the classi-

Parameter Estimates

DRGTYP5		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)		Standardized Logistic Regression Coefficients
								Lower Bound	Upper Bound	Note: These have been edited into the output.
1.000 alcohol	Intercept	5.085	2.463	4.264	1	.039				
	EDF5	.165	.091	3.303	1	.069	1.179	.987	1.409	.209
	BELIEF4	-.271	.070	14.906	1	.000	.763	.665	.875	-.319
	[SEX=1]	.505	.338	2.226	1	.136	1.656	.854	3.214	.075
	[SEX=2]	0(a)	.	.	0	.	.	.	.	
	[ETHN=1]	1.616	.400	16.277	1	.000	5.032	2.295	11.031	.202
	[ETHN=2]	0(a)	.	.	0	.	.	.	.	
2.000 marijuana	Intercept	2.503	2.674	.876	1	.349				
	EDF5	.506	.096	27.544	1	.000	1.659	1.373	2.004	.671
	BELIEF4	-.285	.078	13.319	1	.000	.752	.645	.876	-.350
	[SEX=1]	-.920	.439	4.401	1	.036	.398	.169	.941	-.143
	[SEX=2]	0(a)	.	.	0	.	.	.	.	
	[ETHN=1]	.357	.462	.596	1	.440	1.428	.578	3.531	.047
	[ETHN=2]	0(a)	.	.	0	.	.	.	.	
3.000 drugs	Intercept	.768	3.049	.064	1	.801				
	EDF5	.633	.106	35.976	1	.000	1.883	1.531	2.316	.677
	BELIEF4	-.360	.086	17.515	1	.000	.698	.590	.826	-.357
	[SEX=1]	-2.224	.619	12.893	1	.000	.108	3.211E-02	.364	-.279
	[SEX=2]	0(a)	.	.	0	.	.	.	.	
	[ETHN=1]	2.209	.841	6.901	1	.009	9.104	1.752	47.302	.233
	[ETHN=2]	0(a)	.	.	0	.	.	.	.	

a This parameter is set to zero because it is redundant.

Figure 5.1. (Continued)

fication table has been constructed, indices of predictive efficiency can be calculated as they have been for the Classification table in Figure 5.1, using the procedures described in Chapter 2. It is for polytomous models with nominal dependent variables that the differences between λ_p and τ_p, as opposed to other proposed indices of predictive efficiency, become most evident.

In Figure 5.1, the model works fairly well, as indicated by the statistically significant model χ^2 and the McFadden R^2_L of .28. The explained variance in logit(Y) varies by the category of the dependent variable and is highest for g_3 (polydrug use) and lowest for g_2 (marijuana use). In the overall model, as indicated by the Likelihood Ratio Tests table, all four of the predictors are statistically significant. As indicated at the top of Figure 5.1 in the SPSS NOMREG statement, the dispersion has been corrected using the deviance χ^2 (/scale = deviance). This is because the deviance χ^2 appears to be somewhat lower than the degrees of freedom ($\chi^2 = 341$, df = 447, $\chi^2/\text{df} = .76$), indicating underdispersion. The adjustment for dispersion will affect the statistical significance of the Wald coefficients. For alcohol use, the standardized coefficients (not part of the SPSS output, but like λ_p, τ_p, and R^2 added to the output) indicate that the best predictor is belief that it is wrong to violate the law, followed by ethnicity (white respondents are more likely to use alcohol than nonwhites). Exposure to delinquent friends is marginally significant according to the Wald statistic ($p = .069$), and gender is not statistically significant. For both marijuana and polydrug use, the best predictor is exposure to delinquent friends, followed by belief, then gender. Ethnicity is not a statistically significant predictor for marijuana use, but white respondents are more likely than nonwhite respondents to be polydrug users. Based on the Classification table in Figure 5.1, the indices of predictive efficiency $\lambda_p = .300$ and $\tau_p = .399$ are both statistically significant and moderately strong.

5.2. Polytomous or Multinomial Ordinal Dependent Variables

When the dependent variable is measured on an ordinal scale, many possibilities for analysis exist, including, but by no means limited to, logistic regression analysis. For a more detailed discussion, see Agresti (1990, pp. 318–332), Long (1997, pp. 114–147), or Clogg and Shihadeh (1994). Briefly, the options available include

1. Ignoring the ordering of the categories of the dependent variable and treating it as nominal
2. Treating the variable as though it were measured on a true ordinal scale
3. Treating the variable as though it were measured on an ordinal scale, but the ordinal scale represented crude measurement of an underlying interval/ratio scale
4. Treating the variable as though it were measured on an interval scale.

One possibility consistent with the first option is the use of a multinomial logit or logistic regression model for a nominal categorical dependent variable, as in Figure 5.1. Also possible under option 1 would be the use of discriminant analysis (Klecka, 1980). An example of the second option is the use of a *cumulative logit* model, in which the transformation of the dependent variable incorporates not only each category compared to a reference category, but also a comparison of each category with *all* of the categories with higher (or lower) numeric codes than the present category. The third option, assuming an underlying interval scale, could be implemented in LISREL by using weighted least squares (WLS) analysis of polychoric correlations (Jöreskog & Sörbom, 1988).[23] The fourth option might be implemented by using OLS regression with an ordinal dependent variable.

Selecting one of the options is a matter requiring careful judgment. The fourth option effectively assumes that the data are measured more precisely than they really are, but for ordinal variables with a large number of categories, it may be reasonable. The use of WLS with polychoric correlations appears to be a better option; it can be used with both large and small numbers of categories, and for most ordinal variables. The assumption of imprecise measurement of a quantity that is really continuous (political conservatism, seriousness of drug use) is inherently plausible. Both of these options allow predicted values that lie outside the range of observed values, but under the assumption of imprecise measurement, this may be reasonable.

Mechanical application of options available in existing software packages is *not* recommended. For example, SAS PROC LOGISTIC and SPSS PLUM can calculate polytomous logistic regression models for ordinal dependent variables, but both use a *cumulative logit model* for the dependent variable. This model assumes that the coefficient for each independent variable is invariant across the three equations, that is, $b_{\text{EDF5},1} = b_{\text{EDF5},2} = b_{\text{EDF5},3}$, $b_{\text{SEX},1} = b_{\text{SEX},2} = b_{\text{SEX},3}$, etc. (*parallel slopes*), where the variable in the subscript is the variable to which the coefficient refers, and the number in the subscript is the equation (1, 2, or 3) in which the coefficient appears. For the parallel slopes model, only the intercept is different for the three equations; otherwise, the effects of the independent variables are assumed to be constant across group comparisons. *It is important to emphasize that although this model is easily calculated using SAS PROC LOGISTIC or SPSS PLUM, it may not be the most appropriate model for the relationship between the dependent variable and the predictors.*

Figure 5.2 summarizes the results of analyzing drug user type, DRGTYPE, as an ordinal variable in SAS PROC LOGISTIC. SAS provides a test of the assumption that the slopes are equal, the Score test. For Figure 5.2, the Score test of the null hypothesis that the slopes are equal is 32.066 with 8 degrees of freedom, statistically significant at the .0001 level. Because the Score test is statistically significant, the parallel slopes assumption is rejected, indicating that a model that does not assume parallel slopes would be more appropriate. The reasons for the rejection of the equal slopes model are evident from Figure 5.1: the variation in both the strength and statistical significance of the effects of EDF5 (not statistically significant for alcohol users as opposed to nonusers), SEX (not statistically significant for alcohol users as opposed to nonusers; stronger for polydrug users than for marijuana users as opposed to nonusers), and ETHN (not statistically significant for marijuana users as opposed to nonusers). The pattern of the differences in the coefficients in Figure 5.1 (especially the down-and-up pattern of the coefficients for ethnicity) suggests that treating DRGTYPE as a categorical nominal variable may be the best option.

SPSS PLUM provides much the same information as SAS LOGISTIC, except that SPSS PLUM excludes the information at the bottom of Figure 5.2 (Association of Predicted Probabilities and Observed Responses) and (as in SPSS NOMREG) includes the Pearson and deviance goodness-of-fit χ^2 statistics and the McFadden R^2_{L}, the latter of which (along with R^2 for the overall model and for each of the separate functions) has been edited into the SAS output in Figure 5.2. SPSS PLUM also offers alternatives to the logit distribution for dependent variables that are normally distributed, positively or negatively skewed, or have many extreme values. In both SPSS PLUM and SAS LOGISTIC, it is possible to save predicted values, and to use the predicted and observed values to produce contingency tables (in SAS PROC FREQ or SPSS CROSSTABS) to analyze the accuracy of classification. Doing so for Figure 5.2 would result in $\lambda_{\text{p}} = .229$ ($p = .000$) and $\tau_{\text{p}} = .208$ ($p = .000$), both smaller than in Figure 5.1, further suggesting that the dependent variable may better be treated as nominal rather than ordinal. For an ordinal variable in general, however, the statistics at the bottom of Figure 5.2, particularly the familiar ordinal measures of association Gamma and Tau-a,

```
data;
  infile 'saslr16a.dat' missover linesize=60 firstobs=1 obs=257;
  input ID F66 SEX 8 ETHN 10 USR5 12 PDRUGS5 14-15 PMRJ5 17-18 PALC5 20-21
        DRGTYP5 23-24 EDP5 26-33 BELIEF4 35-42 MEANSCIN 44-51 MEANFAIN 53-60;
  if ethn=1 then white=1; if ethn=2 then white=0; if ethn=3 then white=0;
  if drgtyp5=1 then drgtyp5r=4; if drgtyp5=2 then drgtyp5r=3;
  if drgtyp5=3 then drgtyp5r=2; if drgtyp5=4 then drgtyp5r=1;const5=1;
run;
proc logistic;
  model drgtyp5r=edp5 belief4 sex white;run;

Data Set: WORK.DATA1                    Response Profile
Response Variable: DRGTYP5R
Response Levels: 4                      Ordered
Number of Observations: 227               Value  DRGTYP5R     Count
Link Function: Logit
                                            1         1        31
                                            2         2        50
                                            3         3        87
                                            4         4        59

WARNING: 30 observation(s) were deleted due to missing values for the response or explanatory variables.

                    Score Test for the Proportional Odds Assumption
                      Chi-Square = 32.0660 with 8 DF (p=0.0001)

                Model Fitting Information and Testing Global Null Hypothesis BETA=0

                          Intercept     Intercept
          Criterion          Only     and Covariates   Chi-Square for Covariates      R^2_1 = .004
                                                                                    R^2_2 = .089
          AIC              606.600        485.626                .                  R^2_3 = .151
          SC               616.875        509.600                .
          -2 LOG L         600.600        471.626      128.975 with 4 DF (p=0.0001)  R^2_g = .264
          Score                .              .        97.395 with 4 DF (p=0.0001)  R_L^2 = .215

                  RSquare = .433                  Adjusted RSquare = .467

                        Analysis of Maximum Likelihood Estimates

                                                                                    Standardized
                    Parameter   Standard      Wald        Pr >    Standardized      Odds     coefficient
Variable    DF      Estimate      Error  Chi-Square  Chi-Square  Estimate b_sas     Ratio   b*=(b)(s_x)/s_y

INTERCP1     1       -1.3616     1.4611      0.8684      0.3514             .         .          .
INTERCP2     1        0.6157     1.4513      0.1800      0.6714             .         .          .
INTERCP3     1        2.9884     1.4655      4.1583      0.0414             .         .          .
EDP5         1        0.2701     0.0424     40.5402      0.0001      0.633781     1.310      .343
BELIEF4      1       -0.1774     0.0426     17.3225      0.0001     -0.386429     0.837     -.209
SEX          1       -0.7905     0.2630      9.0312      0.0027     -0.218288     0.454     -.118
WHITE        1        0.8343     0.3167      6.9391      0.0084      0.193729     2.303      .105

                Association of Predicted Probabilities and Observed Responses

                Concordant = 80.5%          Somers' D = 0.623
                Discordant = 18.2%          Gamma     = 0.631
                Tied       =  1.3%          Tau-a     = 0.449
                (18509 pairs)               c         = 0.811
```

Figure 5.2. SAS Output for Ordinal Logistic Regression

may be even more informative than λ_p or τ_p, because the former two measures, unlike the latter two, incorporate information on the ordering of the categories of the dependent variable.

5.3. Conclusion

The principal concern in using logistic regression analysis with polytomous dependent variables is not how to make the model work, but instead whether the logistic regression model is appropriate at all. For ordinal dependent variables, the problems that motivated the development of the logistic regression model (out-of-range predicted values of the dependent variable, heteroscedasticity) may not be present, and other models may be more appropriate than logistic regression, depending on assumptions about the underlying scale of the dependent variable and the functional form (linear, monotonic, nonmonotonic) of the relationship between the dependent variable and the independent variables. If there is an underlying interval scale, and if the relationships appear to be linear or monotonic, weighted least squares with polychoric correlations may be the best option. For nonmonotonic relationships, and especially when there are relatively few categories of the dependent variable, it may be best to treat the dependent variable as though it were nominal. When the dependent variable is nominal, or is an ordinal variable with few categories and is treated as though it were nominal, an alternative worth considering is discriminant analysis (Klecka, 1980). Another alternative, separate logistic regressions (Bess & Grey, 1984; Hosmer & Lemeshow, 1989, pp. 230–232), does not appear to produce results sufficiently consistent with the multinomial logit/polytomous logistic regression model to warrant its use. Only if polytomous logistic regression software were unavailable (an increasingly rare phenomenon, with all major software packages now including polytomous logistic regression routines) would this approach have any merit.

The ease of use, flexibility, broad applicability, and current popularity of logistic regression analysis make it particularly susceptible to misuse. Thoughtless and mechanical applications of logistic regression analysis will be no more fruitful than thoughtless and mechanical applications of linear regression or any other technique. It is important to recognize the weaknesses as well as the strengths of the method. Logistic regression is especially appropriate for the analysis of dichotomous and unordered nominal polytomous dependent variables. For ordinal polytomous dependent variables, it may be possible to use polytomous logistic regression analysis, but other models, including linear regression and weighted least squares with polychoric correlations, also deserve serious consideration. Polytomous ordinal

variables are the dependent variables for which the technical motivation for using logistic regression is weakest and for which alternative methods of analysis are most likely to provide better solutions than logistic regression. Given these qualifications, however, the same ease of use (particularly improvements in logistic regression software, even in the few years since the first edition of this monograph), flexibility, and broad applicability of the logistic regression approach, as mentioned at the beginning of this paragraph, make logistic regression an extremely useful tool for analyzing a broad range of dependent variables for which OLS regression is not appropriate.